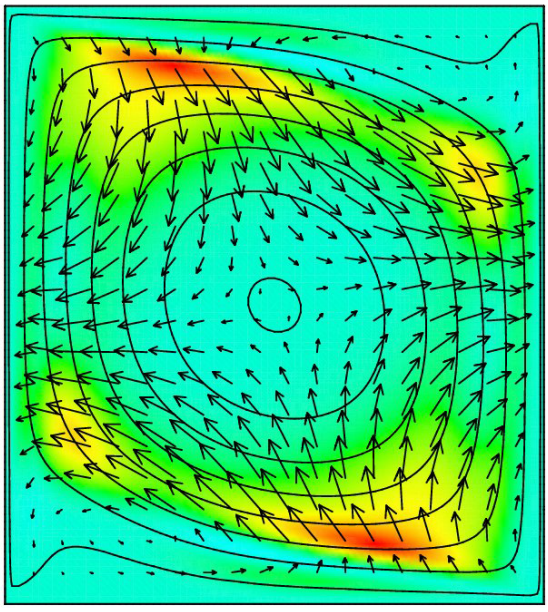
Vorlesung SS21
TU Wien, LVA-Nr. 302.017

NUMERISCHE METHODEN DER STRÖMUNGS- UND WÄRMETECHNIK

Hendrik C. Kuhlmann



Hendrik C. Kuhlmann
Institut für Strömungslehre und Wärmeübertragung
Technische Universität Wien
Getreidemarkt 9
A-1060 Wien
Austria

Das Frontispiz zeigt die Stromlinien in einem Rechteckbehälter mit Breiten-zu-Höhenverhältnis $\Gamma = 0.9$ (Albensoeder and Kuhlmann, 2002). Die Strömung wird auf der rechten und auf der linken Seite durch tangential nach oben bzw. nach unten bewegte Wände angetrieben. Die Strömung befindet sich gerade an der linearen Schwelle zu einer dreidimensionalen Strömung. Die dreidimensionale kritische Mode (Pfeile) und der Energietransfer (Farbe) sind in einem Schnitt dargestellt.

Vorbemerkungen

Ziel des Kurses ist die Vermittlung von Grundlagenwissen über numerische Methoden, die in der Strömungsmechanik zum Einsatz kommen. Dafür gibt es einige hervorragende Bücher. Diese Aufzeichnungen können kein Ersatz für diese Werke sein, in denen umfangreiche weitergehende Informationen zu finden sind. Dieses Skriptum enthält lediglich meine persönlichen Aufzeichnungen zu einer eingeschränkten Auswahl von Themen, wie man sie in einer 2-stündigen Vorlesung behandeln kann, deren Anfänge auf das Sommersemester 2004 zurückgehen. Es hält sich (auch in der Notation) recht eng an die bewährten Bücher von Fletcher (1991*a*) und Ferziger and Perić (2002). Einige Abschnitte wurden jedoch weggelassen, andere vertieft. Als Literatur sind vor allem die beiden o.g. Bücher zu empfehlen. Bei Patankar (1980) steht die Wärmeübertragung im Vordergrund. Weitere Lehrbücher finden sich im Literaturverzeichnis.

Zur Illustration der Techniken werden vor allem einfache partielle Differentialgleichungen wie die Wärmeleitungsgleichung herangezogen. Die Lösung der Navier-Stokes-Gleichung, gekoppelt mit der Wärmetransportgleichung, wird erst im zweiten Teil der Vorlesung (Numerische Methoden der Strömungsmechanik, LVA-Nr. 302.042) behandelt.

H. C. K.
im März 2022

Inhaltsverzeichnis

1. Partielle Differentialgleichungen

Die Strömungsmechanik befaßt sich mit der Dynamik beliebig stark deformierbarer Medien (Fluide). Sie spielt in den Natur- und Ingenieurwissenschaften eine bedeutende Rolle. Die wichtigste Differentialgleichung zur Beschreibung der Dynamik von Fluiden ist die *Navier-Stokes-Gleichung*. Mit ihr können atmosphärische Strömungen, die Aerodynamik von Flugzeugen, der Transport von fluiden Stoffen in verfahrenstechnischen Anlagen und sogar in mikromechanischen Systemen beschrieben werden. Die Navier-Stokes-Gleichung ist *quadratisch nichtlinear* in dem Geschwindigkeitsfeld $\boldsymbol{u}(\boldsymbol{x}, t)$. Deshalb kann die Navier-Stokes-Gleichung nur in Ausnahmefällen analytisch gelöst werden. Aus diesem Grund nimmt die numerische Lösung der Navier-Stokes-Gleichung eine wichtige Stellung ein, insbesondere für Strömungen in komplexen Geometrien. Ähnliches gilt für Strömungen bei hohen Reynoldszahlen, bei denen die Strömung turbulent ist. Dann variiert die Lösung der Navier-Stokes-Gleichung in sehr komplizierter Weise und auf vielen Skalen in Raum und Zeit.

Durch den Einsatz numerischer Methoden konnten in den letzten Jahren immer komplexere Strömungen berechnet und analysiert werden. Eindrucksvolle Beispiele sind die Berechnung der Strömung um ganze Flugzeuge und Raumtransporter (Abb. 1.1a) oder schneller Vorgänge samt chemischer Reaktionen bei der Kraftstoffverbrennung in Motorbrennkammern (Abb. 1.1b). Auch die Wettervorhersage ist durch effiziente numerische Berechnungen der atmosphärischen Strömungen immer weiter verbessert worden, genauso wie die lokale Windvorhersage (Abb. 1.2).

Heutzutage ist die numerische Strömungsmechanik (***C**omputational **F**luid **D**ynamics*, CFD)[1] bei der Auslegung technischer Anlagen, der Vorhersage des Transports von Schadstoffen in der Atmosphäre und der Beantwortung fundamentaler wissenschaftlicher Fragestellungen, um nur einige Beispiele zu nennen, nicht mehr wegzudenken. Trotzdem bedeutet CFD nicht eine Lösung im Handumdrehen. Die Einarbeitung in ein leistungsfähiges CFD-Software-Paket erfordert eine gewisse Zeit. Unabhängig davon, ob man eigene Berechnungsprogramme schreiben oder fertige Software verwenden will, ist es sinnvoll, etwas über die Grundlagen der numerischen Methoden für partielle Differentialgleichungen zu lernen.

1.1. Gleichungen der Strömungsmechanik

Auch wenn wir im Rahmen dieser einführenden zweistündigen Vorlesung nur sehr einfache Gleichungen betrachten können, soll kurz die Struktur der Transportglei-

[1] CFD ist nicht mit *Colorful Fluid Dynamics* zu verwechseln.

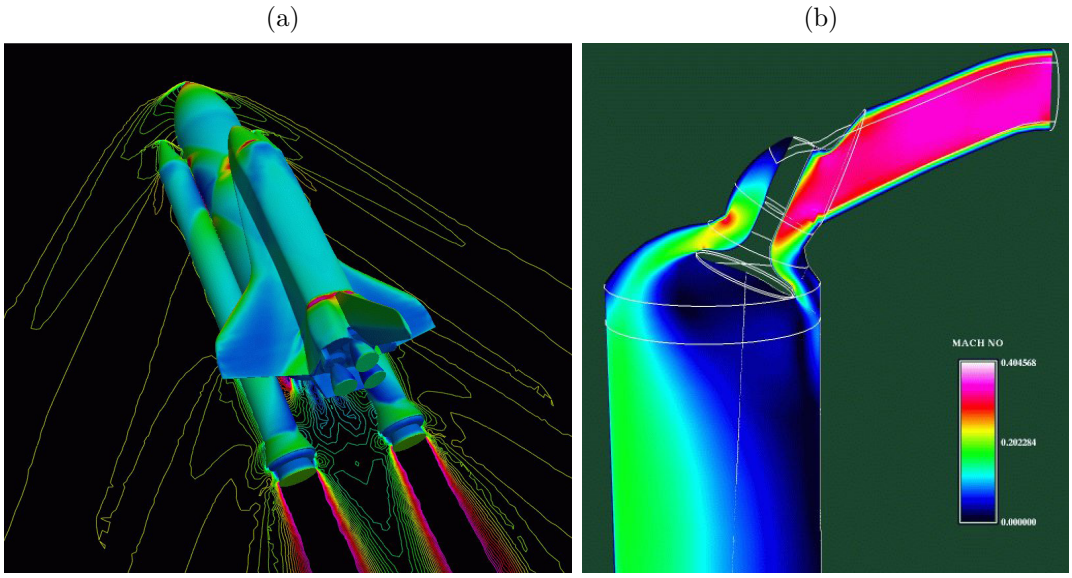


Abbildung 1.1.: Beispiele für Anwendungen numerischer Methoden in der Strömungsmechanik. (a) Druckverteilung auf der Körperoberfläche des Space Shuttle und Konturlinien der Machzahl M $= u/c$ in der Strömung. (b) Machzahl bei der Einströmung in eine Motorbrennkammer (Engineering Sciences Inc.).

chungen rekapituliert werden, mit denen wir es in der Strömungsmechanik zu tun haben. Lösungstechniken für die Navier-Stokes-Gleichungen werden erst in einer weiterführenden Veranstaltung behandelt.

1.1.1. Allgemeine Form der Transportgleichungen

Die allgemeine Form der Transportgleichungen ergibt sich aus dem *Reynoldsschen Transport-Theorem* (siehe z.B. Kuhlmann, 2007). Es führt die zeitliche Änderung der Dichte ϵ einer physikalischen Größe in einem Volumen, das sich mit der Strömung mitbewegt (*substantielles Volumen*), auf Änderungsraten in einem laborfesten Volumen zurück. In differentieller Form erhält man so die *Erhaltungsgleichungen* im Laborsystem in der Form

$$\frac{\partial \epsilon}{\partial t} + \nabla \cdot (\epsilon \boldsymbol{u}) = q. \tag{1.1}$$

Hierbei kann ϵ Dichte der Masse, des Impulses (vektorielle Größe), der Energie, etc. sein. Die zu ϵ gehörige Stromdichte ist $\epsilon \boldsymbol{u}$. Zum Beispiel ist $\rho \boldsymbol{u}$ die Massenstromdichte. Ihr Betrag gibt an, wieviel Masse pro Zeit und Fläche durch eine Fläche senkrecht zu $\boldsymbol{u}$ tritt. Alle Größen hängen i.a. vom Ort $\boldsymbol{x}$ und der Zeit t ab.

Die Dichte ϵ an einem festen Punkt $\boldsymbol{x}$ kann sich ändern, indem die Größe ϵ durch die Strömung $\boldsymbol{u}$ an den Punkt $\boldsymbol{x}$ heran- oder von ihm wegtransportiert wird. Diese Änderung durch Strömung wird durch die negative Divergenz der Stromdichte

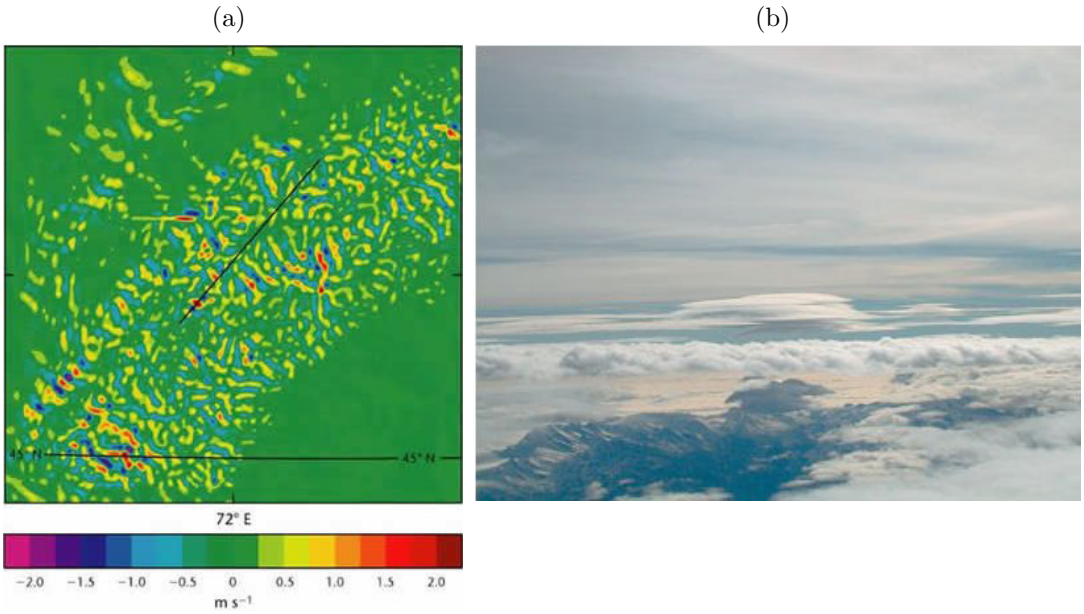


Abbildung 1.2.: (a) Simulation der vertikalen Windgeschwindigkeit in einem Gebiet von 284 km^2 auf einem Höhenniveau von 5 780 m. Die horizontale Auflösung beträgt 1 km. Aufwärtswinde entsprechen positiven Zahlenwerten. Das Muster kommt durch Schwerewellen zustande, die durch die verschiedenen Bergzüge angeregt werden und komplizierte Muster ergeben können. (b) Photographie des betreffenden Gebiets (www.metoffice.com).

$-\nabla \cdot (\epsilon \boldsymbol{u})$ beschrieben. Alle anderen Prozesse, die eine Änderung von ϵ bewirken, sind in der Quelldichte q zusammengefaßt. Zu diesen Prozessen zählt insbesondere der diffusive Transport, der noch zu spezifizieren wäre (für den diffusiven Transport von Impuls und Temperatur siehe (1.10) bzw. (1.13)).

1.1.2. Kontinuitätsgleichung

Wenn wir in (1.1) für ϵ die Massendichte ρ einsetzen und die Massenerhaltung ($q = 0$) berücksichtigen, erhalten wir die *Kontinuitätsgleichung*

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{u}) = 0. \tag{1.2}$$

1.1.3. Impulsbilanz für reibungsfreie Fluide: Eulergleichung

In ähnlicher Weise ergibt sich die Euler-Gleichung für die Impulsänderung, wenn wir in (1.1) für ϵ die Impulsdichte $\rho \boldsymbol{u}$ einsetzen. Hierbei ist jedoch zu beachten, daß die Quelldichte des Impulses nicht verschwindet. Denn wenn sich ein Fluid bewegt, werden auch Druckvariationen erzeugt. Der negative Druckgradient $-\nabla p$ stellt dabei eine Kraft pro Volumen dar und muß als Quellterm für die Impulsdichte

berücksichtigt werden. So erhält man

$$\frac{\partial(\rho \boldsymbol{u})}{\partial t} + \nabla \cdot (\rho \boldsymbol{u}\boldsymbol{u}) = -\nabla p. \tag{1.3}$$

Der Tensor $\rho\boldsymbol{u}\boldsymbol{u}$ stellt einen Beitrag zur *Impulsstromdichte* dar. Die vollständige Impulsstromdichte lautet aber $\rho\boldsymbol{u}\boldsymbol{u} + p\mathsf{I}$,[2] wobei I der Einheitstensor ist. Unter Verwendung der Kontinuitätsgleichung (1.2) kann man die *Euler-Gleichung* auch in der üblichen Form

$$\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{u} = -\frac{1}{\rho}\nabla p \tag{1.4}$$

schreiben. Zusammen mit entsprechenden Anfangs- und Randbedingungen beschreiben die Euler- und die Kontinuitätsgleichung die Dynamik reibungsfreier kompressibler Strömungen. Ob man die Strömung eines realen Fluids durch diese Gleichungen (reibungsfrei) gut approximieren kann, hängt von den Stoffwerten und den Strömungsbedingungen ab (Grenzschichtproblematik).

Für inkompressible Strömungen ist $\rho = \text{const.}$ und die *Kontinuitätsgleichung* vereinfacht sich zu

$$\nabla \cdot \boldsymbol{u} = 0. \tag{1.5}$$

In diesem Fall kann man den Druck durch Bilden der Rotation von (1.4) vollständig eliminieren und erhält die *Helmholtz-Gleichung* für die Vortizität $\boldsymbol{\omega} = \nabla \times \boldsymbol{u}$ eines inkompressiblen reibungsfreien Fluids[3]

$$\frac{\partial \boldsymbol{\omega}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{\omega} = \boldsymbol{\omega} \cdot \nabla \boldsymbol{u}. \tag{1.6}$$

Für den Fall, daß die Vortizität verschwindet ($\boldsymbol{\omega} = 0$), ist die Helmholtz-Gleichung trivial erfüllt. Es bleibt dann nur die Kontinuitätsgleichung zu lösen. Wegen $\boldsymbol{\omega} = \nabla \times \boldsymbol{u} = 0$ kann man die Geschwindigkeit dann aber als Gradienten eines skalaren Potentials ϕ darstellen, $\boldsymbol{u} = \nabla\phi$, was zusammen mit der Kontinuitätsgleichung auf die *Potentialgleichung*

$$\nabla^2 \phi = 0 \tag{1.7}$$

führt.

[2] Auch der Druck p hat die Dimension einer Impulsstromdichte (Impuls pro Fläche und Zeit). Es ist $\nabla \cdot (p\mathsf{I}) = \nabla p$.

[3] Es ist

$$\nabla \times (\boldsymbol{u} \cdot \nabla \boldsymbol{u}) = \nabla \times \left[\nabla\left(\frac{\boldsymbol{u}^2}{2}\right) - \boldsymbol{u} \times \boldsymbol{\omega}\right] = \boldsymbol{u} \cdot \nabla \boldsymbol{\omega} - \boldsymbol{\omega} \cdot \nabla \boldsymbol{u}.$$

Falls das Fluid barotrop ist, d.h. wenn die Dichte $\rho = \rho(p)$ eine eindeutige Funktion des Drucks ist, kann man auch für ein kompressibles Fluid den Druck eliminieren, was auf die kompressible Helmholtz-Gleichung

$$\frac{\partial \boldsymbol{\omega}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{\omega} = \boldsymbol{\omega} \cdot \nabla \boldsymbol{u} - \boldsymbol{\omega} \nabla \cdot \boldsymbol{u}$$

1.1.4. Impulsbilanz für Newtonsche Fluide

Bei realen Fluiden muß die Viskosität berücksichtigt werden. Dann tritt ein weiterer Quellterm auf, der die Änderung der Impulsdichte durch viskose Effekte beschreibt. Man muß dann auf der rechten Seite von (1.3) noch die Divergenz des viskosen Spannungstensors addieren. Die einfachste Form des *viskosen Spannungstensors* ist

$$\mathsf{T}^{\text{viskos}} = \mu \left[\nabla \boldsymbol{u} + (\nabla \boldsymbol{u})^{\mathrm{T}} - \frac{2}{3}(\nabla \cdot \boldsymbol{u})\mathsf{I} \right] + \zeta (\nabla \cdot \boldsymbol{u})\mathsf{I}. \tag{1.8}$$

Fluide, die diesem Materialgesetz genügen, heißen *Newtonsche Fluide*. Wenn wir die Volumenviskosität ζ vernachlässigen, erhalten wir die *Navier-Stokes-Gleichung* in der vereinfachten Form

$$\frac{\partial (\rho \boldsymbol{u})}{\partial t} + \nabla \cdot (\rho \boldsymbol{u}\boldsymbol{u}) = -\nabla p + \nabla \cdot \left\{ \mu \left[\nabla \boldsymbol{u} + (\nabla \boldsymbol{u})^{\mathrm{T}} - \frac{2}{3}(\nabla \cdot \boldsymbol{u})\mathsf{I} \right] \right\} + \rho \boldsymbol{f}. \tag{1.9}$$

In dieser Gleichung ist μ die *dynamische Viskosität*. In (1.9) wurde auch noch eine Volumenkraft pro Masse $\boldsymbol{f}$ aufgenommen, die zum Beispiel die Schwerebeschleunigung sein kann.[4]

Für inkompressible Fluide ($\rho = \text{const.}$) mit konstanter dynamischer Viskosität vereinfachen sich die Kontinuitätsgleichung und die Navier-Stokes-Gleichung erheblich. Dann erhalten wir aus (1.2) die vereinfachte Kontinuitätsgleichung (1.5), $\nabla \cdot \boldsymbol{u} = 0$, und die Navier-Stokes-Gleichung vereinfacht sich zu

$$\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{u} = -\frac{1}{\rho}\nabla p + \nu \nabla^2 \boldsymbol{u} + \boldsymbol{f}, \tag{1.10}$$

wobei $\nu = \mu/\rho$ die *kinematische Viskosität* ist.[5]

1.1.5. Energieerhaltung

Wenn wir in (1.1) für ϵ die Dichte der Gesamtenergie einsetzen, kann man nach einigen Umformungen die folgende häufig verwendete Näherung für die *Energiegleichung*

führt (Saffman, 1992).

[4] Wenn sich ein Fluid konstanter Dichte ρ in einem homogenen Kraftfeld $\rho \boldsymbol{f} = \text{const.}$ befindet (z.B. in einem homogenen Schwerefeld), kann man den modifizierten Druck $p' = p - \rho \boldsymbol{f} \cdot \boldsymbol{x}$ definieren. Dann erhält man die Navier-Stokes-Gleichung mit p' anstelle von p, aber ohne äußere Kräfte. In diesem Fall beeinflußt das Kraftfeld also nur den Druck und nicht das Geschwindigkeitsfeld.

[5] Zur Berechnung der thermischen Konvektion bei geringen Strömungsgeschwindigkeiten kann das Fluid bezüglich der Strömung als inkompressibel aufgefaßt werden, nicht aber bezüglich seiner thermischen Expansion. Man berücksichtigt die Variation der Dichte dann nur in der Auftriebskraft $\rho \boldsymbol{f}$. Diese Näherung wird Oberbeck-Boussinesq-Approximation genannt (siehe z.B. Landau and Lifschitz, 1991, § 56). In dieser Näherung ist $\boldsymbol{f} = -\beta(T - T_0)\boldsymbol{g}$, wobei $\beta = -\rho^{-1}\, \partial\rho/\partial T|_p$ der thermische Ausdehnungskoeffizient ist, $T - T_0$ die Temperaturabweichung von der mittleren Temperatur T_0, $\boldsymbol{g}$ die Schwerebeschleunigung und der konstante Anteil der

chung erhalten,[6] in der die Temperatur T die abhängige Variable ist,

$$\frac{\partial T}{\partial t} + \boldsymbol{u} \cdot \nabla T = \frac{1}{\rho c_p} \nabla \cdot (\lambda \nabla T)\,. \tag{1.11}$$

Hierin wurde nur der wichtigste Quellterm der thermischen Energie, die thermische Diffusion, berücksichtigt und angenommen, daß $\rho c_p = \text{const.}$ ist. Der Diffusionsterm ergibt sich aus der negativen Divergenz der Wärmestromdichte (*Fouriersches Gesetz*)

$$\boldsymbol{j} = -\lambda \nabla T. \tag{1.12}$$

Die Änderungsrate der thermischen Energiedichte (die Energie pro Volumen ist $\rho c_p T$) ist $\partial(\rho c_p T)/\partial t$.

Wenn auch die Wärmeleitfähigkeit λ konstant ist, ergibt sich so

$$\frac{\partial T}{\partial t} + \boldsymbol{u} \cdot \nabla T = \kappa \nabla^2 T, \tag{1.13}$$

wobei wir die *Temperaturleitfähigkeit* (thermische Diffusivität) als $\kappa = \lambda/\rho c_p$ definiert haben.

Mit formal denselben Gleichungen kann man auch den Transport von zusätzlichen Stoffen (Spezies) geringer Konzentration beschreiben, wenn man T durch die jeweilige Konzentration c_i ersetzt und κ durch die zugehörige Diffusivität D_i.

1.1.6. Zustandsgleichnung

Die Dynamik eines Fluids wird durch die Kontinuitätsgleichung (1.2), durch die Impulsbilanzgleichung (1.9) für viskose bzw. (1.4) für reibungsfreie Fluide und durch die Energiegleichung (1.13) beschrieben, wobei wir gewisse Materialparameter als konstant vorausgesetzt hatten. Die Materialeigenschaften hängen aber im allgemeinen vom thermodynamischen Zustand des Fluids ab. Meist kann man davon ausgehen, daß sich das Fluid in einem *lokalen thermodynamischen Gleichgewicht* befindet. Typischerweise wird man dann z.B. die Wärmeleitfähigkeit $\lambda(\rho, T)$ als Funktion der lokalen Dichte und Temperatur angeben.

Wir haben also 5 skalare Gleichungen für die 6 Unbekannten $\boldsymbol{u}$, T, ρ und p. Daher benötigen wir noch eine weitere Gleichung zur vollständigen Beschreibung des Systems. Diese liefert die thermodynamische Zustandsgleichung der Form $f(\rho, p, T) = 0$, deren konkrete Form vom jeweiligen Fluid abhängt. In vielen Fällen kann man die Gleichung für ein ideales Gas

$$p = \rho R T \tag{1.14}$$

verwenden, wobei $R = 8.3144598\ \text{J/(mol K)}$ die universelle Gaskonstante ist. Für

Scherkraft wie in Fußnote 4 in den Druck einbezogen wurde.

[6] Hierbei wurde insbesondere die Erwärmung durch innere Reibung (Dissipation) und die Kom-

inkompressible Fluide ist $\rho = \text{const.}$ und eine Zustandsgleichung wird nicht benötigt. Vielmehr muß p dann so bestimmt werden, daß die inkompressible Kontinuitätsgleichung (1.5) erfüllt ist.

1.1.7. Entdimensionalisierung

Wenn die Materialeigenschaften als konstant angenommen werden können, macht es häufig Sinn, zu einer *dimensionslosen Formulierung* überzugehen. Je nach Größenordnung der Diffusionskonstanten ν und κ und der antreibenden Kräfte $\boldsymbol{f}$ haben entweder der diffusive oder der konvektive Transport einen dominierenden Einfluß auf das Temperatur- bzw. Geschwindigkeitsfeld. Im ersten Fall dominieren die Terme $\kappa\nabla^2 T$ bzw. $\nu\nabla^2\boldsymbol{u}$, im letzteren die Terme $\boldsymbol{u}\cdot\nabla T$ bzw. $\boldsymbol{u}\cdot\nabla\boldsymbol{u}$. Die entsprechenden Lösungen können daher ganz unterschiedliche Eigenschaften aufweisen, wie zum Beispiel bei Grenzschichtströmungen.

Bei Verwendung dimensionsloser Gleichungen lasen sich die verschiedenen Strömungsbereiche besser abschätzen und klassifizieren. Darüber hinaus kann durch eine Transformation auf dimensionslose Größen meist auch die Anzahl der unabhängigen Parameter reduziert werden. Damit kann ein gegebenes Problem effizienter untersucht werden.

Zur Demonstration sei hier die inkompressible Navier-Stokes-Gleichung entdimensionalisiert. Mit den Skalen L, U, L/U und ρU^2 für Ort, Geschwindigkeit, Zeit und Druck definieren wir die dimensionslosen Größen

$$\boldsymbol{x}' = \frac{\boldsymbol{x}}{L}, \quad \boldsymbol{u}' = \frac{\boldsymbol{u}}{U}, \quad t' = \frac{t}{L/U}, \quad p' = \frac{p}{\rho U^2}. \tag{1.15}$$

Wenn wir diese Ausdrücke in (1.10) einsetzen und den Strich ' wieder weglassen, erhalten wir (ohne äußere Kräfte)

$$\frac{\partial\boldsymbol{u}}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{u} = -\nabla p + \frac{1}{\mathrm{Re}}\nabla^2\boldsymbol{u}. \tag{1.16}$$

In dieser Gleichung treten dann nur noch dimensionslose abhängige und unabhängige Variablen auf sowie die *Reynoldszahl*

$$\mathrm{Re} = \frac{UL}{\nu}. \tag{1.17}$$

Sie ist ein dimensionsloser Parameter (*Ähnlichkeitsparameter*), der eine bequeme Klassifikation der Strömung erlaubt, da die Reynoldszahl ein Maß für die Größe der Trägheitskräfte (U^2/L) im Vergleich zur Größe der Reibungskräfte ($\nu U/L^2$) ist. Beachte, daß die Anzahl der unabhängigen Parameter durch die dimensionslose Beschreibung von den vier dimensionsbehafteten Größen ρ, ν, U und L auf eine dimensionslose Größe Re reduziert wurde.

pressionsleistung vernachlässigt.

Bei der Kombination von Strömung und Wärmetransport tritt außerdem die *Prandtlzahl* $\mathrm{Pr} = \nu/\kappa$ auf, bei kompressiblen Strömungen die *Machzahl* $\mathrm{M} = U/c$ (c: Schallgeschwindigkeit) und bei Konvektionsproblemen im Schwerefeld die *Grashofzahl* $\mathrm{Gr} = g\beta\Delta T L^3/\nu^2$.[7] Je nach Problem können andere/weitere Ähnlichkeitsparameter auftreten.

1.1.8. Typische Form der Gleichungen

Alle oben genannten Gleichungen für reale Fluide haben die Form (1.13) einer *Konvektions-Diffusionsgleichung*

$$\frac{\partial T}{\partial t} + \boldsymbol{u} \cdot \nabla T = \kappa \nabla^2 T. \tag{1.18}$$

Sie spielt daher eine überragende Rolle als Prototyp zum Test numerischer Verfahren. Ist $\boldsymbol{u}$ unabhängig von T, dann ist die Konvektions-Diffusionsgleichung linear in T. Der dem *Konvektionsterm* $\boldsymbol{u} \cdot \nabla T$ entsprechende *Advektionsterm* in der Navier-Stokes-Gleichung lautet $\boldsymbol{u} \cdot \nabla \boldsymbol{u}$. Er ist *nichtlinear*. Diese Nichtlinearität ist letztendlich für die ganze Vielfalt der strömungsmechanischen Phänomene verantwortlich.

Die Struktur der inkompressiblen Navier-Stokes-Gleichung (1.10) legt die folgende eindimensionale Gleichung nahe, in der Nichtlinearität und Diffusion in der einfachsten Form enthalten sind,

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \nu \frac{\partial^2 u}{\partial x^2}. \tag{1.19}$$

Dies ist die sogenannte *Burgers-Gleichung*. Wegen der Abwesenheit des Drucks ist diese eindimensionale Gleichung nicht direkt auf eindimensionale Strömungen übertragbar. Sie dient aber als einfaches mathematisches Modell, für welches man analytische Lösungen finden kann. Diese Lösungen können sehr hilfreich sein, um numerische Verfahren zu bewerten, z.B. hinsichtlich ihrer Genauigkeit.

1.1.9. Randbedingungen

Zur Integration der partiellen Differentialgleichungen benötigen wir noch Randbedingungen. Für die inkompressible Navier-Stokes-Gleichung benötigt man 2 Randbedingungen für jede der kartesischen Geschwindigkeitskomponenten $\boldsymbol{u} = (u, v, w)$, da die 3 Gleichungen für die kartesischen Komponenten jeweils von zweiter Ordnung im Raum sind. Die physikalisch motivierte Randbedingung ist die Haftbedingung $\boldsymbol{u} = \boldsymbol{u}_{\mathrm{Wand}}$ an festen Wänden, wobei $\boldsymbol{u}_{\mathrm{Wand}}$ die Geschwindigkeit der festen Wand ist. Falls es sich um eine freie Phasengrenze handelt (z.B. flüssig/gasförmig),

[7] Hierbei ist ΔT eine charakteristische Temperaturdifferenz und L die charakteristische Längenskala.

muß man die Stetigkeit der 3 Geschwindigkeitskomponenten (*kinematische Randbedingung*) und der 3 Komponenten des Spannungsvektors (*dynamische Randbedingung*) fordern. Der Spannungsvektor ist der Kraftvektor pro Flächeneinheit.[8] Im instationären Fall benötigt man noch eine Anfangsverteilung der Geschwindigkeit $\boldsymbol{u}(\boldsymbol{x}, t = 0)$ im gesamten Raumgebiet. Bei inkompressiblen Fluiden muß der Druck so bestimmt werden, daß die Kontinuitätsgleichung erfüllt ist. Dies bedeutet in der Praxis, daß Druck und Geschwindigkeitsfeld simultan berechnet werden müssen, was nur iterativ möglich ist.

Im Falle einer *reibungsfreien* inkompressiblen Strömung reduziert sich die räumliche Ordnung der Differentialgleichung, und man kann auf jeder festen Berandung nur noch die Komponente der Normalgeschwindigkeit vorgeben: $\boldsymbol{u}_\perp = \boldsymbol{u}_{\perp,\text{Wand}}$. Die tangentiale Geschwindigkeit ist frei und ergibt sich erst aus der Lösung des Problems, die von der Anfangsbedingung $[\boldsymbol{u}(\boldsymbol{x}, t = 0), p(\boldsymbol{x}, t = 0)]$ abhängt. Bei durchströmten Gebieten muß man auch noch geeignete Bedingungen für die Ein- und Ausströmung formulieren.

Die Energiegleichung (1.13) ist von zweiter Ordnung in den Koordinaten x, y und z. Daher wird auf jeder Berandung eine Randbedingung erforderlich. Dies kann eine vorgegebene Temperatur T (*Dirichlet-Randbedingung*), ein vorgegebener Wärmestrom $\sim \partial_\perp T$ (*Neumann-Randbedingung*) oder eine andere physikalisch motivierte Randbedingung sein. Eine additive Kombination aus Dirichlet- und Neumann-Randbedingung bezeichnet man auch als *Robin-Randbedingungen*. Bei instationären Problemen muß man noch eine Anfangsbedingung $T(\boldsymbol{x}, t = 0)$ im ganzen Volumen angeben.

Bei kompressiblen Strömungen sind alle o.a. Gleichungen gekoppelt zu lösen. Je nach Problem und numerischer Methode kann die Angabe von Randbedingungen variieren und ist nicht ganz trivial (siehe z.B. Fletcher, 1991*b*). Auch im inkompressiblen Fall gibt es unterschiedliche Formulierungen der Randbedingungen (Gresho, 1991). Die damit zusammenhängenden Fragen brauchen uns hier aber zunächst nicht weiter zu beschäftigen.

1.2. Klassifizierung partieller Differentialgleichungen

Die Differentialgleichungen der Strömungsmechanik sind partielle Differentialgleichungen (PDEs) in Raum und Zeit. Die den unterschiedlichen Strömungstypen (z.B. reibungsfrei oder viskos) entsprechenden PDEs sowie deren Lösungen können prinzipiell unterschiedliche Eigenschaften besitzen. Daher können je nach Typ der Gleichung unterschiedliche numerische Methoden erforderlich sein. Aus diesem Grund ist eine Klassifizierung von PDEs sinnvoll.

Lineare partielle Differentialgleichungen zweiter Ordnung mit konstanten Koeffi-

[8] Bei kapillaren Grenzflächen sind bei der Normalspannung ggf. noch der Drucksprung durch den Laplace-Druck und bei den Tangentialspannungen noch Kräfte zu berücksichtigen, die nur an der Grenzfläche angreifen (z.B. beim *thermokapillaren Effekt*).

zienten können immer in der Form

$$\left(A\partial_x^2 + B\partial_{xy} + C\partial_y^2 + D\partial_x + E\partial_y + F\right) u = -G \tag{1.20}$$

geschrieben werden. Sie werden in drei Klassen eingeteilt: *elliptisch*, *parabolisch* und *hyperbolisch*. Diese Klassifizierung ergibt sich aus dem charakteristischen Verhalten der Lösungen für jeden Typ. Der Typus der PDE hängt von den Koeffizienten der Ableitungsoperatoren ab. Man kann zeigen, daß nur die Koeffizienten A, B und C vor den 3 höchsten (zweiten) Ableitungen den Typ bestimmen.

1.2.1. Motivation

Um die Klassifikation zu motivieren, betrachten wir die lineare PDE

$$\left(A\partial_x^2 + B\partial_{xy} + C\partial_y^2\right) u(x,y) = 0, \tag{1.21}$$

mit reellen Koeffizienten $A, B, C \in \mathbb{R}$. Die Lösung kann man durch *Fourier-Transformation* erhalten. Danach läßt sich die Lösung als *Superposition* (Integral über k_x und k_y) von *Fourier-Moden*[9]

$$u = \hat{u}\mathrm{e}^{\mathrm{i}k_x x}\mathrm{e}^{\mathrm{i}k_y y} \tag{1.22}$$

schreiben. Wenn wir (1.22) in (1.21) einsetzen, erhalten wir

$$-k_x^2 A - k_x k_y B - k_y^2 C = 0 \qquad \Rightarrow \qquad As^2 + Bs + C = 0, \tag{1.23}$$

wobei wir $s = k_x/k_y$ definiert haben. Die quadratische Gleichung in s ist eine Bedingung für das Verhältnis der beiden Wellenzahlen. Sie können nicht unabhängig voneinander gewählt werden. Vielmehr muß für ihr Verhältnis gelten (Lösung der quadratischen Gleichung)

$$s = -\frac{B}{2A} \pm \frac{1}{2A}\sqrt{B^2 - 4AC}. \tag{1.24}$$

[9] Zur Erinnerung: Die Fourier-Transformation und die -Rücktransformation sind definiert durch

$$\hat{u}(k) = \int_{-\infty}^{\infty} u(x)\mathrm{e}^{-\mathrm{i}kx}\,\mathrm{d}x \qquad \text{und} \qquad u(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{u}(k)\mathrm{e}^{\mathrm{i}kx}\,\mathrm{d}k.$$

Dies kann man durch Probe bestätigen

$$u(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \underbrace{\left(\int_{-\infty}^{\infty} u(x')\mathrm{e}^{-\mathrm{i}kx'}\,\mathrm{d}x'\right)}_{\hat{u}(k)} \mathrm{e}^{\mathrm{i}kx}\,\mathrm{d}k = \frac{1}{2\pi}\int_{-\infty}^{\infty} \underbrace{\left(\int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}k(x-x')}\,\mathrm{d}k\right)}_{2\pi\delta(x-x')} u(x')\,\mathrm{d}x'$$
$$= \int_{-\infty}^{\infty} u(x')\delta(x-x')\,\mathrm{d}x' = u(x).$$

Die Klassifizierung der Lösungen (1.22) ergibt sich aus der Diskriminante

$$\begin{aligned} B^2 - 4AC > 0 \quad &\Rightarrow \quad s_1 \neq s_2 \in \mathbb{R} \quad (\textit{hyperbolisch}), \\ B^2 - 4AC = 0 \quad &\Rightarrow \quad s_1 = s_2 \in \mathbb{R} \quad (\textit{parabolisch}), \\ B^2 - 4AC < 0 \quad &\Rightarrow \quad s_1 \neq s_2 \in \mathbb{C} \quad (\textit{elliptisch}). \end{aligned} \tag{1.25}$$

Im hyperbolischen Fall sind die Wellenzahlen k_x und k_y reell. Die zugehörige Lösung hat dann einen rein oszillatorischen Charakter in beiden Koordinatenrichtungen. Im elliptischen Fall führen die Terme mit den höchsten Ableitungen zu gedämpften bzw. angefachten Lösungen. Anschaulich ist klar, daß die Koeffizienten vor etwaigen Ableitungen ersten Ordnung in (1.21) keinen Einfluß auf dieses prinzipielle Verhalten haben, da die Terme erster Ordnung Dämpfungsgliedern entsprechen.

Die Bezeichnungen *hyperbolisch*, *parabolisch* und *elliptisch* rühren daher, daß (1.24) identisch ist mit der Diskriminante für die Lösungen der allgemeinen Gleichung zweiten Grades,[10] die dann entweder Hyperbeln, Parabeln oder Ellipsen sind. Diese Analogie hat aber keine geometrische Bedeutung für die Lösungen der PDEs.

Wenn die Koeffizienten der Differentialgleichung nicht konstant sind, kann der Typ der Gleichung vom Ort abhängen. Eine Klassifizierung kann man dann durch *Einfrieren der Koeffizienten* vornehmen. Als Beispiel wollen wir die entdimensionalisierte stationäre inkompressible Navier-Stokes- (1.16) und Kontinuitätsgleichung (1.5) in zwei Dimensionen betrachten

$$\partial_x u + \partial_y v = 0, \tag{1.26a}$$

$$u\partial_x u + v\partial_y u + \partial_x p - \frac{1}{\mathrm{Re}}\left(\partial_x^2 + \partial_y^2\right) u = 0, \tag{1.26b}$$

$$u\partial_x v + v\partial_y v + \partial_y p - \frac{1}{\mathrm{Re}}\left(\partial_x^2 + \partial_y^2\right) v = 0. \tag{1.26c}$$

Wenn wir die Koeffizienten (rot) in den konvektiven Gliedern formal als konstant annehmen, dann werden die Gleichungen linear und mit dem Fourier-Moden-Ansatz

$$(u, v, p)^{\mathrm{T}} = (\hat{u}, \hat{v}, \hat{p})^{\mathrm{T}} \mathrm{e}^{\mathrm{i}(k_x x + k_y y)} \tag{1.27}$$

erhalten wir das lineare System für die Amplituden $\hat{u}$, $\hat{v}$ und $\hat{p}$

$$\begin{bmatrix} \mathrm{i}k_x & \mathrm{i}k_y & 0 \\ \mathrm{i}\left(uk_x + vk_y\right) + \dfrac{k_x^2 + k_y^2}{\mathrm{Re}} & 0 & \mathrm{i}k_x \\ 0 & \mathrm{i}\left(uk_x + vk_y\right) + \dfrac{k_x^2 + k_y^2}{\mathrm{Re}} & \mathrm{i}k_y \end{bmatrix} \cdot \begin{pmatrix} \hat{u} \\ \hat{v} \\ \hat{p} \end{pmatrix} = 0. \tag{1.28}$$

[10] Die Lösungen der Gleichung

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$$

sind in der (x, y)-Ebene Hyperbeln, Parabeln oder Ellipsen, je nachdem ob $B^2 - 4AC > 0, = 0$ oder < 0 ist.

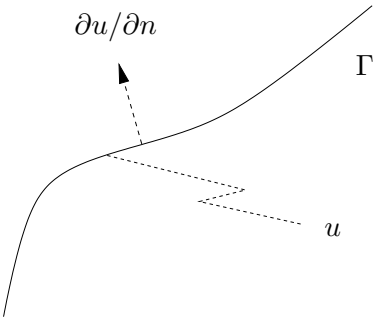


Abbildung 1.3.: Raumkurve, entlang der die Größe u und die senkrechte Ableitung $\partial u/\partial n$ vorgegeben sind.

Die Lösbarkeitsbedingung für dieses homogene lineare System $\det[...] = 0$ führt auf das *charakteristische Polynom*

$$\left(k_x^2 + k_y^2\right)\left[\mathrm{i}\left(uk_x + vk_y\right) + \frac{k_x^2 + k_y^2}{\mathrm{Re}}\right] = 0. \tag{1.29}$$

Es gehen zwar mit $\mathrm{i}\left(uk_x + vk_y\right)$ die ersten Ableitungen ein, aber der Charakter der Lösungen ist durch die höchsten Ableitungen über $\left(k_x^2 + k_y^2\right) = 0$ bestimmt. Die zugehörigen Wurzeln $s = k_x/k_y = \pm\mathrm{i}$ sind rein imaginär, woraus wir folgern, daß die stationären zweidimensionalen inkompressiblen Navier-Stokes-Gleichungen *elliptisch* sind.

1.2.2. Charakteristiken

Die Klassifizierung von PDEs erhält man in der mathematischen Theorie über die sogenannten Charakteristiken (siehe z.B. Fletcher, 1991*a*; Sommerfeld, 1978), die weiter unten erklärt werden. Sie spielen insbesondere für hyperbolische PDEs eine wichtige Rolle. Wir schreiben nun (1.20) in der Form

$$\left(A\partial_x^2 + B\partial_{xy} + C\partial_y^2\right) u = \Phi\left(u, \partial_x u, \partial_y u, x, y\right) \tag{1.30}$$

mit $A, B, C \in \mathbb{R}$ und betrachten nun das folgende Anfangswert- bzw. Randwertproblem:

> Entlang einer Kurve Γ in der (x, y)-Ebene seien u und die Ableitung senkrecht zur Kurve $\partial u/\partial n$ vorgegeben (Abb. 1.3). Mit u ist dann natürlich auch die tangentiale Ableitung $\partial u/\partial s$ entlang Γ bekannt. Neben u sind auf Γ also die beiden ersten partiellen Ableitungen $\partial u/\partial x$ und $\partial u/\partial y$ gegeben. Der Mathematiker fragt sich dann: *Existiert eine Lösung von (1.30), die diesen Randbedingungen genügt?*

Zur positiven Beantwortung dieser Frage muß man zeigen, daß alle höheren Ableitungen von u in der Umgebung von Γ existieren. Denn dann besitzt u eine Potenzreihenentwicklung, und die Lösung kann fortgesetzt werden.

Wenn wir die üblichen Bezeichnungen für die partiellen Ableitungen

$$p = \frac{\partial u}{\partial x}, \quad q = \frac{\partial u}{\partial y}, \quad r = \frac{\partial^2 u}{\partial x^2}, \quad s = \frac{\partial^2 u}{\partial x \partial y}, \quad t = \frac{\partial^2 u}{\partial y^2}, \tag{1.31}$$

verwenden, dann lautet die zu lösende PDE (1.30)

$$Ar + Bs + Ct = \underbrace{\Phi(x, y, u, p, q)}_{\text{auf } \Gamma \text{ bekannt}}. \tag{1.32a}$$

Wir sind an den höheren Ableitungen (2. Ordnung) r, s und t auf Γ interessiert und möchten diese bestimmen. Deshalb brauchen wir zusätzlich zu (1.32a) noch zwei weitere Gleichungen, um diese Ableitungen bestimmen zu können. Diese können wir mit Hilfe der Differentiale $\mathrm{d}p$ und $\mathrm{d}q$ erhalten. Für sie gilt *per definitionem* (und insbesondere auf Γ)

$$\mathrm{d}p := \frac{\partial p}{\partial x}\mathrm{d}x + \frac{\partial p}{\partial y}\mathrm{d}y = r\,\mathrm{d}x + s\,\mathrm{d}y, \tag{1.32b}$$

$$\mathrm{d}q := \frac{\partial q}{\partial x}\mathrm{d}x + \frac{\partial q}{\partial y}\mathrm{d}y = s\,\mathrm{d}x + t\,\mathrm{d}y. \tag{1.32c}$$

Mit (1.32a)–(1.32c) haben wir drei inhomogene lineare Gleichungen zur Bestimmung der drei höheren Ableitungen r, s und t aus den durch die Randbedingung vorgegebenen ersten Ableitungen p und q sowie u selbst. Außerdem hängen diese Gleichungen von der Richtung $(\mathrm{d}x, \mathrm{d}y)^{\mathrm{T}}$ ab, die betrachtet wird. Damit die drei höheren Ableitungen (r, s, t) existieren, muß die Koeffizienten-Determinante des linearen inhomogenen Systems (1.32a)–(1.32c) von Null verschieden sein

$$\Delta = \det \begin{vmatrix} A & B & C \\ \mathrm{d}x & \mathrm{d}y & 0 \\ 0 & \mathrm{d}x & \mathrm{d}y \end{vmatrix} = A\,\mathrm{d}y^2 - B\,\mathrm{d}x\,\mathrm{d}y + C\,\mathrm{d}x^2 \overset{!}{\neq} 0. \tag{1.33}$$

Feste Werte der Differentiale $\mathrm{d}x$ und $\mathrm{d}y$ definieren eine Richtung in der (x, y)-Ebene durch die Steigung $\mathrm{d}y/\mathrm{d}x$. Die Bedingung (1.33) ist offenbar in jedem Punkt (x, y) für fast alle Richtungen erfüllt. Nur in zwei Richtungen ist diese Bedingung nicht erfüllt. Diese beiden Richtungen sind gerade durch $\Delta = 0$ bestimmt. Die Bedingung $\Delta = 0$ liefert die quadratische Gleichung

$$A\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2 - B\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right) + C = 0 \tag{1.34}$$

für die beiden Steigungen $\mathrm{d}y/\mathrm{d}x$ der beiden Richtungen. Entlang der Kurve Γ definieren diese Richtungen zwei (reelle oder konjugiert komplexe) Kurvenscharen, die *Charakteristiken* genannt werden. Ihre Steigungen im Punkt (x, y) sind gegeben durch

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{B}{2A} \pm \frac{1}{2A}\sqrt{B^2 - 4AC} = c_{1,2}. \tag{1.35}$$

Entlang der Charakteristiken, kann man die höheren Ableitungen r, s und t nicht bestimmen. Damit also eine Lösung des obigen Anfangswertproblems in einer Umgebung von Γ existiert, darf Γ nirgendwo ein Bogenelement mit einer Charakteristik gemeinsam haben.[11] Für alle Richtungen mit $\Delta \neq 0$ kann man nun zeigen, daß neben r, s und t auch alle höheren Ableitungen existieren. Für sie gelten exakt dieselben Existenzbedingungen wie (1.33) (siehe z.B. Sommerfeld, 1978). Damit ist die Existenz einer Lösung in der Umgebung von Γ sichergestellt.

Die Charakteristiken besitzen dieselbe Diskriminate wie (1.23), was zu derselben Klassifikation wie schon in (1.25) führt:

$$\begin{aligned} B^2 - 4AC > 0 \quad &\Rightarrow \quad \textit{hyperbolisch}: \\ &\quad \text{2 reelle Scharen von Charakteristiken,} \\ B^2 - 4AC = 0 \quad &\Rightarrow \quad \textit{parabolisch}: \\ &\quad \text{Nur eine Schar von Charakteristiken,} \\ B^2 - 4AC < 0 \quad &\Rightarrow \quad \textit{elliptisch}: \\ &\quad \text{Charakteristiken sind konj. komplex.} \end{aligned} \tag{1.36}$$

Die Charakteristiken definieren die beiden Kurvenscharen[12]

$$y = c_1 x + \psi, \tag{1.37}$$
$$y = c_2 x + \phi, \tag{1.38}$$

wobei ψ und ϕ die Kurven parametrisieren. Seien die Charakteristiken gegeben durch die Funktionen $\psi(x,y) = y - c_1 x = \text{const.}$ und $\phi(x,y) = y - c_2 x = \text{const.}$ Dann kann man die Differentialgleichung (1.30) durch die *Koordinatentransformationen*

$$\xi = \phi(x,y), \qquad \eta = \psi(x,y), \qquad \text{(hyperbolisch)}, \tag{1.39a}$$
$$\xi + \mathrm{i}\eta = \phi(x,y) = \psi(x,y), \qquad \eta = x, \qquad \text{(parabolisch)}, \tag{1.39b}$$
$$\xi + \mathrm{i}\eta = \phi(x,y), \qquad \xi - \mathrm{i}\eta = \psi(x,y), \qquad \text{(elliptisch)}. \tag{1.39c}$$

auf besonders einfache Formen (Normalformen) bringen (Sommerfeld, 1978). Die sogenannten *Normalformen* lauten

$$\frac{\partial^2 u}{\partial \xi \partial \eta} = X\left(u, \frac{\partial u}{\partial \xi}, \frac{\partial u}{\partial \eta}, \xi, \eta\right) \qquad \text{(hyperbolisch)}, \tag{1.40a}$$
$$\frac{\partial^2 u}{\partial \eta^2} = X\left(u, \frac{\partial u}{\partial \xi}, \frac{\partial u}{\partial \eta}, \xi, \eta\right) \qquad \text{(parabolisch)}, \tag{1.40b}$$
$$\frac{\partial^2 u}{\partial \xi^2} + \frac{\partial^2 u}{\partial \eta^2} = X\left(u, \frac{\partial u}{\partial \xi}, \frac{\partial u}{\partial \eta}, \xi, \eta\right) \qquad \text{(elliptisch)}. \tag{1.40c}$$

[11] Denn sonst könnte man noch nicht einmal die höheren Ableitungen auf Γ selbst bestimmen.
[12] Für $A, B, C = \text{const.}$ sind die beiden Kurvenscharen jeweils parallele Geraden.

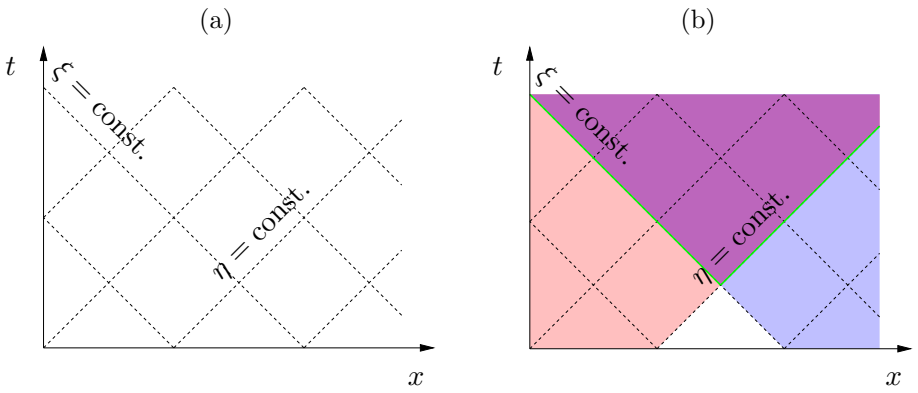


Abbildung 1.4.: (a) Charakteristiken für die Wellengleichung (gestrichelte Linien). (b) Bei Vorgabe der Funktionen f und g entlang der grünen Segmente der Charakterisitiken erhält man die Lösung im violetten Bereich.

Die obige Klassifizierung gilt für PDEs zweiter Ordnung. Man kann aber auch PDEs erster Ordnung in ähnlicher Weise klassifizieren. Im folgenden wird für jeden Typ ein Beispiel gegeben.

1.2.3. Hyperbolische Differentialgleichungen

Eine der einfachsten hyperbolischen PDEs ist die *Wellengleichung* (zum Beispiel für eindimensionale Schallwellen)

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = 0. \tag{1.41}$$

Sie ist linear. Mit dem Fourier-Ansatz $u = \hat{u}\mathrm{e}^{\mathrm{i}(kx-\omega t)}$ erhält man die *Dispersionsrelation*

$$\omega = \pm ck \tag{1.42}$$

zwischen der Kreisfrequenz ω und der Wellenzahl k. Man kann $c = \pm\omega/k$ also mit der *Phasengeschwindigkeit* identifizieren. Die beiden Wurzeln der Dispersionsrelation entsprechen rechts- ($\sim \mathrm{e}^{\mathrm{i}(kx-\omega t)}$) und linkslaufenden Wellen ($\sim \mathrm{e}^{\mathrm{i}(kx+\omega t)}$), wenn man $k > 0 \in \mathbb{R}$ wählt. Unabhängig von der Wellenzahl k besitzen alle Wellen dieselbe Phasengeschwindigkeit $c = \text{const}$. Man sagt dann: die Wellen zeigen *keine Dispersion*.

Mit $A = 1$, $B = 0$ und $C = -c^2$ ist die Diskriminante $B^2 - 4AC = 4c^2 > 0$. Daher ist die Wellengleichung hyperbolisch. Nach (1.35) sind die Charakteristiken durch Geraden-Scharen mit den Steigungen $\mathrm{d}x/\mathrm{d}t = \pm c$ gegeben, also durch $x = ct + \psi$ und $x = -ct + \phi$ (Abb. 1.4a).

Die Gleichungen für die Charakteristiken lauten dann (siehe Abb. 1.4)

$$\phi(x,t) = x + ct = \text{const.} \tag{1.43a}$$
$$\psi(x,t) = x - ct = \text{const.} \tag{1.43b}$$

Mit der Transformation (1.39a), hier $\xi = \phi = x + ct$ und $\eta = \psi = x - ct$ kann man die Wellengleichung (1.41) auf die Normalform (1.40a) bringen[13]

$$\frac{\partial^2 u}{\partial \xi \partial \eta} = 0. \tag{1.44}$$

Die allgemeine Lösung dieser PDE hat die Form[14]

$$\begin{aligned} & u(\xi, \eta) = f(\xi) + g(\eta) \\ \text{bzw.} \qquad & u(x, t) = f(x + ct) + g(x - ct). \end{aligned} \tag{1.45}$$

Wenn man nun die Funktion $f(\xi)$ entlang einer Charakteristik $\eta = \text{const.}$ *und*[15] $g(\eta)$ entlang einer Charakteristik $\xi = \text{const.}$ vorgeben könnte, wäre $u(\xi, \eta)$ mit (1.45) in dem von den Charakteristiken aufgespannten Bereich sofort anzugeben (Abb. 1.4b). Von physikalischem Interesse sind aber die beiden *Anfangsbedingungen* bei $t = 0$ auf einer Länge $\Delta x = L$ der Form (entspricht der Raumkurve Γ aus Abb. 1.3)

$$u|_{t=0} = u_0(x) \qquad \text{und} \qquad \frac{1}{c} \left. \frac{\partial u}{\partial t} \right|_{t=0} = a_0(x). \tag{1.46}$$

Aus diesen Angaben können wir die gesuchten Funktionen f und g konstruieren. Dazu definieren wir

$$F(x) := \frac{1}{2} \left(u_0(x) + \int_0^x a_0(x') \, \mathrm{d}x' \right) \quad \text{und} \quad G(x) := \frac{1}{2} \left(u_0(x) - \int_0^x a_0(x') \, \mathrm{d}x' \right). \tag{1.47}$$

Damit lassen sich die Anfangsbedingungen bei $t = 0$ ausdrücken als

$$u|_{t=0} = F(x) + G(x) \qquad \text{und} \qquad \frac{1}{c} \left. \frac{\partial u}{\partial t} \right|_{t=0} = F'(x) - G'(x). \tag{1.48}$$

---

[13] Mit

$$\frac{\partial}{\partial x} = \underbrace{\frac{\partial \xi}{\partial x}}_{=1} \frac{\partial}{\partial \xi} + \underbrace{\frac{\partial \eta}{\partial x}}_{=1} \frac{\partial}{\partial \eta} = \frac{\partial}{\partial \xi} + \frac{\partial}{\partial \eta} \qquad \text{und} \qquad \frac{\partial}{\partial t} = \underbrace{\frac{\partial \xi}{\partial t}}_{=c} \frac{\partial}{\partial \xi} + \underbrace{\frac{\partial \eta}{\partial t}}_{=-c} \frac{\partial}{\partial \eta} = c \frac{\partial}{\partial \xi} - c \frac{\partial}{\partial \eta}$$

folgt aus (1.41)

$$c^2 \left(\frac{\partial}{\partial \xi} - \frac{\partial}{\partial \eta} \right)^2 u - c^2 \left(\frac{\partial}{\partial \xi} + \frac{\partial}{\partial \eta} \right)^2 u = -4c^2 \frac{\partial}{\partial \xi} \frac{\partial}{\partial \eta} u = 0.$$

[14] Die Integration von (1.44) über ξ bzw. über η ergibt

$$\frac{\partial u}{\partial \eta} = g'(\eta) \qquad \text{und} \qquad \frac{\partial u}{\partial \xi} = f'(\xi).$$

[15] Die Vorgabe entlang einer Charakteristik reicht nicht aus.

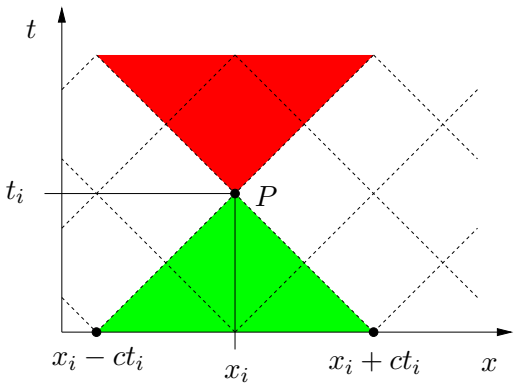


Abbildung 1.5.: Der Wert $u(x_i, t_i)$ im Punkt P hängt nur von u im grünen Bereich ab. Andererseits beeinflußt $u(x_i, t_i)$ im Punkt P nur den roten Bereich in der Zukunft.

Durch einen Vergleich mit (1.45) für $t = 0$ können wir F und G genau mit den Funktionen f und g identifizieren. Somit erhalten wir die Lösung

$$u(x,t) = F(x+ct) + G(x-ct) = \frac{1}{2}\left[u_0(x+ct) + u_0(x-ct) + \int_{x-ct}^{x+ct} a_0(x')\,\mathrm{d}x'\right]. \tag{1.49}$$

An dieser Form sieht man, daß die Feldgröße $u(x_i, t_i)$ am Orte x_i und zur Zeit t_i in eindeutiger Weise durch die Anfangswerte aus dem Bereich $x \in [x_i - ct_i, x_i + ct_i]$ bestimmt ist (Abb. 1.5). Entsprechend kann der Funktionswert $u(x_i, t_i)$ am Punkt x_i nur diejenigen anderen Punkte mit $t > t_i$ beeinflussen, die *innerhalb* der Charakteristiken liegen, die durch den Ausgangspunkt x_i gehen. Die Information breitet sich mit der Phasengeschwindigkeit $\pm c$ aus. Von der anfänglichen Verteilung $u_0(x)$ propagiert die Hälfte nach rechts und die andere Hälfte nach links. Hinzu kommt noch ein Beitrag von der anfänglichen Änderungsrate $a_0(x)$.

In der Verallgemeinerung dieses Resultats erhält man eindeutige Lösungen in Gebieten, die begrenzt sind durch die (im allgemeinen gekrümmten) Charakteristiken[16] und den Linienzug, entlang dem die Rand/Anfangsbedingungen vorgegeben werden (Abb. 1.6a). Auch wenn u und $\partial u/\partial t$ nicht auf derselben Linie vorgegeben sind, sondern auf verschiedenen sich treffenden Linien, ist die Lösung eindeutig. Die wichtigsten Fälle sind in Abb. 1.6b,c gezeigt.

Neben der Wellengleichung sind auch die instationären Eulergleichungen und die Gleichungen für stationäre Überschallströmungen reibungsfreier Fluide hyperbolisch. Die Lösungen hyperbolischer PDEs sind typischerweise ungedämpft (Wellen). Das bedeutet, daß sich Diskontinuitäten in den Anfangsbedingungen im Falle linearer hyperbolischer PDEs entlang den Charakteristiken in das Integrationsgebiet ausbreiten.[17]

[16] Die Charakteristiken sind krummlinig, wenn die Koeffizienten A, B und C der Differentialgleichung nicht konstant sind.

[17] Die *Methode der Charakteristiken* zur Lösung partieller Differentialgleichungen *erster* Ordnung durch Überführung in ein System gewöhnlicher Differentialgleichungen erster Ordnung wird hier nicht behandelt.

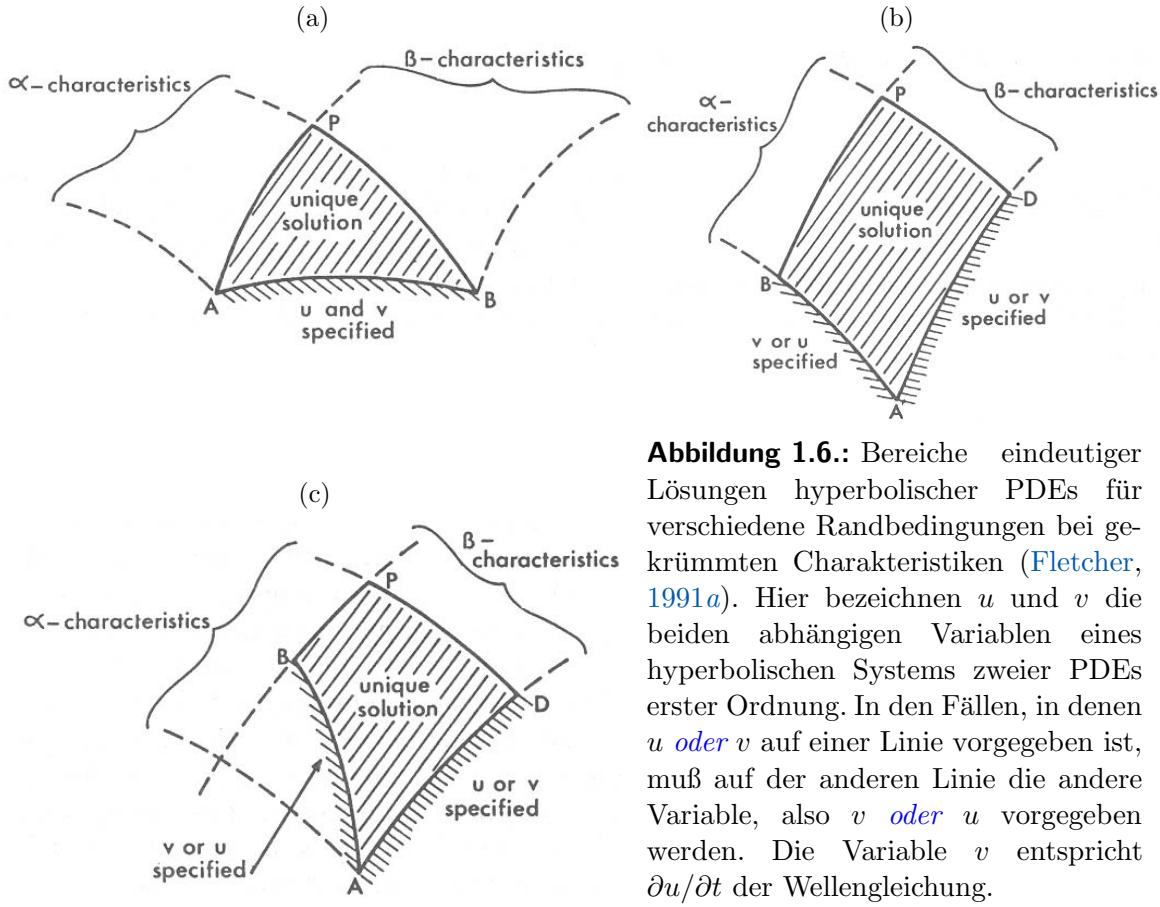


Abbildung 1.6.: Bereiche eindeutiger Lösungen hyperbolischer PDEs für verschiedene Randbedingungen bei gekrümmten Charakteristiken (Fletcher, 1991*a*). Hier bezeichnen u und v die beiden abhängigen Variablen eines hyperbolischen Systems zweier PDEs erster Ordnung. In den Fällen, in denen u *oder* v auf einer Linie vorgegeben ist, muß auf der anderen Linie die andere Variable, also v *oder* u vorgegeben werden. Die Variable v entspricht $\partial u/\partial t$ der Wellengleichung.

1.2.4. Parabolische Differentialgleichungen

Das klassische Beispiel für eine parabolische PDE ist die eindimensionale Wärmeleitungs- bzw. *Diffusionsgleichung*

$$\frac{\partial T}{\partial t} = \kappa \frac{\partial^2 T}{\partial x^2}, \tag{1.50}$$

mit der Temperaturleitfähigkeit (Wärme-Diffusivität) $\kappa > 0$. Wir können sie mit dem *Separationsansatz*

$$T(x,t) = f(x)g(t) \tag{1.51}$$

lösen. Der Ansatz liefert

$$fg' = \kappa f''g \qquad \Rightarrow \qquad \frac{g'(t)}{g(t)} = \kappa \frac{f''(x)}{f(x)} = -\lambda. \tag{1.52}$$

Da die linke Seite der Gleichung eine Funktion nur von t ist und die rechte eine Funktion nur von x, müssen beide Seiten konstant sein mit der Konstanten $-\lambda <$

0.[18] So erhalten wir die Lösungen

$$f_k(x) = A_k \sin(kx) + B_k \cos(kx) \qquad \text{und} \qquad g_k(t) = \mathrm{e}^{-\lambda t}, \tag{1.53}$$

wobei $k^2 = \lambda/\kappa$. Die Konstante λ, bzw. k, kann zunächst beliebig sein. Daher gibt es ein ganzes Spektrum von Moden

$$T_k(x,t) = \mathrm{e}^{-\kappa k^2 t} \left[A_k \sin(kx) + B_k \cos(kx)\right]. \tag{1.54}$$

Die erlaubten Werte von k ergeben sich aus den Randbedingungen.

Wenn wir den Bereich $x \in [0,1]$ betrachten und die *Randbedingungen* $T(x=0) = T(x=1) = 0$ fordern, dann muß $B_k = 0$ sein und $k = n\pi$ mit $n \in \mathbb{N}$, so daß

$$T(x,t) = \sum_{n=1}^{\infty} A_n \mathrm{e}^{-\kappa n^2 \pi^2 t} \sin(n\pi x). \tag{1.55}$$

Die Amplituden A_n ergeben sich aus den *Anfangsbedingungen*. Für $T(t=0) = A\sin(\pi x)$ bleibt nur der Summand $n = 1$ übrig. Die exakte Lösung lautet dann

$$T(x,t) = A\mathrm{e}^{-\kappa \pi^2 t} \sin(\pi x). \tag{1.56}$$

Das zeitlich exponentielle Abklingen steht im klaren Gegensatz zu dem Wellencharakter der Lösungen hyperbolischer PDEs.

Ein Vergleich von (1.50) mit (1.20) ergibt $B = C = 0$, so daß die Diskriminante $B^2 - 4AC = 0$ ist. Daher ist die Diffusionsgleichung (1.50) parabolisch. Nach (1.35) gibt es dann nur eine Schar von Charakteristiken mit Steigung $\mathrm{d}t/\mathrm{d}x = 0$. Sie sind deshalb gegeben durch $t = \text{const}$. Die Charakteristiken haben bei parabolischen Gleichungen keine große Bedeutung. Die Feldgröße an irgendeinem Punkt x_i zum Zeitpunkt t_i beeinflußt *alle* Punkte des Gebiets (hier $x \in [0,1]$) zu einem *späteren* Zeitpunkt $t \geq t_i$ (Abb. 1.7). Sie hat keinen Einfluß auf frühere Zeiten $t < t_i$. Daher kann man die Lösung einer parabolischen PDE im Prinzip dadurch erhalten, daß man das Gebiet einmal in Zeitrichtung durchschreitet (*time-marching*).

Wie wir oben schon gesehen haben, muß zum Zeitpunkt $t = 0$ die Anfangsbedingung (z.B. die Temperatur) vorgegeben werden (Dirichlet-Bedingung). Für die Ränder des räumlichen Gebiets kann man Dirichlet-, Neumann- oder gemischte Bedingungen (Robin-Randbedingungen) vorgeben. Neumann-Bedingungen entsprechen der Vorgabe des Wärmestroms $\sim \partial T/\partial x$ auf dem Rand.

Im Gegensatz zu hyperbolischen PDEs sind bei parabolischen PDEs die Lösungen immer kontinuierlich, wenn man von der einen zu einer anderen Charakteristik übergeht. Auch die Fortsetzung unstetiger Anfangs- oder Randbedingungen in das Integrationsgebiet erfolgt kontinuierlich (glatt). Es können aber sehr hohe Gradienten im Integrationsgebiet auftreten, was zu numerischen Problemen führen kann.

[18] Wenn man $\lambda < 0$ wählen würde, bekäme man Lösungen, die in der Zeit exponentiell wachsen. Die räumlichen Funktionen wären dann sinh und cosh. Mit diesen Lösungen lassen sich aber

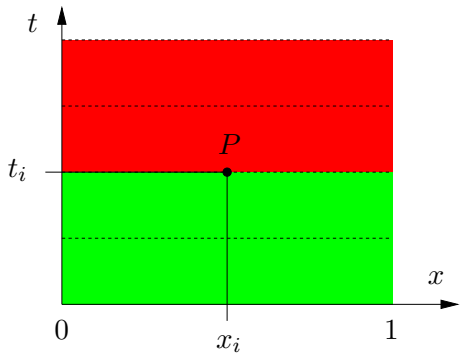


Abbildung 1.7.: Einflußgebiete für parabolische Gleichungen. Das grüne Gebiet ($t < t_i$) kann nicht vom Wert T am Punkt $P = (x_i, t_i)$ beeinflußt werden. Der Wert $u(x_i, t_i)$ beeinflußt jedoch alle Punkte im roten Gebiet ($t \geq t_i$). Die Charakteristiken genügen $t = \text{const.}$ (gestrichelte Linien).

Parabolische Differentialgleichungen besitzen typischerweise einen Propagationscharakter in t- (erste Ableitung) und einen Diffusionscharakter in x-Richtung (zweite Ableitung). Weitere parabolische Gleichungen sind die Navier-Stokes-Gleichung und die stationäre Grenzschichtgleichung. Wenn eine zeitabhängige PDE parabolisch ist (wie hier die Wärmeleitungsgleichung), dann ist die zugehörige stationäre PDE elliptisch.

1.2.5. Elliptische Differentialgleichungen

Das Paradebeispiel für eine elliptische PDE ist die *Potentialgleichung* (Laplace-Gleichung)

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} = 0. \tag{1.57}$$

Sie beschreibt zum Beispiel das Geschwindigkeitspotential von Potentialströmungen (inkompressible, rotationsfreie Strömungen[19]) oder die stationäre Wärmeleitung in zwei Dimensionen.

Eine spezielle und einfache analytische Lösung auf dem Quadrat $(x, y) \in [0, 1] \times [0, 1]$ können wir konstruieren, wenn wir $\phi(x = 0) = \phi(x = 1) = 0$ als Randbedingungen vorgeben und eine Lösung der Form $\phi(x, y) = f(y) \sin(\pi x)$ suchen. Für $f(y)$ folgt dann $f'' - \pi^2 f = 0$, was auf $f(y) = \mathrm{e}^{\pm \pi y}$ führt. Die Lösung ist demnach

$$\phi(x, y) = A\mathrm{e}^{-\pi y} \sin(\pi x), \tag{1.58}$$

wenn wir die Randbedingungen $\phi(y = 0) = A \sin(\pi x)$ und $\phi(y = 1) = A\mathrm{e}^{-\pi} \sin(\pi x)$ annehmen.

Im Gegensatz zu hyperbolischen und parabolischen PDEs hat bei elliptischen PDEs der Wert von ϕ an *jedem* Punkt eine Auswirkung auf *alle* anderen Punkte des Systems. Der Einfluß wird aber mit wachsendem Abstand immer geringer. Daher kann man zur numerischen Lösung elliptischer PDEs kein Verfahren anwenden, bei dem man nur einmal durch das Gebiet gehen muß (wie das *time-marching*

keine physikalisch sinnvollen Randbedingungen (z.B. konstante Temperatur) realisieren.

[19] Mit $\boldsymbol{\omega} = \nabla \times \boldsymbol{u} = 0$ folgt $\boldsymbol{u} = \nabla\phi$. Kontinuität erfordert bei inkompressiblen Fluiden dann $\nabla \cdot \boldsymbol{u} = \nabla \cdot \nabla\phi = 0$.

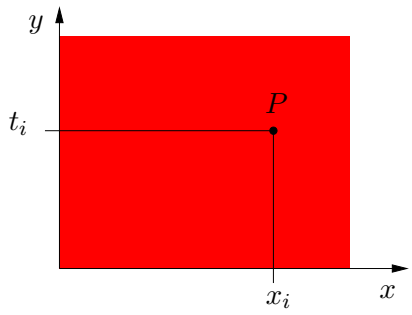


Abbildung 1.8.: Bei elliptischen Gleichungen beeinflußt jeder Punkt P alle anderen Punkte des Volumens.

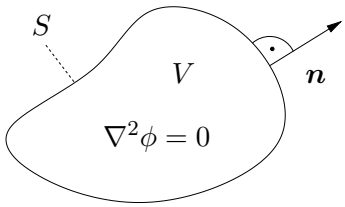


Abbildung 1.9.: Globale Bedingung bei Vorgabe des Flusses durch den Rand S eines Volumens V.

für parabolische PDEs). Ähnlich wie bei parabolischen Gleichungen werden Diskontinuitäten in den Randbedingungen im Innern des Gebiets zu kontinuierlichen Funktionen geglättet.

Bei elliptischen PDEs zweiter Ordnung der Form (1.57) existiert ein Extremal-Prinzip, wonach ϕ sowohl sein Maximum wie auch sein Minimum auf dem Rand des Gebiets annehmen muß (Garabedian, 1964). Diese Bedingung kann man auch zum Test numerischer Lösungen verwenden.

Da jeder Punkt des Gebiets jeden anderen beeinflußt (Abb. 1.8), erfordern elliptische PDEs *Randbedingungen auf allen Rändern*. Dies können Dirichlet Bedingungen durch Spezifizierung von $\phi|_S$ sein, wobei S den Rand des betrachteten Gebiets V bezeichnet. Falls Neumann-Bedingungen $\partial_n\phi|_S$ vorgegeben werden, ist Vorsicht geboten. Denn die Neumann-Randbedingung führt zu der globalen Bedingung (Abb. 1.9)

$$\int_S \boldsymbol{n}\cdot\nabla\phi\,\mathrm{d}S \overset{\text{Gauss}}{=} \int_V \nabla^2\phi\,\mathrm{d}V, \tag{1.59}$$

die konsistent mit der Lösung sein muß. Im Falle einer Potentialströmung ist $\boldsymbol{n}\cdot\nabla\phi$ die Geschwindigkeit senkrecht zum Rand. Dann ist das Integral (1.59) der totale Volumenstrom durch die Berandung aus dem Gebiet hinaus, der bei inkompressiblen Strömungen natürlich verschwinden muß: $\nabla^2\phi = 0 \;\Rightarrow\; \int_V \nabla^2\phi\,\mathrm{d}V = 0$. Mit $A = C = 1$ und $B = 0$ in (1.30) ist die Diskriminante $B^2 - 4AC = -4 < 0$ und die Charakteristiken sind komplex (siehe (1.35)). Für die Differentialgleichungen der Strömungsmechanik haben die komplexen Charakteristiken jedoch keine Bedeutung.

Zusammenfassung: *Hyperbolische* Gleichungen treten typischerweiser bei *Propagationsproblemen* auf, bei denen die Dissipation keine Rolle spielt, wie z.B. bei der ungedämpften Wellenausbreitung. *Parabolische* Gleichungen beschreiben *Ausbreitungsphänomene mit Dissipation* (wie bei der Wärmeleitung). *Elliptische* Gleichun-

gen treten bei *stationären Strömungsproblemen* auf. Da sich der Typ der Differentialgleichung im betrachteten Gebiet ändern kann A, B und C variabel), müssen die Randbedingungen so formuliert werden, daß sie zu dem Typ passen, den die Differentialgleichung am Rand des Gebiets besitzt.

2. Finite Differenzen und einige generelle Betrachtungen

Zur numerischen Lösung einer partiellen Differentialgleichung muß sie zunächst in eine geeignete Form gebracht werden. Danach erfolgt typischerweise eine *Diskretisierung*, bei der die kontinuierlichen Feldgrößen, wie u, v, w, p und T mit unendlich vielen Freiheitsgraden auf eine endliche Anzahl von Unbekannten an den Knotenpunkten eines *Gitters* reduziert werden. Bei der Diskretisierung wird die kontinuierliche Differentialgleichung in ein *System von algebraischen Gleichungen* für die endliche (aber oft große) Zahl von Unbekannten umgeformt. Der dabei auftretende *Diskretisierungsfehler* ist normalerweise der wesentliche Grund (Fehlerquelle) für Abweichungen der *numerischen Lösung* von der *exakten Lösung* der Differentialgleichung. Die Lösung der algebraischen Gleichungen liefert dann die gesuchte Approximation der exakten Lösung der Differentialgleichung. Der Fehler, der bei der numerischen Lösung der algebraischen Gleichungen gemacht wird (Rundungsfehler), ist normalerweise sehr klein im Vergleich zum Diskretisierungsfehler.

Je nach Verfahren kann der Aufwand zur Lösung der algebraischen Gleichungen verschieden sei. Bei stationären elliptischen Problemen, bei denen jeder Punkt alle anderen beeinflußt, wird das algebraische Problem oft so formuliert, daß alle Gleichungen miteinander gekoppelt sind und simultan gelöst werden. Dieses *implizite* Verfahren besteht nur aus einem Schritt. Er ist aber sehr aufwendig. Bei zeitabhängigen Problemen (meist ist der Charakter parabolisch oder hyperbolisch) wird oft eine *explizite* Formulierung verwendet. Die Feldgröße zu einem neuen Zeitpunkt hängt dann nur von einigen benachbarten Punkten ab, die zu einem oder wenigen vorherigen Zeitpunkten berechnet wurden. Bei einer expliziten Formulierung ist die algebraische Lösung einfach und schnell zu erreichen, es sind aber viele Einzelschritte in der Zeit erforderlich.

2.1. Explizite und implizite Diskretisierung

Als Beispiel für die Diskretisierung betrachten wir die zeitabhängige Wärmeleitung in einer Dimension (1.50)

$$\frac{\partial T}{\partial t} = \kappa \frac{\partial^2 T}{\partial x^2}. \tag{2.1}$$

Wir überziehen das Gebiet $x \in [0,1]$ und $t \geq 0$ mit einem äquidistanten Gitter mit *Gitterabständen* Δx und Δt und betrachten die Temperatur T_j^n nur an den

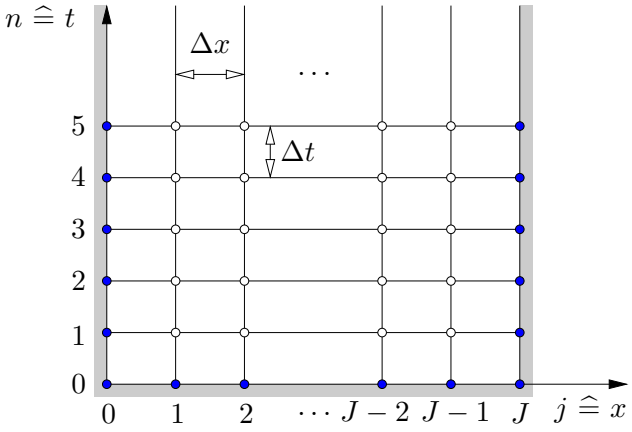


Abbildung 2.1.: Äquidistantes Gitter. Die Funktionswerte an den blauen Punkten werden durch die Rand- und Anfangsbedingungen vorgegeben.

Gitterpunkten (Knotenpunkten), die durch

$$x \longrightarrow x_j = j\Delta x, \qquad j \in [0, J], \qquad \Delta x = 1/J, \tag{2.2a}$$

$$t \longrightarrow t_n = n\Delta t, \qquad n \in [0, \infty), \tag{2.2b}$$

gegeben sind.

Die einfachste Form der Diskretisierung besteht darin, die Differentialquotienten in (2.1) durch die entsprechenden *Differenzenquotienten* zu ersetzen. Dann erhalten wir zum Beispiel am Punkt (j, n)[1]

$$\frac{T_j^{n+1} - T_j^n}{\Delta t} = \kappa \frac{T_{j+1}^n - 2T_j^n + T_{j-1}^n}{\Delta x^2}. \tag{2.3}$$

Diese Gleichung kann man nach T_j^{n+1} auflösen

$$\underbrace{T_j^{n+1}}_{\text{gesucht}} = \underbrace{T_j^n + \frac{\kappa\Delta t}{\Delta x^2}\left(T_{j+1}^n - 2T_j^n + T_{j-1}^n\right)}_{\text{bekannt } n}. \tag{2.4}$$

Wenn man nun die Temperatur an allen Orten j zum alten Zeitpunkt n kennt, kann man die Temperatur zum neuen Zeitpunkt $n+1$ für alle j berechnen. Dabei wurde angenommen, daß die Randwerte T_0^n und T_J^n durch Dirichlet-Bedingungen vorgegeben sind (blaue Punkte in Abb. 2.1).

Das Schema (2.4) wird *FTCS-Algorithmus* genannt (***F**orward in **T**ime and **C**entered in **S**pace*). Hierbei ist die zeitliche Diskretisierung asymmetrisch und die

[1] Es ist

$$\frac{\partial^2 T}{\partial x^2} = \lim_{\Delta x \to 0} \frac{1}{\Delta x}\left(\left.\frac{\partial T}{\partial x}\right|_{j+1/2} - \left.\frac{\partial T}{\partial x}\right|_{j-1/2}\right) = \lim_{\Delta x \to 0} \frac{1}{\Delta x}\left(\frac{T_{j+1} - T_j}{\Delta x} - \frac{T_j - T_{j-1}}{\Delta x}\right)$$
$$= \lim_{\Delta x \to 0} \frac{1}{\Delta x^2}\left(T_{j+1} - 2T_j + T_{j-1}\right).$$

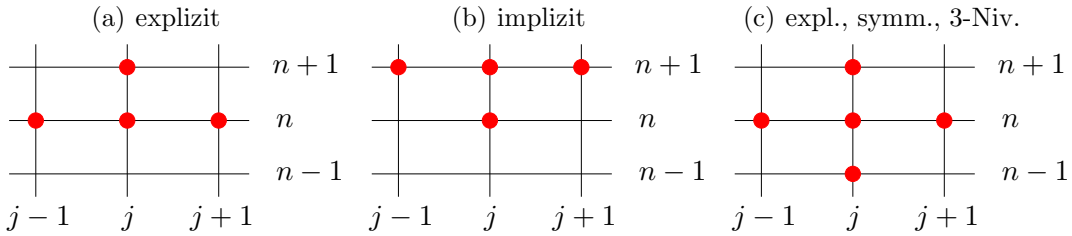


Abbildung 2.2.: Mögliche Diskretisierungen der Wärmeleitungsgleichung (s. Text).

räumliche Diskretisierung symmetrisch.[2] Die FTCS-Diskretisierung ist ein *explizites* Lösungsverfahren (Abb. 2.2a). Diese Lösungsmethode ist konsistent mit dem parabolischen Charakter der Differentialgleichung, wobei die Temperatur zum Zeitpunkt t_n nur von früheren Zeitpunkten und nicht von späteren Zeitpunkten abhängt.[3] Werte zwischen berechneten Gitterpunkten kann man mittels Interpolation erhalten.

Unter Beibehaltung der asymmetrischen zeitlichen Diskretisierung hätten wir anstelle von (2.4) auch schreiben können

$$T_j^{n+1} = T_j^n + \frac{\kappa\Delta t}{\Delta x^2}\left(T_{j+1}^{n+1} - 2T_j^{n+1} + T_{j-1}^{n+1}\right). \tag{2.5}$$

Wenn man alle unbekannten Terme zur Zeit $n+1$ auf die linke Seite bringt, erhält man das *implizite* Schema (Abb. 2.2b)

$$\underbrace{-sT_{j+1}^{n+1} + (2s+1)T_j^{n+1} - sT_{j-1}^{n+1}}_{\text{gesucht}} = \underbrace{T_j^n}_{\text{bekannt}}, \tag{2.6}$$

wobei

$$s = \frac{\kappa\Delta t}{\Delta x^2} \tag{2.7}$$

gesetzt wurde. Wenn wir nun die Temperatur zum neuen Zeitpunkt t_{n+1} berechnen wollen, müssen wir die gekoppelten Gleichungen für alle Werte $j \in [0, J]$ simultan berücksichtigen. Das ist wesentlich aufwendiger als das explizite Verfahren.

Eine dritte Variante besteht darin, den Differentenquotienten auch für die Zeit symmetrisch zu wählen. Mit $\partial T/\partial t \longrightarrow (T_j^{n+1} - T_j^{n-1})/2\Delta t$ erhalten wir

$$T_j^{n+1} = T_j^{n-1} + \frac{2\kappa\Delta t}{\Delta x^2}\left(T_{j+1}^n - 2T_j^n + T_{j-1}^n\right). \tag{2.8}$$

[2] Die räumliche Diskretisierung kann verschieden aussehen, insbesondere bei anderen Verfahren (finite Volumen, finite Elementen oder spektrale Verfahren); siehe dazu Kap. 4. Andere zeitliche Diskretisierungen werden in Kap. 6 behandelt.

[3] Durch die Diskretisierung bleibt innerhalb eines Zeitschritts der Einfluß weit entfernter Raumpunkte T_i^n mit $|i-j| \geq 2$ unberücksichtigt. Dies ist ein gewisses Artefakt der Diskretisierung. Die Information von entfernten Raumpunkten erreicht erst nach entsprechend vielen Zeitschritten den betrachteten Raumpunkt x_j.

Diese Formulierung ist wieder explizit, involviert aber insgesamt drei *Zeitniveaus* (siehe Abb. 2.2c). Der Algorithmus ist komplizierter als FTCS (2.4), aber im Prinzip genauer. In der Praxis ist das Verfahren aber nicht anwendbar, da es für die Wärmeleitungsgleichung instabil ist. Bei anderen Gleichungen kann diese vollständig symmetrische Formulierung jedoch stabil sein.

2.2. Konstruktion von Differenzenformeln

Die Taylorentwicklung einer beliebig oft differenzierbaren Funktion $f(x)$ um den Punkt x_j lautet

$$f(x) = \sum_{m=0}^{\infty} \frac{(x-x_j)^m}{m!} \left(\frac{\partial^m f}{\partial x^m}\right)_{x=x_j} . \tag{2.9}$$

Sie kann verwendet werden, um die in der PDE auftretenden Ableitungen von f am Gitterpunkt (x_j, t_n), zum Beispiel $\partial_t f|_j^n$ oder $\partial_{xx} f|_j^n$, durch die Werte an den benachbarten Gitterpunkten auszudrücken.

Dazu entwickeln wir beispielsweise den Wert $T(x_{j+1}, t_n)$ am Punkt (x_{j+1}, t_n) in eine Taylorreihe um den Entwicklungspunkt (x_j, t_n)

$$T(x_{j+1}, t_n) = \sum_{m=0}^{\infty} \frac{\Delta x^m}{m!} \left(\frac{\partial^m T}{\partial x^m}\right)_j^n = T_j^n + \Delta x \left(\frac{\partial T}{\partial x}\right)_j^n + \frac{\Delta x^2}{2} \left(\frac{\partial^2 T}{\partial x^2}\right)_j^n + O\left(\Delta x^3\right) . \tag{2.10}$$

Für den zeitversetzten Punkt (x_j, t_{n+1}) gilt entsprechend die zeitliche Taylorentwicklung

$$T(x_j, t_{n+1}) = \sum_{m=0}^{\infty} \frac{\Delta t^m}{m!} \left(\frac{\partial^m T}{\partial t^m}\right)_j^n = T_j^n + \Delta t \left(\frac{\partial T}{\partial t}\right)_j^n + \frac{\Delta t^2}{2} \left(\frac{\partial^2 T}{\partial t^2}\right)_j^n + O\left(\Delta t^3\right) . \tag{2.11}$$

Der Ausdruck $O(\Delta x^3)$ bedeutet, daß der Fehler, der bei dem Abbruch der Reihe nach dem letzten aufgeführten Summanden entsteht, von der Größenordnung Δx^3 ist. Dabei wird davon ausgegangen, daß die Faktoren (Ableitungen) bei den Termen $\sim \Delta x^m$ von der Größenordnung 1 sind, also $O(1)$. Der Fehler in (2.10) ist dann im wesentlichen bestimmt durch den größten vernachlässigten Term (mit der niedrigsten Potenz von Δx), hier also durch $a\Delta x^3$, $a = \text{const.} = O(1)$.[4]

Aus (2.10) und (2.11) erhält man durch Auflösen die Ausdrücke für die ersten Ableitungen am Punkt (x_j, t_n)

$$\left(\frac{\partial T}{\partial x}\right)_j^n = \frac{T_{j+1}^n - T_j^n}{\Delta x} + O\left(\Delta x\right) , \tag{2.12a}$$

$$\left(\frac{\partial T}{\partial t}\right)_j^n = \frac{T_j^{n+1} - T_j^n}{\Delta t} + O\left(\Delta t\right) . \tag{2.12b}$$

[4] Genauer gesagt bedeutet $O(\Delta^m)$, daß im Limes $\Delta \to 0$ gilt: $O(\Delta^m) \sim a\Delta^m$ mit $a < \infty$. Im

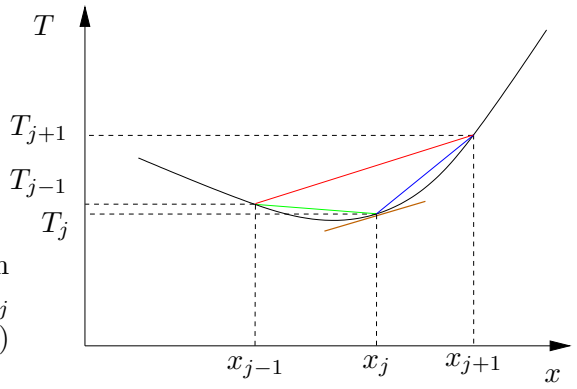


Abbildung 2.3.: Approximation der exakten Ableitung $(\partial T/\partial x)_j$ (braun) im Punkt x_j mittels Vorwärts- (blau), Rückwärts- (grün) und zentralen Differenzen (rot).

Dies sind Vorwärtsdifferenzen-Formeln (*forward Euler*). Die Rückwärts-Euler-Formel (*backward Euler*)

$$\left(\frac{\partial T}{\partial x}\right)_j^n = \frac{T_j^n - T_{j-1}^n}{\Delta x} + O\left(\Delta x\right) \tag{2.13}$$

erhält man durch die Taylor-Entwicklung von $T(x_{j-1}, t_n)$ um den Entwicklungspunkt x_j. Die geometrische Interpretation zeigt Abb. 2.3.

Will man genauere Formeln entwickeln, muß man weitere Gitterpunkte involvieren. Man setzt dann die gesuchte Ableitung als Linearkombination von Funktionswerten an den gewünschten Gitterpunkten an. Für drei symmetrische Punkte lautet dann der Ansatz

$$\left(\frac{\partial T}{\partial x}\right)_j^n = aT_{j-1}^n + bT_j^n + cT_{j+1}^n + O\left(\Delta x^m\right), \tag{2.14}$$

wobei wir a, b und c bestimmen wollen und jetzt noch nicht genau wissen, welche Größenordnung $O(\Delta x^m)$ der Fehler hat. Wenn wir nun für T_{j-1}^n, T_j^n und T_{j+1}^n die Taylorentwicklungen von T um x_j an den Punkten x_{j-1}, x_j und x_{j+1} einsetzen, erhalten wir

$$\begin{aligned} aT_{j-1}^n + bT_j^n + cT_{j+1}^n = \underbrace{(a+b+c)}_{=0} T_j^n + \underbrace{(c-a)\Delta x}_{=1} \left(\frac{\partial T}{\partial x}\right)_j^n + \underbrace{(a+c)}_{=0} \frac{\Delta x^2}{2} \left(\frac{\partial^2 T}{\partial x^2}\right)_j^n \\ + (c-a)\frac{\Delta x^3}{6} \left(\frac{\partial^3 T}{\partial x^3}\right)_j^n + O\left(\Delta x^4\right). \end{aligned} \tag{2.15}$$

Offensichtlich müssen wir $(c-a)\Delta x = 1$ setzen, da sonst die erste Ableitung ganz eliminiert wird. Zur Festlegung der drei Unbekannten a, b, c haben wir dann noch

Limes $\Delta \to 0$ bleibt also nur der angegebene Term übrig und es liegen beliebig viele Größenordnungen zwischen den Summanden mit unterschiedlichen Exponenten $n > m$. Dann werden die Größenordnungen der Beträge dieser Summanden beliebig stark voneinander getrennt. Für hinreichend kleines Δ ist der Fehler daher durch den führenden Term bestimmt.

zwei weitere Bedingungen frei. Diese wählen wir so, daß möglichst viele Koeffizienten vor den größten Fehlertermen verschwinden. Dies führt auf 3 Bedingungen, die wir als lineares Gleichungssystem schreiben können

$$\begin{pmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 0 \\ 1/\Delta x \\ 0 \end{pmatrix}. \tag{2.16}$$

Die Lösung lautet

$$c = -a = \frac{1}{2\Delta x}, \qquad b = 0. \tag{2.17}$$

Als Ergebnis erhalten wir die *zentralen Differenzen* (Abb. 2.3)

$$\left(\frac{\partial T}{\partial x}\right)_j^n = \frac{T_{j+1}^n - T_{j-1}^n}{2\Delta x} + O\left(\Delta x^2\right). \tag{2.18}$$

Der Fehler ist von der Größenordnung $O(\Delta x^2)$, da der erste nicht verschwindende und vernachlässigte Term in (2.15) formal von der Größenordnung $O(\Delta x^3)$ ist, aber noch mit $c - a = \Delta x^{-1}$ multipliziert wird. Im Vergleich zu den Vorwärts- (2.12a) und Rückärtsdifferenzen (2.13) sind die zentralen Differenzen also um eine Ordnung genauer.

