Numerische Methoden für Partielle Differentialgleichungen

Roland Pulch

Institut für Mathematik und Informatik
Universität Greifswald

Inhalt:

1. Beispiele und Klassifikation
2. Elliptische Dgln. (zweiter Ordnung)
3. Parabolische Dgln. (zweiter Ordnung)
4. Hyperbolische Dgln. (zweiter Ordnung)

Literatur:

Ch. Großmann, H.-G. Roos: Numerische Behandlung partieller Differentialgleichungen. (3. Aufl.) Teubner, Wiesbaden 2005. (Kapitel 1-4)

H.R. Schwarz: Numerische Mathematik. (4. Aufl.) Teubner, Stuttgart 1997. (Kapitel 10)

D. Braess: Finite Elemente. (2. Aufl.) Springer, Berlin 2007.

A. Quarteroni, R. Sacco, F. Saleri: Numerische Mathematik 2. Springer, Berlin 2002. (Kapitel 12-13)

Inhaltsverzeichnis

1 Beispiele und Klassifikation **4**

1.1 Beispiele 5

1.2 Klassifikation 11

2 Elliptische Differentialgleichungen **16**

2.1 Maximumprinzip 16

2.2 Finite-Differenzen-Methoden 20

2.3 Sobolev-Räume und Variationsform 36

2.4 Finite-Elemente-Methoden 50

3 Parabolische Differentialgleichungen **69**

3.1 Anfangs-Randwert-Probleme 69

3.2 Finite-Differenzen-Methoden 75

3.3 Stabilitätsanalyse 81

3.4 Semidiskretisierung 87

4 Hyperbolische Differentialgleichungen **94**

4.1 Wellengleichung 94

4.2 Finite-Differenzen-Methoden 98

4.3 Charakteristikenverfahren 105

Kapitel 1

Beispiele und Klassifikation

Diese Vorlesung behandelt die numerische Lösung von *partiellen Differentialgleichungen* (PDGen). Die exakte Lösung hängt von mehreren unabhängigen Veränderlichen ab, die oft die Zeit und Ortskoordinaten darstellen. Verschiedene Typen partieller Dgln. existieren bereits im linearen Fall. Jede Klasse besitzt gewisse Eigenschaften und benötigt daher entsprechende numerische Verfahren. Anfangsbedingungen und/oder Randbedingungen treten dabei auf.

Im Gegensatz dazu können Systeme aus *gewöhnlichen Differentialgleichungen* (GDGen) in der allgemeinen Form

$$y'(x) = f(x, y(x)) \qquad (y : \mathbb{R} \to \mathbb{R}^n,\ f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n)$$

dargestellt werden. Die unabhängige Veränderliche stellt häufig die Zeit dar. Daher sind Anfangswertprobleme $y(x_0) = y_0$ die überwiegende Aufgabenstellung. Eine analytische Lösung ist im allgemeinen nicht möglich. Daher werden numerische Verfahren zur Erzeugung einer Näherungslösung benötigt. Eine konvergente numerische Methode kann prinzipiell alle Systeme aus GDGen (mit gewissen Voraussetzungen) lösen.

1.1 Beispiele

Wir stellen drei wichtige Beispiele vor, welche die drei Typen von Modellen aus partiellen Dgln. demonstrieren.

Poisson-Gleichung

Wir betrachten ein beschränktes Gebiet $\Omega \subset \mathbb{R}^2$, beispielsweise das Einheitsquadrat $\Omega = (0,1) \times (0,1)$. Für $u \in C^2(\Omega)$ ist der Laplace-Operator (in zwei Ortskoordinaten x, y) definiert durch

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}. \tag{1.1}$$

Die *Poisson-Gleichung* lautet dann

$$-\Delta u = f$$

mit vorgegebener Funktion $f : \Omega \to \mathbb{R}$. Der Spezialfall $f \equiv 0$ stellt die *Laplace-Gleichung* dar. Die Lösung der Poisson-Gleichung (1.1) ist stationär, d.h. sie ändert sich nicht mit der Zeit.

Hier befassen wir uns mit Randwertaufgaben. Sei $\partial\Omega$ der Rand von Ω. Randbedingungen des Dirichlet-Typs lauten

$$u(x,y) = g(x,y) \quad \text{für } (x,y) \in \partial\Omega$$

mit vorgegebener Funktion $g : \partial\Omega \to \mathbb{R}$. Randbedingungen des Neumann-Typs beziehen sich auf die Ableitung der Lösung senkrecht auf dem Rand, d.h.

$$\frac{\partial u}{\partial \nu}(x,y) = \langle \nu(x,y), \nabla u(x,y) \rangle = h(x,y) \quad \text{für } (x,y) \in \partial\Omega$$

mit Normalenvektor ν ($\|\nu\|_2 = 1$) und vorgegebener Funktion $h : \partial\Omega \to \mathbb{R}$. Ebenfalls können gemischte Randbedingungen

$$u(x,y) = g(x,y) \text{ für } (x,y) \in \Gamma_D, \quad \frac{\partial u}{\partial \nu}(x,y) = h(x,y) \text{ für } (x,y) \in \Gamma_N$$

auftreten mit $\Gamma_D \cup \Gamma_N = \partial\Omega$ und $\Gamma_D \cap \Gamma_N = \emptyset$.

Wir leiten die Poisson-Gleichung für das elektrische Feld im Fall von drei Raumdimensionen ($x = (x_1, x_2, x_3)$) her. Sei $E : \mathbb{R}^3 \to \mathbb{R}^3$ das elektrische Feld und $\Phi : \mathbb{R}^3 \to \mathbb{R}$ das zugehörige Potential. Es gilt

$$E(x) = -\nabla\Phi(x)$$

mit dem Gradienten $\nabla\Phi = (\frac{\partial\Phi}{\partial x_1}, \frac{\partial\Phi}{\partial x_2}, \frac{\partial\Phi}{\partial x_3})$. Es folgt

$$\mathrm{div}E(x) = -\Delta\Phi(x).$$

Die Ladungsverteilung wird beschrieben durch $\rho : \mathbb{R}^3 \to \mathbb{R}$. Sei $\varepsilon > 0$ die Permittivität. Die erste Maxwell-Gleichung (Gauß'sches Gesetz) lautet

$$\mathrm{div}E(x) = \frac{\rho(x)}{\varepsilon}.$$

Beide Gleichungen zusammen liefern die Poisson-Gleichung

$$-\Delta\Phi(x) = \frac{\rho(x)}{\varepsilon},$$

wobei ρ vorgegeben und Φ unbekannt ist.

Eine Verbindung zur komplexen Analysis (Funktionentheorie) besteht für holomorphe Funktionen. Sei $g : \mathbb{C} \to \mathbb{C}$ holomorph und $\Omega \subset \mathbb{C}$ offen, zusammenhängend und beschränkt. Einerseits liefert die Cauchy'sche Integralformel

$$g(z) = \frac{1}{2\pi\mathrm{i}} \oint_{\partial\Omega} \frac{g(\zeta)}{\zeta - z}\, \mathrm{d}\zeta \qquad \text{für } z \in \Omega.$$

Folglich ist g in Ω bereits eindeutig durch die Funktionswerte auf dem Rand $\partial\Omega$ bestimmt. Andererseits impliziert die Formel der komplexen Differentiation (Cauchy-Riemann-Dgln.)

$$\Delta(\mathrm{Re}\ g) = 0 \qquad \text{und} \qquad \Delta(\mathrm{Im}\ g) = 0.$$

Somit sind Real- und Imaginärteil jeweils Lösung einer Laplace-Gleichung in zwei Veränderlichen. Die Werte von g auf $\partial\Omega$ werden durch Dirichlet-Randbedingungen spezifiziert. Es folgt jeweils die Eindeutigkeit des Real- und Imaginärteils in Ω. Daher stimmen diese beiden theoretischen Konzepte überein.