In derselben Weise kann man auch asymmetrische Differenzenformeln für den Punkt (j, n) entwickeln. Eine *3-Punkt-Formel* für die erste Ableitung erhält man mit dem Ansatz

$$\left(\frac{\partial T}{\partial x}\right)_j^n = aT_j^n + bT_{j+1}^n + cT_{j+2}^n + O\left(\Delta x^m\right). \tag{2.19}$$

Wenn man die Taylorentwicklungen für T_{j+1}^n und T_{j+2}^n um den Punkt (x_j, t_n) einsetzt, erhält man

$$\begin{aligned} aT_j^n + bT_{j+1}^n + cT_{j+2}^n = &\underbrace{(a+b+c)}_{=0} T_j^n + \underbrace{(b+2c)\Delta x}_{=1} \left(\frac{\partial T}{\partial x}\right)_j^n \\ &+ \underbrace{(b+4c)}_{=0} \frac{\Delta x^2}{2} \left(\frac{\partial^2 T}{\partial x^2}\right)_j^n + O\left(\Delta x^3\right). \end{aligned} \tag{2.20}$$

Die drei Bedingungen führen auf

$$a = -\frac{3}{2\Delta x}, \qquad b = \frac{4}{2\Delta x}, \qquad c = -\frac{1}{2\Delta x}, \tag{2.21}$$

wodurch wir die *einseitige Differenzenformel 2. Ordnung*

$$\left(\frac{\partial T}{\partial x}\right)_j^n = \frac{-3T_j^n + 4T_{j+1}^n - T_{j+2}^n}{2\Delta x} + O\left(\Delta x^2\right) \tag{2.22}$$

erhalten.

Man kann nun noch weitere Punkte hinzunehmen und die Funktionswerte an den Stellen mit entsprechenden unbestimmten Vorfaktoren in den Ansatz einbeziehen. Durch Nullsetzen der Koeffizienten der Ableitungen lassen sich weitere (höhere) Ableitungen eliminieren, wodurch der *Abbruchfehler* (*truncation error*) weiter verkleinert wird. Die so entstehenden Differenzenformeln höherer Ordnung werden zwar immer genauer, sie haben aber auch Nachteile: Zum einen sind sie komplizierter und zum anderen führen sie oft zu weniger stabilen Algorithmen. In vielen Fällen stellen daher die einfachen Formeln schon die beste Wahl dar.

2.3. Beispiel: Eindimensionale Wärmeleitung

Als Beispiel kann man die zeitabhängige eindimensionale Wärmeleitung berechnen. Dazu betrachte das Anfangswertproblem (1.50) auf dem Gebiet $(x,t) \in [0,1] \times [0,\infty)$ mit den Randbedingungen $T(x=0) = T(x=1) = T_0 \neq 0$ und der Anfangsbedingung $T(t=0) = 0$. Dies entspricht dem Aufheizen eines Drahtes durch plötzliches Aufprägen der Temperatur T_0 an den Enden.

Man kann nun das *explizite FTCS-Schema* (2.4)

$$T_j^{n+1} = sT_{j-1}^n + (1-2s)T_j^n + sT_{j+1}^n \tag{2.23}$$

in einem Computer-Programm mit $s = 0.5$ implementieren (Übung). Wenn man ein Zeitniveau n berechnet hat, erhält man mit diesem Schema die Temperatur für das darauf folgende Niveau $n+1$ an den inneren Punkten $j \in [2, J-1]$. Die Randwerte an den Punkten $j = 1, J$ sind durch die Randbedingungen festgelegt. Wegen der unstetigen Änderung der Randtemperatur bei $t = 0$ ist es für eine bessere Approximation der exakten Lösung[5] (vgl. (1.55))

$$T(x,t) = T_0 - T_0 \sum_{m=1}^{\infty} \frac{4\sin\left[(2m-1)\pi x\right]}{(2m-1)\pi} \mathrm{e}^{-\kappa(2m-1)^2\pi^2 t}. \tag{2.24}$$

günstig, wenn man $T_1^0 = T_J^0 = T_0/2 \neq 0$ setzt. Das macht den Einschaltvorgang etwas glatter. Die numerische Lösung für verschiedene Zeiten ist in Abb. 2.4 gezeigt. Die gestrichelten Kurven entsprechen der Summe (2.24), die bei einem hohen Wert von m abgebrochen wurde.

2.4. Diskretisierungsfehler

2.4.1. Fehlerordnung bei homogenen Gittern

Der Fehler bei der Diskretisierung der Ableitungen einer Funktion hängt von den vernachlässigten Termen ab. Die Größenordnung dieses Diskretisierungsfehlers ska-

[5] Aus Symmetriegründen brauchen nur ungerade Werte von n in (1.55) berücksichtigt zu werden.

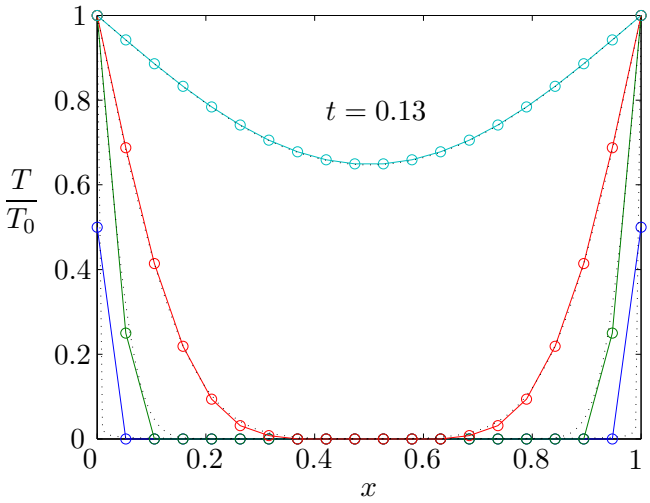


Abbildung 2.4.: Lösung der eindimensionalen Wärmeleitungsgleichung (1.50) mittels FTCS-Schema (2.23) für $s = 0.50$ und $J = 20$ ($\Delta x = 1/19$). Es wurde $\kappa = 1$ gesetzt. Blau: Anfangsbedingung $t = 0$, grün: $t = \Delta t$, rot: $t = 6\Delta t$ und cyan: $t = 94\Delta t \approx 0.13$ ($\kappa = 1$). Die gepunkteten Linien stellen die Galerkin-Lösung $T(x) = T_0 \left\{1 - \sum_{n=1}^{1000} \frac{2}{n\pi}[1-(-1)^n] \mathrm{e}^{-\kappa n^2\pi^2 t} \sin(n\pi x)\right\}$ zu den entsprechenden Zeitpunkten dar, vgl. (2.24) und Kap. 4.5.1.

liert mit der niedrigsten Potenz des Gitterabstands Δx, die in den vernachlässigten Termen auftritt. Die tatsächliche Größe hängt außerdem noch von der entsprechenden exakten (höheren) Ableitung der approximierten Funktion ab.[6]

Im Limes $\Delta x \to 0$ bestimmt ausschließlich der führende vernachlässigte Term den tatsächlichen Fehler ϵ der numerischen Rechnung (z.B. der Strömungsgeschwindigkeit an einem festen Punkt). Deshalb reduziert sich der Fehler bei einer Gitterverfeinerung $\Delta x \to 0$ asymptotisch wie

$$\epsilon = K\,(\Delta x)^m\,. \tag{2.25}$$

Man nennt dann m die *Ordnung* des verwendeten Schemas. In einem logarithmischen Plot des Fehlers (Abb. 2.5a)

$$\ln(\epsilon) = \ln(K) + m\ln(\Delta x) \tag{2.26}$$

kann man an der Steigung die Ordnung m des Verfahrens ablesen.

Die Abhängigkeit irgendeiner lokalen oder integralen Feldgröße f^{num} vom Gitterabstand Δx

$$f^{\mathrm{num}}(\Delta x) = f^{\mathrm{exakt}} + \epsilon \sim f^{\mathrm{exakt}} + K\,(\Delta x)^m \tag{2.27}$$

kann man zu Konvergenztests verwenden, indem man f^{num} als Funktion von $(\Delta x)^m$ aufträgt. Für hinreichend kleinen Gitterabstand Δx sollte die Abweichung vom

Für $t = 0$ erkennt man die Fourier-Reihe der Stufenfunktion.

[6] In Fletcher (1991*a*) finden sich einige Tabellen, in denen die Approximationen mittels finiter

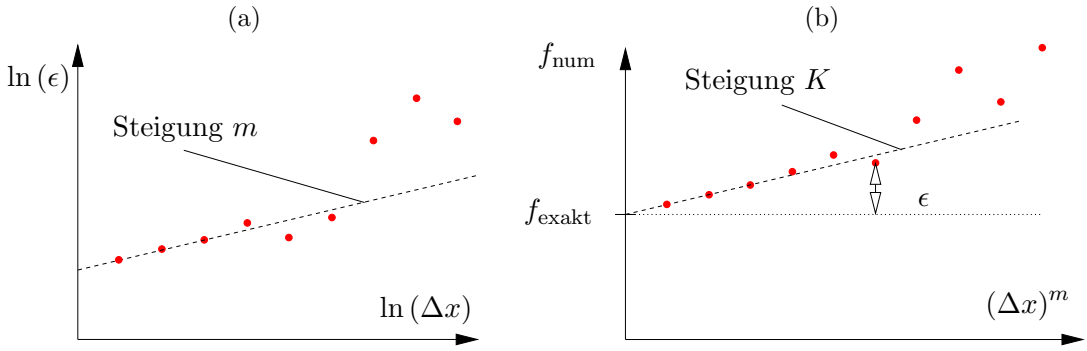


Abbildung 2.5.: Schematische Darstellung eines typischen Konvergenzverhaltens (rote Punkte) als Funktion des Gitterabstands Δx. Die gestrichelten Geraden zeigen das asymptotische Verhalten für $\Delta x \to 0$. (a) Fehler als Funktion von Δx, doppel-logarithmisch. (b) Numerische Feldgröße als Funktion von $(\Delta x)^m$ bei bekannter Ordnung m.

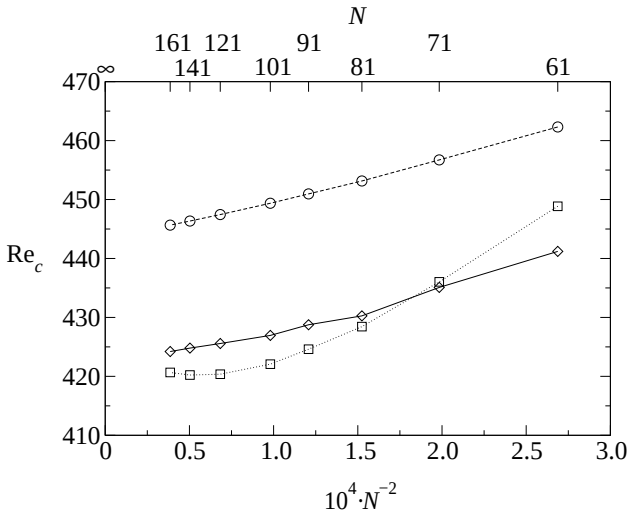


Abbildung 2.6.: Konvergenzverhalten der berechneten kritischen Reynoldszahl Re_c als Funktion der Anzahl der Unbekannten N pro Raumrichtung ($\Delta x \sim N^{-1}$) für den Übergang von einer zwei- zu einer dreidimensionalen Wirbelströmung in einem Rechteckbehälter, in dem die Strömung durch die tangentiale Bewegung einer Seitenwand angetrieben wird (lid-driven cavity, ähnlich wie beim Frontispiz) (nach Albensoeder et al., 2001). Die Symbole bezeichen verschiedenen Höhen-zu-Breiten-Verhältnisse Γ des Behälters: $\Gamma = 2.00$ (◯), $\Gamma = 3.00$ (◇) and $\Gamma = 4.00$ (□). Es wurde ein Verfahren zweiter Ordnung verwendet.

exakten Wert f^{exakt} linear mit $(\Delta x)^m$ variieren (Abb. 2.5b). Im asymptotischen Bereich läßt sich dann die numerisch berechnete Größe damit auf ein unendlich feines Gitter ($\Delta x \to 0$) extrapolieren. Ein konkretes Beispiel ist in Abb. 2.6 gezeigt.

Differenzen mit den exakten Ableitungen (zum Beispiel von e^x) verglichen werden.

Je höher die Fehlerordnung ist, desto genauer ist das numerische Verfahren. Eine Diskretisierung hoher Ordnung involviert aber relativ viele Gitterpunkte. Bei einem groben Gitter können daher schnelle Variationen nicht gut mit einem Verfahren hoher Ordnung wiedergegeben werden. Deshalb erhält man auf groben Gittern genauere Ergebnisse mit Verfahren niedriger Ordnung. Verfahren höherer Ordnung sind denjenigen niedriger Ordnung erst auf sehr feinen Gittern überlegen (siehe auch Fletcher, 1991*a*).

2.4.2. Gitterstreckung

Der Diskretisierungsfehler hängt nicht nur von der Ordnung des in der Taylorentwicklung vernachlässigten führenden Terms ab, sondern auch von der entsprechenden Ableitung der Funktion am den Gitterpunkten (x_j, t_n). Für die Zentrale-Differenzen-Formel (2.18) ist zum Beispiel der führende Fehler auf einem äquidistanten Gitter $(\Delta x^2/6)\partial^3 T/\partial x^3$.

Um eine über das ganze Integrationsgebiet gleichmäßig gute Näherung zu erhalten, ist man daher bestrebt, die Gitterweite Δx dort zu verkleinern, wo die Ableitungen der Unbekannten sehr groß sind. Dies ist bei viskosen Strömungen meist in der Nähe der Wände der Fall. Denn dort entstehen für große Reynoldszahlen dünne viskose Grenzschichten.

Um den Einfluß eines *variablen* Gitters mit *Gitterabstand* $\Delta x_i = x_i - x_{i-1}$ auf den Diskretisierungsfehler zu untersuchen, betrachten wir als Beispiel die Approximation der Ableitung im Punkt x_i durch zentrale Differenzen

$$\left(\frac{\partial T}{\partial x}\right)_i = \frac{T_{i+1} - T_{i-1}}{\Delta x_{i+1} + \Delta x_i} + \epsilon. \tag{2.28}$$

Für den *Diskretisierungsfehler* ϵ erhält man[7]

$$\epsilon = -\frac{\Delta x_{i+1}^2 - \Delta x_i^2}{2\left(\Delta x_{i+1} + \Delta x_i\right)}\left(\frac{\partial^2 T}{\partial x^2}\right)_i - \frac{\Delta x_{i+1}^3 + \Delta x_i^3}{3!\left(\Delta x_{i+1} + \Delta x_i\right)}\left(\frac{\partial^3 T}{\partial x^3}\right)_i + \ldots \tag{2.29}$$

Wenn nun das Gitter homogen ist, dann ist $\Delta x_{i+1} = \Delta x_i$. Damit verschwindet der erste Term und wir erhalten den obigen bekannten Fehler. Ein inhomogenes Gitter führt also formal zu einem größeren Fehler.

Eine Möglichkeit der Verfeinerung eines inhomogenen Gitters besteht darin, zwischen zwei benachbarten Gitterpunkten des inhomogenen Gitters jeweils einen neuen Punkt einzufügen. Wenn wir bei dieser Verfeinerungstrategie sukzessive das Intervall zwischen zwei Gitterpunkten halbieren, hat man nach einigen Verfeinerungen

[7] Den Fehler, den man bei der Approximation der ersten Ableitung mit zentralen Differenzen macht, erhält man aus den Taylorentwicklungen um den Punkt x_i

$$T_{i+1} - T_{i-1} = \left(\Delta x_{i+1} + \Delta x_i\right)\left(\frac{\partial T}{\partial x}\right)_i + \underbrace{\frac{\Delta x_{i+1}^2 - \Delta x_i^2}{2}\left(\frac{\partial^2 T}{\partial x^2}\right)_i + \frac{\Delta x_{i+1}^3 + \Delta x_i^3}{3!}\left(\frac{\partial^3 T}{\partial x^3}\right)_i + \ldots}$$

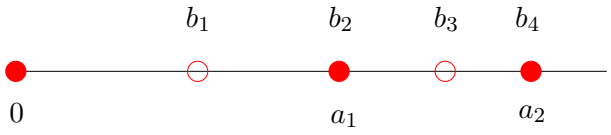


Abbildung 2.7.: Gitterverfeinerung bei einer Gitterstreckung um einen konstanten Faktor r_{2h} zwischen den vollen Punkten a_i und um einen Faktor r_h zwischen allen Punkten b_i.

praktisch ein stückweise homogenes Gitter. Daher wird der Fehler bei einer großen Zahl von Gitterpunkten ähnlich sein wie bei einem vollständig homogenen Gitter.

Eine andere Möglichkeit besteht in einer *systematischen* Gitterverfeinerung mit einem *Streckungsfaktor* r, so daß $\Delta x_{i+1} = r\Delta x_i$. Dann gilt für den führenden Fehler

$$\epsilon \approx -\frac{(r^2-1)\Delta x_i^2}{2(r+1)\Delta x_i}\left(\frac{\partial^2 T}{\partial x^2}\right)_i = \frac{(1-r)\Delta x_i}{2}\left(\frac{\partial^2 T}{\partial x^2}\right)_i. \tag{2.30}$$

Er verschwindet für $r \to 1$ (homogenes Gitter). Wenn man das Gitter nun in jedem Schritt so verfeinert, daß das resultierende Gitter wieder gleichmäßig gestreckt ist (Abb. 2.7), dann nähert sich der Streckungsfaktor $t \to 1$ und der führende Fehlerterm verschwindet im Limes eines unendlich feinen Gitters.

Aus Abb. 2.7 findet man für die Streckungsfaktoren auf dem feinen (r_h) (mittlerer Punktabstand h) und auf dem grobe Gitter (r_{2h}) (mittlerer Punktabstand $2h$)

$$r_{2h} = \frac{a_2 - a_1}{a_1} = \frac{b_4 - b_2}{b_2} = \frac{b_1\left(1 + r_h + r_h^2 + r_h^3\right) - b_1\left(1 + r_h\right)}{b_1\left(1 + r_h\right)} = r_h^2, \tag{2.31}$$

so daß

$$r_h = \sqrt{r_{2h}}. \tag{2.32}$$

Durch das sukzessive Wurzelziehen bei einer systematischen Gitterverfeinerung nähert sich der Streckungsfaktor $r \to 1$ immer mehr demjenigen des homogenen Gitters an.

Das Verhältnis des führenden Fehlers (2.30) auf einem groben zu demjenigen auf einem feinen gestreckten Gitter (mit doppelter Punktzahl) an einem gemeinsamen Punkt x_i beträgt

$$\Lambda = \frac{(1-r_{2h})}{(1-r_h)}\frac{(\Delta x_i)_{2h}}{(\Delta x_i)_h} \overset{(*)}{=} \frac{(1-r_h^2)}{(1-r_h)}\frac{(1+r_h)(\Delta x_i)_h/r_h}{(\Delta x_i)_h} = \frac{(1+r_h)^2}{r_h} \geq 4, \tag{2.33}$$

wobei wir im Schritt (*) beachtet haben: $(\Delta x_i)_{2h} = (\Delta x_i)_h + (\Delta x_{i-1})_h = (1 + r_h)(\Delta x_{i-1})_h = (1 + r_h)(\Delta x_i)_h/r_h$.

Bei einer Verdopplung der Gitterpunkte eines homogenen Gitters mit $r_h = 1$ verringert sich der führende Fehlterterm (zweiter Summand in (2.29)) um einen Faktor 4. Bei einem inhomogenen Gitter verringert sich der führende Fehlterterm (erster Summand in (2.29)) wegen $\Lambda > 4$ jedoch schneller (siehe Abb. 2.8)! Damit verringert sich bei einem so gestreckten Gitter der Fehlerterm *erster* Ordnung (erster Summand in 2.29) schneller als der Fehlerterm *zweiter* Ordnung (zweiter Summand in (2.29)).

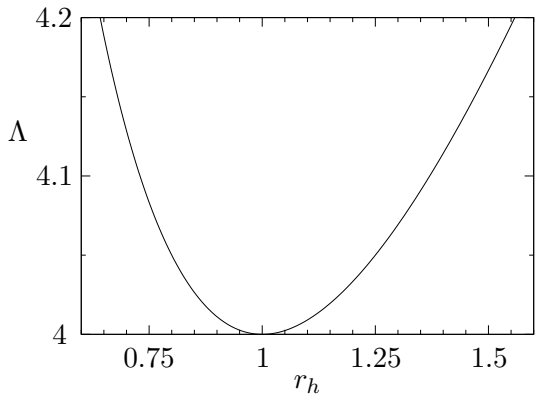


Abbildung 2.8.: Verhältnis Λ des Fehlers auf einem groben zu demjenigen auf einem feinen Gitter bei Verdopplung der Zahl der Gitterpunkte und einem räumlich konstanten Streckungsfaktor r_h.

Auch wenn der Abbruchfehler bei einem gestreckten Gitter formal nur von erster Ordnung ist, so nähert er sich jedoch bei einer sukzessiven Verfeinerung wegen $r_h \to 1$ (2.32) asymptotisch der zweiten Ordnung an. Für eine begrenzte Anzahl von Gitterpunkten ist die Lösung auf einem geeignet gestreckten Gitter fast immer genauer als diejenige auf einem homogenen Gitter. Der optimale Streckungsfaktor hängt von der Anzahl der Gitterpunkte und von den aufzulösenden Gradienten ab. In der Praxis sollte der Streckungsfaktor aber nicht außerhalb des Bereichs $0.85 \lesssim r \lesssim 1.2$ liegen.

2.4.3. Spektrale Betrachtung von Diskretisierungsfehlern

In vielen strömungsmechanischen Problemen treten Wellen auf. Es ist nun instruktiv zu sehen, welche Diskretisierungsfehler hierbei entstehen. Dazu sei zunächst die *Fourier-Darstellung* einer periodischen Funktion rekapituliert.

Diskrete Fourier-Transformation

Die Fourier-Transformation einer *kontinuierlichen* periodischen Funktion $g(x)$ mit *Periode* 2π lautet

$$g(x) = \sum_{m=-\infty}^{\infty} \hat{g}_m \mathrm{e}^{\mathrm{i}mx}, \tag{2.34}$$

mit $m \in \mathbb{Z}$, $x \in \mathbb{R}$ und $\hat{g}_m \in \mathbb{C}$.[8] Die *Amplituden* $\hat{g}_m$ der spektralen Komponenten lauten[9]

$$\hat{g}_m = \frac{1}{2\pi} \int_0^{2\pi} g(x) \mathrm{e}^{-\mathrm{i}mx} \,\mathrm{d}x. \tag{2.35}$$

[8] Für reelle Funktionen $g(x) \in \mathbb{R}$ muß man fordern $\hat{g}_{-m} = \hat{g}_m^*$.

[9] Probe durch Einsetzen:

$$\hat{g}_m = \frac{1}{2\pi} \int_0^{2\pi} \sum_{k=-\infty}^{\infty} \hat{g}_k \mathrm{e}^{\mathrm{i}kx} \mathrm{e}^{-\mathrm{i}mx} \,\mathrm{d}x = \sum_{k=-\infty}^{\infty} \hat{g}_k \underbrace{\frac{1}{2\pi} \int_0^{2\pi} \mathrm{e}^{\mathrm{i}(k-m)x} \,\mathrm{d}x}_{=\delta_{km}} = \sum_{k=-\infty}^{\infty} \hat{g}_k \delta_{km} = \hat{g}_m.$$

In der Numerik hat man in der Regel keine kontinuierlichen Funktionen. Die Funktion ist nur diskret auf den Stützstellen eines Gitters vorgegeben. Bei der *diskreten* Darstellung der 2π-periodischen Funktion auf einem Gitter mit $\Delta x = 2\pi/J$ kann man die diskrete Funktion $g_j = g(x_j = j\Delta x = 2\pi j/J)$ als *diskrete Fourier-Reihe* darstellen (siehe auch Kap. 4.6.1)

$$g_j = \sum_{m=0}^{J-1} \hat{g}_m \mathrm{e}^{\mathrm{i}m \overbrace{j\Delta x}^{x_j}} = \sum_{m=0}^{J-1} \hat{g}_m \mathrm{e}^{\mathrm{i}2\pi mj/J}. \tag{2.36}$$

Man kann leicht nachprüfen, daß die Fourier-Amplituden $\hat{g}_m$ dann durch

$$\hat{g}_m = \frac{1}{J} \sum_{j=0}^{J-1} g_j \mathrm{e}^{-\mathrm{i}m \overbrace{j\Delta x}^{x_j}} = \frac{1}{J} \sum_{j=0}^{J-1} g_j \mathrm{e}^{-\mathrm{i}2\pi mj/J} \tag{2.37}$$

gegeben sind.[10]

Beachte, daß sowohl die Funktionswerte g_j im Ortsraum, wie auch die spektralen Komponenten $\hat{g}_m$ periodisch sind mit Periode J. Für die periodische Funktion im Ortsraum ist das klar. Für die spektralen Komponenten gilt dies auch. Denn für $m \to m + nJ$, $n \in \mathbb{Z}$ gilt

$$\hat{g}_{m+nJ} = \frac{1}{J} \sum_{j=0}^{J-1} g_j \mathrm{e}^{-\mathrm{i}2\pi(m+nJ)j/J} = \frac{1}{J} \sum_{j=0}^{J-1} g_j \mathrm{e}^{-\mathrm{i}2\pi mj/J} \underbrace{\mathrm{e}^{-\mathrm{i}2\pi nj}}_{=1} = \hat{g}_m. \tag{2.38}$$

Darüber hinaus kann man den Summationsindex in (2.36) und (2.37) beliebig um $n \in \mathbb{Z}$ verschieben.[11] Insbesondere kann man ihn auch symmetrisch wählen, von $j = -(J-1)/2$ bis $(J-1)/2$ (J ungerade) oder von $j = -J/2$ bis $J/2 - 1$ (J

[10] Die Probe ergibt

$$\begin{aligned}\hat{g}_m &= \frac{1}{J} \sum_{j=0}^{J-1} \left(\sum_{k=0}^{J-1} \hat{g}_k \mathrm{e}^{\mathrm{i}kj\Delta x} \right) \mathrm{e}^{-\mathrm{i}mj\Delta x} = \frac{1}{J} \sum_{j=0}^{J-1} \sum_{k=0}^{J-1} \hat{g}_k \mathrm{e}^{\mathrm{i}(k-m)j\Delta x} \\ &= \frac{1}{J} \sum_{k=0}^{J-1} \hat{g}_k \underbrace{\sum_{j=0}^{J-1} \mathrm{e}^{\mathrm{i}(k-m)j\Delta x}}_{=J\delta_{km}} = J \frac{1}{J} \underbrace{\sum_{k=0}^{J-1} \hat{g}_k \delta_{km}}_{=\hat{g}_m} = \hat{g}_m.\end{aligned}$$

Hierbei wurde beachtet, daß für $n = k - m \neq 0$ die Summe $\sum_{j=0}^{J-1} \mathrm{e}^{\mathrm{i}nj\Delta x} = 0$ ist, da die J komplexen Zeiger $\mathrm{e}^{\mathrm{i}nj\Delta x} = \mathrm{e}^{\mathrm{i}nj2\pi/J}$ den Einheitskreis (bzw. das n-fache des Einheitskreises) äquidistant abdecken.

[11] Für die periodische Funktion g_j ist das klar. Man kann aber auch den Index der Fouriertransformierten um ein beliebiges $n \in \mathbb{Z}$ verschieben, denn es gilt aufgrund der Periodizität von g_j

$$\hat{g}_m = \frac{1}{J} \sum_{j=0}^{J-1} g_j \mathrm{e}^{-\mathrm{i}2\pi mj/J} \overset{\text{Aufspalten}}{=} \frac{1}{J} \sum_{j=0}^{n-1} g_j \mathrm{e}^{-\mathrm{i}2\pi mj/J} + \frac{1}{J} \sum_{j=n}^{J-1} g_j \mathrm{e}^{-\mathrm{i}2\pi mj/J}$$

gerade).

Eine Index-Verschiebung im Ortsraum um n liefert im spektralen Raum eine Phasenverschiebung der Amplituden $\hat{g}_m$ um $\mathrm{e}^{\mathrm{i}2\pi mn/J}$, denn es gilt

$$\begin{aligned}\hat{g}_m &= \frac{1}{J}\sum_{j=0}^{J-1} g_{j+n}\mathrm{e}^{-\mathrm{i}2\pi mj/J} = \mathrm{e}^{\mathrm{i}2\pi mn/J}\frac{1}{J}\sum_{j=0}^{J-1} g_{j+n}\mathrm{e}^{-\mathrm{i}2\pi m(j+n)/J} \\ &= \mathrm{e}^{\mathrm{i}2\pi mn/J}\frac{1}{J}\sum_{j=n}^{n+J-1} g_j\mathrm{e}^{-\mathrm{i}2\pi mj/J} \overset{\text{Fußnote11}}{=} \mathrm{e}^{\mathrm{i}2\pi mn/J}\frac{1}{J}\sum_{j=0}^{J-1} g_j\mathrm{e}^{-\mathrm{i}2\pi mj/J}. \end{aligned} \tag{2.39}$$

Falls die Funktionswerte $g_j \in \mathbb{R}$ reell sind, hat man J reelle Daten. Die Fouriertransformation liefert aber formal $2J$ Daten, weil die spektralen Komponenten $\hat{g}_m \in \mathbb{C}$ komplex sind. Daher muß die Information in $\hat{g}_m$ redundant sein. In der Tat gilt für reelle Funktionen

$$\hat{g}_{J-m} = \hat{g}_m^*, \tag{2.40}$$

Dies sieht man, wenn man beachtet

$$\hat{g}_{J-m} = \frac{1}{J}\sum_{j=0}^{J-1} g_j\mathrm{e}^{-\mathrm{i}2\pi(J-m)j/J} = \frac{1}{J}\sum_{j=0}^{J-1} g_j\mathrm{e}^{\mathrm{i}2\pi mj/J}\underbrace{\mathrm{e}^{-\mathrm{i}2\pi J/J}}_{=1} \overset{g_j\in\mathbb{R}}{=} \hat{g}_m^*. \tag{2.41}$$

Für reelle Funktionen $g_j \in \mathbb{R}$ ist die Amplitude des konstanten Anteils $\hat{g}_0 \in \mathbb{R}$ auch reell. Bei einer geraden Anzahl J von Stützstellen ist wegen (2.40) die Nyquist-Amplitude (s.u.) $\hat{g}_{J/2} \in \mathbb{R}$ ebenfalls reell.[12]

Fehler bei der Finite-Differenzen-Approximation einer periodischen Funktion

Auf dem diskreten Gitter können keine Wellenlängen wiedergegeben werden, die kleiner sind als die *cut-off-Wellenlänge* $\lambda_c = 2\Delta x$. Die *cut-off*-Wellenlänge wird auch *Nyquist-Wellenlänge* $\lambda_{\text{Nyquist}} = \lambda_c$ genannt. Ihr entspricht die *Nyquist-Wellenzahl*

$$k_{\text{Nyquist}} = \frac{2\pi}{\lambda_c} = \frac{\pi}{\Delta x} \triangleq \frac{J}{2}. \tag{2.42}$$

$$\begin{aligned}&\overset{g\text{ periodisch}}{=} \frac{1}{J}\sum_{j=0}^{n-1} g_{J+j}\mathrm{e}^{-\mathrm{i}2\pi mj/J} + \frac{1}{J}\sum_{j=n}^{J-1} g_j\mathrm{e}^{-\mathrm{i}2\pi mj/J} \\ &= \frac{1}{J}\sum_{j=0}^{n-1} g_{J+j}\mathrm{e}^{-\mathrm{i}2\pi m(j+J)/J} + \frac{1}{J}\sum_{j=n}^{J-1} g_j\mathrm{e}^{-\mathrm{i}2\pi mj/J} \\ &= \frac{1}{J}\sum_{j=J}^{J+n-1} g_j\mathrm{e}^{-\mathrm{i}2\pi mj/J} + \frac{1}{J}\sum_{j=n}^{J-1} g_j\mathrm{e}^{-\mathrm{i}2\pi mj/J} = \sum_{j=n}^{n+J-1} g_j\mathrm{e}^{-\mathrm{i}2\pi mj/J}.\end{aligned}$$

[12] Für gerades J hat man dann die reellen Amplituden $\hat{g}_0$ und $\hat{g}_{J/2}$ sowie die $J/2-1$ nichtredundanten komplexen Amplituden $\hat{g}_1, \ldots, \hat{g}_{J/2-1}$. Dies ergibt zusammen J reelle Koeffizienten.

In diesem Sinne muß jede diskrete Approximationen einer kontinuierlichen Funktionen als eine *langwellige Approximation* verstanden werden. Je kürzer die Wellenlänge ist, desto schlechter werden die entsprechenden Amplituden durch die diskrete Approximation wiedergegeben. Um dies zu quantifizieren, betrachten wir die monochromatische Welle

$$T(x,t) = \cos\left[k\left(x - ct\right)\right], \tag{2.43}$$

mit der exakten (angedeutet durch den Querstrich) Ableitung

$$\frac{\partial \bar{T}}{\partial x} = -k \sin\left[k\left(x - ct\right)\right]. \tag{2.44}$$

Die diskrete Approximation der ersten Ableitung mit der diskreten symmetrischen 3-Punkt-Formel liefert[13]

$$\begin{aligned}\left(\frac{\partial T}{\partial x}\right)_j^n &\approx \frac{T_{j+1}^n - T_{j-1}^n}{2\Delta x} = \frac{\cos\left\{k\left[(x_j + \Delta x) - ct_n\right]\right\} - \cos\left\{k\left[(x_j - \Delta x) - ct_n\right]\right\}}{2\Delta x} \\ &= \frac{-k \sin\left[k\left(x_j - ct_n\right)\right] \sin\left(k\Delta x\right)}{k\Delta x}.\end{aligned} \tag{2.45}$$

Daraus folgt für das Verhältnis der diskreten Approximation der ersten Ableitung zur exakten Ableitung am Gitterpunkt $x = x_j$

$$\frac{\left(\partial T/\partial x\right)_j^n}{\left(\partial \bar{T}/\partial x\right)_j^n} = \frac{\sin\left(k\Delta x\right)}{k\Delta x}. \tag{2.46}$$

Offenbar wird die Phase korrekt wiedergegeben. Außerdem wird die Amplitude für $k\Delta x \ll 1$, d.h. für lange Wellen mit $k \ll (\Delta x)^{-1}$, durch die diskrete Formulierung gut wiedergegeben, denn für $\epsilon \ll 1$ gilt: $\sin(\epsilon) \approx \epsilon$. Aber je kürzer die Wellenlänge ist, desto stärker wird die wirkliche Ableitung gemäß (2.46) unterschätzt. Bei der *Nyquist-Wellenzahl* $k_c = 2\pi/\lambda_c = \pi/\Delta x$ ist das Amplitudenverhältnis auf $\sin\pi/\pi = 0$ abgesunken!

Für die zweite Ableitung ergibt sich ein ähnliches Bild. Zentrale Differenzen zweiter Ordnung liefern

$$\frac{\left(\partial^2 T/\partial x^2\right)_j^n}{\left(\partial^2 \bar{T}/\partial x^2\right)_j^n} = \left(\frac{\sin\left(k\Delta x/2\right)}{k\Delta x/2}\right)^2. \tag{2.47}$$

Unsymmetrische Differenzen liefern außerdem noch einen *Phasenfehler*. Auch Verfahren höherer Ordnung ergeben für kleine Wellenlängen Fehler derselben Größenordnung wie bei den hier betrachteten Verfahren niedriger Ordnung.

[13] Es ist

$$\cos\alpha - \cos\beta = -2\sin\left(\frac{\alpha+\beta}{2}\right)\sin\left(\frac{\alpha-\beta}{2}\right).$$

3. Theoretischer Hintergrund

Idealerweise möchte man verlangen, daß die numerische Approximation eines Problems im Limes eines unendlich feinen Gitters (in Raum und Zeit) gegen die exakte Lösung konvergiert. Die Frage ist nun, welche Forderungen an einen Algorithmus zu stellen sind, damit diese *Konvergenz* eintritt. Leider ist die Konvergenz eines Verfahrens für die typischen Gleichungen der Strömungsmechanik (z.B. die Navier-Stokes-Gleichungen) nicht direkt zu beweisen. Die Mindestanforderungen, die man aber stellen muß, damit die numerische Lösung gegen die exakte Lösung konvergiert, sind die *Konsistenz* der Diskretisierung und die *Stabilität* des Algorithmus.

Eine Formulierung ist konsistent, wenn man aus der diskreten Formulierung der Differentialgleichung in Umkehrung des Diskretisierungsprozesses die Differentialgleichung zurückgewinnen kann. Stabilität bedeutet, daß kleine Fehler in den Anfangsbedingungen, Diskretisierungs- oder Rundungsfehler während der Rechnung nicht verstärkt werden sondern zerfallen.

3.1. Konvergenz

Peter David Lax
1926–

Wegen der Nichtlinearitäten in den Differentialgleichungen der Strömungsmechanik sind Konvergenzbeweise im allgemeinen nicht verfügbar. Außerdem sind nur in seltenen Fällen exakte analytische Lösungen bekannt, die es erlauben, den Fehler der numerischen Rechnung exakt zu quantifizieren und damit die Konvergenz auf die exakte Lösung bei Gitterverfeinerung zu untersuchen.

Zumindest für den einfachen Fall eines *linearen* Anfangswertproblems stellt das *Theorem von Lax* die Konvergenz sicher (Richtmyer and Morton, 1967):

> Ein Verfahren ist konvergent, wenn das entsprechende lineare Anfangswertproblem in konsistenter Weise diskretisiert wurde und der Algorithmus stabil ist.

Dieses Theorem gilt nicht nur für finite Differenzen, sondern auch für andere Diskretisierungen wie z.B. für finite Elemente. Die Bedeutung des Theorems besteht darin, daß es relativ einfach ist, Konsistenz und Stabilität zu zeigen, nicht aber direkt die Konvergenz.

Für *lineare* Anfangswertprobleme ist das Lax-Kriterium *hinreichend* für die Konvergenz. Für die uns interessierenden *nichtlinearen* Probleme der Strömungsmechanik liefert es jedoch nur ein *notwendiges* Kriterium. Um die Konvergenz numerisch zu zeigen, kann man jedoch gewisse lokale oder integrale Testgrößen wie in Kap. 2.4 als Funktion von $(\Delta x)^m$ auftragen und damit Hinweise auf Konvergenz oder Divergenz erhalten.[1]

3.2. Konsistenz

Ein System algebraischer Gleichungen ist konsistent, wenn es sich im Limes verschwindender Gitterweite $(\Delta x, \Delta t) \to 0$ auf die zugrundeliegende PDE reduziert. Um eine diskrete Formulierung auf Konsistenz zu prüfen, setzt man formal die *exakte Lösung* in die algebraischen Gleichungen ein und entwickelt alle diskreten Funktionswerte um ein und denselben Punkt in Taylor-Reihen. Als Ergebnis muß dann die ursprüngliche PDE resultieren plus ein Restglied (Fehlerterm), das im Limes verschwindender Gitterweite gegen Null geht.

Dies soll nun für den FTCS-Algorithmus (2.23) demonstriert werden. Setzt man die exakten Werte (gekennzeichnet durch den Querstrich) ein, lautet er

$$\bar{T}_j^{n+1} = s\bar{T}_{j-1}^n + (1-2s)\bar{T}_j^n + s\bar{T}_{j+1}^n. \tag{3.1}$$

Wenn wir alle Größen um den Punkt (x_j, t_n) entwickeln, erhalten wir aus (3.1)

$$\Delta t \left(\frac{\partial \bar{T}}{\partial t}\right)_j^n + \frac{\Delta t^2}{2}\left(\frac{\partial^2 \bar{T}}{\partial t^2}\right)_j^n + \ldots = 2s\left[\frac{\Delta x^2}{2}\left(\frac{\partial^2 \bar{T}}{\partial x^2}\right)_j^n + \frac{\Delta x^4}{4!}\left(\frac{\partial^4 \bar{T}}{\partial x^4}\right)_j^n\right] + \ldots \tag{3.2}$$

Dabei haben wir ausgenutzt, daß der konstante Term $\bar{T}_j^n$ wegfällt, wie auch die ungeraden Ableitungen auf der rechten Seite. Einsetzen von $s = \kappa\Delta t/\Delta x^2$ liefert

$$\left(\frac{\partial \bar{T}}{\partial t}\right)_j^n - \kappa\left(\frac{\partial^2 \bar{T}}{\partial x^2}\right)_j^n + \epsilon_j^n = 0. \tag{3.3}$$

Bis auf den *Abbruchfehler* (*truncation error*)

$$\epsilon_j^n = \frac{\Delta t}{2}\left(\frac{\partial^2 \bar{T}}{\partial t^2}\right)_j^n - \kappa\frac{\Delta x^2}{12}\underbrace{\left(\frac{\partial^4 \bar{T}}{\partial x^4}\right)_j^n}_{\kappa^{-2}\left(\partial^2\bar{T}/\partial t^2\right)} + O\left(\Delta t^2, \Delta x^4\right) \tag{3.4}$$

wird also die Wärmeleitungsgleichung reproduziert. Der FTCS-Algorithmus ist *konsistent*, da der Abbruch-Fehler im Limes $(\Delta x, \Delta t) \to 0$ verschwindet.

Formal ist der Abbruch-Fehler von der Größenordnung $O(\Delta t, \Delta x^2)$. Wir können

[1] Für einige nichtlineare Gleichungen kann man die Stabilität jedoch mit Hilfe der Energie-Methode beweisen (Richtmyer and Morton, 1967).

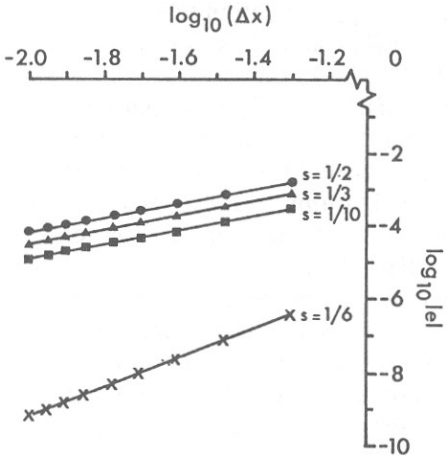


Abbildung 3.1.: Skalierung des numerischen Fehlers e bei der Lösung der eindimensionalen Wärmeleitungsgleichung auf $[0,1]$ mit Anfangswerten $T(x,0) = \sin(\pi x)$ nach Fletcher (1991*a*). Nur für $s = 1/6$ ist die Steigung im log-log-Plot $m = 4$, für alle anderen Werte ist die Steigung $m = 2$. Hier wurde der rms-Wert des Fehlers

$$e = \left[\frac{1}{J} \sum_{j=1}^{J} \left(T_j^n - \bar{T}(x_j, t_n) \right)^2 \right]^{1/2}$$

nach $n = 5000$ Zeitschritten ausgewertet.

ihn aber umformen, wenn wir beachten, daß für die exakte Lösung der Wärmeleitungsgleichung gilt

$$\frac{\partial^2 \bar{T}}{\partial t^2} = \kappa \frac{\partial}{\partial t} \frac{\partial^2 \bar{T}}{\partial x^2} = \kappa \frac{\partial^2}{\partial x^2} \frac{\partial \bar{T}}{\partial t} = \kappa^2 \frac{\partial^4 \bar{T}}{\partial x^4}. \tag{3.5}$$

Damit kann man die 4. Ableitung nach x in (3.4) ersetzen und den führenden Abbruch-Fehler allein durch die zweite Zeitableitung ausdrücken[2]

$$\epsilon_j^n = \left(\frac{\Delta t}{2} - \frac{\Delta x^2}{12\kappa} \right) \left(\frac{\partial^2 \bar{T}}{\partial t^2} \right)_j^n + \ldots = \frac{\Delta x^2}{2\kappa} \left(s - \frac{1}{6} \right) \left(\frac{\partial^2 \bar{T}}{\partial t^2} \right)_j^n + O\left(\Delta t^2, \Delta x^4 \right). \tag{3.6}$$

An dieser Form des Abbruch-Fehlers sieht man, daß der führende Term des Fehlers für $s = 1/6$ verschwindet. Nur für dieses spezielle Schrittweitenverhältnis besitzt das Verfahren den Fehler $O(\Delta t^2)$ in der Zeit und $O(\Delta x^4)$ im Raum. Es ist dann von 2. Ordnung in der Zeit und von 4. Ordnung im Raum. Dies sieht man auch an der Steigung des Fehlers als Funktion der räumlichen Gitterweite Δx im log-log-Plot in Abb. 3.1.

In ähnlicher Weise kann man die Konsistenz der voll impliziten Formulierung (2.6) der Wärmeleitungsgleichung zeigen (Fletcher, 1991*a*). Anders als beim FTCS-Verfahren ist das implizite Verfahren aber immer von erster Ordnung in der Zeit und von zweiter Ordnung im Raum. Es sei betont, daß nicht alle Formulierungen trivialerweise konsistent sind. Um möglichst genaue und stabile Formulierungen zu erreichen, können auch Diskretisierungen resultieren, die möglicherweise inkonsistent sind. Ein Beispiel dafür ist das DuFort-Frankel-Schema (Kap. 6.1.1). Die Konsistenz allein garantiert noch keine Konvergenz. Hierfür muß der Algorithmus außerdem noch stabil sein.

[2] In Fletcher (1991*a*) ist in (4.7) ein Fehler. Es müßte dort heißen α^{-1}.

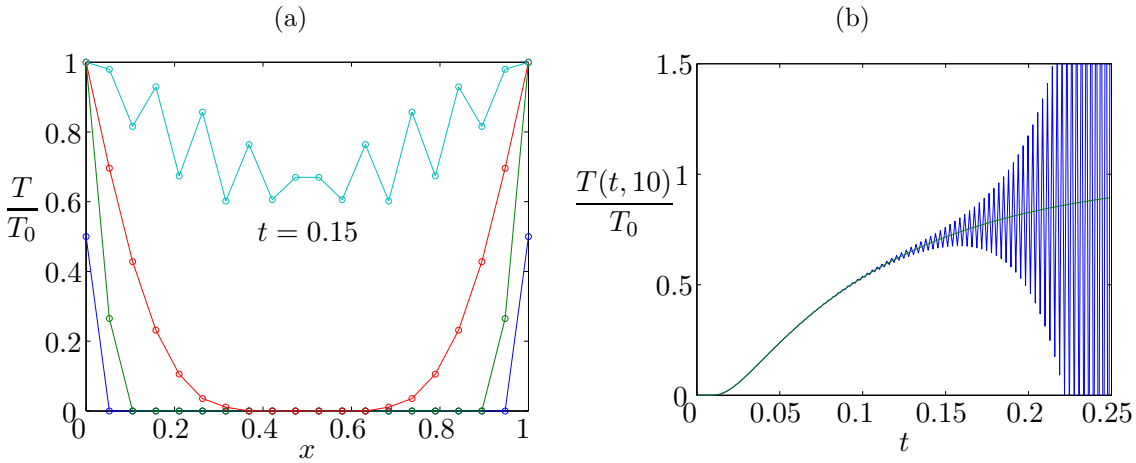


Abbildung 3.2.: Instabilität des FTCS-Algorithmus für die zeitabhängige eindimensionale Wärmeleitungsgleichung mit $s = 0.53$ ($J = 20$) und Anfangs- und Randwerten wie in Abb. 2.4. (a) Räumliches Verhalten der Lösung (vergleiche mit Abb. 2.4 für das stabile Verfahren für $s = 0.50$). Blau: Anfangsbedingung $t = 0$, grün: $t = \Delta t$, rot: $t = 6\Delta t$ und cyan: $t = 94\Delta t \approx 0.15$ ($\kappa = 1$). (b) Zeitliche Entwicklung am Punkt $j = 10$ (blau) im Vergleich zur stabilen Lösung (grün, $s = 0.5$).

3.3. Stabilität

Die Abweichung $T_j^n - \bar{T}(x_j, t_n)$ der numerisch berechneten Lösung T_j^n von der exakten Lösung $\bar{T}(x_j, t_n)$ an den Gitterpunkten x_j und t_n entsteht (a) durch Fehler aufgrund der Diskretisierung und (b) durch Rundungsfehler aufgrund der endlichen Genauigkeit der Zahlendarstellung. Der Diskretisierungsfehler hängt entscheidend von den Gitterweiten $(\Delta x, \Delta t)$ und von den Werten der höheren Ableitungen an den Gitterpunkten ab (siehe Kap. 2.2).

Bei numerischen Berechnungen treten immer spontane Störungen auf, die in der Regel durch *Rundungsfehler* bedingt sind. Diese spontanen Störungen verhindern eine exakte Lösung der mit dem Diskretisierungsfehler behafteten algebraischen Gleichungen. Wenn die spontanen Fehler im Laufe der Rechnung, z.B. bei der zeitlichen Integration, kumulativ signifikant anwachsen, dann ist das numerische Verfahren *instabil*. Falls der kumulative Fehler aber für alle Zeiten vernachlässigbar klein bleibt, ist das Verfahren *stabil*. Ein Beispiel für ein instabiles Verfahren ist der FTCS-Algorithmus für die eindimensionale Wärmeleitungsgleichung mit $s > 0.5$. In diesem Fall hängt die Stabilität vom Verhältnis der Schrittweiten in x- und t-Richtung ab. Dies ist in Abb. 3.2 gezeigt. Die unphysikalischen Oszillationen breiten sich von den Rändern her in das Integrationsgebiet aus und wachsen im Laufe der Zeit exponentiell an.

3.3.1. Matrix-Methode

Um zu entscheiden, ob ein Verfahren numerisch stabil[3] ist oder nicht, muß man untersuchen, ob zufällige Störungen im Laufe der Zeit klein bleiben oder ob sie anwachsen. Dazu schreiben wir die numerisch berechnete Lösung als

$$T_j^n = \tilde{T}_j^n + \xi_j^n, \tag{3.7}$$

wobei $\tilde{T}_j^n$ die *exakte Lösung der algebraischen Gleichungen* bei Abwesenheit von Störungen ist. Durch Rundungsfehler hat man stets Abweichungen ξ_j^n von der exakten Lösung des algebraischen Problems. Wir sind an der Entwicklung dieser Abweichungen (Störungen) interessiert, da ihr Wachstum oder Abklingen über die Stabilität des Verfahrens entscheidet. Dies soll am Beispiel des FTCS-Algorithmus demonstriert werden.

Stabilität des FTCS-Algorithmus

Um zu einer Gleichung zu kommen, welche die Entwicklung der Störung ξ_j^n beschreibt, setzen wir die gesamte numerische Lösung T_j^n in den FTCS-Algorithmus (2.23) ein. Da die Differenzengleichung linear (und homogen) ist, können wir sie schreiben als

$$\mathcal{L}T_j^n = \mathcal{L}\left(\tilde{T}_j^n + \xi_j^n\right) = \underbrace{\mathcal{L}\tilde{T}_j^n}_{=0} + \mathcal{L}\xi_j^n = 0, \tag{3.8}$$

wobei $\mathcal{L}$ der linearer Finite-Differenzen-Operator (in Raum und Zeit) ist. Da die exakte Lösung der algebraischen Gleichungen die Gleichung $\mathcal{L}\tilde{T}_j^n = 0$ erfüllt, muß auch $\mathcal{L}\xi_j^n = 0$ sein. Damit gilt für die Entwicklung der Störungen ξ_j^n dieselbe Gleichung wie für die Lösung T_j^n, nämlich

$$\xi_j^{n+1} = s\xi_{j-1}^n + (1-2s)\xi_j^n + s\xi_{j+1}^n. \tag{3.9}$$

Wenn wir annehmen, daß von den Rändern keine Fehler verursacht werden, dann genügt ξ homogenen Randbedingungen, $\xi_1^n = \xi_J^n = 0$.[4] Wenn wir außerdem den $(J-2)$-dimensionale Vektor der Störungen $\boldsymbol{\xi}^n = (\xi_2^n, \xi_3^n, ..., \xi_{J-2}^n, \xi_{J-1}^n)^{\mathrm{T}}$ an den inneren Stützpunkten zum Zeitpunkt n betrachten, können wir einen Zeitschritt als Matrix-Multiplikation mit

$$\mathsf{A} = \begin{pmatrix} (1-2s) & s & & & & \\ s & (1-2s) & s & & & \\ & \ddots & \ddots & \ddots & & \\ & & s & (1-2s) & s & \\ & & & s & (1-2s) & \end{pmatrix} \tag{3.10}$$

[3] Die *numerische Instabilität* darf man nicht mit der *physikalischen Instabilität* verwechseln.

[4] Für das Lax-Friedrichs-Verfahren für hyperbolische Gleichungen (siehe Fußnote 13 auf Seite 164) ist die Behandlung der Randbedingungen allerdings wichtig.

auffassen. Die Form der $(J-2)\times(J-2)$-dimensionalen Matrix A kann man an (3.9) ablesen. Damit erhalten wir

$$\boldsymbol{\xi}^{n+1} = \mathsf{A}\cdot\boldsymbol{\xi}^{n} = \mathsf{A}\cdot\mathsf{A}\cdot\boldsymbol{\xi}^{n-1} = \ldots = (\mathsf{A})^{n+1}\cdot\boldsymbol{\xi}^{0}. \tag{3.11}$$

Eine *hinreichende Bedingung für Stabilität* ist der monotone Zerfall der Störungen

$$\|\boldsymbol{\xi}^{n+1}\| = \|\mathsf{A}\cdot\boldsymbol{\xi}^{n}\| \leq \|\boldsymbol{\xi}^{n}\|. \tag{3.12}$$

Eine etwas schwächere Bedingung ist der asymptotische Zerfall $\lim_{n\to\infty}\|\boldsymbol{\xi}^{n}\| = 0$. Um diese Bedingung zu untersuchen, benötigen wir die Eigenwerte λ_j der Matrix A. Sie genügen der Eigenwertgleichung

$$\mathsf{A}\cdot\boldsymbol{\psi}_j = \lambda_j\boldsymbol{\psi}_j, \tag{3.13}$$

wobei $\boldsymbol{\psi}_j$ die zu λ_j gehörigen Eigenvektoren sind. Jede beliebige Anfangsstörung

$$\boldsymbol{\xi}^{0} = \sum_{k=1}^{J-2} a_k\boldsymbol{\psi}_k. \tag{3.14}$$

kann man dann als Superposition der Eigenvektoren darstellen. Ein Zeitschritt entspricht dann nur einer Multiplikation des Summanden $\sim\boldsymbol{\psi}_k$ mit λ_k, was nach $n+1$ Zeitschritten auf

$$\boldsymbol{\xi}^{n+1} = \sum_{k=1}^{J-2} (\lambda_k)^{n+1}\, a_k\boldsymbol{\psi}_k. \tag{3.15}$$

führt. Damit $\lim_{n\to\infty}\|\boldsymbol{\xi}^{n}\| = 0$ ist, muß man für alle Eigenwerte verlangen $|\lambda_k| < 1$.[5]

Die Eigenwerte λ_j von A erhält man im Prinzip durch eine Transformation der nichtsingulären Matrix A auf Diagonalform. Dann stehen die Eigenwerte auf der

[5] Wenn man möchte, daß die stärkere Bedingung (3.12) zu *jedem* Zeitschritt erfüllt ist, muß man außerdem verlangen, daß alle Eigenvektoren orthogonal zueinander sind. Wenn die Eigenvektoren nicht orthogonal sind, kann es sein, daß der Betrag (die Norm) der Störung transient (d.h. für eine begrenzte Zeit und abhängig von der Anfangsstörung) anwächst. Der Fehler kann dann zunächst ansteigen und erst später zerfallen. Dies hängt davon ab, ob die Matrix A *normal* oder *nicht-normal* ist. Eine Matrix A ist normal, genau dann wenn $\mathsf{A}^{\dagger}\cdot\mathsf{A} = \mathsf{A}\cdot\mathsf{A}^{\dagger}$, wobei das Kreuz $\dagger$ die adjungierte Matrix (transponierte und konjugiert komplexe Matrix) andeutet. Ist die Matrix nicht normal, dann kann die Länge eines Vektors durch Multiplikation mit A ansteigen, obwohl alle Eigenwerte von A betragsmäßig keiner sind als eins.

Sei zum Beispiel die Anfangsbedingung $\boldsymbol{\xi}^{0} = \boldsymbol{\psi}_1 - \boldsymbol{\psi}_2$ wobei $\boldsymbol{\psi}_1 \approx \boldsymbol{\psi}_2$; d.h. $\|\boldsymbol{\xi}^{0}\| = \delta$, wobei δ sehr klein ist. Wenn nun die Eigenwerte $\lambda_1 \ll \lambda_2$ sehr verschieden sind, dann erhält man nach einem Zeitschritt $\|\boldsymbol{\xi}^{1}\| = \|\lambda_1\boldsymbol{\psi}_1 - \lambda_2\boldsymbol{\psi}_2\| \approx \|\lambda_2\boldsymbol{\psi}_2\| > \delta$ (für hinreichend kleines δ), also ein Wachstum der Störung. Wegen $|\lambda_i| < 1$ gilt trotzdem $\lim_{n\to\infty}\|\boldsymbol{\xi}^{n}\| = \lim_{n\to\infty}\|\lambda_1^n\boldsymbol{\psi}_1 - \lambda_2^n\boldsymbol{\psi}_2\| = 0$.

Für tridiagonale Toeplitz-Matrizen sind die Eigenvektoren dann und nur dann orthogonal, wenn die beiden Nebendiagonalen identisch sind (Noschese et al., 2013). Dies ist hier der Fall. Daher ist A aus (3.10) und (3.11) normal und die Eigenvektoren sind orthogonal. Damit ist die Bedingung (3.12) für $s < 0.5$ erfüllt.

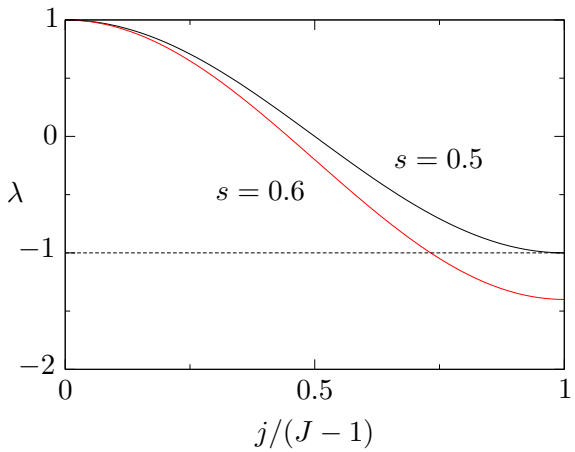


Abbildung 3.3.: Die Eigenwerte der Matrix A (3.10) des Stabilitätsproblems für den FTCS-Algorithmus liegen auf den dargestellten Kurven.

Hauptdiagonalen der transformierten Matrix. Die Eigenwerte von Tridiagonalmatrizen mit konstanten Diagonalen (Toeplitz-Matrix) kann man analytisch berechnen (siehe Anhang A und Abb. 3.3). Für die Eigenwerte der obigen Matrix A gilt demnach

$$\lambda_j = (1-2s) + 2s\cos\left(\frac{j\pi}{J-1}\right), \qquad j = 1, ..., J-2. \tag{3.16}$$

Die Bedingung $|\lambda_j| \leq 1$ lautet daher

$$-1 \leq \overbrace{1 - 2s\underbrace{\left[1-\cos\left(\frac{j\pi}{J-1}\right)\right]}_{\in[0,2]}}^{\lambda_j} \leq 1. \tag{3.17}$$

Die zweite Ungleichheit ist immer erfüllt ($s > 0$). Damit die erste Ungleichheit erfüllt ist, muß gelten

$$s \leq \frac{1}{2}. \tag{3.18}$$

Dies ist die Stabilitätsbedingung für den FTCS-Algorithmus. Da $s \propto \Delta t$ ist, darf also der Zeitschritt nicht zu groß werden. Offensichtlich hat man die größte Wachstumsrate für $j = J-2$. Der zugehörige Eigenvektor reflektiert die räumliche Struktur der anwachsenden Störung. Daraus folgt die Interpretation der gefährlichsten Mode.[6]

[6] Für das Beispiel aus Abb. 3.2 mit $J = 20$ und $s = 0.53$ gibt es zwei Eigenwerte, deren Betrag größer ist als 1, nämlich $\lambda_{J-2} = -1.1055$ und $\lambda_{J-3} = -1.0626$. Das Minuszeichen deutet schon auf den oszillierenden Charakter in der Zeit hin. Der oszillierende Charakter im Raum spiegelt sich in den Eigenvektoren wieder. Der Eigenvektor zu λ_{J-2} lautet

$$\boldsymbol{\psi}_{J-2} = (0.053, -0.105, 0.154, -0.199, 0.239, -0.272, 0.297, -0.315, 0.323, \\ -0.323, 0.315, -0.297, 0.272, -0.239, 0.199, -0.154, 0.105, -0.053)^{\mathrm{T}}.$$

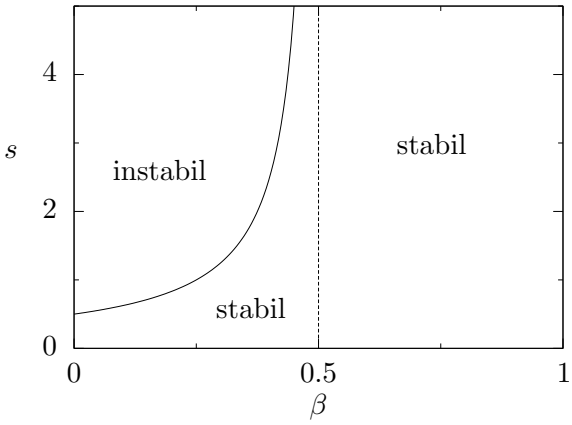


Abbildung 3.4.: Stabilitätsgrenze $s_c(\beta)$ für das semi-implizite Verfahren (3.19) für die Wärmeleitungsgleichung.

Stabilität des allgemeinen 2-Niveau-Schemas für die Wärmeleitungsgleichung

Die Stabilitätsanalyse mittels der Matrix-Methode kann auf das allgemeine (teilimplizite) 2-Niveau-Schema für die Wärmeleitungsgleichung

$$\frac{T_j^{n+1}-T_j^n}{\Delta t}-\beta\kappa\frac{T_{j+1}^{n+1}-2T_j^{n+1}+T_{j-1}^{n+1}}{\Delta x^2}-(1-\beta)\kappa\frac{T_{j+1}^{n}-2T_j^{n}+T_{j-1}^{n}}{\Delta x^2}=0 \quad (3.19)$$

erweitert werden, wobei β ein Maß für den Anteil des Diffusionsterms ist, der implizit behandelt wird ($\beta = 1$: voll implizit, $\beta = 0$: voll explizit). Da dieses teilimplizite Schema, genauso wie das oben behandelte explizite FTCS-Schema, linear und homogen ist, muß das Schema (3.19) auch für den Fehlervektor $\boldsymbol{\xi}$ gelten.

Für die Abweichung $\boldsymbol{\xi}$ von der exakten Lösung $\tilde{T}_j^n$ von (3.19) gilt daher das verallgemeinerte Stabilitätsproblem der Form

$$\mathsf{A}\cdot\boldsymbol{\xi}^{n+1}=\mathsf{B}\cdot\boldsymbol{\xi}^{n}. \quad (3.20)$$

Das Schema (3.19) ist demnach stabil, wenn die Beträge aller Eigenwerte von $\mathsf{A}^{-1}\cdot\mathsf{B}$ kleiner als 1 sind. Wenn man diese Bedingung auswertet (Fletcher, 1991*a*), erhält man:

$$\begin{aligned}\beta<\frac{1}{2} \quad &\Rightarrow \quad (3.19)\text{ ist stabil, falls } s\le\frac{0.5}{1-2\beta},\\ \beta\ge\frac{1}{2} \quad &\Rightarrow \quad (3.19)\text{ ist uneingeschränkt stabil für alle } s.\end{aligned} \quad (3.21)$$

Der Stabilitätsbereich in der (s,β)-Ebene ist in Abb. 3.4 graphisch dargestellt.

Randbedingungen

Die bisherigen Betrachtungen bezogen sich auf Probleme mit Dirichlet-Randbedingungen (T_1^n und T_J^n fest vorgegeben), wobei angenommen wurde, daß die Störung am Rand verschwindet ($\xi_1^n = \xi_J^n = 0$). In Fletcher (1991*a*) werden auch Neumann-Randbedingungen behandelt, bei denen die Ableitung am Rand (entsprechend dem Wärmestrom) vorgegeben ist. Sei zum Beispiel auf der Seite $x = 0$ die

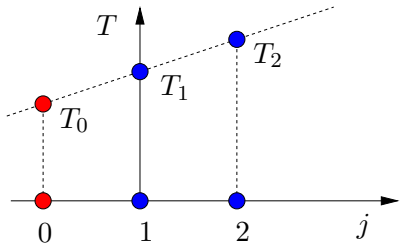


Abbildung 3.5.: Aufprägen eines Wärmestroms bei $j = 1$ durch Anpassung der Temperatur an einem Gitterpunkt $j = 0$ (rot) außerhalb des physikalischen Gebiets.

Randbedingung (konstanter Wärmestrom)

$$\frac{\partial T}{\partial x} = \delta \tag{3.22}$$

vorgegeben. Um diese Bedingung zu erfüllen, kann man einen zusätzlichen Gitterpunkt $j = 0$ außerhalb des eigentlichen Gebiets einführen und verlangen, daß am Rand ($j = 1$) gilt (siehe Abb. 3.5)

$$\left(\frac{\partial T}{\partial x}\right)_1^n = \delta = \frac{T_2^n - T_0^n}{2\Delta x}, \qquad \text{oder} \qquad T_0^n = T_2^n - 2\delta\Delta x. \tag{3.23}$$

Da der aufgeprägte Wärmestrom am Rand vollständig von der *exakten Lösung* der algebraischen Gleichungen erfaßt wird, dürfen die Störungen am Rand ($j = 1$) keinen zusätzlichen Wärmestrom bewirken. Daher muß man fordern: $\xi_0^n = \xi_2^n$. Wenn man nun das explizite FTCS-Verfahren betrachtet und diese Bedingung in die Gleichung für die Entwicklung der Störung (3.9) am Punkt $j = 1$ (der jetzt ein innerer Punkt ist) einsetzt, erhält man

$$\xi_1^{n+1} = s\xi_0^n + (1-2s)\xi_1^n + s\xi_2^n \stackrel{!}{=} (1-2s)\xi_1^n + 2s\xi_2^n. \tag{3.24}$$

Aufgrund dieser zusätzlichen Gleichung für ξ_1^n wird die Ordnung der Matrix A des Stabilitätsproblems um 1 erhöht und sie lautet

$$\mathsf{A} = \begin{pmatrix} (1-2s) & 2s & & & & & \\ s & (1-2s) & s & & & & \\ & s & (1-2s) & s & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & s & (1-2s) & s & \\ & & & & s & (1-2s) \end{pmatrix}. \tag{3.25}$$

Bei der erforderlichen Vorgabe eines verschwindenden Wärmestroms am Rand für die Störungen sind nun die Werte auf den Diagonalen der Tridiagonalmatrix nicht mehr konstant. Durch diese Modifikation ergeben sich allerdings nur geringe Änderungen der Stabilitätseigenschaften (s. Tabelle 4.2 in Fletcher (1991*a*)).[7]

[7] Bei Fletcher (1991*a*) geht in seiner Gleichung (4.28) (zwei-Niveau-Schema) für die Störungen bei $j = 1$ der normierte Wärmestrom am Rand δ ein. Dies entspricht aber nicht der Bedingung

3.3.2. Von-Neumann-Methode

Die gebräuchlichste und einfachste Methode zur Untersuchung der Stabilität ist die *von-Neumann-Methode*. Für *lineare* Anfangswertprobleme mit konstanten Koeffizienten liefert sie eine notwendige und hinreichende Bedingung für die Stabilität. Im allgemeinen Fall *nichtlinearer* Gleichungen und/oder variabler Koeffizienten kann die Neumann-Methode nur lokal bzw. mit eingefrorenen Koeffizienten verwendet werden. Sie liefert dann zwar eine notwendige, aber keine hinreichende Bedingung für die Stabilität. Diese Einschränkung gilt natürlich auch für die obige Matrix-Methode.

Die *Matrix-Methode* zur Analyse der zeitlichen Entwicklung von Störungen basierte auf den Eigenwerten und Eigenvektoren der Matrix A, die einem Zeitschritt entspricht. Dabei hatten wir die zeitliche Entwicklung einer beliebigen Störung (3.15) in dem Raum betrachtet, der von den Eigenvektoren von A aufgespannt wird. Bei der *von-Neumann-Methode* betrachtet man die zeitliche Entwicklung im örtlichen Fourier-Raum. Dazu wird der Fehler als diskrete Fourier-Reihe dargestellt (vgl. Kap. 2.4.3). Wenn wir das Intervall $x \in [0,1]$ mit J Gitterpunkten im Abstand $\Delta x = 1/J$ betrachten, dann läßt sich eine beliebige Anfangsbedingung ξ_j^0 der Störung zum Zeitpunkt $n = 0$ in Analogie zu (2.36) als *diskrete Fourier-Reihe* darstellen[8]

$$\xi_j^0 = \sum_{m=1}^{J} a_m \mathrm{e}^{\mathrm{i}m\pi \overbrace{(j\Delta x)}^{x_j}} = \sum_{m=1}^{J} a_m \mathrm{e}^{\mathrm{i}\theta_m j}, \qquad \text{für} \quad j = 1, \ldots J. \tag{3.26}$$

Hierbei ist $\theta_m = m\pi\Delta x \in \mathbb{R}$ und $a_m \in \mathbb{C}$ (Amplitude und Phasenlage). Der Ort ist offensichtlich $x_j = j\Delta x$, die Wellenzahl $k = m\pi$ und die Wellenlänge $\lambda = 2\pi/k = 2/m$. Der langwelligste Beitrag ($m = 1$) besitzt die Wellenlänge $\lambda = 2$. Damit paßt genau eine halbe Wellenlänge in das betrachtete Gebiet $[0,1]$. Der kurzwelligste Anteil ($m = J$) besitzt die Wellenlänge $\lambda = 2/J$. Bei der kurzwelligsten Mode ist die Störung daher von Punkt zu Punkt alternierend (Nyquist-Wellenlänge, siehe Abb. 3.6).

Für lineare Gleichungen, wie der Wärmeleitungsgleichung, braucht man nur die zeitliche Entwicklung einer einzigen repräsentativen Fourier-Mode zu betrachten, da die Moden linear unabhängig sind und entkoppeln. Die Wellenzahl geht über den Parameter θ_m ein. Die allgemeine Lösung erhält man durch Superposition aller Fouriermoden. Aus Bequemlichkeit schreiben wir im folgenden den Index m nicht mehr mit und ersetzen $\theta_m \to \theta$, beachten aber, daß $\theta \in [0,\pi]$ sein darf.[9]

---

eines vorgegebenen Wärmestroms. Daher erscheint das von Fletcher (1991*a*) in seinem Kap. 4.3.3 behandelte Problem unphysikalisch. Auch fehlt in seiner Tabelle 4.2 eine Angabe über die Anzahl J der Gitterpunkte.

[8] In (2.36) hatten wir eine 2π-periodische Funktion dargestellt. Hier hat die Funktion die Periode 2 ($x \to \pi x$). Dies ergibt sich aus der Forderung, daß die Funktion auf $[0,1]$ nicht unbedingt periodisch sein muß.

[9] Genauer aus der diskreten Menge $\theta = m\pi\Delta x = m\pi/J$ mit $m = 1, \ldots, J$.

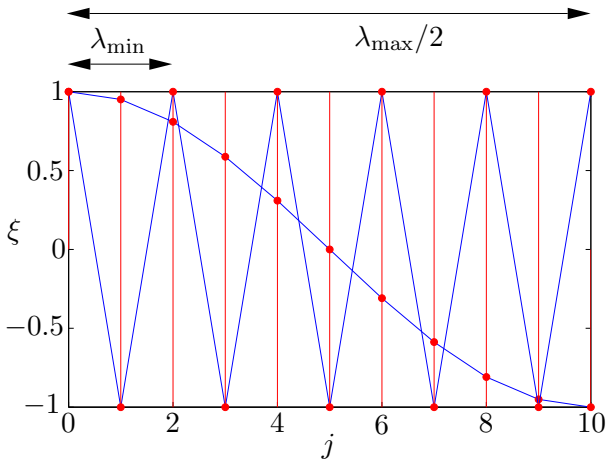


Abbildung 3.6.: Langwelligste und kurzwelligste Mode auf einem Gitter mit $J = 10$ Punkten. Dargestellt sind die Realteile der entsprechenden Summanden in (3.26) mit $a_m = 1$.

Für die zeitliche Entwicklung einer räumlichen Störungsmode $\xi_j^n \sim \mathrm{e}^{\mathrm{i}\theta j}$ machen wir nun den *Separationsansatz*[10]

$$\xi_j^n = G\xi_j^{n-1} = (G)^n\, \xi_j^0 = a\,(G)^n\, \mathrm{e}^{\mathrm{i}\theta j}. \tag{3.27}$$

Die Zeitabhängigkeit steckt im Faktor $(G)^n \in \mathbb{C}$, wobei n in diesem Ausdruck eine Potenz und kein Index ist. Der Faktor G stellt den (komplexen) *Verstärkungsfaktor* bei einem Zeitschritt dar, denn

$$G = \frac{\xi_j^{n+1}}{\xi_j^n}. \tag{3.28}$$

Nach n Zeitschritten hat sich die Amplitude der Mode somit um den Faktor $(G)^n$ verändert (erhöht ($|G| > 1$) oder verringert ($|G| < 1$)). Der Verstärkungsfaktor $G(\theta, s)$ hängt von der Wellenzahl der Fouriermode (bzw. θ) und von etwaigen weiteren Parametern (hier s) ab.

Stabilität des FTCS-Algorithmus mittels Neumann-Methode

Wenn wir die spektrale Komponente (3.27) des Fehlers in das FTCS-Schema der Wärmeleitungsgleichung (2.23) bzw. in die Störungsgleichung (3.9) einsetzen, erhalten wir

$$(G)^{n+1}\, \mathrm{e}^{\mathrm{i}\theta j} = s\,(G)^n\, \mathrm{e}^{\mathrm{i}\theta(j-1)} + (1-2s)\,(G)^n\, \mathrm{e}^{\mathrm{i}\theta j} + s\,(G)^n\, \mathrm{e}^{\mathrm{i}\theta(j+1)}. \tag{3.29}$$

[10] Dieser Ansatz wird auch Normalmoden-Ansatz genannt. Man kann ihn machen, weil wir hier annehmen, daß die Störungsgleichung linear in ξ_j^n ist und konstante Koeffizienten hat, wie in (3.9). Dann bleibt die räumliche Struktur der Fouriermode $\sim \mathrm{e}^{\mathrm{i}\theta j}$ erhalten, und die Zeit faktorisiert (Faktor G pro Zeitschritt).

Nach dem Kürzen durch $(G)^n \, \mathrm{e}^{\mathrm{i}\theta j}$ folgt

$$G = s\mathrm{e}^{-\mathrm{i}\theta} + (1-2s) + s\mathrm{e}^{\mathrm{i}\theta} = (1-2s) + s\left(\mathrm{e}^{\mathrm{i}\theta} + \mathrm{e}^{-\mathrm{i}\theta}\right) = (1-2s) + 2s\cos\theta. \quad (3.30)$$

Dies führt auf auf den Verstärkungsfaktor[11]

$$G(s,\theta) = 1 - 4s\sin^2\left(\frac{\theta}{2}\right). \quad (3.31)$$

Damit der Fehler nicht anwächst, muß gelten: $|G| \leq 1$. Dies ist gleichbedeutend mit

$$-1 \leq 1 - 4s\sin^2\left(\frac{\theta}{2}\right) \leq 1. \quad (3.32)$$

Damit diese Stabilitätsbedingung für alle Wellenzahlen bzw. $\theta \in [0,\pi]$ erfüllt ist, muß $s \leq 1/2$ sein. Die von-Neumann-Stabilitätsbedingung stimmt offenbar mit der oben mittels Matrix-Methode erzielten Bedingung (3.18) exakt überein.

3.4. Numerische Genauigkeit

3.4.1. Richardson-Extrapolation

Die Genauigkeit eines numerischen Schemas wird bei hinreichend hoher Auflösung sehr gut durch den Abbruchfehler (Ordnung des Verfahrens) wiedergegeben. Um eine hohe Genauigkeit zu erzielen, scheinen daher auf den ersten Blick Verfahren hoher Ordnung gegenüber Verfahren niedriger Ordnung von Vorteil zu sein. Ihre Überlegenheit können Verfahren hoher Ordnung jedoch erst dann ausspielen, wenn das Gitter bereits so fein ist, daß auch Verfahren niedriger Ordnung schon typischerweise bis auf $\approx 1\%$ genau sind. Für viele ingenieurmäßige Anwendungen reichen daher Verfahren niedriger Ordnung aus, um mit relativ geringem Aufwand ein hinreichend genaues Ergebnis zu erzielen.

Mit der sogenannten *Richardson-Extrapolation* kann man die Genauigkeit eines Resultats auf einfache Weise verbessern. Die Extrapolation beruht darauf, daß der numerische Fehler bei hinreichend feinem Gitter durch den Abbruchfehler bestimmt ist. Man kann dann die auf verschieden feinen Gittern erzielten Ergebnisse so superponieren, daß sich die beiden Fehlerterme der führenden Ordnung kompensieren.

Wir betrachten zwei numerische Näherungslösungen T_a und T_b (einer nicht notwendigerweise linearen Differentialgleichung) zu den Gitterabständen Δx_a und Δx_b. Für die exakte Lösung gilt dann

$$\bar{T} = T_a + \epsilon_a + \text{h.o.t.}, \quad (3.33a)$$
$$\bar{T} = T_b + \epsilon_b + \text{h.o.t.}, \quad (3.33b)$$

[11] Es ist $\sin^2\alpha = (1-\cos 2\alpha)/2$.

wobei ϵ_a und ϵ_b die führenden Fehlerterme sind (siehe Kap. 2.2). Wir können nun die beiden numerischen Lösungen T_a und T_b so kombinieren, daß der führende Fehler verschwindet. Dazu bilden wir das gewichtete Mittel a(3.33a) + b(3.33b) mit dem Ergebnis

$$(a+b)\,\bar{T} = \underbrace{aT_a + bT_b}_{:=T_{\mathrm{RE}}} + a\epsilon_a + b\epsilon_b + \text{h.o.t.} \tag{3.34}$$

Mit $a + b = 1$ und $T_{\mathrm{RE}} := aT_a + bT_b$ erhalten wir

$$\bar{T} = T_{\mathrm{RE}} + \underbrace{a\epsilon_a + b\epsilon_b}_{\stackrel{!}{=}0} + \text{h.o.t.} \tag{3.35}$$

Der Abbruchfehler setzt sich hier aus zwei Anteilen zusammen. Neben der Bedingung $b = 1 - a$ haben wir noch die Freiheit, a so zu wählen, daß der führende Fehler verschwindet. Dies führt auf

$$a\epsilon_a + (1-a)\epsilon_b = 0 \quad \Rightarrow \quad \begin{aligned} b &= 1 - a, \\ a &= \frac{\epsilon_b}{\epsilon_b - \epsilon_a}. \end{aligned} \tag{3.36}$$

Die führenden Fehler sind bestimmt durch eine Potenz des Gitterabstands und einen *gemeinsamen* Faktor,[12] der von der exakten Ableitung der betreffenden Ordnung abhängt. Der gemeinsame Faktor kürzt sich heraus. Daher kann man a und b allein aus den Gitterabständen bestimmen.

Für den räumlichen Fehler des FTCS-Schemas (siehe (3.6)) gilt zum Beispiel

$$(\epsilon_a)_j^n = \frac{\Delta x_a^2}{2\kappa}\left(s - \frac{1}{6}\right)\left(\frac{\partial^2 \bar{T}}{\partial t^2}\right)_j^n + O\left(\Delta x_a^4\right), \tag{3.37a}$$

$$(\epsilon_b)_j^n = \frac{\Delta x_b^2}{2\kappa}\left(s - \frac{1}{6}\right)\left(\frac{\partial^2 \bar{T}}{\partial t^2}\right)_j^n + O\left(\Delta x_b^4\right). \tag{3.37b}$$

Damit ergibt sich

$$a = \frac{\Delta x_b^2}{\Delta x_b^2 - \Delta x_a^2}, \qquad b = \frac{\Delta x_a^2}{\Delta x_a^2 - \Delta x_b^2}. \tag{3.38}$$

Angenommen, es liegen Rechnungen mit Gitterweiten Δx_a und $\Delta x_b = 2\Delta x_a$ vor. Dann wird $a = 4/3$ und $b = -1/3$, so daß

$$T_{\mathrm{RE}} = \frac{4}{3}T_a - \frac{1}{3}T_b. \tag{3.39}$$

Bei dieser Wahl der Koeffizienten würde also der führende Term des gesamten Abbruchfehlers der Ordnung $O(\Delta x_a^2, \Delta x_b^2)$ von T_{RE} verschwinden. Für ein hinreichend feines Gitter liefert die Richardson-Extrapolation (3.39) für das FTCS-Schema der Wärmeleitungsgleichung damit ein Ergebnis von 4. Ordnung Genauigkeit. Es

[12] Im allgemeinen sind die Faktoren nur nahezu identisch, da die Ableitung noch vom Ort des

ist aber Vorsicht geboten. Denn wenn die Gitter nicht fein genug sind, kann die Richardson-Extrapolation auch eine Verschlechterung des Ergebnisses liefern, verursacht durch die vernachlässigten Terme.

Bei der Richardson-Extrapolation werden lediglich numerische Lösungen superponiert, um Fehlerterme zu eliminieren. Sie ist unabhängig vom Typ der Differentialgleichung. Deshalb kann sie auch für numerische Lösungen nichtlinearer Differentialgleichungen verwendet werden.

Man kann auch mehr als zwei numerische Lösungen zu verschiedenen Gitterweiten additiv überlagern. Die zusätzlichen Wichtungskoeffizienten a, b, c, ... kann man dann derart bestimmen, daß weitere (höhere) Fehlerterme verschwinden.

3.4.2. Numerische Bestimmung der Fehlerordnung

Mit Hilfe numerischer Resultate auf drei verschiedenen Gittern kann man auch die *effektive Ordnung p* eines Verfahrens bestimmen, wenn zum Beispiel ein gestrecktes Gitter verwendet wird. Dazu betrachten wir Gitter mit Gitterweiten h, $2h$ und $4h$. Die exakte Lösung kann dann geschrieben werden als

$$\bar{T} = T_h + \alpha h^p + \text{h.o.t.} \tag{3.40a}$$
$$= T_{2h} + \alpha (2h)^p + \text{h.o.t.} \tag{3.40b}$$
$$= T_{4h} + \alpha (4h)^p + \text{h.o.t.}, \tag{3.40c}$$

wobei $p \in \mathbb{R}$ die unbekannte Fehlerordnung ist. Unter Vernachlässigung der Terme höherer Ordnung hat man mit den drei Gittern die Möglichkeit, die drei Unbekannten $\bar{T}$, α und p zu bestimmen. Man kann leicht überprüfen,[13] daß man für die *Fehlerordnung*

$$p = \frac{\ln\left(\dfrac{T_{4h} - T_{2h}}{T_{2h} - T_h}\right)}{\ln(2)} \tag{3.41}$$

erhält. Die mit Hilfe einer Verdopplung der Zahl der Gitterpunkte aus (3.40a) und (3.40b) verbesserte Lösung ergibt sich dann als

$$T_{\text{RE}} = T_h + \frac{T_h - T_{2h}}{2^p - 1}. \tag{3.42}$$

Für $p = 2$ ist dies konsistent mit (3.39).

3.4.3. Effizienz numerischer Verfahren

Zur Beurteilung der Effizienz eines Verfahrens muß man den Fehler quantifizieren. Man berechnet ihn normalerweise in einer geeigneten *Norm*, meist in der *euklidischen Norm* (l_2-Norm). Als Fehler definiert man die Abweichung des Lösungsvektors

Gitterpunkts abhängt, der für die Auflösungen Δx_a und Δx_b etwas verschieden sein kann.

[13] Eliminiere zuerst $\bar{T}$ und dann α.

$T_j \in \mathbb{R}^J$ von einer Referenzlösung T_j^{ref} in der betreffenden Norm. In der euklidischen Norm lautet der Fehler damit[14]

$$\epsilon = \left\| T_j - T_j^{\text{ref}} \right\|_2 := \left[\sum_{j=1}^{J} \left(T_j - T_j^{\text{ref}} \right)^2 \right]^{1/2} .$$

Die erzielte Genauigkeit kann man als den inversen Fehler ϵ^{-1} definieren. Die Effizienz des numerischen Verfahrens ist die erzielte Genauigkeit ϵ^{-1} im Verhältnis zum numerischen Aufwand. Den numerischen Aufwand kann man durch die CPU-Zeit T_{CPU} ausdrücken. Dabei muß die Zeit für den (langsamen) I/O- Prozeß abgezogen werden. Die Effizienz ist damit $(\epsilon T_{\text{CPU}})^{-1}$.

Für die meisten Probleme stellt sich heraus, daß die Verwendung finiter Differenzen 4. Ordnung oder kubischer finiter Elemente nicht besonders effizient ist. Man sollte jedoch mindestens finite Differenzen 2. Ordnung verwenden, oder finite Elemente mit linearen Ansatzfunktionen. Auf der anderen Seite gibt es Probleme, die eine höhere Genauigkeit als 2. Ordnung erfordern.

Oft ist auch eine grobe Abschätzung des numerischen Aufwands sinnvoll. Dabei bestimmt man zunächst die CPU-Zeit, die für die wesentlichen Operationen (Fließkomma-Operationen, Zuweisungen, etc.) erforderlich sind, multipliziert diese mit der Anzahl der jeweiligen Operationen und bildet die Summe. Damit kann man die zeitkritischen Teile eines Programms ermitteln.

[14] Eine Verallgemeinerung ist die l_p-Norm, die durch $\|x_j\|_p := \left(\sum_{j=1}^{J} \left| x_j^p \right| \right)^{1/p}$ definiert ist.

4. Räumliche Diskretisierung: Gewichtete Residuen

Die Methode der gewichteten Residuen umfaßt eine große Klasse von Verfahren. In diesem Kapitel soll anhand von einfachen Beispielen ein Überblick über die verschiedenen Methoden gewonnen werden. Dabei wird zunächst der Zusammenhang zwischen finiten Volumen, finiten Elementen und spektralen Methoden hergestellt.[1]

Bei der Methode der gewichteten Residuen geht man immer davon aus, daß sich die Lösung — anders als bei den punktweise definierten finiten Differenzen — durch vorgegebene Funktionen approximieren läßt, z.B. in der Form

$$T(x,t) = \sum_{j=1}^{J} a_j(t)\phi_j(x). \tag{4.1}$$

Hierbei sind $\phi_j(x)$ *bekannte Ansatz-Funktionen* (*trial functions*) für die räumliche Abhängigkeit und $a_j(t)$ *unbekannte* zeitabhängige Koeffizienten. Indem man die Lösung in diese Form zwingt, entsteht ein Fehler. Er ist relativ groß, wenn J zu klein ist oder die Menge der Ansatzfunktionen $\{\phi_j(x)\}$ ungünstig gewählt wird. Bei einer geeigneten Wahl der Ansatzfunktionen und hinreichend großem J kann aber eine sehr hohe Genauigkeit erreicht werden.

4.1. Allgemeines Konzept

Ausgangspunkt für ein dreidimensionales zeitabhängiges Problem ist ein Ansatz der Form (hier für eine skalare Funktion T geschrieben)

$$T(x,y,z,t) = T_0(x,y,z,t) + \sum_{j=1}^{J} a_j(t)\phi_j(x,y,z). \tag{4.2}$$

Dabei spaltet man zweckmäßigerweise eine Funktion T_0 ab, die möglichst vielen Rand- und Anfangsbedingungen genügen sollte. Der verbleibenden Teil der Lösung braucht dann oft nur noch homogene Randbedingungen zu erfüllen, was die Wahl des Funktionensystems $\{\phi_j\}$ deutlich erleichtert. Beispiele für Ansatzfunktionen im eindimensionalen Fall sind einfache Polynome (Monome) $\phi_j = x^j$, Chebyshev-

[1] Finlayson (1972) gibt einen Überblick über die verschiedenen Verfahren, aber mehr hinsichtlich einer Approximation mit analytischen Mitteln, weniger mit Bezug zur numerischen Anwendung.

Polynome $\phi_j = T_j(x)$ (siehe Kap. 4.6.2 oder Abramowitz and Stegun, 1972) oder harmonische Funktionen $\phi_j = \sin(j\pi x)$.

Wenn man ein Funktionensystem gewählt hat, möchte man die Koeffizienten $a_j(t)$ so bestimmen, daß der Fehler möglichst klein wird. Um diese Bedingungen zu formulieren, schreiben wir die zu lösende Differentialgleichung in der Form

$$\mathcal{L}\bar{T} = 0, \tag{4.3}$$

wobei der Differentialoperator $\mathcal{L}$ nicht notwendigerweise linear sein muß. $\bar{T}$ ist wieder die exakte Lösung des Problems. Für die eindimensionale Wärmeleitungsgleichung wäre beispielsweise

$$\mathcal{L} = \frac{\partial}{\partial t} - \kappa\frac{\partial^2}{\partial x^2}. \tag{4.4}$$

Wenn wir anstelle von $\bar{T}$ unseren Ansatz T aus (4.2) in (4.3) einsetzen, erhalten wir

$$\mathcal{L}T = R. \tag{4.5}$$

Da die Approximation (4.2) die Differentialgleichung nicht exakt lösen wird, bleibt ein Rest $R \neq 0$ übrig. Man kann den Rest schreiben als $R = \mathcal{L}(T - \bar{T})$. Hieran erkennt man den Zusammenhang zwischen dem Fehler $T - \bar{T}$ und dem *Residuum* R. Das Residuum ist daher ein Maß für den Fehler. Wenn wir erreichen können, daß $R = 0$ wird, dann ist $T = \bar{T}$ und wir haben die exakte Lösung gefunden. Ziel ist es daher, die J Koeffizienten a_j im Ansatz (4.2) so zu bestimmen, daß das Residuum R minimal wird. Dies erfordert J Bedingungen.

Zur Bestimmung der Koeffizienten a_j fordert man deshalb, daß die gewichteten Volumen-Integrale über das Residuum (hier allgemein in drei Raumdimensionen geschrieben)[2]

$$\int_V W_m(\boldsymbol{x})R(\boldsymbol{x},t)\,\mathrm{d}V = 0, \qquad m = 1,\ldots,J \tag{4.6}$$

verschwinden, wobei $\boldsymbol{x} = (x,y,z)^{\mathrm{T}}$. Die Bedingung (4.6) liefert J Gleichungen für die J unbekannten Koeffizienten $a_j(t)$. Falls R nicht von der Zeit abhängig ist, erhalten wir ein algebraisches System zur Bestimmung der Koeffizienten $\{a_j\}$. Andernfalls ergibt sich ein System gewöhnlicher Differentialgleichungen. In der Wahl der *Gewichts-* oder *Testfunktionen* $W_m(\boldsymbol{x})$ ist man zunächst frei. Je nach Wahl von $W_m(\boldsymbol{x})$ erhält man verschiedene Methoden.

4.1.1. Gebietszerlegung (Subdomain Method)

Eine Möglichkeit besteht darin, das betrachtete Raumgebiet V in J Teilgebiete V_m zu zerlegen, die in der Regel nicht überlappen. Dann fordert man (4.6) und wählt

$$W_m(\boldsymbol{x}) = \begin{cases} 1, & \text{falls } \boldsymbol{x} \in V_m, \\ 0, & \text{falls } \boldsymbol{x} \notin V_m. \end{cases} \tag{4.7}$$

[2] Die Abhängigkeit des Residuums R von a_j wird hier nicht extra aufgeschrieben.

Es wird also gefordert, daß der einfache Mittelwert von R in jedem Teilgebiet V_m verschwindet. Die Bestimmungsgleichungen für die Koeffizienten a_j lauten demnach

$$\int_{V_m} R(\boldsymbol{x},t)\,\mathrm{d}V = \int_{V_m} \mathcal{L}T(\boldsymbol{x},t)\,\mathrm{d}V = 0, \qquad m = 1,\ldots,J. \tag{4.8}$$

Diese Methode führt auf die Methode der *finiten Volumen*, die wir später noch genauer betrachten werden. Ein Vorteil dieser Methode besteht darin, daß gewisse Erhaltungssätze (für Masse, Energie, etc.) nicht nur global, sondern auch auf diskreter Ebene in jedem Teilvolumen V_m exakt erfüllt werden können. Dies ist insbesondere bei Strömungen in abgeschlossenen Systemen (Innenströmungen) und bei kompressiblen Strömungen mit Verdichtungsstößen vorteilhaft.

4.1.2. Kollokation

Bei der *Kollokationsmethode* (*collocation method*), die auch Stützpunktverfahren genannt wird, werden als Testfunktionen die *Diracschen δ-Funktionen*

$$W_m(\boldsymbol{x}) = \delta(\boldsymbol{x} - \boldsymbol{x}_m) \tag{4.9}$$

verwendet. Sie haben ihren Peak an J unterschiedlichen Stützpunkten $\boldsymbol{x}_m$. Die δ-Funktionen besitzen die Eigenschaft

$$\int_V R(\boldsymbol{x},t)\delta(\boldsymbol{x} - \boldsymbol{x}_m)\,\mathrm{d}V = R(\boldsymbol{x}_m,t) \overset{(4.6)}{=} 0. \tag{4.10}$$

Die Bedingung (4.6) führt beim Kollokationsverfahren dazu, daß das Residuum an allen J Stützpunkten $\boldsymbol{x}_m$ (Kollokationspunkten) exakt verschwindet. An diesen Stellen ist damit die Differentialgleichung exakt erfüllt. Es ist aber i.a. $T(\boldsymbol{x}_m) \neq \bar{T}(\boldsymbol{x}_m)$.