Wellengleichung

In einer Raumdimension lautet die *Wellengleichung*

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}, \tag{1.2}$$

wobei die Konstante $c > 0$ die Wellengeschwindigkeit darstellt. Die Lösung u hängt sowohl vom Ort als auch von der Zeit ab. Wir lösen die Wellengleichung mit der Methode nach d'Alembert. Neue Variablen werden eingeführt durch

$$\xi = x - ct, \qquad \eta = x + ct.$$

Es folgt

$$\begin{aligned}
\frac{\partial}{\partial x} &= \frac{\partial}{\partial \xi}\frac{\partial \xi}{\partial x} + \frac{\partial}{\partial \eta}\frac{\partial \eta}{\partial x} = \frac{\partial}{\partial \xi} + \frac{\partial}{\partial \eta}, \\
\frac{\partial^2}{\partial x^2} &= \frac{\partial^2}{\partial \xi^2} + 2\frac{\partial^2}{\partial \xi \partial \eta} + \frac{\partial^2}{\partial \eta^2}, \\
\frac{\partial}{\partial t} &= \frac{\partial}{\partial \xi}\frac{\partial \xi}{\partial t} + \frac{\partial}{\partial \eta}\frac{\partial \eta}{\partial t} = c\left(-\frac{\partial}{\partial \xi} + \frac{\partial}{\partial \eta}\right), \\
\frac{\partial^2}{\partial t^2} &= c^2\left(\frac{\partial^2}{\partial \xi^2} - 2\frac{\partial^2}{\partial \xi \partial \eta} + \frac{\partial^2}{\partial \eta^2}\right).
\end{aligned}$$

Wir erhalten die transformierte Dgl.

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

und daher

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

Es ist leicht nachzuprüfen, dass die allgemeine Lösung gegeben ist durch

$$u(\xi, \eta) = \Phi(\xi) + \Psi(\eta)$$

mit Hilfsfunktionen $\Phi, \Psi \in C^2(\mathbb{R})$. Ein Spezielfall ist $\Phi \equiv \Psi$ (nur sofern $\frac{\partial u}{\partial t}(x, 0) \equiv 0$). Wir erhalten

$$u(x, t) = \Phi(x - ct) + \Psi(x + ct).$$

Die Funktionen Φ, Ψ ergeben sich aus Anfangsbedingungen (für $u(x, 0)$ und $\frac{\partial u}{\partial t}(x, 0)$). Zur Interpretation schreiben wir

$$\Phi(x - ct) = \Phi(x + c\Delta t - c(t + \Delta t)) = \Phi(x^* - ct^*)$$

mit $x^* = x + c\Delta t$ and $t^* = t + \Delta t$. In der Zeitdauer Δt wird die Information vom Punkt x zum Punkt x^* mit $\Delta x = c\Delta t$ transportiert. Daher stellt der Term $\Phi(x - ct)$ eine Welle dar, die sich mit Geschwindigkeit c fortbewegt. Entsprechend repräsentiert der Term $\Psi(x + ct)$ eine Welle mit der Geschwindigkeit $-c$. Die Lösung u ist die Überlagerung dieser beiden Wellen.

Mit Anfangswerten $u(x, 0) = u_0(x)$, $\frac{\partial u}{\partial t}(x, 0) = cu_1(x)$ bei $t = 0$ ergibt sich für die exakte Lösung die Formel

$$u(x,t) = \frac{1}{2}\left(u_0(x+ct) + u_0(x-ct) + \int_{x-ct}^{x+ct} u_1(s)\,\mathrm{d}s\right). \tag{1.3}$$

Es folgt, dass die Lösung u an einem Punkt (x^*, t^*) mit $t^* > 0$ nur von Anfangswerten bei $t = 0$ im Interval $x \in [x^* - ct^*, x^* + ct^*]$ abhängt. Deshalb beschreibt die Wellengleichung einen Informationstransport mit endlicher Geschwindigkeit.

Die lineare Dgl. (1.2) zweiter Ordnung kann in ein lineares System aus Dgln. erster Ordnung transformiert werden. Wir definieren $v_1 = \frac{\partial u}{\partial t}$ und $v_2 = \frac{\partial u}{\partial x}$. Sei $u \in C^2$. Somit liefert der Satz von Schwarz

$$\frac{\partial v_1}{\partial x} = \frac{\partial^2 u}{\partial x \partial t} = \frac{\partial v_2}{\partial t}.$$

Die Dgl. (1.2) zeigt

$$\frac{\partial v_1}{\partial t} = c^2 \frac{\partial v_2}{\partial x}.$$

Es folgt das System

$$\frac{\partial}{\partial t}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} + \begin{pmatrix} 0 & -c^2 \\ -1 & 0 \end{pmatrix} \frac{\partial}{\partial x}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{1.4}$$

Die entstehende Matrix besitzt die Eigenwerte $+c$ und $-c$, d.h. die Wellengeschwindigkeiten. Um die Lösung u der Dgl. (1.2) zu erhalten ist noch eine Integration mit v_1, v_2 durchzuführen.

Wärmeleitungsgleichung

In einer Raumdimension lautet die *Wärmeleitungsgleichung*

$$\frac{\partial u}{\partial t} = \lambda \frac{\partial^2 u}{\partial x^2} \tag{1.5}$$

mit der Konstante $\lambda > 0$. Um Lösungen dieser Dgl. zu diskutieren setzen wir o.E.d.A. $\lambda = 1$ und betrachten ein endliches Intervall $x \in [0, \pi]$. Wir verwenden homogene Dirichlet-Randbedingungen

$$u(0,t) = 0, \quad u(\pi,t) = 0 \qquad \text{für alle } \ t \geq 0. \tag{1.6}$$

Die Funktionen

$$v_k(x,t) = \mathrm{e}^{-k^2 t} \sin(kx) \qquad \text{für } \ k \in \mathbb{N} \tag{1.7}$$

sind jeweils Lösungen der Wärmeleitungsgleichung (1.5) und erfüllen die Randbedingungen (1.6).

Gegeben seien nun Anfangsbedingungen $u(x,0) = u_0(x)$ für $x \in [0,\pi]$ mit $u_0(0) = u_0(\pi) = 0$. Dadurch kann u_0 zu einer stetigen ungeraden Funktion auf $[-\pi,\pi]$ fortgesetzt werden. Diese wiederum zu einer stetigen 2π-periodischen Funktion auf ganz $\mathbb{R}$. Wir erhalten die Entwicklung in eine Fourierreihe

$$u_0(x) = \sum_{k=1}^{\infty} a_k \sin(kx)$$

mit Koeffizienten $a_k \in \mathbb{R}$.

Da die Wärmeleitungsgleichung (1.5) linear ist, erhalten wir die Lösung aus einer Superposition der Funktionen (1.7)

$$u(x,t) = \sum_{k=1}^{\infty} a_k \mathrm{e}^{-k^2 t} \sin(kx)$$

für $t \geq 0$.

Alternativ können Randbedingungen des Neumann-Typs vorgegeben werden. Homogene Neumann-Randbedingungen

$$\frac{\partial u}{\partial x}(0,t) = 0, \quad \frac{\partial u}{\partial x}(\pi,t) = 0 \qquad \text{für alle } \ t \geq 0$$

bedeuten, dass kein Wärmefluss durch den Rand stattfindet.

Nun seien Anfangsbedingunen $u(x,0) = u_0(x)$ im gesamten Ortsbereich $x \in \mathbb{R}$ vorgegeben. Es ergibt sich eine Formel für die exakte Lösung der Wärmeleitungsgleichung (mit $\lambda = 1$)

$$u(x,t) = \frac{1}{2\sqrt{\pi t}} \int_{-\infty}^{+\infty} \mathrm{e}^{-\xi^2/4t} u_0(x-\xi)\, \mathrm{d}\xi, \tag{1.8}$$

wobei die Existenz des Integrals vorausgesetzt wird.

Wir stellen fest, dass die Lösung in einem Punkt (x,t) nun von allen Anfangswerten $u_0(\xi)$ mit $\xi \in \mathbb{R}$ abhängt. Daher findet der Informationstransport mit unendlicher Geschwindigkeit statt. Jedoch wird die Größenordnung der Information exponentiell gedämpft bei ansteigenden Entfernungen.

Wir leiten die Wärmeleitungsgleichung in drei Raumdimensionen her. Sei $T : \mathbb{R}^3 \to \mathbb{R}$ die Temperatur, $F : \mathbb{R}^3 \to \mathbb{R}^3$ der Wärmefluss und $\kappa > 0$ die Diffusionskonstante. Es gilt allgemein

$$F = -\kappa \nabla T.$$

Wir erhalten für die Energie $E : \mathbb{R}^3 \to \mathbb{R}$

$$\frac{\partial E}{\partial t} = -\mathrm{div}\, F = \kappa\, \mathrm{div} \nabla T = \kappa\, \Delta T.$$

Sei $\alpha = \frac{\partial E}{\partial T}$ ein konstanter Materialparameter. Es gilt $\frac{\partial E}{\partial t} = \frac{\partial E}{\partial T}\frac{\partial T}{\partial t}$. Somit erhalten wir die Wärmeleitungsgleichung

$$\frac{\partial T}{\partial t} = \frac{\kappa}{\alpha}\, \Delta T$$

mit der Konstanten $\lambda = \frac{\kappa}{\alpha}$.

Black-Scholes-Gleichung

Die bisherigen Beispiele sind aus der Physik und technischen Anwendungen motiviert. Wir stellen kurz ein Beispiel aus der Finanzmathematik vor. Sei S ein Aktienpreis und V der faire Preis einer Europäischen Call-Option auf diese Aktie. Es folgt eine partielle Dgl. mit Lösung V, wobei die unabhängigen Veränderlichen die Zeit t und die (möglichen) Aktienkurse $S \geq 0$ sind. Die berühmte Black-Scholes-Gleichung lautet

$$\frac{\partial V}{\partial t} + \tfrac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS \frac{\partial V}{\partial S} - r\, V = 0 \tag{1.9}$$

mit Konstanten $r, \sigma > 0$. Obwohl die Black-Scholes-Gleichung (1.9) kompliziert aussieht, kann sie in eine Wärmeleitungsgleichung (1.5) transformiert werden. Daher stimmen die Eigenschaften der Black-Scholes-Gleichung mit denen der Wärmeleitungsgleichung überein.

Bemerkung: Die Wellengleichung (1.2), die Wärmeleitungsgleichung (1.5) sowie die Black-Scholes-Gleichung (1.9) sind relativ einfache Dgln., wodurch Formeln für die entsprechenden Lösungen vorliegen, siehe (1.3) und (1.8). Daher sind numerische Verfahren nicht zwingend erforderlich. Jedoch ist eine analytische Lösung oft nicht mehr möglich, wenn ein Quellterm auftritt, d.h.

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} + f(x,t,u) \qquad \text{oder} \qquad \frac{\partial u}{\partial t} = \lambda \frac{\partial^2 u}{\partial x^2} + f(x,t,u).$$

Nun benötigen wir numerische Verfahren zur Erzeugung einer Näherungslösung. Jedoch ändern sich die Eigenschaften einer Dgl. nicht durch das Hinzufügen eines Quellterms.

1.2 Klassifikation

Wir betrachten eine lineare partielle Dgl. zweiter Ordnung

$$\sum_{i,j=1}^{n} a_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} = f\Big(x_1, \ldots, x_n, u, \frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_n}\Big) \tag{1.10}$$

mit $n \geq 2$ unabhängigen Veränderlichen. Die Lösung $u : \Omega \to \mathbb{R}$ erfülle $u \in C^2(\Omega)$ mit einer offenen Menge $\Omega \subseteq \mathbb{R}^n$. Wir verwenden die Abkürzungen $x = (x_1, \ldots, x_n)$ und $\nabla u = (\frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_n})$. Die Typen partieller Dgln. bezüglich des Grads der Linearität lauten:

- *lineare Dgl.*: Die Koeffizienten a_{ij} sind Konstanten oder hängen nur von x ab und die rechte Seite ist linear in u und ∇u ($f = b(x) + c(x)u + d_1(x)\frac{\partial u}{\partial x_1} + \cdots + d_n(x)\frac{\partial u}{\partial x_n}$).
- *semi-lineare Dgl.*: Die Koeffizienten a_{ij} sind Konstanten oder hängen nur von x ab und die rechte Seite ist nichtlinear.

- *quasi-lineare Dgl.*: Die Koeffizienten a_{ij} hängen (auch) von u und/oder ∇u ab. (Die rechte Seite f kann linear oder nichtlinear sein.)

Die Definition eines korrekt gestellten Problems lautet wie folgt.

Definition 1.1 *Eine Dgl. oder ein System aus Dgln. mit zugehörigen Anfangs- und/oder Randbedingungen heißt korrekt gestellt genau dann, wenn eine eindeutige Lösung existiert und die Lösung stetig von den Eingabedaten abhängt. Anderenfalls ist das Problem schlecht gestellt.*

Die Koeffizienten a_{ij} bilden eine Matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$. O.E.d.A. setzen wir die Matrix A symmetrisch voraus, da $u \in C^2$ angenommen wird. Daher ist die Matrix A diagonalisierbar und alle Eigenwerte (EWe) $\lambda_1, \ldots, \lambda_n$ sind reelle Zahlen. Die Klassifikation partieller Dgln. (1.10) beruht auf der Definitheit von A. (Die Klassifikation ist unabhängig von der rechten Seite f.) Im Falle zweier Raumdimensionen ($n = 2$) gilt $\det(A) = \lambda_1\lambda_2$, d.h. die Definitheit ergibt sich aus dem Vorzeichen der Determinante. Die quadratische Form

$$q(z) = z^\top Az \qquad (A \in \mathbb{R}^{2\times 2}, z \in \mathbb{R}^2)$$

kann für eine geometrische Interpretation betrachtet werden. Dies liefert die Namensgebung der Typen zu partiellen Dgln.

1. Fall: A (pos. oder neg.) definit $\rightarrow$ *elliptische Dgl.*
(alle EWe sind positiv oder alle EWe sind negativ)
Elliptische Dgln. modellieren stationäre Zustände. Randbedingungen liefern korrekt gestellte Probleme.

Für $n = 2$ bildet die Menge $\{z \in \mathbb{R}^2 : z^\top Az = \pm 1\}$ eine Ellipse.

Beispiele *Poisson-Gleichung*
In n Dimensionen lautet die Poisson-Gleichung

$$-\Delta u = -\frac{\partial^2 u}{\partial {x_1}^2} - \cdots - \frac{\partial^2 u}{\partial {x_n}^2} = f(x_1, \ldots, x_n). \tag{1.11}$$

Es folgt, dass die Matrix $A \in \mathbb{R}^{n\times n}$ diagonal ist und alle Diagonalelemente (Eigenwerte) sind -1. Somit ist die Dgl. (1.11) elliptisch.

2. Fall: A indefinit, $\det A \neq 0 \rightarrow$ *hyperbolische Dgl.*
(mindestens zwei EWe besizten unterschiedliches Vorzeichen, alle EWe sind ungleich null)
Hyperbolische Dgln. modellieren Transportprozesse. Anfangsbedingungen (möglicherweise zusammen mit Randbedingungen) liefern korrekt gestellte Probleme.

Für $n = 2$ bildet die Menge $\{z \in \mathbb{R}^2 : z^\top Az = 1\}$ eine Hyperbel.