In der Wahl der Stützpunkte $\boldsymbol{x}_m$ ist man im Prinzip frei. Die Wahl der Stützpunkte ist aber hinsichtlich der Genauigkeit sehr entscheidend. Meist gibt es eine optimale Anordnung von $\boldsymbol{x}_m$, die von der Art der verwendeten Ansatzfunktionen abhängt und im Zusammenhang steht mit der Gauß-Quadratur.[3]

4.1.3. Methode der kleinsten Quadrate

Mit

$$W_m(\boldsymbol{x}) = \frac{\partial R(\boldsymbol{x},t)}{\partial a_m} \tag{4.11}$$

erhält man aus der Bedingung (4.6) für das Residuum

$$\int_V R(\boldsymbol{x},t)\frac{\partial R(\boldsymbol{x},t)}{\partial a_m}\,\mathrm{d}V = \frac{1}{2}\frac{\partial}{\partial a_m}\int_V R^2(\boldsymbol{x},t)\,\mathrm{d}V = 0, \qquad m = 1,\ldots,J. \tag{4.12}$$

[3] In einem gewissen Sinne kann man auch die finiten Differenzen als Kollokationsmethode auffas-

Man sucht also ein Extremum (Minimum) der positiven Größe $\int_V R^2(\boldsymbol{x},t)\,\mathrm{d}V$ hinsichtlich einer Variation aller Koeffizienten a_m. Im Idealfall ist das Minimum dieser Größe gleich Null. Diese Methode bezeichnet man als *Methode der kleinsten Quadrate* (*least-squares method*).

4.1.4. Galerkin-Methode

Boris Grigorjewitsch Galjorkin 1871–1945

Bei der *Galerkin-Methode* sind die Testfunktionen identisch mit den Ansatzfunktionen

$$W_m(\boldsymbol{x}) = \phi_m(\boldsymbol{x}). \tag{4.13}$$

Wenn die Ansatzfunktionen für die jeweiligen Randbedingungen ein vollständiges und orthogonales Funktionensystem darstellen,[4] dann entspricht (4.6) der Projektion des Residuums auf die orthogonalen Komponenten des Funktionenraums, in welchem R existiert. Damit wird das Residuum bezüglich der J Basisfunktionen ϕ_j *orthogonalisiert*.

4.2. Ein einfaches Beispiel

Um die verschiedenen Methoden zu demonstrieren, betrachten wir die einfache eindimensionale Differentialgleichung erster Ordnung auf dem Gebiet $[0,1]$

$$\left(\frac{\mathrm{d}}{\mathrm{d}x} - 1\right)\bar{y}(x) = 0, \tag{4.14}$$

mit der Randbedingung $\bar{y}(0) = 1$. Die exakte Lösung ist $\bar{y}(x) = \mathrm{e}^x$.

Um eine Näherungslösung im Sinne der gewichteten Residuen zu erhalten, setzen

sen.

[4] Ein System von Funktionen $\{\phi_n\}$ ist *orthonormal*, wenn es ein Skalarprodukt zwischen den Funktionen gibt, so daß

$$\langle\phi_n|\phi_m\rangle := \int_V w(\boldsymbol{x})\phi_n^*(\boldsymbol{x})\phi_m(\boldsymbol{x})\,\mathrm{d}V = \delta_{n,m},$$

gilt, wobei * das konjugiert Komplexe bedeutet und $w(\boldsymbol{x})$ eine Gewichtsfunktion ist, die vom jeweiligen Funktionensystem abhängt. Das Produkt $\langle\phi_n|\phi_m\rangle$ läßt sich genauso interpretieren wie das Skalarprodukt zwischen gewöhnlichen Vektoren. Der Raum aller Vektoren kann z.B. durch die kartesischen (orthonormalen) Einheitsvektoren $\boldsymbol{e}_n$ aufgespannt werden. In analoger Weise spannen die Funktionen ϕ_n einen Funktionenraum auf. Damit alle Funktionen zu den gegebenen Randbedingungen dargestellt werden können, muß das Funktionensystem zusätzlich noch vollständig sein. Man spricht dann von einem *vollständigen Orthonormalsystem*. Zum Beispiel ist das System $\{\sqrt{2}\sin(n\pi x)\}$ mit $n \in \mathbb{N}$ ein vollständiges Orthonormalsystem für reelle Funktionen $f(x)$ auf $[0,1]$ mit den Randbedingungen $f(0) = f(1) = 0$.

wir die Lösung in Form einfacher Polynome an

$$y = 1 + \sum_{j=1}^{J} a_j x^j. \tag{4.15}$$

Einsetzen in (4.14) liefert das Residuum

$$R(x) = \sum_{j=1}^{J} j a_j x^{j-1} - \left(1 + \sum_{j=1}^{J} a_j x^j\right) = -1 + \sum_{j=1}^{J} a_j \left(j x^{j-1} - x^j\right). \tag{4.16}$$

Zur Bestimmung der Koeffizienten a_j wird (4.6) verwendet. Die eindimensionale Variante lautet

$$\int_0^1 W_m(x) R(x)\,\mathrm{d}x = 0. \tag{4.17}$$

Für $m \in [1, J]$ ergeben sich J Gleichungen zur Bestimmung der unbekannten Koeffizienten a_j. In unserem Fall sind die Gleichungen linear. Wir können sie daher in kompakter Form schreiben als

$$\mathsf{S} \cdot \boldsymbol{a} = \boldsymbol{d}, \tag{4.18}$$

wobei $\boldsymbol{a} = a_j$ der Vektor der unbekannten Koeffizienten ist.

An dieser Stelle können wir uns für eine Wichtungsmethode entscheiden. Wenn wir die *Galerkin-Methode* mit $W_m = x^{m-1}$ und $m \in [1, J]$ verwenden,[5] erhalten wir

$$\begin{aligned}
\int_0^1 W_m(x) R(x)\,\mathrm{d}x &= \int_0^1 x^{m-1} \left[-1 + \sum_{j=1}^{J} a_j \left(j x^{j-1} - x^j\right)\right] \mathrm{d}x \\
&= -\left[\frac{x^m}{m}\right]_0^1 + \sum_{j=1}^{J} a_j \int_0^1 \left(j x^{m+j-2} - x^{m+j-1}\right) \mathrm{d}x \\
&= -\frac{1}{m} + \sum_{j=1}^{J} a_j \left[j \frac{x^{m+j-1}}{m+j-1} - \frac{x^{m+j}}{m+j}\right]_0^1 \\
&= -\underbrace{\frac{1}{m}}_{d_m} + \sum_{j=1}^{J} \underbrace{\left[\frac{j}{m+j-1} - \frac{1}{m+j}\right]}_{S_{mj}} a_j \stackrel{!}{=} 0.
\end{aligned} \tag{4.19}$$

Dies ist das lineares Gleichungssystem (4.18) mit

$$\mathsf{S} = S_{mj} = \frac{j}{m+j-1} - \frac{1}{m+j} \qquad \text{und} \qquad d_m = \frac{1}{m}. \tag{4.20}$$

[5] Eigentlich sollte man $W_m = x^m$ und $m \in [1, J]$ verwenden. Für $J > 3$ liefern beide Methoden praktisch dasselbe Ergebnis. Die hier verwendete Methode mit $W_m = x^{m-1}$ und $m \in [1, J]$ ist für $J \leq 2$ etwas genauer.

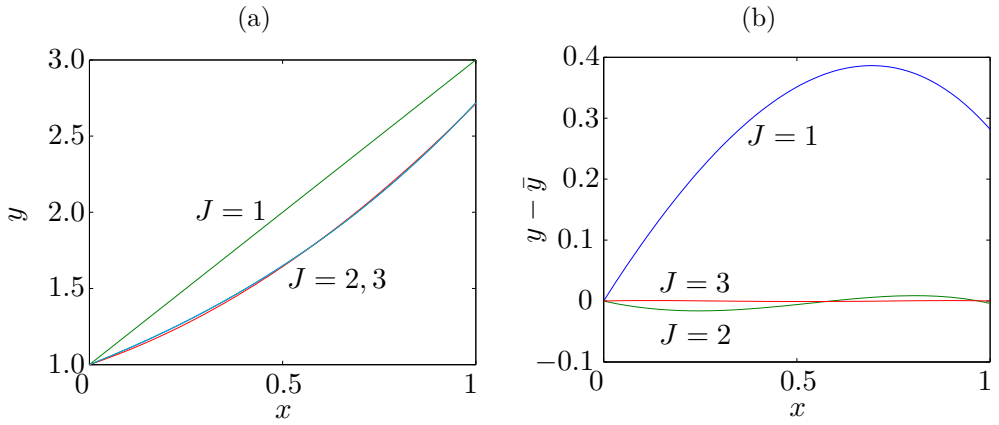


Abbildung 4.1.: (a) Sukzessive Approximationen der Lösung von $(\mathrm{d}/\mathrm{d}x - 1)\bar{y} = 0$ auf dem Intervall $[0, 1]$ mittels Galerkin-Verfahren und dem Ansatz $y = 1 + \sum_{j=1}^{J} a_j x^j$ mit a_j nach (4.21). (b) Fehler $y - \bar{y}$ der verschiedenen Approximationen, als Funktion von x.

Für $J = 1, 2, 3$ lauten die Lösungen von (4.18) (sukzessive verbesserte Näherungen von $\bar{y}$)

$$a^{(J=1)} = 2, \qquad \boldsymbol{a}^{(J=2)} = \begin{pmatrix} 0.8571 \\ 0.8571 \end{pmatrix}, \qquad \boldsymbol{a}^{(J=3)} = \begin{pmatrix} 1.0141 \\ 0.4225 \\ 0.2817 \end{pmatrix}. \tag{4.21}$$

Die Approximationen sind in Abb. 4.1a dargestellt. Mit wachsender Ordnung nimmt der numerische Fehler schnell ab (Abb. 4.1b). Auch das Residuum verringert sich schnell. Da die exakte Lösung und damit auch der Fehler i.a. nicht verfügbar ist, wird normalerweise die Größe des Residuums zur Kontrolle der Konvergenz verwendet.

Die 3×3 Systeme für das Problem (4.14), die man mittels des obigen Galerkin-Verfahren (a), mittels Gebietszerlegung (Drittelung) (b) und mittels Kollokation bei $x = 0$, 0.5, und 1 (c) erhält, lauten

$$\text{(a)} \qquad \begin{pmatrix} 1/2 & 2/3 & 3/4 \\ 1/6 & 5/12 & 11/20 \\ 1/12 & 3/10 & 13/30 \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1/2 \\ 1/3 \end{pmatrix}, \tag{4.22a}$$

$$\text{(b)} \qquad \begin{pmatrix} 5/18 & 8/81 & 11/324 \\ 3/18 & 20/81 & 69/324 \\ 1/18 & 26/81 & 163/324 \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}, \tag{4.22b}$$

$$\text{(c)} \qquad \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 3/4 & 5/8 \\ 0 & 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}. \tag{4.22c}$$

Gebietszerlegung und Kollokation liefern ähnliche Approximationen wie das Galerkin-Verfahren. Die Lösungen (Koeffizienten a_j) sind in Tabelle 4.1 aufgelistet. Der lokale Fehler für die verschiedenen Verfahren ist in Abb. 4.2 dargestellt.

Tabelle 4.1.: Koeffizienten der Näherungslösung $y = 1 + \sum_{j=1}^{J} a_j x^j$ von (4.14) für $J = 3$, die sich aus der Lösung von (4.22) bei Verwendung der verschiedenen Methoden ergeben.

Methode	a_1	a_2	a_3
Galerkin	1.0141	0.4225	0.2817
Gebietszerlegung	1.0156	0.4219	0.2813
Kollokation	1	0.4286	0.2857

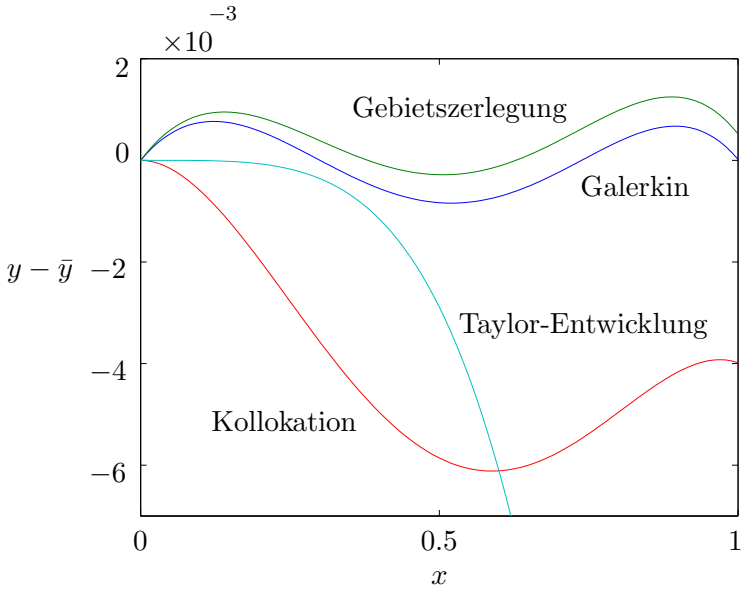


Abbildung 4.2.: Fehler der Lösungen von (4.22), die mittels verschiedener gewichteter Residuen-Verfahren und jeweils drei Ansatzfunktionen ($J = 3$) erzielt wurden. Zusätzlich ist der Fehler der Taylorentwicklung 3. Ordnung um $x = 0$ gezeigt.

Während die gewichten Residuen im gesamten Bereich $[0, 1]$ gute Approximationen liefern, erhält man mit der Taylor-Entwicklung nur in der Nähe des Entwicklungspunkts $x = 0$ gute Ergebnisse. Dort ist sie allerdings unübertroffen gut. Beachte, daß die Kollokationsmethode sensitiv gegenüber der Wahl der Stützpunkte ist.

4.3. Finite Volumen

Die Methode der finiten Volumen ist konzeptionell ähnlich zur oben beschriebenen Gebietszerlegung (Kap. 4.1.1). Zwar integriert man $R = \mathcal{L}T$ (die approximierte Differentialgleichung) wie in (4.8) über ein Teilvolumen V_m, aber man verwendet für T nicht den Ansatz (4.2) mit vorab gewählten Ansatzfunktionen und unbekannten Amplituden. Vielmehr approximiert man die resultierenden Integrale mittels der Funktionswerte von T an gewissen Knotenpunkten, wobei diese Funktionswerte die

zu bestimmenden Größen sind. Die Form und Größe der finiten Volumen können frei gewählt werden. Daher eigenen sie sich sehr gut zur Lösung von Problemen in komplexen Geometrien. Dies ist ein wesentlicher Vorteil gegenüber finiten Differenzen.

Um die Methode der finiten Volumen zu demonstrieren, werden im folgenden Beispiele betrachtet, bei denen die Berandung der finiten Volumen (Gitterlinien) krummlinig sind. Die Berechnung der Lösung erfolgt jedoch in kartesischen Koordinaten.[6]

4.3.1. Gleichungen erster Ordnung

Das Prinzip der finiten Volumen soll an der allgemeinen PDE erster Ordnung in zwei Raumdimensionen

$$\frac{\partial \bar{q}}{\partial t} + \frac{\partial \bar{F}}{\partial x} + \frac{\partial \bar{G}}{\partial y} = 0 \tag{4.23}$$

demonstriert werden. Für den Spezialfall $\bar{q} = \rho$, $\bar{F} = \rho u$ und $\bar{G} = \rho v$ entspricht (4.23) der zweidimensionalen Kontinuitätsgleichung

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{u}) = 0. \tag{4.24}$$

Zunächst wird das gesamte Volumen (hier eine Fläche) in eine (meist große) Anzahl N von kleinen Teilvolumina zerlegt. Eine Integration der PDE über das gesamte Volumen, wobei ein kleines Teilvolumen mit 1 gewichtet wird und das restliche Volumen mit 0, entspricht der einfachen Integration über das zweidimensionale *finite Volumen* (hier über die Fläche S). Wir erhalten dann

$$\int_S \left[\frac{\partial \bar{q}}{\partial t} + \nabla \cdot \begin{pmatrix} \bar{F} \\ \bar{G} \end{pmatrix} \right] \mathrm{d}x\, \mathrm{d}y = 0. \tag{4.25}$$

Anwenden des *Gaußschen Satzes* auf den zweiten Summanden liefert[7]

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_S \bar{q}\, \mathrm{d}x\, \mathrm{d}y + \oint_\Gamma \boldsymbol{n} \cdot \begin{pmatrix} \bar{F} \\ \bar{G} \end{pmatrix} \mathrm{d}s = \frac{\mathrm{d}}{\mathrm{d}t} \int_S \bar{q}\, \mathrm{d}x\, \mathrm{d}y + \oint_\Gamma \begin{pmatrix} \bar{F} \\ \bar{G} \end{pmatrix} \cdot \begin{pmatrix} \mathrm{d}y \\ -\mathrm{d}x \end{pmatrix}$$

[6] Die Beibehaltung der kartesischen Komponenten hat den Vorteil, daß man keine Koordinatentransformation durchführen muß, durch welche meist viele neue Zusatzterme in den Gleichungen entstehen, da bei krummlinigen Koordinaten die Ableitungen der Einheitsvektoren nach den Koordinaten nicht verschwinden, z.B. bei Polarkoordinaten $\partial_\varphi \boldsymbol{e}_r = \boldsymbol{e}_\varphi$.

[7] Mit Hilfe des Gaußschen Satzes wird in drei Dimensionen das Volumenintegral über die Divergenz einer Funktion $\boldsymbol{f}$ in ein Integral über die geschlossenen Oberfläche umgewandelt

$$\int_V \nabla \cdot \boldsymbol{f}\, \mathrm{d}V = \int_S \boldsymbol{n} \cdot \boldsymbol{f}\, \mathrm{d}S = \int_S \boldsymbol{f} \cdot \boldsymbol{n}\, \mathrm{d}S = \int_S \boldsymbol{f} \cdot \mathrm{d}\boldsymbol{S},$$

wobei $\boldsymbol{n}$ der nach außen gerichtete Einheitsvektor senkrecht zur Oberfläche des Volumens ist. In zwei Dimensionen wird ein Flächenintegral über die Divergenz einer Funktion $\boldsymbol{f}$ dann ein

$$= \frac{\mathrm{d}}{\mathrm{d}t} \int_S \bar{q}\,\mathrm{d}x\,\mathrm{d}y + \oint_\Gamma \bar{F}\,\mathrm{d}y - \oint_\Gamma \bar{G}\,\mathrm{d}x = 0, \quad (4.26)$$

wobei Γ die geschlossene Kontur der Fläche bezeichnet und $\boldsymbol{n}$ den nach außen gerichteten Einheitsvektor.[8] Diese Gleichung muß für alle N finiten Volumen erfüllt sein. Damit erhalten wir N Gleichungen mit deren Hilfe wir N Unbekannte bestimmen können. Bei den Unbekannten handelt es sich um die gesuchte Funktion (hier q) an den Knotenpunkten eines Gitters.[9]

Wir überziehen das Gebiet nun mit Gitterlinien, die der Oberfläche des Gesamtgebiets angepaßt sind. In der Regel handelt es sich dabei um ein verzerrtes (nichtkartesisches) Gitter wie in Abb. 4.3 dargestellt, das per Gittergenerierung erzeugt wird. Wir definieren N Unbekannte $q_{j,k}$ an den Knoten des Gitters und wählen als finites Volumen das eingezeichnete rote Viereck mit dem Flächeninhalt $\mathcal{A}$ um einen Knotenpunkt herum.

Nach diesen Festlegungen können wir (4.26) diskretisieren. Das im ersten Summanden von (4.26) auftretende Flächenintegral einer Funktion über die Fläche S ist durch den Mittelwert der Funktion × Flächeninhalt $\mathcal{A}$ gegeben. Wenn man den Mittelwert durch den Wert der Funktion im Zentrum der Fläche approximiert (hier $q_{j,k}$), bezeichnet man dies als *Mittelpunktsregel* (*mid-point rule approximation*).[10]

Die Konturintegrale im zweiten und dritten Summanden lassen sich jeweils durch vier Beiträge darstellen, die aus dem Produkt aus der Länge des Randsegments und dem Mittelwert auf dem Segment bestehen. Damit erhalten wir die Näherung

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathcal{A} q_{j,k}\right) + \sum_{i=AB}^{DA} \left(F_i \Delta y_i - G_i \Delta x_i\right) = 0, \quad (4.27)$$

Integral über die geschlossene Kontur der Fläche

$$\int_S \nabla \cdot \boldsymbol{f}\,\mathrm{d}S = \oint_\Gamma \boldsymbol{f} \cdot \boldsymbol{n}\,\mathrm{d}s.$$

wobei $\mathrm{d}s$ das skalare Bogenelement ist und $\boldsymbol{n}$ der nach außen gerichtete Einheitsvektor senkrecht zur Kontur bzw. senkrecht zum vektoriellen Linienelement $\mathrm{d}\boldsymbol{s}$.

[8] Wenn die Kontur im mathematisch positiven Sinn durchlaufen wird, ergibt sich der nach außen gerichtete Normalenvektor aus dem vektoriellen Linienelement $\mathrm{d}\boldsymbol{s} = (\mathrm{d}x, \mathrm{d}y)^{\mathrm{T}}$ durch Rotation um $\alpha = -90°$

$$\begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \cdot \begin{pmatrix} \mathrm{d}x \\ \mathrm{d}y \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \cdot \begin{pmatrix} \mathrm{d}x \\ \mathrm{d}y \end{pmatrix} = \begin{pmatrix} \mathrm{d}y \\ -\mathrm{d}x \end{pmatrix}.$$

Normierung liefert den Einheitsvektor

$$\boldsymbol{n} = \frac{1}{\sqrt{\mathrm{d}x^2 + \mathrm{d}y^2}} \begin{pmatrix} \mathrm{d}y \\ -\mathrm{d}x \end{pmatrix} = \frac{1}{\mathrm{d}s} \begin{pmatrix} \mathrm{d}y \\ -\mathrm{d}x \end{pmatrix}.$$

Die Orthogonalität $\boldsymbol{n} \cdot \mathrm{d}\boldsymbol{s} = 0$ sieht man sofort.

[9] Es wird angenommen, daß für die anderen abhängigen Variablen F und G jeweils eigene Differentialgleichungen existieren, die in ähnlicher Weise behandelt werden wie die Gleichung für q.

[10] Damit ist $q_{j,k}$ von zweiter Ordnung genau.

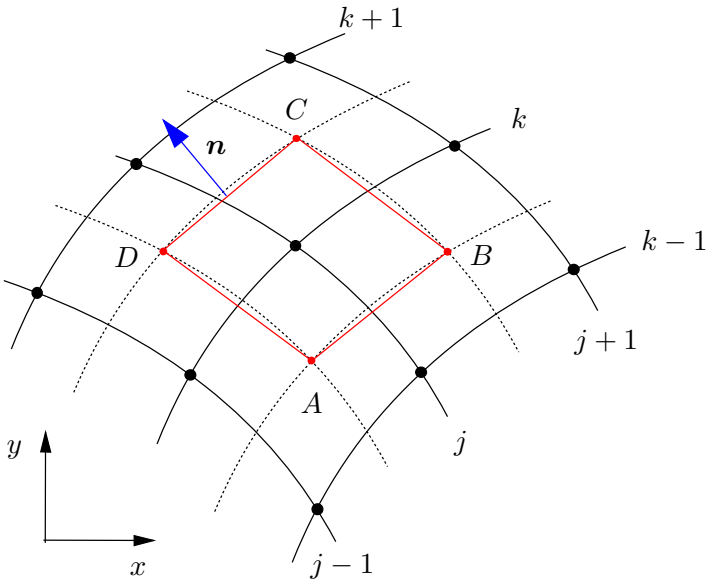


Abbildung 4.3.: Geometrie und Bezeichnungen für zweidimensionale finite Volumen für PDEs erster Ordnung. Die Gitterlinien sind durchgezogen gezeichnet, Hilfslinien gestrichelt. Die Unbekannten sind an den Knotenstellen (j,k) definiert (schwarze Punkte). Das rote Viereck stellt das zum Knoten (j,k) gehörige finite Volumen dar.

mit $\Delta y_{AB} = y_B - y_A$, $\Delta x_{AB} = x_B - x_A$ und den Approximationen für die Mittelwerte $F_{AB} = \frac{1}{2}(F_{j,k-1} + F_{j,k})$, $G_{AB} = \frac{1}{2}(G_{j,k-1} + G_{j,k})$, etc. Wenn wir alle Terme explizit ausschreiben, erhalten wir

$$\begin{aligned}\mathcal{A}\frac{\mathrm{d}q_{j,k}}{\mathrm{d}t} &+ \frac{1}{2}\left(F_{j,k-1} + F_{j,k}\right)\Delta y_{AB} - \frac{1}{2}\left(G_{j,k-1} + G_{j,k}\right)\Delta x_{AB} \\ &+ \frac{1}{2}\left(F_{j,k} + F_{j+1,k}\right)\Delta y_{BC} - \frac{1}{2}\left(G_{j,k} + G_{j+1,k}\right)\Delta x_{BC} \\ &+ \frac{1}{2}\left(F_{j,k} + F_{j,k+1}\right)\Delta y_{CD} - \frac{1}{2}\left(G_{j,k} + G_{j,k+1}\right)\Delta x_{CD} \\ &+ \frac{1}{2}\left(F_{j-1,k} + F_{j,k}\right)\Delta y_{DA} - \frac{1}{2}\left(G_{j-1,k} + G_{j,k}\right)\Delta x_{DA} = 0.\end{aligned} \tag{4.28}$$

Obwohl das Gitter irregulär ist, erhält man auf diese Weise eine Diskretisierung in *kartesischen Koordinaten*. Wenn das Gitter mit den kartesischen Koordinaten übereinstimmt, ist $\Delta y_{AB} = \Delta x_{BC} = \Delta y_{CD} = \Delta x_{DA} = 0$, $\Delta y_{BC} = -\Delta y_{DA} = \Delta y$, $\Delta x_{AB} = -\Delta x_{CD} = \Delta x$ und $\mathcal{A} = \Delta x \Delta y$. Die Gleichung vereinfacht sich dann zu

$$\begin{aligned}\Delta x \Delta y \frac{\mathrm{d}q_{j,k}}{\mathrm{d}t} &- \frac{1}{2}\left(G_{j,k-1} + \underline{G_{j,k}}\right)\Delta x + \frac{1}{2}\left(\underline{F_{j,k}} + F_{j+1,k}\right)\Delta y \\ &+ \frac{1}{2}\left(\underline{G_{j,k}} + G_{j,k+1}\right)\Delta x - \frac{1}{2}\left(F_{j-1,k} + \underline{F_{j,k}}\right)\Delta y = 0.\end{aligned} \tag{4.29}$$

Da sich die unterstrichenen Summanden kompensieren, erhalten wir

$$\frac{\mathrm{d}q_{j,k}}{\mathrm{d}t} + \frac{F_{j+1,k} - F_{j-1,k}}{2\Delta x} + \frac{G_{j,k+1} - G_{j,k-1}}{2\Delta y} = 0. \tag{4.30}$$

Die räumlichen Ableitungen im Rahmen der finiten Volumen sind hier (Mittelpunktsregel) identisch mit zentralen finiten Differenzen.

Die Diskretisierung mittels finiter Volumen hat zwei wichtige *Vorteile*:

1. Die Oberflächenintegrale stellen den jeweiligen diffusiven oder konvektiven Fluß einer Größe dar. Im obigen Beispiel und mit der Massenstromdichte in x-Richtung $\bar{F} = \rho u$ ist $\oint_\Gamma \bar{F}\,\mathrm{d}y$ der Anteil des Massenstroms $\dot{m}$, der das finite Volumen in x-Richtung verläßt.

   Das gewichtete Residuum (4.25) per Gebietszerlegung, d.h. durch Integration, stellt sicher, daß die PDE (Erhaltungsgleichung) im integralen Sinne *für jedes einzelnen finite Volumen* erfüllt ist. Wenn die Oberflächenintegrale zweier benachbarter finiter Volumen identisch gebildet werden, ist der Fluß durch die gemeinsame Grenzfläche aus dem einen Volumen gleich dem Fluß in das andere Volumen hinein. Damit ist die Erhaltungsgleichung auch im gesamten Rechengebiet (gesamtes Volumen) erfüllt. Dies gilt auch für die diskrete Formulierung! Damit ist es z.B. möglich, die Massenerhaltung für jedes einzelne finite Volumen exakt zu gewährleisten.

2. Finite Volumen erlauben eine sehr einfache Diskretisierung auch auf irregulären Gebieten, d.h. auf Gebieten mit komplizierter Berandung.

4.3.2. Gleichungen zweiter Ordnung

Die Diskretisierung einer PDE erster Ordnung war relativ einfach. PDEs zweiter Ordnung sind nicht ganz so einfach im Rahmen allgemeiner finiter Volumen zu diskretisieren. Dies soll am Beispiel der zweidimensionalen *Laplace-Gleichung*

$$\nabla^2 \bar{\phi} = 0 \tag{4.31}$$

demonstriert werden. Integration der Laplace-Gleichung (4.31) über ein finites Volumen und partielle Integration ergibt

$$\int_S \nabla \cdot \nabla \bar{\phi}\,\mathrm{d}S \overset{\text{Gauß}}{=} \oint_\Gamma \boldsymbol{n} \cdot \nabla \bar{\phi}\,\mathrm{d}s = 0. \tag{4.32}$$

In kartesischen Koordinaten erhalten wir mit $\boldsymbol{n}\,\mathrm{d}s = (\mathrm{d}y, -\mathrm{d}x)^{\mathrm{T}}$

$$\oint_\Gamma \left(\frac{\partial \bar{\phi}}{\partial x}\mathrm{d}y - \frac{\partial \bar{\phi}}{\partial y}\mathrm{d}x \right) = 0. \tag{4.33}$$

Im vorherigen Fall einer PDE erster Ordnung hatten wir *Funktionswerte* auf der Kontur des Flächenelements durch Mittelwerte diskreter Werte an den Gitterpunkten ersetzt. Jetzt müssen wir die *ersten Ableitungen* auf der Kontur des betrachteten Flächenelements approximieren.

Wir betrachten zunächst nur das Segment $\overline{AB}$ und approximieren den Mittelwert der Ableitung auf diesem Segment durch den Wert der Ableitung an der benachbarten Zwischengitterstelle (grüner Punkt in Abb. 4.4). Dann erhalten wir für das Segment $\overline{AB}$ (siehe Abb. 4.4)

$$\left[\frac{\partial \phi}{\partial x}\right]_{j,k-1/2} \Delta y_{AB} - \left[\frac{\partial \phi}{\partial y}\right]_{j,k-1/2} \Delta x_{AB}, \tag{4.34}$$

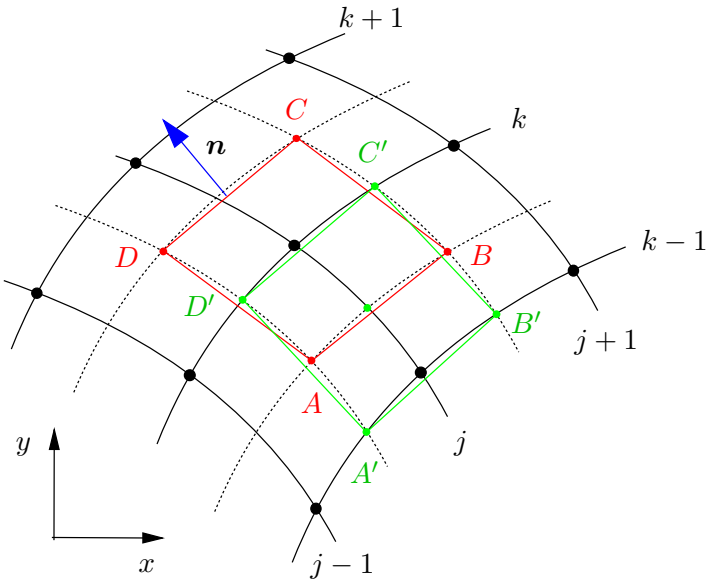


Abbildung 4.4.: Geometrie und Bezeichnungen für zweidimensionale finite Volumen für PDEs zweiter Ordnung. Das finite Volumen um den Punkt (j,k) ist rot gezeichnet. Zur Berechnung der Ableitung im Mittelpunkt $(j,k-\frac{1}{2})$ der Strecke $\overline{AB}$ wird der Mittelwert über der grünen Fläche verwendet.

und entsprechende Beiträge von den anderen drei Segmenten.

Man kann die Ableitung in der Mitte der Strecke $\overline{AB}$ auf verschiedene Arten berechnen. Hier werden wir sie durch Mittelwerte über Flächen approximieren. Wir bezeichnen die grün konturierte Fläche in Abb. 4.4 mit $S_{AB} := S_{A'B'C'D'}$. Dann approximieren wir die Ableitungen über die Mittelwerte

$$\left[\frac{\partial\phi}{\partial x}\right]_{j,k-1/2} = \frac{1}{S_{AB}}\int_{S_{AB}} \frac{\partial\phi}{\partial x}\,\mathrm{d}x\,\mathrm{d}y \overset{(*)}{=} \frac{1}{S_{AB}}\oint_{\Gamma_{AB}} \phi\,\mathrm{d}y, \tag{4.35a}$$

$$\left[\frac{\partial\phi}{\partial y}\right]_{j,k-1/2} = \frac{1}{S_{AB}}\int_{S_{AB}} \frac{\partial\phi}{\partial y}\,\mathrm{d}x\,\mathrm{d}y \overset{(*)}{=} -\frac{1}{S_{AB}}\oint_{\Gamma_{AB}} \phi\,\mathrm{d}x. \tag{4.35b}$$

Beim der Umformung $(*)$ kann man genauso vorgehen wie bei (4.32) und (4.33) und wenn man beachtet, daß in (4.35a) und (4.35b) die Komponenten von $\nabla\phi$ eingehen: $\int_S \nabla\phi\,\mathrm{d}x\mathrm{d}y = \oint_\Gamma \boldsymbol{n}\,\phi\,\mathrm{d}s$ (siehe Fußnote 7 auf S. 62). Alternativ dazu kann man auch den *Stokesschen Satz* verwenden.[11]

Die resultierenden geschlossenen Linienintegrale sind über die Kontur $\Gamma_{AB} := \Gamma_{A'B'C'D'}$ (grün) der Fläche S_{AB} zu nehmen und können z.B. durch

$$\oint_{\Gamma_{AB}} \phi\,\mathrm{d}y \approx \phi_{j,k-1}\Delta y_{A'B'} + \phi_B \Delta y_{B'C'} + \phi_{j,k}\Delta y_{C'D'} + \phi_A \Delta y_{D'A'} \tag{4.36}$$

[11] Der Stokessche Satz lautet

$$\int_S \nabla\times\boldsymbol{u}\cdot\mathrm{d}\boldsymbol{S} = \oint_\Gamma \boldsymbol{u}\cdot\mathrm{d}\boldsymbol{x}.$$

Wir betrachten nun eine Kontur Γ einer Fläche in der (x,y)-Ebene mit Flächenelement $\mathrm{d}\boldsymbol{S} = \boldsymbol{e}_z\,\mathrm{d}x\,\mathrm{d}y$. Mit der speziellen Wahl $\boldsymbol{u} = (0,\phi,0)^\mathrm{T}$ ist $\nabla\times\boldsymbol{u} = (\partial\phi/\partial z, 0, \partial\phi/\partial x)^\mathrm{T}$ und wir erhalten

$$\int_S \begin{pmatrix}\partial\phi/\partial z\\ 0\\ \partial\phi/\partial x\end{pmatrix}\cdot\boldsymbol{e}_z\,\mathrm{d}x\,\mathrm{d}y = \int_S \frac{\partial\phi}{\partial x}\,\mathrm{d}x\,\mathrm{d}y \overset{!}{=} \oint_\Gamma \begin{pmatrix}0\\ \phi\\ 0\end{pmatrix}\cdot\begin{pmatrix}\mathrm{d}x\\ \mathrm{d}y\\ \mathrm{d}z\end{pmatrix} = \oint_\Gamma \phi\,\mathrm{d}y.$$

approximiert werden. Wenn das Gitter nur schwach verzerrt ist, gilt für die Strecken $\Delta y_{A'B'} \approx -\Delta y_{C'D'} \approx \Delta y_{AB}$ und $\Delta y_{B'C'} \approx -\Delta y_{D'A'} \approx \Delta y_{k-1,k}$ (entsprechend für Δx) wodurch wir aus (4.35a)–(4.35b) erhalten

$$\left[\frac{\partial\phi}{\partial x}\right]_{j,k-1/2} = \frac{\Delta y_{AB}\left(\phi_{j,k-1}-\phi_{j,k}\right)+\Delta y_{k-1,k}\left(\phi_B-\phi_A\right)}{S_{AB}} \tag{4.37a}$$

$$\left[\frac{\partial\phi}{\partial y}\right]_{j,k-1/2} = -\frac{\Delta x_{AB}\left(\phi_{j,k-1}-\phi_{j,k}\right)+\Delta x_{k-1,k}\left(\phi_B-\phi_A\right)}{S_{AB}}. \tag{4.37b}$$

Außerdem kann man dann die Fläche S_{AB} aus dem Betrag des Kreuzprodukts der Kantenvektoren erhalten

$$S_{AB} \approx \left|\begin{pmatrix}\Delta x_{AB}\\ \Delta y_{AB}\end{pmatrix}\times\begin{pmatrix}\Delta x_{k-1,k}\\ \Delta y_{k-1,k}\end{pmatrix}\right| = \left|\Delta x_{AB}\Delta y_{k-1,k}-\Delta x_{k-1,k}\Delta y_{AB}\right|. \tag{4.38}$$

Wenn man S_{AB} in (4.37) einsetzt, die Beiträge der drei weiteren Kanten von S_{ABCD} berücksichtigt und außerdem die Funktionswerte von ϕ an den Zwischengitterstellen A, B, C und D durch die Mittelwerte der jeweiligen vier nächsten Nachbarn (reguläre Gitterstellen) ausdrückt, dann sind in dem resultierenden Ausdruck nur noch Funktionswerte $\phi_{j,k}$ für ganzzahlige Gitterindizes involviert. Dies führt schließlich auf (siehe Fletcher, 1991*a*)

$$\begin{aligned}
&\frac{1}{4}\left(P_{CD}-P_{DA}\right)\phi_{j-1,k+1}+\left[Q_{CD}+\frac{1}{4}\left(P_{BC}-P_{DA}\right)\right]\phi_{j,k+1}\\
+&\frac{1}{4}\left(P_{BC}-P_{CD}\right)\phi_{j+1,k+1}+\left[Q_{DA}+\frac{1}{4}\left(P_{CD}-P_{AB}\right)\right]\phi_{j-1,k}\\
-&\left(Q_{AB}+Q_{BC}+Q_{CD}+Q_{DA}\right)\phi_{j,k}+\left[Q_{BC}+\frac{1}{4}\left(P_{AB}-P_{CD}\right)\right]\phi_{j+1,k}\\
+&\frac{1}{4}\left(P_{DA}-P_{AB}\right)\phi_{j-1,k-1}+\left[Q_{AB}+\frac{1}{4}\left(P_{DA}-P_{BC}\right)\right]\phi_{j,k-1}\\
+&\frac{1}{4}\left(P_{AB}-P_{BC}\right)\phi_{j+1,k-1} \;=\; 0,
\end{aligned} \tag{4.39}$$

wobei

$$Q_{AB}=\left(\Delta x_{AB}^2+\Delta y_{AB}^2\right)/S_{AB},\quad P_{AB}=\left(\Delta x_{AB}\Delta x_{k-1,k}+\Delta y_{AB}\Delta y_{k-1,k}\right)/S_{AB}, \tag{4.40a}$$

$$Q_{BC}=\left(\Delta x_{BC}^2+\Delta y_{BC}^2\right)/S_{BC},\quad P_{BC}=\left(\Delta x_{BC}\Delta x_{j+1,j}+\Delta y_{BC}\Delta y_{j+1,j}\right)/S_{BC}, \tag{4.40b}$$

$$Q_{CD}=\left(\Delta x_{CD}^2+\Delta y_{CD}^2\right)/S_{CD},\quad P_{CD}=\left(\Delta x_{CD}\Delta x_{k+1,k}+\Delta y_{CD}\Delta y_{k+1,k}\right)/S_{CD}, \tag{4.40c}$$

$$Q_{DA}=\left(\Delta x_{DA}^2+\Delta y_{DA}^2\right)/S_{DA},\quad P_{DA}=\left(\Delta x_{DA}\Delta x_{j-1,j}+\Delta y_{DA}\Delta y_{j-1,j}\right)/S_{DA}. \tag{4.40d}$$

Das System (4.39) ist eine *9-Punkt-Diskretisierung*, die alle 8 ϕ-Werte der nächsten Nachbarn des Punktes (j,k) miteinander verknüpft. Die Gesamtheit aller derartigen Gleichungen (für alle (j,k)) stellt ein lineares Gleichungssystem für die Werte von $\phi_{j,k}$ an allen Gitterpunkten dar. Dieses Gleichungssystem kann man im allgemeinen nicht direkt lösen. Häufig wird ein iteratives Verfahren zur Lösung benutzt, z.B. die sukzessive Überrelaxation (SOR) (*successive over-relaxation*, siehe Kap. 5.2.3 oder die Vorlesung über *Numerische Methoden der Ingenieurwissenschaften*, LVA-Nr. 322.036).

Das Prinzip der *SOR-Iteration* ist folgendes: Im n-ten Iterationsschritt sei die n-te Approximation $\{\phi_{j,k}^{(n)}\}$ der Lösung bekannt. Die Näherung $\{\phi_{j,k}^{(n)}\}$ erfüllt jedoch nicht genau das Gleichungssystem (4.39). Dann erhalten wir für den Punkt (j,k) hoffentlich eine verbesserte Näherung $\{\phi_{j,k}^{(*)}\}$, wenn wir (4.39) formal nach $\phi_{j,k}$ am zentralen Punkt auflösen und $\{\phi_{j,k}^{(*)}\}$ aus den bekannten Schätzwerten für die umgebenden Punkte berechnen

$$\begin{aligned}\phi_{j,k}^{(*)} = &\left(Q_{AB}+Q_{BC}+Q_{CD}+Q_{DA}\right)^{-1} \\ &\times\Bigg\{\frac{1}{4}\left(P_{CD}-P_{DA}\right)\phi_{j-1,k+1}^{(n)}+\left[Q_{CD}+\frac{1}{4}\left(P_{BC}-P_{DA}\right)\right]\phi_{j,k+1}^{(n)} \\ &+\frac{1}{4}\left(P_{BC}-P_{CD}\right)\phi_{j+1,k+1}^{(n)}+\left[Q_{DA}+\frac{1}{4}\left(P_{CD}-P_{AB}\right)\right]\phi_{j-1,k}^{(n)} \\ &+\left[Q_{BC}+\frac{1}{4}\left(P_{AB}-P_{CD}\right)\right]\phi_{j+1,k}^{(n)} \\ &+\frac{1}{4}\left(P_{DA}-P_{AB}\right)\phi_{j-1,k-1}^{(n)}+\left[Q_{AB}+\frac{1}{4}\left(P_{DA}-P_{BC}\right)\right]\phi_{j,k-1}^{(n)} \\ &+\frac{1}{4}\left(P_{AB}-P_{BC}\right)\phi_{j+1,k-1}^{(n)}\Bigg\}.\end{aligned} \tag{4.41}$$

Die verbesserte Näherung wird für alle Gitterpunkte (j,k) berechnet. Dabei wird $\{\phi_{j,k}^{(*)}\}$ zwar eine Verbesserung von $\{\phi_{j,k}^{(n)}\}$ sein, aber immer noch nicht die exakte Lösung. Deshalb muß man in dieser Art weiter iterieren (Jacobi-Iteration).

Es hat sich herausgestellt, daß man eine schnellere Konvergenz erhält, wenn man nicht die genaue Korrektur $\phi_{j,k}^{(*)}-\phi_{j,k}^{(n)}$ zur vorhandenen Approximation $\{\phi_{j,k}^{(n)}\}$ hinzuaddiert, sondern einen etwas größeren Anteil $\omega>1$

$$\phi_{j,k}^{(n+1)} = \phi_{j,k}^{(n)} + \omega\left(\phi_{j,k}^{(*)}-\phi_{j,k}^{(n)}\right). \tag{4.42}$$

Wenn der *Relaxationsparameter* ω klein ist, wird die Iteration nur langsam konvergieren. Ist er zu groß, kann die Iteration divergieren. Das Optimum ist problemabhängig und liegt für die Laplace-Gleichung etwa bei $\omega_{\text{opt}}\approx 1.8$. Es wird also *über*-relaxiert. Für $\phi_{j,k}^{(*)}=\phi_{j,k}^{(n)}$ hat man offensichtlich Konvergenz.

Die Berechnung für $\boldsymbol{u}=(\phi,0,0)^{\mathrm{T}}$ liefert den Ausdruck $-\int_\Gamma \phi\,\mathrm{d}x$.

Es sei noch einmal betont, daß die obige Formulierung des Problems eine Formulierung in *kartesischen Koordinaten* ist, obwohl die Gitterpunkte auf irgendwelchen krummlinigen Koordinaten liegen können. Diese Flexibilität hat ihren Preis. Wenn man sie aufgibt, vereinfachen sich die Gleichungen. Wählt man zum Beispiel Gitterpunkte auf äquidistanten kartesischen Koordinaten, dann werden $S_{AB} = S_{BC} = S_{CD} = S_{DA} = \Delta x \Delta y$, wodurch sich Q_{AB} bis P_{DA} vereinfachen, und man erhält

$$\frac{\phi_{j-1,k} - 2\phi_{j,k} + \phi_{j+1,k}}{\Delta x^2} + \frac{\phi_{j,k-1} - 2\phi_{j,k} + \phi_{j,k+1}}{\Delta y^2} = 0. \tag{4.43}$$

Das hier verwendete Finite-Volumen-Verfahren wird bei Verwendung von kartesischen Koordinaten also identisch mit finiten Differenzen 2. Ordnung. Dies bedeutet, daß sich die Größenordnung des Fehlers bei einer Halbierung des Gitterabstands um einen Faktor $(1/2)^2 = 1/4$ verkleinert. Bei Verwendung eines nicht-kartesischen Gitters wird die Genauigkeit zweiter Ordnung in dem Maße verringert, in dem das Gitter verzerrt ist (siehe Kap. 2.4.2). Dies kann bei sehr starken Verzerrungen zu erheblichen Problemen führen.[12]

Die Finite-Volumen-Methode wird sehr häufig bei viskosen Strömungen eingesetzt und auch zur Berechnung reibungsfreier transsonischen Strömungen.

4.4. Finite Elemente

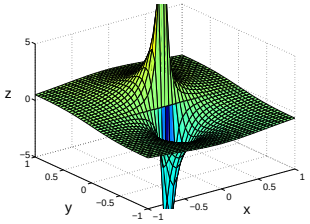
Die Methode der finiten Elemente wurde ursprünglich als ad-hoc Methode zur Berechnung von Spannungen und Verschiebungen in der Strukturmechanik eingeführt. Später wurde eine fundierte mathematische Theorie auf der Basis eines Variationsprinzips der potentiellen Energie entwickelt. Dabei erhält man die Lösung aus einer Gleichgewichtsbedingung, die durch ein *Energie-Minimum* gekennzeichnet ist.

[12] Zur Übung kann das 9-Punkt-Schema (4.39) ähnlich wie in Fletcher (1991*a*) implementiert und mit der analytischen Lösung verglichen werden. Als nicht-kartesisches Volumen wird ein Tortenstück $S = [\epsilon, R] \times [0, \pi/2]$ (Polarkoordinaten) betrachtet. Eine nicht-triviale analytische Lösung der Potentialgleichung in Polarkoordinaten

$$\nabla^2 \bar{\phi} = \left(\frac{1}{r} \frac{\partial}{\partial r} r \frac{\partial}{\partial r} + \frac{1}{r^2} \frac{\partial^2}{\partial \theta^2} \right) \bar{\phi} = 0$$

findet man mit dem Separationsansatz $\bar{\phi} = f(r) \sin(m\theta)$. Es folgt

$$\left(\frac{1}{r} \partial_r r \partial_r - \frac{m^2}{r^2} \right) f(r) = 0,$$



was durch den Potenzansatz $f(r) = r^n$ gelöst wird und auf $n^2 - m^2 = 0$ führt. Wenn wir uns auf $m = 1$ beschränken, liefert die Wurzel $n = 1$ die einfache Lösung $\bar{\phi} = r \sin\theta = y$. Die interessantere Lösung mit einer Singularität bei $r = 0$ entspricht $n = -1$ und lautet $\bar{\phi} = \sin\theta / r = y/r^2$. Sie ist in der nebenstehenden Abbildung dargestellt. Die Aussparung einer ϵ-Umgebung des Ursprungs ist notwendig, um die Divergenz der Lösung bei $r = 0$ auszuklammern. Die analytische Lösung kann man verwenden, um die Randbedingungen auf dem Integrationsgebiet vorzugeben.

Das zugehörige Variationsproblem kann mit Hilfe der Rayleigh-Ritz-Methode gelöst werden (siehe Anhang B).

John William Strutt Lord Rayleigh 1842–1919

Die strömungsmechanischen Gleichungen lassen sich aber im allgemeinen nicht als Variationsproblem formulieren.[13] Eine Alternative ist die *Galerkin-finite-Elemente-Methode*, auf die wir uns im folgenden beschränken werden (Kap. 4.4.3 und 4.4.4). Sie ist in vielen Fällen äquivalent zur Ritz-Methode.

Die Methode der finiten Elemente geht von einer Darstellung der Lösung T eines (hier stationären) Problems aus in der Form

$$T = \sum_{j=1}^{J} T_j \phi_j(x, y, z). \tag{4.44}$$

Hierbei sind die Koeffizienten T_j die gesuchten Funktionswerte an den Gitterpunkten, hier auch *Knoten* genannt. Die Funktionen ϕ_j werden *Test-*, *Ansatz-* oder *Interpolationsfunktionen* genannt, wobei meist Polynome niedriger Ordnung gewählt werden.

Walter Ritz 1878–1909

Außer den *Knoten* gibt es *Elemente*. Elemente sind gewisse kleine Raumgebiete, in denen die Ansatzfunktionen definiert sind. In jedem Element befindet sich eine gewisse (geringe) Anzahl von Knoten. Zu jedem Knoten eines Elements gehört eine Ansatzfunktion, die an dem betreffenden Knoten den Wert 1 annimmt. An allen anderen Knoten desselben Elements nimmt sie den Wert 0 an, und außerhalb des Elements verschwindet sie identisch. Zwischen den Knoten wird die gesuchte Funktion T nun mittels der in dem betreffenden Element definierten Testfunktionen interpoliert. Die Genauigkeit (der Grad) der Interpolation bestimmt die Anzahl der erforderlichen Testfunktionen und damit auch der Knoten in einem Element. Dies wird später an konkreten Beispielen erklärt.[14]

Ein Vorteil dieses Ansatzes finiter Elemente (im Vergleich zur klassischen Galerkin-Methode) besteht darin, daß die Lösung direkt durch die Unbekannten T_j an den Knotenpunkten x_j ausgedrückt wird. Um die finiten Elemente zu verstehen, müssen wir uns zunächst etwas mit der Interpolation beschäftigen.[15]

[13] Es gibt aber etliche Arbeiten in dieser Richtung, siehe z.B. Marner et al. (2019).

[14] Diese Wahl der Testfunktionen führt später zu Matrizen, die nur wenige von Null verschiedene Einträge in der Nähe der Diagonalen haben (diagonal-dominante Matrizen). Das entsprechende lineare Gleichungssystem kann dann ökonomisch gelöst werden.

[15] Siehe auch Kap. 2 der Vorlesung über *Numerische Methoden der Ingenieurwissenschaften*, LVA-Nr. 322.036.

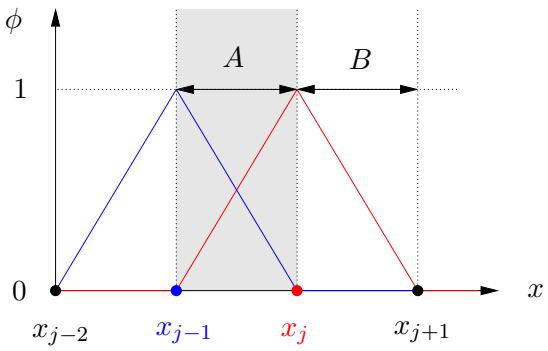


Abbildung 4.5.: Stückweise lineare Testfunktionen ϕ_j (rot) und ϕ_{j-1} (blau). Die Elemente sind mit A, B, usw. bezeichnet.

4.4.1. Eindimensionale Interpolation

Die einfachste Möglichkeit der Approximation besteht in einer linearen Interpolation der gesuchten Funktion zwischen den Knoten. Dazu betrachten wir stückweise lineare Testfunktionen (Abb. 4.5). Die in Abb. 4.5 eingezeichnete Testfunktion ϕ_j um den Knoten x_j (rot) ist definiert durch

$$\phi_j(x) = \begin{cases} 0, & x < x_{j-1}, \\ \dfrac{x - x_{j-1}}{x_j - x_{j-1}}, & x_{j-1} \le x \le x_j, \quad \text{(Element } A\text{)} \\ 1, & x = x_j, \\ \dfrac{x_{j+1} - x}{x_{j+1} - x_j}, & x_j \le x \le x_{j+1}, \quad \text{(Element } B\text{)} \\ 0, & x > x_{j+1}. \end{cases} \tag{4.45}$$

Die Elemente bestehen in diesem Fall aus den Gebieten zwischen je zwei benachbarten Knoten, die in diesem Fall zu *beiden* angrenzenden Elementen gehören. Bei dieser Wahl der Testfunktionen tragen in jedem Element nur zwei Summanden zur Lösung T bei. In den Elementen A und B (Abb. 4.5) sind z.B.

$$\text{in Element } A: \quad T(x) = \sum_{j=1}^{J} T_j \phi_j(x) = T_{j-1}\phi_{j-1}(x) + T_j \phi_j(x), \tag{4.46a}$$

$$\text{in Element } B: \quad T(x) = \sum_{j=1}^{J} T_j \phi_j(x) = T_j \phi_j(x) + T_{j+1}\phi_{j+1}(x). \tag{4.46b}$$

Dabei ist ϕ_j wie oben in (4.45) definiert und in Element A ist

$$\phi_{j-1}(x) = \frac{x_j - x}{x_j - x_{j-1}}, \tag{4.47a}$$

während in Element B gilt

$$\phi_{j+1}(x) = \frac{x - x_j}{x_{j+1} - x_j}. \tag{4.47b}$$

Es ist klar, daß die so definierte Funktion T auf den Rändern des Elements jeweils die Knotenwerte T_j annimmt und dazwischen durch (4.46) linear interpoliert wird.

Bei hinreichend vielen Knoten im Abstand Δx hängt der Fehler einer approximierten Funktion quadratisch ($\sim \Delta x^2$) vom Knotenabstand Δx ab. Dies ist typisch für die lineare Interpolation (für den Fehler sind die quadratischen Terme der Taylor-Entwicklung der Funktion maßgebend).

Die Interpolation mit quadratischen Testfunktionen ist genauer (siehe Fletcher, 1991*a*). Der Fehler ist dann von der Ordnung $O(\Delta x^3)$. Wenn man quadratische Testfunktionen (Polynome) verwenden will, benötigt man drei Punkte, um sie eindeutig festzulegen. Daher muß ein Element bei Verwendung quadratischer Testfunktionen 3 Knoten besitzen. Gleichzeitig benötigen wir auch drei Testfunktionen pro Element. Jede Testfunktion muß an einem Knotenpunkt den Wert 1 annehmen und an den beiden anderen Knotenpunkten verschwinden.

Wir betrachten die Knoten x_{j-1}, x_j und x_{j+1}. Dann ist die quadratische Testfunktion, die bei x_j und x_{j+1} verschwindet, gegeben durch $\phi_{j-1} = c(x - x_j)(x - x_{j+1})$. Die Normierung $\phi_{j-1}(x = x_{j-1}) = 1$ legt die Konstante c fest, so daß

$$\phi_{j-1} = \frac{(x - x_j)}{(x_{j-1} - x_j)} \frac{(x - x_{j+1})}{(x_{j-1} - x_{j+1})}. \tag{4.48a}$$

Entsprechend erhält man die Testfunktionen ϕ_j und ϕ_{j+1}, die bei x_j bzw. x_{j+1} gleich 1 sind und an den jeweils beiden anderen Knoten verschwinden als

$$\phi_j = \frac{(x - x_{j-1})}{(x_j - x_{j-1})} \frac{(x - x_{j+1})}{(x_j - x_{j+1})}, \tag{4.48b}$$

$$\phi_{j+1} = \frac{(x - x_{j-1})}{(x_{j+1} - x_{j-1})} \frac{(x - x_j)}{(x_{j+1} - x_j)}. \tag{4.48c}$$

Diese Testfunktionen sind in Abb. 4.6 dargestellt. Die gesuchte Funktion T wird dann in dem Element, in welchem der Knoten x_j im Zentrum liegt, durch 3 Terme approximiert,

$$T = T_{j-1}\phi_{j-1}(x) + T_j\phi_j(x) + T_{j+1}\phi_{j+1}(x). \tag{4.49}$$

Das Verfahren kann man auch auf kubische und höhere Ansatzfunktionen ausdehnen. Diese sind aber in der Praxis nicht von Bedeutung, da die resultierenden Gleichungen zu komplex werden und nur mit höherem numerischen Aufwand gelöst werden können.[16]

[16] Um die Testfunktionen für den allgemeinen Fall von n Knoten pro Element zu erhalten, kann man *Lagrangesche Interpolationspolynome* verwenden. Man sucht ein Polynom vom Grade $n-1$, das an einem der Knoten x_i ($i \in [1, \ldots, n]$) den Wert 1 annimmt und an den anderen Knoten $k \neq i$ verschwindet. Dazu betrachten wir zunächst das sogenannte fundamentale Polynom vom Grade n, das an *allen* Knoten verschwindet,

$$F_n(x) = (x - x_1)(x - x_2)\ldots(x - x_n).$$

Das *Lagrangesche Interpolationspolynom* Q_i für den Knoten i ergibt sich dann mittels Division

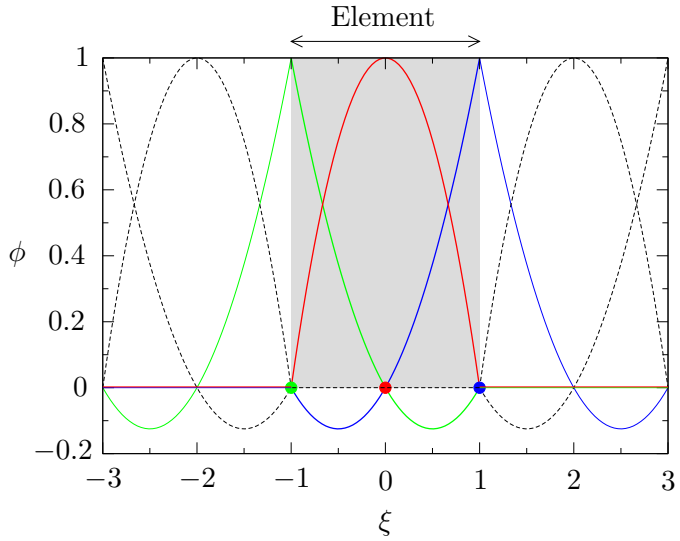


Abbildung 4.6.: Verlauf der quadratischen Testfunktionen innerhalb eindimensionaler Elemente. Die Knoten des grau angedeuteten Elements befinden sich bei $\xi = -1$, 0 und 1. Die Testfunktionen, die zu den drei Knoten dieses Elements gehören, sind farbig kodiert. Die rote Testfunktion zum Konten $\xi = 0$ ist nur in dem grauen Element von Null verschieden, während die beiden anderen Testfunktionen, die zu den Knoten bei $\xi = -1$ (grün) und $\xi = 1$ (blau) gehören, bis in die Nachbarelemente reichen bevor auch sie verschwinden. Die Testfunktionen, die ausschließlich zu den benachbarten Knoten (bei $\xi = \pm 2$ und $\xi = \pm 3$) gehören, sind gestrichelt gezeichnet.

Eine allgemeine Eigenschaft der Testfunktionen ist es, daß ihre Summe in jedem Element gleich 1 ist. Dies sieht man leicht, wenn man die Funktion $f(x) = 1$ darstellt und beachtet, daß die Polynom-Interpolation eindeutig und für jedes Polynom bis zur Ordnung der Interpolationspolynome exakt ist (und damit sicherlich für das Polynom 0-ter Ordnung $f(x) = 1$). Damit gilt

$$1 = f(x) = \sum_{i=1}^{I} f_i \phi_i(x) \overset{f_i \equiv 1}{=} \sum_{i=1}^{I} \phi_i(x). \tag{4.50}$$

Bei den obigen eindimensionalen Elementen mit Knoten auf den Elementgrenzen ist die interpolierte Lösung per constuctionem stetig über die Elementgrenzen hinweg. Wenn man außerdem stetige Differenzierbarkeit der interpolierten Funktionen erreichen will, benötigt man Ansatzfunktionen, die über die Elementgrenzen hinweg stetig differenzierbar sind. Dies läßt sich nur mit Polynomen höheren Grades erreichen.

---

durch den Wurzelfaktor $(x - x_i)$ und Normierung auf 1 (l'Hospital)

$$Q_i(x) = \frac{1}{F_n'(x_i)} \frac{F_n(x)}{x - x_i} = \prod_{\substack{k=0 \\ k \neq i}}^{n} \frac{x - x_k}{x_i - x_k}, \qquad i = 1, \ldots, n.$$

Es verschwindet offenbar an allen Knoten, bis auf den Knoten $k = i$, an dem gilt $Q_i(x = x_i) = 1$.

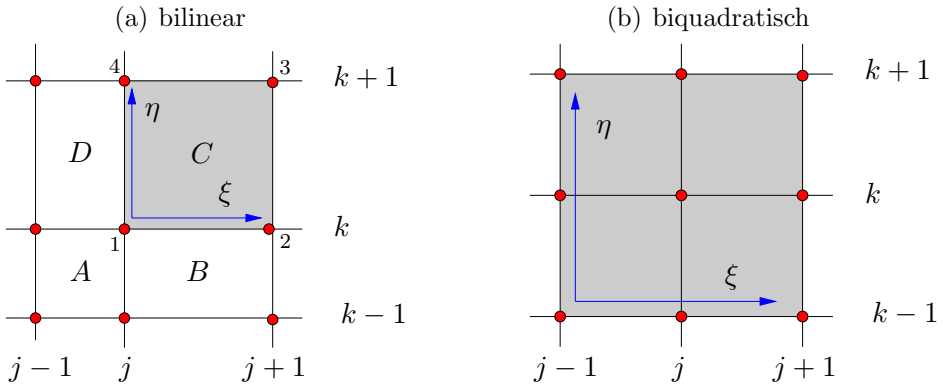


Abbildung 4.7.: (a) Knoten und lokale Koordinaten für vier bilineare Elemente A, B, C und D. Die Numerierung der Knoten bezieht sich auf (4.52). (b) Ein zweidimensionales biquadratisches Element.

4.4.2. Zweidimensionale Interpolation

Bilineare Interpolation Bisher haben wir nur die eindimensionale Interpolation betrachtet. Bei der eindimensionalen linearen Interpolation waren in einem Element zwei Knoten zu berücksichtigen. Für jedes zweidimensionale Element benötigen wir für eine lineare Interpolation 4 Knoten. Wir betrachten nun die Elemente A, B, C und D in Abb. 4.7a. Die gesuchte Funktion wird innerhalb jedes dieser Elemente separat approximiert. In Analogie zum eindimensionalen Fall approximieren wir nun die gesuchte Funktion in jedem Element als

$$T = \sum_{l=1}^{4} T_l \phi_l(\xi, \eta). \tag{4.51}$$

Dabei haben wir in jedem Element die lokalen Koordinaten $(\xi, \eta) \in [-1, 1] \times [-1, 1]$ in gleicher Weise definiert. Der Index l numeriert die Knoten des Elements. In unserem Fall liegen die 4 Knoten bei $(\xi_l, \eta_l) = (\pm 1, \pm 1)$. Die Testfunktionen sollen linear sein in ξ und in η und an allen Knoten bis auf einen verschwinden. Die beiden eindimensionalen Testfunktionen für die Knoten bei $\xi = \pm 1$ lauten $\frac{1}{2}(1 \pm \xi)$. Als bilineare Testfunktionen ergeben sich dann die 4 Produkte dieser Funktionen mit den beiden Funktionen $\frac{1}{2}(1 \pm \eta)$

$$\begin{aligned} \phi_1 &= \frac{1}{4}(1-\xi)(1-\eta), & \phi_2 &= \frac{1}{4}(1+\xi)(1-\eta), \\ \phi_3 &= \frac{1}{4}(1+\xi)(1+\eta), & \phi_4 &= \frac{1}{4}(1-\xi)(1+\eta), \end{aligned} \tag{4.52}$$

Allgemein kann man die bilineare Testfunktion zum Knoten l eines Elements bei $(\xi_l, \eta_l) = (\pm 1, \pm 1)$ darstellen als

$$\phi_l(\xi, \eta) = \frac{1}{4}(1 + \xi_l \xi)(1 + \eta_l \eta). \tag{4.53}$$

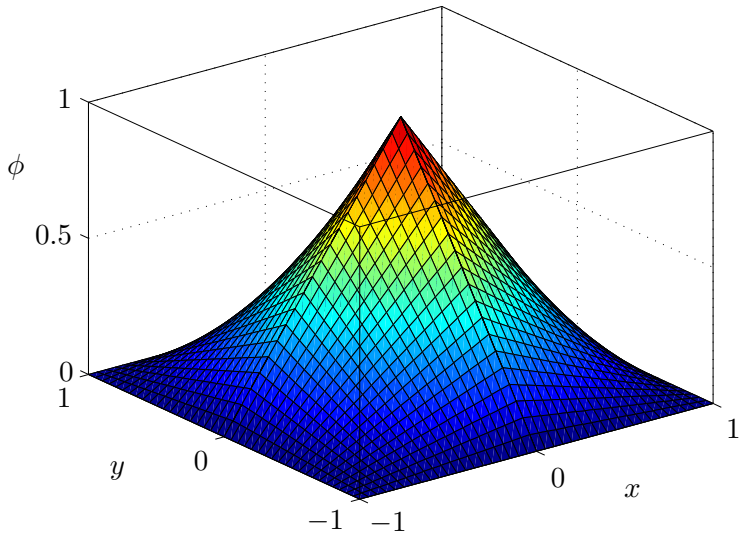


Abbildung 4.8.: Bilineare Testfunktion in der Umgebung eines Knotens (hier bei $x = y = 0$) eines äquidistanten Gitters. Die Basis, auf der die Testfunktion von Null verschieden ist, wird durch 4 Elemente gebildet, die insgesamt 9 Knoten bei $(x, y) = ([0; \pm 1] \times [0; \pm 1])$ enthalten. In jedem Schnitt parallel zu den Koordinatenlinien sind die Interpolationsfunktionen stückweise linear, entlang den Diagonalen sind sie stückweise quadratisch.

Jeder Knoten gehört zu vier Elementen. Die bilineare Testfunktion für einen Knoten besteht daher aus 4 Segmenten in den vier umgebenden Elementen. Zum Beispiel ist die Testfunktion zum Knoten 1 in Abb. 4.7a in den vier Elementen A, B, C und D von Null verschieden. In Abb. 4.8 ist eine bilineare Testfunktion dargestellt. Man sieht, daß die Testfunktionen an den Rändern der Elemente stetig sind, nicht aber ihre ersten Ableitungen.

Biquadratische Interpolation Mit quadratischen Elementen kann man eine höhere Genauigkeit erzielen. Da ein eindimensionales quadratisches Element 3 Knoten umfaßt, benötigt ein zweidimensionales quadratisches Element 9 Knoten (Abb. 4.7b), wobei die gesuchte Funktion in einem Element als

$$T = \sum_{l=1}^{9} T_l \phi_l(\xi, \eta) \tag{4.54}$$

dargestellt wird. Auf dem eindimensionalen Element $\xi \in [-1, 1]$ lauten die eindimensionalen quadratischen Testfunktionen $-(\xi+1)(\xi-1)$, $\frac{1}{2}\xi(\xi-1)$ und $\frac{1}{2}\xi(\xi+1)$. Sie können auch dargestellt werden als

$$(1 - \xi^2) \qquad \text{und} \qquad \frac{1}{2}\xi_l\xi(1 + \xi_l\xi) \quad \text{mit} \quad \xi_l = \pm 1. \tag{4.55}$$

Entsprechendes gilt für die η-Richtung. Wenn man nun alle 9 Produkte der Testfunktionen (4.55) in ξ-Richtung mit den 3 entsprechenden quadratischen Interpolationsfunktionen in η-Richtung bildet, lauten die zweidimensionalen biquadratischen Lagrangeschen Testfunktionen in lokalen Koordinaten $(\xi, \eta) \in [-1, 1] \times [-1, 1]$ innerhalb eines Elements

$$\phi_l(\xi, \eta) = \frac{1}{4}\xi_l\xi\,(1 + \xi_l\xi)\,\eta_l\eta\,(1 + \eta_l\eta)\,, \qquad \text{(4 Eckenknoten)}, \tag{4.56a}$$

$$\phi_l(\xi, \eta) = \frac{1}{2}\left(1 - \xi^2\right)\eta_l\eta\,(1 + \eta_l\eta)\,, \qquad \text{(2 Kantenknoten für } \xi_l = 0), \tag{4.56b}$$

$$\phi_l(\xi, \eta) = \frac{1}{2}\left(1 - \eta^2\right)\xi_l\xi\,(1 + \xi_l\xi)\,, \qquad \text{(2 Kantenknoten für } \eta_l = 0), \tag{4.56c}$$

$$\phi_l(\xi, \eta) = \left(1 - \xi^2\right)\left(1 - \eta^2\right), \qquad \text{(1 zentraler Knoten)}. \tag{4.56d}$$

Die Lagrangesche Interpolation kann in natürlicher Weise auf drei Dimensionen erweitert werden. Werden in zwei Dimensionen 4 Knoten für ein lineares Element und 9 Knoten für ein biquadratisches Element benötigt, so sind es in drei Dimensionen schon 8 Knoten für ein lineares Quader-Element und 27 Knoten für ein triquadratisches Quader-Element.

4.4.3. Eindimensionale Diffusionsgleichung

Die Methode der *Galerkin-finiten-Elemente* soll nun auf die eindimensionale Wärmeleitungsgleichung

$$\frac{\partial \bar{T}}{\partial t} - \kappa\frac{\partial^2 \bar{T}}{\partial x^2} = 0 \tag{4.57}$$

angewandt werden. Der Einfachheit halber verwenden wir lineare Elemente. Wir suchen eine Lösung auf dem Gebiet $x \in [0, 1]$ für $t \geq 0$ zu der Anfangsbedingung $\bar{T}(x, t = 0) = T_0(x)$ und den Randbedingungen $\bar{T}(x = 0, t) = a$ sowie $\bar{T}(x = 1, t) = b$.

Die Lösung wird approximiert in der Form

$$T(x, t) = \sum_{j=1}^{J} T_j(t)\phi_j(x), \tag{4.58}$$

wobei die Zeitabhängigkeit durch die Werte $T_j(t)$ and den nicht äquidistanten Knotenpunkten x_j berücksichtigt wird. Die Testfunktion ϕ_j zum Knotenpunkt x_j ist linear im Element A ($x \in [x_{j-1}, x_j]$) mit Länge $\Delta x_j = x_j - x_{j-1}$ und im Element B ($x \in [x_j, x_{j+1}]$) mit Länge $\Delta x_{j+1} = x_{j+1} - x_j$ (Abb. 4.9). Mit der lokalen Koordinate $\xi \in [-1, 1]$ erhalten wir aus Abb. 4.9 die lineare Beziehung zwischen ξ und x im Intervall $[x_j, x_{j+1}]$ (Element B)

$$x - x_j = \frac{\Delta x_{j+1}}{2}\,(1 + \xi) \qquad \Rightarrow \qquad \xi = \frac{2x - x_j - x_{j+1}}{\Delta x_{j+1}}. \tag{4.59}$$

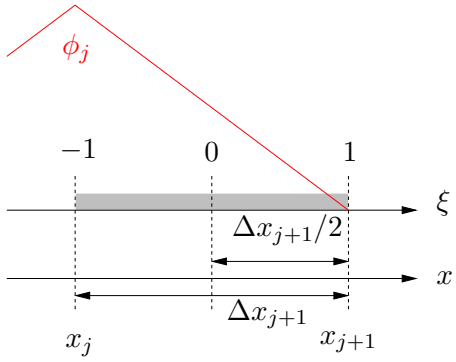


Abbildung 4.9.: Beziehung zwischen der globalen Koordinate x und der lokalen Koordinate ξ zwischen zwei Knoten x_j und x_{j+1} im Abstand Δx_{j+1} innerhalb des Elements B (grau angedeutet).

Die Transformation für Element A ergibt sich aus (4.59) durch $j \to j-1$. Damit können wir die zum Knoten x_j gehörige Testfunktion (rote Kurve in Abb. 4.5 bzw. 4.9) schreiben als

$$\text{Element A, } [x_{j-1}, x_j]: \quad \phi_j(\xi) = \frac{1}{2}(1+\xi), \quad \text{mit } \xi = \frac{2x - (x_{j-1} + x_j)}{\Delta x_j}, \tag{4.60a}$$

$$\text{Element B, } [x_j, x_{j+1}]: \quad \phi_j(\xi) = \frac{1}{2}(1-\xi), \quad \text{mit } \xi = \frac{2x - (x_j + x_{j+1})}{\Delta x_{j+1}}. \tag{4.60b}$$

Wenn wir nun unseren Ansatz (4.58) in die exakte Gleichung einsetzen, erhalten wir

$$\frac{\partial T}{\partial t} - \kappa \frac{\partial^2 T}{\partial x^2} = R, \tag{4.61}$$

mit dem Residuum $R(x,t)$. Dieses Residuum müssen wir minimieren, indem wir die Unbekannten T_j geeignet bestimmen. Dazu wird die Methode der gewichteten Residuen nach (4.6) verwendet. Bei der Galerkin-Methode sind die Gewichtsfunktionen W_m identisch mit den Testfunktionen ϕ_m für alle Knoten. Durch Multiplikation mit den Testfunktionen ϕ_m und Integration über das *gesamte* Volumen erhalten wir somit

$$\begin{aligned}\int_0^1 \phi_m R \,\mathrm{d}x &= \int_0^1 \phi_m \left(\frac{\partial T}{\partial t} - \kappa \frac{\partial^2 T}{\partial x^2} \right) \mathrm{d}x \\ &= \int_0^1 \left(\phi_m \frac{\partial T}{\partial t} + \kappa \frac{\partial \phi_m}{\partial x} \frac{\partial T}{\partial x} \right) \mathrm{d}x - \kappa \left[\phi_m \frac{\partial T}{\partial x} \right]_0^1 \overset{!}{=} 0,\end{aligned} \tag{4.62}$$

wobei die zweite Ableitung durch partielle Integration reduziert wurde.[17]

Da die Werte von $\bar{T}$ am Rand vorgegeben sind, benötigen wir keine Gleichungen

[17] Ganz allgemein gilt für die partielle Integration im $\mathbb{R}^n$

$$\int_V \boldsymbol{v} \cdot \nabla u \,\mathrm{d}\boldsymbol{x}^n = \int_{\partial V} \boldsymbol{v} \cdot \boldsymbol{n} u \,\mathrm{d}\boldsymbol{x}^{n-1} - \int_V u \nabla \cdot \boldsymbol{v} \,\mathrm{d}\boldsymbol{x}^n,$$

wobei $\boldsymbol{v}(\boldsymbol{x}) \in \mathbb{R}^n$ und $u(\boldsymbol{x})$ eine skalare Funktion des n-dimensionalen Vektors $\boldsymbol{x}$ ist, $\mathrm{d}\boldsymbol{x}^n$ das n-dimensionale Volumenelement und $\boldsymbol{n}$ der nach außen zeigende Einheitsvektor auf der Oberfläche

für die Knoten $m = 0$ bei $x = 0$ und für $m = J+1$ bei $x = 1$. Wir betrachten nur die inneren Knoten $m \in [1, J]$. Am Rand $x = 0, 1$ verschwinden die Ansatzfunktionen aller inneren Knoten mit $m \in [1, J]$. Insbesondere gilt $\phi_1(x = 0) = \phi_J(x = 1) = 0$. Damit entfällt der ausintegrierte Anteil in den Gleichungen für die inneren Knoten und es verbleibt

$$\int_0^1 \phi_m \frac{\partial T}{\partial t}\,\mathrm{d}x + \kappa \int_0^1 \frac{\partial \phi_m}{\partial x}\frac{\partial T}{\partial x}\,\mathrm{d}x = 0, \qquad \text{für } 1 \le m \le J. \tag{4.63}$$

Wenn wir nun unseren Ansatz (4.58) einsetzen, erhalten wir

$$\sum_{j=1}^{J} \left(\frac{\partial T_j}{\partial t} \underbrace{\int_0^1 \phi_m \phi_j\,\mathrm{d}x}_{A_{mj}} + \kappa T_j \underbrace{\int_0^1 \frac{\partial \phi_m}{\partial x}\frac{\partial \phi_j}{\partial x}\,\mathrm{d}x}_{B_{mj}} \right) = 0, \qquad \text{für } 1 \le m \le J. \tag{4.64}$$

Dieses System von Gleichungen hat die Form

$$\mathsf{A} \cdot \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{T} + \kappa\,\mathsf{B} \cdot \boldsymbol{T} = 0, \tag{4.65}$$

wobei $\boldsymbol{T} = T_j$ der Vektor der Unbekannten ist. Die Matrizen A und B lauten

$$\mathsf{A} = A_{mj} = \int_0^1 \phi_m \phi_j\,\mathrm{d}x, \tag{4.66a}$$

$$\mathsf{B} = B_{mj} = \int_0^1 \frac{\partial \phi_m}{\partial x}\frac{\partial \phi_j}{\partial x}\,\mathrm{d}x. \tag{4.66b}$$

Fast alle Integrale (Matrixelemente) sind Null. Nur diejenigen Integrale sind von Null verschieden, bei denen sich die Indizes m und j um maximal 1 unterscheiden. Denn andernfalls besitzen die Testfunktionen ϕ_m und ϕ_j bzw. ihre Ableitungen keine Überlappung. Daher sind die Matrizen A und B tridiagonal.

∂V von V. In einer Dimension erhalten wir

$$\int_a^b vu'\,\mathrm{d}x = [vu]_a^b - \int_a^b v'u\,\mathrm{d}x.$$

Ein anderer wichtiger Fall ist der dreidimensionale. Dann erhalten wir aus der allgemeinen Formel

$$\int_V \boldsymbol{v}\cdot\nabla u\,\mathrm{d}V = \int_A \boldsymbol{v}\cdot\boldsymbol{n}u\,\mathrm{d}A - \int_V u\nabla\cdot\boldsymbol{v}\,\mathrm{d}\boldsymbol{V} = \int_A u\boldsymbol{v}\cdot\mathrm{d}\boldsymbol{A} - \int_V u\nabla\cdot\boldsymbol{v}\,\mathrm{d}\boldsymbol{V},$$

wobei A die zweidimensionale Oberfläche des Volumens V ist. Wenn man $u = 1$ setzt, erhält man

$$0 = \int_A \boldsymbol{v}\cdot\mathrm{d}\boldsymbol{A} - \int_V \nabla\cdot\boldsymbol{v}\,\mathrm{d}\boldsymbol{V}.$$

Dies ist nichts anderes als der *Gaußsche Satz*. In zwei Dimensionen folgt der *Stokessche Satz* entsprechend.

Als Beispiel betrachten wir die Diagonalelemente von A, wobei wir die Integration über x auf die Integration über ξ zurückführen. Mit $\mathrm{d}x = (\Delta x_j/2)\mathrm{d}\xi$ für Element A (Element B entsprechend, siehe (4.60)) ergibt sich

$$\begin{aligned} A_{j,j} &= \int_0^1 \phi_j^2(x)\,\mathrm{d}x = \underbrace{\frac{\Delta x_j}{2}\int_{-1}^{1}\phi_j^2(\xi)\,\mathrm{d}\xi}_{\text{Beitrag von Element } A} + \underbrace{\frac{\Delta x_{j+1}}{2}\int_{-1}^{1}\phi_j^2(\xi)\,\mathrm{d}\xi}_{\text{Beitrag von Element } B} \\ &= \frac{\Delta x_j}{8}\int_{-1}^{1}(1+\xi)^2\,\mathrm{d}\xi + \frac{\Delta x_{j+1}}{8}\int_{-1}^{1}(1-\xi)^2\,\mathrm{d}\xi \\ &= \frac{\Delta x_j}{24}\left[(1+\xi)^3\right]_{-1}^{1} - \frac{\Delta x_{j+1}}{24}\left[(1-\xi)^3\right]_{-1}^{1} \\ &= \frac{1}{3}\left(\Delta x_j + \Delta x_{j+1}\right). \end{aligned} \tag{4.67}$$

In gleicher Weise erhält man die Matrixelemente der beiden ersten Nebendiagonalen, so daß insgesamt

$$A_{j,j-1} = \frac{\Delta x_j}{6}, \tag{4.68a}$$

$$A_{j,j} = \frac{1}{3}\left(\Delta x_j + \Delta x_{j+1}\right), \tag{4.68b}$$

$$A_{j,j+1} = \frac{\Delta x_{j+1}}{6}, \tag{4.68c}$$

sowie

$$B_{j,j-1} = -\frac{1}{\Delta x_j}, \tag{4.69a}$$

$$B_{j,j} = \frac{1}{\Delta x_j} + \frac{1}{\Delta x_{j+1}}, \tag{4.69b}$$

$$B_{j,j+1} = -\frac{1}{\Delta x_{j+1}}. \tag{4.69c}$$

Alle anderen Matrix-Elemente verschwinden.

Im Spezialfall eines äquidistanten Gitters ($\Delta x_j = \Delta x = \text{const.}$) nehmen die Diagonalen in A und B konstant und aus (4.65) es ergibt sich das System (nach Division durch Δx und Index-Umbenennung $m \to j$)

$$\frac{1}{6}\left[\frac{\mathrm{d}T}{\mathrm{d}t}\right]_{j-1} + \frac{2}{3}\left[\frac{\mathrm{d}T}{\mathrm{d}t}\right]_{j} + \frac{1}{6}\left[\frac{\mathrm{d}T}{\mathrm{d}t}\right]_{j+1} - \frac{\kappa\left(T_{j-1} - 2T_j + T_{j+1}\right)}{\Delta x^2} = 0. \tag{4.70}$$

Die Zeitableitung, die wir bisher noch nicht diskretisiert haben, wird offenbar an verschiedenen räumlichen Stellen ausgewertet. Die Wichtungsfaktoren sind $\frac{1}{6}$, $\frac{2}{3}$ und $\frac{1}{6}$ für die Knoten bei $j-1$, j und $j+1$. Diese räumliche *Verschmierung* der Zeitableitung ist typisch für lineare finite Elemente auf äquidistanten Gittern. Der

letzte Summand in (4.70) entspricht der zweiten räumlichen Ableitung und hat hier genau dieselbe Gestalt wie bei finiten Differenzen.

Zur Diskretisierung der Zeit kann man die Zeitableitung $\mathrm{d}T/\mathrm{d}t$ durch die diskrete Version ersetzen: $\mathrm{d}T/\mathrm{d}t \to \Delta T^{n+1}/\Delta t$, mit $\Delta T^{n+1} = T^{n+1} - T^n$ und $T^n = T(t_n)$. Wenn man darüber hinaus die zweite räumliche Ableitung als gewichtetes Mittel zu den Zeitpunkten n und $n+1$ ansetzt, erhält man den *teilimpliziten Algorithmus*

$$\begin{aligned} &\frac{1}{6}\left[\frac{\Delta T_{j-1}^{n+1}}{\Delta t}\right] + \frac{2}{3}\left[\frac{\Delta T_{j}^{n+1}}{\Delta t}\right] + \frac{1}{6}\left[\frac{\Delta T_{j+1}^{n+1}}{\Delta t}\right] \\ -\kappa &\left[\beta\frac{T_{j-1}^{n+1} - 2T_j^{n+1} + T_{j+1}^{n+1}}{\Delta x^2} + (1-\beta)\,\frac{T_{j-1}^{n} - 2T_j^{n} + T_{j+1}^{n}}{\Delta x^2}\right] = 0. \end{aligned} \tag{4.71}$$

Der Parameter $\beta \in [0,1]$ gibt den Grad der Implizität an. Der Unterschied dieses Algorithmus zum teilimpliziten Finite-Differenzen-Verfahren (3.19) besteht lediglich darin, daß hier bei der Zeitableitung auch die dem zentralen Knoten j benachbarten Knoten $j \pm 1$ beteiligt sind.

Symbolisch kann man das Gleichungssystem kompakt schreiben als

$$M_x \frac{\Delta T_j^{n+1}}{\Delta t} = \kappa\left[\beta L_{xx} T_j^{n+1} + (1-\beta)\, L_{xx} T_j^n\right], \tag{4.72}$$

mit dem sogenannten *Massenoperator* M_x und dem *Ableitungsoperator* L_{xx}

$$M_x = \left(\frac{1}{6}, \frac{2}{3}, \frac{1}{6}\right), \qquad L_{xx} = \left(\frac{1}{\Delta x^2}, -\frac{2}{\Delta x^2}, \frac{1}{\Delta x^2}\right), \tag{4.73}$$

welche die räumliche Wichtung der ihnen folgenden Terme an den Knoten um den Punkt x_j herum beschreiben. Die Finite-Differenzen-Version (3.19) erhält man demnach mit dem Massenoperator $M_x = (0,1,0)$.

Wenn man $\Delta T^{n+1} = T^{n+1} - T^n$ in (4.72) einsetzt und die Terme zu den Zeitpunkten n und $n+1$ trennt, erhält man die implizite Diskretisierung in der Form

$$(M_x - \Delta t\kappa\beta L_{xx})\, T_j^{n+1} = [M_x + \Delta t\kappa\,(1-\beta)\, L_{xx}]\, T_j^n. \tag{4.74}$$

Um einen Zeitschritt zu berechnen, muß man jeweils ein tridiagonales Gleichungssystem lösen. Dazu kann man den Thomas-Algorithmus verwenden.

Für $\beta = 0$ und finite Differenzen ($M_x = (0,1,0)$) ist dieses System explizit. Das Finite-Elemente-Verfahren ist jedoch wegen der Form des Massenoperators M_x selbst für $\beta = 0$ nicht explizit. Aufgrund der Symmetrien kann man aus (4.74) für $\beta \neq 0$ ein explizites Verfahren erhalten, wenn man den Zeitschritt $\Delta t = \Delta x^2/6\kappa\beta$ wählt. Dann verschwinden nämlich auf der linken Seite von (4.74) die Außerdiagonalterme mit den Indizes $j \pm 1$.

Schließlich kann man leicht zeigen, daß (4.74) eine konsistente Diskretisierung von (4.57) ist. Für $\beta = 0.5$ ist der Abbruchfehler $O(\Delta t^2, \Delta x^2)$. Weiter ist der Algorithmus (4.74) für $\beta \geq 0.5$ uneingeschränkt stabil.

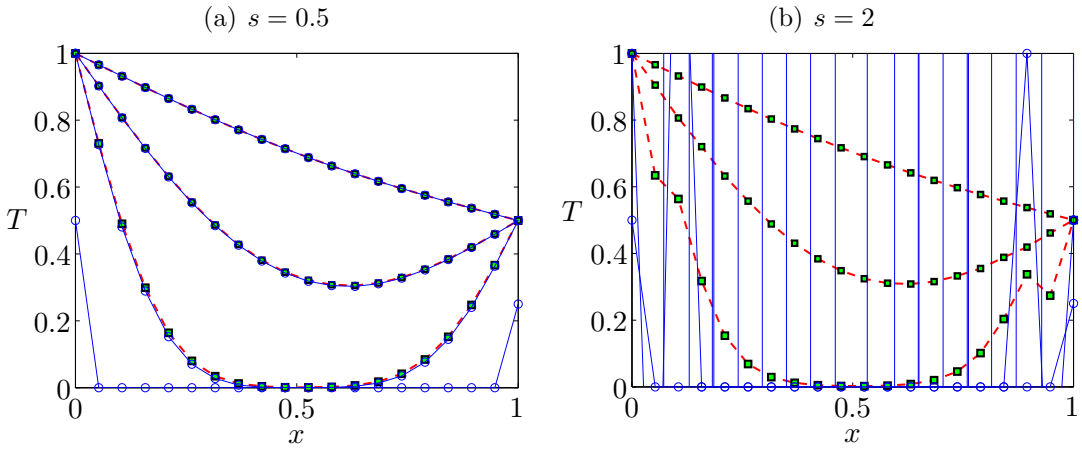


Abbildung 4.10.: Zeitliche Entwicklung der Temperatur in einer Dimension, berechnet mittels FTCS-Verfahren (blaue Linien und Kreise) und mittels des teilimpliziten Galerkin-finite-Elemente-Verfahrens mit $\beta = 0.5$ (rot gestrichelt und grüne Quadrate) nach (4.74) zu gewissen Zeitpunkten. Die Randtemperaturen sind $T(x = 0) = 1$ und $T(x = 1) = 0.5$. Für das FTCS-Verfahren wurde bei Anfangsbedingung $T(t = 0) = 0$ wie in Abb. 2.4 am Rand etwas nachgeholfen (unterste blaue Kurve), während für das finite-Elemente-Verfahren die Randtemperaturen von $T(t = 0) = 0$ plötzlich auf den Endwert erhöht wurden. Gezeigt sind Resultate für die Zeitschrittweiten $s = 0.5$ (a) und $s = 2$ (b). Für $s = 2$ ist das FTCS-Verfahren instabil (senkrechte Streifen in (b)).

Beispielrechnungen sind in Abb. 4.10 gezeigt. Man sieht, daß für die gewählten Parameter die Ergebnisse für den einfachen FTCS-Algorithmus und das teilimplizite finite-Elemente-Verfahren (4.74) fast identisch sind, wenn der Schrittweitenparameter $s = 0.5$ auf der Stabilitätsgrenze des FTCS-Verfahrens liegt. Wird die Zeitschrittweite aber auf $s = 2$ erhöht, ist der FTCS-Algorithmus instabil. Das teilimplizite Verfahren bleibt stabil, obwohl anfänglich gewisse unphysikalische Überschwinger auftreten, die mit der rapiden Änderung der Randwerte von Null auf einen endlichen Wert zusammenhängen. Ansonsten wird die Dynamik und auch der asymptotische Endzustand (in der Abbildung noch nicht ganz erreicht) auch bei der höheren Zeitschrittweite $s = 2$ korrekt wiedergegeben.

Bei der Implementierung von (4.74) ist zu beachten, daß die Gleichung nur für die inneren Punkte des Gebiets gilt. In die erste ($j = 1$) und die letzte Gleichung ($j = J$) von (4.74) gehen die Randwerte T_0 und T_{J+1} ein. Da diese Werte per Randbedingung vorgegeben und bekannt sind, können die entsprechenden Terme der rechten Seite der Gleichung zugeschlagen werden.

4.4.4. Laminare Durchströmung eines Rechteckkanals

Um die Anwendung finiter Elemente auf ein zweidimensionales Problem zu demonstrieren, betrachten wir die laminare stationäre Strömung eines inkompressiblen Fluids durch einen geraden und (unendlich) langen Kanal mit rechteckiger

Querschnittsfläche $2a \times 2b$. Die Randbedingung (Haftbedingung) $\boldsymbol{u}(x, y, z) = 0$ auf $x = \pm a$ und $y = \pm b$ zusammen mit der Navier-Stokes-Gleichung (1.10) erlaubt eine Lösung der Form

$$\boldsymbol{u} = w(x, y)\boldsymbol{e}_z, \tag{4.75a}$$

$$p = p(z). \tag{4.75b}$$

Mit diesem Ansatz verschwindet der nichtlineare konvektive Term ($\boldsymbol{u} \cdot \nabla \boldsymbol{u} = 0$) und wir müssen lediglich das lineare stationäre Problem (z-Komponente der Navier-Stokes-Gleichung)

$$\underbrace{\frac{1}{\rho}\frac{\partial p}{\partial z}}_{f(z)} = \underbrace{\nu \left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} \right)}_{g(x,y)} \tag{4.76}$$

lösen.[18] Physikalisch bedeutet diese Gleichung, daß der durch den Druck verursachte Impulsstrom (p ist eine Impulsstromdichte) in Stromrichtung (z-Richtung) abnimmt, denn $\partial p/\partial z < 0$. Das Defizit des Impulsstroms kommt durch eine seitliche Diffusion des Impulses zu den Rändern zustande, welcher durch die viskosen Terme beschrieben wird.

Da die linke Seite der Gleichung nur von z und die rechte nur von x und y abhängt, müssen beide Seiten der Gleichung konstant sein. Wenn man nun die halben Kanalweiten a und b als Längenskalen verwendet und die Geschwindigkeit w mit Hilfe des konstanten Druckgradienten $\partial p/\partial z$ skaliert

$$x' = \frac{x}{a}, \qquad y' = \frac{y}{b}, \qquad w' = \frac{\rho\nu}{b^2 \partial p/\partial z} w, \tag{4.77}$$

können wir das Problem in dimensionsloser Form als *Poisson-Gleichung* schreiben (der Strich ' an den dimensionslosen Variablen wurde wieder weggelassen)[19]

$$\Gamma^2 \frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} = 1. \tag{4.78}$$

Durch die unterschiedliche Skalierung von x und y ist das Problem nun auf dem Quadrat $[-1, 1] \times [-1, 1]$ zu lösen. Die Geometrie ist durch das sogenannte *Aspektverhältnis* $\Gamma = b/a$ des Kanals charakterisiert, welches als Parameter in der Differentialgleichung auftaucht. An den Wänden des Kanals muß die Geschwindigkeit w die *Haftbedingung* (*no-slip condition*)

$$w(x = \pm 1, y) = w(x, y = \pm 1) = 0 \tag{4.79}$$

erfüllen.

[18] Wir haben hier den Querstrich ($\bar{w}$) zur Andeutung der exakten Lösung aus Bequemlichkeit weggelassen.

[19] Diese Poisson-Gleichung ist äquivalent zur Gleichung für die stationäre zweidimensionale Wärmeleitung mit einer homogenen Verteilung von Wärmequellen und konstanter Randtemperatur.

Zur Approximation der exakten Lösung verwenden wir den Ansatz

$$w = \sum_{i=1}^{I} w_i \phi_i(x, y), \tag{4.80}$$

wobei wir bilineare Elemente nach (4.53) verwenden und die Summe über alle Knoten i läuft. Die Projektion der Differentialgleichung (4.78) auf die Testfunktionen $\phi_m(x, y)$ liefert (S ist die Querschnittsfläche des Kanals)

$$\int_S \left(\Gamma^2 \phi_m \frac{\partial^2 w}{\partial x^2} + \phi_m \frac{\partial^2 w}{\partial y^2} \right) \mathrm{d}S = \int_S \phi_m \, \mathrm{d}S. \tag{4.81}$$

Nach partieller Integration und Einsetzen des Ansatzes (4.80) erhalten wir das lineare System

$$\mathsf{B} \cdot \boldsymbol{w} = \boldsymbol{G} \qquad \text{bzw.} \qquad B_{m,i} w_i = G_m, \tag{4.82}$$

wobei $\boldsymbol{w} = w_i = w_{j,k}$ der Vektor der unbekannten Geschwindigkeiten an den Knotenpunkten $i = (j, k)$ ist. In der Gleichung sind

$$\mathsf{B} = B_{m,i} = -\int_{-1}^{1} \int_{-1}^{1} \left(\Gamma^2 \frac{\partial \phi_m}{\partial x} \frac{\partial \phi_i}{\partial x} + \frac{\partial \phi_m}{\partial y} \frac{\partial \phi_i}{\partial y} \right) \mathrm{d}x \, \mathrm{d}y, \tag{4.83}$$

und

$$\boldsymbol{G} = G_m = \int_{-1}^{1} \int_{-1}^{1} \phi_m \, \mathrm{d}x \, \mathrm{d}y. \tag{4.84}$$

Da die Knotenwerte auf dem Rand vorgegeben sind und verschwinden, ($w_{i,\text{Rand}} = 0$), müssen nur die inneren Knoten betrachtet werden. Bei bilinearen Elementen verschwinden aber die Ansatzfunktionen der inneren Knoten auf dem Rand. Damit verschwinden auch die bei der partiellen Integration auftretenden ausintegrierten Anteile. Die Indizes i und m in (4.82) laufen also nur über alle internen Knoten mit $2 \le j \le N_x - 1$ und $2 \le k \le N_y - 1$.