Beispiel: *Wellengleichung*
In n Raumdimensionen lautet die Wellengleichung

$$\frac{\partial^2 u}{\partial t^2} - c^2 \Delta u = \frac{\partial^2 u}{\partial t^2} - c^2 \left(\frac{\partial^2 u}{\partial {x_1}^2} + \cdots + \frac{\partial^2 u}{\partial {x_n}^2} \right) = 0 \qquad (1.12)$$

mit Wellengeschwindigkeit $c > 0$. Wieder ist die Koeffizientenmatrix $A \in \mathbb{R}^{(n+1)\times(n+1)}$ diagonal. Es tritt ein einfacher EW $+1$ und ein n-facher EW $-c^2$ auf. Somit ist die Wellengleichung (1.12) eine hyperbolische Dgl.

Oft besitzt ein EW ein anderes Vorzeichen als alle anderen EWe (die Zeit verhält sich qualitativ anders als die Ortskoordinaten). Wenn mindestens zwei EWe positiv und mindestens zwei EWe negativ sind, dann spricht man von einer *ultrahyperbolischen Dgl.* ($n \geq 4$ notwendig). Jedoch werden wir in dieser Vorlesung keine ultrahyperbolischen Dgln. betrachten.

3. Fall: $\det A = 0 \rightarrow$ *parabolische Dgl.*
(mindestens ein EW ist null)
Parabolische Dgln. modellieren z.B. Diffusionsprozesse. Anfangsbedingungen zusammen mit Randbedingungen liefern korrekt gestellte Probleme.

Für $n = 2$ bildet die Menge $\{z \in \mathbb{R}^2 : z^\top Az = 1\}$ ein Geradenpaar. Falls lineare Terme der quadratischen Form hinzugefügt werden ($q(z) = z^\top Az + b^\top z$), dann kann eine Parabel entstehen.

Beispiel: *Wärmeleitungsgleichung*
In n Raumdimensionen lautet die Wärmeleitungsgleichung

$$\frac{\partial u}{\partial t} - \lambda \Delta u = \frac{\partial u}{\partial t} - \lambda \left(\frac{\partial^2 u}{\partial {x_1}^2} + \cdots + \frac{\partial^2 u}{\partial {x_n}^2} \right) = 0 \qquad (1.13)$$

mit der Konstanten $\lambda > 0$. Die Koeffizientenmatrix $A \in \mathbb{R}^{(n+1)\times(n+1)}$ ist diagonal. Ein einfacher EW null und ein n-facher EW $-\lambda$ tritt auf. Somit ist die Dgl. (1.13) parabolisch.

Die Klassifikation ist im Falle konstanter Koeffizienten a_{ij} eindeutig. Bei $a_{ij}(x)$ kann dieselbe Dgl. (1.10) unterschiedliche Typen in verschiedenen Teilmengen von Ω aufweisen. Bei $a_{ij}(u)$ kann der Typ der Dgl. sogar von der Lösung u abhängen. Jedoch tritt dies nur selten in der Praxis auf.

Skalierung

Multiplikation der Dgl. (1.10) mit einer Konstanten $\alpha \neq 0$ verändert die Koeffizientenmatrix A zu $\tilde{A} = \alpha A$. Die Unterschiede der Vorzeichen bei den Eigenwerten bleiben dabei bestehen. Daher ist der Typ der Dgl. invariant bezüglich einer Skalierung.

Basistransformationen

Nun untersuchen wir die Invarianz des Typs einer Dgl. (1.10) bezüglich einer Basistransformation im Definitionsbereich Ω. Wir betrachten konstante Koeffizienten a_{ij}, da eine Verallgemeinerung auf nichtkonstante Koeffizienten leicht möglich ist. Sei $y = Bx$ mit einer regulären Matrix $B = (b_{ij}) \in \mathbb{R}^{n\times n}$. In neuen Koordinaten y ist die Lösung $\tilde{u}(y) = \tilde{u}(Bx) = u(x)$. Die Kettenregel der mehrdimensionalen Differentiation liefert

$$\begin{aligned} \frac{\partial u(x)}{\partial x_i} = \frac{\partial \tilde{u}(Bx)}{\partial x_i} &= \sum_{k=1}^{n} \frac{\partial \tilde{u}(Bx)}{\partial y_k} b_{ki}, \\ \frac{\partial^2 u(x)}{\partial x_i \partial x_j} = \frac{\partial^2 \tilde{u}(Bx)}{\partial x_j \partial x_i} &= \sum_{k=1}^{n}\sum_{\ell=1}^{n} \frac{\partial^2 \tilde{u}(Bx)}{\partial y_\ell \partial y_k} b_{ki} b_{\ell j}. \end{aligned}$$

Es folgt

$$\sum_{i,j=1}^{n} a_{ij} \frac{\partial^2 u(x)}{\partial x_j \partial x_i} = \sum_{i,j=1}^{n} a_{ij} \sum_{k,\ell=1}^{n} \frac{\partial^2 \tilde{u}(Bx)}{\partial y_\ell \partial y_k} b_{ki} b_{\ell j} = \sum_{k,\ell=1}^{n} \left[\sum_{i,j=1}^{n} a_{ij} b_{ki} b_{\ell j} \right] \frac{\partial^2 \tilde{u}(y)}{\partial y_\ell \partial y_k}.$$

Seien $\tilde{A} = (\tilde{a}_{k\ell})$ die Koeffizienten in der neuen Basis. Es gilt

$$\tilde{A} = BAB^\top. \tag{1.14}$$

Somit ist die Matrix $\tilde{A}$ stets symmetrisch. Die Matrix B is regulär vorausgesetzt. Der Trägheitssatz von Sylvester impliziert, dass die Anzahl der positiven Eigenwerte sowie die Anzahl der negativen Eigenwerte jeweils in A und $\tilde{A}$ übereinstimmen. Es folgt, dass der Typ einer Dgl. invariant unter allen Basistransformationen ist.

Desweiteren ist eine symmetrische Matrix diagonalisierbar und es existiert eine Basis aus orthogonalen Eigenvektoren. Daher gibt es eine orthogonale Matrix S ($S^{-1} = S^\top$), so dass $D = SAS^\top$ eine Diagonalmatrix ist. Wir können daher die Dgl. jeweils in Diagonalform transformieren, wodurch die gemischten Ableitungen verschwinden.

Der Trägheitssatz von Sylvester wird dargestellt in z.B.:
G. Fischer: Lineare Algebra, (18. Aufl.) Springer Spektrum 2014. (S. 323/324)
J. Liesen, V. Mehrmann: Lineare Algebra, (2. Aufl.) Springer Spektrum 2015. (S. 304)

Kapitel 2

Elliptische Differentialgleichungen

In diesem Kapitel behandeln wir die numerische Lösung von Randwertproblemen zu elliptischen Differentialgleichungen. Dabei stellt die Poisson-Gleichung das Musterbeispiel dar. Es gibt zwei Klassen aus numerischen Verfahren für diese Aufgabenstellung: Finite-Differenzen-Methoden und Finite-Elemente-Methoden.

2.1 Maximumprinzip

Wir betrachten die Poisson-Gleichung

$$-\Delta u(x) = -\sum_{i=1}^{n} \frac{\partial^2 u}{\partial x_i^2}(x) = f(x) \tag{2.1}$$

mit $x = (x_1, \ldots, x_n)$ in $n \geq 2$ Raumdimensionen. Sei $\Omega \subset \mathbb{R}^n$ beschränktes Gebiet. Die Randbedingungen vom Dirichlet-Typ lauten

$$u(x) = g(x) \qquad \text{für } x \in \partial\Omega \tag{2.2}$$

mit einer vorgegebenen Funktion $g : \partial\Omega \to \mathbb{R}$. Wir setzen die Existenz einer Lösung $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$ voraus. (Entsprechende Existenzsätze benötigen gewisse Annahmen und sind schwierig zu beweisen.) Sowohl die Eindeutigkeit als auch die stetige Abhängigkeit von den Eingabedaten folgen aus dem Maximumprinzip.

Satz 2.1 (Maximumprinzip) *Sei $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$. Es gilt*

(i) *Maximumprinzip: Falls $-\Delta u \leq 0$ in Ω, dann besitzt u ein Maximum auf dem Rand $\partial\Omega$.*

(ii) *Minimumprinzip: Falls $-\Delta u \geq 0$ in Ω, dann besitzt u ein Minimum auf dem Rand $\partial\Omega$.*

(iii) *Vergleich: Falls $v \in C^2(\Omega) \cap C^0(\bar{\Omega})$ und $-\Delta u \leq -\Delta v$ in Ω sowie $u \leq v$ auf $\partial\Omega$, dann folgt $u \leq v$ in Ω.*

Beweis:

Wir zeigen zuerst die Eigenschaft (i). Sei $-\Delta u < 0$ in Ω. Wenn ein $\xi \in \Omega$ existiert mit

$$u(\xi) = \sup_{x \in \Omega} u(x) > \sup_{x \in \partial\Omega} u(x),$$

dann ist ξ auch ein lokales Maximum. Es folgt $\nabla u(\xi) = 0$ und die Hesse-Matrix $\nabla^2 u(\xi) = (u_{x_i,x_j}(\xi))$ ist negativ semi-definit. Insbesondere sind die Einträge auf der Diagonalen nicht positiv. Daher gilt

$$-(u_{x_1x_1}(\xi) + \cdots + u_{x_nx_n}(\xi)) \geq 0.$$

Dies ist ein Widerspruch zu $-\Delta u < 0$. Daher muss ein Maximum auf dem Rand $\partial\Omega$ liegen.

Nun sei $-\Delta u \leq 0$ und $\eta \in \Omega$ mit

$$u(\eta) = \sup_{x \in \Omega} u(x) > \sup_{x \in \partial\Omega} u(x).$$

Wir definieren die Funktionen $h(x) := (\eta_1 - x_1)^2 + \cdots + (\eta_n - x_n)^2$ und $w(x) := u(x) + \delta \cdot h(x)$ mit einer reellen Zahl $\delta > 0$. Da $h \in C^2(\Omega) \cap C^0(\bar{\Omega})$ gilt, besitzt die Funktion w ein Maximum in Ω für hinreichend kleines δ. Es folgt

$$-\Delta w(x) = -\Delta u(x) - \delta \Delta h(x) = -\Delta u - 2\delta n < 0.$$

Wieder tritt ein Widerspruch auf. Daher muss hier ein Maximum auf dem Rand $\partial\Omega$ liegen.

Die Eigenschaft (ii) folgt aus (i) durch das Maximumprinzip angewendet auf die Funktion $v := -u$.

Um die Eigenschaft (iii) zu zeigen setzen wir $w := v - u$. Es folgt

$$-\Delta w = -\Delta v + \Delta u \geq 0$$

durch die Voraussetzungen im Satz. Es gilt $w \geq 0$ auf dem Rand $\partial\Omega$. Das Minimumprinzip liefert $w(x) \geq 0$ für alle $x \in \Omega$. □

Mit dem Maximumprinzip erhalten wir folgende Abschätzung.

Satz 2.2 *Für $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$ gilt*

$$|u(x)| \leq \sup_{z \in \partial\Omega} |u(z)| + c \sup_{z \in \Omega} |\Delta u(z)|. \tag{2.3}$$

für jedes $x \in \Omega$ mit einer Konstanten $c \geq 0$.

Beweis:

Das beschränkte Gebiet Ω befindet sich in einer Kugel mit Radius R und Mittelpunkt bei $x = 0$. Wir definieren

$$w(x) := R^2 - \sum_{i=1}^{n} x_i^2.$$

Es folgt $w_{x_i,x_j} = -2\delta_{ij}$. Desweiteren gilt $-\Delta w = 2n$ und $0 \leq w \leq R^2$ in Ω. Nun setzen wir

$$v(x) := \sup_{z \in \partial\Omega} |u(z)| + w(x) \cdot \tfrac{1}{2n} \sup_{z \in \Omega} |\Delta u(z)| \geq 0.$$

Mit dieser Konstruktion ergibt sich $-\Delta v \geq |\Delta u|$ in Ω und $v \geq |u|$ auf $\partial\Omega$. Satz 2.1 (iii) liefert $-v(x) \leq u(x) \leq +v(x)$ in Ω, d.h. $|u(x)| \leq v(x)$ in Ω. Da $w \leq R^2$ gilt, folgt (2.3) mit $c := \frac{R^2}{2n}$. □

Seien u_1 und u_2 zwei Lösungen eines Randwertproblems (2.1), (2.2), d.h. es gilt $-\Delta u_1 = f_1$, $-\Delta u_2 = f_2$ in Ω und $u_1 = g_1$, $u_2 = g_2$ auf $\partial\Omega$. Einsetzen der Differenz $u_1 - u_2$ in (2.3) liefert

$$|u_1(x) - u_2(x)| \leq \sup_{z \in \partial\Omega} |g_1(z) - g_2(z)| + c \sup_{z \in \Omega} |f_1(z) - f_2(z)| \tag{2.4}$$

für alle $x \in \Omega$. Daher hängen die Lösungen Lipschitz-stetig von den Eingabedaten ab. Zudem ist die Lösung eines Randwertproblems vom Dirichlet-Typ (2.1), (2.2) eindeutig (setze $f_1 \equiv f_2$, $g_1 \equiv g_2$).

Randwertprobleme des Dirichlet-Typs sind korrekt gestellt, siehe Definition 1.1, da (2.4) gilt (nur die Existenz wurde nicht gezeigt sondern vorausgesetzt). Wir diskutieren ein Beispiel um zu zeigen, dass Anfangswertprobleme nicht korrekt gestellt sind. Wir betrachten die Laplace-Gleichung $\Delta u = 0$ im Gebiet $\Omega = \{(x, y) \in \mathbb{R}^2 : y \geq 0\}$. Seien Anfangswerte vorgegeben bei $y = 0$ durch

$$u(x, 0) = \tfrac{1}{n} \sin(nx), \quad \tfrac{\partial u}{\partial y}(x, 0) = 0.$$

Es ergibt sich eine eindeutige Lösung

$$u(x, y) = \tfrac{1}{n} \cosh(ny) \sin(nx),$$

welche für $y \to \infty$ wie e^{ny} anwächst. Es gilt $|u(x, 0)| \leq \frac{1}{n}$, während u immer größer wird bei $y = 1$ für $n \to \infty$. Im Grenzfall $u(x, 0) = 0$ ist die Lösung $u \equiv 0$. Somit hängen die Lösungen nicht stetig von den Anfangswerten ab.

Nun betrachten wir einen allgemeinen Differentialoperator des elliptischen Typs.

Definition 2.3 *Der lineare Differentialoperator* $L : C^2(\Omega) \to C^0(\Omega)$

$$L = -\sum_{i,j=1}^{n} a_{ij}(x) \frac{\partial^2}{\partial x_i \partial x_j} \tag{2.5}$$

heißt elliptisch in Ω*, wenn die Matrix* $A = (a_{ij})$ *positiv definit ist für jedes* $x \in \Omega$*, d.h. es gilt* $\xi^\top A(x)\xi > 0$ *für alle* $\xi \in \mathbb{R}^n \backslash \{0\}$*. Der Operator (2.5) heißt gleichmäßig elliptisch in* $\Omega \subset \mathbb{R}^n$*, wenn eine Konstante* $\alpha > 0$ *existiert mit*

$$\xi^\top A(x)\xi \geq \alpha \|\xi\|_2^2 \quad \textit{für alle } \xi \in \mathbb{R}^n \textit{ und alle } x \in \Omega. \tag{2.6}$$

Das Maximumprinzip aus Satz 2.1 gilt ebenso für L anstelle von $-\Delta$ im Falle eines allgemeinen elliptischen Operators (2.5). Die Abschätzung aus Satz 2.2 bleibt bestehen mit L anstelle von $-\Delta$ im Falle eines gleichmäßig elliptischen Operators (2.5).