Da bei den bilinearen Lagrangeschen Testfunktionen $\phi_i = (1 + \xi_i \xi)(1 + \eta_i \eta)/4$ (4.53) die x- und y-Abhängigkeiten faktorisieren, lassen sich die zweidimensionalen Integrale als Produkte eindimensionaler Integrale berechnen.

Im folgenden betrachten wir einen repräsentativen Knoten m und fragen, welche Summanden zu der m-ten Gleichung des linearen Systems (4.82) beitragen. Dabei verwenden wir ein äquidistantes Gitter mit den Gitterweiten Δx und Δy. Hierfür gilt nach (4.60) $\mathrm{d}x/\mathrm{d}\xi = \Delta x/2$, $\mathrm{d}y/\mathrm{d}\eta = \Delta y/2$. Mit der Faktorisierung

$$\phi_m(x, y) = \phi_m^x(x) \phi_m^y(y) \tag{4.85}$$

erhalten wir für den Knoten m (Abb. 4.11)

$$\begin{aligned} G_m &= \left(\int_{-1}^{1} \phi_m^x \, \mathrm{d}x \right) \left(\int_{-1}^{1} \phi_m^y \, \mathrm{d}y \right) \\ &= \frac{\Delta x}{2} \underbrace{\left(\int_{\substack{2 \text{ Elem.} \\ \text{um } m}} \phi_m^\xi \, \mathrm{d}\xi \right)}_{=2} \frac{\Delta y}{2} \underbrace{\left(\int_{\substack{2 \text{ Elem.} \\ \text{um } m}} \phi_m^\eta \, \mathrm{d}\eta \right)}_{=2} = \Delta x \Delta y. \end{aligned} \tag{4.86}$$

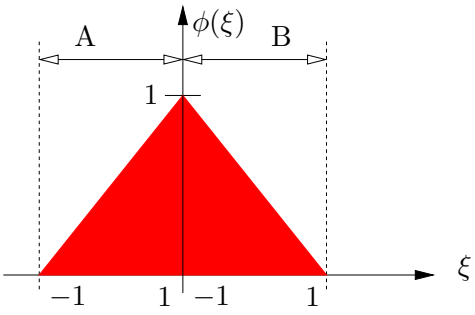


Abbildung 4.11.: Schnitt in ξ-Richtung durch die lineare Testfunktion, die sich in ξ-Richtung über zwei Elemente (gestrichelt) erstreckt. Die Werte der jeweiligen lokalen Koordinaten ξ sind für beide Elemente angezeigt.

Die Integration über ξ und η läuft hier jeweils über die beiden Elemente A und B, die zum Knoten m gehören. Graphisch entspricht jedes der Integrale der roten Fläche unter der Testfunktion in Abb. 4.11 mit Flächeninhalt 2. Der Vektor $\boldsymbol{G}$ besitzt also konstante Einträge $G_m = \Delta x \Delta y$, unabhängig vom Index m.

Um die Summanden in der m-ten Gleichung zu berechnen, die von der Matrix B stammen, betrachten wir zunächst den zweiten Summanden in (4.83). Für ihn gilt

$$\int_{-1}^{1}\int_{-1}^{1} \frac{\partial\phi_m}{\partial y}\frac{\partial\phi_i}{\partial y}\,\mathrm{d}x\,\mathrm{d}y \overset{(4.85)}{=} \left(\int_{-1}^{1} \phi_m^x \phi_i^x\,\mathrm{d}x\right)\left(\int_{-1}^{1} \frac{\partial\phi_m^y}{\partial y}\frac{\partial\phi_i^y}{\partial y}\,\mathrm{d}y\right). \tag{4.87}$$

Die Integrale ergeben nur dann einen von Null verschiedenen Wert, wenn sich die Gebiete überlappen, auf denen die Testfunktionen ϕ_m von Knoten m und ϕ_i von Knoten i von Null verschieden sind (Grundfläche in Abb. 4.8). Da sich die zweidimensionalen Testfunktionen über jeweils vier Elemente erstrecken, müssen insgesamt 9 Fälle berücksichtigt werden: Der Fall $m = i$ und 8 Fälle, in denen die Knoten m und i benachbart sind (Abb. 4.12). Die 5 Fälle, bei denen die Knoten m und i beide auf der x oder beide auf der y-Achse liegen, entsprechen wegen der o.a. Faktorisierung dem eindimensionalen Fall (siehe Abb. 4.12a).

Tatsächlich haben wir die eindimensionalen Integrale schon berechnet (siehe (4.66)). Für ein äquidistantes Gitter erhalten wir deshalb aus (4.68) und (4.69) mit $\mathrm{d}y = (\Delta y/2)\mathrm{d}\eta$

$$\frac{1}{\Delta y}\int_{-1}^{1} \frac{\partial\phi_m^y}{\partial y}\frac{\partial\phi_i^y}{\partial y}\,\mathrm{d}y = \frac{2}{\Delta y^2}\int \frac{\partial\phi_m^\eta}{\partial\eta}\frac{\partial\phi_i^\eta}{\partial\eta}\,\mathrm{d}\eta \overset{(4.69)}{=} \frac{1}{\Delta y^2}\,(-1,2,-1)^{\mathrm{T}} =: -L_{yy}, \tag{4.88a}$$

$$\frac{1}{\Delta x}\int_{-1}^{1} \phi_m^x \phi_i^x\,\mathrm{d}x = \frac{1}{2}\int \phi_m^\xi \phi_i^\xi\,\mathrm{d}\xi \overset{(4.68)}{=} \left(\frac{1}{6},\frac{2}{3},\frac{1}{6}\right) =: M_x. \tag{4.88b}$$

Das vorletzte Gleichheitszeichen gilt für je ein Element des Vektors, je nachdem wie der Knoten i relativ zu zum Knoten m liegt. Die Elemente des Spaltenvektors L_{yy} beziehen sich also auf die Punkte N, P und S (in geographischer Notation), und diejenigen des Zeilenvektors M_x auf die Punkte W, P und E. Neben diesen 5 Fällen treten aber noch 4 andere Fälle auf, bei denen sich die Testfunktionen der Knoten m und i überlappen. Einer dieser Fälle ist in Abb. 4.12b skizziert.

Da die Integration über die Fläche bei den verwendeten bilinearen Elementen immer in zwei eindimensionale Integrationen über x und y bzw. über ξ und η sepa-

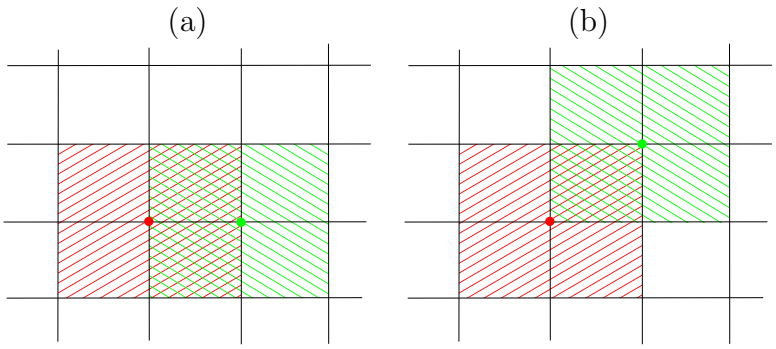


Abbildung 4.12.: Zwei von neun Möglichkeiten der Überlappung der Grundflächen (schraffiert) von bilinearen Testfunktionen. Der Knoten m ist rot und der Knoten i ist grün angedeutet.

riert, kann man alle Fälle durch das *dyadische Produkt*[20] von M_x und L_{yy} erhalten. Dieses *Tensorprodukt* schreibt man gelegentlich auch als $L_{yy} \otimes M_x$. Demnach ist

$$L_{yy} \otimes M_x = -\frac{1}{\Delta y^2} \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix} \otimes \left(\frac{1}{6}, \frac{2}{3}, \frac{1}{6} \right) \tag{4.89}$$

$$= -\frac{1}{\Delta y^2} \begin{pmatrix} (-1) \times \frac{1}{6} & (-1) \times \frac{2}{3} & (-1) \times \frac{1}{6} \\ 2 \times \frac{1}{6} & 2 \times \frac{2}{3} & 2 \times \frac{1}{6} \\ (-1) \times \frac{1}{6} & (-1) \times \frac{2}{3} & (-1) \times \frac{1}{6} \end{pmatrix} = \frac{1}{\Delta y^2} \begin{pmatrix} 1/6 & 2/3 & 1/6 \\ -1/3 & -4/3 & -1/3 \\ 1/6 & 2/3 & 1/6 \end{pmatrix}.$$

Für den betrachteten zweiten Summanden in (4.83) erhalten wir damit für jeden Punkt m des Integrationsgebiets denselben Satz von 9 von Null verschiedenen Koeffizienten in Form des dyadischen Produkts $L_{yy} \otimes M_x$ (bis auf das Vorzeichen). Die Position des jeweiligen Koeffizienten innerhalb der 3×3 Matrix (4.89) gibt die relative Lage der beteiligten Knoten an.

In analoger Weise kann man mit dem ersten Summanden in (4.83) verfahren. Da in diesem Summanden nur die Ableitung bezüglich der anderen Koordinatenrichtung (nach x) auftritt, erhalten wir alle Integrale des ersten Summanden durch $M_y \otimes L_{xx}$ (Vertauschung von x und y). Wenn wir nun Gleichung (4.82) durch $\Delta x \Delta y$ dividieren, erhalten wir (der Knotenindex i entspricht den Gitterindizes j und k: $w_i \to w_{j,k}$)

$$\underbrace{\left(\Gamma^2 M_y \otimes L_{xx} + L_{yy} \otimes M_x\right)}_{\mathsf{B}/\Delta x \Delta y} w_{j,k} = 1, \tag{4.90}$$

[20] Als Ergebnis des dyadischen Produkts (oder auch Tensorprodukt bzw. äußeres Produkt) zweier Vektoren (einfach indizierte Objekte) erhält man eine Matrix (zweifach indiziertes Objekt): $\boldsymbol{ab} = a_i b_j = C_{ij}$. Man schreibt auch $\boldsymbol{ab} = \boldsymbol{a} \otimes \boldsymbol{b}$. Die besondere Kennzeichnung durch $\otimes$ ist nicht erforderlich, wenn man das Skalarprodukt immer, wie hier, mit $\cdot$ schreibt. Beachte: $\boldsymbol{a} \otimes \boldsymbol{b} = a_i b_j \neq b_i a_j = \boldsymbol{b} \otimes \boldsymbol{a} = (\boldsymbol{a} \otimes \boldsymbol{b})^{\mathrm{T}}$.

mit

$$L_{xx} = \left(\frac{1}{\Delta x^2}, -\frac{2}{\Delta x^2}, \frac{1}{\Delta x^2} \right), \quad L_{yy} = \left(\frac{1}{\Delta y^2}, -\frac{2}{\Delta y^2}, \frac{1}{\Delta y^2} \right)^{\mathrm{T}}, \tag{4.91a}$$

und

$$M_x = \left(\frac{1}{6}, \frac{2}{3}, \frac{1}{6} \right), \quad M_y = \left(\frac{1}{6}, \frac{2}{3}, \frac{1}{6} \right)^{\mathrm{T}}. \tag{4.91b}$$

Damit lautet

$$M_y \otimes L_{xx} = \frac{1}{\Delta x^2} \underbrace{\begin{pmatrix} 1/6 \\ 2/3 \\ 1/6 \end{pmatrix}}_{\substack{\text{Index } k \\ \text{in } y\text{-Richtung}}} \otimes \underbrace{(1, -2, 1)}_{\substack{\text{Index } j \\ \text{in } x\text{-Richtung}}} = \frac{1}{\Delta x^2} \begin{pmatrix} 1/6 & -1/3 & 1/6 \\ 2/3 & -4/3 & 2/3 \\ 1/6 & -1/3 & 1/6 \end{pmatrix}. \tag{4.92}$$

Beachte, daß in (4.82) bzw. (4.90) über alle Knoten mit $i = (j, k)$, d.h. über j und k summiert wird.[21] Daher resultiert aus dem Produkt von $M_y \otimes L_{xx}$ mit $w_{j,k}$ eine Summe über alle $w_{j,k}$ der 9 beteiligten Elemente, gewichtet mit den entsprechenden Koeffizienten von $M_y \otimes L_{xx}$. Gleichung (4.90) ist also eine skalare Gleichung für den Knoten m. Entsprechend dieser Betrachtungen erhalten wir im ersten Summanden von (4.90)

$$L_{xx} w_{j,k} = \frac{w_{j-1,k} - 2w_{j,k} + w_{j+1,k}}{\Delta x^2}. \tag{4.93}$$

Dann wird

$$\begin{aligned} M_y \otimes L_{xx} w_{j,k} &= M_y \otimes \frac{w_{j-1,k} - 2w_{j,k} + w_{j+1,k}}{\Delta x^2} \\ &= \frac{1}{6} \left(\frac{w_{j-1,k-1} - 2w_{j,k-1} + w_{j+1,k-1}}{\Delta x^2} \right) + \frac{2}{3} \left(\frac{w_{j-1,k} - 2w_{j,k} + w_{j+1,k}}{\Delta x^2} \right) \\ &\quad + \frac{1}{6} \left(\frac{w_{j-1,k+1} - 2w_{j,k+1} + w_{j+1,k+1}}{\Delta x^2} \right). \end{aligned} \tag{4.94}$$

und für den zweiten Summanden von (4.90) (vgl. (4.92))

$$\begin{aligned} L_{yy} \otimes M_x w_{j,k} &= \frac{w_{j,k-1} - 2w_{j,k} + w_{j,k+1}}{\Delta y^2} \otimes M_x \\ &= \frac{1}{6} \left(\frac{w_{j-1,k-1} - 2w_{j-1,k} + w_{j-1,k+1}}{\Delta y^2} \right) + \frac{2}{3} \left(\frac{w_{j,k-1} - 2w_{j,k} + w_{j,k+1}}{\Delta y^2} \right) \\ &\quad + \frac{1}{6} \left(\frac{w_{j+1,k-1} - 2w_{j+1,k} + w_{j+1,k+1}}{\Delta y^2} \right). \end{aligned} \tag{4.95}$$

Gleichung (4.90) ist eine 9-Punkt-Formel. Die vergleichbare 5-Punkt-Formel für

[21] Die Notation ist daher hier nicht ganz sauber, denn in (4.90) hat man ein doppeltes Skalarpro-

zentrale finite Differenzen lautet

$$\Gamma^2 \frac{w_{j-1,k} - 2w_{j,k} + w_{j+1,k}}{\Delta x^2} + \frac{w_{j,k-1} - 2w_{j,k} + w_{j,k+1}}{\Delta y^2} = 1. \tag{4.96}$$

Man kann sie formal aus (4.90) reproduzieren, wenn man $M_x = M_y = (0, 1, 0)$ setzt.

Die Operatoren M_x und L_{xx} traten auch schon bei der Lösung der eindimensionalen Wärmeleitungsgleichung (4.72) auf. Es ist daher interessant zu bemerken, daß diese Operatoren bei dem zweidimensionalen Problem im Rahmen der finiten Volumen in Form des Tensorprodukts auftreten.

In Gleichung (4.90) sind jeweils 9 Unbekannte miteinander verkoppelt. Man kann das für alle Gitterpunkte (j, k) resultierende Gleichungssystem bequem durch *sukzessive Überrelaxation* (SOR) lösen (siehe Vorl. *Numerische Methoden der Ingenieurwissenschaften*, LVA-Nr. 322.036). Diese Methode wurde auch schon im Zusammenhang mit der Lösung der Laplace-Gleichung mittels finiter Volumen (4.39) erwähnt. Zur Übung könnte man dies auch einmal programmieren. Eine Implementierung eines entsprechenden FORTRAN-Programms ist in Fletcher (1991*a*) beschrieben. Durch Gitterverfeinerung kann man zeigen, daß das Finite-Elemente-Verfahren (4.90) von zweiter Ordnung ist (der Diskretisierungsfehler skaliert mit Δx^2 und Δy^2) und eine Genauigkeit besitzt, die mit finiten Differenzen vergleichbar ist.

4.4.5. Verzerrte Gebiete

Ein Hauptanwendungsgebiet finiter Elemente ist die Integration von Differentialgleichungen auf irregulären Gebieten. Für Dreieckselemente in 2D und Tetraeder in 3D ist dies sofort einsichtig. Aber auch viereckige bzw. kubische (Hexaeder-) Elemente lassen sich zur Berechnung auf komplexen Gebieten verwenden. Vierecks- bzw. Hexaeder-Elemente werden gerne verwendet, weil man für sie leichter ein Gitter generieren kann als für Dreiecks- bzw. Tetraeder-Elemente.[22]

Wenn man nun ein komplexes Gebiet hat, könnte man die Testfunktionen für die finiten Elemente im physikalischen Raum definieren und auch dort die Integrationen durchführen. Dieses Verfahren läßt sich aber nicht leicht verallgemeinern, wenn beispielsweise das Gebiet (und damit das Gitter) nur ein wenig geändert wird. Die Idee ist daher, eine einfache Transformation zu finden, mit deren Hilfe jedes Element eines gegebenen viereckigen (kubischen) Gitters und seine Knotenpunkte auf ein und dasselbe Quadrat $(\xi, \eta) \in [-1, 1] \times [-1, 1]$ (bzw. auf den Kubus $[-1, 1]^3$) transformiert wird (Abb. 4.13).

Für lineare zweidimensionale Viereckselemente mit $n = 4$ Knoten wird dies mit

dukt (doppelte Verjüngung) wischen den Dyaden und $w_{j,k}$.

[22] Die Gittergenerierung ist ein Thema für sich und wird ansatzweise im zweiten Teil der Vorlesung, *Numerische Methoden der Strömungsmechanik*, LVA-Nr. 302.042, behandelt.

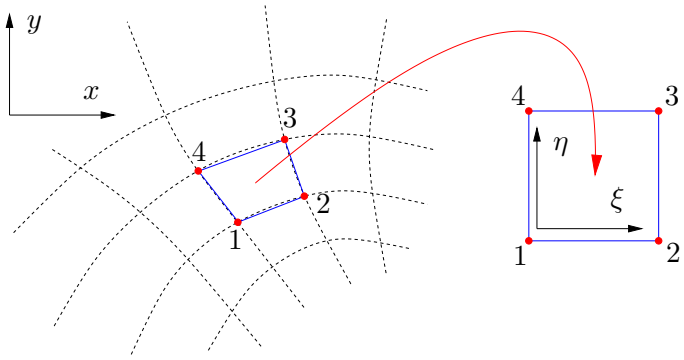


Abbildung 4.13.: Transformation eines Elements eines krummlinigen Gitters (Koordinaten x und y) auf ein universelles Quadrat mit Koordinaten (ξ, η).

der *isoparametrischen Transformation*[23] $(x, y) \leftrightarrow (\xi, \eta)$

$$x(\xi,\eta) = \sum_{l=1}^{4} x_l \phi_l(\xi,\eta), \qquad y(\xi,\eta) = \sum_{l=1}^{4} y_l \phi_l(\xi,\eta) \tag{4.97}$$

erreicht. Dies ist eine bilineare Interpolation zwischen den Knotenpunkten im physikalischen Raum. In Vektorform lautet sie

$$\boldsymbol{x}(\xi,\eta) = \sum_{l=1}^{4} \phi_l(\xi,\eta)\boldsymbol{x}_l. \tag{4.98}$$

Dabei sind die vier Vektoren $\boldsymbol{x}_l = (x_l, y_l)^{\mathrm{T}}$ die physikalischen Koordinaten der vier Knoten eines linearen Elements und $\phi_l(\xi, \eta)$ sind die linearen Testfunktionen (4.53) auf $[-1, 1] \times [-1, 1]$.

Um die entscheidenden Schritte bei der Anwendung zu zeigen, betrachten wir die Poisson-Gleichung $\nabla^2 w = 1$ auf einem irregulären Gebiet. Bei der Projektion des Residuums erhält man ein lineares System wie (4.82) der Form

$$\mathsf{B} \cdot \boldsymbol{w} = \boldsymbol{G}. \tag{4.99}$$

Die Matrix B entsteht hierbei durch die Projektion von $\nabla^2 w$ auf die Testfunktionen ϕ_m. Nach partieller Integration ergibt sich für B die Form

$$B_{m,j} = -\int_S \left(\frac{\partial \phi_m}{\partial x}\frac{\partial \phi_j}{\partial x} + \frac{\partial \phi_m}{\partial y}\frac{\partial \phi_j}{\partial y} \right) \mathrm{d}x\,\mathrm{d}y \tag{4.100}$$

hat. Die Integration erstreckt sich hierbei wie immer über das gesamte physikalische Integrationsgebiet S.

Die Integration über das irreguläre Gitter kann man nun in eine Integration über ein reguläres Gitter überführen. Dazu betrachten wir exemplarisch den ersten Summanden in (4.100)

$$I = \int_S \frac{\partial \phi_m}{\partial x}\frac{\partial \phi_j}{\partial x}\,\mathrm{d}x\,\mathrm{d}y \tag{4.101}$$

[23] *Isoparametrisch* bedeutet, daß für die Koordinatentransformation dieselben Funktionen gewählt

und führen alle Ableitungen und Integrationen bezüglich x und y auf die Variablen ξ und η entsprechend (4.97) zurück. Für die partiellen Ableitungen gilt allgemein

$$\begin{pmatrix} \frac{\partial}{\partial \xi} \\ \frac{\partial}{\partial \eta} \end{pmatrix} \phi_i \left[x(\xi,\eta), y(\xi,\eta)\right] = \underbrace{\begin{pmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial y}{\partial \xi} \\ \frac{\partial x}{\partial \eta} & \frac{\partial y}{\partial \eta} \end{pmatrix}}_{J} \cdot \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{pmatrix} \phi_i(x,y), \tag{4.102}$$

mit der Jacobi-Matrix J. Die Elemente von J können mittels (4.97) berechnet werden. Unter Verwendung von (4.53), d.h. $\phi_l(\xi,\eta) = (1+\xi_l\xi)\,(1+\eta_l\eta)\,/4$, ist beispielsweise

$$\frac{\partial y}{\partial \xi} = \sum_{l=1}^{4} \frac{\partial \phi_l}{\partial \xi} y_l = \frac{1}{4} \sum_{l=1}^{4} \xi_l \,(1+\eta_l\eta)\, y_l. \tag{4.103}$$

Mit der Umkehrung von (4.102)[24]

$$\begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{pmatrix} \phi_i = J^{-1} \cdot \begin{pmatrix} \frac{\partial}{\partial \xi} \\ \frac{\partial}{\partial \eta} \end{pmatrix} \phi_i = \underbrace{\frac{1}{\det(J)} \begin{pmatrix} \frac{\partial y}{\partial \eta} & -\frac{\partial y}{\partial \xi} \\ -\frac{\partial x}{\partial \eta} & \frac{\partial x}{\partial \xi} \end{pmatrix}}_{J^{-1}} \cdot \begin{pmatrix} \frac{\partial}{\partial \xi} \\ \frac{\partial}{\partial \eta} \end{pmatrix} \phi_i \tag{4.104}$$

und $\mathrm{d}x\,\mathrm{d}y = \det(J)\,\mathrm{d}\xi\,\mathrm{d}\eta$ können wir das gesuchte Integral daher schreiben als eine Summe von Integralen über einzelne reguläre Elemente

$$I = \sum_{\text{Elemente}} \int_{-1}^{1}\int_{-1}^{1} \frac{1}{\det(J)} \left(\frac{\partial y}{\partial \eta}\frac{\partial \phi_m}{\partial \xi} - \frac{\partial y}{\partial \xi}\frac{\partial \phi_m}{\partial \eta}\right)\left(\frac{\partial y}{\partial \eta}\frac{\partial \phi_j}{\partial \xi} - \frac{\partial y}{\partial \xi}\frac{\partial \phi_j}{\partial \eta}\right) \mathrm{d}\xi\,\mathrm{d}\eta, \tag{4.105}$$

wobei die Summe über alle Elemente zur erstrecken ist, die zum Integral I über S beitragen. Dies hängt von der Lage der Knoten und der Reichweite der zugehörigen Interpolationsfunktionen ab (siehe Abb. 4.12). Der Vorteil der isoparametrischen Transformation besteht darin, daß alle Terme in (4.105) Funktionen von ξ und η sind, und daß die Elemente regulär (quadratisch) sind. Die Integration kann dann relativ einfach numerisch durchgeführt werden, z.B. durch Gauß-Quadratur.

4.5. Spektrale Methoden

Wie alle Methoden der gewichteten Residuen gehen auch spektrale Methoden vom Ansatz (4.1) aus. Allerdings sind die Ansatzfunktionen und die Gewichtsfunktionen

werden wie für die Ansatzfunktionen.

[24] Inversion einer 2×2-Matrix:

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \qquad \Rightarrow \qquad A^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

nicht um einen Knoten lokalisiert wie bei den finiten Elementen. Vielmehr sind die Ansatzfunktionen auf dem gesamten Volumen definiert.

Bei Problemen mit einfachen Geometrien und glatten Randbedingungen,[25] für welche auch die Lösungen hinreichend glatt sind, besitzen spektrale Methode entscheidende Vorteile gegenüber anderen Methoden. In diesen Fällen kann man genaue Näherungen schon mit einer sehr geringen Anzahl von Ansatzfunktionen (Moden) erhalten. Speziell für Probleme auf periodischen Gebieten mit unendlich oft stetig differenzierbaren Lösungen kann man zeigen, daß der Fehler spektraler Methoden von der Ordnung $O[(1/N)^m]$ ist. Hierbei ist N die Anzahl der Ansatzfunktionen und m steht im Zusammenhang mit der Anzahl der kontinuierlichen Ableitungen der zu approximierenden Funktion. Bei der spektralen Approximation einer unendlich oft stetig differenzierbaren Funktion ist die Fehlerordnung damit geringer als jede Potenz von $(1/N)$.[26] Diese Eigenschaft wird auch *exponentielle Konvergenz* genannt (Canuto et al., 1988). Bei glatten Problemen in einfachen Geometrien ist die Genauigkeit spektraler Verfahren daher nicht zu übertreffen. Sie sind jedoch nicht so leicht anzuwenden, wenn das Gebiet eine komplexe Geometrie aufweist oder die Randbedingungen nicht glatt sind.

Bei den klassischen spektralen Methoden wie dem Galerkin-Verfahren rechnet man im wesentlich mit den Fourieramplituden. Sie sind die zu bestimmenden Unbekannten. Daher rührt die Bezeichnung *spektral*. Es zeigt sich aber, daß die Berechnung nichtlinearer Terme bei einer rein spektralen Diskretisierung sehr rechenintensiv ist, da sie bei der Transformation in den spektralen Raum in Faltungsintegrale übergehen.[27] Mit der Einführung (seit 1965) von schnellen Transformationen (z.B. der *Fast Fourier Transform* FFT) vom spektralen Raum in den Ortsraum und zurück wurde es möglich, die Faktoren der nichtlinearen Terme in den Ortsraum zu transformieren, erst dort miteinander zu multiplizieren und das Produkt dann wieder in den spektralen Raum zurückzutransformieren. Diese Vorgehensweise nennt man *pseudospektral*. Die pseudospektrale Methode ist bei großen Problemen we-

[25] Glatt bedeutet, daß die Randbedingungen hinreichend oft (am besten beliebig oft) stetig differenzierbar sind. Ein Sprung in der Randtemperatur wäre z.B. nicht glatt.

[26] Die Anzahl N der Ansatzfunktionen entspricht hierbei der Anzahl N der Gitterpunkte (gleiche Anzahl von Unbekannten). $1/N$ entspricht dann dem Gitterabstand Δx.

[27] Wenn wir den nichtlinearen Term $f(x)g(x)$ durch die Fouriertransformierten $\hat{f}(k)$ und $\hat{g}(k)$ darstellen, dann gilt mit der Variablen-Transformation von k' nach l, $k' = l - k$ mit $\mathrm{d}l = \mathrm{d}k'$, im Schritt $(*)$

$$
\begin{aligned}
f(x)g(x) &= \left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\hat{f}(k)\mathrm{e}^{-\mathrm{i}kx}\mathrm{d}k\right)\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\hat{g}(k')\mathrm{e}^{-\mathrm{i}k'x}\mathrm{d}k'\right)\\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\hat{f}(k)\hat{g}(k')\mathrm{e}^{-\mathrm{i}(k+k')x}\mathrm{d}k\,\mathrm{d}k' \overset{(*)}{=} \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\hat{f}(k)\hat{g}(l-k)\mathrm{e}^{-\mathrm{i}lx}\mathrm{d}k\,\mathrm{d}l\\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left[\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\hat{f}(k)\hat{g}(l-k)\mathrm{d}k\right]\mathrm{e}^{-\mathrm{i}lx}\mathrm{d}l.
\end{aligned}
$$

Die eckige Klammer nennt man Faltungsintegral. wir können es als Fouriertransformierte des Produkts $f(x)g(x)$ identifizieren.

sentlich schneller als die reine spektrale Methode. Das hat der (pseudo)spektralen Behandlung der Navier-Stokes-Gleichung zu einem gewissen Durchbruch verholfen. Bei Verwendung von Kollokation kann man praktisch vollständig im Ortsraum rechnen, man nutzt aber die Eigenschaften der Ansatzfunktionen, um geeignete Gitter zu wählen und die Ableitungen der Funktionen optimal zu berechnen. Je nach Verfahren geht man also vom spektralen oder vom Orts-Raum aus.

Die Ansatzfunktionen sind bei spektralen Methoden oft orthogonal zueinander.[28] Ein System von Funktionen $\{\phi_i(x,y,z)\}$ ist *orthogonal*, wenn für beliebige Elemente ϕ_i und ϕ_j des Funktionensystems gilt[29]

$$\langle \phi_i | \phi_j \rangle := \int_V w(\boldsymbol{x}) \phi_i^*(\boldsymbol{x}) \phi_j(\boldsymbol{x}) \,\mathrm{d}V = N_i \delta_{ij} = \begin{cases} N_i \neq 0, & i = j, \\ 0, & \text{sonst}, \end{cases} \tag{4.106}$$

wobei δ_{ij} das *Kronecker-Symbol* ist, und $w(\boldsymbol{x})$ eine Gewichtsfunktion, die von der Art der Ansatzfunktionen abhängt. Das integrale Produkt $\langle \phi_i | \phi_j \rangle$ definiert ein *Skalarprodukt* zwischen den Funktionen ϕ_i des Funktionenraums. Bei *orthonormalen* Funktionen ist der Normierungsfaktor $N_i = 1$. Beispiele für orthogonale Funktionen sind Harmonische (Fourier-Reihen), Chebyshev-Polynome oder Legendre-Polynome.

4.5.1. Galerkin-Methode

Zur Demonstration der spektralen Methode betrachten wir wieder das Diffusionsproblem (1.50) auf $x \in [0,1]$

$$\frac{\partial \bar{T}}{\partial t} - \kappa \frac{\partial^2 \bar{T}}{\partial x^2} = 0. \tag{4.107}$$

Spezielle Anfangsbedingungen

Wir wählen nun die speziellen Rand- und Anfangsbedingungen

$$\bar{T}(0,t) = 0, \qquad \bar{T}(1,t) = 1 \qquad \text{und} \qquad \bar{T}(x,0) = x + \sin \pi x. \tag{4.108}$$

Für diese Bedingungen läßt sich die exakte Lösung leicht angeben (vgl. (1.56))

$$\bar{T}(x,t) = x + \mathrm{e}^{-\kappa \pi^2 t} \sin(\pi x). \tag{4.109}$$

Um die Diffusionsgleichung (4.107) spektral zu lösen, verwenden wir harmonische Funktionen und setzen die Lösung als Fourier-Reihe mit zeitabhängigen Koeffizienten an

$$T(x,t) = x + \sin \pi x + \sum_{j=1}^{J} a_j(t) \sin(j \pi x). \tag{4.110}$$

[28] Bei der klassischen Galerkin-Methode ist dies im allgemeinen nicht der Fall.
[29] Die Schreibweise $\langle | \rangle$ wird auch *bra-ket*-Notation genannt, motiviert durch das englische *bracket*.

Es ist sinnvoll, die beiden ersten Summanden so zu wählen, daß die Randbedingungen erfüllt sind. Unserer Wahl $x + \sin(\pi x)$ genügt außerdem noch der Anfangsbedingung. Für den Rest suchen wir eine Funktion, die am Rand und zum Zeitpunkt $t = 0$ verschwindet. Die homogenen Randbedingungen werden von dem Funktionensystem $\{\sin(j\pi x)\}$ erfüllt. Für sie gilt das Skalarprodukt (siehe (4.106) oder Fußnote 4 auf Seite 58)

$$\begin{aligned}\langle \sin(i\pi x)|\sin(j\pi x)\rangle &= \int_0^1 \sin(i\pi x)\,\sin(j\pi x)\,\mathrm{d}x \\ &= \frac{1}{2}\int_0^1 \{\cos[(i-j)\pi x] - \cos[(i+j)\pi x]\}\,\mathrm{d}x \qquad (4.111)\\ &= \left\{ \begin{array}{ll} i \neq j: & \frac{1}{2}\left[\frac{\sin[(i-j)\pi x]}{(i-j)\pi} - \frac{\sin[(i+j)\pi x]}{(i+j)\pi}\right]_0^1 = 0, \\ i = j: & \frac{1}{2}\int_0^1 \mathrm{d}x - \frac{1}{2}\left[\frac{\sin[(i+j)\pi x]}{(i+j)\pi}\right]_0^1 = \frac{1}{2} \end{array} \right\} = \frac{\delta_{ij}}{2}.\end{aligned}$$

Für homogene Randbedingungen bilden die Funktionen $\sin(j\pi x)$ ein vollständiges Funktionensystem auf dem Intervall $[0, 1]$.

Wenn man den Ansatz (4.110) in die Diffusionsgleichung einsetzt, erhält man das Residuum

$$R(x,t) = \sum_{j=1}^{J}\left(\frac{\mathrm{d}a_j}{\mathrm{d}t} + \kappa j^2\pi^2 a_j\right)\sin(j\pi x) + \kappa\pi^2\sin(\pi x). \qquad (4.112)$$

Zur Bestimmung der unbekannten Koeffizienten $a_j(t)$ wird nun das Residuum auf die Gewichtsfunktionen ϕ_m projiziert, die hier identisch mit den Ansatzfunktionen $\phi_m = \sin(m\pi x)$ gewählt werden, und zu Null gesetzt. Wir verlangen also

$$\begin{aligned}0 \overset{!}{=} \langle \sin(m\pi x)|R(x,t)\rangle &= \int_0^1 R(x)\sin(m\pi x)\,\mathrm{d}x \qquad (4.113)\\ &= \sum_{j=1}^{J}\left(\frac{\mathrm{d}a_j}{\mathrm{d}t} + \kappa j^2\pi^2 a_j\right)\underbrace{\langle \sin(m\pi x)|\sin(j\pi x)\rangle}_{\frac{1}{2}\delta_{mj}} + \kappa\pi^2\underbrace{\langle \sin(m\pi x)|\sin(\pi x)\rangle}_{\frac{1}{2}\delta_{m1}}.\end{aligned}$$

Für lineare PDEs erster Ordnung in der Zeit ergibt sich so im allgemeinen ein lineares System gewöhnlicher Differentialgleichungen erster Ordnung der Form

$$\mathsf{A}\cdot\frac{\mathrm{d}\boldsymbol{a}}{\mathrm{d}t} + \mathsf{B}\cdot\boldsymbol{a} + \boldsymbol{r} = 0 \qquad (4.114)$$

für die unbekannten Koeffizienten $\boldsymbol{a}(t) = a_m(t)$. Bei linearen Differentialgleichungen (wie hier) erkennt man den Vorteil orthogonaler Funktionen: Es tragen zur Summe

Der Stern $*$ bezeichnet das konjugiert Komplexe für den Fall, daß komplexe Funktionen ver-

in (4.113) nur Integrale mit $j = m$ bei, so daß in den Matrizen A und B nur die Diagonalen besetzt sind. Vom zweiten Integral (Skalarprodukt) bleibt nur der Beitrag mit $m = 1$ übrig. Damit erhalten wir (der Normierungsfaktor $N_i = 1/2$ fällt heraus)

$$\frac{\mathrm{d}a_m}{\mathrm{d}t} + \kappa m^2\pi^2 a_m + r_m = 0, \tag{4.115}$$

mit $\boldsymbol{r} = r_m = \kappa\pi^2\delta_{m1} = (\kappa\pi^2, 0, 0, ...)^{\mathrm{T}}$. Die Lösung $\boldsymbol{a}(t)$ kann man sofort ablesen

$$\boldsymbol{a}(t) = a_m(t) = \begin{cases} \hat{a}_m \mathrm{e}^{-\kappa m^2\pi^2 t}, & \text{falls } m \neq 1, \\ \hat{a}_1 \mathrm{e}^{-\kappa\pi^2 t} - 1, & \text{falls } m = 1. \end{cases} \tag{4.116}$$

Die konstanten Amplituden $\hat{a}_m$ werden durch die Anfangsbedingung ($t = 0$) festgelegt. Mit der Anfangsbedingung $a_m(t = 0) = 0$ gilt daher $\hat{a}_m = 0$ für $m \neq 1$ und $\hat{a}_1 = 1$. Dies führt auf die Lösung

$$T(x,t) = x + \sin(\pi x) + \left(\mathrm{e}^{-\kappa\pi^2 t} - 1\right)\sin(\pi x) = \mathrm{e}^{-\kappa\pi^2 t}\sin(\pi x) + x. \tag{4.117}$$

Tatsächlich ist diese Lösung, die wir für jede Abbruchordnung $J \geq 1$ erhalten, identisch mit der exakten Lösung (4.109). Dies liegt hier an der besondere Wahl der Anfangsbedingung. Mit den Anfangsbedingungen hatten wir nämlich gerade die exakte asymptotische Lösung $\bar{T}(x, t \to \infty) = x$ für $t \to \infty$ gewählt, plus einem Beitrag $\sin(\pi x)$, der gerade der ersten Mode $j = 1$ entspricht. Da die Wärmeleitungsgleichung linear ist, wird diese Mode einfach exponentiell gedämpft, was man an (4.117) sieht. Für eine allgemeinere Anfangsbedingung erhält man natürlich keine exakte Lösung, wenn man den Ansatz (4.110) bei einer endlichen Ordnung J abbricht.

Allgemeinere Anfangsbedingungen

Wir wählen nun die algebraische Anfangsbedingung $T(x, 0) = 5x - 4x^2$ bei gleichen Randbedingungen. Wir erwarten, daß die Abweichung $4x - 4x^2$ dieser Anfangsbedingung von der asymptotischen Lösung $\bar{T}(x, t \to \infty) = x$ exponentiell mit der Zeit zerfallen wird. Dazu können wir uns $4x - 4x^2$ in die spektralen Bestandteile zerlegt vorstellen, wobei jede spektrale Komponente unabhängig von den anderen Komponenten (lineares System) mit seiner charakteristischen Zeitkonstante abklingt. Die Abklingrate $\kappa m^2\pi^2$ steigt quadratisch mit der Wellenzahl $m\pi$ der spektralen Komponente an, eine typische und wichtige Eigenschaft des diffusiven Terms.

Wir setzen nun den Ansatz

$$T(x,t) = 5x - 4x^2 + \sum_{j=1}^{J} a_j(t)\sin(j\pi x) \tag{4.118}$$

in die Wärmeleitungsgleichung ein. Vom algebraischen Anteil des Ansatzes bleibt

dann nur noch $-\kappa\partial_x^2(5x-4x^2)=8\kappa$ übrig, und das Residuum lautet

$$R(x,t)=\sum_{j=1}^{J}\left(\frac{\mathrm{d}a_j}{\mathrm{d}t}+\kappa j^2\pi^2 a_j\right)\sin(j\pi x)+8\kappa. \tag{4.119}$$

Die Projektion $\langle\sin(m\pi x)|R(x,t)\rangle=0$ liefert uns

$$0=\sum_{j=1}^{J}\left(\frac{\mathrm{d}a_j}{\mathrm{d}t}+\kappa j^2\pi^2 a_j\right)\underbrace{\langle\sin(m\pi x)|\sin(j\pi x)\rangle}_{\delta_{mj}/2}+\underbrace{\langle\sin(m\pi x)|8\kappa\rangle}_{(*)}. \tag{4.120}$$

Mit der Auswertung des letzten Integrals $(*)$[30] in (4.120) ergeben sich dann die Bestimmungsgleichungen für $a_m(t)$

$$\left(\frac{\mathrm{d}a_m}{\mathrm{d}t}+\kappa m^2\pi^2 a_m\right)+\frac{16\kappa}{m\pi}\left[1-(-1)^m\right]=0. \tag{4.121}$$

Die Amplituden $a_m(t)$ der Moden zerfallen exponentiell zu 0 (m gerade) oder zu $-32/m^3\pi^3$ (m ungerade)

$$a_m(t)=\begin{cases}\hat{a}_m\mathrm{e}^{-\kappa m^2\pi^2 t}, & \text{falls } m \text{ gerade},\\ \hat{a}_m\mathrm{e}^{-\kappa m^2\pi^2 t}-\dfrac{32}{m^3\pi^3}, & \text{falls } m \text{ ungerade}.\end{cases} \tag{4.122}$$

Damit die Anfangsbedingung $a_m(t=0)=0$ realisiert wird, muß $\hat{a}_m=0$ sein (m

wendet werden.

[30]

$$\begin{aligned}\langle\sin(m\pi x)|8\kappa\rangle &= 8\kappa\int_0^1\sin(m\pi x)\,\mathrm{d}x=\frac{8\kappa}{m\pi}\left[-\cos(m\pi x)\right]_0^1\\ &=\frac{8\kappa}{m\pi}\left[1-(-1)^m\right]=\begin{cases}\dfrac{16\kappa}{m\pi}, & m \text{ ungerade},\\ 0, & m \text{ gerade}.\end{cases}\end{aligned}$$

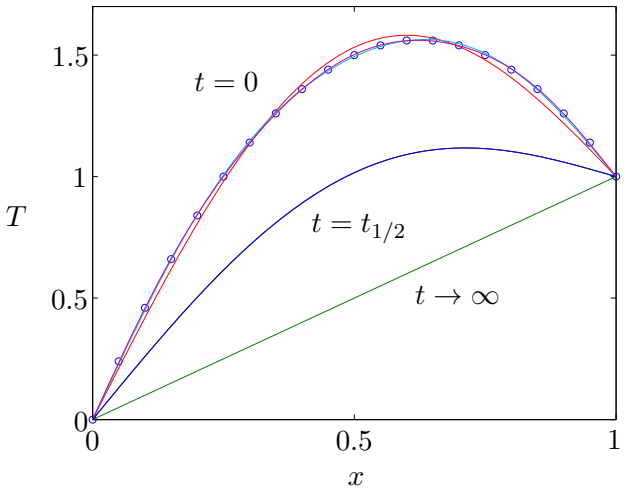


Abbildung 4.14.: Die Lösung der eindimensionalen Wärmeleitungsgleichung in verschiedenen Approximationen. Die Symbole (∘) deuten die exakten Anfangsbedingungen an. In unmittelbarer Nachbarschaft dieser Punkte finden sich die Galerkin-Approximationen (4.123) mit $J = 1$, 2 und 3 für $t = 0$. Die mittlere Kurve zeigt die nicht mehr voneinander zu unterscheidenden Approximationen mit $J = 1$, 2 und 3 für $t = t_{1/2}$ (siehe Text). Außerdem ist das asymptotische lineare Profil für $t \to \infty$ eingezeichnet (grün).

gerade) und $\hat{a}_m = 32/m^3\pi^3$ (m ungerade). Also lautet die Lösung[31]

$$\begin{aligned} T(x,t) &= 5x - 4x^2 + \sum_{\substack{j=1 \\ j \text{ ungerade}}}^{J} \frac{32}{j^3\pi^3}\left(\mathrm{e}^{-\kappa j^2\pi^2 t} - 1\right)\sin(j\pi x) \\ &\overset{j\to 2j-1}{=} 5x - 4x^2 + 32\sum_{j=1}^{J} \frac{\mathrm{e}^{-\kappa(2j-1)^2\pi^2 t} - 1}{(2j-1)^3\pi^3}\sin\left[(2j-1)\pi x\right] \\ &\overset{(*)}{=} x + 32\sum_{j=1}^{J} \frac{\mathrm{e}^{-\kappa(2j-1)^2\pi^2 t}}{(2j-1)^3\pi^3}\sin\left[(2j-1)\pi x\right]. \end{aligned} \tag{4.123}$$

Dieselbe Lösungsstruktur hatten wir auch schon in Kap. 1.2.4 durch Separation der Variablen erhalten.

Die Fourier-Reihe (4.123) konvergiert mit wachsendem J extrem schnell (Abb. 4.14). Zum Zeitpunkt $t_{1/2} = \ln(2)/(\kappa\pi^2)$, zu dem die Amplitude der Mode $j = 1$ auf die Hälfte der Anfangsgröße (Faktor 1/2) abgefallen ist, sind die drei ersten

[31] Beim letzten Gleichheitszeichen (∗) wurde die Fourierreihendarstellung

$$4x(1-x) = \sum_{\substack{j=1 \\ j \text{ ungerade}}}^{J} \frac{32}{j^3\pi^3}\sin(j\pi x)$$

ausgenutzt. Beachte, daß $4x - 4x^2$ auf dem Rand des Gebiets $[0,1]$ verschwindet und daher durch eine reine Sinus-Reihe dargestellt werden kann.

Näherungen in Abb. 4.14 schon nicht mehr von einander zu unterscheiden, da die höheren Moden noch schneller abklingen (mit Zerfallsraten $\propto j^2$).

Für die hier verwendete Diskretisierung der eindimensionalen Wärmeleitungsgleichung mittels harmonischer Funktionen konnten wir die Zeitintegration analytisch durchführen. Um verschiedene Fehlerquellen zu vergleichen, kann man in (4.121) $\mathrm{d}a_m/\mathrm{d}t$ auch durch Vorwärtsdifferenzen approximieren. Dann lassen sich die beiden Fehlerquellen aus dem räumlichen spektralen und dem zeitlichen finiten Differenzen-Verfahren leicht identifizieren (siehe Tabelle 5.13 in Fletcher (1991*a*)).

In den vorangegangenen Beispielen konnten wir in den Ansätzen (4.110) und (4.118) je eine Funktion abspalten, welche die Randbedingungen exakt erfüllt. Das ist aber nicht immer in dieser Form möglich. Wenn z.B. die J Ansatzfunktionen die Randbedingungen nicht erfüllen, kann man die sogenannte ***τ-Methode*** verwenden. Dabei wird das Residuum nur auf $J - M$ Moden projiziert, wobei M die Anzahl der nicht erfüllten Randbedingungen ist. Um das System eindeutig zu lösen, nimmt man zu diesen $J - M$ Gleichungen dann noch M Gleichungen hinzu, die sich durch das Einsetzen des Ansatzes in die Randbedingungen ergeben. Dies wird anhand der Diffusionsgleichung mit Neumann-Randbedingungen in Fletcher (1991*a*) demonstriert.

4.6. Pseudospektrale Methode

Bei einer rein spektralen Methode, wie zum Beispiel der Galerkin-Methode, wird die gesamte Rechnung (Zeitintegration) im spektralen Raum durchgeführt (z.B. im Fourier-Raum oder im Chebyshev-Raum). Diese Vorgehensweise ist bei nichtlinearen Termen rechenintensiv, da man zumindest Doppelsummen (quadratische Nichtlinearität) auswerten muß.[32] Es stellt sich heraus, daß man den Rechenaufwand reduzieren kann, wenn man nur teilweise im spektralen Raum rechnet und gewisse Ausdrücke im Ortsraum berechnet. Eine solche Methode wird *pseudospektral* genannt. Um effizient pseudospektral rechnen zu können, werden schnelle Transformationen zwischen dem Ortsraum und dem spektralen Raum benötigt. Die Entdeckung der schnellen Fourier-Transformation (*Fast Fourier Transform*, FFT) war ein entscheidender Fortschritt, mit dem die pseudospektrale Methode gegenüber anderen (nicht-spektralen) Methoden konkurrenzfähig wurde. Im folgenden sollen einige Elemente pseudospektraler Methoden vorgestellt werden.

4.6.1. Fourier-Transformation

Da die schnellen Transformationen zwischen Orts- und Spektralraum eine Schlüsselstellung einnehmen, soll beispielhaft die schnelle Fourier-Transformation behandelt werden. Entsprechende schnelle Transformationen gibt es auch für andere Funktionensysteme, insbesondere auch für Chebyshev-Polynome, die im nächsten Unterkapitel vorgestellt werden.

[32] Siehe dazu Fußnote 27 auf S. 90 und auch weiter unten.

Diskrete Fouriertransformation

Die schnelle Fourier-Transformation (FFT) basiert auf der *diskreten Fourier-Transformation (DFT)*. Wir betrachten eine 2π-periodische Funktion $u(x)$, die an N äquidistanten Punkten x_j ausgewertet (ge*sampled*) wird. Die Intervallänge ist dann $\Delta x = 2\pi/N$ und die Stützpunkte befinden sich bei

$$x_j = \frac{2\pi j}{N}, \qquad \text{mit} \quad j = 1, ..., N. \tag{4.124}$$

Die Werte der betrachteten Funktion an den Stützstellen im Ortsraum sind

$$u_j = u(x_j). \tag{4.125}$$

Dann stellen wir die Funktion $u(x)$ durch die diskrete *Fourier-Reihe*[33] dar (siehe auch (2.36) und (2.37) in Kap. 2.4.3)

$$u_j = u(x_j) = \sum_{k=-K}^{K} \hat{u}_k \mathrm{e}^{\mathrm{i}kx_j}, \quad j = 1, ..., N. \tag{4.126}$$

Diese N Gleichungen bieten Informationen zur Bestimmung von $N = 2K+1$ unbekannten Amplituden $\hat{u}_k$. Sie sind gegeben durch die *diskrete Fourier-Transformation*

$$\hat{u}_k = \frac{1}{N} \sum_{m=1}^{N} u_m \mathrm{e}^{-\mathrm{i}kx_m}, \quad k = -K, ..., K. \tag{4.127}$$

Um diese Relation beweisen zu können, benötigen wir die *diskrete Orthogonalitätsrelation* für die harmonischen Funktionen $\mathrm{e}^{\mathrm{i}kx_j}$ des Fourier-Systems

$$\sum_{j=1}^{N} \mathrm{e}^{-\mathrm{i}m(2\pi j/N)} \mathrm{e}^{\mathrm{i}k(2\pi j/N)} = \sum_{j=1}^{N} \mathrm{e}^{\mathrm{i}(k-m)2\pi j/N} = \begin{cases} N, & \text{falls } k-m = nN,\ n \in \mathbb{Z} \\ 0, & \text{sonst.} \end{cases} \tag{4.128}$$

Diese Orthogonalitätsrelation kann man sich leicht klarmachen. Denn wenn $k-m = nN$ ist, dann ist der Exponent ein ganzzahliges Vielfaches von $2\pi\mathrm{i}$ und damit lauten alle Summanden $\mathrm{e}^{\mathrm{i}nj2\pi} = 1$. Der Summand mit Wert 1 taucht dann genau N-mal auf. Falls jedoch $k - m = q \in \mathbb{Z} \neq nN$ ist, dann liegen die Summanden $\mathrm{e}^{\mathrm{i}q2\pi(j/N)}$ für $j = 1, ..., N$ gleichmäßig verteilt auf dem Einheitskreis in der komplexen Ebene. Die komplexen Zeiger überstreichen dabei den Winkelbereich von $2\pi q/N$ $(j = 1)$

[33] Beachte, daß man die Summationsbereiche $k = -K, ..., K$ und $j = 1, ..., N$ beliebig verschieben kann, wenn die Funktionswerte u_j periodisch in j fortgesetzt werden (mit Periode N), was wir hier annehmen. Denn der Faktor $\mathrm{e}^{\mathrm{i}kx_j} = \mathrm{e}^{\mathrm{i}kj2\pi/N}$ ist periodisch in k mit Periode N, weil $k \to k+N$ nur den Zusatzfaktor $\mathrm{e}^{\mathrm{i}2\pi j} = 1$ liefert. Außerdem ist auch die Fourier-Transformierte $\hat{u}_k$ N-periodisch in k (siehe (4.127)). Dies erklärt den nur formalen Unterschied zwischen (4.126) und (2.36).

bis $2\pi q$ ($j = N$) mit dem Inkrement $2\pi q/N$. Der Einheitskreis wird also q-mal überstrichen. Wegen der Gleichverteilung der Zeiger kompensieren sie sich zu Null.

Die *Orthogonalitätsrelation* (4.128) können wir nun verwenden, um die diskrete Fourier-Transformation zu beweisen. Wenn wir (4.127) in (4.126) einsetzen, erhalten wir[34]

$$\begin{aligned} u_j &= \sum_{k=-K}^{K} \left(\frac{1}{N} \sum_{m=1}^{N} u_m \mathrm{e}^{-\mathrm{i}kx_m} \right) \mathrm{e}^{\mathrm{i}kx_j} = \frac{1}{N} \sum_{m=1}^{N} u_m \sum_{k=-K}^{K} \mathrm{e}^{\mathrm{i}k(x_j - x_m)} \\ &= \frac{1}{N} \sum_{m=1}^{N} u_m \underbrace{\sum_{k=-K}^{K} \mathrm{e}^{2\pi \mathrm{i}k(j-m)/N}}_{=N\delta_{j,m},\ (4.128)} = \sum_{m=1}^{N} u_m \delta_{j,m} = u_j. \end{aligned} \tag{4.129}$$

Damit haben wir bewiesen, daß (4.127) die Umkehrung von (4.126) ist.

Wenn $u_j \in \mathbb{R}$ reell ist, dann ist $u_j = u_j^*$ und es gilt

$$\hat{u}_{-k} = \frac{1}{N} \sum_{m=1}^{N} u_m \mathrm{e}^{+\mathrm{i}kx_m} \overset{u_m = u_m^*}{=} \left(\frac{1}{N} \sum_{m=1}^{N} u_m \mathrm{e}^{-\mathrm{i}kx_m} \right)^* = \hat{u}_k^*. \tag{4.130}$$

Damit sind nur noch $K+1$ Amplituden unabhängig voneinander, was den Rechenaufwand verringert.

Diskrete Fouriertransformation als lineare Operation

Wenn wir die Gitterpunkte (4.124) einsetzen, erhalten wir aus (4.126) für den Vektor der Funktionswerte

$$\boldsymbol{u} = u_j = \sum_{k=-K}^{K} \hat{u}_k \underbrace{\mathrm{e}^{\mathrm{i}kj2\pi/N}}_{G_{jk}} = \mathsf{G} \cdot \hat{\boldsymbol{u}}. \tag{4.131}$$

Wir sehen also, daß die Abbildung $\hat{\boldsymbol{u}} \to \boldsymbol{u}$ der Amplituden auf die Funktionswerte (Transformation aus dem Fourier-Raum in den Ortsraum) eine lineare Transformation ist, die durch eine Multiplikation mit der symmetrischen Matrix

$$\mathsf{G} = G_{jk} = \mathrm{e}^{\mathrm{i}kj2\pi/N} \tag{4.132}$$

erreicht wird. Diese Transformation kostet $O(N^2)$ Operationen (Multiplikationen). Entsprechend kann man auch die diskrete Fouriertransformation (4.127) als Multiplikation einer Matrix mit dem Vektor der Funktionswerte schreiben

$$\hat{\boldsymbol{u}} = \mathsf{F} \cdot \boldsymbol{u} \tag{4.133}$$

[34] Beachte, daß der Summationsbereich in der Orthogonalitätsrelation keine Rolle spielt, solange er über N zusammenhängende ganze Zahlen geht.

mit der Transformationsmatrix (vgl. (4.127))

$$\mathsf{F} = F_{km} = \frac{1}{N} \mathrm{e}^{-\mathrm{i}km2\pi/N}. \tag{4.134}$$

Mit Hilfe der diskreten Orthogonalitätsrelation (4.128) kann man zeigen, daß F die Inverse von G ist: $\mathsf{F} = \mathsf{G}^{-1}$. Denn es gilt[35]

$$\mathsf{G} \cdot \mathsf{F} = \frac{1}{N} \sum_{k=-K}^{K} \mathrm{e}^{2\pi\mathrm{i}kj/N} \mathrm{e}^{-2\pi\mathrm{i}km/N} = \frac{1}{N} \sum_{k=1}^{N} \mathrm{e}^{2\pi\mathrm{i}k(j-m)/N} = \delta_{j,m} = \mathsf{I}. \tag{4.135}$$

Schnelle Fourier-Transformation

Wie wir gesehen haben, kostet die Berechnung der Fourier-Transformierten $\hat{\boldsymbol{u}}$ aus den Funktionswerten $\boldsymbol{u}$ (bzw. umgekehrt) eine Anzahl von $O(N^2)$ Operationen (Matrix-Multiplikation), wenn man die normale Matrix-Vektor-Multiplikation verwendet. Die Anzahl der Operationen kann man reduzieren, wenn man anders vorgeht.

Bei der *schnellen Fouriertransformation* (FFT) wird die ursprüngliche Fouriertransformation für N Punkte sukzessive auf Fourier-Transformationen mit jeweils der halben Anzahl von Punkten zurückgeführt ($N \to N/2 \to N/4 \to ...$). Diese Aufspaltung wird solange fortgesetzt bis nur noch die triviale Fouriertransformation für einen einzigen Punkt übrig bleibt.

Mit Hilfe der FFT kann man die Anzahl der erforderlichen Operationen auf $O(N \log_2 N)$ reduzieren. Bei großen Werten von N, sagen wir $N = 128$, ist die FFT damit um einen Faktor von $N/\log_2 N = 128/7 \approx 18$ schneller als die Matrix-Multiplikation.[36] In drei Dimensionen ergibt sich damit schon der beträchtliche Faktor $18^3 \approx 6 \times 10^3$. Je größer N desto größer ist die relative Ersparnis.

Um die Vorteile der FFT nutzen zu können, darf N keine Primzahl sein. Die größte Beschleunigung erhält man, wenn N eine Potenz von 2 ist: $N = 2^p$, $p \in \mathbb{N}$.[37] Dann kann man die diskrete Fourier-Transformation für N Stützstellen auf zwei diskrete Fourier-Transformationen mit jeweils $N/2$ Stützstellen zurückführen, wenn wir die Summe in zwei Teilsummen spalten, die sich nur über die geraden bzw. die ungeraden Indizes erstrecken. Wenn wir die Summation von 0 bis $N-1$ laufen lassen und den Faktor N^{-1} vorerst weglassen, erhalten wir mit $x_j = 2\pi j/N$ aus (4.127) die diskrete Fourier-Transformierte

$$\hat{u}_k = \sum_{j=0}^{N-1} u_j \mathrm{e}^{-\mathrm{i}k2\pi j/N} = \sum_{j=0}^{N/2-1} u_{2j} \mathrm{e}^{-\mathrm{i}k2\pi(2j)/N} + \sum_{j=0}^{N/2-1} u_{2j+1} \mathrm{e}^{-\mathrm{i}k2\pi(2j+1)/N}$$

[35] Der Summationsindex kann verschoben werden, da die Matrizen in jedem Index periodisch sind mit Periode N.

[36] Dies gilt nur im Prinzip. In der Praxis kommen noch andere Faktoren hinzu, vgl. Tab. 4.3.

[37] In (4.126) war N ungerade. Man kann aber auch N gerade wählen und die Summe in (4.126) von 1 bis N oder von 0 bis $N-1$ nehmen, vgl. Fußnote 33 auf S. 97.

$$= \underbrace{\sum_{j=0}^{N/2-1} u_{2j} \mathrm{e}^{-\mathrm{i}k2\pi j/(N/2)}}_{:=\hat{u}_k^{\text{even}}:=\hat{u}_k^{(0)},\ k=0,\ldots,N/2-1} + \underbrace{\mathrm{e}^{-\mathrm{i}k2\pi/N}}_{:=W^k} \underbrace{\sum_{j=0}^{N/2-1} u_{2j+1} \mathrm{e}^{-\mathrm{i}k2\pi j/(N/2)}}_{:=\hat{u}_k^{\text{odd}}:=\hat{u}_k^{(1)},\ k=0,\ldots,N/2-1}, \tag{4.136}$$

mit $W := \mathrm{e}^{-2\pi\mathrm{i}/N}$.[38]

Man sieht, daß man die Fourier-Transformierte $\hat{u}_k$ als Summe darstellen kann, die aus den Fourier-Transformierten der Funktionswerte mit geraden Indizes und derjenigen mit ungeraden Indizes besteht. Beide Fourier-Transformationen haben dann jeweils die halbe Länge $N/2$. Dadurch haben wir schon eine Ersparnis: Die beiden Fourier-Transformationen der Länge $N/2$ kosten nur $2 \times (N/2)^2 = N^2/2$ Operationen. Hinzu kommen N Multiplikationen mit einem komplexen Exponentialfaktor und N Additionen (wenn man sie mitrechnen will); in Summe also

$$\frac{N^2}{2} + 2N < N^2, \quad \text{falls } N > 4. \tag{4.137}$$

Bei dieser Betrachtung haben wir ausgenutzt, daß wir die Produkte (in den Summen über j) nur für die halbe Anzahl $N/2$ der k-Werte bilden müssen, und nicht für alle k-Werte aus dem eigentlichen Bereich $[0, N-1]$. Dies ist so, weil die Summanden in den FTs der halben Länge $\hat{u}_k^{\text{odd}}$ und $\hat{u}_k^{\text{even}}$ periodisch in k sind mit der Periode $N/2$. Denn für festes j gilt

$$\mathrm{e}^{-\mathrm{i}k2\pi j/(N/2)} \overset{k\to k+N/2}{=} \mathrm{e}^{-\mathrm{i}k2\pi j/(N/2)} \underbrace{\mathrm{e}^{-\mathrm{i}2\pi j}}_{=1}. \tag{4.138}$$

Wenn man dies beachtet, erhält man für $k = 0, 1, 2, ..., N/2-1$

$$\hat{u}_k = \hat{u}_k^{(0)} + W^k \hat{u}_k^{(1)}, \tag{4.139a}$$

$$\hat{u}_{k+N/2} = \hat{u}_k^{(0)} - W^k \hat{u}_k^{(1)}. \tag{4.139b}$$

Das Minuszeichen vor W^k in der zweiten Gleichung kommt von $W^{k+N/2} = \mathrm{e}^{-2\pi\mathrm{i}(k+N/2)/N} = \mathrm{e}^{-2\pi\mathrm{i}k/N}\mathrm{e}^{-\mathrm{i}\pi} = -W^k$.

Wir können nun die Zerlegung weiterführen und die beiden Summen in (4.136) wieder in zwei Summen über die geraden und die ungeraden Indizes aufspalten. Als Beispiel sei die zweite Unterteilung durchgeführt,

$$\hat{u}_k \overset{(4.136)}{=} \underbrace{\sum_{j=0}^{N/4-1} u_{2(2j)} \mathrm{e}^{-\mathrm{i}k2\pi(2j)/(N/2)} + \sum_{j=0}^{N/4-1} u_{2(2j+1)} \mathrm{e}^{-\mathrm{i}k2\pi(2j+1)/(N/2)}}_{\hat{u}_k^{(0)}}$$

[38] Beachte: $W^k = \left(\mathrm{e}^{-2\pi\mathrm{i}/N}\right)^k = \mathrm{e}^{-2\pi\mathrm{i}k/N}$.

Tabelle 4.2.: Zuordnung der Werte an den Gitterpunkten j zu den einzelnen FFTs bei sukzessiver Halbierung der Punktzahl der FFTs. Eine 0 steht für gerade Punkte und eine 1 für ungerade Punkte. Bei jeder Aufteilung in zwei Fouriertransformationen für die geraden und die ungeraden Punkte, aber der halben Länge, wird dem Index (Bezeichnung der FFT) eine 0 (gerade) oder eine 1 (ungerade) angehängt, je nachdem, ob es sich bei der neuen Anordnung um eine gerade oder eine ungerade Position in der momentanen Fourierreihe handelt.

Gitterpunkt j	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
	0	1	0	1	0	1	0	1
FFT-Zuordnung	00	10	01	11	00	10	01	11
	000	100	010	110	001	101	011	111
Bit-Inversion	**000**	**001**	**010**	**011**	**100**	**101**	**110**	**111**

$$
\begin{aligned}
&+ W^k \underbrace{\left(\sum_{j=0}^{N/4-1} u_{2(2j)+1} \mathrm{e}^{-\mathrm{i}k2\pi(2j)/(N/2)} + \sum_{j=0}^{N/4-1} u_{2(2j+1)+1} \mathrm{e}^{-\mathrm{i}k2\pi(2j+1)/(N/2)} \right)}_{\hat{u}_k^{(1)}} \\
&= \sum_{j=0}^{N/4-1} u_{2(2j)} \mathrm{e}^{-\mathrm{i}k2\pi j/(N/4)} + \sum_{j=0}^{N/4-1} u_{2(2j+1)} \mathrm{e}^{-\mathrm{i}k(2j+1)\pi/(N/4)} \\
&+ W^k \left(\sum_{j=0}^{N/4-1} u_{2(2j)+1} \mathrm{e}^{-\mathrm{i}k2j\pi/(N/4)} + \sum_{j=0}^{N/4-1} u_{2(2j+1)+1} \mathrm{e}^{-\mathrm{i}k(2j+1)\pi/(N/4)} \right) \\
&= \underbrace{\underbrace{\sum_{j=0}^{N/4-1} u_{2(2j)} \mathrm{e}^{-\mathrm{i}k2\pi j/(N/4)}}_{=\hat{u}_k^{(00)},\ k=0,\ldots,N/4-1} + W^{2k} \underbrace{\sum_{j=0}^{N/4-1} u_{2(2j+1)} \mathrm{e}^{-\mathrm{i}k2j\pi/(N/4)}}_{=\hat{u}_k^{(01)},\ k=0,\ldots,N/4-1}}_{\hat{u}_k^{(0)}} \\
&+ W^k \underbrace{\left(\underbrace{\sum_{j=0}^{N/4-1} u_{2(2j)+1} \mathrm{e}^{-\mathrm{i}k2j\pi/(N/4)}}_{=\hat{u}_k^{(10)},\ k=0,\ldots,N/4-1} + W^{2k} \underbrace{\sum_{j=0}^{N/4-1} u_{2(2j+1)+1} \mathrm{e}^{-\mathrm{i}k2j\pi/(N/4)}}_{=\hat{u}_k^{(11)},\ k=0,\ldots,N/4-1} \right)}_{\hat{u}_k^{(1)}} \\
&= \ldots
\end{aligned}
\tag{4.140}
$$

Hierbei können die Exponentialfaktoren in den Summen ebenfalls als Potenzen von W ausgedrückt werden. Wenn $N = 2^p$ eine ganzzahlige Potenz von 2 ist, kann man diese Aufteilung bis zu dem Punkt durchführen, bei dem die Summe nur noch über einen einzigen Funktionswert zu bilden ist. Diese Prozedur ist anschaulich in Tab. 4.2 dargestellt. Bei jeder Zerlegung in gerade bzw. ungerade Terme erhält der zur Kennzeichnung hochgestellte Index von $\hat{u}_k$ die Ziffer 0 (gerade) oder 1

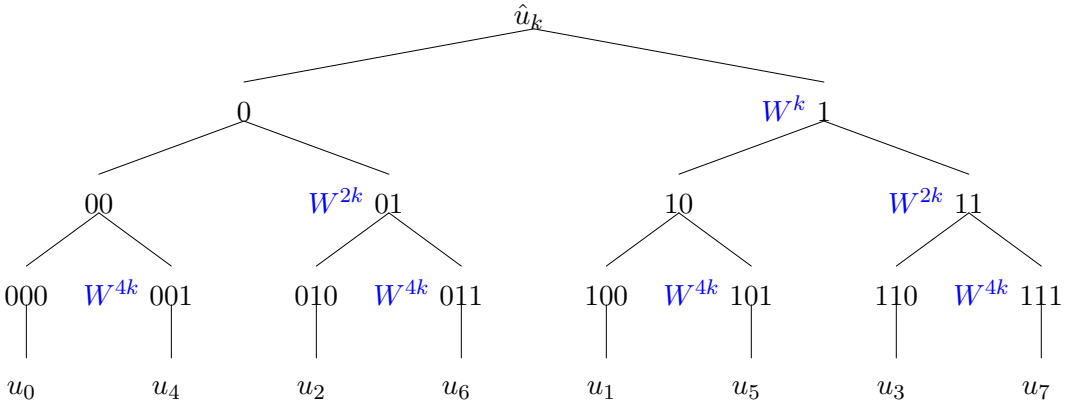


Abbildung 4.15.: Baumdiagramm der Aufspaltung einer FFT für $N = 8$. Auf jeder Stufe muß man eine gewichtete Summe über die beiden Terme nehmen, die von der Stelle verzweigen. Die Anzahl der k-Werte, für die man das machen muß, halbiert sich aber mit jeder Stufe. Die Ausgangsreihenfolge ergibt sich durch die Binärinversion des Gitterindex.

(ungerade) angehängt. Bei $N = 8$ muß man die Fouriertransformation dreimal ($= \log_2(8)$) in gerade und ungerade Beiträge zerlegen. In der Tabelle ist die jeweilige Zuordnung jedes einzelnen Summanden entweder zu den geraden oder ungeraden Punkten der jeweiligen Transformation dargestellt. Wenn man auf dem niedrigsten Niveau angekommen ist, besteht die Fourier-Transformation nur noch aus einem einzigen Punkt. Wenn man die Kodierung (den hochgestellten Index von $\hat{u}_k$) der resultierenden 1-Punkt-FFTs bitweise invertiert (*Bit-Inversion*), erhält man gerade die Binär-Darstellung des Gitter-Index j des jeweiligen Funktionswertes (Tab. 4.2). Der FFT-Algorithmus besteht dann in einer sukzessiven gewichteten Addition der im Baumdiagramm Abb. 4.15 benachbarten Werte.[39]

In rekursiver Form kann man den Algorithmus (nach A. Bekele) wie folgt darstellen:

```
procedure FFT(A)

Input: An array of complex values which has a size of 2^m for m ≥ 0
Output: An array of complex values which is the DFT of the input

N := A.length
   if N = 1
      return A
   else
      W_N := e^(2πi/N)
```

[39]Die immer wiederkehrenden Operationen können auch durch das sogenannte *Butterfly-*

```
    W := 1
    A_even := (A_0, A_2, ...., A_{N-2})
    A_odd := (A_1, A_3, ...., A_{N-1})
    Y_even := FFT(A_even)
    Y_odd := FFT(A_odd)
    for j := 0 to N/2 - 1
        Y[j] = Y_even[j] + W * Y_odd[j]
        Y[j + N/2] = Y_even[j] - W * Y_odd[j]
        W := W * W_N
    end
    return Y
end
```

Um den numerischen Aufwand abzuschätzen, beachten wir, daß wir in jedem Schritt N Werte berechnen müssen, zum Beispiel

$$W_{00}\hat{u}_k^{(00)} + W_{01}\hat{u}_k^{(01)} + W_{10}\hat{u}_k^{(10)} + W_{11}\hat{u}_k^{(11)} \tag{4.141}$$

Dazu benötigt man bei *jeder* Unterteilung $O(N)$ Operationen, denn die Anzahl der Summanden erhöht sich zwar mit der Tiefe der Unterteilung, aber in demselben Maße nimmt die Zahl der Werte k, für welche man die Summen berechnen muß, ab (wegen der Periodizität des Ausdrucks). In dem Beispiel haben wir $O(4)$ Summanden, die wiederum aus Summen bestehen, die aber nur für $N/4$ k-Werte berechnet werden müssen. Da wir $\log_2 N$ Unterteilungen haben, beträgt der Gesamtaufwand zur Berechnung von $\hat{u}_k$ nur *$O(N \log_2 N)$ Operationen*. Dies muß mit den $O(N^2)$ Operationen für die DFT verglichen werden.

4.6.2. Chebyshev Polynome

Definition und allgemeine Eigenschaften

Zur Beschreibung von Problemen mit periodischen Randbedingungen sind Fouriermoden am besten geeignet. In der Strömungsmechanik treten aber meist nichtperiodische Randbedingungen auf, wenn man zum Beispiel an die Strömung in einem Kanal denkt. In transversaler Richtung (senkrecht zur Hauptströmungsgrichtung) müssen wir die Variablen auf einem endlichen Intervall beschreiben, wobei in Randnähe häufig große Gradienten auftreten (Grenzschichten). Für eine spektrale Behandlung bieten sich dann Chebyshev-Polynome als Ansatzfunktionen an.

Die *Chebyshev-Polynome* $\mathrm{T}_k(x)$, $k \in \mathbb{N}_0$ sind auf $x \in [-1, 1]$ definiert und lassen sich mit den ersten Chebyshev-Polynomen $\mathrm{T}_0(x) = 1$ und $\mathrm{T}_1(x) = x$ aus der *Rekursionsformel* (*recurrence relation*) gewinnen (Abramowitz and Stegun, 1972)

$$\mathrm{T}_{k+1}(x) = 2x\mathrm{T}_k(x) - \mathrm{T}_{k-1}(x). \tag{4.142}$$

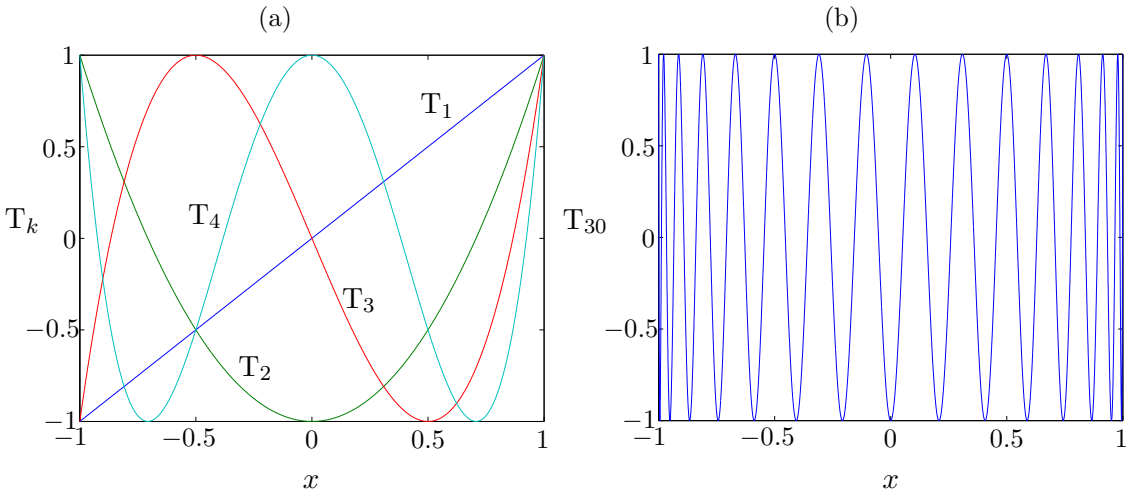


Abbildung 4.16.: Chebyshev-Polynome $\mathrm{T}_k(x)$ für $k = 1, 2, 3, 4$ (a) und $k = 30$ (b).

Die Chebyshev-Polynome sind *orthogonal* mit dem *Skalarprodukt*

$$\langle \mathrm{T}_n | \mathrm{T}_m \rangle := \int_{-1}^{1} \frac{\mathrm{T}_n(x)\mathrm{T}_m(x)}{\sqrt{1-x^2}}\,\mathrm{d}x = \frac{\pi}{2} c_m \delta_{n,m}, \quad \text{wobei} \quad c_m = \begin{cases} 2, & m = 0, \\ 1, & m \neq 0. \end{cases} \tag{4.143}$$

Mit $x = \cos\theta$ lassen sich die Chebyshev-Polynome auch in der Form

$$\mathrm{T}_k(x) = \cos(k\theta) = \cos\left[k \arccos(x)\right] \tag{4.144}$$

darstellen, was man mit Hilfe der Rekursion (4.142) beweisen kann. Diese Darstellung ist hilfreich bei der Berechnung bestimmter Funktionswerte (Extrema und Nullstellen). Die gute Randauflösung der Chebyshev-Polynome rührt daher, daß die Chebyshev-Polynome mit wachsender Ordnung n immer schneller in der Nähe von $x = \pm 1$ oszillieren (Abb. 4.16b). Mit $\arccos(1) = 0$ und $\arccos(-1) = \pi$ gilt an den Rändern $\mathrm{T}_k(1) = \cos(0) = 1$ und $\mathrm{T}_k(-1) = \cos(k\pi) = (-1)^k$ (Abb. 4.16a).

Chebyshev-Kollokation

Die spektrale Darstellung einer Funktion $u(x,t)$ mit Hilfe von Chebyshev-Polynomen lautet

$$u(x,t) = \sum_{k=0}^{N} \hat{u}_k(t)\mathrm{T}_k(x). \tag{4.145}$$

Hierbei wollen wir die Zeit t noch nicht diskretisieren. Im Rahmen der Kollokationsmethode betrachten wir die Funktion u im Ortsraum an den $N+1$ diskreten *Kollokationspunkten* x_j, $j = 0, 1, ..., N$, wobei wir definieren: $u_j = u(x_j)$. Die Kollokationspunkte kann man im Prinzip frei wählen. Man kann aber zeigen, daß es gewisse

Diagramm symbolisiert werden.

optimale Kollokationspunkte gibt, die den Diskretisierungsfehler minimieren.[40] Die optimalen Stützstellen (d.h. Kollokationspunkte) hängen von den Ansatzfunktionen ab. Für Chebyshev-Polynome werden meistens die sogenannten *Gauß-Lobatto-Punkte* verwendet. Sie sind durch die Extrema des höchsten Chebyshev-Polynoms bestimmt, d.h. durch $\mathrm{T}'_N(x_j) = -N\sin(N\theta_j)\theta'_j = 0$ inklusive der Randpunkte, und lauten[41]

$$x_j = \cos\left(\frac{\pi j}{N}\right), \qquad j \in [0, ..., N]. \tag{4.146}$$

Das Gitter ist also nicht homogen, sondern gemäß der cos-Funktion zum Rand hin verdichtet. Damit lautet die diskrete Darstellung der unbekannten Funktion an den Kollokationspunkten

$$u_j(t) = \boldsymbol{u} = \sum_{k=0}^{N} \hat{u}_k(t) \underbrace{\mathrm{T}_k(x_j)}_{\mathsf{T}} \qquad \text{bzw.} \quad \boldsymbol{u} = \mathsf{T}\cdot\hat{\boldsymbol{u}}. \tag{4.147}$$

Hierbei sind $\hat{u}_k(t)$ die Amplituden der spektralen Darstellung von $\boldsymbol{u}$. An der Darstellung (4.147) sieht man, daß die Transformation vom spektralen Raum $\hat{u}_k(t)$ in den Ortsraum $u_j(t)$ als Matrixmultiplikation der Amplituden $\hat{u}_k$ mit der Matrix T aufgefaßt werden kann, ganz analog zur diskreten Fouriertransformation (vgl. (4.131)). Die Komponenten der Matrix T sind bekannt. Sie ergeben sich einfach aus den Chebyshev-Polynomen und den gegebenen Kollokationspunkten.

Die erste und zweite Ableitung an den Kollokationspunkten x_j lautet dann

$$\left.\frac{\partial u}{\partial x}\right|_{x_j} = \boldsymbol{u}^{(1)} = \sum_{k=0}^{N} \hat{u}_k(t) \underbrace{\left.\frac{\partial \mathrm{T}_k(x)}{\partial x}\right|_{x_j}}_{\text{1. Abl.–Mat.}} = \mathsf{D}^{(1)}\cdot\hat{\boldsymbol{u}}, \tag{4.148a}$$

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_{x_j} = \boldsymbol{u}^{(2)} = \sum_{k=0}^{N} \hat{u}_k(t) \underbrace{\left.\frac{\partial^2 \mathrm{T}_k(x)}{\partial x^2}\right|_{x_j}}_{\text{2. Abl.–Mat.}} = \mathsf{D}^{(2)}\cdot\hat{\boldsymbol{u}}. \tag{4.148b}$$

Die Ableitungen erhält man auch wieder durch eine einfache Matrix-Multiplikation der Amplituden mit der jeweiligen Ableitungsmatrix $\mathsf{D}^{(1)}$ bzw. $\mathsf{D}^{(2)}$. Die Matrizen T, $\mathsf{D}^{(1)}$, $\mathsf{D}^{(2)}$, etc. brauchen nur einmal zu Beginn einer jeden Simulation berechnet werden. Ganz analog kann man auch beim Fourier-System die Ableitungen als Matrix-Multiplikationen darstellen.

Transformation in den spektralen Raum

Für die Transformation der Werte u_j in den spektralen Raum müssen die Amplituden $\hat{u}_k$ berechnet werden. Nach (4.147) erfordert dies im Prinzip die Inversion der

[40] Siehe dazu das Kapitel über *Gauß-Quadratur* im Skriptum der Vorlesung *Numerische Methoden der Ingenieurwissenschaften* (LVA-Nr. 322.036).

[41] Man kann auch die sogenannten Chebyshev-Gauß-Punkte $z_j = \cos\left[(j+1/2)\,\pi/N\right]$ mit $j \in$

dichtbesetzten Matrix $\mathsf{T} = \mathrm{T}_k(x_j)$. Zwar kann man das lineare Gleichungssystem (4.147) mit Hilfe der *Gauß-Elimination* (siehe Kap. 5.1.1) lösen, aber bei $N+1$ Unbekannten kostet dieses Verfahren $O\left[(N+1)^3\right]$ Rechenoperationen.

Wenn man die Chebyshev-Polynome jedoch an den Gauß-Lobatto-Punkten (4.146) auswertet, kann man die Transformationsmatrix

$$\mathsf{T} = \mathrm{T}_k(x_j) = \cos[k \arccos(x_j)] \overset{(4.146)}{=} \cos\left\{k \arccos\left[\cos\left(\frac{\pi j}{N}\right)\right]\right\} = \cos\left(\frac{k\pi j}{N}\right) \tag{4.149}$$

leicht invertieren. Zur Berechnung der Inversen benötigt man (wie beim Fourier-System) die diskrete Version der Orthogonalitätsbedingung (4.143). Dazu wird das Integral in der kontinuierlichen Orthogonalitätsrelation (4.143) mittels Gauß-Quadratur für eine endliche Anzahl $N+1$ von Stützstellen approximiert. Man kann zeigen, daß diese Approximation bei Verwendung der Gauß-Lobatto-Punkte (4.146) exakt wird (Peyret, 2002). Man erhält dann die *diskrete Orthogonalitätsrelation*

$$\sum_{j=0}^{N} \frac{1}{\bar{c}_j} \mathrm{T}_k(x_j)\mathrm{T}_l(x_j) = \frac{\bar{c}_k}{2} N \delta_{k,l}, \qquad \text{mit} \quad \bar{c}_k = \begin{cases} 2, & k = 0, N, \\ 1, & 1 \le k \le N-1. \end{cases} \tag{4.150}$$

Damit kann man die Inverse von T berechnen und erhält[42]

$$\mathsf{T}^{-1} = \frac{2\cos\left(k\pi j/N\right)}{\bar{c}_k \bar{c}_j N}. \tag{4.151}$$

Die inverse Transformation braucht nur einmal zu Beginn einer Rechnung ermittelt werden. Mit Kenntnis der Inversen kostet die Berechnung von $\hat{u}_k = \hat{\boldsymbol{u}}$ (d.h. die Transformation in den spektralen Raum) mittels Matrixmultiplikation dann nur noch $O[(N+1)^2]$ Operationen. Für $N \gtrsim 100$ ist aber eine schnelle Chebyshev-Transformation günstiger (discrete Chebyshev Transformation: DCT). Dazu nutzt man die Darstellung der Chebyshev Polynome durch Cosinus-Funktionen (4.144), was zusammen mit den Gauss-Lobatto-Punkten (4.146) auf eine diskrete Cosinus-

$[0, N-1]$ verwenden. An ihnen besitzt das höchste Polynom seine Nullstellen: $\mathrm{T}_N(z_j) = 0$.

[42] Wenn man (4.147) mit $\mathrm{T}_l(x_j)/\bar{c}_j$ multipliziert und über $j = 0$ bis N summiert, wird die Komponente (Amplitude) $\hat{u}_l$ herausprojiziert. Man erhält dann

$$\sum_{j=0}^{N} \frac{\mathrm{T}_l(x_j)}{\bar{c}_j} u_j(t) = \sum_{k=0}^{N} \hat{u}_k(t) \underbrace{\sum_{j=0}^{N} \frac{\mathrm{T}_k(x_j)\mathrm{T}_l(x_j)}{\bar{c}_j}}_{(4.150)} = \sum_{k=0}^{N} \hat{u}_k(t) \frac{\bar{c}_k}{2} N \delta_{k,l} = \hat{u}_l(t) \frac{\bar{c}_l}{2} N.$$

Also ist mit $\mathrm{T}_l(x_j) = \cos(l\pi j/N)$

$$\hat{u}_l(t) = \frac{2}{N\bar{c}_l} \sum_{j=0}^{N} \frac{\mathrm{T}_l(x_j)}{\bar{c}_j} u_j(t) = \sum_{j=0}^{N} \underbrace{\frac{2\cos(l\pi j/N)}{\bar{c}_l \bar{c}_j N}}_{\mathsf{T}^{-1}} u_j(t).$$

Reihe führt, die man mittels FFT transformieren kann. Man benötigt zur Berechnung von $\hat{\boldsymbol{u}}$ dann nur noch $O\left[(N+1)\times\log_2(N+1)\right]$ Operationen (Canuto et al., 1988; Peyret, 2002).

4.6.3. Ableitungsoperatoren für Chebyshev-Kollokation

Vollständige Berechnung der Ableitungen im Ortsraum

In (4.148) hatten wir die räumliche Ableitung ausgedrückt durch eine lineare Transformation (Matrixmultiplikation) der spektralen Komponenten $\hat{u}_k$ (Chebyshev-Amplituden). Da auch die Transformation von $\boldsymbol{u}$ in den spektralen Raum ($\hat{\boldsymbol{u}}$) einer Matrixmultiplikation entspricht, kann man die Ableitung vollständig im Ortraum als Matrixmultiplikation schreiben. Mit (4.148) und (4.147) erhalten wir

$$\boldsymbol{u}^{(1)} \overset{(4.148)}{=} \mathsf{D}^{(1)}\cdot\hat{\boldsymbol{u}} \overset{(4.147)}{=} \underbrace{\mathsf{D}^{(1)}\cdot\mathsf{T}^{-1}}_{\mathcal{D}}\cdot\boldsymbol{u}. \tag{4.152}$$

Damit ist

$$\boldsymbol{u}^{(1)} = \mathcal{D}\cdot\boldsymbol{u}. \tag{4.153}$$

Nach einigen algebraischen Umformungen (siehe Peyret (2002)) erhält man dann den Ableitungsoperator für den Ortsraum (und für die Gauß-Lobatto-Punkte)

$$\mathcal{D} = \begin{cases} \mathcal{D}_{i,j} &= \dfrac{\bar{c}_i}{\bar{c}_j}\dfrac{(-1)^{i+j}}{x_i - x_j}, & \text{falls } i\neq j,\ 0\le(i,j)\le N,\\ \mathcal{D}_{i,i} &= -\dfrac{x_i}{2\left(1-x_i^2\right)}, & \text{falls } 1\le i\le N-1,\\ \mathcal{D}_{0,0} &= \dfrac{2N^2+1}{6} = -\mathcal{D}_{N,N}. \end{cases} \tag{4.154}$$

Die zweite Ableitung ergibt sich dementsprechend als

$$\boldsymbol{u}^{(2)} = \mathcal{D}\cdot(\mathcal{D}\cdot\boldsymbol{u}) = \mathcal{D}^2\cdot\boldsymbol{u}. \tag{4.155}$$

Die Matrizen sind in Canuto et al. (1988) und Peyret (2002) angegeben. Ganz analog kann man die lineare Operation der Ableitung auch vollständig in spektralen Raum formulieren (siehe Anhang C). Bei einem festen Gitter braucht die Matrix $\mathcal{D}$ nur einmal zu Beginn der Rechnung ermittelt werden.

Als Beispiel ist in Abb. 4.17 die Approximation der ersten Ableitung mittels Matrix-Multiplikation mit $\mathcal{D}$ entsprechend (4.154) dargestellt. Für $N=8$ ist die Ableitung sehr fehlerhaft. Für $N=16$ ergibt sich für das verwendete Beispiel schon eine sehr genaue Darstellung der ersten Ableitung.

Rekursive Berechnung der Ableitungen

Der numerische Aufwand zur Berechnung von Ableitungen hängt davon ab, ob die Ableitungen im Orts- oder im spektralen Raum durchgeführt werden. Die direkte

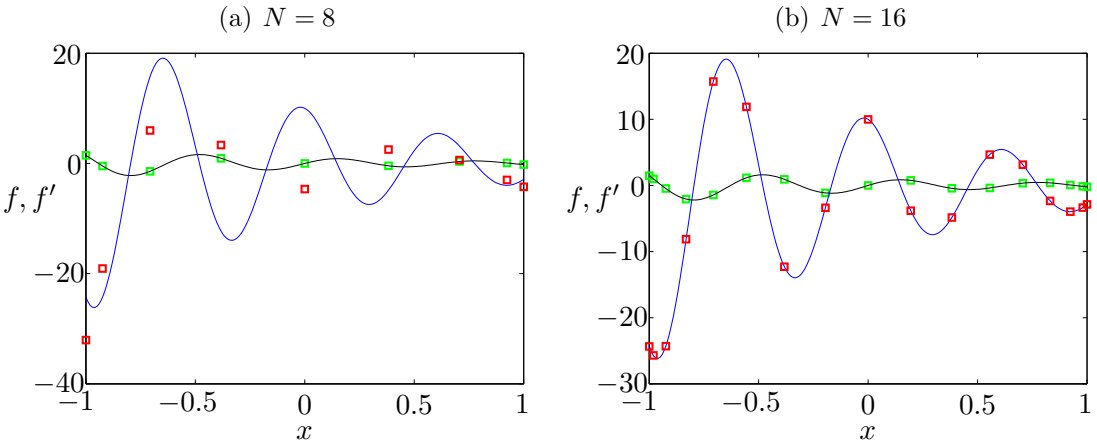


Abbildung 4.17.: Kollokationsableitung von $f(x) = \mathrm{e}^{-x}\sin(10x)$ mittels Matrix-Multiplikation nach (4.153) und der Ableitungsmatrix (4.154). Die durchgezogenen Kurven zeigen die exakte Funktion $f(x)$ (schwarz) und ihre erste Ableitung $f'(x)$ (blau). Die grünen Quadrate zeigen $\boldsymbol{f} = f_i = f(x_i)$ an den Gauß-Lobatto-Punkten (4.146) und die roten Quadrate zeigen die Chebshev-Ableitung $\boldsymbol{f}^{(1)} = \mathcal{D} \cdot \boldsymbol{f}$ für $N = 8$ (a) und $N = 16$ (b).

Berechnung der ersten bzw. zweiten Ableitung im Ortsraum an allen $N + 1$ Kollokationspunkten nach (4.148) für gegebene Amplituden kostet $O[(N + 1)^2]$ Operationen. Die Ableitungsmatrizen sind voll besetzt. Dies gilt für die Fourier- wie auch die Chebyshev-Methode. Wenn die Ableitungen für die Fourier-Methode aber vollständig im Fourier-Raum durchgeführt werden, sind nur $O(N)$ Operationen nötig, denn die Ableitung im Fourier-Raum ist nur eine Multiplikation mit ik. Die zugehörige Ableitungsmatrix ist diagonal!

Für Chebyshev-Polynome ist die Ableitungsmatrix im Ortsraum voll besetzt und die Ableitung kostet $O[(N+1)^2]$ Operationen. Die Ableitungsmatrix im Chebyshev-Raum ist jedoch eine voll besetzte obere Dreiecksmatrix. Für $N \lesssim 100$ verwendet man deshalb meistens (4.148), während man für größere Werte von N die Funktion u vor der Ableitung in den Chebyshev-Raum transformiert, dann ableitet und dann wieder in den Ortsraum zurücktransformiert. Da die Ableitungsmatrix im Chebyshev-Raum eine obere Dreiecksform hat, kann man die Chebyshev-Amplituden der Ableitung über eine Rekursion berechnen, die nur $O(N)$ Operationen erfordert. Der Aufwand skaliert also wie beim Fourier-System. Wenn man daher die Ableitung im Ortsraum benötigt, verwendet man eine schnelle Transformation in den Chebyshev-Raum, bildet dort die Ableitung und transformiert zurück in den Ortsraum. Dies kostet dann nur $O(N + 1) + 2 \times O[(N + 1)\log_2(N + 1)] = O[(N + 1)\log_2(N + 1)]$ Operationen im Vergleich zu den Operationen $O[(N + 1)^2]$ bei der Matrix-Multiplikation.

Zur Anwendung der Methode stellt man die Ableitungen als Summe über

Chebyshev-Polynome dar

$$\frac{\partial u}{\partial x} = \sum_{k=0}^{N-1} \hat{u}_k^{(1)} \mathrm{T}_k(x), \qquad \text{bzw.} \qquad \frac{\partial^2 u}{\partial x^2} = \sum_{k=0}^{N-2} \hat{u}_k^{(2)} \mathrm{T}_k(x). \tag{4.156}$$

Die Koeffizienten der ersten und der zweiten Ableitung $\hat{u}_k^{(1)}$ und $\hat{u}_k^{(2)}$ kann man dann aus der allgemeinen Rekursionsformel[43]

$$c_k \hat{u}_k^{(p)} = \hat{u}_{k+2}^{(p)} + 2(k+1)\hat{u}_{k+1}^{(p-1)}, \qquad k \geq 0 \tag{4.157}$$

für die Koeffizienten der p-ten Ableitung erhalten. Zusammen mit zwei Startwerten kann man alle Koeffizienten rekursiv berechnen. Für die Koeffizienten der ersten und der zweiten Ableitung erhält man so[44]

$$c_k \hat{u}_k^{(1)} = \hat{u}_{k+2}^{(1)} + 2(k+1)\hat{u}_{k+1}, \qquad \text{für} \quad 0 \leq k \leq N-2, \tag{4.158a}$$

$$c_k \hat{u}_k^{(2)} = \hat{u}_{k+2}^{(2)} + 2(k+1)\hat{u}_{k+1}^{(1)}, \qquad \text{für} \quad 0 \leq k \leq N-3. \tag{4.158b}$$

Mit den Startwerten $\hat{u}_N^{(1)}$ und $\hat{u}_{N-1}^{(1)}$ und mit $\hat{\boldsymbol{u}}$ (für die 1. Ableitung) bzw. $\hat{u}_{N-1}^{(2)}$ und $\hat{u}_{N-2}^{(2)}$ und $\hat{\boldsymbol{u}}^{(1)}$ (für die 2. Ableitung) (vgl. die Rekursionsformel (C.2) in Anhang C)

$$\hat{u}_N^{(1)} = 0, \qquad \hat{u}_{N-1}^{(1)} = 2N\hat{u}_N, \tag{4.158c}$$

$$\hat{u}_N^{(2)} = \hat{u}_{N-1}^{(2)} = 0, \qquad \hat{u}_{N-2}^{(2)} = 2(N-1)\hat{u}_{N-1}^{(1)} = 4N(N-1)\hat{u}_N, \tag{4.158d}$$

lassen sich alle Chebyshev-Koeffizienten der ersten und zweiten Ableitung in (4.156) rekursiv aus (4.158) berechnen. Damit kostet die Berechnung z.B. der Koeffizienten $\hat{u}_k^{(2)}$ aus $\hat{u}_k$ nur $O(N+1)$ Operationen.

Um die Ableitungen im Ortsraum zu erhalten, muß man das Ergebnis noch einer schnellen Transformation unterziehen. Die Hin- und Rücktransformationen kosten $O[2(N+1)\log_2(N+1)]$ Operationen. Die rekursive Berechnung der Ableitung im Chebyshev-Raum mit $O(N)$ Operationen fällt also praktisch nicht ins Gewicht. Ein Vergleich der beiden Methoden (Matrix-Multiplikation nach (4.148) versus Transformation und Rekursion) ist in Tabelle 4.3 gezeigt.[45]

[43] Die Ableitungsmatrix (C.7) bzw. (C.6) im rein spektralen Raum $\tilde{\mathcal{D}}$ ist eine obere Dreiecksmatrix. Hieraus kann man Rekursionsformeln für die Koeffizienten $\hat{u}_k^{(p)}$ gewinnen (Peyret, 2002).

[44] Fletcher (1991*a*) scheint hier mal wieder einen Druckfehler zu haben.

[45] Cyber waren in den 70er und 80er Jahren Supercomputer der Firma *Control Data Corporation* (CDC). Die Cyber 205 kam 1979 auf den Markt (Freon-Kühlung) und erreichte theoretisch 400 64-bit MFlops.

Ganz allgemein führen Vektorprozessoren eine Operation gleichzeitig auf vielen Daten (einem Vektor oder Array) aus, die in einer Pipe geladen sind. Bei Matrizenoperationen sind Vektorprozessoren reinen Allzweck-Prozessoren (z. B. x86), die jedes Datum nacheinander bearbeiten, weit überlegen. Die legendären Cray-Supercomputer nutzten Vektorprozessoren (auch der Earth Simulator mit Vektorprozessoren von NEC). Erhebliche Performancegewinne ergeben sich auch

Tabelle 4.3.: Vergleich der erforderlichen Rechenzeiten in ms (auf zwei verschiedenen Rechnern) zur Berechnung der Ableitung an den Kollokationspunkten mittels der Matrix-Multiplikation und mittels der Transformationsmethode (nach Canuto et al., 1988).

	Cyber 855 (skalar)		Cyber 205 (Vektor)	
N	Matrix	Transformation	Matrix	Transformation
8	0.11	0.18	0.0010	0.0018
16	0.38	0.39	0.0032	0.0044
32	1.54	0.91	0.0119	0.0091
64	5.74	1.92	0.0456	0.0215
128	24.02	4.20	0.1782	0.0442
256	93.49	9.40	0.7061	0.1015

Da die Ableitungskoeffizienten mit größtem Index (z.B. $\hat{u}^{(1)}_{N-1}$, $\hat{u}^{(1)}_{N-2}$, etc.) normalerweise sehr klein sind, die anderen Koeffizienten aber rekursiv aus ihnen berechnet werden, können sich kleine Fehler in den Koeffizienten mit hohem Index u.U. stark auf die Koeffizienten mit niedrigerem Index auswirken. Um dies zu verhindern, gibt es verschiedene Korrekturmöglichkeiten, auf die hier aber nicht eingegangen werden kann (siehe z.B. Peyret, 2002).

4.6.4. Nichtlineare Terme

Wenn man im Ortsraum arbeitet, ist die Berechnung der nichtlinearen Terme kein Problem. Das einfache Produkte zweier diskreter Größen kostet nur $O(N+1)$ Operationen.

Rein spektrale Behandlung

Wenn man alle Variablen spektral betrachtet, müssen zur Berechnung von Ortsraum-Produkten *Faltungs-Summen* (*convolution sums*) berechnet werden. Als Beispiel betrachten wir die diskrete Fourier-Darstellung

$$u(x_k) = \sum_{n=-N/2}^{N/2} \hat{u}_n \mathrm{e}^{\mathrm{i}nx_k}, \qquad \text{mit} \quad k = 1, ..., N, \tag{4.159}$$

durch vektorisierende Algorithmen. Vergleiche beispielsweise den Unterschied bei vektorisierten numerischen Operationen in MATLAB gegenüber herkömmlicher Numerik. Vektorrechner haben in den letzten Jahren große Konkurrenz durch massiv parallel aufgebaute Rechencluster bekommen, die aus vielen Tausend Standardprozessoren aufgebaut sind. Deshalb sind Vektorrechner heute ausgestorben.

Wegen der Vorteile, die sich durch die gleichzeitige Ausführung einer Rechenoperation auf mehreren Daten ergeben (Single Instruction, Multiple Data, SIMD) haben auch Standardprozessoren seit den 1990er Jahren Erweiterungen der jeweiligen Architektur erfahren, um diese Art von Berechnungen zu beschleunigen (z.B. der x86-Prozessor). Dies trifft besonders auf graphische Anwendungen zu (viele Matrizenoperationen auf 3D-Koordinaten, Antialiasing der Bildschirmausgabe, große Datenmengen) weshalb heutige Grafikprozessoren große Ähnlichkei-

mit den Stützstellen $x_k = 2\pi k/N$ für 2π-periodische Funktionen. Dann lautet das Produkt zweier Funktionen u und v im Ortsraum

$$u(x_k)v(x_k) = \left(\sum_{n=-N/2}^{N/2} \hat{u}_n \mathrm{e}^{\mathrm{i}nx_k}\right)\left(\sum_{m=-N/2}^{N/2} \hat{v}_m \mathrm{e}^{\mathrm{i}mx_k}\right) = \sum_{\substack{n=-N/2\\ m=-N/2}}^{N/2} \hat{u}_n\hat{v}_m \mathrm{e}^{\mathrm{i}(n+m)x_k}$$

$$= \sum_{l=-N}^{N}\left(\sum_{\substack{|m|,|n|\le N/2\\ \text{mit } n+m=l}} \hat{u}_n\hat{v}_m\right)\mathrm{e}^{\mathrm{i}lx_k} \tag{4.160}$$

und man erhält für die Amplituden des Produkts

$$\hat{w}_l = \sum_{\substack{|m|,|n|\le N/2\\ n+m=l}} \hat{u}_n\hat{v}_m. \tag{4.161}$$

Allein zur Berechnung des Produkts ($\hat{w}_l$ im spektralen Raum) sind also $O\left(N^2\right)$ Operationen erforderlich. Auch bei Chebyshev-Kollokation ist ein entsprechender Aufwand nötig.

Als weiteres Problem erkennt man in (4.160) das *Aliasing* (siehe unten): Durch das Produkt werden spektrale Komponenten außerhalb des betrachteten Definitionsbereichs $-N/2 \le l \le N/2$ generiert.

Pseudospektrale Behandlung

Bei der pseudospektralen Methode geht man anders vor. Hier werden die beiden Faktoren des nichtlinearen Terms zuerst in den Ortsraum transformiert, dann dort das Produkt berechnet und schließlich wieder in den Fourier- bzw. Chebyshev-Raum zurücktransformiert. Die pseudospektrale Prozedur für den nichtlinearen Term $u\partial u/\partial x$, z.B. in der Burgers-Gleichung, lautet dann

1. Gegeben sei $\hat{u}_k$. Bestimme die erste Ableitung $\hat{u}_k^{(1)}$ im spektralen Raum (z.B. über Rekursion (4.158a)).
2. Berechne aus $\hat{u}_k$ und $\hat{u}_k^{(1)}$ die Größen u_j und $\left(\partial u/\partial x\right)_j$ im Ortsraum.
3. Berechne das Produkt $u_j\left(\partial u/\partial x\right)_j$ punktweise im Ortsraum.
4. Transformiere das Produkt zurück in den Chebyshev-Raum und berechne den nächsten Zeitschritt.

Bei der pseudospektralen Auswertung benötigt man nur 3 FFTs: Transformation von $\hat{u}_k$ und $\hat{u}_k^{(1)}$ in den Ortsraum und Rücktransformation des Produkts $u\partial u/\partial x$ in den spektralen Raum. Das kostet nur $O[3(N+1)\log_2(N+1)]$ Operationen sowie eine Produktbildung im Orts-Raum mit $O(N+1)$-Operationen, die gegenüber denn FFTs zu vernachlässigen ist.

Konsequenz Der Unterschied im Rechenaufwand zwischen der Produktbildung im spektralen Raum und derjenigen im Ortsraum (inklusive der FFTs) macht sich bei großen Werten von N entscheidend bemerkbar. Man muß diesen Aufwand jedoch im Vergleich zu den nur $O(N+1)$ Operationen sehen, die man zur Berechnung mittels finiter Differenzen benötigt.

Die Technik der Transformation zwischen Orts- und spektralem Raum ist nicht auf die Kollokations-Methode beschränkt. Auch beim Galerkin-Verfahren wird die pseudospektrale Methode verwendet. Dort beschränken sich die Ortsraumberechnungen aber nur auf die Berechnung der nichtlinearen Terme.

Bei der Lösung der Navier-Stokes-Gleichungen durch zeitliche Simulation wird jedoch oft der nichtlineare Term nicht implizit behandelt, sondern explizit. Für einen Integrationsschritt verwendet man zumindest einen Faktor des nichtlinearen Produkts aus dem alten Integrationsschritt. Damit wird die Gleichung linearisiert. Falls der nichtlineare Term vollständig explizit behandelt wird, taucht er als eine bekannte Größe auf der rechten Seite der Gleichungen auf (siehe auch Kap. 4.6.6).

4.6.5. Aliasing

Mit der pseudospektralen Behandlung nichtlinearer Terme wie auch bei der Berechnung von Ausdrücken, deren Koeffizienten von den unbekannten Feldgrößen abhängen, sind spezielle Probleme verbunden. Denn bei der Produktbildung werden spektrale Komponenten generiert, deren Wellenzahlen Summen und Differenzen der Wellenzahlen der Ausgangsfaktoren sind.[46] Bei einem Galerkin-Verfahren, bei dem man ja *kontinuierliche* Ansatz- und Wichtungsfunktionen hat, ist dies kein Problem. Denn bei Verwendung einer orthogonalen Basis wird durch die Projektion des nichtlinearen Terms auf eine Wichtungsfunktion genau eine spektrale Komponente herausprojiziert. Die Komponenten, deren Wellenzahlen oberhalb des Cut-off liegen, d.h. deren Index $k > N$ ist, haben keinen Einfluß auf die Projektion der tatsächlich berücksichtigten Moden mit Wellenzahlen $k \leq N$. Bei einem pseudospektralen Verfahren mit einem *diskreten* Gitter im Ortsraum können aber nicht alle Amplituden der relevanten Moden mit $k \leq N$ korrekt dargestellt werden. Insbesondere werden Amplituden von Moden mit hohen Wellenzahlen, aber immer noch $k \leq N$, verfälscht. Dieses Phänomen wird *Aliasing* genannt. Es wird in Abb. 4.18 motiviert. Im Anhang D werden die bei Fourier-Kollokation durch Aliasing erzeugten zusätzlichen Terme explizit berechnet.

Eine Möglichkeit, das Aliasing zu eliminieren, ist die sogenannte *3/2-Regel* (Peyret, 2002). Dabei wird zur Berechnung der nichtlinearen Terme und von Termen mit nichtkonstanten Koeffizienten die Anzahl der betrachteten Moden nur für

ten zu reinen Vektorprozessoren aufweisen (siehe Rechnen auf GPUs).

[46] Dies wird bei Harmonischen besonders klar:

$$2\cos(k_1 x)\cos(k_2 x) = \cos[(k_1 + k_2)x] + \cos[(k_1 - k_2)x].$$

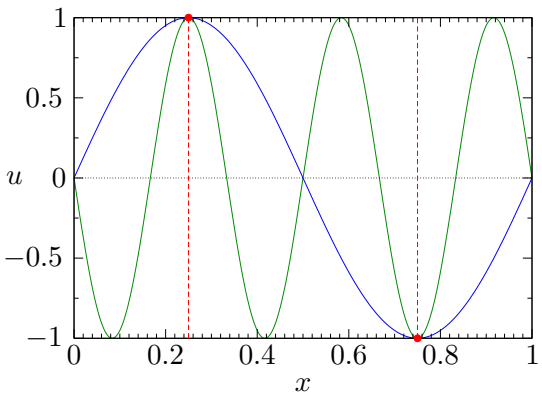


Abbildung 4.18.: Eine Mode mit hoher Wellenzahl (grün) erscheint mit einer von Null verschiedenen Amplitude bei einer niedrigeren Wellenzahl (blau), wenn das Gitter (rot gestrichelt) die Mode mit hoher Wellenzahl nicht auflösen kann. Durch diesen Effekt wird die eigentliche Amplitude der blauen Mode verfälscht.

diese Terme von N auf $N' = 3N/2$ erweitert. Da diese Erweiterung der Modenzahl nur für den nichtlinearen Term notwendig ist, bleibt der zusätzliche Rechenaufwand moderat.

Das Aliasing ist besonders störend bei der Simulation von Systemen mit geringer Dissipation, da es numerische Instabilitäten des Zeitintegrationsschemas verursachen kann. Bei ingenieurmäßigen Anwendungen für viskose Fluide ist dieser Aliasing-Effekt oft nicht so wichtig.

4.6.6. Typische Strategien am Beispiel der Burgers-Gleichung

Eine einfache Modellgleichung mit einer Nichtlinearität wie in der Navier-Stokes-Gleichung ist die eindimensionale Konvektions-Diffusionsgleichung

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} - \nu\frac{\partial^2 u}{\partial x^2} = 0. \tag{4.162}$$

Diese Gleichung heißt *Burgers-Gleichung*.[47] Als Anfangsbedingung kann man $u(x, t = 0) = u_0(x)$ annehmen. Je nach Situation können die Randbedingungen periodisch in x, homogen oder inhomogen sein.

Explizite Behandlung des nichtlinearen Terms

Nichtlineare Gleichungen können i.a. nicht analytisch integriert werden. Für die numerische Behandlung kann man als Zeitintegrationsschema zum Beispiel ein einfaches halbimplizites Verfahren mit Vorwärtsdifferenzen verwenden, bei dem der diffusive Term implizit und der konvektive Term explizit ausgewertet wird. Dann

[47] Die Burgers-Gleichung sieht so ähnlich aus wie die eindimensionale Navier-Stokes-Gleichung. Ein wichtiger Unterschied ist jedoch die Abwesenheit des Druckgradienten-Terms.

erhält man

$$\frac{u^{n+1}-u^n}{\Delta t}+u^n\frac{\partial u^n}{\partial x}-\nu\frac{\partial^2 u^{n+1}}{\partial x^2}=0, \tag{4.163}$$

oder nach Sortieren der Zeitniveaus

$$\frac{u^{n+1}}{\Delta t}-\nu\frac{\partial^2 u^{n+1}}{\partial x^2}=\frac{u^n}{\Delta t}-u^n\frac{\partial u^n}{\partial x}. \tag{4.164}$$

Diese Gleichung ist vom Typus einer eindimensionalen *Helmholtz-Gleichung*,[48][49] die noch um eine Inhomogenität und einen Term der ersten Ableitung erweitert wurde. Wenn man den Zeitindex n unterdrückt besitzt (4.164) die Struktur

$$-\nu u''+au'+bu=f. \tag{4.165}$$

Bezogen auf (4.164) ist dann $b=\Delta t^{-1}$, $a=0$ und alle expliziten Terme werden der Inhomogenität zugeschlagen: $f=u^n/\Delta t-u^n\partial u^n/\partial x$.

Wenn wir den Definitionsbereich $-1\le x\le 1$ betrachten, benötigen wir für diese Gleichung zweiter Ordnung in x zwei Randbedingungen (neben der erforderlichen Anfangsbedingung). Hier betrachten wir Robin-Randbedingungen

$$\alpha_- u(-1)+\beta_- u'(-1)=g_-, \tag{4.166a}$$
$$\alpha_+ u(1)+\beta_+ u'(1)=g_+. \tag{4.166b}$$

Wenn wir nun Gauß-Lobatto-Punkte (4.146) betrachten, erhalten wir die Chebyshev-Kollokations-Näherung gemäß (4.10) dadurch, daß wir das Residuum an den Kollokationspunkten zu Null setzen. Für die inneren Gauß-Lobatto-Stützpunkte liefert dies

$$-\nu u_i''+au_i'+bu_i=f_i,\quad i=1,\dots,N-1. \tag{4.167}$$

Für die beiden Randpunkte ergeben sich die zwei weiteren Gleichungen

$$\alpha_- u_N+\beta_- u_N'=g_-, \tag{4.168a}$$
$$\alpha_+ u_0+\beta_+ u_0'=g_+. \tag{4.168b}$$

Wenn wir die Ableitungen an den Kollokationspunkten x_i mit Hilfe von (4.153) und (4.155) durch die Funktionswerte an den Punkten ausdrücken, erhalten wir

$$\sum_{j=0}^{N}\left(-\nu\mathcal{D}_{ij}^{(2)}+a\mathcal{D}_{ij}^{(1)}\right)u_j+bu_i=f_i,\quad i=1,\dots,N-1 \tag{4.169a}$$

[48] Die allgemeine Form der Helmholtz-Gleichung ist

$$\nabla^2 u+\gamma u=0.$$

Sie beschreibt die Wellenausbreitung oder schwingende Membranen.

[49] Auch spektrale Verfahren für die Navier-Stokes-Gleichung werden in vielen Fällen auf die Lösung

$$\alpha_- u_N + \beta_- \sum_{j=0}^{N} \mathcal{D}_{Nj}^{(1)} u_j = g_-, \tag{4.169b}$$

$$\alpha_+ u_0 + \beta_+ \sum_{j=0}^{N} \mathcal{D}_{0j}^{(1)} u_j = g_+, \tag{4.169c}$$

wobei $\mathcal{D}^{(1)} = \mathcal{D}$ die Matrix der ersten Ableitung ist und $\mathcal{D}^{(2)} = \mathcal{D}^2$ diejenige der zweiten Ableitung. Dieses System können wir in Matrixform schreiben als

$$\mathsf{A} \cdot \boldsymbol{u}^{n+1} = \boldsymbol{f}^n, \tag{4.170}$$

wobei $\boldsymbol{u} = (u_0, \ldots, u_N)^{\mathrm{T}}$, $\boldsymbol{f} = (g_+, f_1, \ldots, f_{N-1}, g_-)^{\mathrm{T}}$ und n der Zeitindex ist. Die $(N+1) \times (N+1)$-Matrix A kann man aus der Matrix

$$\mathcal{Q} = -\nu \mathcal{D}^{(2)} + a\mathcal{D}^{(1)} + b\,\mathsf{I} \tag{4.171}$$

gewinnen, indem man die erste und die letzte Zeile von $\mathcal{Q}$ durch die Einträge ersetzt, die von den Randbedingungen stammen (τ-Methode). Damit erhalten wir A für $j = 0, \ldots, N$

$$A_{ij} = \mathcal{Q}_{ij} = -\nu \mathcal{D}_{ij}^{(2)} + a\mathcal{D}_{ij}^{(1)} + b\,\delta_{ij}, \qquad i = 1, \ldots, N-1, \tag{4.172a}$$

$$A_{0j} = \alpha_+ + \beta_+ \sum_{j=0}^{N} \mathcal{D}_{0j}^{(1)}, \tag{4.172b}$$

$$A_{Nj} = \alpha_- + \beta_- \sum_{j=0}^{N} \mathcal{D}_{Nj}^{(1)}. \tag{4.172c}$$

Das System (4.170) kann man mit unterschiedlichen Methoden lösen. Es ist jedoch zu beachten, daß dieses diskrete Helmholtz-Problem für jeden einzelnen Zeitschritt erneut gelöst werden muß, da sich die rechte Seite bei jedem Zeitschritt ändert. Deshalb ist ein effizienter Löser sehr wichtig. Alle Operationen, die man nur einmal durchführen muß, sollten vorab ausgeführt werden. Auch sollte eine effiziente Methoden die Eigenschaften der Matrix ausnutzen. Leider ist die vorliegende Matrix voll besetzt, sie ist weder symmetrisch noch schief-symmetrisch und sie ist außerdem schlecht konditioniert.

Da die Matrix A jedoch konstant ist und sich im Laufe der Simulation nicht ändert, ist es sinnvoll, sie vorab zu invertieren und die Inverse im Speicher zu halten. Dann kann man jeden Zeitschritt berechnen, indem man lediglich eine Matrix-Vektor-Multiplikation durchführt

$$\boldsymbol{u}^{n+1} = \mathsf{A}^{-1} \cdot \boldsymbol{f}^n. \tag{4.173}$$

von Helmholtz-Gleichungen zurückgeführt.

Diese Operation läßt sich sehr schnell mit Vektorrechnern ausführen. Damit diese Methode funktioniert, benötigt man eine sehr genaue Inversion von A (z.B. mit den Subroutinen LINRG von IMSL oder F04AEF von NAG).[50]

Dies allein reicht oft auch nicht aus, weil A schlecht konditioniert ist. Mittels einer Vorkonditionierung mit einer einfachen Matrix A_0, die man leicht invertieren kann und die eine gute Näherung von A sein sollte, erhält man

$$\underbrace{\mathsf{A}_0^{-1} \cdot \mathsf{A}}_{\mathsf{B}} \cdot \boldsymbol{u}^{n+1} = \mathsf{A}_0^{-1} \cdot \boldsymbol{f}^n. \tag{4.174}$$

Für A_0 kann man zum Beispiel die Matrix der Finite-Differenzen-Approximation des Ableitungsoperators A (auf dem Kollokationsgitter der Gauß-Lobatto-Punkte) nehmen.[51] Die Matrix $\mathsf{B} = \mathsf{A}_0^{-1} \cdot \mathsf{A}$ ist dann gut konditioniert und kann leicht invertiert werden. Man erhält dann

$$\boldsymbol{u}^{n+1} = \mathsf{B}^{-1} \cdot \mathsf{A}_0^{-1} \cdot \boldsymbol{f}^n. \tag{4.175}$$

Die Effizienz des Schemas hängt auch davon ab, wie schnell die rechte Seite $\boldsymbol{f}^n$ berechnet werden kann. Die rechte Seite involviert die Berechnung von $\partial_x u_i^n$ des nichtlinearen Terms. Hier kommen die beiden Möglichkeiten in Betracht, die wir in Kap. 4.6.3 diskutiert hatten, entweder Matrix-Multiplikation von u_i^n mit $\mathcal{D}$ ($N \lesssim 100$) oder Transformation in den Chebyshev-Raum, rekursive Berechnung der Ableitung und Rücktransformation in den Ortsraum ($N \gtrsim 100$). Es sei jedoch bemerkt, daß die Matrix-Multiplikation genauer sein kann und — auf Vektorrechnern — auch schneller ist.

Halbimplizite Behandlung des nichtlinearen Terms

Da implizite Verfahren in der Regel stabiler sind als explizite Verfahren und damit eine größere Zeitschrittweite erlauben, ist man bestrebt, zumindest einen Teil des nichtlinearen Terms implizit zu behandeln. Dies kann man erreichen, wenn sich u aufspalten läßt in einen zeitunabhängigen und einen zeitabhängigen Anteil

$$u(x,t) = \bar{u}(x) + \tilde{u}(x,t). \tag{4.176}$$

[50] NAG Numerical Libraries ist eine umfassende Software-Unterprogrammbibliothek für numerische und statistische Problemstellungen. Sie wird von der Numerical Algorithms Group Ltd (NAG) aus Oxford vertrieben. Die Firma wurde 1971 als Nottingham Algorithms Group gegründet. Heute ist NAG für Fortran und C verfügbar.

Auch IMSL (International Mathematical and Statistical Library) erfüllt denselben Zweck und wird als Lizenz von der Firma Visual Numerics, Inc. aus Houston, Texas vergeben. Die ursprünglich in Fortran implementierte Bibliothek enthält etwa 1000 Unterprogramme. Sie ist eine der ältesten und sicher eine der umfassendsten Programmbibliotheken. Da die Ursprünge der Programme schon aus den 70er Jahren stammen, gelten sie als äußerst robust und gut getestet. Die IMSL Numerikbibliotheken sind auch für C, Java, C#.NET erhältlich. Für die Programmierung in Python wurde durch PyIMSL Studio eine Schnittstelle zur Anwendung der IMSL C-Bibliothek geschaffen.

[51] Die Matrix der Finite-Differenzen-Approximation des Ableitungsoperators A ist tridiagonal,

Dann kann man $u\partial_x u$ zerlegen in $[\bar{u}(x)\partial_x u]^{(n+1)} + [\tilde{u}(x,t)\partial_x u]^{(n)}$, wobei der erste Term implizit und der zweite Term explizit behandelt wird. Die Kollokationsnäherung der Burgers-Gleichung lautet dann an den inneren Punkten

$$\frac{u_i^{n+1}}{\Delta t} - \nu\frac{\partial^2 u_i^{n+1}}{\partial x^2} + \bar{u}_i\frac{\partial u_i^{n+1}}{\partial x} = \frac{u_i^n}{\Delta t} - \tilde{u}_i^n\frac{\partial u_i^n}{\partial x}, \quad i = 1,\ldots,N-1. \tag{4.177}$$

Dieses Verfahren ist (bzgl. kleiner Störungen) uneingeschränkt stabil, wenn $|\bar{u}| \geq |\tilde{u}|$. Da von dem impliziten konvektiven Term nur der zeitunabhängige Term $\bar{u}_i$ in die Matrix A eingeht, kann man wie bei der expliziten Methode verfahren und A vorab invertieren.

Implizite Behandlung des nichtlinearen Terms

Manchmal läßt sich der nichtlineare Term nicht explizit behandeln oder wie oben zerlegen. Dann muß man ihn implizit behandeln und entweder durch $u^n(x)\partial_x u^{n+1}$ oder $u^{n+1}(x)\partial_x u^{n+1}$ darstellen. Wir betrachten hier den ersten Fall

$$\frac{u_i^{n+1}}{\Delta t} - \nu\frac{\partial^2 u_i^{n+1}}{\partial x^2} + u_i^n\frac{\partial u_i^{n+1}}{\partial x} = \frac{u_i^n}{\Delta t}, \quad i = 1,\ldots,N-1. \tag{4.178}$$

Dann ist das resultierende System von Gleichungen zwar linear, aber die Matrix $\mathsf{A}(\boldsymbol{u}^n)$ ist nun über u^n zeitabhängig

$$\mathsf{A}(\boldsymbol{u}^n)\cdot\boldsymbol{u}^{n+1} = \boldsymbol{f}^n. \tag{4.179}$$

Um die Inversion der Matrix A in jedem Zeitschritt zu vermeiden, kann man ein iterative Verfahren verwenden. Eine Möglichkeit besteht in dem Schema

$$\mathsf{A}_0\cdot\bar{\boldsymbol{u}}^{m+1} = \boldsymbol{f}^n - \mathsf{A}(\boldsymbol{u}^n)\cdot\boldsymbol{u}^{n+1,m}, \tag{4.180a}$$
$$\boldsymbol{u}^{n+1,m+1} = \boldsymbol{u}^{n+1,m} + \alpha\bar{\boldsymbol{u}}^{m+1}, \tag{4.180b}$$

wobei m der Iterationsindex ist. Hierbei ist A_0 der Vorkonditionierer und α ein Relaxationsparameter. Die richtige Wahl der Vorkonditionierung ist sehr wichtig für einen effizienten Algorithmus. Offenbar ist die rechte Seite in (4.180a) das Residuum im m-ten Schritt. Aus diesem Residuum wird mittels der Approximation A_0 von A eine Korrektur $\bar{\boldsymbol{u}}^{m+1}$ berechnet und in (4.180b) zum alten Wert addiert.

aber die Inverse A_0^{-1} ist vollbesetzt. Für derartige einfache Matrizen kann man die Inverse aber explizit berechnen.

5. Lösung stationärer Probleme

In der Strömungsmechanik sind oft stationäre Probleme zu lösen, die meist elliptisch sind. Nach der Diskretisierung hat man dann große lineare oder nichtlineare algebraische Gleichungssysteme zu lösen. Die Nichtlinearitäten stammen meist von den konvektiven Termen $\boldsymbol{u} \cdot \nabla \boldsymbol{u}$ oder $\boldsymbol{u} \cdot \nabla T$. Sie können auch durch Koeffizienten bedingt sein, die von den Feldgrößen abhängen, wie zum Beispiel bei einer temperaturabhängigen Viskosität $\nu(T)$. Nach der Diskretisierung durch finite Differenzen, finite Volumen, finite Elemente oder mittels einer spektralen Methode erhält man das *allgemeine nichtlineare Problem*

$$\mathsf{A}(\boldsymbol{x}) \cdot \boldsymbol{x} = \boldsymbol{b}. \tag{5.1}$$

Hierbei ist $\boldsymbol{x}$ der Vektor der unbekannten Feldgrößen an den Gitterpunkten oder der Amplituden bei spektralen Galerkin-Verfahren. A ist eine Matrix, die durch die Diskretisierung zustande kommt und deren Elemente von den Feldgrößen abhängen können. Die Inhomogenität $\boldsymbol{b}$ ist durch die Randbedingungen und durch äußere Kräfte bestimmt.

Wenn die Gleichungen nichtlinear sind, müssen sie durch eine iterative Methode gelöst werden. Dazu werden Gleichungen in jedem Schritt der Iteration um die Zwischenlösung $\boldsymbol{x}^{(n)}$ linearisiert. Das heißt, im Schritt $n+1$ muß das *lineare Problem*

$$\mathsf{A}\left(\boldsymbol{x}^{(n)}\right) \cdot \boldsymbol{x}^{(n+1)} = \boldsymbol{b} \tag{5.2}$$

gelöst werden. Die Iteration wird so lange fortgesetzt, bis die Folge $\boldsymbol{x}^{(n)}$ hinreichend konvergiert ist. Für numerische Probleme in der Strömungsmechanik sind drei Klassen von Matrizen wichtig:

1. *Dünn besetzte* Matrizen: Die meisten Matrixelemente besitzen den Wert 0.

2. *Dünn besetzte* und *bandstrukturierte* Matrizen: Die überwiegende Zahl der Elemente hat den Wert 0. Die von Null verschiedenen Elemente befinden sich in der Nähe der Hauptdiagonalen.

3. *Voll besetzte* Matrizen: Die überwiegende Anzahl der Elemente ist von Null verschieden.

Je nach Typ der Matrix kann man mehr oder weniger effiziente Verfahren verwenden, um das System (5.1) zu lösen. Verfahren für vollbesetzte Matrizen können natürlich auch für dünn besetzte Matrizen verwendet werden. Für dünn besetzte

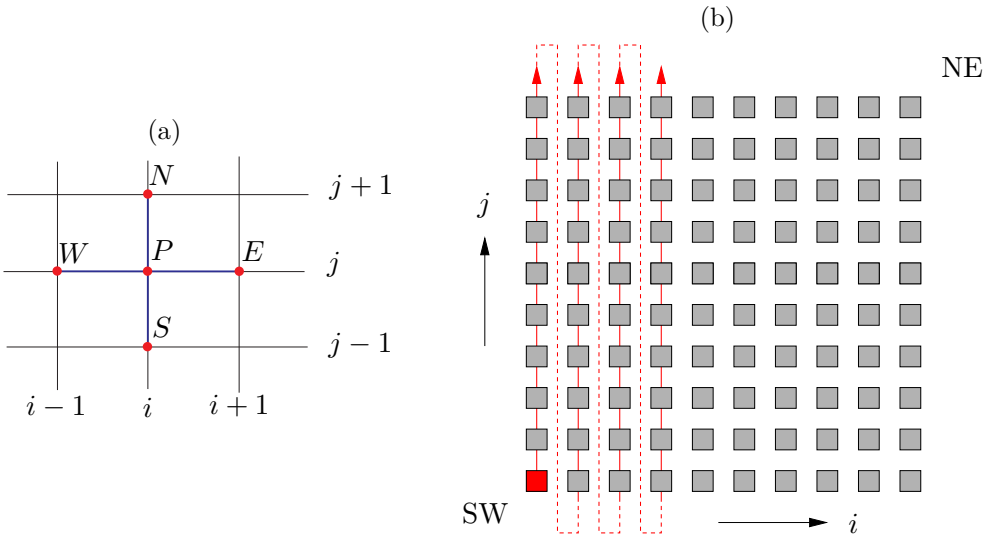


Abbildung 5.1.: (a) 5-Punkt-Stern und geographische Bezeichnungen. (b) Anordnung der Unbekannten innerhalb des Vektors $\boldsymbol{x}$. Die Gitterpunkte werden von unten links beginnend schnell von Süd (S) nach Nord (N) durchlaufen und dann (langsamer Index) von West (W) nach Ost (E).

Fälle gibt es aber effizientere Methoden, welche die spezielle Struktur der jeweiligen Matrix ausnutzen.

Für lineare Systeme (5.2) gibt es iterative und direkte Verfahren der Lösung. Oft sind iterative Verfahren effektiver als direkte, die bei einer exakten Arithmetik die exakte Lösung liefern würden, da es keinen Sinn macht, ein Gleichungssystem genauer zu lösen als nötig. Denn schon das lineare System ist ja aufgrund der Diskretisierung fehlerbehaftet (Diskretisierungsfehler).

Für dünn besetzte Matrizen, in denen nur wenige Diagonalen von Null verschieden sind, macht es Sinn, nicht alle Matrixelemente zu speichern, sondern nur diejenigen, die von Null verschieden sind. Dann braucht man nur alle Diagonalen zu speichern und nicht alle Matrixelemente.[1]

Eine typische Struktur der Matrix, wie sie bei einer Diskretisierung mittels zentralen finiten Differenzen auf einem zweidimensionalen Gebiet nach Abb. 5.1a und bei einer Anordnung der Variablen nach Abb. 5.1b entsteht, ist in Abb. 5.2 gezeigt. Neben den drei Hauptdiagonalen sind nur noch 2 weitere Nebendiagonalen besetzt, die sich in einem gewissen Abstand von der Hauptdiagonalen befindet. Die Matrix ist dünn besetzt und hat eine *Bandstruktur*. Die genaue Struktur der Matrix hängt davon ab, wie man die Gleichungen für jeden Knoten (Gitterpunkt, Unbekannte) ordnet. Im Gegensatz dazu sind die Matrizen bei spektralen Verfahren typischerweise voll besetzt.

[1] In zwei Dimensionen mit $N_i \times N_j$ Gitterpunkten sind dies bei 5 Diagonalen nur $5N_iN_j$ Werte im Gegensatz zu $(N_iN_j)^2$ Werten bei einer Speicherung der gesamten Matrix. Bei dreidimensionalen Problemen ist die Speicherplatzersparnis noch eklatanter.

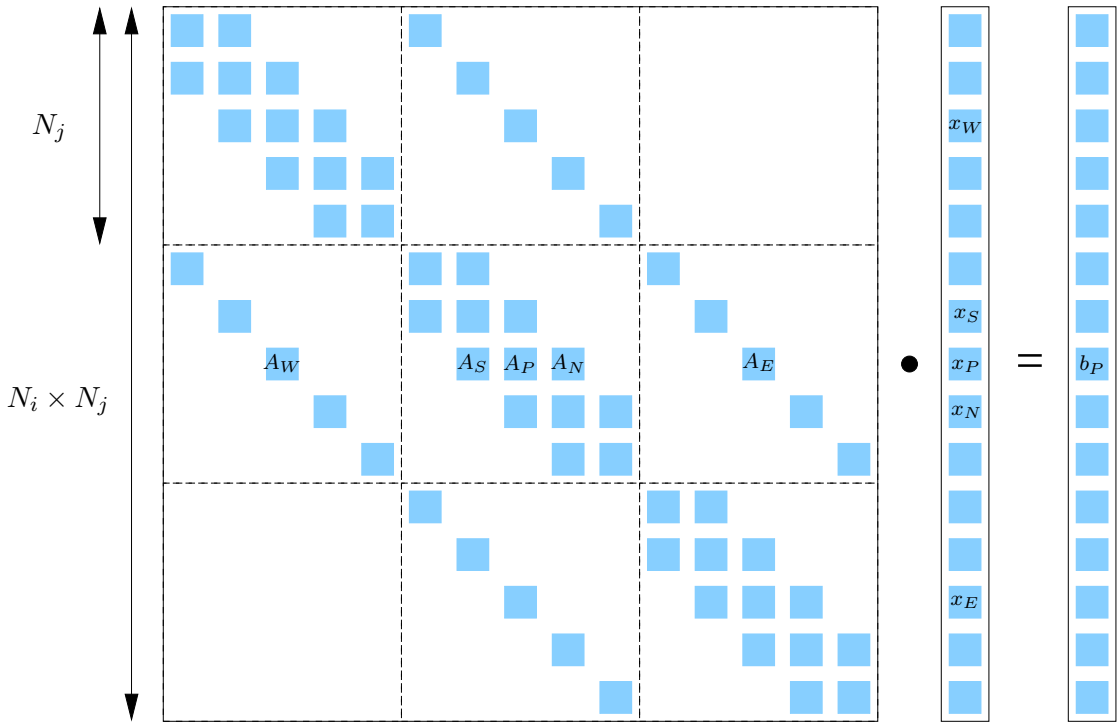


Abbildung 5.2.: Anordnung der Koeffizienten in der Matrix A für zentrale finite Differenzen wie in Abb. 5.1a bei Verwendung der geographischen Notation und Anordnung der Variablen wie in Abb. 5.1b.

Für die Numerierung der Gitterpunkte (Gleichungen) hat sich die *geographische Notation* bewährt. Dabei werden die Nachbarpunkte des zentralen Punkts P nach den Himmelsrichtungen bezeichnet (Abb. 5.1a). Dann bestimmen die Indizes N, W, S und E die relative Lage der betreffenden Punkte bezogen auf den Punkt P.

Wenn man nun die Gleichungen der Reihe nach aufstellt und die Konvention verwendet, daß dabei die Gitterpunkte wie in Abb. 5.1b von Süd (S) nach Nord (N) durchlaufen werden (schneller Index j) und dann von West (W) nach Ost (E) (langsamer Index i), ergibt sich bei einem 5-Punkt-Stern die Struktur der Matrix wie in Abb. 5.2. Im dreidimensionalen Fall wird anschließend auch noch von unten (B, bottom) nach oben (T, top) durchnumeriert. Die Gleichung für den Punkt P mit den konventionellen Indizes (i, j, k) erscheint dann als l-te Gleichung, wobei

$$l = (k-1)N_iN_j + (i-1)N_j + j, \tag{5.3}$$

und alle Indizes bei 1 beginnen. Der schnellste Index ist j, der langsamste ist k. Der zusammengesetzte globale Index l numeriert also die Gleichungen und entspricht dem Zeilenindex der Matrix. Er ist aber auch der Index der Unbekannten x_P im Vektor $\boldsymbol{x}$ aller Unbekannten. Die Position der Unbekannten an den benachbarten Gitterpunkten von P innerhalb des Spaltenvektors ist in Tabelle 5.1 aufgelistet.

Die Gleichung für den Punkt P entsprechend dem globalen Index l kann man nun in zwei Dimensionen in der konventionellen Form aufschreiben (l-te Gleichung,

Tabelle 5.1.: Zusammenhang zwischen den Indizes (i, j, k) der konventionellen Gitterpunkte, der geographischer Notation, und dem Index im Spaltenvektor aller Unbekannten.

Gitterpunkt	geogr. Notation	Index im Spaltenvektor
i, j, k	P	$l = (k-1)N_iN_j + (i-1)N_j + j$
$i, j+1, k$	N	$l+1$
$i, j-1, k$	S	$l-1$
$i+1, j, k$	E	$l+N_j$
$i-1, j, k$	W	$l-N_j$
$i, j, k+1$	T	$l+N_iN_j$
$i, j, k-1$	B	$l-N_iN_j$

siehe auch Abb. 5.2)

$$A_{l,l-N_j}x_{l-N_j} + A_{l,l-1}x_{l-1} + A_{l,l}x_l + A_{l,l+1}x_{l+1} + A_{l,l+N_j}x_{l+N_j} = b_l, \tag{5.4}$$

oder in der geographischen Notation

$$A_W(l)x_W + A_S(l)x_S + A_P(l)x_P + A_N(l)x_N + A_E(l)x_E = b_P(l), \tag{5.5}$$

wobei neben l die Indizes W, S, P, N und E die Numerierung übernehmen indem sie die *relative* Lage zum Punkt P angeben.

Jede Diagonale der Matrix wird in einem Vektor der Länge $N_iN_jN_k$ gespeichert. Der Index $l \in [1, N_iN_jN_k]$ numeriert die Zeilen der Matrix. Im zweidimensionalen Fall ist $l \in [1, N_iN_j]$. Am südwestlichen Anfangspunkt der Numerierung ($l = 1$) gibt es nur einen nördlichen und einen östlichen Nachbarn unter den Unbekannten. Die Werte von x_W und x_S liegen außerhalb des Rechengebiets (außerhalb der in Abb. 5.1b gezeigten inneren Punkte) und müssen z.B. durch Dirichlet-Randbedingungen bereitgestellt werden. Daher sind in der ersten Zeile der Matrix keine Einträge für $A_W(l = 1)$ und $A_S(l = 1)$ vorhanden. Vielmehr werden die zugehörigen Terme A_Wx_W und A_Sx_S (mit den bekannten Werten x_W und x_S) der rechten Seite $\boldsymbol{b}$ zugeschlagen. In analoger Weise wird mit allen vorgegebenen Randwerten verfahren. Deshalb treten die ersten N_j Einträge der Nebendiagonale A_W in der Matrix nicht auf. Weiter tragen entlang der östlichen Seite die Punkte E nicht bei, entlang der Nordkante die Punkte N, und entlang der Südkante die Punkte S. All diese Randwerte finden sich in $\boldsymbol{b}$.

5.1. Direkte Verfahren für lineare Systeme

5.1.1. Gauß-Verfahren

Das *Gauß-Verfahren* ist ein universelles und direktes Verfahren zur Lösung *voll besetzter linearer Gleichungssysteme*

$$\mathsf{A} \cdot \boldsymbol{x} = \boldsymbol{b}. \tag{5.6}$$

Das Gauß-Verfahren beruht auf der Idee, das Problems sukzessive auf kleinere Systeme zurückzuführen. Dazu betrachten wir die $N \times N$ Matrix

$$\mathsf{A} = \begin{pmatrix} A_{11} & A_{12} & A_{13} & . & . & . & A_{1N} \\ A_{21} & A_{22} & A_{23} & . & . & . & A_{2N} \\ . & . & . & & & & . \\ . & . & . & . & & & . \\ . & . & . & & . & & . \\ . & . & . & & & . & . \\ A_{N1} & A_{N2} & A_{N3} & . & . & . & A_{NN} \end{pmatrix} \begin{array}{l} \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{l} \\ \times(-A_{21}/A_{11}) \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{l} \\ + \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{l} \\ \times(-A_{31}/A_{11}) \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{l} \\ \ldots \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{l} \\ \\ \\ \\ \\ \\ + \\ \end{array} \quad . \tag{5.7}$$

Die der Gleichung (5.6) zugrundeliegenden Gleichungen können beliebig vertauscht, mit Faktoren multipliziert oder voneinander subtrahiert werden. Daher können wir die Matrix durch Zeilenoperationen in eine andere Matrix überführen ohne daß sich die Lösung $\boldsymbol{x}$ ändert. Dieselben Operationen müssen natürlich auch auf den Vektor $\boldsymbol{b}$ angewandt werden.

Wie in (5.7) angedeutet, überführen wir A in eine neue Matrix, in der alle Elemente in der Spalte unterhalb von A_{11} gleich Null sind. Damit ist das System mit N Unbekannten in ein System der Dimension $N-1$ überführt worden, denn die Gleichungen $n = 2, \ldots, N$ enthalten nun nicht mehr die Unbekannte x_1. Dieses Verfahren wird nun weitergeführt, wobei im zweiten Schritt mit Hilfe der zweiten Zeile die Matrix-Elemente $A_{32}, ..., A_{N2}$ auf Null gebracht werden. Wenn man dies fortsetzt, hat man am Ende ein Problem, in dem nur die *obere Dreiecksmatrix* besetzt ist:

$$\mathsf{A} = \begin{pmatrix} A_{11} & A_{12} & \ldots & A_{1N} \\ & A_{22} & \ldots & A_{2N} \\ & & \ddots & \vdots \\ & & & A_{NN} \end{pmatrix} = \mathsf{U}. \tag{5.8}$$

Hierbei wurden für die obere Dreiecksmatrix U dieselben Symbole A_{ij} verwendet, obwohl nur die erste Zeile von A und die Komponente b_1 von $\boldsymbol{b}$ gleich geblieben sind. Alle anderen Elemente wurden modifiziert. Da man die ursprünglichen Elemente aber nicht mehr benötigt, können die entsprechenden Speicherplätze mit den neuen Einträgen überschrieben werden. Dieser Teil des Algorithmus wird *forward elimination* genannt.

Nun können wir das Problem leicht lösen, denn aus der letzten Gleichung von (5.8) ergibt sich

$$x_N = \frac{b_N}{A_{NN}}. \tag{5.9}$$

Dieses Resultat kann in Gleichung $N-1$ eingesetzt werden, um die Gleichung $N-1$ nach x_{N-1} aufzulösen. Wenn man dies sukzessive von $n = N$ rückwärts bis $n = 1$ weiterführt, erhalten wir aus der n-ten Gleichung

$$A_{nn}x_n + \underbrace{\sum_{k=n+1}^{N} A_{nk}x_k}_{\text{bekannt}} = b_n \tag{5.10}$$

die Lösung

$$x_n = \frac{b_n - \sum_{k=n+1}^{N} A_{nk}x_k}{A_{nn}}. \tag{5.11}$$

Alle Koeffizienten sind aus der *forward elimination* bekannt und die Komponenten der Lösung x_k mit $k > N$ wurden in den vorherigen Schritten berechnet. Dieser Teil des Algorithmus wird *back substitution* genannt.

Die *forward elimination* kostet $\sim N^3/3$ Operationen[2] und ist damit wesentlich aufwendiger als die *back substitution* mit $\sim N^2/2$ Operationen. Die Gauß-Elimination ist also sehr rechenintensiv, aber für vollbesetzte Matrizen gibt es praktisch kein anderes Verfahren, das wesentlich besser wäre.

Bei der Diskussion des Gauß-Algorithmus haben wir angenommen, daß die Diagonalelemente von Null verschieden sind. Falls dies einmal nicht der Fall sein sollte, kann man die betreffende Zeile n mit einer anderen Zeile $m > n$ vertauschen. Wenn die Matrix regulär ist, gibt es dort immer ein von Null verschiedenes Matrixelement. Die Zeilenvertauschung entspricht nur einer Vertauschung der Gleichungen.

Wenn die Gauß-Elimination auf große voll besetzte Matrizen angewandt wird, kann es zu einer Akkumulation von Fehlern kommen (Golub and van Loan, 1989). Um dies zu vermeiden, sollten die Diagonalelemente (*Pivot-Elemente*) betragsmäßig so groß wie möglich sein.[3] Denn dann sind die Multiplikatoren in Gauß-Verfahren (5.7) möglichst klein und Rundungsfehler werden minimiert. Um die Pivot-Elemente betragsmäßig zu maximieren, werden vor der Eliminierung einer Spalte n die Zeilen mit $m \geq n$ unterhalb des aktuellen Diagonalelements so vertauscht, daß auf der Diagonale das betragsmäßig größtmögliche Element steht. Diese Strategie nennt man *Teilpivotisierung*. Bei der *Totalpivotisierung* werden Zeilen *und* Spalten der jeweilige Untermatrix vertauscht, so daß das betragsmäßig größte Element der Untermatrix auf die Diagonalposition kommt. Die vollständige Pivotisierung ist aber mit einem großen Mehraufwand verbunden. Deshalb wir darauf oft verzichtet.

Gauß-Elimination läßt sich nicht gut vektorisieren oder parallelisieren und wird daher relativ selten bei CFD-Problemen verwendet.

[2] Die Anzahl der Operationen für große N ist $\propto \sum_{k=1}^{N} k^2 \approx \int_0^N k^2 \, \mathrm{d}k = N^3/3$.

[3] Ansonsten können Differenzen sehr großer Zahlen auftreten, die sich fast kompensieren, was zu den besagten numerischen Fehlern führt.

5.1.2. LU-Zerlegung

Eine wichtige Variante der Gauß-Elimination ist die *LU-Zerlegung*. Sie ist vorteilhaft, wenn man mehrere Lösungen von (5.6) für verschiedene rechte Seiten $\boldsymbol{b}$ berechnen muß.

Den erste Schritt der Gauß-Elimination kann man auch als Matrix-Multiplikation $\mathsf{L}_1 \cdot \mathsf{A}$ schreiben, wobei

$$\mathsf{L}_1 = \begin{pmatrix} 1 & & & \\ -l_{21} & 1 & & \\ \vdots & & \ddots & \\ -l_{N1} & & & 1 \end{pmatrix} \tag{5.12}$$

eine *untere Dreiecksmatrix* ist mit $l_{m1} = A_{m1}/A_{11}$. Im n-ten Zwischenschritt kann man die Elimination der n-ten Teilspalte als Matrix-Multiplikation mit der unteren Dreiecksmatrix

$$\mathsf{L}_n = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & -l_{n+1,n} & 1 & & \\ & & \vdots & & \ddots & \\ & & -l_{N,n} & & & 1 \end{pmatrix} \tag{5.13}$$

mit $l_{mn} = A_{mn}/A_{nn}$ und $n < m \leq N$ auffassen. Die Gauß-Elimination, welche auf eine obere Dreiecksmatrix führt, läßt sich also als eine Sequenz von Multiplikationen mit unteren Dreiecksmatrizen darstellen

$$\underbrace{\mathsf{L}_{N-1} \cdot \ldots \cdot \mathsf{L}_2 \cdot \mathsf{L}_1}_{\hat{\mathsf{L}}} \cdot \mathsf{A} = \mathsf{U}. \tag{5.14}$$

Alle Matrizen L_n sind regulär, da ihre Determinanten gleich 1 sind. Demnach ist ihr Produkt $\hat{\mathsf{L}}$ auch regulär. Wenn wir beachten, daß ein Produkt von unteren Dreiecksmatrizen wieder eine untere Dreiecksmatrix ist, und daß die Inverse einer unteren Dreiecksmatrix auch wieder eine untere Dreiecksmatrix ist, dann läßt sich A schreiben als

$$\mathsf{A} = \hat{\mathsf{L}}^{-1} \cdot \mathsf{U} = \mathsf{L} \cdot \mathsf{U}. \tag{5.15}$$

Da alle Dreicksmatrizen L_n Einsen auf der Diagonale haben, hat auch $\hat{\mathsf{L}}$ Einsen auf der Diagonale. Man kann zeigen (Trefethen and Bau, III, 1997), daß

$$\mathsf{L} = \mathsf{L}_1^{-1} \cdot \mathsf{L}_2^{-1} \cdot \ldots \cdot \mathsf{L}_{N-1}^{-1} = \begin{pmatrix} 1 & & & & \\ l_{21} & 1 & & & \\ l_{31} & l_{32} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ l_{N1} & l_{N2} & \ldots & l_{N,N-1} & 1 \end{pmatrix}. \tag{5.16}$$

Damit kostet die LU-Zerlegung genausoviel wie die Gauß-Elimination, denn die Gauß-Elimination liefert U und ganz nebenbei L ohne zusätzlichen Aufwand. In der Praxis werden die Matrizen L_n nie aufgestellt, auch nicht L und U. Denn man kann für L und U die Speicherplätze der Matrix A verwenden, weil die Zwischenergebnisse nicht explizit gespeichert werden müssen.

Mit der LU-Zerlegung kann man das lineare Problem (5.6) in zwei Stufen lösen

$$\mathsf{A}\cdot\boldsymbol{x} = \mathsf{L}\cdot\underbrace{\mathsf{U}\cdot\boldsymbol{x}}_{:=\boldsymbol{y}} = \boldsymbol{b} \tag{5.17a}$$

$$\Leftrightarrow \quad \mathsf{L}\cdot\boldsymbol{y} = \boldsymbol{b}, \tag{5.17b}$$

$$\mathsf{U}\cdot\boldsymbol{x} = \boldsymbol{y}. \tag{5.17c}$$

Die beiden letzten Gleichungen können nun sukzessive mittels *back substitution* effizient gelöst werden, erst (5.17b) dann (5.17c). Wenn man Lösungen für verschiedene Werte von $\boldsymbol{b}$ sucht, ist die LU-Zerlegung vorteilhaft. Denn die LU-Zerlegung muß nur einmal durchgeführt werden, da die Kenntnis von $\boldsymbol{b}$ dazu nicht erforderlich ist. Dann erhält man die verschiedenen Lösungen jeweils mit nur $O(N^2)$ Operationen.

Nicht jede Matrix A kann LU-zerlegt werden. Wenn aber A regulär ist, d.h. $\det|\mathsf{A}| \neq 0$, dann existiert eine eindeutige LU-Zerlegung (Golub and van Loan, 1989).[4]

5.1.3. Tridiagonale Systeme

Eine besonders effiziente Lösung existiert für *tridiagonale Systeme*. Diese ergeben sich zum Beispiel bei der Diskretisierung einer gewöhnlichen Differentialgleichung (eindimensionales Problem) mittels zentraler finiter Differenzen (vgl. (2.4)),

$$\begin{pmatrix} b_1 & c_1 & & & & & & \\ a_2 & b_2 & c_2 & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & a_{n-1} & b_{n-1} & c_{n-1} & & & \\ & & & a_n & b_n & c_n & & \\ & & & & \ddots & \ddots & \ddots & \\ & & & & & a_{N-1} & b_{N-1} & c_{N-1} \\ & & & & & & a_N & b_N \end{pmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \end{matrix} \cdot\boldsymbol{x} = \boldsymbol{d}. \tag{5.18}$$

Dirichlet-Randbedingungen gehen dabei in die Elemente d_1 und d_N ein. In der Matrix sind nur die Hauptdiagonale und die beiden ersten Nebendiagonalen besetzt. Die untere Nebendiagonale a_n kann man mit dem Gauß-Schema eliminieren. Wenn der Algorithmus schon bis zur Zeile $n-1$ durchgeführt wurde, d.h. wenn $a_2 = \ldots =$

[4] Wenn lediglich die Determinanten aller Untermatrizen $\det|\mathsf{A}(1:n, 1:n)| \neq 0$ nicht verschwinden für alle $n \in [1, N-1]$, dann ist A LU-zerlegbar (hinreichende Bedingung). Die Zerlegung muß aber nicht eindeutig sein.

$a_{n-1} = 0$ schon eliminiert wurden, dann wird in der nächsten Zeile a_n durch

$$\text{Zeile}(n) \quad \longrightarrow \quad \text{Zeile}(n) \; - \; \frac{a_n}{b_{n-1}} \times \text{Zeile}(n-1) \tag{5.19}$$

eliminiert. Die neuen Matrix-Elemente der Zeile n und d_n ergeben sich dann zu

$$a_n \quad \longrightarrow \quad a_n - \frac{a_n}{b_{n-1}} b_{n-1} = 0, \tag{5.20}$$

$$b_n \quad \longrightarrow \quad b_n - \frac{a_n}{b_{n-1}} c_{n-1}, \tag{5.21}$$

$$c_n \quad \longrightarrow \quad c_n - \frac{a_n}{b_{n-1}} \times 0 = c_n, \tag{5.22}$$

$$d_n \quad \longrightarrow \quad d_n - \frac{a_n}{b_{n-1}} d_{n-1}. \tag{5.23}$$

Dieser sogenannte *forward sweep* führt auf die einfache obere Dreiecksmatrix[5]

$$\begin{pmatrix} b_1 & c_1 & & & & & & \\ & b_2 & c_2 & & & & & \\ & & \dots & & & & & \\ & & & b_{n-1} & c_{n-1} & & & \\ & & & & b_n & c_n & & \\ & & & & & \dots & & \\ & & & & & & b_{N-1} & c_{N-1} \\ & & & & & & & b_N \end{pmatrix} \cdot \boldsymbol{x} = \boldsymbol{d}, \tag{5.24}$$

wodurch das Gleichungssystem mit *back substitution* gelöst werden kann. Wenn $x_N, x_{N-1}, ..., x_{n+1}$ schon berechnet wurden, dann ergibt sich x_n aus

$$x_n = \frac{d_n - c_n x_{n+1}}{b_n}. \tag{5.25}$$

Dieses Verfahren wird auch *Thomas-Algorithmus* oder *TDMA* (***t**ri-**d**iagonal **m**atrix **a**lgorithm*) genannt. Der Algorithmus ist besonders ökonomisch, denn die erforderlichen Operationen skalieren linear mit der Anzahl N der Unbekannten. Effizienter geht es praktisch nicht. Wenn immer möglich wird daher dieser Algorithmus bzw. Varianten davon benutzt.

In dem Spezialfall, daß alle Elemente einer Diagonale denselben Wert besitzen, kann man das System noch effizienter lösen. Dies kann man mit *zyklischer Reduktion* erreichen, die nur $O(\log_2 N)$ Operationen benötigt.

Der Thomas-Algorithmus kann leicht auf Systeme mit mehreren Diagonalen erweitert werden. Außerdem kann man damit Systeme lösen, deren Matrizen eine *Block-Tridiagonalstruktur* besitzen (Fletcher, 1991*a*). Dann sind alle obigen Operationen vektoriell zu verstehen.

[5] Beachte, daß die modifizierten Werte auch hier wieder dieselbe Bezeichnung haben.

5.2. Iterative Lösung linearer Gleichungssysteme

Jedes lineare System kann im Prinzip mit dem Gauß-Algorithmus bzw. mit LU-Zerlegung gelöst werden. Die Dreiecksmatrizen der LU-Zerlegung sind normalerweise voll besetzt. Dies ist auch dann der Fall, wenn die Matrix A dünn besetzt ist. Außerdem ist der Diskretisierungsfehler viel größer als der Fehler der Computerarithmetik. Es bringt also nichts, wenn man das lineare System genauer löst, als durch den Diskretisierungsfehler vorgegeben.

Generell sind iterative Methoden zur Lösung nichtlinearer Systeme erforderlich. Es ist aber aus den obigen Gründen sinnvoll, sie auch für lineare Probleme zu verwenden, deren Matrizen dünn besetzt sind. Iterative Methoden sind immer dann effektiv, wenn jede Iteration billig ist und man nicht zu viele Iterationen benötigt, um zu einer hinreichenden Genauigkeit der Lösung zu kommen. Das Buch von Saad (2003) enthält einen sehr guten Überblick über iterative Verfahren.

5.2.1. Allgemeines Konzept

Wir betrachten hier wieder das lineare System[6]

$$\mathsf{A} \cdot \boldsymbol{x} = \boldsymbol{b} \tag{5.26}$$

und spalten die Matrix A auf als

$$\mathsf{A} = \mathsf{M} - \mathsf{N}. \tag{5.27}$$

Damit lautet das Problem

$$\mathsf{M} \cdot \boldsymbol{x} = \mathsf{N} \cdot \boldsymbol{x} + \boldsymbol{b}. \tag{5.28}$$

Um zu einem allgemeinen linearen Iterations-Schema zu kommen setzen wir nun

$$\mathsf{M} \cdot \boldsymbol{x}^{(n+1)} = \mathsf{N} \cdot \boldsymbol{x}^{(n)} + \boldsymbol{b} \tag{5.29}$$

Die Frage ist nur, wie man A in M und N aufspalten muß, um eine schnelle Konvergenz von $\boldsymbol{x}^{(n)}$ auf die Lösung von (5.26) zu erreichen.

Für eine effiziente Lösung sollte das Iterationsschema (5.29) mit möglichst wenigen Operationen möglichst schnell konvergieren. Dies erfordert

1. eine *schnelle* Berechnung der rechten Seite von (5.29),
2. eine *einfache* Lösung des verbleibenden linearen Systems, und
3. eine Konvergenz nach *wenigen Iterationen*.

[6] Ggf. ist eine Vorkonditionierung nötig. Dies wird durch die Transformation (Multiplikation) mit einer Matrix P erreicht $\mathsf{A} \cdot \boldsymbol{x} = \boldsymbol{b} \;\rightarrow\; \mathsf{P} \cdot \mathsf{A} \cdot \boldsymbol{x} = \mathsf{P} \cdot \boldsymbol{b}$. Mit dieser Vorkonditionierung kann man eine Akkumulation von Rundungsfehlern vermeiden und u.U. die Konvergenz des iterativen Verfahrens beschleunigen. Allerdings ist es nicht immer einfach, eine geeignete Form von P zu

Da wir davon ausgehen, daß A dünn besetzt ist, sind auch M und N dünn besetzt und $\mathsf{N} \cdot \boldsymbol{x}^{(n)}$ ist schnell zu berechnen (Punkt 1). Um das System schnell lösen zu können (Punkt 2), sollte M schnell zu invertieren sein. Daher sollte M möglichst diagonal oder tridiagonal sein, oder Dreiecksform haben. Um mit wenigen Iterationen auszukommen (Punkt 3), sollte schließlich M eine möglichst gute Approximation von A darstellen, wodurch $\mathsf{N} \cdot \boldsymbol{x}$ in einem gewissen Sinne *klein* wird. Im Grenzfall $\mathsf{N} \cdot \boldsymbol{x} = 0$ wäre man ja nach nur einer Iteration fertig.

Bevor wir uns einer theoretischen Konvergenzbetrachtung zuwenden, wollen wir noch einige weitere Größen definieren und deren Beziehung untereinander klären. Nach n Iterationen haben wir im allgemeinen nur eine Näherung der exakten Lösung $\boldsymbol{x}$ erhalten. Dann ist (5.26) nicht exakt erfüllt, und es bleibt ein *Residuum* $\boldsymbol{\rho}^{(n)}$ übrig

$$\mathsf{A} \cdot \boldsymbol{x}^{(n)} = \boldsymbol{b} - \boldsymbol{\rho}^{(n)}. \tag{5.30}$$

Wenn wir dies von der exakten Gleichung abziehen, erhalten wir

$$\mathsf{A} \cdot \underbrace{\left(\boldsymbol{x} - \boldsymbol{x}^{(n)}\right)}_{\boldsymbol{\epsilon}^{(n)}} = \mathsf{A} \cdot \boldsymbol{\epsilon}^{(n)} = \boldsymbol{\rho}^{(n)}. \tag{5.31}$$

Hierbei haben wir den *Fehler* $\boldsymbol{\epsilon}^{(n)} := \boldsymbol{x} - \boldsymbol{x}^{(n)}$ definiert. Zwischen dem Fehler und dem Residuum besteht ein linearer Zusammenhang. Wenn das Residuum verschwindet, dann verschwindet auch der Fehler.

Eine weitere nützliche Größe ergibt sich, wenn wir $\mathsf{M} \cdot \boldsymbol{x}^{(n)}$ von (5.29) abziehen. Dann erhalten wir eine Gleichung für die *Korrektur* $\boldsymbol{\delta}^{(n)} := \boldsymbol{x}^{(n+1)} - \boldsymbol{x}^{(n)}$

$$\mathsf{M} \cdot \underbrace{\left(\boldsymbol{x}^{(n+1)} - \boldsymbol{x}^{(n)}\right)}_{\boldsymbol{\delta}^{(n)}} = \underbrace{\underbrace{(\mathsf{N} - \mathsf{M})}_{-\mathsf{A}} \cdot \boldsymbol{x}^{(n)} + \boldsymbol{b}}_{\boldsymbol{\rho}^{(n)}}. \tag{5.32}$$

Für die Korrektur gilt also

$$\mathsf{M} \cdot \boldsymbol{\delta}^{(n)} = \boldsymbol{\rho}^{(n)}. \tag{5.33}$$

5.2.2. Konvergenz iterativer Löser

Hier wollen wir mit einfachen Betrachtungen untersuchen, worauf es ankommt, wenn man eine schnelle Konvergenz eines iterativen Verfahrens erhalten möchte. Dazu suchen wir zunächst eine Gleichung für den Iterationsfehler $\boldsymbol{\epsilon}^{(n)} = \boldsymbol{x} - \boldsymbol{x}^{(n)}$. Wenn man (5.28) von der Iterationsgleichung (5.29) subtrahiert, erhält man

$$\mathsf{M} \cdot \boldsymbol{\epsilon}^{(n+1)} = \mathsf{N} \cdot \boldsymbol{\epsilon}^{(n)} \qquad \Rightarrow \qquad \boldsymbol{\epsilon}^{(n+1)} = \mathsf{M}^{-1} \cdot \mathsf{N} \cdot \boldsymbol{\epsilon}^{(n)}. \tag{5.34}$$

finden.

Für Konvergenz muß nun gelten: $\lim_{n\to\infty} \boldsymbol{\epsilon}^{(n)} = 0$. Für den Grenzwert der Iteration (5.34) spielen nun die *Eigenwerte* λ und die *Eigenvektoren* $\boldsymbol{\psi}$ der Matrix $\mathsf{M}^{-1} \cdot \mathsf{N}$ eine entscheidende Rolle. Sei daher

$$\mathsf{M}^{-1} \cdot \mathsf{N} \cdot \boldsymbol{\psi}^{(k)} = \lambda_k \cdot \boldsymbol{\psi}^{(k)}, \tag{5.35}$$

mit $k \in [1, K]$, wobei K die Anzahl der Gleichungen (Gitterpunkte) ist. Wenn wir annehmen, daß die Eigenvektoren ein *vollständiges Funktionensystem* bilden, dann können wir jeden beliebigen Anfangsfehler als Superposition von Eigenvektoren darstellen

$$\boldsymbol{\epsilon}^{(0)} = \sum_{k=1}^{K} a_k \boldsymbol{\psi}^{(k)}. \tag{5.36}$$

Wenn man dies in (5.34) einsetzt, sieht man, daß jeder Iterationsschritt lediglich einen Faktor λ_k vor $\boldsymbol{\psi}^{(k)}$ ausspuckt. Somit erhalten erhalten wir

$$\boldsymbol{\epsilon}^{(n)} = \sum_{k=1}^{K} a_k \left(\lambda_k\right)^n \boldsymbol{\psi}^{(k)}. \tag{5.37}$$

Hieran sieht man: Wenn $\boldsymbol{\epsilon}^{(n)}$ für $n \to \infty$ gegen Null gehen soll, muß für *alle* Eigenwerte gelten $|\lambda_k| < 1$, denn die Eigenvektoren sind linear unabhängig voneinander.

Nach einigen Iterationen wird derjenige Term in (5.37) dominieren, der zum betragsmäßig größten Eigenwert gehört. Dies sei o.B.d.A. λ_1. Dann ist

$$\boldsymbol{\epsilon}^{(n)} \approx a_1 \left(\lambda_1\right)^n \boldsymbol{\psi}^{(1)}. \tag{5.38}$$

Wenn wir die Iteration bei einer *Toleranzschwelle* $|\boldsymbol{\epsilon}^{(n)}| = O(\delta)$ für den Iterationsfehler abbrechen, erhalten wir als Abbruchbedingung [es sei $|\boldsymbol{\psi}^{(1)}| = O(1)$]

$$\delta \sim a_1 \left|\lambda_1\right|^n, \tag{5.39}$$

woraus wir durch Logarithmieren die Anzahl n der erforderlichen Iterationen erhalten

$$n \sim \frac{\ln\left(\delta / a_1\right)}{\ln\left|\lambda_1\right|}. \tag{5.40}$$

Wenn also der *spektrale Radius* (größter Betrag eines Eigenwerts $R = \max_k |\lambda_k|$; hier $R = |\lambda_1|$) gegen 1 geht, wird n sehr groß und die Konvergenz ist langsam. Umgekehrt ist die Konvergenz schnell, wenn der Betrag des betragsmäßig größten Eigenwerts von $\mathsf{M}^{-1} \cdot \mathsf{N}$ möglichst klein wird. Dies entspricht dem Fall, in dem $\mathsf{N} \cdot \boldsymbol{x}$ möglichst klein und $\mathsf{M} \approx \mathsf{A}$ ist. Denn für $\mathsf{M} = \mathsf{A}$ und damit $\mathsf{N} = 0$ wäre man ja schon nach einer einzigen Iteration fertig.

Um entscheiden zu können, wann man eine Iteration abbrechen sollte, muß man den Iterationsfehler $\boldsymbol{\epsilon}^{(n)}$ kennen. Da man die Eigenwerte λ_k aber meist nicht kennt oder sie nur mit großem Aufwand berechnen kann, muß der Iterationsfehler anders abgeschätzt werden.

5.2.3. Einige elementare Methoden

Jacobi-Iteration

Eine der einfachsten iterativen Methoden ist die *Jacobi-Iteration*. Hierbei ist M diagonal und die Diagonalelemente werden identisch mit denjenigen von A gewählt.

Betrachte zum Beispiel die stationäre inhomogene Diffusionsgleichung (*Poisson-Gleichung*)

$$\nabla^2\phi = b \tag{5.41}$$

auf einem Quadrat. In beiden Koordinatenrichtungen werden jeweils K Gitterpunkte verwendet. Die Diskretisierung mittels finiter Differenzen (4.43) kann man dann schreiben als

$$A_P\phi_P + \sum_{l=W,S,N,E} A_l\phi_l = b_P, \tag{5.42}$$

wobei der Index P alle $K \times K$ Gitterpunkte durchläuft und die Summe nur über die jeweiligen 4 physikalischen Nachbarpunkte des aktuellen Punkts P zu nehmen ist. Wir ordnen nun die K^2 Variablen wie in Abb. 5.1b in der Form an

$$\boldsymbol{\phi} = (\phi_{1,1} : \phi_{1,K},\ \phi_{2,1} : \phi_{2,K},\ \ldots,\ \phi_{K,1} : \phi_{K,K})^{\mathrm{T}}, \tag{5.43}$$

wobei der zweite (schnelle) Index vertikal von unten (S) nach oben (N) läuft, und der erste (langsame) Index von links (W) nach rechts (E). Wenn man nun die K^2 Gleichungen für alle Punkte aufstellt, erhält man ein lineares Gleichungssystem, in dem die Matrix Matrix A pentadiagonal ist, so wie in Abb. 5.2 dargestellt. Bei einem äquidistanten Gitter mit $\Delta x = \Delta y$ sind die Diagonalen von A konstant und haben die Werte $A_P = 4$ sowie $A_W = A_S = A_N = A_E = -1$.[7]

Wenn nun $\mathsf{M} = \mathrm{diag}(\mathsf{A})$ identisch mit der Diagonalen von A gewählt wird (siehe Abb. 5.3a), kann man das Iterationsschema (5.29) sofort nach den Diagonalelementen auflösen und erhält

$$\phi_P^{(n+1)} = \frac{1}{A_P}\left(b_P - \sum_{l=W,S,N,E} A_l\phi_l^{(n)}\right). \tag{5.44}$$

Dies nennt man *Jacobi-Iteration*. Man kann zeigen, daß die Anzahl N der für die Konvergenz erforderlichen Iterationen proportional ist zur Anzahl der Unbekannten: $N \sim K^2$. Dies sind sehr viele Iterationen. Zusammen mit dem Aufwand $O(N)$ für jeden Iterationsschritt benötigt man damit mehr Operationen (also $\sim N^2$), als man für einen direkten Löser nach Art von TDMA ($\sim N$) benötigen würde. Daher wird die Jacobi-Iteration selten verwendet.

Gauß-Seidel-Iteration

Bei der *Gauß-Seidel-Iteration* wird für M die untere Dreiecksmatrix von A verwendet (Abb. 5.3b). Diese stellt eine bessere Approximation von A dar als die Diagonale.

[7] Dies gilt, nachdem man die Gleichung mit $-\Delta x^2$ multipliziert hat, vgl. (4.96).

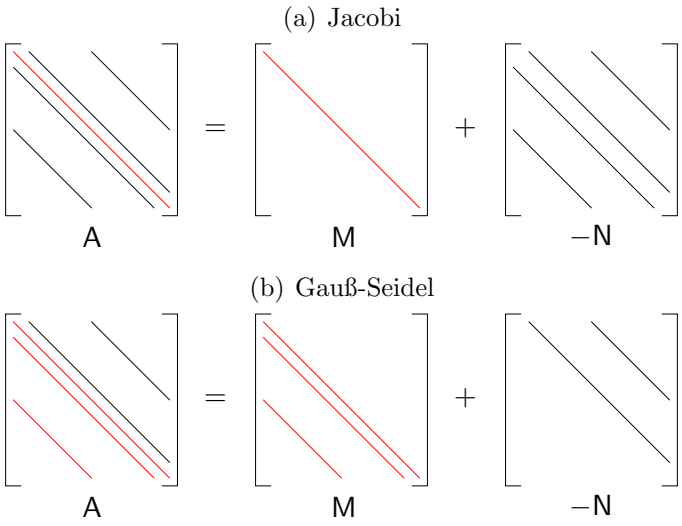


Abbildung 5.3.: Symbolische Darstellung der Zerlegung der Matrix A in die Iterationsmatrix M und einen Teil −N, der der rechten Seite der Gleichung zugeschlagen wird (vgl. 5.29). (a) Jacobi-Iteration, (b) Gauß-Seidel-Iteration.

Wir erwarten also eine Reduktion der Anzahl der erforderlichen Iterationen im Vergleich zum Jacobi-Verfahren. Damit lautet die *Gauß-Seidel-Iteration* (vergleiche Abb. 5.2)

$$\sum_{l=W,S,P} A_l \phi_l^{(n+1)} = b_P - \sum_{l=N,E} A_l \phi_l^{(n)}. \tag{5.45}$$

Wenn man die Gleichungen beginnend mit der linken unteren Ecke (SW-Ecke) löst, d.h. für aufsteigenden globalen Index l, sind für jeden Punkt P die Werte an den Punkten S und W jeweils schon zuvor berechnet worden und daher bekannt. Dann kann man (5.45) nach $\phi_P^{(n+1)}$ auflösen

$$\phi_P^{(n+1)} = \frac{1}{A_P}\left(b_P - A_W \phi_W^{(n+1)} - A_S \phi_S^{(n+1)} - A_N \phi_N^{(n)} - A_E \phi_E^{(n)}\right). \tag{5.46}$$

Wie man in Abb. 5.4 sieht, konvergiert die Gauß-Seidel-Iteration doppelt so schnell wie die Jacobi-Iteration, benötigt aber immer noch $\sim N$ Iterationen und einen Gesamtaufwand von $O(N^2)$ Operationen. D.h., auch die Gauß-Seidel-Iteration ist noch zu langsam.[8]

[8] Alternativ zu (5.45) bzw. (5.46) kann man in M auch die oberen Diagonalen von A verwenden (anstelle der unteren) und

$$\sum_{l=N,E,P} A_l \phi_l^{(n+1)} = b_P - \sum_{l=W,S} A_l \phi_l^{(n)}$$

für *absteigende* Werte von l lösen.

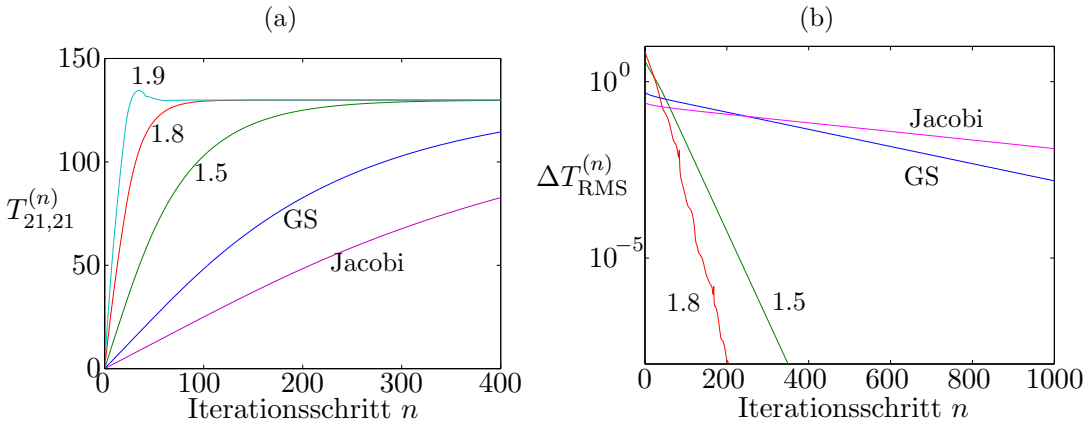


Abbildung 5.4.: (a) Konvergenz des Jacobi-, Gauß-Seidel- (GS) und des SOR-Verfahrens in Abhängigkeit vom Relaxationsfaktor ω (als Zahlen angegeben) für die Poisson-Gleichung $\nabla^2 T = -1$, die auf $K_i \times K_j = 41 \times 41$ (inneren) Punkten mit zentralen finiten Differenzen auf einem Quadrat mit Anfangsbedingung $T(t=0) \equiv 0$ diskretisiert wurde (siehe (4.78) und (4.96) mit $\Gamma = 1$). Für $\omega = 1.9$ ist man schon oberhalb des Optimums, da die Konvergenz nicht mehr monoton ist. Das (langsame) Gauß-Seidel-Verfahren (GS) entspricht $\omega = 1$. (b) Logarithmische Auftragung des mittleren RMS-Inkrements $\Delta T^{(n)}_{\mathrm{RMS}} := \sqrt{(K_i K_j)^{-1} \sum_{l=1}^{K_i K_j} \left(T_l^{(n+1)} - T_l^{(n)} \right)^2}$ (Maß für den Korrekturvektor $\boldsymbol{\delta}^{(n)}$).

SOR-Methode

Durch eine Modifikation der Gauß-Seidel-Methode kann man die Konvergenz wesentlich beschleunigen. Die Iteration mittels *sukzessiver Überrelaxation* (***S**uccessive **O**ver-**R**elaxation*, SOR) lautet (wieder beginnend im Südwesten)

$$\begin{aligned} \phi_P^{(n+1)} &= \phi_P^{(n)} + \omega \left({}^{\mathrm{GS}}\phi_P^{(n+1)} - \phi_P^{(n)} \right) \\ &= \frac{\omega}{A_P} \left(b_P - A_W \phi_W^{(n+1)} - A_S \phi_S^{(n+1)} - A_N \phi_N^{(n)} - A_E \phi_E^{(n)} \right) + (1-\omega)\, \phi_P^{(n)}. \end{aligned} \tag{5.47}$$

Hierbei ist ω der *Überrelaxationsfaktor*. Für $\omega = 1$ erhält man die Gauß-Seidel-Iteration (5.46). Für $\omega > 1$ wird der GS-artige Iterationsterm übermäßig gewichtet (Beiträge von S, W, N und E), was durch einen alten Beitrag $(1-\omega)\, \phi_P^{(n)}$ vom zentralen Punkt P kompensiert wird. Wenn man $\omega > 1$ wählt, wird die Konvergenz beschleunigt. Für einfache Probleme wie die Laplace-Gleichung kann man den optimalen Wert ω_{opt} von ω theoretisch berechnen.[9] Für kompliziertere Probleme

[9] Fletcher (1991*a*) gibt an

$$\omega_{\mathrm{opt}} = \frac{2}{1 + \sqrt{1 - \mu^2}},$$

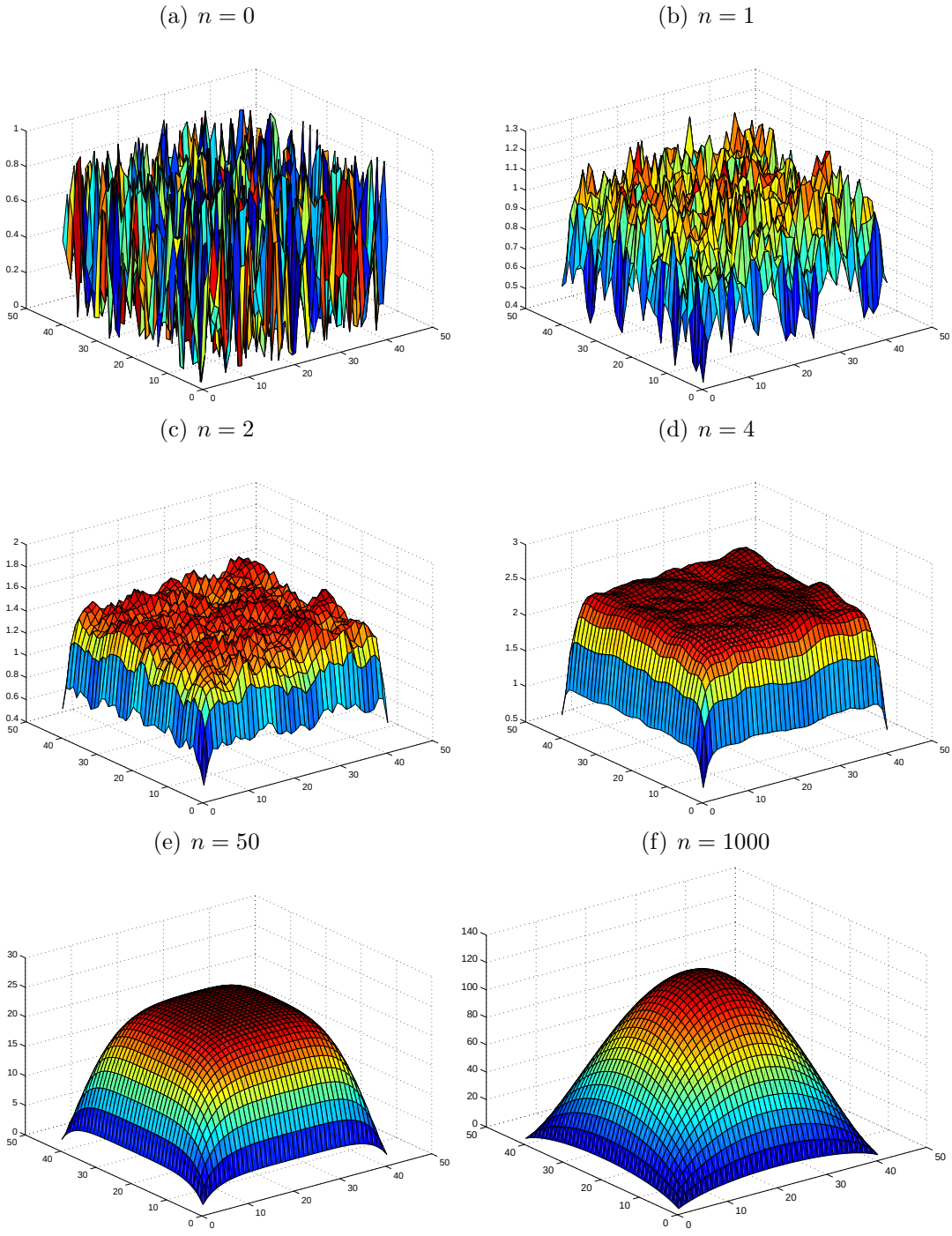


Abbildung 5.5.: Gauß-Seidel-Iteration von $\nabla^2 T = -1$ auf einem Quadrat mit einem Gitter von 41×41 inneren Punkten. Gezeigt sind die zufällige Anfangsbedingung ($n = 0$) mit $T_l \in [0, 1]$ auf dem Definitionsbereich $[0, 1] \times [0, 1]$ (a), sowie die Approximationen der Lösung T nach $n = 1$ (b), $n = 2$ (c), $n = 4$ (d), $n = 50$ (e) und $n = 1000$ (f) Iterationen. Man erkennt die guten Glättungseigenschaften der Gauß-Seidel-Iteration nach nur wenigen Iterationsschritten. Beachte die unterschiedlichen Skalen.

kann man ihn nur empirisch finden. Dabei hilft es zu wissen, daß die Konvergenz monoton ist für $\omega < \omega_{\text{opt}}$ und oszillatorisch für $\omega > \omega_{\text{opt}}$. Für $\omega = \omega_{\text{opt}}$ ist die erforderliche Zahl der Iterationen proportional zur Anzahl $\sim \sqrt{N}$ der Unbekannten, was eine substantielle Verbesserung gegenüber der Gauß-Seidel-Iteration darstellt, bei der die Anzahl der Iterationen ja linear mit der Anzahl der Unbekannten ansteigt.

Das typische Konvergenzverhalten der drei elementaren Methoden ist in Abb. 5.4 für die Lösung von $\nabla^2 T = -1$ auf dem Einheitsquadrat dargestellt. Abbildung 5.5 zeigt die Entwicklung der iterierten Lösung für das Gauß-Seidel-Verfahren. Insbesondere erkennt man, daß kurzwellige Schwankungen schnell gedämpft werden, was den Gauß-Seidel-Operator auch zu einem guten *Glättungsoperator* macht (siehe Kap. 5.2.6).

5.2.4. Unvollständige LU-Zerlegung und SIP-Algorithmus

Wir haben in Kap. 5.1.2 gesehen, daß die LU-Zerlegung ein sehr guter direkter Löser für lineare Probleme mit dichtbesetzter Matrix A ist. Da man bei der Iteration immer wieder dasselbe lineare Gleichungssystem lösen muß, wäre auch bei der iterativen Lösung eine LU-Zerlegung von A wünschenswert. Wenn A nun eine Bandstruktur besitzt, sollte man annehmen, daß sich die Bandstruktur auch auf die LU-Zerlegung der Matrix $\mathsf{A} = \mathsf{L} \cdot \mathsf{U}$ günstig auswirkt. Leider ist dies nicht der Fall und L und U besitzen keine Bandstruktur.

Auf der anderen Seite ist M eine gute Iterationsmatrix, wenn $\mathsf{M} \approx \mathsf{A}$ ist. Man könnte daher eine Iterationsmatrix M durch Approximation von A konstruieren, indem man

$$\mathsf{M} = \mathsf{L} \cdot \mathsf{U} = \mathsf{A} + \mathsf{N} \tag{5.48}$$

wählt, wobei L und U nur die näherungsweise LU-Zerlegung von A darstellen. Die Näherung besteht darin, daß in L und U nur diejenigen Diagonalen besetzt sind, die auch in A besetzt sind.

Wie aber sollten die Matrixelemente auf den Diagonalen gewählt werden? Eine Möglichkeit besteht darin, die Einträge auf den Diagonalen von L und U identisch mit denjenigen der exakten LU-Zerlegung von A zu wählen und die restlichen Elemente auf Null zu setzen. Damit wären M, L und U nur auf wenigen Diagonalen besetzt. Leider konvergiert diese Methode nur langsam.

Unvollständige LU-Zerlegung

Eine Modifikation mit wesentlich besserer Konvergenz wurde von Stone (1968) entwickelt. Wir nehmen zunächst an, daß wir schon geeignete Dreiecksmatrizen L und U haben, in denen nur jeweils die Diagonalen besetzt sind, die auch in A besetzt sind. Die genauen Einträge auf den Diagonalen werden wir später bestimmen. Für eine eindeutige Zerlegung wird jedoch die Hauptdiagonale von U auf 1 gesetzt. Die

wobei μ der größte Eigenwert von $\mathsf{I} - \text{diag}(\mathsf{A})^{-1} \cdot \mathsf{A}$ ist. Die Lösung dieses Problems kann aber genau so aufwendig sein, wie das Ausgangsproblem selbst.

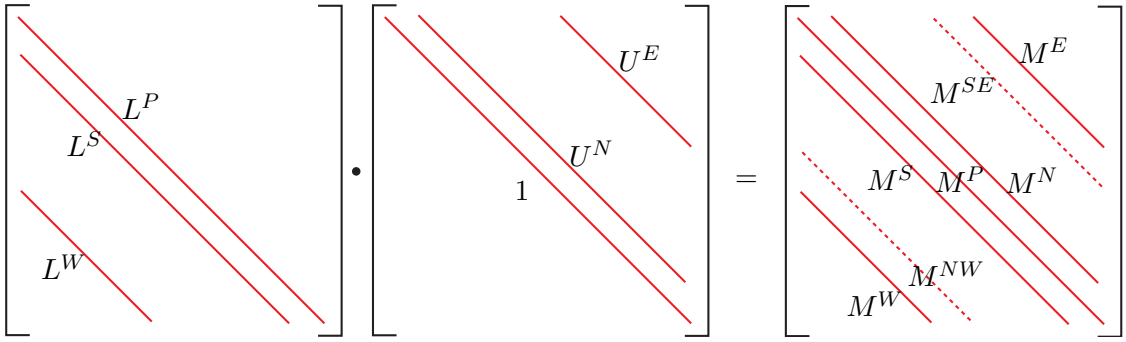


Abbildung 5.6.: Darstellung der Matrix-Multiplikation $\mathsf{L} \cdot \mathsf{U} = \mathsf{M}$. Die gestrichelten Diagonalen sind in A nicht besetzt.

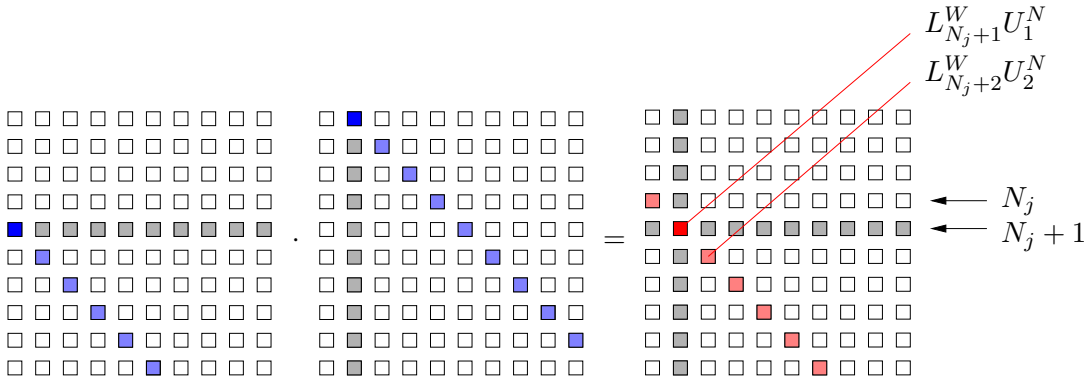


Abbildung 5.7.: Veranschaulichung der Multiplikation der Matrizen, die nur aus der Diagonalen L^W_l bzw. U^N_l bestehen (blau). In L^W_l sind die Elemente $1, ..., N_j$ trivialerweise Null. In U^N_l ist nur das Element N trivialerweise Null. Die Multiplikation ergibt $M^{NW}_l = L^W_l U^N_{l-N_j}$ (rot), wobei die Elemente $1, ..., N_j - 1$ trivialerweise Null sind, aber auch das Element $l = N_j$ ist Null.

Iterationsmatrix M ergibt sich dann aus dem Produkt von $\mathsf{M} = \mathsf{L} \cdot \mathsf{U}$. Für die 5-Punkt-Diskretisierung mittels zentraler Differenzen wie in Abb. 5.1a ist dieses Produkt in Abb. 5.6 graphisch dargestellt.

Um zu verstehen, wie die Diagonalen in M von denjenigen in L und U abhängen, betrachten wir das Produkt von L und U. Offensichtlich werden die Diagonalen L^W und L^S bei der Multiplikation mit der Einheits-Hauptdiagonalen U^P reproduziert. Zur Hauptdiagonale in M tragen bei L^P sowie die beiden Paare der symmetrisch gelegenen Diagonalen L^S und U^N sowie L^W und U^E. Die Diagonale M^{NW} kommt durch das Produkt von L^W mit U^N zustande. Für das erste von Null verschiedene Element von M^{NW} ist dies in Abb. 5.7 dargestellt. Die restlichen Diagonalen ergeben sich in analoger Weise.

Das Produkt können wir formal darstellen. Wenn wir beachten, daß der Zeilenindex der Matrix U identisch ist mit dem globalen Index l, der die Variablen numeriert (vgl. (5.3)), erhalten wir

$$\mathsf{M} = M_{l,k} = L_{l,j} U_{j,k}$$

$$
\begin{aligned}
&= \left(L_l^P\delta_{l,j} + L_l^S\delta_{l,j+1} + L_l^W\delta_{l,j+N_j}\right)\left(\delta_{j,k} + U_j^N\delta_{j,k-1} + U_j^E\delta_{j,k-N_j}\right)\\
&= \underbrace{\delta_{l,k}\left(L_l^P + L_l^SU_{l-1}^N + L_l^WU_{l-N_j}^E\right)}_{M^P} + \underbrace{\delta_{l,k+1}L_l^S}_{M^S} + \underbrace{\delta_{l,k-1}L_l^PU_l^N}_{M^N}\\
&\quad + \underbrace{\delta_{l,k+N_j}L_l^W}_{M^W} + \underbrace{\delta_{l,k-N_j}L_l^PU_l^E}_{M^E} + \underbrace{\delta_{l,k+N_j-1}L_l^WU_{l-N_j}^N}_{M^{NW}} + \underbrace{\delta_{l,k-N_j+1}L_l^SU_{l-1}^E}_{M^{SE}},
\end{aligned}
\tag{5.49}
$$

wobei wir die einzelnen Diagonalen leicht identifizieren können. Es sind also

$$
\begin{aligned}
M_l^W &= L_l^W, && \text{(5.50a)}\\
M_l^{NW} &= L_l^WU_{l-N_j}^N, \qquad (=0 \text{ in } \mathsf{A}), && \text{(5.50b)}\\
M_l^S &= L_l^S, && \text{(5.50c)}\\
M_l^P &= L_l^P + L_l^SU_{l-1}^N + L_l^WU_{l-N_j}^E, && \text{(5.50d)}\\
M_l^N &= L_l^PU_l^N, && \text{(5.50e)}\\
M_l^{SE} &= L_l^SU_{l-1}^E, \qquad (=0 \text{ in } \mathsf{A}), && \text{(5.50f)}\\
M_l^E &= L_l^PU_l^E. && \text{(5.50g)}
\end{aligned}
$$

Die Frage ist nun, wie man L und U am besten wählt. Damit M eine gute Approximation der pentadiagonalen Matrix A ist, sollten die in Abb. 5.6 gestrichelt gezeichneten Diagonalen nicht in M vorkommen: Deshalb könnte man sie N zuschlagen und L und U so bestimmen, daß die restlichen Diagonalen von M identisch sind mit denjenigen von A. Dann besäße N nur die beiden Diagonalen SE und NW. Dieses Vorgehensweise ist möglich und entspricht der standardmäßigen ILU-Methode. Doch es zeigt sich, daß dieses Verfahren schlecht konvergiert.