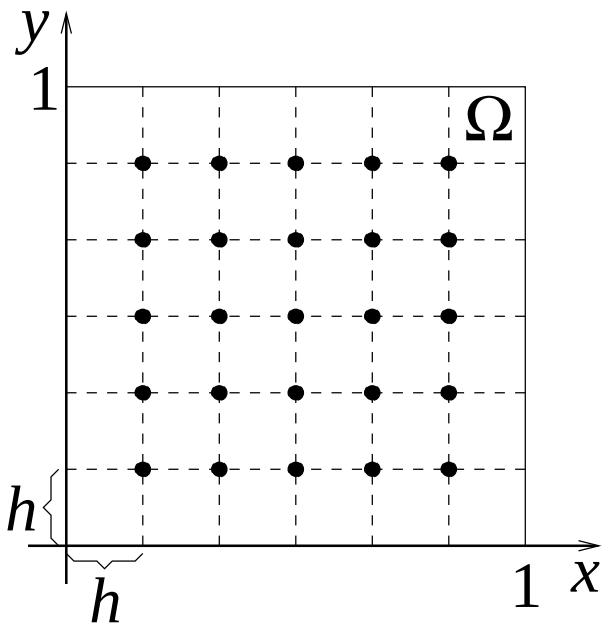


Abbildung 1: Gitter in Finiter-Differenzen-Methode.

2.2 Finite-Differenzen-Methoden

Wir führen die Klasse der Finiten-Differenzen-Verfahren für Randwertprobleme zu elliptischen Dgln. in zwei Raumdimensionen ein.

Laplace-Operator auf dem Quadrat

Als Musterbeispiel betrachten wir ein Randwertproblem vom Dirichlet-Typ auf dem Einheitsquadrat $\Omega := \{(x, y) : 0 < x, y < 1\}$ für die Poisson-Gleichung

$$\begin{aligned} -(\Delta u)(x,y) = -\frac{\partial^2 u}{\partial x^2}(x,y) - \frac{\partial^2 u}{\partial y^2}(x,y) &= f(x,y), \quad (x,y) \in \Omega \\ u(x,y) &= 0, \qquad (x,y) \in \partial\Omega. \end{aligned} \tag{2.7}$$

Wir führen ein uniformes Gitter im Definitionsbereich Ω ein mit Schrittweite $h = \frac{1}{M+1}$ für ein $M \in \mathbb{N}$

$$\Omega_h := \{(x_i, y_j) = (ih, jh) : i, j = 1, \ldots, M\}, \tag{2.8}$$

siehe Abbildung 1. Wir leiten eine Differenzenformel durch Taylor-Entwicklung her. Für $u \in C^4(\Omega)$ gilt

$$\begin{aligned} u(x+h,y) &= u(x,y) + hu_x(x,y) + h^2\tfrac{1}{2}u_{xx}(x,y) + h^3\tfrac{1}{6}u_{xxx}(x,y) \\ &\quad + h^4\tfrac{1}{24}u_{xxxx}(x+\vartheta_1 h, y) \\ u(x-h,y) &= u(x,y) - hu_x(x,y) + h^2\tfrac{1}{2}u_{xx}(x,y) - h^3\tfrac{1}{6}u_{xxx}(x,y) \\ &\quad + h^4\tfrac{1}{24}u_{xxxx}(x-\vartheta_2 h, y) \end{aligned}$$

mit $0 < \vartheta_1, \vartheta_2 < 1$. Es folgt

$$u(x+h,y) + u(x-h,y) = 2u(x,y) + h^2 u_{xx}(x,y) + h^4\tfrac{1}{12}u_{xxxx}(x+\vartheta h, y)$$

mit $-1 < \vartheta < 1$ und daher

$$\frac{\partial^2 u}{\partial x^2}(x,y) = \frac{u(x+h,y) - 2u(x,y) + u(x-h,y)}{h^2} - \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(x+\vartheta h, y)$$

$$\frac{\partial^2 u}{\partial y^2}(x,y) = \frac{u(x,y+h) - 2u(x,y) + u(x,y-h)}{h^2} - \frac{h^2}{12}\frac{\partial^4 u}{\partial y^4}(x,y+\eta h)$$

mit $-1 < \vartheta, \eta < 1$. Diese Differenzenformel besitzt die Ordnung 2. Wir ersetzen die Ableitungen in (2.7) durch diese Differenzenformeln in den Gitterpunkten (2.8). Sei $u_{i,j} := u(x_i, y_j)$, $f_{i,j} := f(x_i, y_j)$ und

$$(\Delta_h u)_{i,j} := \frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h^2} + \frac{u_{i,j-1} - 2u_{i,j} + u_{i,j+1}}{h^2}$$

Weglassen der Restterme ergibt ein lineares Gleichungssystem

$$4u_{i,j} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1} = h^2 f_{i,j} \tag{2.9}$$

für $i, j = 1, \ldots, M$. Die Diskretisierung kann in Form eines Fünf-Punkte-Sterns dargestellt werden, siehe Abbildung 2. Die homogenen Randbedingungen liefern

$$u_{0,j} = u_{M+1,j} = u_{i,0} = u_{i,M+1} = 0 \quad \text{für alle } \ i, j.$$

Wir ordnen die Unbekannten und die Auswertungen der rechten Seite f an in der Form

$$\begin{aligned} U_h &= (u_{1,1}, u_{2,1}, \ldots, u_{M,1}, u_{1,2}, \ldots, u_{M,2}, \ldots, u_{1,M}, \ldots, u_{M,M})^\top \\ F_h &= (f_{1,1}, f_{2,1}, \ldots, f_{M,1}, f_{1,2}, \ldots, f_{M,2}, \ldots, f_{1,M}, \ldots, f_{M,M})^\top. \end{aligned} \tag{2.10}$$

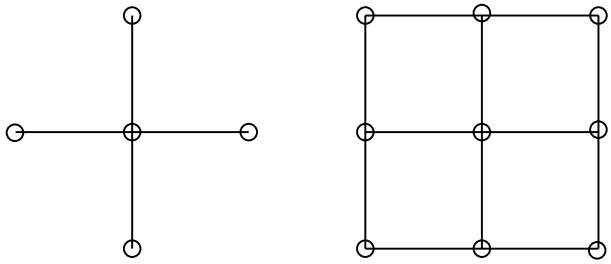


Abbildung 2: Fünf-Punkte-Stern (links) und Neun-Punkte-Stern (rechts).

Wir erhalten damit ein lineares Gleichungssystem $A_h U_h = F_h$ der Dimension $n = M^2$. Die Koeffizientenmatrix A_h ist eine Bandmatrix

$$A_h = \frac{1}{h^2} \begin{pmatrix} C & -I & & \\ -I & C & \ddots & \\ & \ddots & \ddots & -I \\ & & -I & C \end{pmatrix} \quad \text{mit} \quad C = \begin{pmatrix} 4 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{pmatrix}. \quad (2.11)$$

Zudem ist die Matrix A_h dünnbesetzt, weil jede Zeile höchstens fünf Einträge ungleich null besitzt. Offensichtlich ist die Matrix A_h symmetrisch. Es kann nachgewiesen werden, dass A_h stets positiv definit ist. Somit ist die Matrix regulär. Die zugehörige Lösung $U_h = A_h^{-1} F_h$ stellt eine Näherung für die Lösung u der Dgl. in den Gitterpunkten dar.