Eine bessere Konvergenz erhält man, wenn man auch in N alle sieben Diagonalen zuläßt. Um eine schnelle Iteration zu erhalten, sollte $\mathsf{N}\cdot\boldsymbol{x} \approx 0$ sein. Diese Forderung verwenden wir nun, um L und U zu bestimmen. Formal gilt[10]

$$
\begin{aligned}
\mathsf{N}\cdot\boldsymbol{x} &= (\mathsf{M}-\mathsf{A})\cdot\boldsymbol{x}\\
&= N^Px_P + N^Nx_N + N^Sx_S + N^Ex_E + N^Wx_W + \underline{N^{NW}x_{NW}} + \underline{N^{SE}x_{SE}}\\
&\overset{(5.50)}{=} \left(L_l^W - A_l^W\right)x_W + \underline{\left(L_l^WU_{l-N_j}^N\right)x_{NW}}\\
&\quad + \left(L_l^S - A_l^S\right)x_S + \left(L_l^WU_{l-N_j}^E + L_l^SU_{l-1}^N + L_l^P - A_l^P\right)x_P\\
&\quad + \left(U_l^NL_l^P - A_l^N\right)x_N + \underline{\left(L_l^SU_{l-1}^E\right)x_{SE}} + \left(U_l^EL_l^P - A_l^E\right)x_E.
\end{aligned}
\tag{5.51}
$$

An dieser Gleichung können wir die Diagonalen von N identifizieren und durch L, U und A ausdrücken. Damit $\mathsf{N}\cdot\boldsymbol{x} \approx 0$ wird, wäre es gut, wenn wir die bisher noch nicht bestimmten 5 unbekannten Diagonalen der Matrizen L und U so wählen könnten, daß alle 7 Koeffizienten vor den Unbekannten x_l verschwinden. Daran hindern uns

[10] Bei der Schreibweise N^Sx_S gibt S immer den Index relativ zu P an. Mit $P \triangleq l$ ist dann $S \triangleq l-1$

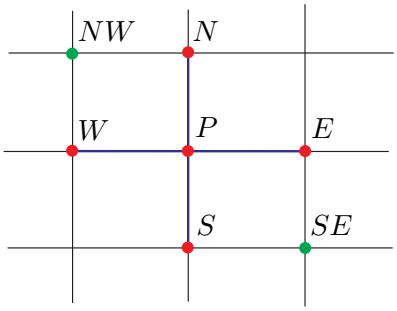


Abbildung 5.8.: Lage der Knoten NW und SE (grün) relativ zu den Punkten, die im Differenzenschema miteinander in Beziehung stehen (rot).

noch die 2 Terme proportional zu x_{NW} und x_{SE}. Deshalb drücken wir sie durch die Werte an den *normalen* Punkten aus

$$x_{NW} \approx \alpha \left(x_W + x_N - x_P \right), \tag{5.52a}$$
$$x_{SE} \approx \alpha \left(x_S + x_E - x_P \right), \tag{5.52b}$$

wobei $\alpha \approx 1$ sein sollte. Für $\alpha = 1$ entspricht dies einer linearen Interpolation (Abb. 5.8).[11] Eingesetzt erhalten wir

$$\begin{aligned} \mathsf{N} \cdot \boldsymbol{x} \approx & \left(L_l^W - A_l^W + \alpha L_l^W U_{l-N_j}^N \right) x_W + \left(L_l^S - A_l^S + \alpha L_l^S U_{l-1}^E \right) x_S \\ & + \left[L_l^W U_{l-N_j}^E + L_l^S U_{l-1}^N + L_l^P - A_l^P - \alpha \left(L_l^W U_{l-N_j}^N + L_l^S U_{l-1}^E \right) \right] x_P \\ & + \left(U_l^N L_l^P - A_l^N + \alpha L_l^W U_{l-N_j}^N \right) x_N + \left(U_l^E L_l^P - A_l^E + \alpha L_l^S U_{l-1}^E \right) x_E . \end{aligned} \tag{5.53}$$

Es ist $\mathsf{N} \cdot \boldsymbol{x} \approx 0$, wenn alle 5 Koeffizienten vor den gekoppelten Variablen verschwinden. Also setzen wir die Koeffizienten formal gleich Null und lösen die resultierenden 5 Bedingungen nach den 5 Diagonalen von U und L (rot in (5.53)) auf. Wir erhalten so

$$L_l^W = \frac{A_l^W}{1 + \alpha U_{l-N_j}^N}, \tag{5.54a}$$
$$L_l^S = \frac{A_l^S}{1 + \alpha U_{l-1}^E}, \tag{5.54b}$$
$$L_l^P = A_l^P + \alpha \left(L_l^W U_{l-N_j}^N + L_l^S U_{l-1}^E \right) - L_l^W U_{l-N_j}^E - L_l^S U_{l-1}^N, \tag{5.54c}$$
$$U_l^N = \frac{A_l^N - \alpha L_l^W U_{l-N_j}^N}{L_l^P}, \tag{5.54d}$$
$$U_l^E = \frac{A_l^E - \alpha L_l^S U_{l-1}^E}{L_l^P}. \tag{5.54e}$$

etc., gemäß Tabelle 5.1.

[11] An dieser Stelle setzen wir voraus, daß $\boldsymbol{x}$ *kontinuierlich* variiert und in dieser Form approxi-

Dies ist der Kern des *SIP-Verfahrens* (***s**trongly **i**mplicit **p**rocedure*) von Stone (1968) (siehe auch Leister and Perić, 1994).[12] Der Clou an der Sache ist, daß die Diagonalen sequentiell berechnet werden können, wenn man für jedes l (aufsteigend $1 \uparrow N$) die Sequenz (5.54) der Reihe nach löst, d.h. von (5.54a) nach (5.54e). Auf den rechten Seiten stehen nämlich nur bekannte Matrix-Elemente: A ist sowieso bekannt, und die Matrixelemente von L und U auf der rechten Seite sind entweder Null oder sie wurden schon vorher berechnet (der Index ist kleiner als l). Auch die Terme mit demselben Index l sind bekannt, wenn man die Gleichungen in der angegebenen Reihenfolge löst. Damit haben wir die unvollständige LU-Zerlegung erreicht. Wir müssen jetzt nur noch die Iteration durchführen.

SIP-Iteration

Der Zusammenhang (5.33) zwischen Korrektur $\boldsymbol{\delta}^{(n)}$ und Residuum $\boldsymbol{\rho}^{(n)}$ lautet mit $\mathsf{M} = \mathsf{L} \cdot \mathsf{U}$

$$\mathsf{L} \cdot \mathsf{U} \cdot \boldsymbol{\delta}^{(n)} = \boldsymbol{\rho}^{(n)}, \tag{5.55}$$

was nach Multiplikation mit L^{-1} auf

$$\mathsf{U} \cdot \boldsymbol{\delta}^{(n)} = \mathsf{L}^{-1} \cdot \boldsymbol{\rho}^{(n)} =: \boldsymbol{R}^{(n)} \tag{5.56}$$

führt. Man kann $\boldsymbol{R}^{(n)}$ berechnen. Dazu betrachten wir $\mathsf{L} \cdot \boldsymbol{R}^{(n)} = \boldsymbol{\rho}^{(n)}$, was man in der Form $\mathsf{L} \cdot \boldsymbol{R} = L_l^P R_l + L_l^S R_{l-1} + L_l^W R_{l-N_j} = \rho_l$ schreiben kann. Dann folgt

$$R_l = \frac{\rho_l - L_l^S R_{l-1} - L_l^W R_{l-N_j}}{L_l^P}. \tag{5.57}$$

Dieses Gleichungssystem können wir wieder sequentiell von $l = 1$ *aufsteigend* lösen. Wenn wir $\boldsymbol{R}$ berechnet haben, dann folgt aus $\mathsf{U} \cdot \boldsymbol{\delta} = \delta_l + U_l^N \delta_{l+1} + U_l^E \delta_{l+N_j} = R_l$ die Korrektur

$$\delta_l = R_l - U_l^N \delta_{l+1} - U_l^E \delta_{l+N_j}. \tag{5.58}$$

Diese Gleichung muß in Richtung *absteigender* Werte von l gelöst werden.

Abbildung 5.9 zeigt den Fortschritt der Iteration mittels SIP (zufällige Anfangsbedingungen $T \in [0, 1]$) und SOR (Anfangsbedingung $T \equiv 0$). Der Beginn der Instabilität des SIP-Verfahrens für zu hohe Werte von α ist in Abb. 5.9b zu erkennen. Die besonders zu Beginn der Iteration auftretende Asymmetrie der SOR-Iteration (Abb. 5.9c) resultiert aus der Asymmetrie von (5.47). Sie kann vermieden werden, wenn man alternierend (5.47) und deren Modifikation für absteigendes l verwendet.

Bemerkungen

miert werden kann. Aus diesem Grund ist der Algorithmus auch nur für Probleme geeignet, die hinreichend glatte Lösungen besitzen, also typischerweise Lösungen elliptischer PDEs.

[12] Die Bezeichnung SIP rührt daher, daß $\mathsf{M} \simeq \mathsf{A}$ ist.

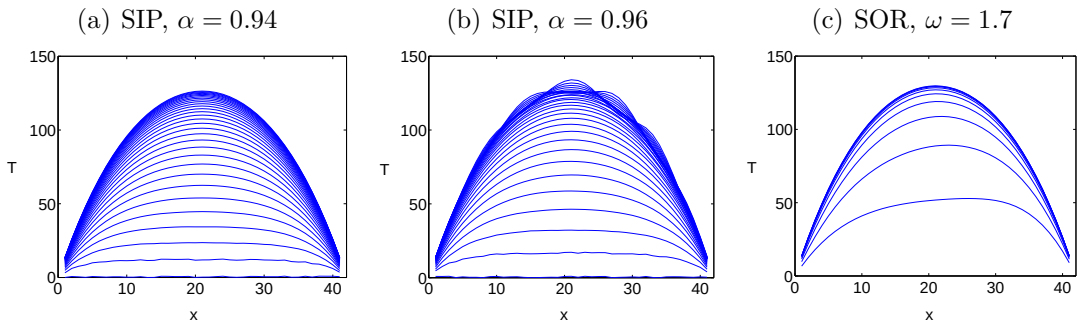


Abbildung 5.9.: Iterative Lösung von $\nabla^2 T = -1$ mittels finiten Differenzen auf 41×41 internen Punkten. Die Iterierte $T(i,j)$ ist in dem Schnitt $y = y_j$ mit $j = 21$ gezeigt. (a) SIP mit $\alpha = 0.94$ für $N = 0, 1, 2, ..., 30$, (b) instabiles SIP-Verfahren für $\alpha = 0.96$ und (c) SOR-Verfahren mit $\omega = 1.7$ für $N = 20, 40, ..., 200$.

- Für gegebenen Wert von α muß die LU-Faktorisierung nur einmal durchgeführt werden. Bei der Iteration werden immer nur dieselbe Matrizen L und U verwendet. Bei jedem Iterationsschritt muß zunächst das Residuum $\boldsymbol{\rho}$ nach (5.30) berechnet werden, dann $\boldsymbol{R}$ und dann die Korrektur $\boldsymbol{\delta}$, woraus sich die neue Approximation $\boldsymbol{x}$ ergibt. All diese Schritte erfordern lediglich die Lösung von tridiagonalen Systemen, was sehr schnell geht (Kap. 5.1.3).

- In der Praxis werden meist nur ein paar wenige Iterationen benötigt. Die Konvergenz ist sehr schnell, wenn α richtig gewählt wird. Für $\alpha = 1$ entspricht (5.52) einer linearen Interpolation in zwei Dimensionen. Der SIP-Algorithmus ist für $\alpha = 1$ instabil (siehe Abb. 5.9b). Deshalb muß man $\alpha < 1$ wählen.

- Die SIP-Methode kann auch auf Matrizen mit weiteren Diagonalen erweitert werden. Schemata höherer Ordnung führen zum Beispiel in zwei Dimensionen auf einen 9-Punkt-Stern, ein dreidimensionales Schema mit zentralen Differenzen führt auf 7 Diagonalen.

- Der SIP-Algorithmus kann auch im Zusammenhang mit *CG-Methoden* (***c**onjugate-**g**radient methods*) als Vorkonditionierer und bei Mehrgitter-Verfahren als Glättungsoperator verwendet werden.

5.2.5. ADI-Methode

Elliptische Probleme, wie zum Beispiel $\nabla^2 T = 0$, können dadurch gelöst werden, daß man die erste Ableitung von T nach einer anderen Variablen τ addiert. Die Variable τ kann man als künstliche Zeitskala interpretieren. Dadurch wird die Gleichung parabolisch. Man integriert dann $\partial T/\partial \tau = \nabla^2 T$ in τ-Richtung bis man einen stationären Zustand ($\partial T/\partial \tau = 0$) erreicht hat. Da der künstliche Term im stationären Zustand verschwindet, hat er keinen Einfluß auf diesen Zustand. Um einen

stabilen Algorithmus mit hinreichend großer Schrittweite zu erhalten, muß ein solches Verfahren (teil-)implizit sein (Kap. 3.3). Das kann aber numerisch sehr teuer sein (zu viele Operationen), da für jeden Zeitschritt ein elliptisches Problem gelöst werden muß. Mit der *ADI-Methode* (***a**lternating **d**irection **i**mplicit method*) kann die Anzahl der erforderlichen Operationen aber sehr stark reduziert werden.

Mit der *künstlichen Zeitskala* τ wandelt sich die Laplace-Gleichung in eine Diffusionsgleichung der Form

$$\frac{\partial T}{\partial \tau} = \kappa \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} \right). \tag{5.59}$$

Wenn wir diese Gleichung teilimplizit mit $\beta = 0.5$ (vgl. (3.19)) aufschreiben,[13] erhalten wir

$$\frac{T^{n+1} - T^n}{\Delta\tau} = \frac{\kappa}{2} \left[\left(\frac{\delta^2 T^n}{\delta x^2} + \frac{\delta^2 T^n}{\delta y^2} \right) + \left(\frac{\delta^2 T^{n+1}}{\delta x^2} + \frac{\delta^2 T^{n+1}}{\delta y^2} \right) \right], \tag{5.60}$$

mit dem Verständnis, daß $T^n = T^n_{i,j}$ und der Notation

$$\frac{\delta^2 T}{\delta x^2} = \frac{T_{i+1,j} - 2T_{i,j} + T_{i-1,j}}{\Delta x^2}, \tag{5.61}$$

sowie entsprechend in y-Richtung. Trennung der beiden Zeitniveaus liefert

$$\left(1 - \frac{\kappa\Delta\tau}{2} \frac{\delta^2}{\delta x^2} - \frac{\kappa\Delta\tau}{2} \frac{\delta^2}{\delta y^2} \right) T^{n+1} = \left(1 + \frac{\kappa\Delta\tau}{2} \frac{\delta^2}{\delta x^2} + \frac{\kappa\Delta\tau}{2} \frac{\delta^2}{\delta y^2} \right) T^n \tag{5.62}$$

Man kann nun die räumlichen Ableitungsoperatoren faktorisieren, indem man die Gleichung in der Form schreibt

$$\begin{aligned} &\left(1 - \frac{\kappa\Delta\tau}{2} \frac{\delta^2}{\delta x^2} \right) \left(1 - \frac{\kappa\Delta\tau}{2} \frac{\delta^2}{\delta y^2} \right) T^{n+1} \\ &= \left(1 + \frac{\kappa\Delta\tau}{2} \frac{\delta^2}{\delta x^2} \right) \left(1 + \frac{\kappa\Delta\tau}{2} \frac{\delta^2}{\delta y^2} \right) T^n + \frac{(\kappa\Delta\tau)^2}{4} \frac{\delta^2}{\delta x^2} \frac{\delta^2}{\delta y^2} \left(T^{n+1} - T^n \right). \end{aligned} \tag{5.63}$$

Hierbei kompensiert der letzte Summand gerade diejenigen Terme, die aus den Produkten der zweiten Ableitungen entstehen, aber nicht in (5.62) enthalten sind.

Wegen $T^{n+1} - T^n \approx \Delta\tau\, \partial T/\partial\tau = O(\Delta\tau)$ ist der letzte Summand von der Größenordnung $O(\Delta\tau^3)$ und kann für die hier verwendeten finiten Differenzen zweiter Ordnung vernachlässigt werden, wenn $\Delta\tau$ hinreichend klein ist. Die verbleibende Gleichung vierter Ordnung ist identisch erfüllt, wenn man einen Zwischenwert T^* in der folgenden Weise einführt:

$$\left(1 - \frac{\kappa\Delta\tau}{2} \frac{\delta^2}{\delta x^2} \right) \underbrace{\left(1 - \frac{\kappa\Delta\tau}{2} \frac{\delta^2}{\delta y^2} \right) T^{n+1}}_{:= \left(1 + \frac{\kappa\Delta\tau}{2} \frac{\delta^2}{\delta x^2} \right) T^*} = \left(1 + \frac{\kappa\Delta\tau}{2} \frac{\delta^2}{\delta x^2} \right) \underbrace{\left(1 + \frac{\kappa\Delta\tau}{2} \frac{\delta^2}{\delta y^2} \right) T^n}_{:= \left(1 - \frac{\kappa\Delta\tau}{2} \frac{\delta^2}{\delta x^2} \right) T^*}. \tag{5.64}$$

[13] Dies entspricht der Trapezregel in der Zeit $(f^n + f^{n+1})\Delta\tau/2$. Bei partiellen Differentialgleichun-

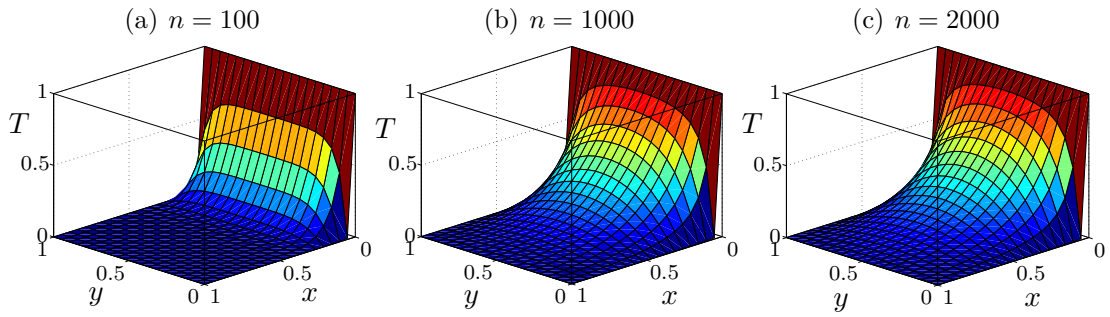


Abbildung 5.10.: Berechnung von $\nabla^2 T = 0$ auf dem Einheitsquadrat $[0,1] \times [0,1]$ zu den Randbedingungen $T(x=0,y) = 1$ und $T(x=1,y) = T(x,y=0) = T(x,y=1) = 0$. Die Anfangsbedingung war $T(x,y) = 0$. Die Parameter sind $\kappa = 1$, $\Delta\tau = 10^{-4}$. Gezeigt ist die Lösung durch ADI auf einem 20×20 Gitter nach n Iterationen (Zeitschritten) wie in den Überschriften angezeigt.

Die beiden Definitionen führen zu den beiden Gleichungen zweiter Ordnung

$$\left(1 - \frac{\kappa\Delta\tau}{2}\frac{\delta^2}{\delta x^2}\right) T^* = \left(1 + \frac{\kappa\Delta\tau}{2}\frac{\delta^2}{\delta y^2}\right) T^n, \tag{5.65a}$$

$$\left(1 - \frac{\kappa\Delta\tau}{2}\frac{\delta^2}{\delta y^2}\right) T^{n+1} = \left(1 + \frac{\kappa\Delta\tau}{2}\frac{\delta^2}{\delta x^2}\right) T^*. \tag{5.65b}$$

In beiden Gleichungen treten lediglich tridiagonale Matrizen auf. Die Gleichungen können daher schnell mit dem Thomas-Algorithmus ohne irgendeine Iteration gelöst werden ($T^n \to T^* \to T^{n+1}$). Man kann zeigen, daß dieser Algorithmus uneingeschränkt stabil und von zweiter Ordnung Genauigkeit $O(\Delta x^2, \Delta y^2)$ ist. Trotz der uneingeschränkten Stabilität gibt es eine optimale Zeitschrittweite $\Delta\tau$. Denn die Schrittweite darf nicht zu groß werden, damit die Fehler durch die vernachlässigten Terme der Ordnung $O(\Delta\tau^3)$ hinreichend klein bleiben. Ist andernfalls die Schrittweite zu klein, wird die Konvergenz verlangsamt. In Abb. 5.10 ist ein Beispiel gezeigt.[14]

Im ersten Halbschritt (5.65a) ist die rechte Seite (T^n) aus dem vorherigen Zeitschritt bekannt. Es wird lediglich in x-Richtung integriert, und zwar implizit. Die Zeitschrittweite ist $\Delta\tau/2$. Im zweiten Halbschritt (5.65b) wird dann in y-Richtung integriert, wobei die zweite Ableitung in x-Richtung aus dem ersten Halbschritt (T^*) verwendet wird. Daher stammt die Bezeichnung ADI. Beim ADI-Verfahren wird das zweidimensionale Problem auf die Lösung eindimensionaler Probleme zu-

gen wird dies auch Crank-Nicolson-Schema genannt.

[14] Die Strömung $\boldsymbol{u} = w(x,y)\boldsymbol{e}_z$ in einem unendlich langen Kanal mit quadratischem Querschnitt, die durch eine tangential Bewegung einer Wand mit konstanter Geschwindigkeit in z-Richtung angetrieben wird, genügt auch einer Diffusionsgleichung für $w(x,y)$ (viskose Impulsdiffusion) mit Randbedingungen wie im Beispiel Abb. 5.10.

rückgeführt.[15]

5.2.6. Mehrgitterverfahren

Die Konvergenz von Iterationsverfahren hängt ab vom *spektralen Radius* $|\lambda_1|$ (betragsmäßig größter Eigenwert) der *Iterationsmatrix* $\mathsf{M}^{-1}\cdot\mathsf{N}$ (Kap. 5.2.2). Die Struktur des Fehlers $\boldsymbol{\epsilon}$ wird nach hinreichend vielen Iterationen bestimmt durch den zugehörigen Eigenvektor $\boldsymbol{\psi}_1$ von $\mathsf{M}^{-1}\cdot\mathsf{N}$. Bei einigen Verfahren (Gauß-Seidel, SIP) ist der Fehler $\boldsymbol{\epsilon}$ schon nach wenigen Iterationen eine glatte Funktion des Orts (siehe Abb. 5.5).[16] In diesem Fall ist auch das Residuum glatt. Damit ist es möglich, den glatten Fehler sehr effizient weiter zu reduzieren, indem man die *Korrektur-Gleichung* (5.33) $\mathsf{M}\cdot\boldsymbol{\delta}^{(n)} = \boldsymbol{\rho}^{(n)}$ kostengünstig auf einem relativ *groben Gitter* löst. Der Vorteil eines groben Gitters besteht in einer drastischen Reduktion der Anzahl der Unbekannten. Bei einer Verdopplung der Gitterweite hat man in 3 Dimensionen schon eine Reduktion der Anzahl der Variablen um einen Faktor $(1/2)^3$. Darüber hinaus konvergieren iterative Verfahren auf gröberen Gittern schneller. Bei Verdopplung des Gitterabstands macht dies bei der Gauß-Seidel-Iteration einen weiteren Faktor von $1/4$ aus.

Die Strategie bei *Mehrgitterverfahren* zielt daher darauf, den Fehler auf einem feinen Gitter zunächst zu glätten und ihn dann auf einem groben Gitter sehr effizient zu reduzieren, um schließlich wieder zu dem feinen Gitter zurückzukehren. Dazu müssen die Gitter und die Operatoren, die zwischen den Gittern vermitteln, geeignet definiert werden. Speziell muß das Residuum $\boldsymbol{\rho}$ von dem feinen auf das grobe Gitter übertragen werden (*Restriktion*) und umgekehrt muß die Korrektur der Lösung $\boldsymbol{\delta}$ vom groben auf das feine Gitter gebracht werden (*Prolongation*). Dabei gibt es viele Möglichkeiten, die zugehörigen Operatoren zu definieren. Typischerweise wird bei finiten Differenzen bei einer Vergröberung des Gitters nur jeder zweite Gitterpunkt verwendet. Bei finiten Volumen wird das Kontrollvolumen in jeder Dimension verdoppelt (Abb. 5.11).

[15] Auf dieser Idee beruht die ganze Klasse der *Splitting-Methoden*. In der Strömungsmechanik wird speziell die Druck-Korrektur-Gleichung (Poisson-Gleichung) oft mit der ADI-Methode gelöst. Bei der Implementierung der Randbedingungen muß man aufpassen, damit die Genauigkeit zweiter Ordnung erhalten bleibt; siehe Kap. 6.2.1.

[16] Bei ADI und SOR hat der Fehler eine relativ komplizierte Struktur.

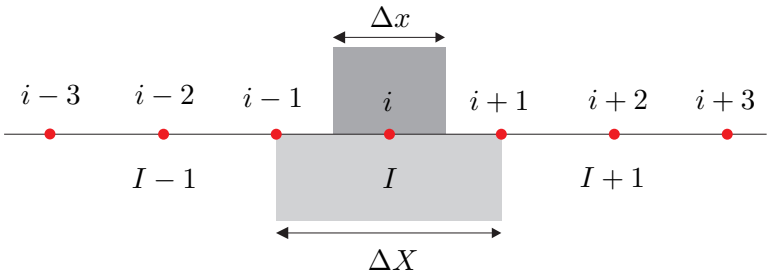


Abbildung 5.11.: Feine und grobe Gitterpunkte für finite Volumen in einer Dimension.

Im folgenden soll ein eindimensionales Beispiel gegeben werden.[17] Betrachte

$$\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2} = f(x). \tag{5.66}$$

Die Approximation durch finite Differenzen lautet

$$\frac{\phi_{i+1} - 2\phi_i + \phi_{i-1}}{\Delta x^2} = f_i. \tag{5.67}$$

Nach N Iterationen auf dem feinen Gitter mit Abstand Δx erhält man dann die Näherungslösung $\phi^{(n)}$ und das Residuum $\rho^{(n)}$. Für sie gilt

$$\frac{\phi_{i+1}^{(n)} - 2\phi_i^{(n)} + \phi_{i-1}^{(n)}}{\Delta x^2} = f_i - \rho_i^{(n)}. \tag{5.68}$$

Wenn man dies von der Gleichung für die exakte Lösung (5.67) subtrahiert, erhält man für den Fehler nach der n-ten Iteration $\epsilon_i^{(n)} = \phi_i - \phi_i^{(n)}$ die Gleichung

$$\frac{\epsilon_{i+1}^{(n)} - 2\epsilon_i^{(n)} + \epsilon_{i-1}^{(n)}}{\Delta x^2} = \rho_i^{(n)}. \tag{5.69}$$

Diese Gleichung soll nun auf dem groben Gitter mit $\Delta X = 2\Delta x$ weiter iteriert werden. Auf dem groben Gitter gilt formal (siehe Abb. 5.11)

$$\frac{\epsilon_{I+1} - 2\epsilon_I + \epsilon_{I-1}}{\Delta X^2} = \rho_I, \tag{5.70}$$

wobei I die Punkte des groben Gitters numeriert. Um ϵ_I und ρ_I durch die entsprechenden Größen des feinen Gitters ϵ_i und ρ_i auszudrücken, kann man zur Gleichung (5.69) für das feine Gitter jeweils die Hälfte der Gleichungen für die beiden benachbarten feinen Gitterpunkte addieren (Abb. 5.11). Wenn wir den Iterationsindex n unterdrücken, führt dies auf

$$\frac{\epsilon_{i+1} - 2\epsilon_i + \epsilon_{i-1}}{\Delta x^2} + \frac{1}{2}\left(\frac{\epsilon_{i+2} - 2\epsilon_{i+1} + \epsilon_i}{\Delta x^2} + \frac{\epsilon_i - 2\epsilon_{i-1} + \epsilon_{i-2}}{\Delta x^2}\right) = \rho_i + \frac{\rho_{i+1} + \rho_{i-1}}{2}, \tag{5.71}$$

wobei sich die roten und die grünen Terme kompensieren. Nach Division durch 2 erhalten wir

$$\underbrace{\frac{1}{(2\Delta x)^2}}_{\Delta X^{-2}} \underbrace{(\epsilon_{i+2} - 2\epsilon_i + \epsilon_{i-2})}_{\epsilon_{I+1} - 2\epsilon_I + \epsilon_{I-1}} = \underbrace{\frac{1}{4}(\rho_{i+1} + 2\rho_i + \rho_{i-1})}_{\rho_I}. \tag{5.72}$$

Hieran kann man sofort die Beziehung zwischen den Fehlern und den Residuen auf dem groben und auf dem feinen Gitter durch Vergleich mit (5.70) ablesen. Wir

[17] Die eindimensionale Poisson-Gleichung kann man am effizientesten mit TDMA lösen, daher

erhalten also konsistent dasselbe Differenzenschema. Insbesondere stellt

$$\rho_I = \frac{\rho_{i+1} + 2\rho_i + \rho_{i-1}}{4} \tag{5.73}$$

eine *Glättung* bzw. *Filterung* des Residuums ρ_i des dem feinen Gitters dar, und definiert damit den *Glättungsoperator* (*restriction operator*). Mit Kenntnis des Residuums (5.73) auf dem groben Gitter kann man nun die Korrekturgleichung (5.33) für das Inkrement δ_I auf dem groben Gitter iterieren. Um die auf dem groben Gitter nach einigen Iterationen erhaltene Lösung ϕ_I wieder auf das feine Gitter zu übertragen (*prologation*) ist es am einfachsten, eine lineare Interpolation zu verwenden, wobei die Werte an kongruenten Punkten direkt übernommen werden. Ein 2-Gitter-Verfahren würde dann folgendermaßen ablaufen:

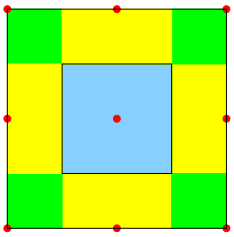
1. Iteriere auf dem feinen Gitter mit einer Methode, die eine glatte Fehlerfunktion liefert,
2. berechne das Residuum ρ_i auf dem feinen Gitter,
3. führe eine Restriktion des Residuums auf das grobe Gitter durch $(\rho_i \to \rho_I)$,
4. iteriere die Korrekturgleichung (5.33) $M_{IJ}\delta_J = \rho_I$ auf dem groben Gitter,
5. interpoliere die Korrektur δ_I der Lösung vom groben auf das feine Gitter $(\delta_I \to \delta_i)$,
6. berechne die korrigierte Lösung $\phi_i + \delta_i$ auf dem feinen Gitter,
7. wiederhole die Prozedur bis zur gewünschten Genauigkeit.

Dieses 2-Gitter-Schema wird zweckmäßigerweise sukzessive auf immer gröbere Gitter erweitert. Auf dem gröbsten Gitter kann das Residuum in wenigen Schritten bei minimaler Anzahl von Operationen exakt zu Null gemacht werden. Durch die Korrektur der langwelligen Fehler auf den groben Gittern (billig) wird auch die Anzahl der erforderlichen Operationen auf dem feinsten Gitter (teuer) stark vermindert.[18] Die für eine gegebene Genauigkeit erforderliche Anzahl von Iterationen auf dem feinsten Gitter ist bei Mehrgitterverfahren proportional zur Anzahl

macht MG im Eindimensionalen in der Praxis wenig Sinn. Der eindimensionale Fall dient hier nur zur Demonstration.

[18] An Abb. 5.5 sieht man, daß die Gauß-Seidel-Iteration schon nach wenigen (< 10) Iterationen die kurzwelligen Fluktuationen eliminiert hat. Andererseits dauert es mindestens 1000 Iterationen bis man auf dem gewählten (feinen) Gitter eine gute Approximation der Lösung erhalten hat.

N der Unbekannten. Dies ist ein signifikanter Vorteil gegenüber den Standard-Iterationsverfahren. Die Beschleunigung macht sich besonders bei dreidimensionalen Rechnungen mit hoher Auflösung bezahlt. Auch ist es oft sinnvoll, mit der Iteration auf dem gröbsten Gitter zu beginnen, damit man für die Fein-Gitter-Iteration schon eine gute Näherung hat. Dies nennt man *FMG* (***f****ull* ***m****ulti* ***g****rid*).



Mehrgitter ist eher eine Strategie als ein eigenes Verfahren. Voraussetzung ist eine Iterationsmatrix, die gute Glättungseigenschaften hat. Außer GS und SIP gibt es noch andere Möglichkeiten.

In zwei Dimensionen gibt es viele Möglichkeiten der Restriktion auf das grobe Gitter. Die obige Methode der anteilmäßigen Addition der Gleichungen für die Nachbarpunkte führt auf ein 9-Punkt-Schema für $\rho_{I,J}$. Wenn die Gewichtung der Punkte entsprechend der Fläche erfolgt, denen sie zugeordnet werden können (1/4,1/8,1/16), erhält man

$$\begin{aligned}\rho_{I,J} = \frac{1}{4}\rho_{i,j} + \frac{1}{8}\left(\rho_{i,j+1} + \rho_{i,j-1} + \rho_{i+1,j} + \rho_{i-1,j}\right) \\ + \frac{1}{16}\left(\rho_{i+1,j+1} + \rho_{i+1,j-1} + \rho_{i-1,j+1} + \rho_{i-1,j-1}\right).\end{aligned} \tag{5.74}$$

Jedoch liefert auch das vereinfachte Schema

$$\rho_{I,J} = \frac{1}{8}\left(4\rho_{i,j} + \rho_{i+1,j} + \rho_{i-1,j} + \rho_{i,j+1} + \rho_{i,j-1}\right) \tag{5.75}$$

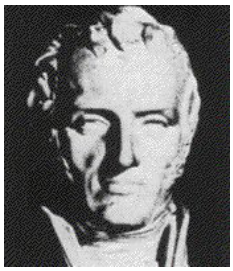
sehr gute Ergebnisse. Für die Prolongation in zwei Dimensionen kann man eine bilineare Interpolation verwenden. Für Details der Mehrgitter-Verfahren sei auf Hackbusch (1985) verwiesen.

6. Zeitliche Diskretisierung: Konvektions-Diffusionsgleichungen

In der Strömungsmechanik realer Fluide hat man es fast immer mit Gleichungen des Typs

$$\underbrace{\frac{\partial T}{\partial t}}_{\text{Änderungsrate}} + \underbrace{\boldsymbol{u}\cdot\nabla T}_{\text{Konvektion}} = \underbrace{\kappa\nabla^2 T}_{\text{Diffusion}} + \underbrace{F}_{\text{Quellen}} \tag{6.1}$$

zu tun.[1] Wenn wir zum Beispiel T als Temperatur auffassen, beschreibt die Gleichung die Entwicklung der Temperatur an allen Punkten $\boldsymbol{x}$ des betrachteten Gebietes und für alle Zeiten $t > t_0$ nach der Präparation eines Anfangszustandes $T_0(\boldsymbol{x}) = T(\boldsymbol{x}, t_0)$. Dann ist (6.1) eine Wärmetransportgleichung. Die zeitliche Änderung der Temperatur an jedem festen Ort $\boldsymbol{x}$ zur Zeit t ist gegeben durch $\partial T/\partial t$. Der Term $\boldsymbol{u}\cdot\nabla T$, in einer Dimension $u\partial T/\partial x$, beschreibt die negative Änderungsrate der Temperatur durch *Konvektion* mit der Geschwindigkeit $\boldsymbol{u}$ (Abb. 6.1a). Die Rate der Änderung durch *Diffusion* $\kappa\nabla^2 T$ (Abb. 6.1b) hatten wir ja schon kennengelernt. Die Größe $F(\boldsymbol{x}, t)$ beschreibt die Quellen bzw. Senken des Feldes $T(\boldsymbol{x}, t)$. Dies können zum Beispiel Wärmequellen aufgrund chemischer Reaktionen (Verbrennung) sein oder Verluste aufgrund von Strahlung.



Claude Louis Marie Henri Navier 1785–1836

Die Konvektion von T wird vom Geschwindigkeitsfeld $\boldsymbol{u}$ bewirkt, das seinerseits durch die Navier-Stokes-Gleichung (hier für ein Newtonsches Fluid)

$$\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{u} = -\frac{1}{\rho}\nabla p + \nu\nabla^2\boldsymbol{u} + \boldsymbol{F}(T) \tag{6.2}$$

bestimmt ist. Die Navier-Stokes-Gleichung ist im wesentlichen auch eine Konvektions-Diffusionsgleichung.[2] Nur wird hier die vektorielle Größe $\boldsymbol{u}$ (Impuls pro Masse) transportiert.[3] Der konvektive Transportterm $\boldsymbol{u}\cdot\nabla\boldsymbol{u}$ stellt einen nichtlinearen Term dar, der das Verhalten der Strömung ganz entscheidend beeinflussen

[1] Eigentlich treten noch weitere Terme auf, die aber sehr oft vernachlässigt werden können.

[2] In der Navier-Stokes-Gleichung tritt jedoch zusätzlich noch der Druck auf, der einer besonderen Behandlung bedarf.

[3] Bei einem inkompressiblen Fluid ist $\rho = \text{const.}$, so daß der Geschwindigkeitsvektor proportional

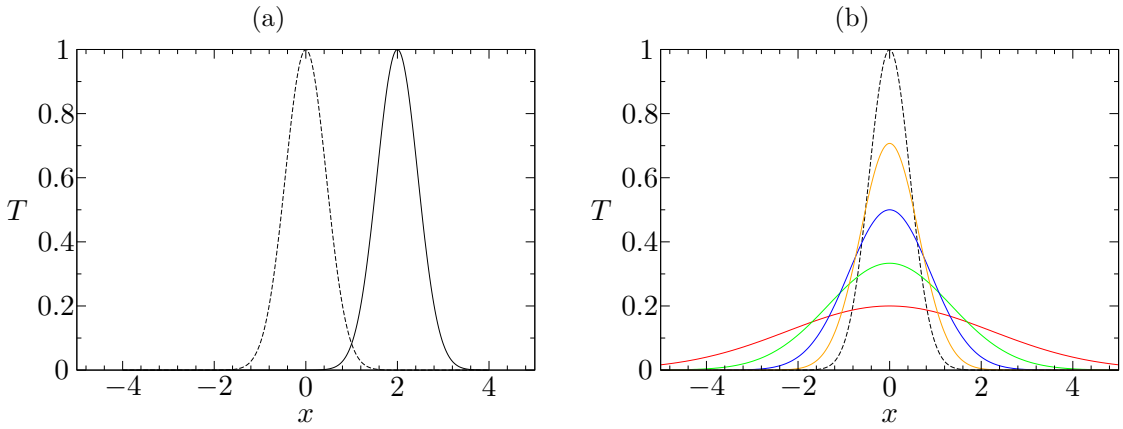
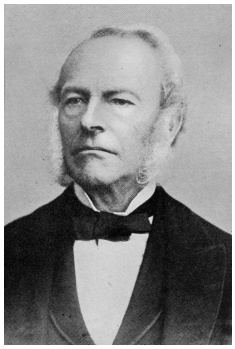


Abbildung 6.1.: Illustration der zeitlichen Entwicklung einer Größe T durch Konvektion (a) oder durch Diffusion (b). Hier wurde als Ausgangsfunktion das Gaußpaket $T(x,t_0) = \exp(-2.5\,x^2)$ gewählt. Als Besonderheit bleibt das Gauß-Profil bei der Diffusion (b) erhalten, nur die Skalen ändern sich. Denn $T(x,t) = (4\pi\kappa t)^{-1/2}\exp\{-x^2/4\kappa t\}$ ist eine exakte Lösung der Diffusionsgleichung für eine konzentrierte Wärmequelle $T(x,0) = \delta(x)$ bei $x = 0$ zum Zeitpunkt $t = 0$ (Landau and Lifschitz, 1991). Die Lösungen sind dargestellt für $t = t_0$ (gestrichelt) und $t = 2/u$ für reine Konvektion (a) sowie für $t = 2t_0$ (orange), $t = 4t_0$ (blau), $t = 9t_0$ (grün) und $t = 25t_0$ (rot) für reine Diffusion (b).

kann. Physikalisch bedeutet der gesamte Term $\mathrm{d}\boldsymbol{u}/\mathrm{d}t = \partial\boldsymbol{u}/\partial t + \boldsymbol{u}\cdot\nabla\boldsymbol{u}$ die Trägheitskraft pro Masse. Aber auch $\boldsymbol{u}\cdot\nabla T$ in der Wärmetransportgleichung (6.1) ist ein nichtlinearer Term, wenn $\boldsymbol{u}$ in irgendeiner Weise von T abhängt (gekoppelte Gleichungen). Falls dies nicht der Fall ist, wenn also $\boldsymbol{u}$ fest vorgegeben ist, dann ist $\boldsymbol{u}\cdot\nabla T$ ein linearer Term.

Bevor wir die Lösung der gekoppelten Navier-Stokes und Wärmetransportgleichung unter Einbeziehung aller Terme betrachten, ist es zweckmäßig, verschiedene zeitliche Diskretisierungen an einfachen Modellproblemen zu testen und die Eigenschaften (Genauigkeit) der Verfahren durch einen Vergleich mit exakten Lösungen zu untersuchen. Für die Diffusionsgleichung hatten wird dies ja teilweise schon in Kap. 2 durchgeführt.

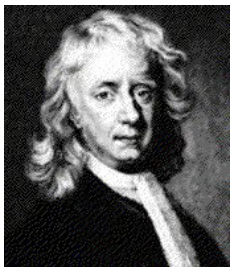

Sir George Gabriel Stokes
1819–1903

Für die Analyse von Zeitintegrationsschemata ist es am einfachsten, eindimensionale Modellgleichungen zu verwenden, da sie das prinzipielle Verhalten in den meisten Fällen sehr gut widerspiegeln. Zunächst werden wir Verfahren für die *parabolische* Gleichung (*Diffusionsgleichung* (1.50))

$$\frac{\partial T}{\partial t} - \kappa\frac{\partial^2 T}{\partial x^2} = 0 \tag{6.3}$$

betrachten. Der zugrundeliegende physikalische Prozeß ist die Realisierung eines thermodynamischen Gleichgewichts durch einen Entropiestrom von Gebieten mit

hoher Temperatur zu Gebieten mit niedriger Temperatur. Im Mikroskopischen geschieht dies durch Austausch der kinetischen Energie zwischen den Molekülen. Im Fall der Konzentration c einer Spezies ($T \to c$) wird das thermodynamische Gleichgewicht durch die Vermischung unterschiedlicher Moleküle aufgrund des chemischen Potentials und der thermischen Bewegung realisiert. Einfach gesagt, wirkt die Diffusion glättend. Je schärfer die Spitzen, d.h. je größer $|\partial^2 T/\partial^2 x|$, desto schneller werden sie abgebaut.


Sir Isaac Newton
1642–1727

Danach werden wir die *Advektionsgleichung*[4]

$$\frac{\partial T}{\partial t} + u\frac{\partial T}{\partial x} = 0 \tag{6.4}$$

betrachten, wobei $u > 0$ konstant vorgegeben sei. Sie ist vom *hyperbolischen* Typ. Bei gegebener Anfangsbedingung $T_0(x, t_0)$ kann man ihre Lösung sofort angeben. Sie lautet $T(x, t) = T_0(x - ut, t_0)$. Die Anfangsverteilung wird lediglich in positiver x-Richtung verschoben. Es handelt sich um eine reine Translation, bei der keine Änderung der Form des Profils $T(x, *)$ stattfindet.

Schließlich werden wir uns auch mit dem kombinierten Effekt im Rahmen der Konvektions-Diffusionsgleichung (Kap. 6.4) und mit Nichtlinearitäten beschäftigen (Kap. 6.5) bevor wir Lösungsmethoden für die Navier-Stokes-Gleichung (6.2) betrachten.

6.1. Diffusion

Zunächst betrachten wir die eindimensionale Diffusionsgleichung (1.50). Zur Lösung kommen explizite oder implizite Verfahren in Betracht.

6.1.1. Explizite Verfahren

FTCS-Algorithmus

Das einfachste explizite Verfahren für die Diffusionsgleichung ist das FTCS-Schema (siehe Kap. 2.1). Es verwendet räumlich zentrale Differenzen und zeitlich einseitige Vorwärtsdifferenzen

$$\frac{T_j^{n+1} - T_j^n}{\Delta t} = \kappa \frac{T_{j+1}^n - 2T_j^n + T_{j-1}^n}{\Delta x^2}. \tag{6.5}$$

Aufgelöst nach der unbekannten Größe T_j^{n+1} erhalten wir den *FTCS-Algorithmus* (2.23)

$$T_j^{n+1} = sT_{j+1}^n + (1 - 2s)T_j^n + sT_{j-1}^n, \tag{6.6}$$

ist zum Vektor der Impulsdichte: $\boldsymbol{u} \sim \rho\boldsymbol{u}$.

[4]Man kann diese Gleichung auch Konvektionsgleichung nennen. Nur in der Meteorologie wird

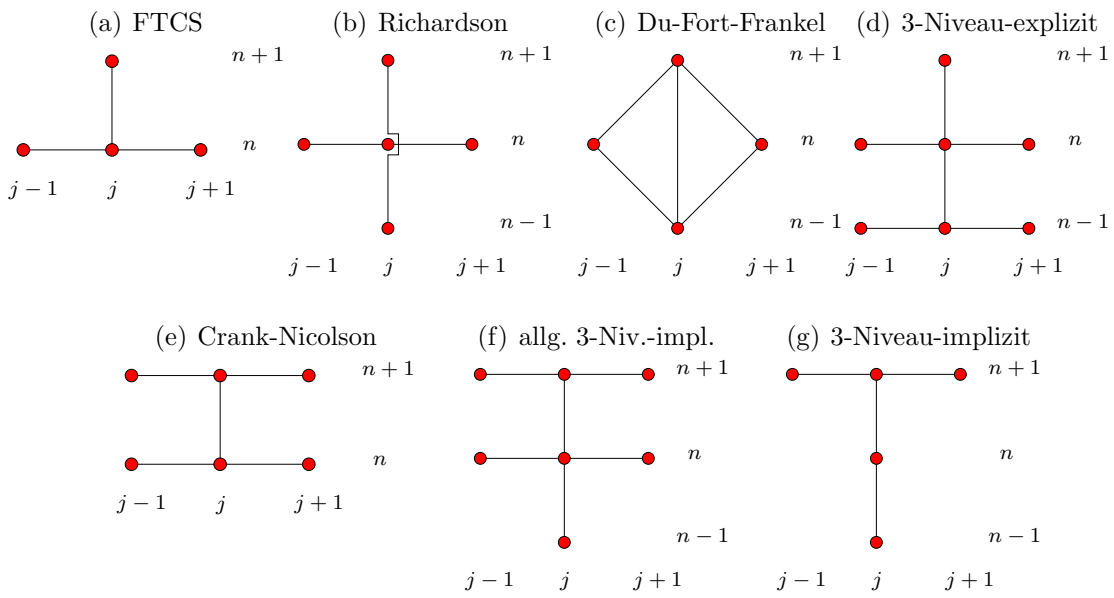


Abbildung 6.2.: Differenzensterne verschiedener expliziter (a–d) und impliziter (e–g) Integrationsschemata zur Lösung der Diffusionsgleichung.

wobei $s = \kappa\Delta t/\Delta x^2$ ist. Die involvierten Gitterpunkte sind in Abb. 6.2a dargestellt. In Kap. 3.2 hatten wir gesehen, daß das Verfahren konsistent, von erster Ordnung in der Zeit $O(\Delta t)$ und von zweiter Ordnung $O(\Delta x^2)$ im Raum ist. Für $s = 1/6$ ist der Fehler sogar nur von der Größenordnung $O(\Delta t^2, \Delta x^4)$. Die *von-Neumann-Stabilitätsanalyse* lieferte für die Diffusionsgleichung und kleine Störungen den Verstärkungsfaktor[5]

$$G = 1 - 4s \sin^2\left(\frac{\theta}{2}\right). \tag{6.7}$$

Für Stabilität muß $|G| \leq 1$ sein, was auf die Bedingung $s \leq 1/2$ führt.

Heuristisches Stabilitätskriterium Einen Hinweis auf die *Stabilität* liefern auch die Koeffizienten von T^n_{j-1}, T^n_j und T^n_{j+1} auf der rechten Seite von (2.23): Aus physikalischen Gründen sollte eine Störung, welche nur die Temperatur von T^n_{j-1}, T^n_j oder T^n_{j+1} erhöht, auf jeden Fall auch zu einer Erhöhung von T^{n+1}_j führen. Falls einer der Koeffizienten von T^n_{j-1}, T^n_j oder T^n_{j+1} in (2.23) negativ ist, kann diese positive Störung jedoch unter Umständen eine Erniedrigung von T^{n+1}_j bewirken, was keinen Sinn machen würde. Daher ist ein negativer Koeffizient in (2.23) ein

zwischen Konvektion und Advektion unterschieden. Konvektion bezeichnet dort das *vertikale* Aufsteigen von Warmluft. Advektion bezeichnet hingegen das *horizontale* Aufgleiten von Warmluft über kältere Luftschichten.

[5] Zur Erinnerung: Bei der von Neumann-Stabilitätsanalyse setzt man für die Abweichungen ξ^n_j von der exakten diskreten Lösung auf dem Bereich $[0, 1]$ den Ansatz $\xi^n_j \sim (G)^n \, \mathrm{e}^{\mathrm{i}\theta j}$ mit $\theta = m\pi\Delta x$ (falls $x \in [0, 1]$) in die lineare Differentialgleichung ein, um den komplexen Verstärkungsfaktor G für einen Schritt der Zeitintegration zu erhalten (siehe Kap. 3.3.2). Dabei ist $m\pi$

Warnzeichen für eine mögliche Instabilität des Verfahrens. In der Tat verschwindet der Koeffizient von T_j^n genau auf der von-Neumann-Stabilitätsgrenze.

Richardson- und DuFort-Frankel-Schema

Beim FTCS-Schema ist die zeitliche Diskretisierung mit einseitigen Vorwärts-Differenzen nur von erster Ordnung. Symmetrische Differenzen für die erste zeitliche Ableitung würden von zweiter Ordnung sein. Daher wäre das *Richardson-Schema* für die Diffusionsgleichung naheliegend (Abb. 6.2b)

$$\frac{T_j^{n+1} - T_j^{n-1}}{2\Delta t} = \kappa \frac{T_{j+1}^n - 2T_j^n + T_{j-1}^n}{\Delta x^2}. \tag{6.8}$$

Mittels Neumann-Stabilitätsanalyse kann man aber zeigen, daß dieses Schema für alle $s > 0$ instabil ist. Das bedeutet aber nicht, daß das Richardson-Schema auch für andere Differentialgleichungen instabil sein muß.[6]

Durch eine leichte Modifikation kann man das Richardson-Verfahren stabilisieren. Dazu wird T_j^n durch den zeitlichen Mittelwert $\left(T_j^{n+1} + T_j^{n-1}\right)/2$ ersetzt. Damit erhalten wir das *DuFort-Frankel-Schema* (Abb. 6.2c, Du Fort and Frankel, 1953)

$$\frac{T_j^{n+1} - T_j^{n-1}}{2\Delta t} = \kappa \frac{T_{j+1}^n - \left(T_j^{n+1} + T_j^{n-1}\right) + T_{j-1}^n}{\Delta x^2} \tag{6.9}$$

oder, aufgelöst nach T_j^{n+1},

$$T_j^{n+1} = \frac{2s}{1+2s}\left(T_{j+1}^n + T_{j-1}^n\right) + \frac{1-2s}{1+2s} T_j^{n-1}. \tag{6.10}$$

Das DuFort-Frankel-Schema umfaßt 3 Zeitniveaus. Nur für $s = 0.5$ entfällt das Niveau bei $n-1$. Dann ist das DuFort-Frankel-Schema identisch mit dem FTCS-Schema. Bei Verwendung eines 3-Niveau-Schemas ist es erforderlich, jeweils zwei vorhergehenden Niveaus im Speicher zu halten. Außerdem muß man zu Beginn der Rechnung zuerst ein zweites Niveau mittels eines 2-Niveau-Schemas berechnen.

Die von Neumann-Stabilitätsanalyse liefert Wachstumsfaktoren $|G| \leq 1$ für alle θ und $s > 0$. Das DuFort-Frankel-Schema ist damit uneingeschränkt stabil. Eine Prüfung der Konsistenz durch Taylorentwicklung aller auftretenden Terme $\bar{T}_j^n$ um den zentralen Punkt (n, j) nach Kap. 3.2 liefert für das DuFort-Frankel-Schema

$$\underbrace{\left(\frac{\partial \bar{T}}{\partial t}\right)_j^n - \kappa \left(\frac{\partial^2 \bar{T}}{\partial x^2}\right)_j^n}_{\text{entspr. Differentialgleichung}} + \underbrace{\kappa \left(\frac{\Delta t}{\Delta x}\right)^2 \left(\frac{\partial^2 \bar{T}}{\partial t^2}\right)_j^n + O\left(\Delta t^2, \Delta x^2\right)}_{\text{Fehler } E_j^n} = 0. \tag{6.11}$$

die Wellenzahl und $j\Delta x$ die Ortskoordinate.

[6] Für die Advektionsgleichung kann man zum Beispiel ein stabiles Verfahren erhalten; siehe Kap. 6.3.3.

Damit der Fehler $\sim \kappa\Delta t^2/\Delta x^2 = s\Delta t$ nicht zu groß wird, muß trotz uneingeschränkter Stabilität man die Zeitschrittweite Δt bzw. s begrenzen. Wenn wir $\partial^2\bar{T}/\partial t^2$ mit Hilfe der Diffusionsgleichung wie in Kap. 3.2 allein durch die Ortsableitung ausdrücken, finden wir den führenden Fehler

$$s\Delta t \left(\frac{\partial^2 \bar{T}}{\partial t^2}\right)_j^n + O\left(\Delta t^2, \Delta x^2\right) = \kappa \Delta x^2 \left(s^2 - \frac{1}{12}\right) \underbrace{\left(\frac{\partial^4 \bar{T}}{\partial x^4}\right)_j^n}_{\kappa^{-2}\left(\partial^2\bar{T}/\partial t^2\right)} + O\left(\Delta t^2, \Delta x^4\right). \tag{6.12}$$

Für den speziellen Wert $s = 1/\sqrt{12} \approx 0.2887$ haben wir damit ein Verfahren 4. Ordnung im Raum und 2. Ordnung in der Zeit.

Allgemeines explizites 3-Niveau-Schema

Das allgemeine *explizite 3-Niveau-Schema* (Abb. 6.2d) für die Diffusionsgleichung kann man schreiben als

$$aT_j^{n+1} + bT_j^n + cT_j^{n-1} = dL_{xx}T_j^n + eL_{xx}T_j^{n-1}, \tag{6.13}$$

wobei L_{xx} der räumliche *Ableitungsoperator* zweiter Ordnung mittels zentraler Differenzen ist, d.h.

$$L_{xx}T_j := \frac{T_{j+1} - 2T_j + T_{j-1}}{\Delta x^2}. \tag{6.14}$$

Zur Bestimmung der Koeffizienten a, b, c, d und e muß man alle diskreten Terme T_i^m als Taylor-Entwicklung um den zentralen Punkt (n, j) schreiben und verlangen, daß die resultierende Gleichung konsistent mit der Diffusionsgleichung ist (Kap. 2.2) und daß möglichst viele Fehlerterme verschwinden. Wenn man so verfährt, kann man zeigen, daß sich die Anzahl der unbekannten Koeffizienten auf 2 reduziert. Man erhält dann die Form

$$(1+\gamma)\frac{T_j^{n+1} - T_j^n}{\Delta t} - \gamma\frac{T_j^n - T_j^{n-1}}{\Delta t} = \kappa\left[(1-\beta)L_{xx}T_j^n + \beta L_{xx}T_j^{n-1}\right] \tag{6.15}$$

oder, nach T_j^{n+1} aufgelöst,

$$T_j^{n+1} = \left(\frac{1+2\gamma}{1+\gamma}\right)T_j^n - \left(\frac{\gamma}{1+\gamma}\right)T_j^{n-1} + \left(\frac{s\Delta x^2}{1+\gamma}\right)\left[(1-\beta)L_{xx}T_j^n + \beta L_{xx}T_j^{n-1}\right]. \tag{6.16}$$

Der *Abbruchfehler* beträgt

$$E_j^n = \kappa s \Delta x^2 \frac{\partial^4 \bar{T}}{\partial x^4}\left(\frac{1}{2} + \beta + \gamma - \frac{1}{12s}\right) + O\left(\Delta t^2, \Delta x^4\right). \tag{6.17}$$

Damit hat man für

$$\beta = -\frac{1}{2} - \gamma + \frac{1}{12s} \tag{6.18}$$

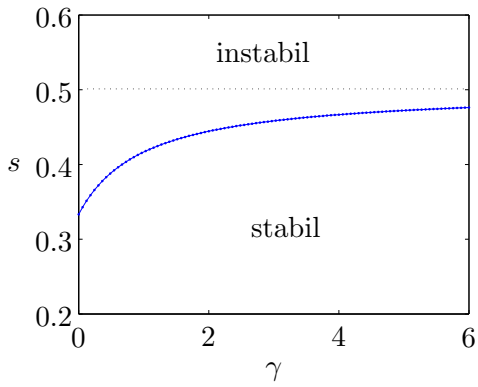


Abbildung 6.3.: Stabilitätsbereich (unterhalb der blauen Kurve) für das explizite 3-Niveau-Schema (6.15) berechnet durch Lösen von (6.19) unter der Bedingung (6.18). Die gepunktete Linie deutet die Stabilitätsgrenze für $\gamma \to \infty$ an.

ein Verfahren 4. Ordnung. Eine Stabilitätsanalyse nach von Neumann liefert eine quadratische Gleichung für den Verstärkungsfaktor $G(\beta, \gamma, s, \theta)$

$$(1+\gamma)G^2 - [1 + 2\gamma + 2s\,(1-\beta)\,(\cos\theta - 1)]\,G + [\gamma - 2\beta s\,(\cos\theta - 1)] = 0. \quad (6.19)$$

Für β nach (6.18) erhält man für $\gamma \to 0$ einen stabilen Algorithmus, wenn $s \leq 1/3$ (Abb. 6.3). Für $\gamma \to \infty$ erhalten wir einen stabilen Algorithmus für $s \leq 0.5$, was der Stabilitätsgrenze des FTCS-Schema entspricht. Tatsächlich sieht man leicht, daß für $\gamma \to \infty$ (6.15) die Summe zweier aufeinanderfolgender FTCS-Schritte ist (wegen (6.18) geht in diesem Limes $\beta \to -\infty$). Das explizite 3-Niveau-Schema mit (6.18) ist genauer als das FTCS-Schema, besitzt aber einen etwas geringeren Stabilitätsbereich.

Man kann nun die obigen expliziten Verfahren implementieren und für spezielle Fälle mit einer exakten Lösung der Diffusionsgleichung vergleichen (Hausaufgabe). Die Effizienz der Verfahren hängt von dem Schrittweitenparameter s ab. Für $s = 0.3$ findet Fletcher (1991*a*), daß das DuFort-Frankel-Verfahren am genauesten ist, was daran liegt, daß $s = 0.3$ nahe an dem Wert $s = 1/\sqrt{12} = 0.2887$ liegt, an dem das DuFort-Frankel-Verfahren von 4. Ordnung ist. Bei dem größerem Wert $s = 0.41$ ist das 3-Niveau-Schema 4. Ordnung (d.h. (6.15) mit (6.18) und $\gamma = 1$) überlegen, insbesondere bei höherer räumlicher Auflösung.

6.1.2. Implizite Verfahren

Bei impliziten Verfahren werden die räumlichen Ableitung zumindest teilweise auf dem jeweils neu zu berechnenden Zeitniveau $n+1$ gebildet. Damit sind die Unbekannten miteinander gekoppelt und man muß in jedem Fall ein lineares Gleichungssystem lösen.

Voll-implizites Verfahren

Das einfachste implizite Verfahren ist die voll-implizite Diskretisierung

$$\frac{T_j^{n+1} - T_j^n}{\Delta t} = \kappa \frac{T_{j+1}^{n+1} - 2T_j^{n+1} + T_{j-1}^{n+1}}{\Delta x^2}. \quad (6.20)$$

Wenn man nach dem Zeitniveau $n+1$ auflöst, ergibt sich

$$-sT_{j+1}^{n+1} + (1+2s)T_j^{n+1} - sT_{j-1}^{n+1} = T_j^n, \tag{6.21}$$

wobei wie üblich $s = \kappa\Delta t/\Delta x^2$. Per Taylor-Entwicklung kann man wieder zeigen, daß das Verfahren von 2. Ordnung im Raum ist, für $s = 1/6$ sogar von 4. Ordnung. Aufgrund des Verstärkungsfaktors

$$G = \frac{1}{1 + 2s\,(1 - \cos\theta)} \tag{6.22}$$

ist immer $|G| \leq 1$, weshalb das Schema uneingeschränkte Stabilität besitzt.

Wenn man das Gleichungssystem nach (6.21) aufstellt, ergibt sich für die internen Punkte T_2^{n+1} bis T_{J-1}^{n+1}

$$\begin{pmatrix} (1+2s) & -s & & & & \\ -s & (1+2s) & -s & & & \\ & -s & \cdot & -s & & \\ & & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & \cdot \\ & & & & -s & (1+2s) \end{pmatrix} \cdot \begin{pmatrix} T_2^{n+1} \\ T_3^{n+1} \\ T_4^{n+1} \\ \dots \\ \dots \\ T_{J-1}^{n+1} \end{pmatrix} = \begin{pmatrix} T_2^n + sT_1^{n+1} \\ T_3^n \\ T_4^n \\ \dots \\ \dots \\ T_{J-1}^n + sT_J^{n+1} \end{pmatrix}. \tag{6.23}$$

Die rot gekennzeichneten Randwerte T_1^{n+1} und T_J^{n+1} sind über die Randbedingungen fest vorgegeben. Da das Gleichungssystem tridiagonal ist, kann es leicht mit dem *Thomas-Algorithmus* (Kap. 5.1.3) gelöst werden.

Die Lösung dieses Systems mit Hilfe des Thomas-Algorithmus kostet nur ungefähr den zweifachen Rechenaufwand wie die Lösung der Diffusionsgleichung mittels explizitem FTCS-Verfahren. Beim impliziten Verfahren kann aber die Zeitschrittweite größer gewählt werden. Der Vergrößerung der Zeitschrittweite sind jedoch Grenzen gesetzt. Denn bei großen Zeitschritten bleibt das Verfahren zwar stabil, es wird aber ungenau.[7]

Crank-Nicolson-Verfahren

Das *Crank-Nicolson-Schema* (Abb. 6.2e) ist teilimplizit, wobei die räumlichen Ableitungen bei $n+1$ und n gleichstark gewichtet werden,

$$\frac{T_j^{n+1} - T_j^n}{\Delta t} = \frac{\kappa}{2}\left(L_{xx}T_j^{n+1} + L_{xx}T_j^n\right). \tag{6.24}$$

John Crank
1916–2006

Die Konsistenzprüfung mittels Taylor-Entwicklung um $(j, n+\frac{1}{2})$ liefert einen Abbruchfehler der Größenordnung $O\left(\Delta t^2, \Delta x^2\right)$. Damit ist das Crank-Nicolson-Verfahren immer von 2. Ordnung in der Zeit, was einen signifikanten Vorteil gegenüber dem voll-impliziten und dem FTCS-Verfahren

[7] Diese Ungenauigkeiten spielen keine Rolle, wenn man nur an der stationären Lösung für $t \to \infty$

darstellt. Außerdem ist es uneingeschränkt stabil. Aus diesen Gründen ist das Crank-Nicolson-Verfahren sehr populär. Durch Trennen der Zeitniveaus ergibt sich der Algorithmus

$$-\frac{s}{2}T_{j+1}^{n+1} + (1+s)T_j^{n+1} - \frac{s}{2}T_{j-1}^{n+1} = \frac{s}{2}T_{j+1}^{n} + (1-s)T_j^{n} + \frac{s}{2}T_{j-1}^{n}. \tag{6.25}$$

Dies führt wiederum auf ein tridiagonales System, das man effizient mit dem Thomas-Algorithmus lösen kann.

Phyllis Nicolson
1917– 1968

Das Crank-Nicolson-Schema kann man nun verallgemeinern, indem man die Wichtung der beiden Zeitniveaus der räumlichen Ableitung variiert (siehe Fletcher (1991*a*)). Wenn sich das System, d.h. die exakte Lösung, auf stark unterschiedlichen Zeitskalen entwickelt[8] (steifes System), besitzt das Crank-Nicolson-Verfahren den Nachteil, daß es unphysikalische, oszillierende Lösungen liefert. In solchen Fällen sind einige 3-Niveau-Schemata besser.

Verallgemeinertes 3-Niveau-Schema

In Anlehnung an das explizite 3-Niveau-Schema (6.15) kann man das implizite 3-Niveau-Schema (Abb. 6.2f)

$$(1+\gamma)\,\frac{T_j^{n+1}-T_j^n}{\Delta t} - \gamma\frac{T_j^n - T_j^{n-1}}{\Delta t} = \kappa\left[\beta L_{xx}T_j^{n+1} + (1-\beta)\,L_{xx}T_j^n\right] \tag{6.26}$$

konstruieren, das bei Fletcher (1991*a*) 3LFI (*3-level-fully-implicit*) genannt wird. Ein besonders effizientes Verfahren erhält man für $\gamma = 1/2$ und $\beta = 1$ (Abb. 6.2g). Bei dieser Wahl der Parameter ist das Verfahren, wie das Crank-Nicolson-Verfahren, von zweiter Ordnung in der Zeit $O\left(\Delta t^2, \Delta x^2\right)$. Es besitzt aber bessere Eigenschaften bei steifen Problemen. Unphysikalische Oszillationen werden gedämpft. Auch die Gleichungen für das 3LFI-Schema kann man mit dem Thomas-Algorithmus lösen. Für

$$\beta = \frac{1}{2} + \gamma + \frac{1}{12s} \tag{6.27}$$

erhält man sogar eine Genauigkeit 4. Ordnung im Raum.

Fazit: Wenn man die obigen impliziten Schemata implementiert, findet man, daß alle Verfahren auf *groben* Gittern in etwa gleich gut bzw. gleich schlecht sind. Auf *feinen* Gittern wird aber die Überlegenheit der Verfahren höherer (4.) Ordnung deutlich sichtbar.

interessiert ist. Die Dynamik wird aber nicht richtig wiedergegeben.

[8] Ein Problem wird *steif* genannt, wenn die Lösung zeitweise extrem schnell variiert und dann wieder sehr langsam. Dann wird die maximale Zeitschrittweite durch die schnellste Änderungsrate bestimmt, was sehr restriktiv sein kann.

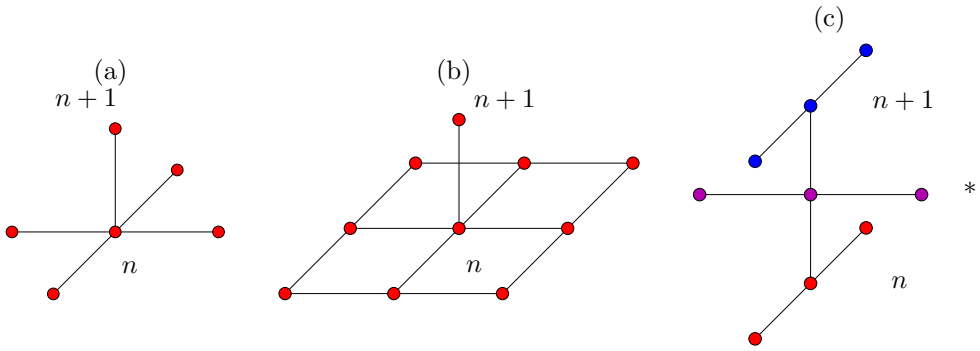


Abbildung 6.4.: Differenzenstern für das 2D-FTCS-Verfahren (a), dessen Erweiterung nach (6.32) (b) und für das ADI-Verfahren (6.37) (c).

Ähnliches gilt für den Vergleich von expliziten und impliziten Verfahren. Auf groben Gittern sind sie etwa gleich gut bzw. schlecht. Auf feinen Gittern besitzen jedoch die impliziten Verfahren in der Regel eine höhere Genauigkeit.

6.2. Mehrdimensionale Diffusion

Meist lassen sich reale Problemstellungen nicht auf eine Raumdimension reduzieren. Dann muß man zwei- oder dreidimensionale Rechnungen durchführen. Für die zweidimensionale Diffusionsgleichung

$$\frac{\partial T}{\partial t} = \kappa \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} \right) \tag{6.28}$$

liefert die direkte Erweiterung des eindimensionalen FTCS-Verfahrens das explizite Verfahren (Abb. 6.4a)

$$\frac{T_{j,k}^{n+1} - T_{j,k}^{n}}{\Delta t} = \kappa \left(L_{xx} T_{j,k}^{n} + L_{yy} T_{j,k}^{n} \right), \tag{6.29}$$

was nach Auflösen als

$$\begin{aligned} T_{j,k}^{n+1} = & \quad T_{j,k}^{n} + \kappa \Delta t \left(\frac{T_{j+1,k}^{n} - 2T_{j,k}^{n} + T_{j-1,k}^{n}}{\Delta x^2} + \frac{T_{j,k+1}^{n} - 2T_{j,k}^{n} + T_{j,k-1}^{n}}{\Delta y^2} \right) \\ = & \quad (1 - 2s_x - 2s_y)\, T_{j,k}^{n} + s_x T_{j+1,k}^{n} + s_x T_{j-1,k}^{n} + s_y T_{j,k+1}^{n} + s_y T_{j,k-1}^{n}. \end{aligned} \tag{6.30}$$

geschrieben werden kann, mit $s_x = \kappa \Delta t / \Delta x^2$ und $s_y = \kappa \Delta t / \Delta y^2$.[9] Der Abbruchfehler ist von der Ordnung $O(\Delta t, \Delta x^2, \Delta y^2)$.

Die Stabilität dieses Verfahrens ist gegenüber der eindimensionalen Version re-

[9] Hier könnte man auch unterschiedliche Diffusionskonstanten κ_x und κ_y in x- und y-Richtung einbauen.

duziert, denn man muß die Bedingung

$$s_x + s_y \leq \frac{1}{2} \tag{6.31}$$

erfüllen. Für $s = s_x = s_y$ entspricht dies $s \leq 1/4$. Im dreidimensionalen Fall reduziert sich der stabile Bereich der Schrittweite sogar auf $s \leq 1/6$.

Eine Modifikation des FTCS-Verfahrens für $s = s_x = s_y$ durch einen Zusatzterm stellt das 9-Punkt-Verfahren dar (Abb. 6.4b)

$$\frac{T_{j,k}^{n+1} - T_{j,k}^{n}}{\Delta t} = \kappa \left(L_{xx} T_{j,k}^{n} + L_{yy} T_{j,k}^{n} \right) + \underbrace{\kappa^2 \Delta t L_{xx} L_{yy} T_{j,k}^{n}}_{\to 0 \text{ für } \Delta t \to 0} . \tag{6.32}$$

Nach $T_{j,k}^{n+1}$ aufgelöst

$$T_{j,k}^{n+1} = T_{j,k}^{n} + \kappa \Delta t \left(L_{xx} T_{j,k}^{n} + L_{yy} T_{j,k}^{n} \right) + \kappa^2 \Delta t^2 L_{xx} L_{yy} T_{j,k}^{n} \tag{6.33}$$

$$= (1 + \kappa \Delta t L_{xx}) \underbrace{(1 + \kappa \Delta t L_{yy}) T_{j,k}^{n}}_{:= T_{j,k}^{*}} . \tag{6.34}$$

Jetzt erkennt man den Grund für die Wahl des Zusatzterms: Die rechte Seite kann faktorisiert werden. Dadurch kann die Gleichung sehr einfach in zwei Schritten

$$T_{j,k}^{*} = (1 + \kappa \Delta t L_{yy}) T_{j,k}^{n}, \tag{6.35a}$$

$$T_{j,k}^{n+1} = (1 + \kappa \Delta t L_{xx}) T_{j,k}^{*}, \tag{6.35b}$$

implementiert werden. In jedem Teilschritt jeweils nur 3 Werte benötigt. Außerdem besitzen die so modifizierten Gleichungen den erweiterten Stabilitätsbereich $s \leq 1/2$.

Bisher waren alle Schritte explizit. Die naheliegendste implizite Diskretisierung besteht darin, alle räumlichen Ableitungen bei $n+1$ zu berechnen (vgl. (6.30)), was auf

$$(1 + 2s_x + 2s_y) T_{j,k}^{n+1} - s_x T_{j+1,k}^{n+1} - s_x T_{j-1,k}^{n+1} - s_y T_{j,k+1}^{n+1} - s_y T_{j,k-1}^{n+1} = T_{j,k}^{n} \tag{6.36}$$

führt. Hierbei sind alle $N_x N_y$ Variablen gekoppelt zu lösen. Man kann die Variablen im Vektor der Unbekannten so anordnen, daß die drei Hauptdiagonalen besetzt sind. Aus der Ableitung in der zweiten Raumrichtung resultieren aber noch zwei weiter außen liegende Diagonalen im Abstand N_x bzw. N_y von der Hauptdiagonalen, wie in Abb. 5.1 (Kap. 5.1). Daher kommt eine Lösung mit dem einfachen Thomas-Algorithmus nicht in Frage. Andere direkte Verfahren, wie etwa das Gauß-Verfahren, sind zu teuer. Zu erwägen ist dann die näherungsweise Lösung mit dem SIP-Verfahren (Kap. 5.2.4).

6.2.1. Splitting-Methoden

Um die starke Kopplung der Variablen beim einfachen impliziten Verfahren (6.36) zu verringern und um tridiagionale Matrizen zu erhalten, kann man versuchen, das implizite Problem ähnlich wie im expliziten Fall (6.33) in zwei Teilschritten zu lösen. Dabei wird im ersten zeitlichen Teilschritt nur die Ableitung in einer Raumrichtung implizit behandelt, im darauf folgenden zweiten Teilschritt die andere Raumrichtung, und so weiter alternierend. Bei dieser Vorgehensweise kann man für jeden einzelnen Teilschritt den Thomas-Algorithmus nutzen.

ADI

Die am häufigsten verwendete *Splitting-Methode* ist die ***A**lternating **D**irections **I**mplicit Method* (*ADI-Methode*). Diese Methode hatten wir schon in Kap. 5.2.5 betrachtet. Sie lautet

$$\frac{T^{*}_{j,k} - T^{n}_{j,k}}{\Delta t/2} = \kappa \left(L_{xx} T^{*}_{j,k} + L_{yy} T^{n}_{j,k} \right), \tag{6.37a}$$

$$\frac{T^{n+1}_{j,k} - T^{*}_{j,k}}{\Delta t/2} = \kappa \left(L_{xx} T^{*}_{j,k} + L_{yy} T^{n+1}_{j,k} \right). \tag{6.37b}$$

Hier wird im ersten Schritt die x-Richtung und im zweiten Schritt die y-Richtung implizit behandelt. Die Gleichungen für die beiden Halbschritte kann man ausführlich schreiben als

$$-\frac{s_x}{2} T^{*}_{j+1,k} + (1+s_x) T^{*}_{j,k} - \frac{s_x}{2} T^{*}_{j-1,k} = \frac{s_y}{2} T^{n}_{j,k+1} + (1-s_y) T^{n}_{j,k} + \frac{s_y}{2} T^{n}_{j,k-1}, \tag{6.38a}$$

$$-\frac{s_y}{2} T^{n+1}_{j,k+1} + (1+s_y) T^{n+1}_{j,k} - \frac{s_y}{2} T^{n+1}_{j,k-1} = \frac{s_x}{2} T^{*}_{j+1,k} + (1-s_x) T^{*}_{j,k} + \frac{s_x}{2} T^{*}_{j-1,k}. \tag{6.38b}$$

Es treten nur tridiagonale Matrizen auf.

Der Verstärkungsfaktor G von Störungen für das ADI-Verfahren ergibt sich als Produkt der Verstärkungsfaktoren G_1 und G_2 für die beiden Teilschritte. Mit dem Ansatz für die Störungen $\xi^{n}_{jk} = G_1^n G_2^n \mathrm{e}^{\mathrm{i}\theta_x j} \mathrm{e}^{\mathrm{i}\theta_y k}$ und $\theta_x = l\pi\Delta x$ und $\theta_y = m\pi\Delta y$ erhält man aus den Von-Neumann-Stabilitätsgleichungen

$$G = G_1 G_2 = \underbrace{\left[\frac{1 - 2s_y \sin^2(\theta_y/2)}{1 + 2s_x \sin^2(\theta_x/2)} \right]}_{G_1} \underbrace{\left[\frac{1 - 2s_x \sin^2(\theta_x/2)}{1 + 2s_y \sin^2(\theta_y/2)} \right]}_{G_2}. \tag{6.39}$$

Wegen

$$|G| = \left| \frac{1 - 2s_y \sin^2(\theta_y/2)}{1 + 2s_y \sin^2(\theta_y/2)} \right| \left| \frac{1 - 2s_x \sin^2(\theta_x/2)}{1 + 2s_x \sin^2(\theta_x/2)} \right| \leq 1 \tag{6.40}$$

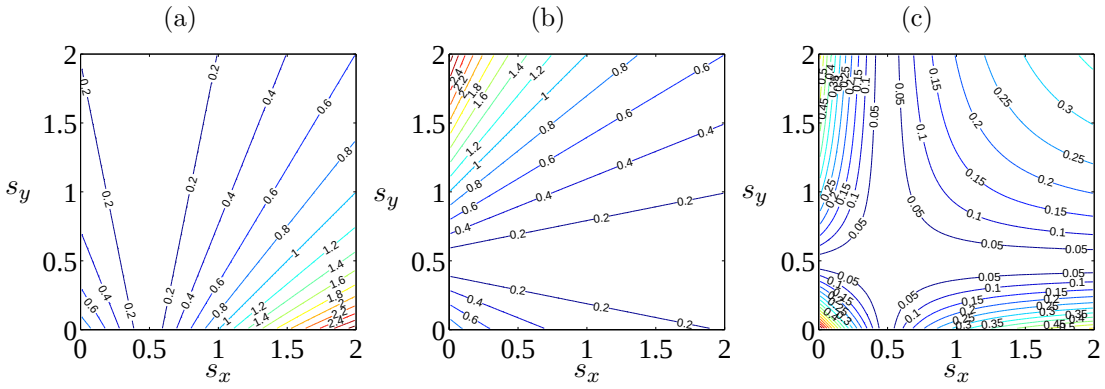


Abbildung 6.5.: Verstärkungsfaktoren $\max_{\theta_x,\theta_y} |G_1|$ (a), $\max_{\theta_x,\theta_y} |G_2|$ (b) und $\max_{\theta_x,\theta_y} |G_1 G_2|$ (c) nach (6.39) für das ADI-Schema (6.37).

ist das gesamte ADI-Verfahren für alle s_x, s_y stabil. Jedoch ist jeder Teilschritt nur bedingt stabil (siehe Abb. 6.5). Das ADI-Verfahren besitzt eine Fehlerordnung von $O(\Delta t^2, \Delta x^2, \Delta y^2)$. Die Genauigkeit zweiter Ordnung resultiert aus der zeitlich symmetrischen Formulierung (siehe auch Abb. 6.4c), ähnlich wie beim Crank-Nicolson-Schema aus Kap. 6.1.2.

Damit diese Genauigkeit aber tatsächlich realisiert wird, muß man auch die Randbedingungen für die Zwischenwerte $T^*_{j,k}$ so formulieren, daß sie kompatibel mit dem Algorithmus sind. Wenn zum Beispiel die Funktion bei $j = N_x$ durch $T^n_{N_x,k} = b^n_k$ vorgegeben ist, dann kann man nicht einfach $T^*_{N_x,k} = (b^{n+1}_k + b^n_k)/2$ verwenden. Der Fehler wäre dann $O(\Delta t)$. Die korrekte Randbedingung erhält man vielmehr durch Einsetzen von $T^n_{j,k} = T^n_{N_x,k} = b^n_k$ in (6.37a) und (6.37b) und Subtraktion der beiden Gleichungen (siehe auch Abb. 6.6). Dann fällt der Term $L_{xx} T^*_{jk}$ auf der rechten Seite weg und man erhält[10]

$$T^*_{N_x,k} = \frac{b^{n+1}_k + b^n_k}{2} - \frac{1}{4}\Delta t \kappa L_{yy} \left(b^{n+1}_k - b^n_k\right). \tag{6.41}$$

Man kann das ADI-Schema auch in naheliegender Weise auf drei Dimensionen erweitern, wobei man drei Teilschritte zu jeweils $\Delta t/3$ macht und in jedem Schritt nur je eine Raumrichtung implizit behandelt. Die Genauigkeit zweiter Ordnung und die schnelle Lösbarkeit der Gleichungen bleiben erhalten, nur ist das ADI-Verfahren in drei Dimensionen nicht mehr uneingeschränkt stabil. Es gilt dann die Beschränkung $s_x, s_y, s_z \leq 3/2$.

Verallgemeinertes 2-Niveau-Schema

Beim ADI-Verfahren hatten wir einfach *ad hoc* ein zusätzliches Zeit-Niveau bei $n + 1/2$ eingeführt, wobei der explizite und der implizite Anteil in jedem Schritt

[10] Wenn man in (6.37) zuerst die y- und dann die x-Richtung implizit behandelt, also in umgekehrter Reihenfolge vorgeht, dann erhält man für die Randwerte bei $k = N_y$ einen entsprechen-

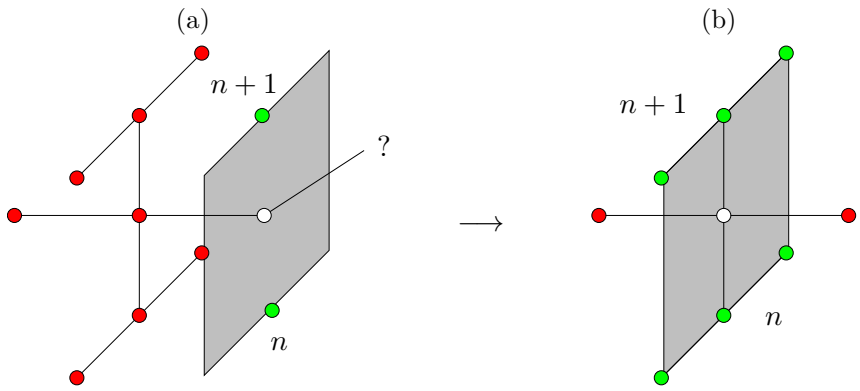


Abbildung 6.6.: Den unbekannten Randwert ∘ in (a) erhält man aus den anderen Randpunkten (grün) durch Differenzbildung der ADI-Gleichungen (6.37a) und (6.37b) für die Punkte in (b), d.h. für $j = N_x$. Dadurch werden die roten Punkte eliminiert (L_{xx} fällt heraus).

gleich gewichtet wurden. Wir wollen nun ein allgemeines 2-Niveau-Schema ableiten, wobei wir den Algorithmus für das Inkrement

$$\Delta T_{j,k}^{n+1} = T_{j,k}^{n+1} - T_{j,k}^{n} \tag{6.42}$$

der gesuchten Funktion schreiben. Wir gehen zunächst von dem allgemeinen teilimpliziten Verfahren

$$\frac{\Delta T_{j,k}^{n+1}}{\Delta t} = \beta\kappa \left(L_{xx} T_{j,k}^{n+1} + L_{yy} T_{j,k}^{n+1}\right) + (1-\beta)\,\kappa \left(L_{xx} T_{j,k}^{n} + L_{yy} T_{j,k}^{n}\right) \tag{6.43}$$

aus. Mittels (6.42) eliminieren wir nun auf der rechten Seite $T_{j,k}^{n+1}$ zugunsten von $\Delta T_{j,k}^{n+1}$. Dann ergibt sich

$$\frac{\Delta T_{j,k}^{n+1}}{\Delta t} = \beta\kappa \left(L_{xx} \Delta T_{j,k}^{n+1} + L_{yy} \Delta T_{j,k}^{n+1}\right) + \kappa \left(L_{xx} T_{j,k}^{n} + L_{yy} T_{j,k}^{n}\right). \tag{6.44}$$

Wenn wir nun alle Terme zur Zeit $n+1$ auf die linke Seite bringen, erhalten wir

$$\left[1 - \Delta t\beta\kappa \left(L_{xx} + L_{yy}\right)\right] \Delta T_{j,k}^{n+1} = \Delta t\kappa \left(L_{xx} + L_{yy}\right) T_{j,k}^{n}. \tag{6.45}$$

Um die linke Seite in Faktoren zerlegen zu können, machen wir nun einen kleinen Fehler – ähnlich wie in (6.32) – und schreiben

$$\left(1 - \Delta t\beta\kappa L_{xx}\right) \underbrace{\left(1 - \Delta t\beta\kappa L_{yy}\right) \Delta T_{j,k}^{n+1}}_{:=\Delta T_{j,k}^{*}} = \Delta t\kappa \left(L_{xx} + L_{yy}\right) T_{j,k}^{n}. \tag{6.46}$$

den Ausdruck, nur mit $T_{j,N_y}^{*} = (b_j^{n+1} + b_j^n)/2 - (\Delta t/4)\kappa L_{xx}\left(b_j^{n+1} - b_j^n\right)$, wobei $b_j^n = T_{j,N_y}^n$ der Randwert auf $k = N_y$ ist.

Hierbei haben wir den Fehler $\Delta t^2 \beta^2 \kappa^2 L_{xx} L_{yy} \Delta T_{j,k}^{n+1}$ gemacht. Er ist von der Ordnung $O(\Delta t^2)$. Das macht aber nichts, solange das Verfahren selbst nur von erster Ordnung in der Zeit ist.[11] In der Tat ist das aus (6.46) resultierende Verfahren lediglich für $\beta = 1/2$ von zweiter Ordnung.

Die Differenzengleichung (6.46) wird nun in zwei Schritten gelöst

$$(1 - \Delta t \beta \kappa L_{xx})\, \Delta T_{j,k}^{*} = \Delta t \kappa\, (L_{xx} + L_{yy})\, T_{j,k}^{n}, \tag{6.47a}$$

$$(1 - \Delta t \beta \kappa L_{yy})\, \Delta T_{j,k}^{n+1} = \Delta T_{j,k}^{*}. \tag{6.47b}$$

Beide Teilschritte können mit dem Thomas-Algorithmus effizient ausgeführt werden.

Die Methode der Faktorisierung des impliziten Operators auf der linken Seite der Gleichung kann auch auf drei Dimensionen erweitert werden. In drei Dimensionen hat das allgemeine Zwei-Niveau-Schema gegenüber dem normalen ADI-Verfahren (entsprechend $\beta = 1/2$) den Vorteil, daß es uneingeschränkt stabil ist, wenn der implizite Anteil überwiegt, also für $\beta \geq 1/2$. Darüber hinaus kann man die Strategie des *Splitting* durch Addition eines kleine Fehlerterms auch auf Schemata mit drei Zeitniveaus (beispielsweise (6.26)) anwenden (siehe Fletcher (1991*a*)).

Bei den oben beschriebenen *Splitting-Schemata* wurde die Differentialgleichung zuerst diskretisiert. Danach wurde versucht, die Operatoren, welche die verschiedenen Raumrichtungen miteinander koppeln, zu faktorisieren. Die Faktorisierung erlaubte dann ein schrittweise Berechnung, wobei in jedem Teilschritt jeweils nur eine einzige Raumrichtung behandelt wird. Damit lassen sich die Gleichungssysteme ökonomisch lösen.

Eine gewisse Modifikation des *Splitting* ist die *Fractional-Step-Methode*. Hierbei wird schon die Differentialgleichung, z.B. die zweidimensionale Diffusionsgleichung

$$\frac{\partial T}{\partial t} = \kappa \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial x^2} \right) T \tag{6.48}$$

in zwei Gleichungen aufgespalten. Erst danach erfolgt die Diskretisierung. Dazu schreibt man

$$\frac{1}{2} \frac{\partial T}{\partial t} = \kappa \frac{\partial^2 T}{\partial x^2}, \tag{6.49a}$$

$$\frac{1}{2} \frac{\partial T}{\partial t} = \kappa \frac{\partial^2 T}{\partial y^2}. \tag{6.49b}$$

In dieser Form sind die Gleichungen natürlich nicht korrekt, da sie im allgemeinen, d.h. für $\partial^2 T/\partial x^2 \neq \partial^2 T/\partial y^2$, widersprüchlich sind. Deshalb sind diese Gleichungen lediglich algorithmisch zu verstehen, im Sinne einer sukzessiven Lösung. Wenn

[11] Wenn das Verfahren $O(\Delta t)$ ist, dann ist der Fehler in der Gleichung (6.46) für das Inkrement $O(\Delta t^2)$. Der durch die Faktorisierung zusätzlich gemachte Fehler ist demnach von derselben Fehlerordnung wie das ursprüngliche Verfahren ohne Splitting. Die Differenzengleichung bleibt

Table 7.1. Algebraic (discretised) schemes for the diffusion equation $\partial \bar{T}/\partial t - \alpha \partial^2 \bar{T}/\partial x^2 = 0$

Scheme	Algebraic form	Truncation error[a] (E) (leading term)	Amplification factor $G(\theta = m\pi \Delta x)$	Stability restrictions	Remarks
FTCS	$\dfrac{\Delta T_j^{n+1}}{\Delta t} - \alpha L_{xx} T_j^n = 0$	$\alpha(\Delta x^2/2)\left(s - \dfrac{1}{6}\right)\dfrac{\partial^4 T}{\partial x^4}$	$1 - 4s \sin^2\left(\dfrac{\theta}{2}\right)$	$s \leqq 0.5$	$s = \alpha \dfrac{\Delta t}{\Delta x^2}$
					$L_{xx} = \dfrac{1}{\Delta x^2}[1, -2, 1]$
DuFort-Frankel	$\dfrac{T_j^{n+1} - T_j^{n-1}}{2\Delta t} - \dfrac{\alpha}{\Delta x^2}[T_{j-1}^n - (T_j^{n-1} + T_j^{n+1}) + T_{j+1}^n] = 0$	$\alpha \Delta x^2 \left(s^2 - \dfrac{1}{12}\right)\dfrac{\partial^4 T}{\partial x^4}$	$\dfrac{2s\cos\theta + (1 - 4s^2 \sin^2\theta)^{1/2}}{(1+2s)}$	None	$\Delta T_j^{n+1} = T_j^{n+1} - T_j^n$
Crank-Nicolson	$\dfrac{\Delta T_j^{n+1}}{\Delta t} - \alpha L_{xx}\left(\dfrac{T_j^n + T_j^{n+1}}{2}\right) = 0$	$-\alpha\left(\dfrac{\Delta x^2}{12}\right)\dfrac{\partial^4 T}{\partial x^4}$	$\dfrac{1 - 2s\sin^2(\theta/2)}{1 + 2s\sin^2(\theta/2)}$	None	
Three-level fully implicit	$\dfrac{3}{2}\dfrac{\Delta T_j^{n+1}}{\Delta t} - \dfrac{1}{2}\dfrac{\Delta T_j^n}{\Delta t} - \alpha L_{xx} T_j^{n+1} = 0$	$-\alpha\left(\dfrac{\Delta x^2}{12}\right)\dfrac{\partial^4 T}{\partial x^4}$	$\dfrac{1 \pm \frac{4}{3} i\,[\frac{3}{16} + s(1 - \cos\theta)]^{1/2}}{2\,[1 + \frac{4}{3}s(1 - \cos\theta)]}$	None	$\Delta T_j^n = T_j^n - T_j^{n-1}$
Linear F.E.M. /Crank-Nicolson	$M_x \dfrac{\Delta T_j^{n+1}}{\Delta t} - \alpha L_{xx}\left(\dfrac{T_j^n + T_j^{n+1}}{2}\right) = 0$	$\alpha\left(\dfrac{\Delta x^2}{12}\right)\dfrac{\partial^4 T}{\partial x^4}$	$\dfrac{(2 - 3s) + \cos\theta(1 + 3s)}{(2 + 3s) + \cos\theta(1 - 3s)}$	None	$M_x = \{\frac{1}{6}, \frac{2}{3}, \frac{1}{6}\}$

[a] The truncation error has been expressed solely in terms of Δx and x-derivatives as in the modified equation method (Section 9.2.2). Thus the algebraic scheme is equivalent to $\partial T/\partial t - \alpha \partial^2 T/\partial x^2 + E(T) = 0$

Abbildung 6.7.: Verfahren für die Diffusionsgleichung (Fletcher, 1991*a*).

man die beiden Gleichungen in diesem Sinne diskretisiert, kann man das explizite Verfahren (6.35) erhalten oder, bei impliziter Diskretisierung, das ADI-Verfahren (6.38).

Abschließend sei noch einmal betont, daß es wichtig ist, auch die Randbedingungen konsistent mit derselben Fehlerordnung zu implementieren, wie die des Integrationsverfahrens. Ansonsten wird die Genauigkeit der Verfahrens generell reduziert. Eine Übersicht über die wichtigsten Eigenschaften der gängigsten Integrationsverfahren für die Diffusionsgleichung ist in Abb. 6.7 gezeigt.

6.3. Advektion

In diesem Abschnitt wollen wir die Eigenschaften verschiedener Integrationsverfahren für hyperbolische Gleichungen am Beispiel der für die Strömungsmechanik wichtigen Advektionsgleichung (6.4) untersuchen. Wie zu Beginn dieses Kapitels erwähnt, erhält man die Lösung der Advektionsgleichung für ein unendlich ausgedehntes Gebiet einfach durch die Translation der Anfangsbedingung um die Länge $L = u(t - t_0)$. Die Kenntnis dieser exakten Lösung ermöglicht eine leichte Überprüfung der numerischen Ergebnisse verschiedener Verfahren. Wenn zusätzlich auch Diffusionsterme in der Differentialgleichung auftreten (siehe Kap. 6.4), ist die Lösung natürlich nicht mehr so einfach darzustellen.

Zunächst werden wir die lineare Advektionsgleichung betrachten. Die nichtlineare Advektionsgleichung wird später behandelt. Die Advektionsgleichung ist von erster Ordnung im Raum. Wir werden sehen, daß eine symmetrische 3-Punkt-Berechnung der ersten räumlichen Ableitung zu unphysikalischen Oszillationen führt. Andererseits wird die Ordnung des Verfahrens meist verringert, wenn man eine stabilere asymmetrische räumliche Diskretisierung verwendet.

6.3.1. FTCS-Schema

Wir betrachten die *Advektionsgleichung* (6.4)

$$\frac{\partial T}{\partial t} + u\frac{\partial T}{\partial x} = 0 \tag{6.50}$$

auf der reellen Achse $x \in [-\infty, \infty]$ mit $u = \text{const.} > 0$. Die einfachste Diskretisierung stellt das explizite FTCS-Schema dar

$$\frac{T_j^{n+1} - T_j^n}{\Delta t} + u\frac{T_{j+1}^n - T_{j-1}^n}{2\Delta x} = 0. \tag{6.51}$$

Wir schreiben es in der Form

$$T_j^{n+1} = T_j^n - \frac{\mathrm{C}}{2}\left(T_{j+1}^n - T_{j-1}^n\right). \tag{6.52}$$

Hierbei ist

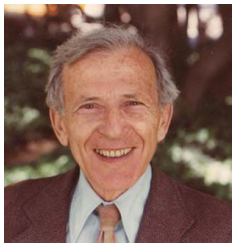
$$\mathrm{C} = u\frac{\Delta t}{\Delta x} \tag{6.53}$$

Kurt Otto Friedrichs 1901–1982

die *Courant-Zahl*.[12] Die Courant-Zahl ist das Verhältnis der Geschwindigkeit u des tatsächlichen physikalischen Transports der Größe T zu der Geschwindigkeit $\Delta x/\Delta t$, mit der sich eine Information bei einem *expliziten* Verfahren auf dem Gitter $(\Delta x, \Delta t)$ ausbreiten kann. Aus dieser Betrachtung folgt, daß Werte $\mathrm{C} > 1$ keinen Sinn machen. Denn für $\mathrm{C} > 1$ reicht die numerisch maximal mögliche Ausbreitungsgeschwindigkeit nicht aus, um die die Größe T mit der erforderlichen physikalischen Geschwindigkeit u auf dem gegebenen Gitter zu transportieren. Wir müssen also verlangen

$$\mathrm{C} \le 1. \tag{6.54}$$

Diese Bedingung wird *Courant-Friedrichs-Lewy-Bedingung* (CFL-Bedingung) genannt. Sie besagt, daß sich ein physikalisches Fluidelement innerhalb eines Zeitschritts Δt nicht weiter bewegen darf als die räumliche Gitterweite Δx.



Hans Lewy 1904–1988

Der Abbruchfehler des FTCS-Verfahrens ist $O(\Delta t, \Delta x^2)$. Eine Untersuchung der Von-Neumann-Stabilität ergibt den Verstärkungsfaktor von Störungen innerhalb eines Zeitschritts

$$G = 1 - \mathrm{i}\mathrm{C}\sin\theta \qquad \Rightarrow \qquad |G| = \sqrt{1 + \mathrm{C}^2\sin^2\theta} \ge 1. \tag{6.55}$$

Wegen $|G| \ge 1$ ist das FTCS-Schema für alle Werte von C uneingeschränkt instabil.[13] Die Auswirkung der Instabilität auf die Entwicklung von T sind in Abb. 6.8a,b zu sehen. In der Nähe von starken Gradienten kommt es zu Oszillationen, deren Amplitude explodiert. Aus diesem Grund ist das FTCS-Schema für reine Advektionsgleichungen nicht zu gebrauchen.

6.3.2. Upwind-Verfahren

Die Instabilität tritt bei einseitigen Differenzen nicht auf. Wenn man räumliche Rückwärtsdifferenzen bildet, ergibt sich

$$T_j^{n+1} = T_j^n - \mathrm{C}\left(T_j^n - T_{j-1}^n\right) = (1-\mathrm{C})\,T_j^n + \mathrm{C}T_{j-1}^n. \tag{6.56}$$

---

also konsistent.

[12] Sie wird auch Courant-Friedrichs-Lewy-Parameter genannt.

[13] Für die Diffusionsgleichung war das FTCS-Schema bedingt stabil. Durch die Modifikation $T_j^n \to (T_{j+1}^n + T_{j-1}^n)/2$ kann das FTCS-Schema stabilisiert werden. Dies führt auf das Lax-Friedrichs-Schema

$$T_j^{n+1} = \frac{1}{2}\left(T_j^{n+1} + T_j^{n-1}\right) - \frac{\mathrm{C}}{2}\left(T_{j+1}^n - T_{j-1}^n\right).$$

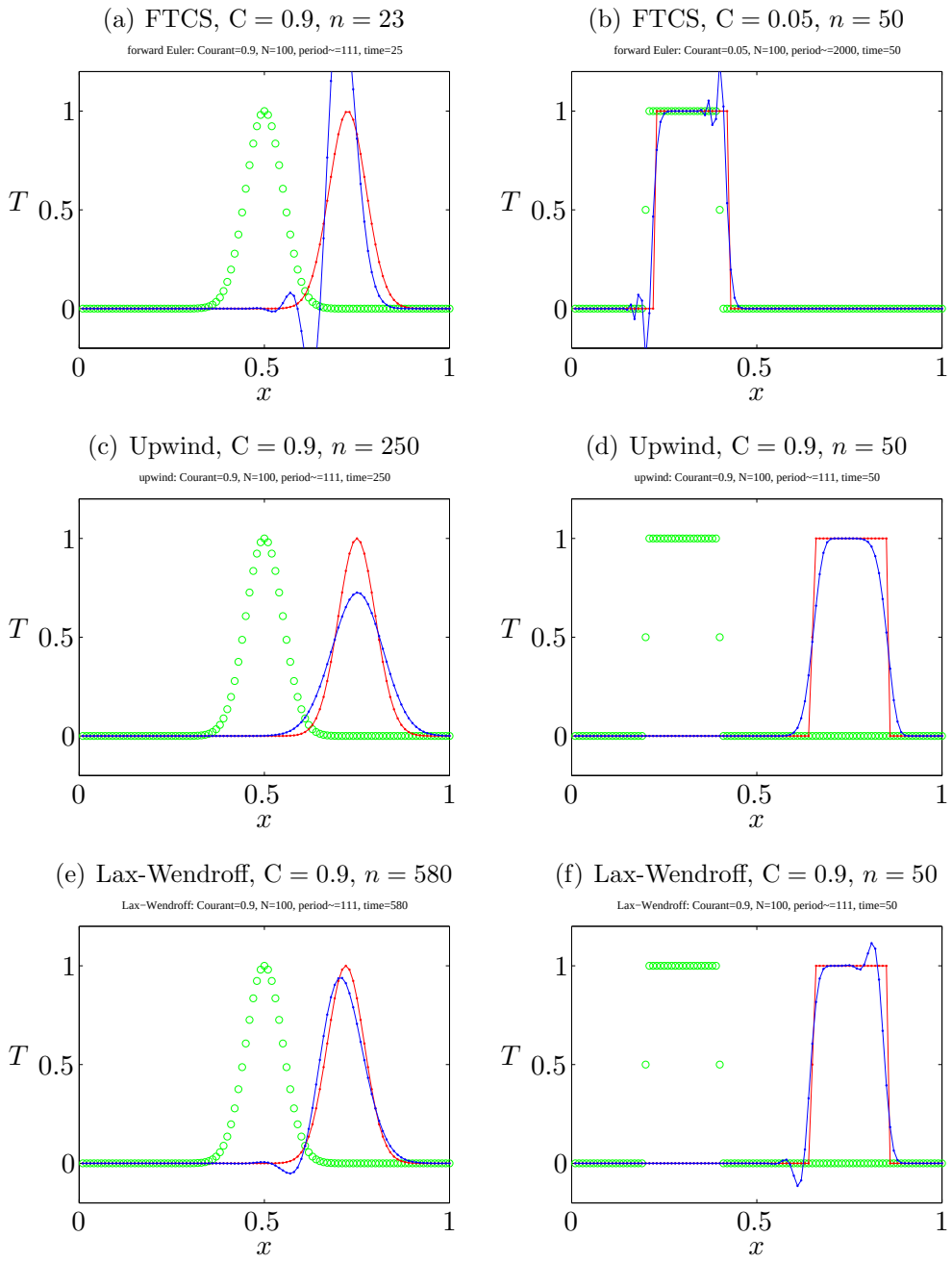
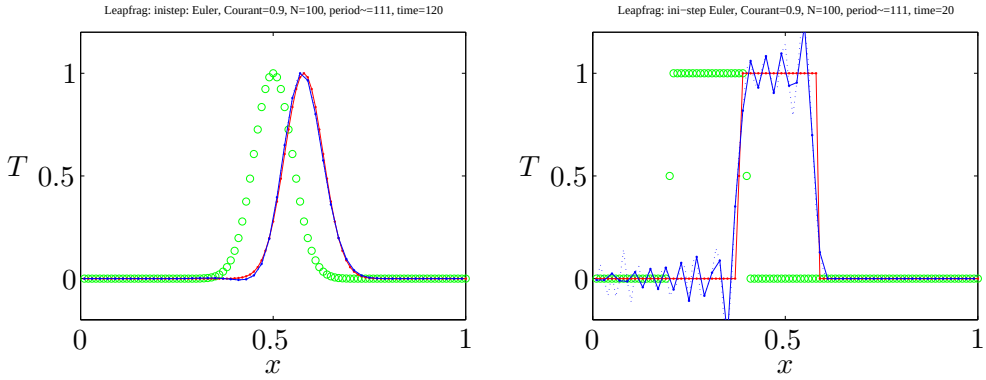
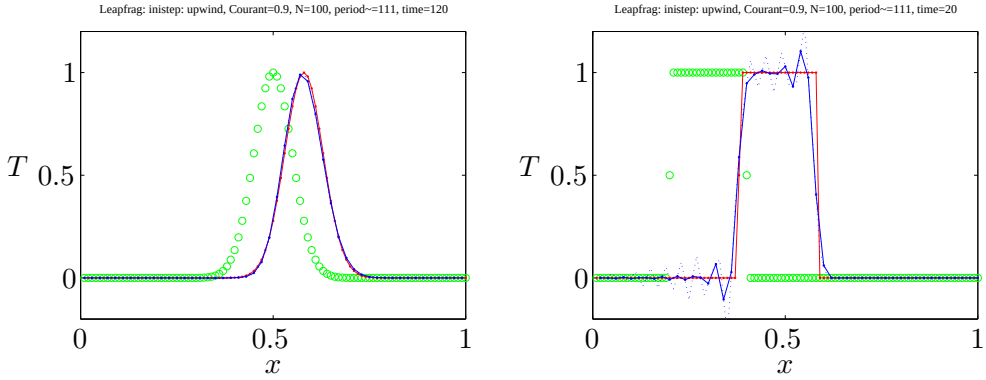


Abbildung 6.8.: Numerisch berechnete Lösungen (blau) im Vergleich mit der exakten Lösung (rot) für die beiden Anfangsbedingungen (grün) eines Gauß-Pakets und eines Rechteckpulses. Gezeigt sind Ergebnisse für das FTCS-Schema (a,b), das *Upwind*-Schema (c,d) und das Lax-Wendroff-Schema (e,f) für den angegebenen Zeitpunkt n. Mit der räumlichen Periodenlänge $L = 1$ und mit $\Delta x = 1/N$ ist die zeitliche Periode $T = L/u = 1/u = \Delta t/(\Delta x \mathrm{C}) = (N/\mathrm{C})\Delta t$. Nach $n_T = N/C$ Zeitschritten nimmt die exakte Lösung also wieder die Ausgangskonfiguration an. In allen Fällen ist $N = 100$ die Anzahl der räumlichen Stützstellen. Damit ist für alle Fälle $n_T = 111$, nur für (b) ist $n_T = 2000$.

(g) Leapfrog (Euler), C = 0.9, $n = 120$ (h) Leapfrog (Euler), C = 0.9, $n = 20$



(i) Leapfrog (Upwind), C = 0.9, $n = 120$ (j) Leapfrog (Upwind), C = 0.9, $n = 20$



Noch Abbildung 6.8.: Beschreibung siehe oben. Hier sind die Ergebnisse für das *Leapfrog*-Verfahren gezeigt, wobei in (g,h) zu Beginn der Rechnung ein FTCS-Schritt gemacht wurde, und in (i,j) ein *Upwind*-Schritt. Die zwei unabhängigen numerischen Lösungen sind durch gestrichelte bzw. durchgezogene blaue Linien gekennzeichnet.

Wegen $u > 0$ werden die Differenzen also zur Luv-Seite gebildet. Man spricht daher auch von *Upwind-Verfahren*.[14] Falls $u < 0$ ist, muß man die räumliche Ableitung nach rechts nehmen

$$T_j^{n+1} = T_j^n - \mathrm{C}\left(T_{j+1}^n - T_j^n\right) \overset{u\le 0}{=} T_j^n - |\mathrm{C}|\left(T_j^n - T_{j+1}^n\right) = (1-|\mathrm{C}|)\,T_j^n + |\mathrm{C}|T_{j+1}^n. \tag{6.57}$$

Die Von-Neumann-Analyse von (6.56) liefert

$$G = 1 - \mathrm{C} + \mathrm{C}\mathrm{e}^{-\mathrm{i}\theta}. \tag{6.58}$$

Dies ergibt

$$|G|^2 = 1 - 2\mathrm{C}\underbrace{(1-\mathrm{C})}_{\ge 0 \text{ für } \mathrm{C}\le 1}\underbrace{(1-\cos\theta)}_{\in[0,2]}. \tag{6.59}$$

[14] Manchmal wird das Upwind-Verfahren auch *Upstream*-Verfahren oder *Donor-Cell*-Schema ge-


Richard Courant
1888–1972

Wir erhalten also einen stabilen Algorithmus für $\mathrm{C} \leq 1$. Das ist gerade die Courant-Friedrichs-Levy-Bedingung. In der Regel findet man diese Stabilitätsbedingung für alle expliziten Verfahren für hyperbolische Systeme. Interessanterweise erhält man für $\mathrm{C} = 1$ das Schema $T_j^{n+1} = T_{j-1}^n$. Dies entspricht gerade der exakten Lösung der Advektionsgleichung.

Zur Bestimmung der Fehlerordnung des *Upwind*-Verfahren setzen wir die Taylorentwicklungen der exakten Lösung $\bar{T}$ um den Punkt (n, j)

$$\bar{T}_j^{n+1} = \bar{T}_j^n + \Delta t \frac{\partial \bar{T}}{\partial t} + \frac{\Delta t^2}{2} \frac{\partial^2 \bar{T}}{\partial t^2} + O\left(\Delta t^3\right), \tag{6.60a}$$

$$\bar{T}_{j-1}^n = \bar{T}_j^n - \Delta x \frac{\partial \bar{T}}{\partial x} + \frac{\Delta x^2}{2} \frac{\partial^2 \bar{T}}{\partial x^2} + O\left(\Delta x^3\right), \tag{6.60b}$$

in das *Upwind*-Schema (6.56) ein und erhalten

$$\frac{\partial \bar{T}}{\partial t} + u \frac{\partial \bar{T}}{\partial x} = -\frac{\Delta t}{2} \frac{\partial^2 \bar{T}}{\partial t^2} + u \frac{\Delta x}{2} \frac{\partial^2 \bar{T}}{\partial x^2} + O\left(\Delta t^2, \Delta x^2\right). \tag{6.61}$$

Das *Upwind*-Verfahren ist also von erster Ordnung $O(\Delta t, \Delta x)$. Durch abermalige Zeitableitung der Advektionsgleichung (6.50) folgt $\partial^2 \bar{T}/\partial t^2 = u^2 \partial^2 \bar{T}/\partial x^2$. Daher kann man den Fehler auch schreiben als

$$\frac{\partial \bar{T}}{\partial t} + u \frac{\partial \bar{T}}{\partial x} = \underbrace{u \frac{\Delta x}{2} (1 - \mathrm{C}) \frac{\partial^2 \bar{T}}{\partial x^2}}_{\text{Diffusion}} + O\left(\Delta t^2, \Delta x^2\right). \tag{6.62}$$

Da der führende Fehlerterm proportional zur zweiten Ortsableitung ist, hat er dieselbe Wirkung wie ein Diffusionsterm. Diese *künstliche Diffusion* wird durch die Diskretisierung verursacht. Die zugehörige Diffusivität ist $\kappa' = u \Delta x (1 - \mathrm{C})/2$. Die diffusive Wirkung des Fehlers kann man sehr gut in Abb. 6.8c,d sehen: Der Gauß-Peak wird im Laufe der zeitlichen Entwicklung abgebaut und die Ecken des Rechteckimpulses werden verschmiert. Für $\mathrm{C} \to 1$ verschwindet die künstliche Diffusion. Dann ist die Lösung des *Upwind*-Schemas ja auch identisch mit der exakten Lösung. Bei Strömungsproblemen kann man leider nicht einfach $\mathrm{C} = 1$ setzen, da die Geschwindigkeit $u(x, t)$ im allgemeinen im Raum variiert. Die Courant-Zahl nimmt dann lokal unterschiedliche Werte an.

6.3.3. Leapfrog-Verfahren

Um ein stabiles Verfahren für die Advektionsgleichung zu erhalten, kann man alternativ versuchen, den zeitlichen Operator des FTCS-Verfahrens zu modifizieren. Es liegt dann nahe, die zeitlichen Differenzen wie beim Richardson-Verfahren für

nannt.

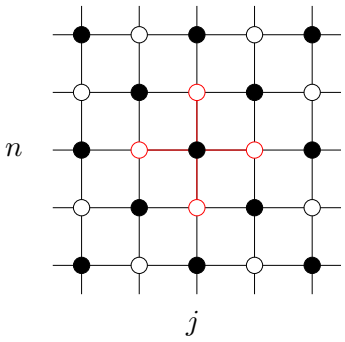


Abbildung 6.9.: Schachbrettmuster der beim Leapfrog-Verfahren miteinander in Verbindung stehenden Punkte.

die Diffusionsgleichung (Kap. 6.1.1) symmetrisch zu bilden, was gleichzeitig die Genauigkeit in der Zeit erhöhen würde. Mit dieser Änderung erhalten wir

$$\frac{T_j^{n+1} - T_j^{n-1}}{2\Delta t} + u\frac{T_{j+1}^n - T_{j-1}^n}{2\Delta x} = 0, \tag{6.63}$$

was auf das *Leapfrog-Verfahren* führt

$$T_j^{n+1} = T_j^{n-1} - \mathrm{C}\left(T_{j+1}^n - T_{j-1}^n\right). \tag{6.64}$$

Wenn man die Konsistenz überprüft, stellt man fest, daß dieses Verfahren von zweiter Ordnung $O\left(\Delta t^2, \Delta x^2\right)$ ist. An dem Verstärkungsfaktor von Störungen

$$G = -\mathrm{iC}\sin\theta \pm \sqrt{1 - \mathrm{C}^2\sin^2\theta} \tag{6.65}$$

sieht man sofort, daß das Leapfrog-Verfahren für $\mathrm{C} = 1$ neutral stabil ist, denn es ist $|G(\mathrm{C} = 1)| = 1$. Das Verfahren ist aber auch für alle $C \leq 1$ neutral stabil, was recht vorteilhaft ist. Denn neutrale Stabilität bedeutet, daß kleine Abweichungen weder verstärkt noch gedämpft werden. Dies entspricht genau dem Verhalten der exakten Advektionsgleichung. Sie ist trivialerweise energieerhaltend.

Wenn man sich das raum-zeitliche Gitter schachbrettartig (schwarz/weiß, Abb. 6.9) vorstellt, erkennt man, daß durch das Leapfrog-Verfahren (6.64) nur die Gitterstellen einer Farbe miteinander in Beziehung stehen. Die weißen Gitterpunkte entwickeln sich völlig unabhängig von den schwarzen. Daher kann es sein, daß sich die Lösungen auf den beiden Teilgittern völlig unterschiedlich entwickeln (Abb. 6.8g–j).[15] Dies kann man vermeiden, wenn die beiden Lösungen in gewissen periodischen Zeitabständen schwach miteinander gekoppelt werden (z.B. durch Mittelung über zwei Zeitschritte).

Man kann analytisch zeigen, daß das Leapfrog-Verfahren (6.64) zwei Lösungen besitzt, eine physikalisch sinnvolle und eine unphysikalische Lösung (*spurious mode*).[16] Die unphysikalische Lösung kann durch eine geeignete Wahl der Anfangsbedingung

[15] In Abb. 6.8 wurden die beiden unabhängigen Funktionswerte mit durchgezogenen beziehungsweise gestrichelten Linien verbunden.

[16] Der Grund dafür sind die drei Zeitniveaus.

auf den beiden ersten Zeitniveaus vermieden werden. Falls die unphysikalische Lösung bei der Implementierung der Anfangsbedingungen nicht eliminiert wird, entwickeln sich die Lösungen auf den weißen und schwarzen Raum-Zeit-Punkten unterschiedlich und die Differenz muß zum Beispiel durch die oben genannte Mittelung ausgeglichen werden.

Beim Leapfrog-Verfahren erfolgt die Berechnung auf 3 Zeitniveaus. Daher braucht man zu Beginn der Rechnung zunächst ein zweites Zeitniveau, das man zum Beipiel mit dem Euler-Verfahren (FTCS) oder dem *Upwind*-Verfahren berechnen kann.

6.3.4. Lax-Wendroff-Verfahren

Eine andere Möglichkeit zur Konstruktion eines Verfahrens zweiter Ordnung in der Zeit besteht darin, den Fehlerterm zweiter Ordnung, welcher die zweite Zeitableitung enthält, durch die zweite Ortsableitung auszudrücken und ins Differenzenschema einzubeziehen. Dazu betrachten wir die Taylorentwicklung von T_j^{n+1} um den Punkt (n, j), aufgelöst nach $\partial\bar{T}/\partial t$

$$\frac{\partial \bar{T}}{\partial t} = \frac{T_j^{n+1} - T_j^n}{\Delta t} - \frac{\Delta t}{2}\frac{\partial^2 \bar{T}}{\partial t^2} + O\left(\Delta t^2\right) \stackrel{!}{=} \frac{T_j^{n+1} - T_j^n}{\Delta t} - u^2\frac{\Delta t}{2}\frac{\partial^2 \bar{T}}{\partial x^2} + O\left(\Delta t^2\right). \quad (6.66)$$

Damit erhält man das *Lax-Wendroff-Verfahren*

$$T_j^{n+1} = T_j^n - \frac{\mathrm{C}}{2}\left(T_{j+1}^n - T_{j-1}^n\right) + \frac{\mathrm{C}^2}{2}\left(T_{j+1}^n - 2T_j^n + T_{j-1}^n\right). \quad (6.67)$$

Es ist von der Fehlerordnung $O\left(\Delta t^2, \Delta x^2\right)$ und stabil für $\mathrm{C} \leq 1$. Für $\mathrm{C} = 1$ wird das Lax-Wendroff-Verfahren exakt. Numerische Ergebnisse sind in Abb. 6.8e,f gezeigt.

Burton Wendroff (Burt) 1930–

Man kann noch viele weitere Schemata untersuchen. Insbesondere kann man auch implizite Verfahren, wie das Crank-Nicolson-Schema, betrachten. Für hyperbolische Gleichungen, wie der Advektionsgleichung, bieten die impliziten Verfahren aber keine wesentlichen Vorteile gegenüber den expliziten. Das liegt daran, daß sich numerische Fehler bei impliziten Verfahren innerhalb nur eines Zeitschritts (also sofort) über das gesamte Raumgebiet ausbreiten. Aus physikalischen Gründen propagieren Störungen bei hyperbolischen Gleichungen aber nur mit endlicher Geschwindigkeit. Obwohl die impliziten Verfahren in der Regel sehr stabil sind (auch für $\mathrm{C} > 1$), produzieren sie bei großen Zeitschritten recht große Fehler. Bei kleinen Zeitschritten ($\mathrm{C} \leq 1$) kann man dann auch die schnelleren expliziten Verfahren verwenden. Im übrigen sollte man nicht vergessen, daß das Upwind- und auch das Lax-Wendroff-Schema für $\mathrm{C} = 1$ die exakte Lösung der Advektionsgleichung reproduzieren. Implizite Methoden sind dazu nicht in der Lage.

Generell kann man sagen, daß es wesentlich einfacher ist, genaue Lösungen für parabolische Gleichungen (Diffusion) zu erhalten, als für hyperbolische Gleichungen (Advektion).

Table 9.1. Algebraic (discretised) schemes for the convection equation $\frac{\partial \bar{T}}{\partial t}+u\frac{\partial \bar{T}}{\partial x}=0$

Scheme	Algebraic form	Truncation error[a] (E) (leading terms)	Amplification factor G ($\theta=m\pi\Delta x$)	Stability restrictions	Remarks
FTCS	$\frac{\Delta T_j^{n+1}}{\Delta t}+uL_x T_j^n=0$	$Cu(\Delta x/2)\frac{\partial^2 T}{\partial x^2}$	$1-\mathrm{i}C\sin\theta$	unstable	$C=u\frac{\Delta t}{\Delta x}$
		$+u\left(\frac{\Delta x^2}{6}\right)(1+2C^2)\frac{\partial^3 T}{\partial x^3}$			$L_x=\frac{1}{2\Delta x}\{-1,0,1\}$
Upwind	$\frac{\Delta T_j^{n+1}}{\Delta t}+u\frac{(T_j^n-T_{j-1}^n)}{\Delta x}=0$	$-u\left(\frac{\Delta x}{2}\right)(1-C)\frac{\partial^2 T}{\partial x^2}$	$1-C(1-\cos\theta)-\mathrm{i}C\sin\theta$	$C\leqq 1$	$\Delta T_j^{n+1}=T_j^{n+1}-T_j^n$
		$+u\left(\frac{\Delta x^2}{6}\right)(1-3C+2C^2)\frac{\partial^3 T}{\partial x^3}$			
Leapfrog	$\frac{T_j^{n+1}-T_j^{n-1}}{2\Delta t}+uL_x T_j^n=0$	$u\left(\frac{\Delta x^2}{6}\right)(1-C^2)\frac{\partial^3 T}{\partial x^3}$	$-\mathrm{i}C\sin\theta\pm(1-C^2\sin^2\theta)^{\frac{1}{2}}$	$C\leqq 1$	

Abbildung 6.10.: **(a)** Verfahren für die Advektionsgleichung (Fletcher, 1991*a*).

Scheme	Algebraic form	Truncation error[a] (E) (leading terms)	Amplification factor G ($\theta = m\pi\Delta x$)	Stabiity restrictions	Remarks.
Lax-Wendroff •—•—• •—•—•	$\dfrac{\Delta T_j^{n+1}}{\Delta t} + uL_x T_j^n - 0.5uC\Delta x L_{xx} T_j^n = 0$	$u\left(\dfrac{\Delta x^2}{6}\right)(1-C^2)\dfrac{\partial^3 T}{\partial x^3} + uC\left(\dfrac{\Delta x^3}{8}\right)(1-C^2)\dfrac{\partial^4 T}{\partial x^4}$	$1 - iC\sin\theta - 2C^2\sin^2\left(\dfrac{\theta}{2}\right)$	$C \leqq 1$	$L_{xx} = \left\{\dfrac{1}{\Delta x^2}, -\dfrac{2}{\Delta x^2}, \dfrac{1}{\Delta x^2}\right\}$
Crank-Nicolson •—•—• •—•—•	$\dfrac{\Delta T_j^{n+1}}{\Delta t} + uL_x\left(\dfrac{T_j^n + T_j^{n+1}}{2}\right) = 0$	$u\left(\dfrac{\Delta x^2}{6}\right)(1+0.5C^2)\dfrac{\partial^3 T}{\partial x^3}$	$\dfrac{(1-0.5iC\sin\theta)}{(1+0.5iC\sin\theta)}$	None	
Three-level fully implicit •—•—• •—•—•	$\dfrac{3}{2}\dfrac{\Delta T_j^{n+1}}{\Delta t} - \dfrac{1}{2}\dfrac{\Delta T_j^n}{\Delta t} + uL_x T_j^{n+1} = 0$	$u\left(\dfrac{\Delta x^2}{6}\right)(1+2C^2)\dfrac{\partial^3 T}{\partial x^3}$	$\dfrac{1 \pm \frac{1}{3}i(3+i8C\sin\theta)^{\frac{1}{2}}}{2\left(1+i\dfrac{2C}{3}\sin\theta\right)}$	None	
Linear F.E.M./ Crank-Nicolson •—•—• •—•—•	$M_x\dfrac{\Delta T_j^{n+1}}{\Delta t} + uL_x\left(\dfrac{T_j^n + T_j^{n+1}}{2}\right) = 0$	$C^2u\left(\dfrac{\Delta x^2}{12}\right)\dfrac{\partial^3 T}{\partial x^3}$	$\dfrac{(2+\cos\theta-1.5iC\sin\theta)}{(2+\cos\theta+1.5iC\sin\theta)}$	None	$M_x = \left\{\dfrac{1}{6}, \dfrac{2}{3}, \dfrac{1}{6}\right\}$

[a] The truncation error (E) has been expressed in terms of Δx and x-derivatives as in the modified equation approach (Sect. 9.2.2). Thus the algebraic scheme is equivalent to $\partial T/\partial t + u\partial T/\partial x + E(T) = 0$.

Noch Abbildung 6.10.: (b) Verfahren für die Advektionsgleichung (Fletcher, 1991*a*).

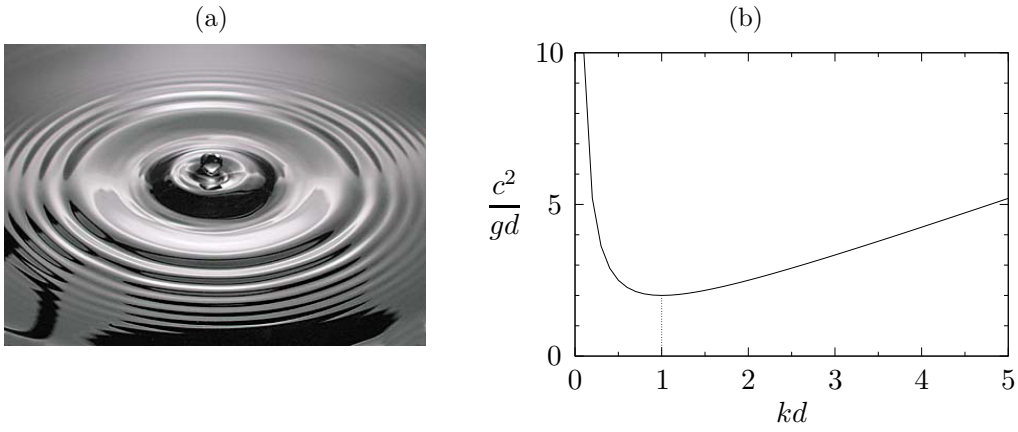


Abbildung 6.11.: (a) Wellen verschiedener Wellenlänge auf einer Wasseroberfläche (Photographie: Andrew Davidhazy) und (b) Dispersionsrelation für Oberflächenwellen für Bondzahl $\mathrm{Bo} = \rho g d^2/\sigma = 1$, wobei d die Längenskala ist. Die Bondzahl ist ein Maß für die relative Bedeutung von Gravitationskräften zu Oberflächenspannungskräften.

6.3.5. Dispersion und Dissipation

Als *Dispersion* bezeichnet man die Tatsache, daß die spektralen Komponenten (Fourierkomponenten) eines Signals unterschiedlich schnell propagieren. Nichtmonochromatische Wellen ändern sich deshalb mit der Zeit. Neben der reinen Formänderung wird die Abschwächung eines Signals als *Dissipation* bezeichnet.[17]

Die Dispersion kann man zum Beispiel beobachten, wenn man einen Stein ins Wasser wirft. Für Wasserwellen unter dem Einfluß von Kapillarität und Schwerkraft (*capillary-gravity waves*) gilt unter Vernachlässigung der Dichte der Luft die *Dispersionsrelation* (Landau and Lifschitz, 1991)

$$c^2 = \left(\frac{\omega}{k}\right)^2 = \frac{g}{k} + \frac{\sigma k}{\rho}. \tag{6.68}$$

Für Wasser/Luft hat die Phasengeschwindigkeit c ein Minimum bei $k_{\min} = \sqrt{\rho g/\sigma} \approx 3.66\,\mathrm{cm}^{-1}$ ($\sigma_{\mathrm{H_2O}} = 72.8 \times 10^{-5}\,\mathrm{N/cm}$, $\rho_{\mathrm{H_2O}} = 1\,\mathrm{g/cm}^3$, $g = 9.8\,\mathrm{m/s}^2$). Dies entspricht einer Wellenlänge von $\lambda_{\min} \approx 1.71\,\mathrm{cm}$. Für Wellenlängen $\lambda < \lambda_{\min}$ dominieren kapillare Effekte und man hat Kapillarwellen. Im Bereich der Kapillarwellen nimmt die Phasengeschwindigkeit mit der Wellenlänge ab. Dies ist der normale Fall bei einem Steinwurf ins Wasser. Für Wellen mit $\lambda > \lambda_{\min}$ nimmt die Phasengeschwindigkeit mit der Wellenlänge langsam wieder zu. In diesem Bereich hat man Schwerewellen. In der Seefahrt sind die sehr langwelligen *freak waves* so gefährlich, weil sie am schnellsten propagieren.

[17] Die Dissipation im strengen strömungsmechanischen Sinn bezeichnet die Umwandlung von kinetischer Energie in Wärmeenergie. Sie ist mathematisch genau definiert.

Ausbreitung von Wellen in linearer Näherung

Wenn Wellen eine kleine Amplitude besitzen, kann man die zugrundeliegenden Gleichung bezüglich der Amplitude linearisieren. Dies ermöglicht eine analytische Behandlung. Dispersion und Dissipation sollen daher an zwei Beispielen für die lineare Wellenausbreitung demonstriert werden. Wir betrachten einerseits die *Konvektions-Diffusions-Gleichung* (Transportgleichung)

Diederik Johannes Korteweg 1848–1941

$$\frac{\partial T}{\partial t} + u\frac{\partial T}{\partial x} - \kappa\frac{\partial^2 T}{\partial x^2} = 0, \tag{6.69}$$

und andererseits die linearisierte *Korteweg-De Vries-Gleichung*

$$\frac{\partial T}{\partial t} + u\frac{\partial T}{\partial x} + \beta\frac{\partial^3 T}{\partial x^3} = 0. \tag{6.70}$$

In beiden Fällen können wir die Lösung als Fourier-Reihe darstellen

$$T(x,t) = \sum_k \hat{T}_k \mathrm{e}^{\mathrm{i}kx+(\sigma-\mathrm{i}\omega)t} + \text{c.c.}, \tag{6.71}$$

mit $k, \sigma, \omega \in \mathbb{R}$ und c.c. als Bezeichnung für das konjugiert Komplexe. Da die linearen Wellen entkoppeln, brauchen wir nur eine repräsentative Fourier-Komponente zu betrachten.[18]

Gustav de Vries 1866–1934

Wenn wir diesen Ansatz in die obigen Differentialgleichungen einsetzen, erhalten wir für die Transportgleichung

$$\sigma - \mathrm{i}\omega + \mathrm{i}uk + \kappa k^2 = 0 \qquad \Rightarrow \qquad \begin{cases} \sigma(k) &= -\kappa k^2, \\ \omega(k) &= uk. \end{cases} \tag{6.72}$$

Die *Wachstumsrate* $\sigma \leq 0$ ist negativ. Je kleiner die Wellenlänge ist, desto größer ist die *Dämpfungsrate* $-\sigma$ (Dissipation). Die Frequenz der Welle ist proportional zur Wellenzahl. Damit ist die *Phasengeschwindigkeit* $c = \omega/k = u$ konstant und unabhängig von k. Dies wird auch oft als *Dispersionsfreiheit* bezeichnet. Die Welle wird also im Laufe der Zeit nur abgeschwächt.

Für die linearisierte Korteweg-De Vries-Gleichung erhalten wir

$$\sigma - \mathrm{i}\omega + \mathrm{i}uk - \mathrm{i}\beta k^3 = 0 \qquad \Rightarrow \qquad \begin{cases} \sigma(k) &= 0, \\ \omega(k) &= uk - \beta k^3. \end{cases} \tag{6.73}$$

[18] Eine mögliche Form der Korteweg-De Vries-Gleichung ist

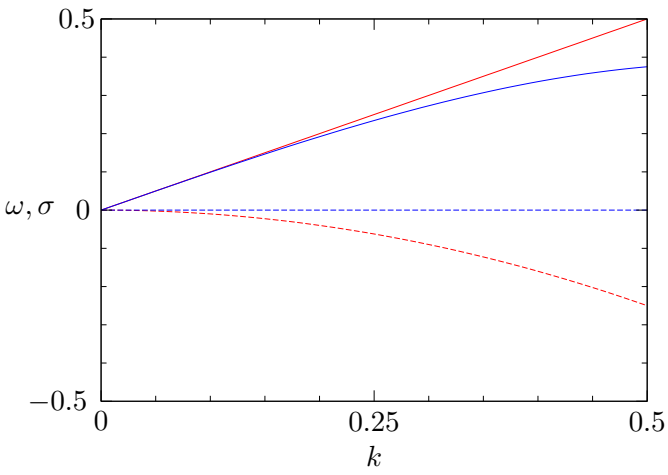


Abbildung 6.12.: Dispersion (ω, durchgezogen) und Dissipation (σ, gestrichelt) für die Transportgleichung (rot) und die linearisierte KdV-Gleichung (blau). Es wurde $u = \kappa = \beta = 1$ gesetzt.

Wellen erfahren hier keine Verstärkung/Dämpfung: $\sigma = 0$. Die lineare KdV-Gleichung besitzt jedoch eine kubische Dispersion. Die Welle zerfließt. Die entsprechenden Kurven sind in Abb. 6.12 gezeigt.

> Aufgrund des Zusammenhangs $\partial_x \to \mathrm{i}k$ tragen die geradzahligen Ortsableitungen immer zur Dissipation, ungeradzahlige Orstableitungen immer zur Dispersion bei.

In Abb. 6.13 ist die Wirkung von Dispersion und Dissipation auf einen Rechteckpuls gezeigt. Angenommen wurden die Relationen (6.72) und (6.73) mit verschiedenen Werten für β und κ.

In analoger Weise kann man die Dispersion und Dissipation der diskretisierten Form der Gleichungen untersuchen. Die Dispersion und Dissipation sollten derjenigen der exakten Gleichungen möglichst nahekommen. Wir wollen dies am Beispiel des Leapfrog-Verfahrens demonstrieren.

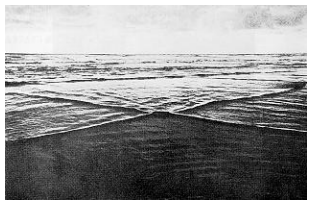
Dispersion und Dissipation des Leapfrog-Verfahrens

Es ist leicht zu sehen, daß die Advektionsgleichung keine Dispersion besitzt, d.h. alle Fourierkomponenten einer beliebigen Anfangsbedingung propagieren mit derselben Phasengeschwindigkeit u. Auch ist die Advektionsgleichung frei von Dissipation, die Amplituden der Fourierkomponenten sind ungedämpft. Um das Verhalten des

---

$$\frac{\partial f}{\partial t} + (1+f)\frac{\partial f}{\partial x} + \frac{\partial^3 f}{\partial x^3} = 0.$$

Sie beschreibt die Ausbreitung nichtlinearer Wellen in einer flachen Flüssigkeitsschicht (Flachwasser). Man kann die KdV-Gleichung auch noch in verschiedenen anderen Formen aufschreiben (Drazin, 1983). Insbesondere erlaubt sie trotz der Dispersion Wellen, die ihre Form nicht verändern (Solitonen und Cnoidal-Wellen). Hierbei wird die Dispersion gerade durch die Nichtlinearität kompensiert.



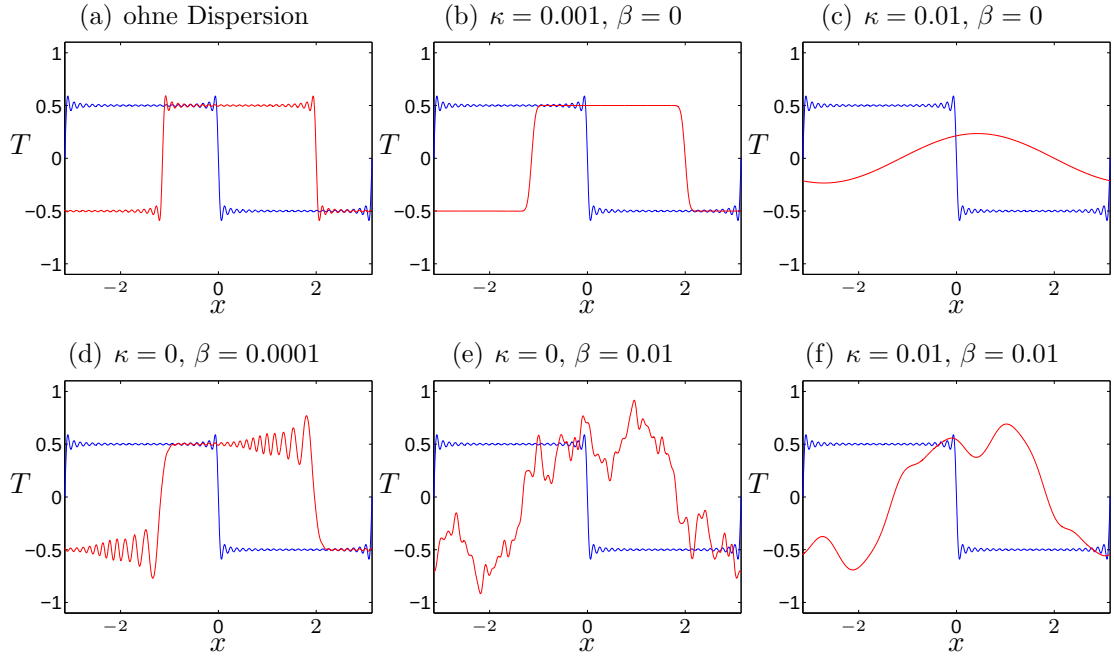


Abbildung 6.13.: Zeitliche Entwicklung eines 2π-periodischen Rechteckpulses nach (6.71) unter dem Einfluß von Disperion $\omega = uk - \beta k^3$ (6.73) und von Dissipation $\sigma = -\kappa k^2$ (6.72). Der anfängliche Puls (blaue Kurve) wurde durch $T(x, t = 0) = \Re[\mathrm{i} \sum_{m=1}^{49} \exp(\mathrm{i}mx)/m]$ (m ungerade) approximiert. Die roten Kurven zeigen den Puls nach $\Delta t = 2$ und für $u = 1$. Alle anderen Parameter sind in den Überschriften angegeben.

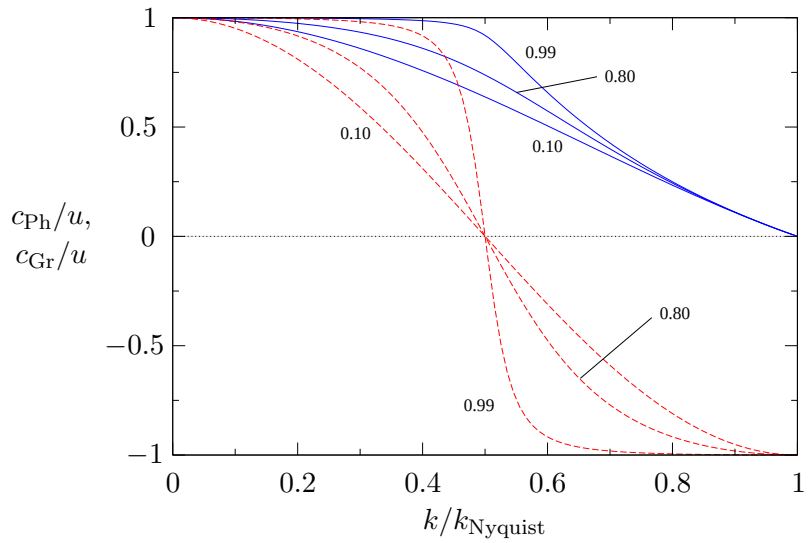


Abbildung 6.14.: Dispersion des Leapfrog-Verfahrens (6.64). Gezeigt ist die normierte Phasengeschwindigkeit c_{Ph} (blau) und die normierte Gruppengeschwindigkeit c_{Gr} (rot). Die Courant-Zahl C ist als Parameter angegeben. Die Wellenzahlen sind durch die Grenzwellenzahl $k_{\mathrm{Nyquist}} = \pi/\Delta x$ skaliert.

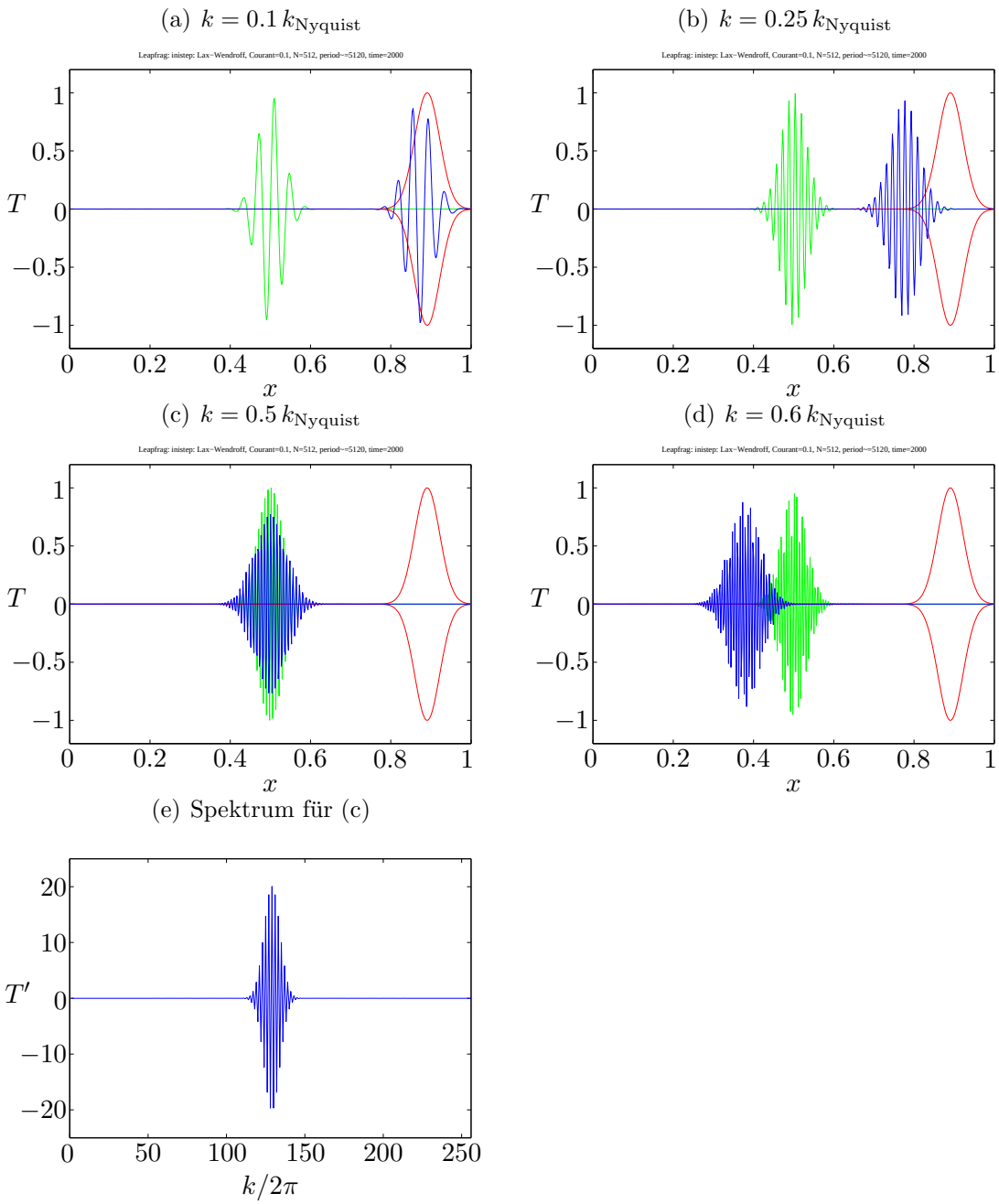


Abbildung 6.15.: Advektion eines Wellenpakets bei periodischen Randbedingungen mittels Leapfrog-Verfahren und C = 0.1. Die Anfangsbedingung (grün) ist $T_0 = \exp\left[-N(x-1/2)^2\right]\sin\left[sk_{\text{Nyquist}}(x-1/2)\right]$, mit Nyquist-Wellenzahl $k_{\text{Nyquist}} = N\pi = \pi/\Delta x$ (vgl. Abb. 6.14). Der erste Zeitschritt wurde mit Lax-Wendroff durchgeführt. Die Auflösung beträgt $N = 512$. Die Einhüllende der exakten Lösung (rot) und die numerische Lösung (blau) sind zur Zeit $n = 2000$ gezeigt. Die Gruppengeschwindigkeit hängt von der Wellenzahl ab: $s = 0.1$ (a), $s = 0.25$ (b), $s = 0.5$ (c) und $s = 0.6$ (d). (e) zeigt das Spektrum für den Fall (c).

Leapfrog-Schemas zu untersuchen, setzen wir den Ansatz (6.71) in (6.64) ein und erhalten für die Welle $T_j^n \sim \mathrm{e}^{\mathrm{i}kx_j+(\sigma-\mathrm{i}\omega)t_n}$

$$\mathrm{e}^{(\sigma-\mathrm{i}\omega)\Delta t} = \mathrm{e}^{-(\sigma-\mathrm{i}\omega)\Delta t} - \mathrm{C}\left(\mathrm{e}^{\mathrm{i}k\Delta x} - \mathrm{e}^{-\mathrm{i}k\Delta x}\right). \tag{6.74}$$

Trennen von Real und Imaginärteil liefert[19]

$$\mathrm{e}^{\sigma\Delta t}\cos(\omega\Delta t) = \mathrm{e}^{-\sigma\Delta t}\cos(\omega\Delta t), \tag{6.75a}$$
$$-\mathrm{e}^{\sigma\Delta t}\sin(\omega\Delta t) = \mathrm{e}^{-\sigma\Delta t}\sin(\omega\Delta t) - 2\mathrm{C}\sin(k\Delta x). \tag{6.75b}$$

Harry Nyquist
1889–1976

Aus dem Realteil folgt $\sigma = 0$. Das Leapfrog-Schema besitzt also keine Dissipation, genau wie die Advektionsgleichung. Mit $\sigma = 0$ folgt aus dem Imaginärteil

$$\sin(\omega\Delta t) = \mathrm{C}\sin(k\Delta x). \tag{6.76}$$

Die Dispersionsrelation lautet also

$$\omega(k) = \frac{1}{\Delta t}\arcsin\left[\mathrm{C}\sin(k\Delta x)\right]. \tag{6.77}$$

Daraus folgt für die Phasen- (c_{Ph}) und für die Gruppengeschwindigkeit (c_{Gr})[20]

$$c_{\mathrm{Ph}} = \frac{\omega}{k} = \frac{1}{k\Delta t}\arcsin\left[\mathrm{C}\sin(k\Delta x)\right] \overset{\mathrm{C}\to 1}{\longrightarrow} \frac{\Delta x}{\Delta t} \overset{\mathrm{C}=1}{=} u, \tag{6.78a}$$
$$c_{\mathrm{Gr}} = \frac{\partial\omega}{\partial k} = \frac{\Delta x}{\Delta t}\frac{\mathrm{C}\cos(k\Delta x)}{\sqrt{1-\mathrm{C}^2\sin^2(k\Delta x)}} \overset{\mathrm{C}\to 1}{\longrightarrow} \frac{\Delta x}{\Delta t} \overset{\mathrm{C}=1}{=} u. \tag{6.78b}$$

Für $\mathrm{C} \neq 1$ hat das Leapfrog-Verfahren eine Dispersion, die nicht in der Advektionsgleichung vorhanden ist. Diese *numerische Dispersion* ist in Abb. 6.14 durch Kurven für Phasen- und Gruppengeschwindigkeit illustriert. Die kleinste auf einem Gitter darstellbare Wellenlänge ist $\lambda_{\min} = 2\Delta x$. Dies entspricht einer maximalen Wellenzahl, der *Nyquistwellenzahl* $k_{\mathrm{Nyquist}} = \pi/\Delta x$. Für $\mathrm{C} \to 1$ fällt die Phasengeschwindigkeit für $k/k_{\mathrm{Nyquist}} = k\Delta x/\pi > 1/2$ stark ab und verschwindet für $k = k_{\mathrm{Nyquist}}$. Für Wellenlängen, die kleiner sind als $\lambda < 4\Delta x$, wird die Gruppengeschwindigkeit sogar negativ. Dieses Verhalten ist konkret in Abb. 6.15 gezeigt. Die langwelligen Strukturen werden gut transportiert. Wenn sich die Wellenzahl der Trägerwelle erhöht, bleibt das Wellenpaket aber hinter der exakten Lösung zurück. Bei der halben Nyquist-Wellenzahl bleibt das Wellenpaket sogar stehen. Eine weitere Erhöhung von k führt dann zu einer Propagation in die entgegengesetzte Richtung.

[19] Es gilt $\mathrm{e}^{\mathrm{i}\varphi} = \cos\varphi + \mathrm{i}\sin\varphi$.

[20] Beachte: $\arcsin'(x) = \left(1-x^2\right)^{-1/2}$.

6.3.6. Erhaltungsgrößen

Meist gibt es bei partiellen Differentialgleichungen gewisse Größen, im Laufe der zeitlichen Entwicklung erhalten bleiben. Die diskrete Version der PDE sollte natürlich auch möglichst viele dieser Größen erhalten. Dies ist besonders wichtig bei der Advektionsgleichung.

Eine offensichtliche Erhaltungsgröße der Advektionsgleichung ist die Fläche unter der Funktion $T(x,t)$. Wir können diese als *Masse* auffassen. Als Beispiel betrachten wir das FTCS-Verfahren (6.52). Wenn wir das Integral mittels Trapezregel approximieren, erhalten wir (den gemeinsamen Faktor Δx können wir kürzen)

$$\sum_j T_j^{n+1} = \sum_j T_j^n - \frac{\mathrm{C}}{2}\sum_j \left(T_{j+1}^n - T_{j-1}^n\right) = \sum_j T_j^n - \frac{\mathrm{C}}{2}\underbrace{\left(\sum_{j'} T_{j'}^n - \sum_{j''} T_{j''}^n\right)}_{=0}. \tag{6.79}$$

Damit bleibt $\sum_j T_j^n$ von einem Zeitniveau zum anderen konstant. Das FTCS-Verfahren ist als massenerhaltend.

Natürlich kann es noch weitere Erhaltungsgrößen geben (zum Beispiel den Impuls). Bei der KdV-Gleichung existieren sogar unendlich viele Erhaltungsgrößen, deren Erhaltung man beim Diskretisieren jeweils überprüfen sollte.

6.4. Lineare Konvektions-Diffusionsgleichungen

6.4.1. Stationäre Konvektion und Diffusion

In vielen strömungsmechanischen Problemen treten *Grenzschichten* auf. Die mathematische Ursache hierfür ist das Auftreten eines kleinen Koeffizienten vor der höchsten Ableitung in der Differentialgleichung. Beim Impuls- bzw. Wärmetransport sind dies die Diffusivitäten von Impuls (ν) bzw. Wärme (κ). Wenn diese klein sind, ist die Diffusion in der Nähe derjenigen Ränder besonders groß, an denen der Impuls (Geschwindigkeit) bzw. die Temperatur aufgeprägt werden.

Als einfachstes Beispiel für ein derartiges Verhalten kann man die *stationäre Konvektions-Diffusions-Gleichung*

$$u\frac{\mathrm{d}\bar{T}}{\mathrm{d}x} = \kappa\frac{\mathrm{d}^2\bar{T}}{\mathrm{d}x^2} \tag{6.80}$$

betrachten. Hierbei steht der konvektive Transport von T in positiver x-Richtung ($u > 0$, linke Seite) im Gleichgewicht mit dem diffusiven Transport von T (rechte Seite). Für die Randbedingungen $\bar{T}(0) = 0$ und $\bar{T}(1) = 1$ kann man leicht eine analytische Lösung dieser linearen gewöhnlichen Differentialgleichung finden.[21] Mit

[21] Diese Randbedingungen entsprechen der homogene Durchströmung eines Gebiets, bei dem die permeablen Wände bei $x = 0, 1$ auf verschiedenen Temperaturen gehalten werden.

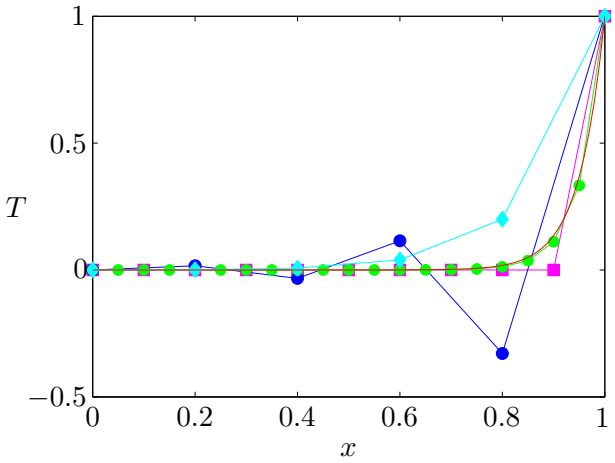


Abbildung 6.16.: Lösungen der stationären Konvektions-Diffusions-Gleichung für $u/\kappa = 20$. Gezeigt sind die exakte Lösung (rot) und numerische Lösungen mittels zentralen Differenzen auf Gittern mit $\Delta x = 0.05$ ($\mathrm{Re_G} = 1$, grün), $\Delta x = 0.1$ ($\mathrm{Re_G} = 2$, magenta) und $\Delta x = 0.2$ ($\mathrm{Re_G} = 4$, blau). Zusätzlich ist auch die Lösung für $\Delta x = 0.2$ ($\mathrm{Re_G} = 4$, cyan) dargestellt, wobei der advektive Terme mit Upwind-Verfahren diskretisiert wurde.

$\bar{T} \sim \mathrm{e}^{\lambda x}$ erhalten wir

$$u\lambda - \kappa\lambda^2 = 0, \tag{6.81}$$

mit den Wurzeln $\lambda_1 = 0$ und $\lambda_2 = u/\kappa$. Die Linearkombination der beiden Moden $\mathrm{e}^{\lambda_i x}$, welche die beiden Randbedingungen erfüllt, lautet

$$\bar{T}(x) = \frac{\mathrm{e}^{(u/\kappa)x} - 1}{\mathrm{e}^{u/\kappa} - 1}. \tag{6.82}$$

Diese exakte Lösung ist in Abb. 6.16 als rote Kurve gezeigt.

Bei Verwendung von zentralen Differenzen für dieses elliptische Problem erhalten wir

$$u\frac{T_{j+1} - T_{j-1}}{2\Delta x} = \kappa\frac{T_{j+1} - 2T_j + T_{j-1}}{\Delta x^2}. \tag{6.83}$$

Dies kann man auch schreiben als

$$\frac{1}{2}\underbrace{\frac{u\Delta x}{\kappa}}_{:=\mathrm{Re_G}}(T_{j+1} - T_{j-1}) = T_{j+1} - 2T_j + T_{j-1}, \tag{6.84}$$

bzw.

$$\left(1 + \frac{\mathrm{Re_G}}{2}\right)T_{j-1} - 2T_j + \left(1 - \frac{\mathrm{Re_G}}{2}\right)T_{j+1} = 0, \tag{6.85}$$

wobei

$$\mathrm{Re_G} = \frac{u\Delta x}{\kappa} \tag{6.86}$$

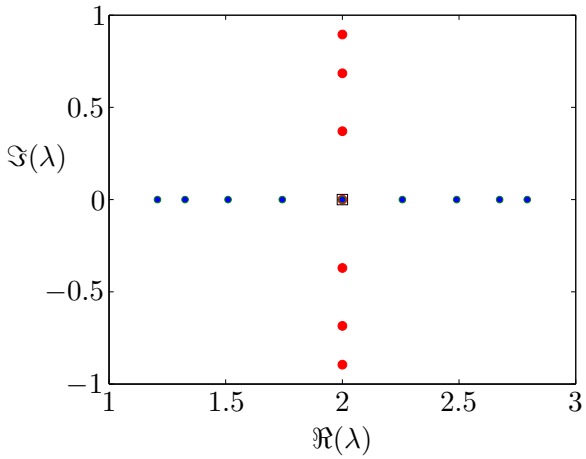


Abbildung 6.17.: Eigenwerte λ der Matrix des linearen Systems (6.85) für $u/\kappa = 20$. Für die unterkritische Bedingung $N = 11$ (blau) sind alle Eigenwerte reell. Bei der kritischen Bedingung $N = 10$ entsprechend $\mathrm{Re_G} = 2$ (schwarzes Quadrat) sind alle Eigenwerte entartet und es sind alle $\lambda_j = 2$. Für überkritische Verhältnisse $N = 9$ (rot) sind alle Eigenwerte komplex.

die *Gitter-Reynoldszahl* ist.[22] Man kann leicht nachprüfen, daß die Rekursionsformel (6.85) die Lösung

$$T_j = A + B\left(\frac{1 + \mathrm{Re_G}/2}{1 - \mathrm{Re_G}/2}\right)^j \tag{6.87}$$

besitzt. An der Form dieser exakten Lösung der diskretisierten Gleichung sieht man, daß T_j oszilliert, wenn $\mathrm{Re_G} > 2$ ist.[23]

Aber man kann auch das lineare Gleichungssystem für T_j aufstellen, das (6.85) entspricht. Dann erhält man

$$\begin{pmatrix} b & c & & & \\ a & b & c & & \\ & \cdot & \cdot & \cdot & \\ & & a & b & c \\ & & & a & b \end{pmatrix} \cdot \begin{pmatrix} T_2 \\ T_3 \\ \dots \\ T_{J-2} \\ T_{J-1} \end{pmatrix} = \begin{pmatrix} -aT_1 \\ 0 \\ \dots \\ 0 \\ -cT_J \end{pmatrix}, \tag{6.88}$$

mit $a = -(1 + \mathrm{Re_G}/2)$, $b = 2$ und $c = -(1 - \mathrm{Re_G}/2)$. Die Eigenwerte der Tridiagonalmatrix lassen sich sofort angeben (siehe Anhang A)

$$\lambda_j = b + 2\sqrt{ac}\,\cos\left(\frac{\pi j}{J-1}\right), \quad j = 1, ..., J-2. \tag{6.89}$$

[22] Dies ist eine Reynoldszahl, die mit der Gitterweite Δx als Längenskala gebildet wurde. Eigentlich ist es eine Gitter-Peclet-Zahl, da wir die nicht weiter spezifizierte Diffusivität κ verwenden, und nicht etwa die kinematische Viskosität ν.

[23] Der Grenzfall $\mathrm{Re_G} = 2$ ist singulär. Dann fällt T_{j+1} aus (6.85) heraus und es bleibt $T_j = 0$, bis auf den Randwert $T_N = 1$.

Falls $ac \leq 0$ wird, werden alle Eigenwerte komplex (Abb. 6.17). Der Einsatz oszillatorischen Verhaltens für

$$\mathrm{Re_G} > 2 \tag{6.90}$$

fällt offenbar mit dem Auftreten von komplexen Eigenwerten zusammen.

Die Bedingung $\mathrm{Re_G} \leq 2$ an die Gitter-Reynoldszahl zur Vermeidung von oszillatorischem Verhalten kann man auch bei vielen anderen Methoden für Grenzschichtprobleme finden. Es ist aber nicht allgemein zwingend, daß für $\mathrm{Re_G} > 2$ die numerische Lösung oszilliert. Sie oszilliert zwar für die abgebremste Strömung *vor* einem Hindernis, nicht aber z.B. für die beschleunigte Strömung *hinter* einem Hindernis.

Wenn man die zentralen Differenzen für die erste Ableitung in (6.85) durch unsymmetrische Differenzen ersetzt, erhält man das Upwind-Verfahren

$$\mathrm{Re_G}\,(T_j - T_{j-1}) = T_{j+1} - 2T_j + T_{j-1}, \tag{6.91}$$

und somit

$$-(1+\mathrm{Re_G})T_{j-1} + 2\left(1 + \frac{\mathrm{Re_G}}{2}\right)T_j - T_{j+1} = 0. \tag{6.92}$$

Mit $a = -(1+\mathrm{Re_G})$, $b = 2(1+\mathrm{Re_G}/2)$ und $c = -1$ in (6.88) hat die Matrix nur reelle Eigenwerte und wir erwarten die Abwesenheit von Oszillationen, was durch die Rechnung bestätigt wird. Das Verfahren ist dann jedoch nur noch von erster Ordnung und deshalb nicht sehr genau (siehe Abb. 6.16).

Die Taylor-Entwicklung der exakten Lösung um den Punkt j zeigt, daß der in der ersten Ableitung beim Upwind-Schema vernachlässigte Term ein diffusiver Term ist. Das Upwind-Schema (6.92) entspricht dann einem Differenzenschema zweiter Ordnung der Differentialgleichung

$$u\frac{\mathrm{d}\bar{T}}{\mathrm{d}x} = \kappa\left(1 + \frac{\mathrm{Re_G}}{2}\right)\frac{\mathrm{d}^2\bar{T}}{\mathrm{d}x^2}. \tag{6.93}$$

Daraus kann man schließen, daß das Upwind-Schema erster Ordnung eine künstliche Diffusivität der Größe $\kappa_{\mathrm{Diff}} = (\mathrm{Re_G}/2)\kappa$ besitzt (vgl. auch (6.62)). Dies erklärt auch den Verlauf der numerischen Lösung (cyan) in Abb. 6.16, der wesentlich flacher ist als derjenige der exakten Lösung.

6.4.2. Zeitabhängige Konvektion und Diffusion

Als nächstes betrachten wir die sogenannte *Transportgleichung*

$$\frac{\partial\bar{T}}{\partial t} + u\frac{\partial\bar{T}}{\partial x} = \kappa\frac{\partial^2\bar{T}}{\partial x^2}. \tag{6.94}$$

Die Größe $\bar{T}$ (z.B. die Temperatur) wird durch das Geschwindigkeitsfeld u transportiert und sie diffundiert. Für $\kappa = 0$ hat man die Advektionsgleichung und $\bar{T}$ wäre

dann ein *passiver Skalar*.[24] Diese Art von Gleichung ist typisch für den Transport von Temperatur (Wärme) und Konzentrationen (Stoffen). Sie ist parabolisch, und man muß Anfangs- und Randbedingungen vorgeben.

Jean Claude Eugène Péclet 1793–1857

Die mit der Längenskala L, Geschwindigkeitsskala U und der Zeitskala L/U dimensionslos gemachte Gleichung lautet, wenn wir dieselben Symbole für die dimensionslosen Variablen verwenden,

$$\frac{\partial \bar{T}}{\partial t} + u\frac{\partial \bar{T}}{\partial x} = \frac{1}{\mathrm{Pe}}\frac{\partial^2 \bar{T}}{\partial x^2}, \tag{6.95}$$

wobei $\mathrm{Pe} = UL/\kappa$ die *Peclet-Zahl* ist. Für $\mathrm{Pe} \gg 1$ dominiert der konvektive Transport (U) über den diffusiven Transport (κ/L). Dann erwarten wir ein ähnliches Verhalten wie bei der Advektionsgleichung (Wellencharakter) sowie ein Grenzschichtverhalten. Umgekehrt sollte für $\mathrm{Pe} \ll 1$ das Verhalten von $\bar{T}$ nahezu diffusiv sein.

Explizite Verfahren

FTCS Die einfachste Diskretisierung ist das FTCS-Verfahren. Angewandt auf (6.94) und mit $u = \text{const.}$ lautet es

$$\frac{T_j^{n+1} - T_j^n}{\Delta t} + u\frac{T_{j+1}^n - T_{j-1}^n}{2\Delta x} - \kappa\frac{T_{j+1}^n - 2T_j^n + T_{j-1}^n}{\Delta x^2} = 0. \tag{6.96}$$

Auflösen nach T_j^{n+1} ergibt

$$\begin{aligned} T_j^{n+1} &= T_j^n - u\Delta t\frac{T_{j+1}^n - T_{j-1}^n}{2\Delta x} + \kappa\Delta t\frac{T_{j+1}^n - 2T_j^n + T_{j-1}^n}{\Delta x^2} \\ &= \left(s - \frac{\mathrm{C}}{2}\right)T_{j+1}^n + (1-2s)\,T_j^n + \left(s + \frac{\mathrm{C}}{2}\right)T_{j-1}^n, \end{aligned} \tag{6.97}$$

mit $s = \kappa\Delta t/\Delta x^2$ und $\mathrm{C} = u\Delta t/\Delta x$. Die Gitter-Reynoldszahl ist von diesen beiden Größen abhängig: $\mathrm{Re_G} = C/s$. Durch Taylor-Entwicklung kann man leicht prüfen, daß dieses Verfahren von der Ordnung $O(\Delta t, \Delta x^2)$ ist. Der führende Fehlerterm ist $\frac{1}{2}\Delta t\partial^2\bar{T}/\partial t^2$. Das heißt, daß durch das FTCS-Verfahren (6.97) die Gleichung

$$\frac{\partial \bar{T}}{\partial t} + u\frac{\partial \bar{T}}{\partial x} - \kappa\frac{\partial^2 \bar{T}}{\partial x^2} + \frac{\Delta t}{2}\frac{\partial^2 \bar{T}}{\partial t^2} = 0 \tag{6.98}$$

in Ordnung $O\left(\Delta t^2, \Delta x^2\right)$ gelöst wird. Man kann nun die zweite zeitliche Ableitung mit dem kleinen Vorfaktor durch räumliche Ableitungen ausdrücken, indem man die Transportgleichung (6.95) nach der Zeit ableitet und in den gemischten Ableitungen

[24] Die häufig zur Visualisierung von Strömungen eingesetzten kleinen Partikel sollten idealerweise diese Eigenschaft besitzen.

$\partial\bar{T}/\partial t$ entsprechend (6.95) ersetzt. Dann erhält man

$$\frac{\partial\bar{T}}{\partial t}+u\frac{\partial\bar{T}}{\partial x}-\left(\kappa-\frac{\Delta t u^2}{2}\right)\frac{\partial^2\bar{T}}{\partial x^2}-\Delta t\kappa u\frac{\partial^3\bar{T}}{\partial x^3}+\frac{\Delta t\kappa^2}{2}\frac{\partial^4\bar{T}}{\partial x^4}+O\left(\Delta t^2\right)=0. \quad (6.99)$$

Die im Vergleich zur Transportgleichung (6.94) auftretenden Terme aus dem führenden Fehler der zeitlichen Diskretisierung sind rot dargestellt. Man sieht, daß das FTCS-Schema eine verfälschte (verringerte) Diffusion liefert. Außerdem wird ein dispersiver Term ($\sim \partial^3\bar{T}/\partial x^3$) generiert und eine *Diffusion höherer Ordnung* ($\sim \partial^4\bar{T}/\partial x^4$). Um diese Effekte zu minimieren sollte die Zeitschrittweite möglichst gering sein, genauer gesagt $\Delta t u^2/2 \ll \kappa$ bzw. $\mathrm{C}^2 \ll 2s$.

Eine Stabilitätsanalyse zeigt, daß genau dann $|G| \leq 1$ ist (siehe Abb. 6.18), wenn

$$\underbrace{\mathrm{C}^2}_{\sim\Delta t^2} \leq \underbrace{2s}_{\sim\Delta t} \leq 1. \quad (6.100)$$

Damit wird die Stabilitätsgrenze des FTCS-Verfahrens für die Diffusionsgleichung reproduziert. Für die reine Advektion war FTCS immer instabil. Im Fall der Transportgleichung erhalten wir zumindest ein stabiles Verfahren, wenn C hinreichend klein ist. Die Stabilitätsbedingung kann man auch durch die Gitter-Reynoldszahl ausdrücken

$$\mathrm{Re_G}=\frac{u\Delta x}{\kappa}=\frac{\mathrm{C}}{s}\leq\frac{2}{\mathrm{C}} \quad \left(\overset{\mathrm{C}\leq 1}{\geq} 2\right). \quad (6.101)$$

DuFort-Frankel Um eine höhere Genauigkeit in der Zeit zu erhalten, sind auch symmetrische Differenzen bezüglich t von Interesse. Wenn man das Richardson-Verfahren mit Zeitniveaus bei $n+1$ und $n-1$ (siehe (6.8)) auf die Transportgleichung anwendet, erhält man wie bei der Diffusiongsgleichung ein instabiles Verfahren für $s > 0$. Für $s = 0$ mit $\mathrm{C} \neq 0$ wird die Diffusion vollständig verhindert man erhält in diesem Limes das Leapfrog-Verfahren (6.63) für die Advektionsgleichung.

Abhilfe schafft eine Modifikation des Richardson-Verfahrens zum DuFort-Frankel-Verfahren, bei welchem T_j^n im diffusiven Term durch den Mittelwert $\left(T_j^{n+1}+T_j^{n-1}\right)/2$ ersetzt wird, wie beim DuFort-Frankel-Verfahren für die reine Diffusionsgleichung. Dann erhält man

$$\frac{T_j^{n+1}-T_j^{n-1}}{2\Delta t}=-u\frac{T_{j+1}^n-T_{j-1}^n}{2\Delta x}+\kappa\frac{T_{j+1}^n-\left(T_j^{n+1}+T_j^{n-1}\right)+T_{j-1}^n}{\Delta x^2}. \quad (6.102)$$

Solange $\mathrm{C} \leq 1$ ist, ist das DuFort-Frankel-Verfahren für alle s stabil. Da der Abbruchfehler von der Größenordnung $O(\mathrm{C}^2)$ ist, muß die Zeitschrittweite hinreichend klein gewählt werden, $\kappa\Delta t^2 \ll \Delta x^2$ bzw. $\mathrm{C}^2 \ll 1$ (bei $\kappa, u = O(1)$), damit das Verfahren konsistent ist.

Upwind Das Upwind-Verfahren ($u > 0$)

$$T_j^{n+1}=sT_{j+1}^n+(1-2s-\mathrm{C})\,T_j^n+(s+\mathrm{C})\,T_{j-1}^n \quad (6.103)$$

besitzt nur eine Genauigkeit von $O(\Delta t, \Delta x)$. Es hat dieselbe künstliche Diffusion wie das Upwind-Verfahren (6.62) für die Advektionsgleichung. Die zusätzliche Diffusivität ist $u\Delta x\,(1-\mathrm{C})\,/2$. Für hinreichend genaue Lösungen muß diese Diffusivität klein sein gegenüber der wirklichen Diffusivität κ. Diese Bedingung entspricht

$$\mathrm{Re_G} = \frac{u\Delta x}{\kappa} \ll \frac{2}{1-\mathrm{C}}. \tag{6.104}$$

Die Stabilitätsanalyse liefert

$$2s + \mathrm{C} \leq 1, \tag{6.105}$$

was eine wesentlich stärkere Einschränkung der Zeitschrittweite darstellt als für die Diffusionsgleichung ($2s \leq 1$).

Auch diese Stabilitätsgrenze bestätigt wieder das heuristische Argument, das wir schon in Kap. 6.1.1 benutzt hatten, wonach kein Koeffizient auf der rechten Seite negativ werden sollte, wenn Stabilität gewährleistet sein soll. Denn auch eine positive Störung von T^n_{j-1} oder T^n_j im Advektionsterm sollte bei $u > 0$ (Upwind) zu einer Erhöhung von T^{n+1}_j führen.[25]

Lax-Wendroff Um ein Verfahren zweiter Ordnung zu erhalten, könnte man versuchen, wie bei der Advektionsgleichung in (6.67) die zweite Zeitableitung im Fehler von $\partial\bar{T}/\partial t$ entsprechend (6.99) zu berücksichtigen. Wegen der Terme höherer Dissipation und Dispersion gelingt dies nicht. Zumindest kann man aber den Fehler zweiter Ordnung in der normalen Dissipation eliminieren, wenn man ihn wie in der Lax-Wendroff-Methode für die Diffusionsgleichung mitnimmt. Man muß also die zu geringe Diffusion des FTCS-Verfahren entsprechend erhöhen. Dies führt auf

$$\begin{aligned} T^{n+1}_j &= T^n_j - \frac{\mathrm{C}}{2}\left(T^n_{j+1} - T^n_{j-1}\right) + \frac{\mathrm{C}^2}{2}\left(T^n_{j+1} - 2T^n_j + T^n_{j-1}\right) \\ &\quad + s\left(T^n_{j+1} - 2T^n_j + T^n_{j-1}\right) \\ &= T^n_j - \frac{\mathrm{C}}{2}\left(T^n_{j+1} - T^n_{j-1}\right) + \left(s + \frac{\mathrm{C}^2}{2}\right)\left(T^n_{j+1} - 2T^n_j + T^n_{j-1}\right). \end{aligned} \tag{6.106}$$

Der korrigierende diffusive Term ist rot dargestellt. Das Verfahren hat formal dieselben Stabilitätsgrenzen wie das FTCS-Verfahren, nur mit der modifizierten Diffusivität $\kappa^* = \kappa + \Delta t u^2/2$.

Implizite Verfahren: Crank-Nicolson

Einfaches Crank-Nicolson-Verfahren Eines der effizientesten impliziten Verfahren für die Diffusionsgleichung ist das Crank-Nicolson-Verfahren. Wenn wir das Konzept auf die Transportgleichung übertragen und alle räumlichen Ableitungen

[25] Bei Diskretisierung des Konvektionsterms mit zentralen Differenzen (6.97) kann man das nicht fordern, weshalb die Stabilitätsbedingung (6.100) nicht mit der Nullstelle von $s - \mathrm{C}/2$ zusammenfällt.

Table 9.3. Algebraic (discretised) schemes for the transport equation $\partial\bar{T}/\partial t + u\partial\bar{T}/\partial x - \alpha\partial^2\bar{T}/\partial x^2 = 0$

Scheme	Algebraic form	Truncation error[a] (E) (leading terms)	Amplification factor G ($\theta = m\pi\Delta x$)	Stability Restrictions	Remarks
FTCS	$\frac{\Delta T_j^{n+1}}{\Delta t} + uL_x T_j^n - \alpha L_{xx} T_j^n = 0$	$Cu(\Delta x/2)\frac{\partial^2 T}{\partial x^2}$	$1 - 2s(1-\cos\theta) - \mathrm{i}C\sin\theta$	$0 \leqq C^2 \leqq 2s \leqq 1$	$R_{\text{cell}} \ll 2/C$ for accuracy
		$-[C\alpha\Delta x - u(\Delta x^2/6)(1+2C^2)]\frac{\partial^3 T}{\partial x^3}$			
Upwind	$\frac{\Delta T_j^{n+1}}{\Delta t} + u\frac{(T_j^n - T_{j-1}^n)}{\Delta x} - \alpha L_{xx} T_j^n = 0$	$-u(\Delta x/2)(1-C)\frac{\partial^2 T}{\partial x^2} - [C\alpha\Delta x$	$1 - (2s+C)(1-\cos\theta) - \mathrm{i}C\sin\theta$	$C + 2s \leqq 1$	$R_{\text{cell}} \ll 2/(1-C)$ for accuracy
		$-u(\Delta x^2/6)(1-3C+2C^2)]\frac{\partial^3 T}{\partial x^3}$			
DuFort-Frankel	$(T_j^{n+1} - T_j^{n-1})/2\Delta t + uL_x T_j^n$	$\alpha C^2\frac{\partial^2 T}{\partial x^2} + (1-C^2)[u\Delta x^2/6$	$\frac{B \pm [B^2 - 8s(1+2s)]^{\frac{1}{2}}}{(2+4s)}$	$C \leqq 1$	$C^2 \ll 1$ for accuracy
	$-\frac{\alpha}{\Delta x^2}\{T_{j-1}^n - (T_j^{n-1} + T_j^{n+1})$	$-2\alpha^2 C^2/u]\frac{\partial^3 T}{\partial x^3}$	where $B = 1 + 4s\cos\theta - \mathrm{i}2C\sin\theta$		
	$+ T_{j+1}^n\} = 0$				

Abbildung 6.18.: **(a)** Verfahren für die Transportgleichung (Fletcher, 1991*a*).

Lax-Wendroff	$\frac{\Delta T_j^{n+1}}{\Delta t}+uL_x T_j^n-\alpha^* L_{xx} T_j^n=0$	$-[C\alpha\,\Delta x-u(\Delta x^2/6)(1-C^2)]\frac{\partial^3 T}{\partial x^3}$	$1-2s^*(1-\cos\theta)-\mathrm{i}C\sin\theta$	$0\leqq C^2\leqq 2s^*\leqq 1$	$R_{\text{cell}}\leqq 2$ to avoid spatial oscillations
	where $\alpha^*=\alpha+0.5uC\,\Delta x$	$+[C\alpha^2/u(\Delta x/2)-\alpha\Delta x^2/12$	where $s^*=\alpha^*\Delta t/\Delta x^2$		
		$-uC(\Delta x^3/8)(C^2-1)]\frac{\partial^4 T}{\partial x^4}$			
Crank-Nicolson	$\frac{\Delta T_j^{n+1}}{\Delta t}$	$u(\Delta x^2/6)(1+0.5C^2)\frac{\partial^3 T}{\partial x^3}$	$\frac{1-s(1-\cos\theta)-\mathrm{i}\,0.5C\sin\theta}{1+s(1-\cos\theta)+\mathrm{i}0.5C\sin\theta}$	None	$R_{\text{cell}}\leqq 2$ to avoid spatial oscillations
	$+\{uL_x-\alpha L_{xx}\}\left\{\frac{T_j^n+T_j^{n+1}}{2}\right\}=0$	$-\alpha(\Delta x^2/12)(1+3C^2)\frac{\partial^4 T}{\partial x^4}$			
Three-level fully implicit	$\frac{3}{2}\frac{\Delta T_j^{n+1}}{\Delta t}-\frac{1}{2}\frac{\Delta T_j^n}{\Delta t}$	$u(\Delta x^2/6)(1+2C^2)\frac{\partial^3 T}{\partial x^3}$	$\frac{1\pm\frac{1}{3}\mathrm{i}[3+16s(1-\cos\theta)+\mathrm{i}8C\sin\theta]^{\frac{1}{2}}}{2(1+\frac{2}{3}[2s(1-\cos\theta)+\mathrm{i}C\sin\theta])}$	None	$R_{\text{cell}}\leqq 2$ to avoid spatial oscillations
	$+\{uL_x-\alpha L_{xx}\}T_j^{n+1}=0$	$-\alpha(\Delta x^2/12)(1+12C^2)\frac{\partial^4 T}{\partial x^4}$			
Linear F.E.M./ Crank-Nicolson	$M_x\frac{\Delta T_j^{n+1}}{\Delta t}+uL_x\left\{\frac{T_j^n+T_j^{n+1}}{2}\right\}$	$uC^2(\Delta x^2/12)\frac{\partial^3 T}{\partial x^3}$	$\frac{2+3\cos\theta-3s(1-\cos\theta)-\mathrm{i}1.5C\sin\theta}{2+3\cos\theta+3s(1-\cos\theta)+\mathrm{i}1.5C\sin\theta}$	None	$R_{\text{cell}}\leqq 2$ to avoid spatial oscillations
	$-\alpha L_{xx}\left\{\frac{T_j^n+T_j^{n+1}}{2}\right\}=0$	$+\alpha(\Delta x^2/12)(1-3C^2)\frac{\partial^4 T}{\partial x^4}$			

[a] The algebraic scheme is equivalent to $\partial T/\partial t+u\partial T/\partial x-\alpha\partial^2 T/\partial x^2+E(T)=0$

$L_x=\frac{1}{2\Delta x}\{-1,0,1\}$, $L_{xx}=\frac{1}{\Delta x^2}\{1,2,1\}$, $M_x=\{\frac{1}{6},\frac{2}{3},\frac{1}{6}\}$, $C=u\Delta t/\Delta x$, $s=\alpha\Delta t/\Delta x^2$, $R_{\text{cell}}=C/s=u\Delta x/\alpha$

Noch Abbildung 6.18.: (b) Verfahren für die Transportgleichung (Fletcher, 1991*a*).

über die Zeitniveaus $n+1$ und n mitteln, erhalten wir aus dem FTCS-Schema das implizite Crank-Nicolson-Verfahren für die Transportgleichung

$$\frac{T_j^{n+1}-T_j^n}{\Delta t}+\frac{u}{2}\left(\frac{T_{j+1}^{n+1}-T_{j-1}^{n+1}}{2\Delta x}+\frac{T_{j+1}^n-T_{j-1}^n}{2\Delta x}\right)$$
$$-\frac{\kappa}{2}\left(\frac{T_{j+1}^{n+1}-2T_j^{n+1}+T_{j-1}^{n+1}}{\Delta x^2}+\frac{T_{j+1}^n-2T_j^n+T_{j-1}^n}{\Delta x^2}\right)=0. \tag{6.107}$$

Wenn wir alle Terme zum neuen Zeitpunkt $n+1$ auf die linke Seite bringen, erhalten wir

$$-\left(s-\frac{\mathrm{C}}{2}\right)T_{j+1}^{n+1}+2\left(1+s\right)T_j^{n+1}-\left(s+\frac{\mathrm{C}}{2}\right)T_{j-1}^{n+1}$$
$$=\left(s-\frac{\mathrm{C}}{2}\right)T_{j+1}^{n}+2\left(1-s\right)T_j^{n}+\left(s+\frac{\mathrm{C}}{2}\right)T_{j-1}^{n}. \tag{6.108}$$

Durch Taylorentwicklung kann man leicht zeigen, daß das Verfahren von der Ordnung $O(\Delta t^2,\Delta x^2)$ ist, also nicht die starke künstliche Diffusion zeigt, wie die Verfahren erster Ordnung. Auch findet man, daß das Crank-Nicolson-Verfahren für alle Werte von C und s stabil ist. Um Oszillationen zu vermeiden, muß jedoch $\mathrm{Re_G}\leq 2$ sein.

Im allgemeinen verhalten sich die impliziten Verfahren für die Transportgleichung gutmütig. Die starke Dispersion, die bei der Advektionsgleichung insbesondere für sehr kleine Wellenlängen viele Probleme macht, kommt bei der Transportgleichung nicht in dem Umfang zum Tragen, da die spektralen Komponenten mit kleiner Wellenlängen durch die reale Diffusion sehr schnell gedämpft werden. Trotzdem können starke Überschwinger auftreten, wenn die Gitter-Reynoldszahl zu groß ist (siehe Abb. 6.19 unten).

Crank-Nicolson-Verfahren mit Massenoperator Das einfache Crank-Nicolson-Verfahren kann erweitert werden, so daß die mit der Dispersion verbundenen Überschwinger bei starker Variation von T (siehe Abb. 6.19) verringert werden. Ein Möglichkeit besteht darin, die Zeitableitung räumlich zu verteilen. Die geschieht mit dem *Massenoperator* $M_x=(\delta,1-2\delta,\delta)$, wobei wir definieren

$$M_x f_j := \delta f_{j-1}+(1-2\delta)\,f_j+\delta f_{j+1}. \tag{6.109}$$

Eine derartige Verschmierung der Zeitableitung tritt in natürlicher Weise bei gewichteten Residuen auf (finiten Elementen, siehe Kap. 4.4.3 in Teil I). Damit lautet das modifizierte CN-Verfahren

$$M_x\left(\frac{T_j^{n+1}-T_j^n}{\Delta t}\right)+\frac{u}{2}\left(\frac{T_{j+1}^{n+1}-T_{j-1}^{n+1}}{2\Delta x}+\frac{T_{j+1}^n-T_{j-1}^n}{2\Delta x}\right)$$
$$-\frac{\kappa}{2}\left(\frac{T_{j+1}^{n+1}-2T_j^{n+1}+T_{j-1}^{n+1}}{\Delta x^2}+\frac{T_{j+1}^n-2T_j^n+T_{j-1}^n}{\Delta x^2}\right)=0. \tag{6.110}$$

Und nach dem Sortieren der Terme erhalten wir

$$\begin{aligned} &-\left(s-\frac{\mathrm{C}}{2}-2\delta\right)T_{j+1}^{n+1}+2\left(1-2\delta+s\right)T_j^{n+1}-\left(s+\frac{\mathrm{C}}{2}-2\delta\right)T_{j-1}^{n+1} \\ &=\left(s-\frac{\mathrm{C}}{2}+2\delta\right)T_{j+1}^{n}+2\left(1-2\delta-s\right)T_j^{n}+\left(s+\frac{\mathrm{C}}{2}+2\delta\right)T_{j-1}^{n}. \end{aligned} \tag{6.111}$$

Man kann zeigen, daß dieses Verfahren stabil ist für $\delta \leq 1/4$. Formal ist (6.111) von zweiter Ordnung. Aber wenn man $\delta = 1/6 + \mathrm{C}^2/12$ wählt, kann man den führenden Term im Dispersionsfehler eliminieren.[26] Wenn man δ so wählt, erfordert die Stabilitätsbedingung $\delta \leq 1/4$ jedoch $\mathrm{C} \leq 1$.

Crank-Nicolson-Verfahren mit Upwind-Verfahren höherer Ordnung Ein besseres Dispersionsverhalten kann man auch erhalten, wenn man das Crank-Nicolson-Verfahren mit einem *Upwind-Verfahren höherer Ordnung* für den advektiven Term kombiniert. Die Vier-Punkt-Upwind-Formulierung der ersten Ableitung lautet[27]

$$\frac{\partial T^n}{\partial x} = L_x^{(4)} T_j^n = \begin{cases} \dfrac{T_{j+1}^n - T_{j-1}^n}{2\Delta x} + q\dfrac{T_{j-2}^n - 3T_{j-1}^n + 3T_j^n - T_{j+1}^n}{3\Delta x}, & u > 0, \\ \dfrac{T_{j+1}^n - T_{j-1}^n}{2\Delta x} - q\dfrac{T_{j+2}^n - 3T_{j+1}^n + 3T_j^n - T_{j-1}^n}{3\Delta x}, & u < 0. \end{cases} \tag{6.112}$$

Der Faktor $q \in [0, 0.5]$ reguliert zwischen reinen zentralen Differenzen ($q = 0$) und reinem Upwind-Verfahren ($q = 0.5$). Für $q \neq 0.5$ ist dieses Schema von zweiter Ordnung. Im Grenzfall $q = 0.5$ hat man genau das 4-Punkt-Upwind-Verfahren, was die Fehlerordnung $O\left(\Delta x^3\right)$ besitzt. In Kombination mit dem Crank-Nicolson-Verfahren erhält man

$$\frac{T_j^{n+1} - T_j^n}{\Delta t} + \frac{1}{2}\left(uL_x^{(4)} - \kappa L_{xx}\right)\left(T_j^{n+1} + T_j^n\right) = 0. \tag{6.113}$$

Sortieren der Terme liefert das implizite Verfahren ($u > 0$)

$$\frac{\mathrm{C}q}{3}T_{j-2}^{n+1} - \left(s + \frac{\mathrm{C}}{2} + \mathrm{C}q\right)T_{j-1}^{n+1} + 2\left(1 + s + \frac{\mathrm{C}q}{2}\right)T_j^{n+1}$$

[26] Es bleibt aber noch ein Term höherer Dissipation ($\sim \partial_x^4$) übrig.

[27] Man erhält dies Schema wie in Kap. 2.2 von Teil I, indem man ansetzt $\partial T/\partial x = aT_{j-2} + bT_{j-1} + cT_j + dT_{j+1}$, alle Terme um den Punkt x_j entwickelt und fordert, daß der Koeffizient vor $\partial T/\partial x$ gleich 1 ist und weitere drei Koeffizienten verschwinden. Dies liefert vier Gleichungen für a, b, c und d mit dem Ergebnis

$$\begin{aligned} \frac{\partial T^n}{\partial x} &= \frac{T_{j-2}^n - 6T_{j-1}^n + 3T_j^n + 2T_{j+1}^n}{6\Delta x} + O(\Delta x^3) \\ &= \frac{T_{j+1}^n - T_{j-1}^n}{2\Delta x} + \frac{T_{j-2}^n - 3T_{j-1}^n + 3T_j^n - T_{j+1}^n}{6\Delta x} + O(\Delta x^3). \end{aligned}$$

$$-\left(s-\frac{\mathrm{C}}{2}+\frac{\mathrm{C}q}{3}\right)T_{j+1}^{n+1} = -\frac{\mathrm{C}q}{3}T_{j-2}^{n}+\left(s+\frac{\mathrm{C}}{2}+\mathrm{C}q\right)T_{j-1}^{n} +2\left(1-s-\frac{\mathrm{C}q}{2}\right)T_{j}^{n}+\left(s-\frac{\mathrm{C}}{2}+\frac{\mathrm{C}q}{3}\right)T_{j+1}^{n}. \quad (6.114)$$

Dies führt auf Matrizen mit vier Diagonalen. Zur Lösung muß daher die vierte Diagonale mit einem zusätzlichen Sweep eliminiert werden (siehe Kap. 5.1, Teil I), damit die Matrix tridiagonal wird. Danach kann man dann den Thomas-Algorithmus verwenden. Für $q = 0$ erhält man das normale Crank-Nicolson-Schema.

Beipielrechnungen

Benchmark und analytische Lösung Um einige der obigen Verfahren zu testen, betrachten wir den eindimensionalen Temperaturtransport. Gegeben sei eine anfängliche Temperaturverteilung in Form der Stufenfunktion

$$T(x,0) = \begin{cases} 1, & x < 0, \\ 0, & x \geq 0. \end{cases} \quad (6.115)$$

Diese scharfe Stufe wird im Laufe der Zeit durch Advektion verschoben und gleichzeitig durch Diffusion verschmiert.

Mit der Längenskala L und $X = x - ut$ lautet die exakte Lösung (siehe Anhang E)

$$\bar{T} = \frac{1}{2} - \sum_{k=1}^{\infty} \frac{2}{(2k-1)\pi} \sin\left[(2k-1)\pi X/L\right] \mathrm{e}^{-\kappa(2k-1)^2\pi^2 t/L^2}. \quad (6.116)$$

Parameter In der Regel sind die physikalischen Parameter u und κ gegeben. Außerdem gibt man Δx und Δt vor. Daraus berechnen sich die wesentlichen numerischen Parameter s und C. Von ihnen hängt auch $\mathrm{Re_G} = \mathrm{C}/s$ ab.

Bei den Rechnungen geben wir die konvektive Geschwindigkeit $u = 1$ vor, damit wir den Sprung zum Zeitpunkt $t = 1$ immer an der Stelle $x = 1$ haben. Damit die Front innerhalb dieser Zeit nicht zu sehr verschmiert, wählen wir $\kappa = 0.001$. Mit dem Parameter s steuern wir die Zeitschrittweite Δt. s muß bei den gegebenen physikalischen Parametern und Auflösung hinreichend klein sein, da sonst die Courant-Zahl $\mathrm{C} = u\Delta t/\Delta x$ größer wird als 1 (siehe (6.54)).