Laplace-Operator auf allgemeinem Gebiet

Nun betrachten wir ein beliebiges beschränktes Gebiet $\Omega \subset \mathbb{R}^2$, siehe Abbildung 3. Die Anwendung der Finiten-Differenzen-Methode erfordert die Konstruktion eines Gitters. Wir definieren ein (unendliches) Hilfsgitter

$$G_h = \{(x, y) = (ih, jh) : i, j \in \mathbb{Z}\}.$$

mit Schrittweite $h > 0$. Nun lautet das zu verwendende (endliche) Gitter

$$\Omega_h = G_h \cap \Omega.$$

Seien $\Omega_h = \{z_1, \ldots, z_R\}$ die Gitterpunkte. Randbedingungen treten in den Gitterpunkten

$$\partial\Omega_h = (\{(ih, y) : i \in \mathbb{Z}, y \in \mathbb{R}\} \cup \{(x, jh) : j \in \mathbb{Z}, x \in \mathbb{R}\}) \cap \partial\Omega.$$

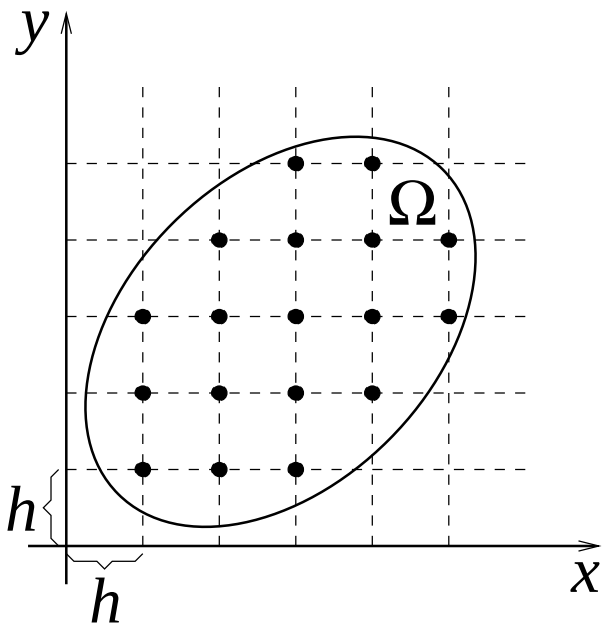


Abbildung 3: Gitter auf allgemeinem Gebiet Ω.

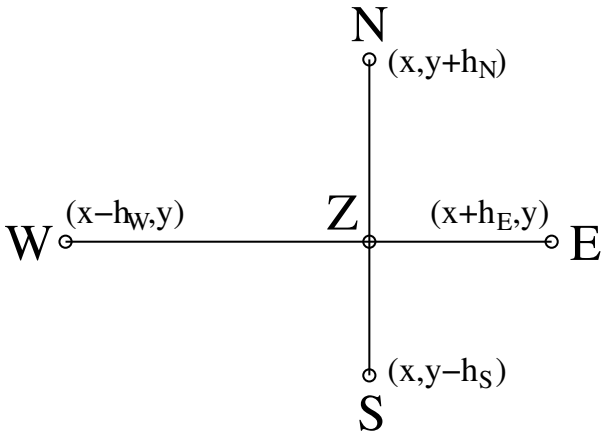


Abbildung 4: Fünf-Punkte-Stern mit variablen Schrittweiten.

auf. Die Differenzenformeln nahe des Rands enthalten jetzt variable Schrittweiten.

Wir wenden wieder den Fünf-Punkte-Stern an, siehe Abbildung 4. Taylor-Entwicklung liefert

$$\begin{aligned}
\left.\frac{\partial^2 u}{\partial x^2}\right|_Z &= \frac{2}{h_E(h_E+h_W)}u_E - \frac{2}{h_E h_W}u_Z + \frac{2}{h_W(h_E+h_W)}u_W + \mathcal{O}(h) \\
\left.\frac{\partial^2 u}{\partial y^2}\right|_Z &= \frac{2}{h_N(h_N+h_S)}u_N - \frac{2}{h_N h_S}u_Z + \frac{2}{h_S(h_N+h_S)}u_S + \mathcal{O}(h)
\end{aligned}$$

unter der Voraussetzung $u \in C^3(\Omega)$. Dabei treten vier (möglicherweise) unterschiedliche Schrittweiten auf. Sei $h = \max\{h_E, h_W, h_N, h_S\}$. Dieses Schema zur Einbeziehung der Randwerte wird auch *Shortley-Weller-Stern* genannt.

Im allgemeinen wird ein Fünf-Punkte-Stern und ein Neun-Punkte-Stern festgelegt durch seine Koeffizienten

$$\begin{bmatrix} & \alpha_N & \\ \alpha_W & \alpha_Z & \alpha_E \\ & \alpha_S & \end{bmatrix} \qquad \text{bzw.} \qquad \begin{bmatrix} \alpha_{NW} & \alpha_N & \alpha_{NE} \\ \alpha_W & \alpha_Z & \alpha_E \\ \alpha_{SW} & \alpha_S & \alpha_{SE} \end{bmatrix},$$

vergleiche Abbildung 2. Eine Diskretisierung der Poisson-Gleichung (2.7) mit einem beliebigen Fünf-Punkte-Stern lautet

$$\sum_{\ell=\mathrm{Z,E,S,W,N}} \alpha_\ell U_\ell = f_Z$$

für jedes $Z \in \Omega_h$. Eine allgemeine Differenzenformel kann in der Gestalt

$$L_h U = \sum_\ell \alpha_\ell U_\ell$$

geschrieben werden, wobei die Summe über alle Koeffizienten α_ℓ ungleich null läuft. Der Differenzenoperator L_h hängt von den Schrittweiten ab. Wir verwenden die Notation

$$L_h u = \sum_\ell \alpha_\ell u(Z_\ell),$$

wobei eine Funktion $u : \bar{\Omega} \to \mathbb{R}$ (üblicherweise eine Lösung der Dgl.) an den Gitterpunkten $Z_\ell \in \Omega_h \cup \partial\Omega_h$ ausgewertet wird.

Wir skizzieren einen Algorithmus der Finiten-Differenzen-Methode für ein Dirichlet-Randwertproblem der Poisson-Gleichung auf allgemeinem zweidimensionalen Ortsgebiet Ω.

Algorithmus: *Finite-Differenzen-Methode für Dirichlet-Problem*

1. Wähle eine Schrittweite $h > 0$ und bestimme Ω_h sowie $\partial\Omega_h$.
2. Wähle eine Nummerierung der Unbekannten U_Z für $Z \in \Omega_h$.
3. Stelle die Differenzenformeln

   $$\alpha_{\mathrm{Z}} U_{\mathrm{Z}} + \alpha_{\mathrm{E}} U_{\mathrm{E}} + \alpha_{\mathrm{W}} U_{\mathrm{W}} + \alpha_{\mathrm{N}} U_{\mathrm{N}} + \alpha_{\mathrm{S}} U_{\mathrm{S}} = f_{\mathrm{Z}}$$

   für jedes $Z \in \Omega_h$ auf.
4. Wenn Randwerte U_R mit $R \in \partial\Omega_h$ in der linken Seite einer Differenzenformel auftreten, dann ersetze U_R durch g_R und verschiebe den Term zur rechten Seite.
5. Löse das entstandene lineare Gleichungssystem

   $$A_h U_h = F_h$$

   mit der gewählten Nummerierung der Unbekannten $U_h = (U_{Z_i})$ für $Z_i \in \Omega_h$.

Es verbleibt noch nachzuweisen, dass die Koeffizientenmatrix A_h regulär ist.