Ergebnisse sind in Abb. 6.19 für verschiedene Verfahren dargestellt. Für $\mathrm{C} = 0.1$ und $s = 0.001$ (Abb. 6.19a) sind alle Verfahren ungenau und zeigen oszillatorisches Verhalten ($\mathrm{Re_G} = 100$). Am schlimmsten wirkt sich dies aus beim FTCS-Verfahren, dem einfachem Crank-Nicolson, dem Lax-Wendroff und dem explizitem Upwind-Verfahren 4. Ordnung. Eine Erhöhung der Zeitschrittweite um einen Faktor 10 zu $\mathrm{C} = 1$ und $s = 0.01$ (Abb. 6.19b) führt dazu, daß das implizite

An der letzten Form sieht man, daß man einen Term abspalten kann, der genau den zentralen Differenzen entspricht. Wenn man den Rest weg läßt, ist das Verfahren 2. Ordnung, ansonsten

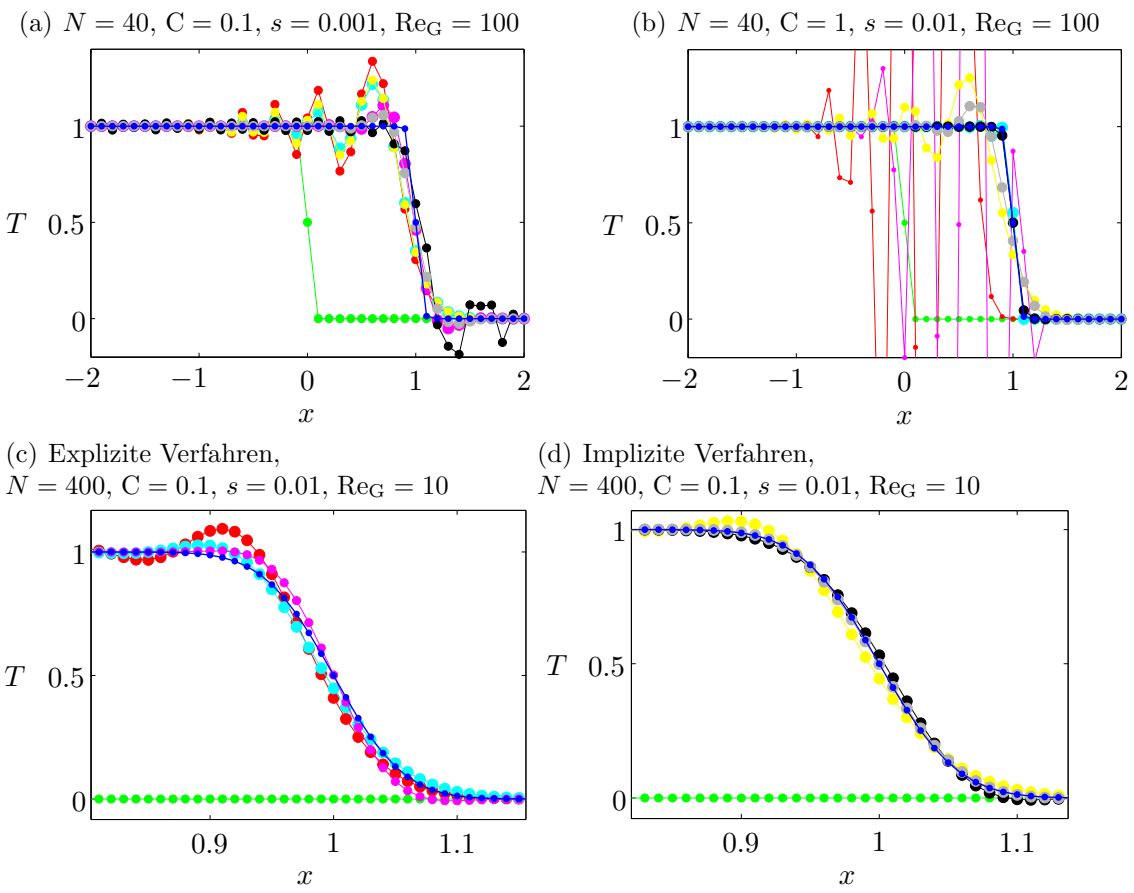


Abbildung 6.19.: Vergleich verschiedenen Verfahren zur Lösung der Transportgleichung $\partial_t T + u\partial_x T - \kappa\partial_{xx} T = 0$ für eine anfänglich stufenförmige Front bei $x = 0$ (grün). Die exakte Lösung nach $t = 1$ ist blau dargestellt.

Crank-Nicolson-Verfahren mit Massenoperator und $\delta = 0.25$ (schwarz) die Front sehr genau wiedergibt. Dies ist gerade die Stabilitätsgrenze für dieses Verfahren. Auch das Lax-Wendroff-Verfahren (cyan) zeigt nur einen kleinen Überschwinger. Alle anderen Verfahren zeigen jedoch inakzeptable Oszillationen.

Wenn man die räumliche Gitterweite um einen Faktor 10 verringert, werden alle Verfahren genauer und wegen der reduzierten Gitter-Reynoldszahl nehmen die Oszillationen ab. Explizite (Abb. 6.19c) und implizite Verfahren (Abb. 6.19d) sind separat dargestellt. Auch hier schneidet FTCS am schlechtesten ab. Bei den expliziten Verfahren ist Lax-Wendroff am genauesten. Implizite Verfahren sind deutlich

3. Ordnung. Den Korrekturterm kann man deshalb mit einem Faktor $2q$ wichten mit $q \in [0, 0.5]$.

besser. Hier schneidet das Crank-Nicolson-Verfahren mit dritter Ordnung Upwind am besten ab. Es ist aber auch das aufwendigste Verfahren (4 Diagonalen). Eine Übersicht über Verfahren zur Diskretisierung der Transportgleichung ist in Abb. 6.18a,b gezeigt.

6.5. Nichtlineare Effekte: Die Burgers-Gleichung

Bisher hatten wir die *lineare* Transportgleichung betrachtet. Der Transport des Impulses, der durch die Euler- bzw. die Navier-Stokes-Gleichung beschrieben wird, ist jedoch nichtlinear (siehe Kap. 1.1, Teil I). Dies kann man der substantiellen Ableitung der Geschwindigkeit $\mathrm{d}\boldsymbol{u}/\mathrm{d}t = \partial\boldsymbol{u}/\partial t + \boldsymbol{u}\cdot\nabla\boldsymbol{u}$ sofort ansehen.

Ein Paradebeispiel für den nichtlinearen Transport ist die Burgers-Gleichung in einer Dimension. Im reibungsfreien Fall lautet sie

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0. \tag{6.117}$$

Johannes Martinus Burgers 1895–1981

Die Nichtlinearität ist quadratisch. Die Lösungen der reibungsfreien Burgers-Gleichung besitzen offenbar die Form $u = u(x - ut)$, bei welcher Funktion und Argument implizit gekoppelt sind. Dies sieht man leicht, wenn man die partiellen Ableitungen bildet und in (6.117) einsetzt.

Physikalisch bedeutet (6.117), daß sich die Geschwindigkeit u eines substantiellen Fluidelements entlang seiner Trajektorie nicht ändert ($\mathrm{d}u/\mathrm{d}t = 0$). Jedes Fluidelement behält seine ursprüngliche Geschwindigkeit bei und bewegt sich mit konstanter Geschwindigkeit. Falls nun die anfängliche Geschwindigkeit $u(x, 0)$ räumlich variiert, dann werden Fluidelemente mit einer hohen Geschwindigkeit diejenigen mit einer niedrigeren Geschwindigkeit ein- und eventuell sogar überholen können.

Dieser Prozeß führt dazu, daß sich eine Welle $u(x, t)$ im Laufe der zeitlichen Entwicklung aufsteilt und, falls Fluidelemente überholt werden, sogar mehrdeutig wird. Dieser Prozeß entspricht einer Generierung von Fourierkomponenten mit sehr hoher Wellenzahl im Signal $u(x, t)$. Ursache hierfür ist die Nichtlinearität der Gleichung (6.117). Als Beispiel ist die zeitliche Entwicklung von u für die Anfangsbedingung $u(x, 0) = \sin(x)$ in Abb. 6.20 gezeigt. Bei $t = 1$ besitzt die Welle an den Stellen $x = (2n - 1)\pi$ eine unendliche Steigung. Für $t > 1$ wird die Welle mehrdeutig.[28]

Die eigentliche *Burgers-Gleichung* enthält noch einen viskosen Term. Sie lautet

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \nu\frac{\partial^2 u}{\partial x^2}. \tag{6.118}$$

Die Diffusion der Geschwindigkeit u mit Diffusivität ν bewirkt eine Dämpfung aller Variationen von u und verhindert gleichzeitig das Entstehen der Mehrdeutigkeit.

[28] Die Aufsteilung kann man auch bei Wellen sehen, die auf einen flachen Strand auflaufen bevor sie

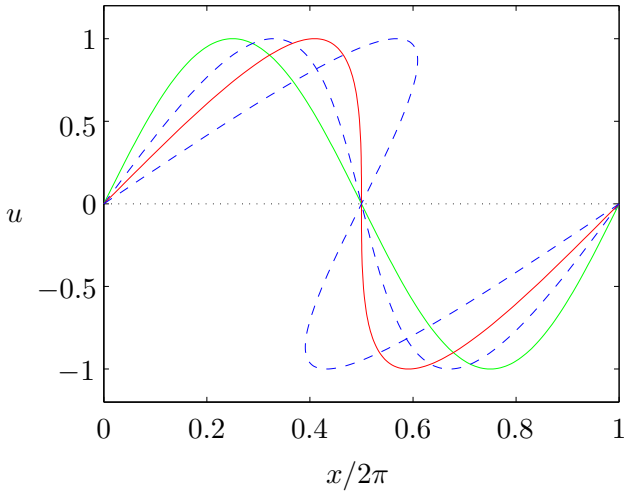


Abbildung 6.20.: Aufsteilung und Entwicklung der Mehrdeutigkeit einer anfänglich harmonischen Welle als Lösung der nichtlinearen reibungsfreien Burgers-Gleichung für $t = 0$ (grün), $t = 0.5$ (blau gestrichelt), $t = 1$ (rot) und $t = 2$ (blau gestrichelt).

Denn je höher jedoch die Wellenzahl k ist, desto stärker wird die entsprechende spektrale Komponente durch Diffusion gedämpft. Bei der Fourier-Transformation geht $\nu\partial_x^2$ in $-\nu k^2$ über.

Oft ist es sinnvoll, die Burgers-Gleichung in sogenannter *konservativer Form* zu schreiben

$$\frac{\partial u}{\partial t} + \frac{\partial F}{\partial x} = \nu \frac{\partial^2 u}{\partial x^2}. \tag{6.119}$$

Hierbei ist $F = u^2/2$.

Eine wichtige Konsequenz eines nichtlinearen Terms ist das *Aliasing*, das wir schon in Kap. 4.6.5 angesprochen hatten. Das Problem des *Aliasing* wird durch den dissipativen Term verringert, wenn die Dämpfung von Moden bei der Nyquist-Wellenzahl hinreichend stark ist. Dann werden die spektralen Komponenten auf großen Längenskalen nur unwesentlich verfälscht.

Die Lösung des Anfangswertproblems für die Burgers-Gleichung (6.118) kann man glücklicherweise exakt angeben

$$u(x,t) = \frac{\displaystyle\int_{-\infty}^{\infty} \left(\frac{x-s}{t}\right) G(s) \exp\left\{-(x-s)^2/4\nu t\right\} \mathrm{d}s}{\displaystyle\int_{-\infty}^{\infty} G(s) \exp\left\{-(x-s)^2/4\nu t\right\} \mathrm{d}s}. \tag{6.120}$$

Wie man dieses Ergebnis erhält, ist in Anhang F beschrieben. Für die weiter unten verwendete Anfangsbedingung $u(x,0) = [1 - \mathrm{sign}(x)]/2$, also für eine Stufe bei $x = 0$, ist $G(x) = \mathrm{e}^{-x/2\nu}$ falls $x < 0$, und 1 für $x \geq 0$. Die exakte Lösung ist

sich brechen. Dabei ist u nicht die Geschwindigkeit eines substantiellen Fluidelements, sondern die Amplitude einer Welle.

sehr hilfreich bei der Untersuchung der verschiedenen numerischen Verfahren. Im folgenden sollen nun die schon bekannten Verfahren auf die nichtlineare Burgers-Gleichung angewandt und ggf. modifiziert werden.

6.5.1. Explizite Verfahren

Das FTCS-Verfahren läßt sich problemlos auf die eindimensionale Burgers-Gleichung anwenden. Es ergibt sich

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}+\frac{u_j^n\left(u_{j+1}^n-u_{j-1}^n\right)}{2\Delta x}-\nu\frac{u_{j+1}^n-2u_j^n+u_{j-1}^n}{\Delta x^2}=0. \tag{6.121}$$

Die konservative Variante lautet

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}+\frac{F_{j+1}^n-F_{j-1}^n}{2\Delta x}-\nu\frac{u_{j+1}^n-2u_j^n+u_{j-1}^n}{\Delta x^2}=0. \tag{6.122}$$

Leider kann man die Stabilität hier nicht so einfach mit der von-Neumann-Methode untersuchen wie bei linearen Gleichungen, denn durch die Nichtlinearität sind alle Fourierkomponenten miteinander gekoppelt. Daher kann man die Wellenzahl nicht als einen einfachen Parameter behandeln. Ein Ausweg besteht in der Näherung, den Faktor u vor der ersten Ableitung $\partial u/\partial x$ als konstant zu betrachten (*frozen coefficient*). Diese Vorgehensweise liefert oft eine brauchbare Näherung für die Stabilitätsgrenze. Da der eingefrorene Koeffizient aber örtlich und zeitlich variiert, ist die Stabilität *lokal* zu verstehen.

Oft erhält man eine Verbesserung der FTCS-Diskretisierung, wenn man die erste Ableitung mittels 4-Punkt-Upwinding nach (6.112) diskretisiert. Der nichtlineare Term ist dann für $q \neq 0.5$ von zweiter Ordnung Genauigkeit und für $q = 0.5$ von dritter Ordnung $O\left(\Delta x^3\right)$.

Da die reibungsfreie Burgers-Gleichung der Advektionsgleichung ähnelt, ist man geneigt, das Lax-Wendroff-Schema auf die Burgers-Gleichung zu übertragen, da Lax-Wendroff für die Advektionsgleichung auf jeden Fall besser ist als FTCS. Kern des Lax-Wendroff-Verfahren ist es, den führenden Term im Fehler in der ersten Zeitableitung bei Diskretisierung in Vorwärtsrichtung $(-(\Delta t/2)\partial^2\bar{u}/\partial t^2)$ durch äquivalente Raumableitungen auszudrücken und mit Zweiter-Ordnung-Diskretisierung in das Differenzenschema aufzunehmen (siehe (6.66)). Für die nichtlineare reibungsfreie Burgers-Gleichung ist es jedoch zunächst unklar, wie man die zweite Zeitableitung durch Ortsableitungen ausdrücken kann. Dies ist aber folgendermaßen möglich.

Wir betrachten die reibungsfreie Burgers-Gleichung in konservativer Form $\partial_t\bar{u}+\partial_x\bar{F}=0$ mit $\bar{F}=\bar{F}[\bar{u}(x,t)]=\bar{u}^2/2$. Für die zweite Zeitableitung gilt

$$\begin{aligned}\frac{\partial^2\bar{u}}{\partial t^2}&=-\frac{\partial}{\partial x}\frac{\partial\bar{F}}{\partial t}=-\frac{\partial}{\partial x}\underbrace{\frac{\partial\bar{F}}{\partial\bar{u}}}_{:=A}\frac{\partial\bar{u}}{\partial t}=\frac{\partial}{\partial x}\left(A\frac{\partial\bar{F}}{\partial x}\right)\\&\xrightarrow[\text{(sym.)}]{\text{Diskretisierung}}\frac{1}{\Delta x}\left[A_{j+1/2}\left(\frac{F_{j+1}^n-F_j^n}{\Delta x}\right)-A_{j-1/2}\left(\frac{F_j^n-F_{j-1}^n}{\Delta x}\right)\right].\end{aligned} \tag{6.123}$$

Wenn wir diesen Teil des Fehlers (mit dem Koeffizienten $-\Delta t/2$) wie beim Lax-Wendroff-Verfahren in die Differenzengleichung für den reibungsfreien Fall einbeziehen, erhalten wir

$$\frac{u_j^{n+1}-u_j^n}{\Delta t} = -\frac{F_{j+1}^n - F_{j-1}^n}{2\Delta x} + \frac{\Delta t}{2\Delta x^2}\left[A_{j+1/2}\left(F_{j+1}^n - F_j^n\right) - A_{j-1/2}\left(F_j^n - F_{j-1}^n\right)\right]. \tag{6.124}$$

Für die Burgers-Gleichung ist $A_{j+1/2} = u_{j+1/2} = (u_j + u_{j+1})/2$. Das Schema besitzt den Abbruch-Fehler $O(\Delta x^2, \Delta t^2)$ und es ist stabil für $u_{\max}\Delta t/\Delta x \leq 1$ entsprechend der Courant-Friedrichs-Levi-Bedingung (6.53) für die Advektionsgleichung.

Um die Auswertung von A an den räumlichen Zwischenstellen $j \pm 1/2$ zu vermeiden, kann man die obige Gleichung durch die äquivalente Form

$$u_{j+1/2}^* = \frac{1}{2}\left(u_j^n + u_{j+1}^n\right) - \frac{\Delta t}{2\Delta x}\left(F_{j+1}^n - F_j^n\right), \tag{6.125a}$$

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x}\left(F_{j+1/2}^* - F_{j-1/2}^*\right). \tag{6.125b}$$

ersetzen, die etwas ökonomischer ist. Die Äquivalenz kann man leicht durch Einsetzen überprüfen. Mit $F = u$ ist dieses Schema identisch mit dem Lax-Wendroff-Verfahren (6.67) für die Advektionsgleichung.

Man kann das Verfahren auch auf den viskosen Fall erweitern. Das entsprechende Schema lautet dann[29]

$$u_{j+1/2}^* = \frac{1}{2}\left(u_j^n + u_{j+1}^n\right) - \frac{\Delta t}{2\Delta x}\left(F_{j+1}^n - F_j^n\right) + \frac{s}{2}\left[\frac{1}{2}\left(u_{j-1}^n - 2u_j^n + u_{j+1}^n\right) + \frac{1}{2}\left(u_j^n - 2u_{j+1}^n + u_{j+2}^n\right)\right], \tag{6.126a}$$

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x}\left(F_{j+1/2}^* - F_{j-1/2}^*\right) + s\left(u_{j-1}^n - 2u_j^n + u_{j+1}^n\right). \tag{6.126b}$$

Hierbei ist wie üblich $s = \nu\Delta t/\Delta x^2$. Die angegebene viskose Version des Verfahrens hat jedoch den Nachteil, daß es nur noch $O(\Delta t, \Delta x^2)$ ist.[30] Stationäre Lösungen sind zwar von zweiter Ordnung Genauigkeit. Um aber auch die Dynamik (Zeitabhängigkeit) genau zu berechnen, muß der Zeitschritt sehr klein sein. Darüber hinaus benötigt man auch noch eine weitere Randbedingung, da der viskose Term vier Gitterpunkte involviert. Für Stabilität kann man Bedingung

$$\Delta t\left(A^2\Delta t + 2\nu\right) \leq \Delta x^2 \tag{6.127}$$

ableiten (Fletcher, 1991*a*). Für praktische Berechnungen schlagen Peyret and Taylor (1983) das Kriterium $\Delta t \leq \Delta x^2/(2\nu + |A|\Delta x)$ vor.

[29] Im ersten Term auf der rechten Seite der ersten Gleichung hat Fletcher (1991*a*) einen kleinen Druckfehler (u_j^{n+1}). Außerdem scheint dieses Lax-Wendroff-Schema gar nicht so schlecht zu sein, wie in Fletcher (1991*a*) implizit behauptet. Die abweichenden numerischen Werte in Tabelle 10.3 von Fletcher (1991*a*) können so nicht nachvollzogen werden. Siehe auch Abb. 6.21 und 6.22 .

[30] Offenbar wird in dieser Version der Fehlerterm in der Zeitableitung nicht mit zweiter Ordnung

6.5.2. Implizite Verfahren

Wenn wir nach der Crank-Nicolson-Methode vorgehen, werden sowohl der konvektive wie auch der diffusive Term je zur Hälfte dem Zeitniveau n und $n+1$ zugeschlagen. Für die Burgers-Gleichung ergibt sich so

$$\frac{u_j^{n+1}-u_j^n}{\Delta t} = -\frac{1}{2}L_x\left(F_j^{n+1}+F_j^n\right)+\frac{\nu}{2}L_{xx}\left(u_j^{n+1}+u_j^n\right)=0. \tag{6.128}$$

Um diese Gleichung lösen zu können, würden wir sie gerne in ein *lineares* tridiagonales System überführen. Der Term F_j^{n+1} ist für die Burgers-Gleichung jedoch *quadratisch* in u_j^{n+1}. Man kann ihn aber mit Hilfe einer zeitlichen Taylor-Entwicklung auf einen bezüglich des Zeitpunkts t_{n+1} linearen Term zurückführen. Dazu entwickeln wir F_j^{n+1} um den Zeitpunkt t_n und erhalten mit $F=u^2/2$ und $(\partial u/\partial t)_j^n = \Delta u_j^{n+1}/\Delta t + O(\Delta t)$

$$\begin{aligned} F_j^{n+1} &= F_j^n + \Delta t\left(\frac{\partial F}{\partial t}\right)_j^n + O\left(\Delta t^2\right) = F_j^n + u_j^n \underbrace{\Delta t\left(\frac{\partial u}{\partial t}\right)_j^n}_{\Delta u_j^{n+1}+O(\Delta t^2)} + O\left(\Delta t^2\right) \\ &= F_j^n + u_j^n\left(u_j^{n+1}-u_j^n\right)+O\left(\Delta t^2\right). \end{aligned} \tag{6.129}$$

Dieser Term ist linear in u_j^{n+1}. Damit können wir die Burgers-Gleichung schreiben als

$$\frac{u_j^{n+1}-u_j^n}{\Delta t} = -\frac{1}{2}L_x\underbrace{\left[2F_j^n+u_j^n\left(u_j^{n+1}-u_j^n\right)\right]}_{u_j^n u_j^{n+1}}+\frac{\nu}{2}L_{xx}\left(u_j^{n+1}+u_j^n\right)=0. \tag{6.130}$$

Wenn wir nun die Zeitniveaus trennen, erhalten wir

$$u_j^{n+1}+\frac{\Delta t}{2}\left[L_x\left(u_j^n u_j^{n+1}\right)-\nu L_{xx}u_j^{n+1}\right] = u_j^n+\frac{\nu\Delta t}{2}L_{xx}u_j^n. \tag{6.131}$$

Mit den Operatoren[31] $L_x=(-1,0,1)/2\Delta x$ und $L_{xx}=(1,-2,1)/\Delta x^2$ erhalten wir dann die gewünschte tridiagonale Form

$$a_j^n u_{j-1}^{n+1}+b_j^n u_j^{n+1}+c_j^n u_{j+1}^{n+1}=d_j^n, \tag{6.132}$$

mit

$$a_j^n = -\frac{1}{2}\left(\frac{u_{j-1}^n\Delta t}{2\Delta x}+s\right), \tag{6.133a}$$

$$b_j^n = 1+s, \tag{6.133b}$$

$$c_j^n = \frac{1}{2}\left(\frac{u_{j+1}^n\Delta t}{2\Delta x}-s\right), \tag{6.133c}$$

$$d_j^n = \frac{s}{2}u_{j-1}^n+(1-s)u_j^n+\frac{s}{2}u_{j+1}^n. \tag{6.133d}$$

Genauigkeit diskretisiert.

[31] Hier bedeuten z.B. $(-1,0,1)$ die Wichtung der räumlichen Komponenten $j-1$, j und $j+1$ (in

Das Verfahren ist von der Ordnung $O(\Delta t^2, \Delta x^2)$ und es ist uneingeschränkt stabil.

Dieses Crank-Nicolson-Verfahren kann jetzt wie in Kap. 6.4.2 weiter verbessert werden; zum Beispiel durch Verteilung der zeitlichen Ableitung auf die benachbarten Raumpunkte mittels Massenoperator $M_x = (\delta, 1-2\delta, \delta)$ oder/und durch Verwendung des 4-Punkt-Upwind-Verfahrens ($L_x^{(4)}$ nach (6.112)). Ein allgemeines Crank-Nicolson-Verfahren, das beide Modifikationen erhält, kann man schreiben als

$$M_x \left(\frac{u_j^{n+1} - u_j^n}{\Delta t} \right) = -\frac{1}{2} L_x^{(4)} \left(u_j^n u_j^{n+1} \right) + \frac{\nu}{2} L_{xx} \left(u_j^{n+1} + u_j^n \right) = 0. \tag{6.134}$$

Man erhält so das quadridiagonale System (für $u > 0$)

$$e_j^n u_{j-2}^{n+1} + a_j^n u_{j-1}^{n+1} + b_j^n u_j^{n+1} + c_j^n u_{j+1}^{n+1} = d_j^n, \tag{6.135}$$

mit den Koeffizienten

$$e_j^n = \frac{q}{6} \frac{\Delta t}{\Delta x} u_{j-2}^n, \tag{6.136a}$$

$$a_j^n = -\frac{1}{2} \left[\left(\frac{1}{2} + q \right) \frac{u_{j-1}^n \Delta t}{\Delta x} + s - 2\delta \right], \tag{6.136b}$$

$$b_j^n = 1 - 2\delta + s + \frac{q u_j^n \Delta t}{2\Delta x}, \tag{6.136c}$$

$$c_j^n = \frac{1}{2} \left[\left(\frac{1}{2} - q \right) \frac{u_{j+1}^n \Delta t}{3\Delta x} - s + 2\delta \right], \tag{6.136d}$$

$$d_j^n = \left(\delta + \frac{s}{2} \right) u_{j-1}^n + (1 - 2\delta - s) u_j^n + \left(\delta + \frac{s}{2} \right) u_{j+1}^n. \tag{6.136e}$$

Die Lösung erfolgt wieder durch Gauß-Elimination der durch e_j^n beschriebenen vierten Diagonale durch einen *Sweep* und anschließendem Thomas-Algorithmus für das verbleibende tridiagonale System.

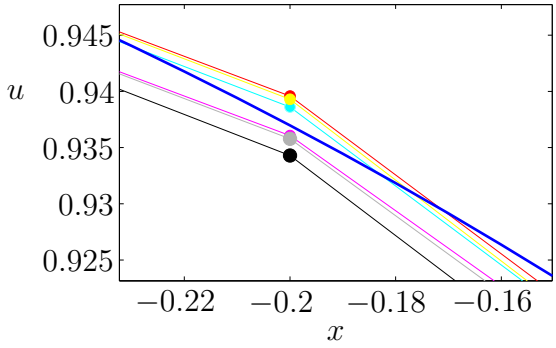
6.5.3. Numerische Ergebnisse

Numerische Ergebnisse sind in Abb. 6.21–6.23 dargestellt. Bei moderaten Parametern, wie in Abb. 6.21, sind alle untersuchten Verfahren gut zu gebrauchen. Eine genauere Inspektion zeigt, daß bei den expliziten Verfahren das FTCS- und bei den impliziten Verfahren das Crank-Nicolson-Verfahren mit 4-Punkt-Upwinding am genauesten sind.

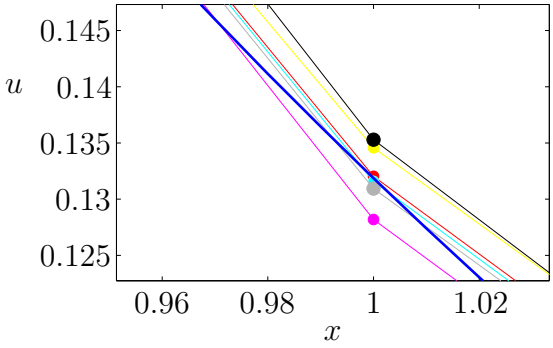
Wenn die Gitter-Reynoldszahl anwächst, werden zunächst die expliziten Verfahren ungenau (Abb. 6.22). Auch das einfache Crank-Nicolson-Verfahren zeigt Schwächen. Die Crank-Nicolson-Verfahren mit Massenoperator bzw. 4-Punkt-Upwind bieten hier noch die genauesten Ergebnisse.

Für sehr hohe Gitter-Reynoldszahlen kommt es auch bei Verwendung von impliziten Crank-Nicolson-Verfahren zu Oszillationen in der Nähe der Front. Um diese

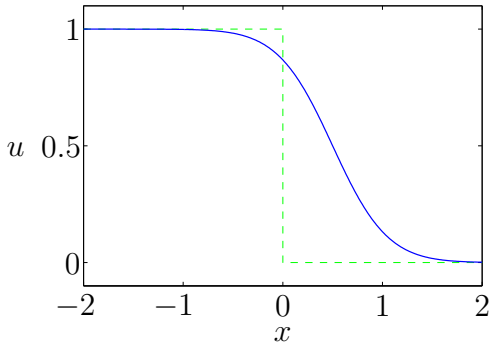
(a) $J = 20$, $\Delta x = 0.2$, $s = 0.25$, $\mathrm{Re_G} = 1$, $\nu = 0.2$, $t = 1$



(b) $J = 20$, $\Delta x = 0.2$, $s = 0.25$, $\mathrm{Re_G} = 1$, $\nu = 0.2$, $t = 1$



(c) exakte Lösung, $\nu = 0.2$, $t = 1$



(d) Legende

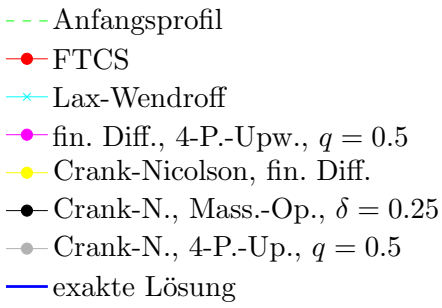


Abbildung 6.21.: Vergleich verschiedenen Verfahren für die Burgers-Gleichung mit $\nu = 0.2$ und $t = 1$. Die Gitter-Reynoldszahl, gebildet mit der Referenzgeschwindigkeit $u = 1$, ist $\mathrm{Re_G} = 1$. Auf der vollen Skala (c) sind die Ergebnisse nicht zu unterscheiden. In (a) und (b) sind Vergrößerungen gezeigt.

Oszillationen zu vermeiden, wird manchmal eine zusätzliche *künstliche Diffusion* eingeführt. Die künstliche Diffusion soll die Oszillationen dämpfen, die durch das numerische Verfahren enstehen. Fletcher (1991*a*) schlägt vor, den künstlichen Diffusionsterm

$$\frac{\nu_a}{2}\Delta t L_{xx}\left(F_j^n + F_j^{n+1}\right) \overset{F=u^2/2}{=} \frac{\nu_a}{2}\Delta t L_{xx}\left(u_j^{n+1}u_j^n\right) \tag{6.137}$$

zur rechten Seite von (6.134) zu addieren. Dies führt auf

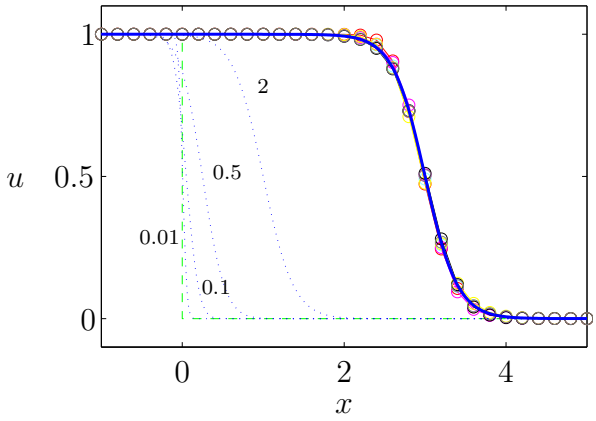
$$M_x\left(\frac{u_j^{n+1}-u_j^n}{\Delta t}\right) = -\frac{1}{2}L_x^{(4)}\left(u_j^n u_j^{n+1}\right) + \frac{\nu}{2}L_{xx}\left(u_j^{n+1}+u_j^n\right) + \frac{\nu_a}{2}\Delta t L_{xx}\left(u_j^{n+1}u_j^n\right) = 0 \tag{6.138}$$

mit entsprechender Modifikation der Koeffizienten (6.136).

Die Ergebnisse, die damit erzielt werden können, sind in Abb. 6.23 gezeigt. Für alle Crank-Nicolson-Verfahren wurde hier $\nu_a = 0.25$ verwendet, so daß sich für $\mathrm{C} = 1$ der Wert $s_a = \nu_a \mathrm{C}^2 = 0.25$ ergibt, wobei hier $\mathrm{C} = \Delta t/\Delta x$ eine Referenz-

dieser Reihenfolge).

(a) $J = 30$, $\Delta x = 0.2$, $s = 0.2$, $\mathrm{Re_G} = 2$, $\nu = 0.1$, $t = 6$

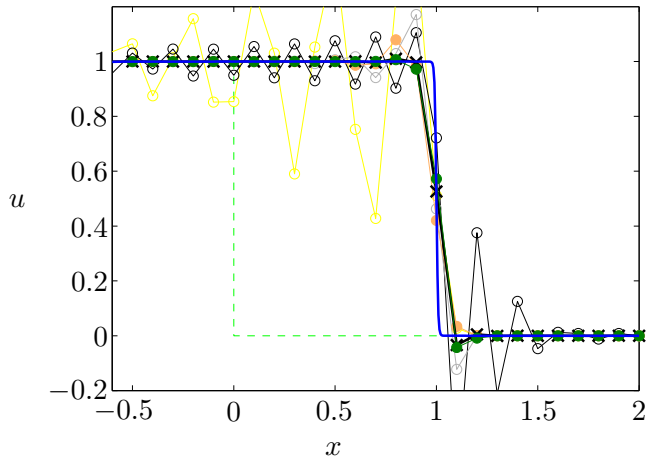


(b) Legende

- Anfangsprofil
- FTCS
- Lax-Wendroff
- fin. Diff., 4-P.-Upw., $q = 0.5$
- Crank-Nicolson, fin. Diff.
- Crank-N., Mass.-Op., $\delta = 0.25$
- Crank-N., 4-P.-Up., $q = 0.5$
- exakte Lösung

Abbildung 6.22.: Vergleich verschiedenen Verfahren für die Burgers-Gleichung mit $\nu = 0.1$ und $t = 6$. Die Gitter-Reynoldszahl ist $\mathrm{Re_G} = 2$ ($u = 1$). Die blau gestrichelten Kurven zeigen die zeitliche Entwicklung der exakten Lösung. Die Zeiten sind als Zahlen angegeben.

(a) $J = 40$, $\Delta x = 0.1$, $s = 0.02$, C $= 1$, $\mathrm{Re_G} = 50$, $\nu = 0.002$, $\nu_a = 0.25$, $t = 2$



(b) Legende

- Crank-Nicolson, fin. Diff.
- Crank-N., Mass.-Op., $\delta = 0.25$
- Crank-N., 4-P.-Up., $q = 0.5$
- Crank-N., fin. Diff., $s_a = 0.25$
- CN, MO, $\delta = 0.12$, $s_a = 0.25$
- CN, 4-PU, $q = 0.5$, $s_a = 0.25$
- exakte Lösung

Abbildung 6.23.: Einfuß der künstlichen Diffusion auf die verschiedenen Varianten des Crank-Nicolson-Verfahrens für die Burgers-Gleichung bei kleiner Viskosität, entsprechend einer hohen Gitter-Reynoldszahl von $\mathrm{Re_G} = 50$. Die Ergebnisse ohne künstliche Diffusion sind durch offene Symbole gekennzeichnet.

Courant-Zahl ist, die mit dem Referenzwert $u = 1$ gebildet wird. Für das Crank-Nicolson-Verfahren mit Massenoperator wurde $\delta = 0.12$ verwendet und für das Crank-Nicolson-Verfahren mit 4-Punkt-Upwind $q = 0.5$. Man sieht, daß man die Oszillationen durch eine geeignete Wahl von ν_a sehr stark unterdrücken kann, ohne daß sich die Front zu sehr verschmiert. Im betrachteten Fall funktioniert die künstliche Viskosität recht gut, weil die Funktion u außerhalb des Sprungs konstant ist.

Im konstanten Bereich macht eine zusätzliche Diffusion natürlich nichts aus. Man kann auch Schemata konstruieren, bei denen die künstliche Diffusion erst ab einem gewissen Gradienten der Funktion u wirksam wird.

6.6. Ausbreitung eines Verdichtungsstoßes

Als strömungsmechanisches Beispiel betrachten wir Ausbreitung eines senkrechten Verdichtungsstoßes unter der Annahme einer reibungsfreien Strömung. Bei Abwesenheit äußerer Kräfte und Wärmequellen gelten hierfür die Kontinuitäts-, Euler- und Energie-Gleichungen

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{u}) = 0, \tag{6.139a}$$

$$\frac{\partial (\rho \boldsymbol{u})}{\partial t} + \nabla \cdot (\rho \boldsymbol{u}\boldsymbol{u}) = -\nabla p, \tag{6.139b}$$

$$\frac{\partial}{\partial t}\left[\rho\left(\frac{\boldsymbol{u}^2}{2} + e\right)\right] + \nabla \cdot \left[\rho \boldsymbol{u}\left(\frac{\boldsymbol{u}^2}{2} + h\right)\right] = 0. \tag{6.139c}$$

Mit der Zustandsgleichung für ein ideales Gas

$$h = e + \frac{p}{\rho} = c_p T = \frac{\gamma}{\gamma - 1}\frac{p}{\rho}, \tag{6.140}$$

wobei $\gamma = c_p/c_v = 1.4$ das Verhältnis der spezifischen Wärmen ist, erhalten wir für den eindimensionalen Fall

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) = 0, \tag{6.141a}$$

$$\frac{\partial (\rho u)}{\partial t} + \frac{\partial}{\partial x}\left(\rho u^2 + p\right) = 0, \tag{6.141b}$$

$$\frac{\partial}{\partial t}\left(\frac{\rho u^2}{2} + \frac{p}{\gamma - 1}\right) + \frac{\partial}{\partial x}\left[u\left(\frac{\rho u^2}{2} + \frac{\gamma}{\gamma - 1}p\right)\right] = 0. \tag{6.141c}$$

Alle Gleichungen sind Advektionsgleichungen, die im vorliegenden Fall nichtlinear und gekoppelt sind.

Aus der integralen Form der Gleichungen werden normalerweise die Rankine-Hugoniot-Relationen hergeleitet, welche die Propagation von Verdichtungsstößen beschreiben. Hier gehen wir davon aus, daß sich das Gas auf der Seite 1 in Ruhe befindet ($u_1 = 0$) und bei dem Druck p_1 die Dichte ρ_1 besitzt. Aus der Zustandsgleichung folgt die Temperatur $T_1 = c_v(\gamma - 1)^{-1} p_1/\rho_1$. Aus einem Gebiet 2 mit einem Druck $p_2 > p_1$ breite sich nun eine Verdichtungswelle in das Gebiet 1 des ruhenden Fluids aus. Die Stärke des damit verbundenen Stoßes kann man durch den Drucksprung

$$z = \frac{p_2 - p_1}{p_1} > 0, \tag{6.142}$$

auch *Stoßstärke* genannt, charakterisieren. Für den senkrechten Verdichtungsstoß lassen sich dann aus den Erhaltungsgleichungen für Masse, Impuls- und Energie die Dichte ρ_2, die Geschwindigkeit u_2 und die Geschwindigkeit U, mit der der Stoß propagiert, als alleinige Funktionen des Parameters z ausdrücken (siehe z.B. Chapman, 2000)

$$\frac{\rho_2}{\rho_1} = \frac{1 + \left(\frac{\gamma+1}{2\gamma}\right) z}{1 + \left(\frac{\gamma-1}{2\gamma}\right) z}, \tag{6.143a}$$

$$\frac{u_2}{c_1} = \frac{z/\gamma}{\sqrt{1 + \frac{\gamma+1}{2\gamma} z}}, \tag{6.143b}$$

$$\frac{U}{c_1} = \sqrt{1 + \frac{\gamma+1}{2\gamma} z}. \tag{6.143c}$$

Hierbei ist $c_1 = \sqrt{\gamma p_1/\rho_1}$ die Schallgeschwindigkeit im ruhenden Medium 1.

Zur Darstellung ist es bequem, den Druck p, die Dichte ρ und die Geschwindigkeit u auf die Größen p_1, ρ_1 und c_1 des ruhenden Mediums zu beziehen. Wenn wir außerdem die Längenskala L einführen, ergeben sich die dimensionslosen Variablen

$$p' = \frac{p}{p_1}, \qquad \rho' = \frac{\rho}{\rho_1}, \qquad u' = \frac{u}{c_1}, \qquad x' = \frac{x}{L}, \qquad t' = \frac{t}{L/c_1}. \tag{6.144}$$

Wenn wir den Strich wieder weglassen, lauten die skalierten Gleichungen

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) = 0, \tag{6.145a}$$

$$\frac{\partial (\rho u)}{\partial t} + \frac{\partial}{\partial x}\left(\rho u^2 + \frac{p}{\gamma}\right) = 0, \tag{6.145b}$$

$$\frac{\partial}{\partial t}\left(\frac{\rho u^2}{2} + \frac{p}{\gamma(\gamma-1)}\right) + \frac{\partial}{\partial x}\left[u\left(\frac{\rho u^2}{2} + \frac{p}{\gamma-1}\right)\right] = 0. \tag{6.145c}$$

Man kann die Gleichungen nun als eine einzige Advektionsgleichung der Form

$$\frac{\partial \boldsymbol{q}}{\partial t} + \frac{\partial \boldsymbol{F}}{\partial x} = 0 \tag{6.146}$$

schreiben, und zwar für die vektoriellen Größen

$$\boldsymbol{q} = \begin{pmatrix} \rho \\ \rho u \\ \frac{\rho u^2}{2} + \frac{p}{\gamma(\gamma-1)} \end{pmatrix} \qquad \text{und} \qquad \boldsymbol{F} = \begin{bmatrix} \rho u \\ \rho u^2 + p/\gamma \\ u\left(\frac{\rho u^2}{2} + \frac{p}{\gamma-1}\right) \end{bmatrix}. \tag{6.147}$$

Beachte, daß (6.146) nichtlinear ist, da $\boldsymbol{F}$ und $\boldsymbol{q}$ keine unabhängigen Größen sind. Vielmehr sind sie über ρ, u und p nichtlinear miteinander gekoppelt.

Eine Lösungsmöglichkeit besteht in der Anwendung des zweistufigen *Lax-Wendroff-Schemas* (6.125). Für die hier auftretenden vektoriellen Größen lautet es

$$\boldsymbol{q}^*_{j+1/2} = \frac{1}{2}\left(\boldsymbol{q}^n_j + \boldsymbol{q}^n_{j+1}\right) - \frac{\Delta t}{2\Delta x}\left(\boldsymbol{F}^n_{j+1} - \boldsymbol{F}^n_j\right), \tag{6.148a}$$

$$\boldsymbol{q}^{n+1}_j = \boldsymbol{q}^n_j - \frac{\Delta t}{\Delta x}\left(\boldsymbol{F}^*_{j+1/2} - \boldsymbol{F}^*_{j-1/2}\right). \tag{6.148b}$$

Da $\boldsymbol{F}$ und $\boldsymbol{q}$ über ρ, u und p zusammenhängen, muß man in jedem Teilschritt ρ, u und p aus $\boldsymbol{q}$ berechnen und daraus $\boldsymbol{F}$ bestimmen (und umgekehrt).

Robert W. MacCormack

Alternativ dazu wird auch das *MacCormack-Schema*[32]

$$\boldsymbol{q}^*_j = \boldsymbol{q}^n_j - \frac{\Delta t}{\Delta x}\left(\boldsymbol{F}^n_{j+1} - \boldsymbol{F}^n_j\right), \tag{6.149a}$$

$$\boldsymbol{q}^{n+1}_j = \frac{1}{2}\left(\boldsymbol{q}^n_j + \boldsymbol{q}^*_j\right) - \frac{\Delta t}{2\Delta x}\left(\boldsymbol{F}^*_j - \boldsymbol{F}^*_{j-1}\right). \tag{6.149b}$$

eingesetzt. Im Gegensatz zu Lax-Wendroff werden bei MacCormack einseitige Differenzen verwendet. Der dadurch verursachte führende Fehler der Ordnung $O(\Delta x)$ hebt sich aber bei den beiden Schritten gerade auf, so daß das Schema ebenfalls zweiter Ordnung im Raum und Zeit ist. Für lineare Probleme ($\boldsymbol{F} = \boldsymbol{q}$) reduziert sich das MacCormack-Schema auf das einstufige Lax-Wendroff-Verfahren (6.67).

Als Stabilitätsbedingung kann man für das Lax-Wendroff- und das MacCormack-Verfahren

$$|\lambda_k| \frac{\Delta t}{\Delta x} \le 1 \tag{6.150}$$

ableiten, wobei λ_k die Eigenwerte der Jacobi-Matrix $\partial F_i/\partial q_j$ sind. Für das vorliegende Problem erhält man die 3 Eigenwerte u, $u + c$ und $u - c$ (siehe Richtmyer and Morton, 1967). Daher lautet die Stabilitätsbedingung hier

$$(|u| + c) \frac{\Delta t}{\Delta x} \le 1. \tag{6.151}$$

Ein Beispiel für die Stoßstärke $z = 1.5$ ist in Abb. 6.24 gezeigt. Dabei wurde die Skalierung (6.144) verwendet. Beide Schemata zeigen einen starken Dispersionsfehler durch die hohen Harmonischen, die in der spektralen Darstellung des Sprungs enthalten sind.

Man kann den Dispersionsfehler glätten, indem man eine künstliche Viskosität einführt. Diese sollte nur dort wirken, wo sehr hohe Gradienten der Feldgrößen auftreten. Dazu hat sich die Erweiterung der Grundgleichungen auf

$$\frac{\partial \boldsymbol{q}}{\partial t} + \frac{\partial \boldsymbol{F}}{\partial x} - \nu \Delta x^2 \frac{\partial}{\partial x}\left(\left|\frac{\partial \boldsymbol{q}}{\partial x}\right| \frac{\partial \boldsymbol{q}}{\partial x}\right) = 0 \tag{6.152}$$

[32] Gleichungen (14.49) und (14.50) in Fletcher (1991*b*) enthalten Druckfehler.

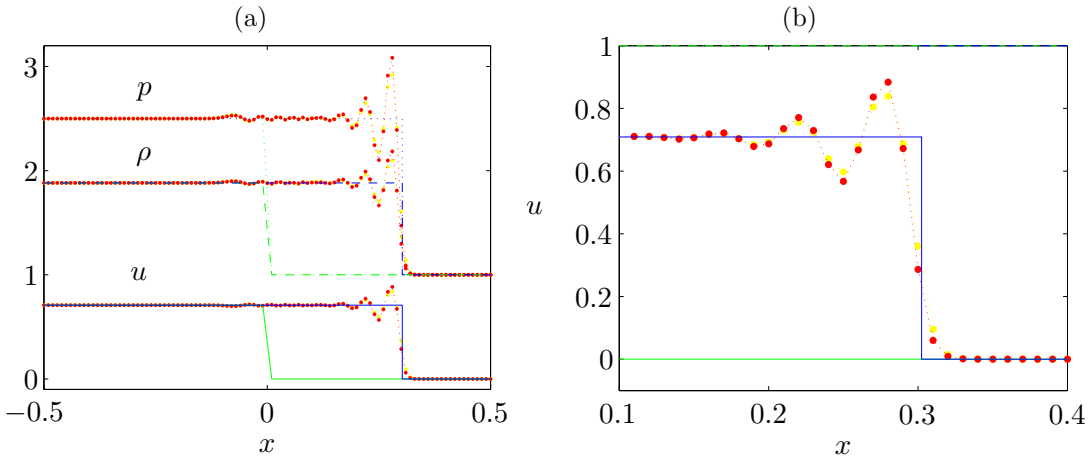


Abbildung 6.24.: Entwicklung eines Verdichtungsstoßes für Stoßstärke $z = 1.5$. Die Anfangsbedingung zur Zeit $t = 0$ ist grün und die exakte Lösung bei $t = 0.2$ ist blau dargestellt. Numerische Ergebnisse für $\Delta x = 0.01$ und $\Delta t = 0.002$ sind in rot (Lax-Wendroff) und in gelb (MacCormack) gezeigt. (a) zeigt den Überblick, (b) Details für u.

bewährt. Wegen der Betragsstriche hängt die künstliche Viskosität nicht von der Richtung des Sprungs ab. Außerdem wird der Gradient quadratisch *bestraft*. Bei der Implementierung der künstlichen Viskosität wird diese dem Lax-Wendroff- oder MacCormack-Verfahren als dritter Schritt nachgeschaltet. Aus dem Lax-Wendroff-Verfahren erhält man so das modifizierte Schema

$$\boldsymbol{q}^{*}_{j+1/2} = \frac{1}{2}\left(\boldsymbol{q}^{n}_{j} + \boldsymbol{q}^{n}_{j+1}\right) - \frac{\Delta t}{2\Delta x}\left(\boldsymbol{F}^{n}_{j+1} - \boldsymbol{F}^{n}_{j}\right), \tag{6.153a}$$

$$\boldsymbol{q}^{**}_{j} = \boldsymbol{q}^{n}_{j} - \frac{\Delta t}{\Delta x}\left(\boldsymbol{F}^{*}_{j+1/2} - \boldsymbol{F}^{*}_{j-1/2}\right), \tag{6.153b}$$

$$\begin{aligned}\boldsymbol{q}^{n+1}_{j} &= \boldsymbol{q}^{**}_{j} + \nu\frac{\Delta t}{\Delta x}\Delta\left\{\left|\Delta\boldsymbol{q}^{**}_{j+1}\right|\Delta\boldsymbol{q}^{**}_{j+1}\right\} \\ &= \boldsymbol{q}^{**}_{j} + \nu\frac{\Delta t}{\Delta x}\left[\left|\boldsymbol{q}^{**}_{j+1} - \boldsymbol{q}^{**}_{j}\right|\left(\boldsymbol{q}^{**}_{j+1} - \boldsymbol{q}^{**}_{j}\right) - \left|\boldsymbol{q}^{**}_{j} - \boldsymbol{q}^{**}_{j-1}\right|\left(\boldsymbol{q}^{**}_{j} - \boldsymbol{q}^{**}_{j-1}\right)\right]\end{aligned} \tag{6.153c}$$

Nach Richtmyer and Morton (1967) wird durch die künstliche Viskosität der stabile Bereich der zeitlichen Schrittweite aber weiter eingeschränkt (siehe auch Fletcher, 1991*b*). Der Effekt der künstlichen Viskosität ist für das obige Beispiel an Hand des Lax-Wendroff-Verfahrens in Abb. 6.25 für die Werte $\nu = 0$, 0.5 und 1 illustriert. Hierbei wurde die künstliche Viskosität auf alle drei Komponenten von $\boldsymbol{q}$ angewandt. Man sieht, daß die künstliche Viskosität einen geringen Einfluß auf die Strömung weit weg vom Stoß hat. In der Dichte werden jedoch einige der weiter entfernten künstlichen Oszillationen noch etwas verstärkt.

Für starke Stöße liefert das sogenannte fluß-korrigierte Transport-Schema (FCT-Schema, ***f**lux **c**orrected **t**ransport scheme*) bessere Ergebnisse. Dieser Ansatz beruht darauf, in einem ersten Schritt eine relativ große künstliche Diffusion zu verwenden und dies dann im zweiten Schritt durch eine fast gleich große *Anti-Diffusion* wieder

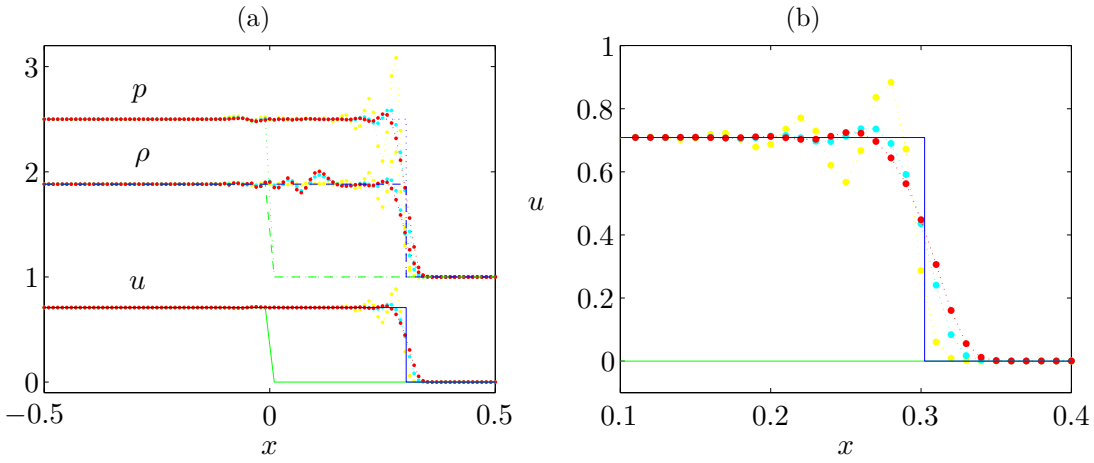


Abbildung 6.25.: Einfluß der künstlichen Viskosität auf die numerische Lösung des Verdichtungsstoß-Problems mittels des nach (6.153) modifizierten Lax-Wendroff-Verfahrens. Der Wert der künstlichen Viskosität beträgt $\nu = 0$ (gelb), $\nu = 0.5$ (cyan) und $\nu = 1$ (rot). Die Parameter sind identisch mit denen aus Abb. 6.24.

zu kompensieren. Dabei wird die Anti-Diffusion so gestaltet, daß sie limitiert ist und daß keine neuen Extrema auftreten können. Dies ist in Kap. 14.2.6 und 14.2.7 von Fletcher (1991*b*) genauer beschrieben.

A. Eigenwerte einer tridiagonalen Matrix

Zur Berechnung von Eigenwerten muß man Determinanten berechnen. Daher betrachten wir zunächst die Determinante D_n einer tridiagonalen Matrix A der Ordnung n

$$D_n = \det|\mathsf{A}| = \det\begin{vmatrix} a & b & & & & \\ c & a & b & & & \\ & & \cdot & & & \\ & & & \cdot & & \\ & & & c & a & b \\ & & & & c & a \end{vmatrix}. \tag{A.1}$$

Wenn man die Determinante nach Spalten entwickelt, erhält man

$$D_n = aD_{n-1} - bcD_{n-2}. \tag{A.2}$$

Hierbei sind D_{n-1} und D_{n-2} Unterdeterminanten der Ordnung $n-1$ und $n-2$. Sie haben dieselbe Form wie D_n, nur mit verringerter Ordnung. Gleichung (A.2) ist eine Rekursionsformel für D_n, wobei wir die Anfangsbedingungen

$$D_0 = 1 \qquad \text{und} \qquad D_1 = a \tag{A.3}$$

beachten müssen.[1] Die Rekursionsgleichung (A.2) kann man mit dem Potenzansatz

$$D_n = \mu^n \tag{A.4}$$

lösen. Eingesetzt erhalten wir für μ

$$\mu^2 = a\mu - bc. \tag{A.5}$$

Die Lösungen dieser quadratischen Gleichung lauten

$$\mu_\pm = \frac{a}{2} \pm \sqrt{\frac{a^2}{4} - bc}. \tag{A.6}$$

Da wir zwei Wurzel haben, lautet der vollständige Ansatz für D_n

$$D_n = \alpha\mu_+^n + \beta\mu_-^n, \tag{A.7}$$

[1] Die bekannteste derartige Rekursionsformel ist $F_n = F_{n-1} + F_{n-2}$. Mit $F_0 = F_1 = 1$ generiert

wobei α und β aus den Anfangsbedingungen bestimmt werden müssen. Man kann leicht verifizieren, daß die Anfangsbedingungen (A.3) von der Lösung

$$D_n = \frac{\mu_+^{n+1} - \mu_-^{n+1}}{\mu_+ - \mu_-} \tag{A.8}$$

erfüllt werden.

Wir betrachten nun das eigentliche Eigenwertproblem

$$\mathsf{A} - \lambda\,\mathsf{I} = 0. \tag{A.9}$$

Die Lösbarkeitsbedingung erfordert

$$D_n = \det \begin{vmatrix} a-\lambda & b & & & & \\ c & a-\lambda & b & & & \\ & & \cdot & & & \\ & & & \cdot & & \\ & & & c & a-\lambda & b \\ & & & & c & a-\lambda \end{vmatrix} = 0. \tag{A.10}$$

Nach der allgemeinen Lösung für die Determinate tridiagonaler Matrizen (A.8) folgt dann

$$\mu_+^{n+1} = \mu_-^{n+1}. \tag{A.11}$$

Mit $\mu_\pm$ entsprechend (A.6) ($a \to a - \lambda$) ist diese Bedingung äquivalent zu

$$\left[\frac{(a-\lambda) + \sqrt{(a-\lambda)^2 - 4bc}}{(a-\lambda) - \sqrt{(a-\lambda)^2 - 4bc}}\right]^{n+1} = 1. \tag{A.12}$$

Mit der Substitution

$$\cos\varphi = \frac{a-\lambda}{2\sqrt{bc}} \tag{A.13}$$

lautet die Lösbarkeitsbedingung dann

$$\Big[\ldots\Big]^{n+1} = \left(\frac{\cos\varphi + \mathrm{i}\sin\varphi}{\cos\varphi - \mathrm{i}\sin\varphi}\right)^{n+1} = \left(\frac{\mathrm{e}^{\mathrm{i}\varphi}}{\mathrm{e}^{-\mathrm{i}\varphi}}\right)^{n+1} = \mathrm{e}^{2\mathrm{i}\varphi(n+1)} = 1. \tag{A.14}$$

Diese Gleichung hat die Wurzeln

$$\varphi = \frac{k\pi}{n+1}, \qquad k = 1, \ldots, n. \tag{A.15}$$

Wenn wir nun (A.13) nach λ auflösen, erhalten wir die Eigenwerte[2]

$$\lambda_k = a + 2\sqrt{bc}\cos\left(\frac{k\pi}{n+1}\right), \qquad k = 1, \ldots, n. \tag{A.16}$$

sie die Fibonacci-Zahlen und steht im Zusammenhang mit dem goldenen Schnitt, der durch die positive Lösung von (A.6) für $a = b = -c = 1$ als $\mu_+ = (1+\sqrt{5})/2$) gegeben ist.

[2]Das Vorzeichen vor dem zweiten Summanden spielt keine Rolle, da die Eigenwerte symmetrisch

bzgl. $a \pm ...$ liegen.

B. Ritz-Verfahren

Bei dem *Ritzschen Verfahren* such man ein Minimum einer skalaren Größe $\mathcal{E}$, die von einer Funktion $f(\boldsymbol{x})$ (die auch vektorwertig sein kann) abhängt, wobei die Funktion als Variable fungiert. Eine derartige Größe nennt man *Funktional* $\mathcal{E}[f(\boldsymbol{x})]$. Die Abhängigkeit von der Funktion wird in eckigen Klammern geschrieben. Insbesondere will man wissen, wie die Funktion $f_0(\boldsymbol{x})$ aussieht, bei der das Funktional $\mathcal{E}$ minimal wird.

Für das Rechnen mit Funktionalen gelten ähnliche Regeln wie für die normale Differentiation (Produkt- und Kettenregel). Insbesondere gilt für die *erste Variation* (die Variation wird mit dem Symbol δ bezeichnet)

$$\delta\mathcal{E}(f) = \frac{\mathrm{d}\mathcal{E}}{\mathrm{d}f}\delta f \tag{B.1}$$

Die formale Ableitung $\mathrm{d}\mathcal{E}/\mathrm{d}f$ nennt man *Funktionalableitung*. Für ein Minimum von $\mathcal{E}$ muß die erste Variation des Funktionals $\delta\mathcal{E} = 0$ verschwinden. Dies führt im allgemeinen zu den *Euler-Lagrange-Gleichungen* für das Minimierungsproblem. Die Euler-Lagrange-Gleichungen sind Differentialgleichungen für f, deren Lösung die gesuchte Funktion $f_0(\boldsymbol{x})$ ist.

Wenn aber die Variation δf eingeschränkt ist, indem man *fest vorgegebene* Ansatzfunktionen $\phi_i(\boldsymbol{x})$ verwendet mit

$$f_A(\boldsymbol{x}) = \sum_{i=1}^{N} a_i\phi_i(\boldsymbol{x}) \tag{B.2}$$

und nur die Parameter a_i variiert, erhält man mit

$$\delta f_A = \delta\sum_{i=1}^{N} a_i\phi_i(\boldsymbol{x}) = \sum_{i=1}^{N}\phi_i(\boldsymbol{x})\delta a_i \tag{B.3}$$

die Extremal-Bedingung

$$\delta\mathcal{E}(f_A) = \frac{\mathrm{d}\mathcal{E}}{\mathrm{d}f_A}\delta f_A = \frac{\mathrm{d}\mathcal{E}}{\mathrm{d}f_A}\sum_{i=1}^{N}\phi_i(\boldsymbol{x})\delta a_i = 0. \tag{B.4}$$

Da die Parameter a_i unabhängig voneinander gewählt werden können, muß jeder Summand einzeln gleich Null sein. Deshalb erhalten wir N Gleichungen zur Bestimmung der a_i (und damit der minimierenden Funktion) in der Form

$$\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}f_A}\phi_i(\boldsymbol{x}) = 0. \tag{B.5}$$

Hierbei ist zu beachten, daß $\mathrm{d}\mathcal{E}/\mathrm{d}f_A$ natürlich von den vorgegebenen Ansatzfunktionen $\{\phi_i\}$ und den zu bestimmenden Koeffizienten $\{a_i\}$ abhängt.

Das obige Gleichungssystem (B.5) ist ein spezieller Fall, der für den Ansatz (B.2) gilt. Im allgemeinen müssen die Unbekannten nicht in Form von Koeffizienten in einer Linearkombination von Ansatzfunktionen auftreten. Die Unbekannten könnten z.B. auch Exponenten sein. Welche Koeffizienten man bestimmen will hängt einzig und allein von dem jeweiligen Ansatz ab, bei dem man eine große Freiheit hat.

Das Variationsproblem $\delta\mathcal{E} = 0$, bei dem die Funktion f variiert wird, führt mit dem Ansatz $f = f_A$ bei bekannten Ansatzfunktionen und variablen (zu bestimmenden) Parametern a_i zu den allgemeinen Bestimmungsgleichungen der Form

$$\delta_{a_i}\mathcal{E}[f_A] = \frac{\mathrm{d}\mathcal{E}[f_A]}{\mathrm{d}a_i}\mathrm{d}a_i = 0, \tag{B.6}$$

bei unabhängigen Parametern a_i also auf

$$\frac{\mathrm{d}\mathcal{E}[f_A]}{\mathrm{d}a_i} = 0. \tag{B.7}$$

C. Ableitungsoperatoren im Chebyshev-Raum

Wenn wir die Chebyshev-Polynome ableiten, erhalten wir

$$\mathrm{T}_k'(x) = \frac{\mathrm{d}}{\mathrm{d}x}\left[\cos(k\theta)\right] = \frac{\mathrm{d}\cos(k\theta)}{\mathrm{d}\theta}\frac{\mathrm{d}\theta}{\mathrm{d}x} = k\sin(k\theta)\underbrace{\frac{\mathrm{d}\arccos x}{\mathrm{d}x}}_{\sin^{-1}x} = k\frac{\sin(k\theta)}{\sin x}. \tag{C.1}$$

Wegen $\sin[(k+1)\theta] - \sin[(k-1)\theta] = 2\sin\theta\cos(k\theta)$ kann man die Rekursionsformel

$$\frac{\mathrm{T}_{k+1}'}{k+1} - \frac{\mathrm{T}_{k-1}'}{k-1} = 2\mathrm{T}_k \tag{C.2}$$

leicht verifizieren. Wenn man diese Rekursionsformel nach T_{k+1}' auflöst und umnumeriert ($k \to k-1$), dann erhält man für die Ableitung

$$\mathrm{T}_k'(x) = 2k\left(\mathrm{T}_{k-1} + \frac{1}{2}\frac{1}{k-2}\underbrace{\mathrm{T}_{k-2}'}_{=2(k-2)[\mathrm{T}_{k-3}+\ldots]}\right) = 2k\sum_{n=0}^{(k-1)/2}\frac{\mathrm{T}_{k-1-2n}(x)}{c_{k-1-2n}}. \tag{C.3}$$

Nun schreiben wir die Ableitung unserer unbekannten Funktion formal als

$$u'(x) = \sum_{k=0}^{N}\hat{u}_k\mathrm{T}_k'(x) = \sum_{k=0}^{N}\hat{u}_k^{(1)}\mathrm{T}_k(x) \tag{C.4}$$

und setzen (C.3) für $\mathrm{T}_k'(x)$ ein. Dann erhalten wir, bei Umbenennung des Laufindex $k \to p$,

$$\sum_{k=0}^{N}\hat{u}_k^{(1)}\mathrm{T}_k(x) = \sum_{p=0}^{N}\hat{u}_p\mathrm{T}_p'(x) \overset{(C.3)}{=} \sum_{p=0}^{N}\hat{u}_p 2p\sum_{n=0}^{(p-1)/2}\frac{\mathrm{T}_{p-1-2n}(x)}{c_{p-1-2n}}. \tag{C.5}$$

Wenn wir jetzt die Koeffizienten von $\mathrm{T}_k(x)$ vergleichen, erhalten wir die Ableitungskoeffizienten $\hat{u}_k^{(1)}$. Insbesondere liefert auf der rechten Seite der Gleichung für festes p nur der Wert von n einen Beitrag, für den $p-1-2n = k$ ist. Also fällt die innere Summe weg, und wir müssen für p fordern: $p = k+1+2n$. Dies führt auf die Darstellung

$$\hat{u}_k^{(1)} = \sum_{p=k+1,\ p+k\ \text{odd}}^{N}\frac{2p}{c_k}\hat{u}_p. \tag{C.6}$$

Dieses Gleichungssystem entspricht aber einer Darstellung der Koeffizienten der ersten Ableitung mittels einer Matrixmultiplikation der Koeffizienten $\hat{u}_k$ selbst. Dies definiert die Ableitungsmatrix $\mathcal{D}$. Sie besitzt eine obere Dreiecksstruktur. Damit erhalten wir die Ableitung im spektralen Raum als einfache Matrixmultiplikation

$$\hat{\boldsymbol{u}}^{(1)} = \tilde{\mathcal{D}} \cdot \hat{\boldsymbol{u}}. \tag{C.7}$$

Die zweite Ableitung ergibt sich dementsprechend als

$$\hat{\boldsymbol{u}}^{(2)} = \tilde{\mathcal{D}} \cdot \left(\tilde{\mathcal{D}} \cdot \hat{\boldsymbol{u}} \right) = \tilde{\mathcal{D}}^2 \cdot \hat{\boldsymbol{u}}. \tag{C.8}$$

D. Aliasing bei Fourier-Kollokation

Um das Aliasing zu demonstrieren, betrachten wir wieder den nichtlinearen Term $u\partial_x u$. Der Einfachheit halber wählen wir hier die Fourierdarstellung von u und folgen der Darstellung von Peyret (2002)

$$u(x,t) = \sum_{k=-K}^{K} \hat{u}_k(t) \mathrm{e}^{\mathrm{i}kx}. \tag{D.1}$$

Bei der Galerkin-Behandlung der in der Zeit diskretisierten Burgers-Gleichung (4.164) hätten wir im Fourier-Raum für jede Fouriermode $k = -K, ..., K$ die Gleichung

$$\left(\frac{1}{\Delta t} + \nu k^2\right) \hat{u}_k^{n+1} = \frac{\hat{u}_k^n}{\Delta t} - \hat{w}_k \tag{D.2}$$

zu lösen, wobei $\hat{w}_k$ die spektrale Darstellung des nichtlineare Terms $w = u\partial_x u$ ist. Im Ortsraum lautet der Term

$$w(x) := u(x)\frac{\partial u(x)}{\partial x} = \left(\sum_{p=-K}^{K} \hat{u}_p(t)\mathrm{e}^{\mathrm{i}px}\right)\left(\sum_{q=-K}^{K} \mathrm{i}q\hat{u}_q(t)\mathrm{e}^{\mathrm{i}qx}\right). \tag{D.3}$$

Galerkin-Methode Im Fourier-Raum lautet der nichtlineare Term (Fourier-Transformation) im Rahmen einer *Galerkin-Darstellung*, bei der die Nichtlinearität auf die *kontinuierlichen* Gewichts- bzw. Ansatzfunktionen $\mathrm{e}^{\mathrm{i}kx}$ projiziert werden,

$$\hat{w}_k = \frac{1}{2\pi}\int_0^{2\pi} u\frac{\partial u}{\partial x}\,\mathrm{e}^{-\mathrm{i}kx}\,\mathrm{d}k = \sum_{p=-K}^{K}\sum_{q=-K}^{K} \mathrm{i}q\hat{u}_p\hat{u}_q \underbrace{\frac{1}{2\pi}\int_0^{2\pi} \mathrm{e}^{\mathrm{i}(p+q-k)x}\,\mathrm{d}k}_{\delta_{k,p+q}}, \tag{D.4}$$

also

$$\hat{w}_k = \sum_{\substack{p,q=-K \\ p+q=k}}^{K} \mathrm{i}q\hat{u}_p\hat{u}_q. \tag{D.5}$$

Bei der Produktbildung werden Produkte und Differenzen der Wellenzahlen generiert ($\mathrm{e}^{\mathrm{i}(p+q)x}$). Die Wellenzahlen, die außerhalb des betrachteten Spektrums liegen ($p+q > K$ oder $p+q < -K$) werden beim kontinuierlichen Galerkin-Verfahren herausgefiltert. Beim diskreten pseudospektralen Verfahren können diese Moden aber Amplituden von Moden innerhalb des betrachteten Spektrums $-K \le k \le K$ verfälschen.

Kollokation Um dieses **Aliasing** zu demonstrieren, betrachten wir nun die diskrete Darstellung von w. Da $w(x_j)$ durch die pseudospektrale Methode verfälscht wird, nennen wir die pseudospektralen Amplituden $\tilde{w}_k$, im Gegensatz zu den korrekten Galerkin-Amplituden $\hat{w}_k$. Es sei also

$$w(x_j) = \sum_{k=-K}^{K} \tilde{w}_k \mathrm{e}^{\mathrm{i}kx_j} \tag{D.6}$$

die diskrete Fourierdarstellung von $w(x)$ mit $w(x_j) = u(x_j)\partial_x u(x_j)$. Um zu sehen, woher der Fehler in $\tilde{w}_k$ kommt, gehen wir die Berechnung der pseudospektralen Methode rückwärts durch. Die umgekehrte diskrete Fouriertransformation lautet (siehe auch (2.36) und (2.37))

$$\tilde{w}_k \overset{(1)}{=} \frac{1}{N}\sum_{j=1}^{N} w(x_j)\mathrm{e}^{-\mathrm{i}kx_j} \tag{D.7}$$

$$\overset{(2)}{=} \frac{1}{N}\sum_{j=1}^{N}\left(\sum_{p=-K}^{K}\hat{u}_p\mathrm{e}^{\mathrm{i}px_j}\right)\left(\sum_{q=-K}^{K}\mathrm{i}q\hat{u}_q\mathrm{e}^{\mathrm{i}qx_j}\right)\mathrm{e}^{-\mathrm{i}kx_j} \tag{D.8}$$

$$\overset{(3)}{=} \frac{\mathrm{i}}{N}\sum_{j=1}^{N}\sum_{p=-K}^{K}\sum_{q=-K}^{K} q\hat{u}_q\hat{u}_p\mathrm{e}^{\mathrm{i}(p+q-k)x_j}. \tag{D.9}$$

Dabei ist (1) die diskrete Fourier-Transformation von $w(x_j)$, (2) die Produktbildung von $u(x_j)$ und $\partial_x u(x_j)$ im Ortsraum und (3) die Transformation in den Fourier-Raum. Mit der diskreten Orthogonalitätsrelation ($x_j = 2\pi j/N$)

$$\sum_{j=1}^{N}\mathrm{e}^{\mathrm{i}(k-l)2\pi j/N} = \begin{cases} N, & k-l = mN, m\in\mathbb{Z}, \\ 0, & \text{sonst.}\end{cases} \tag{D.10}$$

erhalten wir

$$\tilde{w}_k = \underbrace{\sum_{\substack{p,q=-K\\ p+q=k}}^{K}\mathrm{i}q\hat{u}_p\hat{u}_q}_{\hat{w}_k} + \underbrace{\sum_{\substack{p,q=-K\\ p+q=k+N}}^{K}\mathrm{i}q\hat{u}_q\hat{u}_p + \sum_{\substack{p,q=-K\\ p+q=k-N}}^{K}\mathrm{i}q\hat{u}_q\hat{u}_p}_{\text{aliasing}}. \tag{D.11}$$

Durch Vergleich mit (D.5) können die Aliasing-Terme identifiziert werden. Sie können mit Hilfe der 3/2-Regel eliminiert werden Peyret (2002). Dazu wird das Spektren nur für die Terme im nichtlinearen Ausdruck formal bis auf K' erweitert:

$$\hat{u}'_q = \begin{cases} \hat{u}_q & \text{if} \quad |q| \le K, \\ 0 & \text{if} \quad K < |q| \le K'. \end{cases} \tag{D.12}$$

Dann laufen die Summen über p und q in den Aliasing-Termen in (D.11) von $-K'$ bis K' mit $p+q = k \pm N'$ und $N' = 2K'+1$. Man kann sich dann überlegen, daß immer ein Faktor $\hat{u}'_q$ oder $\hat{u}'_p$ in (D.11) Null ist, wenn $K' \ge 3K/2$ ist (siehe S. 35 in Peyret, 2002).

E. Exakte Lösung der eindimensionalen Transportgleichung

Um eine exakte Lösung der Transportgleichung (6.94) zu finden, nehmen wir eine konstante Advektionsgeschwindigkeit $u = \text{const.}$ an und transformieren die Transportgleichung

$$\frac{\partial \bar{T}}{\partial t} + u\frac{\partial \bar{T}}{\partial x} = \kappa \frac{\partial^2 \bar{T}}{\partial x^2} \tag{E.1}$$

zunächst in das mit u mitbewegte Koordinatensystem

$$X = x - ut, \qquad \tau = t. \tag{E.2}$$

Im bewegten Koordinatensystem ist die Konvektionsgleichung nur noch eine Diffusionsgleichung. Man erhält[1]

$$\frac{\partial \bar{T}}{\partial \tau} - \kappa \frac{\partial^2 \bar{T}}{\partial X^2} = 0. \tag{E.3}$$

Mit dem Separationsansatz

$$\bar{T}(X, \tau) = f(\tau) g(X) \tag{E.4}$$

erhalten wir nach Division durch fg

$$\frac{f'(\tau)}{f(\tau)} = \kappa \frac{g''(X)}{g(X)}. \tag{E.5}$$

Jetzt hängt die linke Seite der Gleichung nur von τ und die rechte nur von X ab. Damit diese Gleichung aber für beliebige Werte von X und τ gelten muß, müssen beide Seiten konstant und identisch sein. Dann erhalten wir für $f(\tau)$ die Form

$$f(t) = \mathrm{e}^{\lambda \tau}. \tag{E.6}$$

[1] Beachte, daß sich die Ableitungen nach den alten Koordinaten (x, t) dann folgendermaßen durch die Ableitungen nach den neuen (X, τ) ausdrücken lassen

$$\frac{\partial}{\partial x} = \underbrace{\left(\frac{\partial \tau}{\partial x}\right)}_{=0} \frac{\partial}{\partial \tau} + \underbrace{\left(\frac{\partial X}{\partial x}\right)}_{=1} \frac{\partial}{\partial X} = \frac{\partial}{\partial X},$$

Die resultierende Gleichung für $g(X)$ lautet

$$\kappa g'' = \lambda g. \tag{E.7}$$

Der Exponentialansatz $g(X) = \mathrm{e}^{\mathrm{i}p\pi X/L}$ (p : Zahlenwert, L : Referenzlänge) ergibt $\lambda = -\kappa\pi^2 p^2/L^2$. Aus Symmetriegründen können wir hier auf $g(X) = \sin(p\pi X)$ spezialisieren. Damit lautet die Lösung für eine räumliche Fourier-Mode mit Wellenzahl $p\pi/L$

$$\bar{T} = A(p)\mathrm{e}^{-\kappa p^2\pi^2\tau/L^2}\sin(p\pi X/L) = A(p)\mathrm{e}^{-\kappa p^2\pi^2 t/L^2}\sin\left[p\pi(x-ut)/L\right]. \tag{E.8}$$

Wegen der Linearität der Diffusionsgleichungen können wir nun beliebige Fourier-Moden superponieren, um die Lösung für unser anfängliches Stufenprofil zu erhalten. Dazu nutzen wir die spektrale Darstellung des Stufenprofils (6.115)

$$\bar{T}(X,0) = \frac{1}{2} + \sum_{\substack{p=1\\ p\ \text{odd}}}^{\infty} \underbrace{\left(-\frac{2}{p\pi}\right)}_{A_p}\sin(p\pi X/L) = \frac{1}{2} - \sum_{k=1}^{\infty}\frac{2}{(2k-1)\pi}\sin\left[(2k-1)\pi X/L\right]. \tag{E.9}$$

Wenn wir dies einsetzen, erhalten wir die Lösung der Transportgleichung für ein anfängliches Stufenprofil

$$\bar{T} = \frac{1}{2} - \sum_{k=1}^{\infty}\frac{2}{(2k-1)\pi}\sin\left[(2k-1)\pi X/L\right]\mathrm{e}^{-\kappa(2k-1)^2\pi^2 t/L^2}. \tag{E.10}$$

$$\frac{\partial}{\partial t} = \underbrace{\left(\frac{\partial\tau}{\partial t}\right)}_{=1}\frac{\partial}{\partial\tau} + \underbrace{\left(\frac{\partial X}{\partial t}\right)}_{=-u}\frac{\partial}{\partial X} = \frac{\partial}{\partial\tau} - u\frac{\partial}{\partial X}.$$

F. Exakte Lösungen der Burgers-Gleichung

F.1. Anfangswertproblem in einer Dimension

Die Ableitung folgt dem Buch von Kevorkian and Cole (1981) bzw. Whitham (1974). Die Burgers-Gleichung ist die einfachste partielle Differentialgleichung, in der Nichtlinearität und Dispersion zusammenwirken. Sie ist ein elementares Beispiel auch für Struktur von Verdichtungsstößen. Sie besonders von Interesse, da man ihre Lösungen explizit angeben kann.

Eine Lösung der Burgers-Gleichung

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} - \nu\frac{\partial^2 u}{\partial x^2} = 0 \tag{F.1}$$

kann man erhalten, wenn man sie auf eine Diffusionsgleichung transformiert. Zunächst schreiben wir

$$u = \frac{\partial \phi}{\partial x}. \tag{F.2}$$

Einsetzen und Integration der Gleichung liefert dann die Gleichung

$$\frac{\partial \phi}{\partial t} + \frac{1}{2}\left(\frac{\partial \phi}{\partial x}\right)^2 - \nu\frac{\partial^2 \phi}{\partial x^2} = 0 \tag{F.3}$$

für ϕ. Den nichtlinearen Term kann man nun durch die Transformation

$$\phi = -2\nu \ln v \tag{F.4}$$

eliminieren. Wenn man dann

$$\frac{\partial \phi}{\partial t} = -\frac{2\nu}{v}\frac{\partial v}{\partial t}, \qquad \frac{\partial \phi}{\partial x} = -\frac{2\nu}{v}\frac{\partial v}{\partial x}, \qquad \frac{\partial^2 \phi}{\partial x^2} = -\frac{2\nu}{v}\frac{\partial^2 v}{\partial x^2} + \frac{2\nu}{v^2}\left(\frac{\partial v}{\partial x}\right)^2, \tag{F.5}$$

in (F.3) einsetzt, vereinfacht sich die Differentialgleichung zu der Diffusionsgleichung

$$\frac{\partial v}{\partial t} - \nu\frac{\partial^2 v}{\partial x^2} = 0. \tag{F.6}$$

Der Zusammenhang zwischen u und v ist dabei gegeben durch

$$u = -\frac{2\nu}{v}\frac{\partial v}{\partial x} = -2\nu\frac{\partial}{\partial x}\ln v. \tag{F.7}$$

Dies nennt man auch die Cole-Hopf-Transformation. Wenn nun die Anfangswerte für u in der Form $u(x,0) = F(x)$ auf $x \in [-\infty, \infty]$ gegeben sind, dann erhält man $v(x,0)$ durch Integration von (F.7) zu

$$v(x,0) = \exp\left(-\frac{1}{2\nu}\int_0^x F(s)\mathrm{d}s\right) := G(x). \tag{F.8}$$

Die Integrationskonstante wurde dabei so gewählt, daß $v(0,0) = 1$ ist. Diese willkürliche Wahl macht aber nichts aus, denn die Lösung $u(x,t)$ hängt nicht davon ab. Nun können wir die bekannte Lösung des Anfangswertproblems für die Diffusionsgleichung[1]

$$v(x,t) = \frac{1}{\sqrt{4\pi\nu t}}\int_{-\infty}^{\infty} G(s)\mathrm{e}^{-(x-s)^2/4\nu t}\mathrm{d}s \tag{F.9}$$

verwenden, um daraus mittels (F.7) die gesuchte Lösung

$$u(x,t) = \frac{\displaystyle\int_{-\infty}^{\infty}\left(\frac{x-s}{t}\right)G(s)\exp\left\{-(x-s)^2/4\nu t\right\}\mathrm{d}s}{\displaystyle\int_{-\infty}^{\infty} G(s)\exp\left\{-(x-s)^2/4\nu t\right\}\mathrm{d}s}. \tag{F.10}$$

zu erhalten.

F.1.1. Stationäre Lösung in einer Dimension

In einer Dimension kann man stationäre Lösungen der Burgers-Gleichung dadurch finden, indem man sie einmal integriert

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{u^2}{2} - \nu\frac{\mathrm{d}u}{\mathrm{d}x}\right) = 0. \tag{F.11}$$

[1] Die Lösung der Diffusionsgleichung zu der Anfangsbedingung $v(x,0) = \delta(x-s)$ (anfänglicher Delta-Peak bei $x = s$) lautet (siehe z.B. Landau and Lifschitz, 1991)

$$v(x,t) = \frac{1}{\sqrt{4\pi\nu t}}\mathrm{e}^{-(x-s)^2/4\nu t}.$$

Wir hatten sie in Abb. 6.1 schon verwendet. Eine allgemeinere Anfangsbedingung kann man darstellen als

$$v(x,0) = G(x) = \int_{-\infty}^{\infty} G(s)\delta(x-s)\mathrm{d}s.$$

Aufgrund der Linearität der Diffusionsgleichung kann man daher die obige Lösung für einen anfänglichen Delta-Peak mit Wichtungsfaktoren $G(s)$ zu der allgemeinen Lösung superponieren. Denn im Limes $t \to 0$ erhält man aus der allgemeinen Lösung (F.9) unter Beachtung der Darstellung der Delta-Funktion

$$\lim_{t\to 0}\frac{1}{\sqrt{4\pi\nu t}}\int_{-\infty}^{\infty} G(s)\mathrm{e}^{-(x-s)^2/4\nu t}\mathrm{d}s = \lim_{t\to 0}\int_{-\infty}^{\infty} G(s)\underbrace{\frac{\mathrm{e}^{-(x-s)^2/4\nu t}}{\sqrt{4\pi\nu t}}}_{\to\delta(x-s)}\mathrm{d}s = \int_{-\infty}^{\infty} G(s)\delta(x-s)\mathrm{d}s.$$

ν	$\sqrt{2c}$	c
10^{-2}	1	0.5
10^{-1}	1.00009	0.50009
1	1.5434	1.1910
10	4.50975	10.1689
100	14.1539	100.166

Tabelle F.1.: Werte für die Integrationskonstante c in Abhängigkeit von ν für die stationären Lösung der 1D Burgers-Gleichung zu den Randbedingungen $u(0) = 1$ und $u(1) = 0$.

Eine Möglichkeit, diese Gleichung zu erfüllen ist $u = \text{const}$. Die andere Möglichkeit besteht darin, daß der Ausdruck in der Klammer konstant ist. Dies führt auf

$$\frac{\mathrm{d}u}{\mathrm{d}x} = \frac{u^2}{2\nu} - \frac{c}{\nu} \quad \text{oder} \quad \frac{\mathrm{d}x}{2\nu} = \frac{\mathrm{d}u}{u^2 - 2c}, \tag{F.12}$$

mit der Integrationskonstante c. Wenn man dies integriert, erhält man

$$\frac{x - x_0}{2\nu} = -\frac{1}{\sqrt{2c}} \operatorname{arctanh}\left(\frac{u}{\sqrt{2c}}\right) \tag{F.13}$$

bzw.

$$u = \sqrt{2c} \tanh\left[\frac{\sqrt{2c}}{2\nu}\,(x_0 - x)\right]. \tag{F.14}$$

Hierbei sind c und x_0 zunächst beliebige Integrationskonstanten. Sie müssen so festgelegt werden, daß die Randbedingungen erfüllt sind. Sei zum Beispiel $u(0) = 1$ und $u(1) = 0$, dann folgt $x_0 = 1$ und c ist Lösung der transzendenten Gleichung

$$\frac{1}{\sqrt{2c}} = \tanh\left[\frac{\sqrt{2c}}{2\nu}\right]. \tag{F.15}$$

Die Lösungen sind für einige Werte von ν in Tabelle F.1 angegeben. Einige der zugehörigen stationären Lösungen sind in Abb. F.1 gezeigt. Man sieht, daß für $\nu \gg 1$ die Diffusion dominiert und das Profil im Limes $\nu \to \infty$ zu einem lineare Profil wird. Umgekehrt ist das Profil konvektiv dominiert, wenn $\nu \ll 1$ ist.

F.2. Stationäre Lösung in zwei Dimensionen

Die Methode der Cole-Hopf-Transformation kann man auch auf die zweidimensionale Burgers-Gleichung

$$\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{u} - \nu \nabla^2 \boldsymbol{u} = 0 \tag{F.16}$$

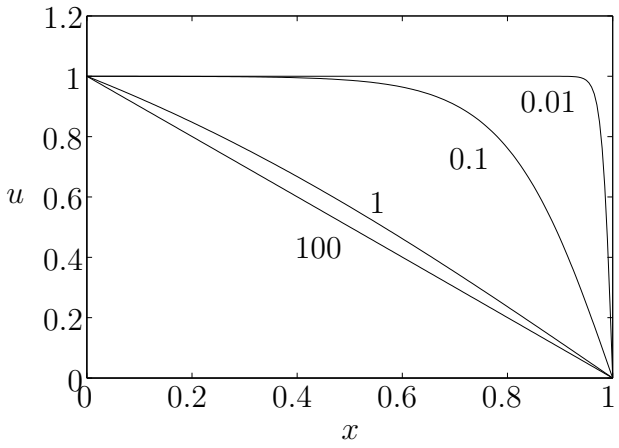


Abbildung F.1.: Exakte Lösungen der stationären Burgers-Gleichung zu den Randbedingungen $u(0) = 1$ und $u(1) = 0$. Die Werte für ν sind als Parameter angegeben..

mit $\boldsymbol{u} = (u, v)^{\mathrm{T}}$ erweitern. Dann führt die Transformation

$$u = -\frac{2\nu}{\Phi}\frac{\partial \Phi}{\partial x}, \qquad v = -\frac{2\nu}{\Phi}\frac{\partial \Phi}{\partial y} \tag{F.17}$$

und führt auf die 2D-Diffusionsgleichung

$$\frac{\partial \Phi}{\partial t} - \nu\left(\frac{\partial^2 \Phi}{\partial x^2} + \frac{\partial^2 \Phi}{\partial y^2}\right) = 0. \tag{F.18}$$

Man kann jetzt durch Kenntnis exakter Lösungen der 2D-Diffusionsgleichung wieder exakte Lösungen der Burgers-Gleichung finden. Hier beschränken wir uns auf den stationären Fall

$$\frac{\partial^2 \Phi}{\partial x^2} + \frac{\partial^2 \Phi}{\partial y^2} = 0. \tag{F.19}$$

Durch den Separationsansatz $\Phi = f(x)g(y)$ kann man leicht die allgemeine Lösung

$$\Phi = a_1 + a_2 x + a_3 y + a_4 xy + a_5 \left[\mathrm{e}^{\lambda(x-x_0)} + \mathrm{e}^{\lambda(x-x_0)}\right] \cos(\lambda y) \tag{F.20}$$

finden. Die Lösung für $\boldsymbol{u}$ ergibt sich dann als

$$\begin{pmatrix} u \\ v \end{pmatrix} = -2\nu \begin{pmatrix} \dfrac{a_2 + a_4 y + \lambda a_5 \left[\mathrm{e}^{\lambda(x-x_0)} - \mathrm{e}^{\lambda(x-x_0)}\right] \cos(\lambda y)}{a_1 + a_2 x + a_3 y + a_4 xy + a_5 \left[\mathrm{e}^{\lambda(x-x_0)} + \mathrm{e}^{\lambda(x-x_0)}\right] \cos(\lambda y)} \\ \dfrac{a_3 + a_4 x - \lambda a_5 \left[\mathrm{e}^{\lambda(x-x_0)} + \mathrm{e}^{\lambda(x-x_0)}\right] \sin(\lambda y)}{a_1 + a_2 x + a_3 y + a_4 xy + a_5 \left[\mathrm{e}^{\lambda(x-x_0)} + \mathrm{e}^{\lambda(x-x_0)}\right] \cos(\lambda y)} \end{pmatrix}. \tag{F.21a}$$

Abbildung F.2 zeigt das Beispiel, das auch in Fletcher (1991*a*) abgedruckt ist. Es

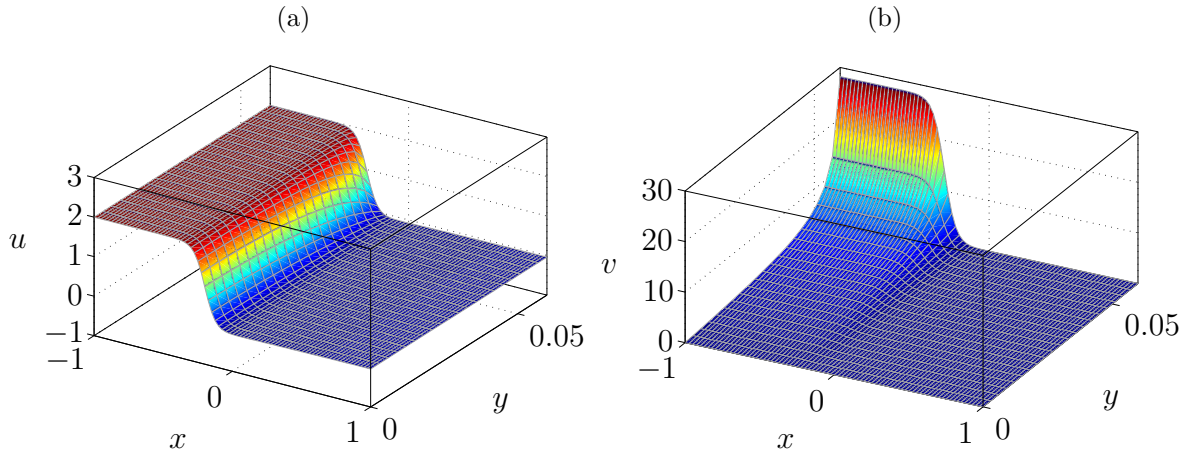


Abbildung F.2.: Lösungen der zwei-dimensionalen Burgers-Gleichung nach (F.21). Die Parameter sind identisch mit denjenigen, die Fletcher (1991*a*) verwendet: $a_1 = a_2 = 1.3 \times 10^{13}$, $a_3 = a_4 = 0$, $a_5 = 1$, $\lambda = 25$, $x_0 = 1$ und $\nu = 0.04$.

wurden dieselben Parameter verwendet.[2] Für größere Werte von y als die in der Abbildung gezeigten erhält man Singularitäten und eine dramatische Variation von u und v die sich in Abb. F.2b schon andeutet. Dies hängt mit den Nullstellen des Nenners in (F.21) zusammen. Beachte auch, daß es für diese Lösungen keine Kontinuitätsgleichung gibt. Insbesondere ist (F.21) nicht divergenzfrei.

[2]Trotzdem besteht bei der Skala von u eine Diskrepanz (Faktor 2).

Literaturverzeichnis

Abramowitz, M. and Stegun, I. A. (1972), *Handbook of Mathematical Functions*, Dover.

Albensoeder, S. and Kuhlmann, H. C. (2002), 'Linear stability of rectangular cavity flows driven by anti-parallel motion of two facing walls', *J. Fluid Mech.* **458**, 153–180.

Albensoeder, S., Kuhlmann, H. C. and Rath, H. J. (2001), 'Three-dimensional centrifugal-flow instabilities in the lid-driven cavity problem', *Phys. Fluids* **13**, 121–135.

Canuto, C., Hussaini, M. Y., Quarteroni, A. and Zhang, T. A. (1988), *Spectral Methods in Fluid Dynamics*, Springer.

Chapman, C. J. (2000), *High speed flow*, Cambridge University Press, Cambridge.

Drazin, P. G. (1983), *Solitons*, Vol. 85 of *London Mathematical Society Lecture Notes*, Cambridge University Press, Cambridge.

Du Fort, E. C. and Frankel, S. P. (1953), 'Stability conditions in the numerical treatment of parabolic differential equations', *J. Math. Tables and other Aids to Comput. (former title of Math. Comput.)* **7**, 135–152.

Ferziger, J. H. and Perić, M. (2002), *Computational Methods for Fluid Dynamics*, Springer, Berlin.

Finlayson, B. A. (1972), *The Method of Weighted Residuals and Variational Principles*, Academic Press.

Fletcher, C. A. J. (1991*a*), *Computational Techniques for Fluid Dynamics*, Vol. I of *Springer Series in Computational Physics*, Springer.

Fletcher, C. A. J. (1991*b*), *Computational Techniques for Fluid Dynamics*, Vol. II of *Springer Series in Computational Physics*, Springer.

Garabedian, P. R. (1964), *Partial differential equations*, Wiley, New York.

Golub, H. G. and van Loan, H. G. (1989), *Matrix Computations*, Johns Hopkins University Press.

Gresho, P. M. (1991), 'Incompressible fluid dynamics: some fundamental formulation issues', *Annu. Rev. Fluid Mech.* **23**, 413–453.

Hackbusch, W. (1985), *Multi-grid methods and applications*, Springer, Berlin.

Kevorkian, J. and Cole, J. D. (1981), *Perturbation methods in applied mathematics*, Vol. 34 of *Applied Mathematical Sciences*, Springer, Heidelberg.

Kuhlmann, H. (2007), *Strömungsmechanik*, Pearson Studium, München.

Landau, L. D. and Lifschitz, E. M. (1991), *Hydrodynamik*, Vol. VI of *Lehrbuch der Theoretischen Physik*, Akademie Verlag.

Leister, H.-J. and Perić, M. (1994), 'Vectorized strongly implicit solving procedure for a seven-diagonal coefficient matrix', *Int. J. Num. Meth. Heat Fluid Flow* **4**, 159–172.

Marner, F., Scholle, M., Herrmann, D. and Gaskell, P. H. (2019), 'Competing Lagrangians for incompressible and compressible viscous flow', *R. Soc. Open Sci.* **6**, 181595 (14pp).

Noschese, S., Pasquini, L. and Reichel, L. (2013), 'Tridiagonal toeplitz matrices: properties and novel applications', *Numer. Linear Algebra Appl.* **20**(2), 302–326.

Patankar, S. V. (1980), *Numerical Heat Transfer and Fluid Flow*, Hemisphere.

Peyret, R. (2002), *Spectral methods for incompressible viscous flow*, Vol. 148 of *Applied Mathematical Sciences*, Springer, New York, Berlin.

Peyret, R. and Taylor, T. D. (1983), *Computational Methods for Fluid Flow*, Springer Series in Computational Physics, Springer, Berlin, Heidelberg.

Richtmyer, R. D. and Morton, K. W. (1967), *Difference Methods for Initial Value Problems*, Wiley, New York.

Saad, Y. (2003), *Iterative methods for sparse linear systems*, Society for Industrial and Applied Mathematics, Philadelphia.

Saffman, P. G. (1992), *Vortex Dynamics*, Cambridge University Press.

Sommerfeld, A. (1978), *Partielle Differentialgleichungen in der Physik*, Harri Deutsch, Thun, Frankfurt/M.

Stone, H. L. (1968), 'Iterative solution of implicit approximations of multidimensional partial differential equations', *SIAM J. Num. Anal.* **5**, 530–558.

Trefethen, L. N. and Bau, III, D. (1997), *Numerical linear algebra*, SIAM, Philadelphia.

Whitham, G. B. (1974), *Linear and nonlinear waves*, Wiley, New York.

Index