Allgemeiner Differentialoperator

Differenzenschemata existieren ebenfalls für gemischte Ableitungen. Beispielsweise gilt

$$\frac{\partial^2 u}{\partial x \partial y}(x_i, y_j) = \frac{u_{i+1,j+1} - u_{i-1,j+1} - u_{i+1,j-1} + u_{i-1,j-1}}{4h^2} + \mathcal{O}(h^2). \quad (2.12)$$

Wir bestätigen diese Formel durch (mehrdimensionale) Taylor-Entwicklung in den Nachbarpunkten des zentralen Punkts $u \equiv u_{i,j}$.

$$\begin{aligned}
u_{i+1,j+1} &= u + hu_x + hu_y + \tfrac{1}{2}h^2(u_{xx} + 2u_{xy} + u_{yy}) \\
&\quad + \tfrac{1}{6}h^3(u_{xxx} + 3u_{xxy} + 3u_{xyy} + u_{yyy}) + \mathcal{O}(h^4) \\
u_{i-1,j+1} &= u - hu_x + hu_y + \tfrac{1}{2}h^2(u_{xx} - 2u_{xy} + u_{yy}) \\
&\quad + \tfrac{1}{6}h^3(-u_{xxx} + 3u_{xxy} - 3u_{xyy} + u_{yyy}) + \mathcal{O}(h^4) \\
u_{i+1,j-1} &= u + hu_x - hu_y + \tfrac{1}{2}h^2(u_{xx} - 2u_{xy} + u_{yy}) \\
&\quad + \tfrac{1}{6}h^3(u_{xxx} - 3u_{xxy} + 3u_{xyy} - u_{yyy}) + \mathcal{O}(h^4) \\
u_{i-1,j-1} &= u - hu_x - hu_y + \tfrac{1}{2}h^2(u_{xx} + 2u_{xy} + u_{yy}) \\
&\quad + \tfrac{1}{6}h^3(-u_{xxx} - 3u_{xxy} - 3u_{xyy} - u_{yyy}) + \mathcal{O}(h^4)
\end{aligned}$$

$$\Rightarrow \quad u_{i+1,j+1} - u_{i-1,j+1} - u_{i+1,j-1} + u_{i-1,j-1} = 4h^2 u_{xy} + \mathcal{O}(h^4)$$

Nun können wir einen beliebigen Differentialoperator zweiter Ordnung

$$Lu := a\,\frac{\partial^2 u}{\partial x^2} + 2b\,\frac{\partial^2 u}{\partial x \partial y} + c\,\frac{\partial^2 u}{\partial y^2} \quad (2.13)$$

mit $a, b, c \in \mathbb{R}$ für $n = 2$ diskretisieren. Der Operator (2.13) ist elliptisch genau dann, wenn $ac > b^2$ gilt. Die Ableitungen $\frac{\partial^2 u}{\partial x^2}, \frac{\partial^2 u}{\partial y^2}$ werden durch Differenzenformeln für Δ_h und die gemischten Ableitungen $\frac{\partial^2 u}{\partial x \partial y}$ werden durch die Differenzenformeln (2.12) ersetzt. Es folgt der (diskrete) Differenzenoperator

$$\begin{aligned}
L_h u \;=\;& \tfrac{a}{h^2}[u_{i-1,j} - 2u_{i,j} + u_{i+1,j}] \\
+\;& \tfrac{b}{2h^2}[u_{i+1,j+1} - u_{i-1,j+1} - u_{i+1,j-1} + u_{i-1,j-1}] \\
+\;& \tfrac{c}{h^2}[u_{i,j-1} - 2u_{i,j} + u_{i,j+1}].
\end{aligned} \quad (2.14)$$

Dieses Differenzenschema entspricht einem Neun-Punkte-Stern, siehe Abbildung 2.

Konsistenz, Stabilität und Konvergenz

Wir sind an der Konvergenz einer Finiten-Differenzen-Methode interessiert, d.h. ob der globale Fehler gegen null konvergiert für abnehmende Schrittweite. Die Konsistenz einer Differenzenformel bezüglich des Differentialoperators allein ist nicht hinreichend für die Konvergenz. Wir benötigen zusätzlich eine Stabilitätseigenschaft um die Konvergenz zu erhalten. Die Konzepte sind ähnlich wie bei der numerischen Lösung von Anfangswertproblemen zu gewöhnlichen Dgln. mittels eines Mehrschrittverfahrens.

Wir definieren einen lokalen Fehler und einen globalen Fehler.

Definition 2.4 (lokaler und globaler Fehler) *Gegeben sei $\Omega \subset \mathbb{R}^n$ und ein (endliches) Gitter $\Omega_h \subset \Omega$. Sei L ein Differentialoperator und L_h ein Differenzenoperator. Für eine hinreichend glatte Funktion $u : \Omega \to \mathbb{R}$ ist der lokale Fehler festgelegt durch $\tau(h) := Lu - L_h u$ auf Ω_h. Falls u eine Lösung einer Dgl. $Lu = f$ und $U \in \mathbb{R}^{|\Omega_h|}$ eine numerische Lösung auf Ω_h ist, dann lautet der globale Fehler $\eta(z_i) := u(z_i) - U_i$ für jedes $z_i \in \Omega_h$.*

Nun erfolgt die Definition der Konvergenz über den globalen Fehler.

Definition 2.5 (Konvergenz) *Ein numerisches Verfahren mit dem Differenzenopertor L_h heißt konvergent, falls für den globalen Fehler gilt*

$$\lim_{h \to 0} \max_{z_i \in \Omega_h} |\eta(z_i)| = 0.$$

Ein Verfahren heißt konvergent mit Ordnung (mindestens) p, falls gilt

$$\max_{z_i \in \Omega_h} |\eta(z_i)| = \mathcal{O}(h^p).$$

Im Grenzfall $h \to 0$ geht die Anzahl der Gitterpunkte üblicherweise gegen unendlich, d.h. $|\Omega_h| \to \infty$. Um ein konvergentes Verfahren zu erhalten stellt die Konsistenz eine wesentliche Eigenschaft dar.

Definition 2.6 (Konsistenz) *Ein Differenzenoperator L_h heißt konsistent bezüglich eines Differentialoperators $L : V \to W$ falls für den lokalen Fehler*

$$\lim_{h \to 0} \tau(h) = 0 \qquad \textit{gleichmäßig auf } \Omega_h$$

für alle Funktionen $u \in V$ gilt.
Das Verfahren heißt konsistent mit Ordnung (mindestens) p, falls

$$\tau(h) = \mathcal{O}(h^p) \qquad \textit{gleichmäßig auf } \Omega_h$$

für alle Funktionen $u \in V$ gilt.

Beispielsweise ist der Differenzenoperator (2.14) bezüglich des Differentialoperators (2.13) mit $V = C^4(\bar{\Omega})$ konsistent mit Ordnung $p = 2$.

Wir untersuchen die Konvergenz für das Dirichlet-Randwertproblem der Poisson-Gleichung, d.h. eine Diskretisierung Δ_h des Laplace-Operators Δ. Es gilt

$$A_h U_h = F_h, \qquad A_h \hat{U}_h = F_h + R_h,$$

wobei $\hat{U}_h = (u(z_i))$ die Werte der exakten Lösung in den Gitterpunkten darstellen. Es ist R_h ein Vektor, welcher die lokalen Fehler $\tau(h)$ enthält. Da die Differenzenformel konsistent mit Ordnung $p = 2$ ist, folgt in der Vektornorm $\|R_h\|_\infty = \mathcal{O}(h^2)$. Wir verwenden die Maximumnorm, da die Länge der Vektoren von der Schrittweite h abhängt. Unter der Annahme, dass A_h regulär für alle $h > 0$ ist, erhalten wir

$$U_h - \hat{U}_h = A_h^{-1} F_h - A_h^{-1}(F_h + R_h) = -A_h^{-1} R_h$$

und somit

$$\|U_h - \hat{U}_h\|_\infty \leq \left\|A_h^{-1}\right\|_\infty \|R_h\|_\infty \leq C \left\|A_h^{-1}\right\|_\infty h^2$$

mit einer Konstanten $C > 0$ und hinrechend kleinem $h > 0$. Um die Konvergenz zu garantieren benötigen wir jetzt eine Bedingung wie

$$\left\|A_h^{-1}\right\|_\infty \leq K \qquad \text{oder} \qquad \left\|A_h^{-1}\right\|_\infty \leq \frac{K}{h}$$

gleichmäßig für alle $h < h_0$ mit einer Konstanten $K > 0$. Eine derartige Bedingung entspricht der Stabilität der Finiten-Differenzen-Methode.

Wir erhalten ein allgemeines Kriterium für die Stabilität aus folgendem theoretischen Resultat. Dabei benötigen wir die Annahme, dass das Gitter Ω_h zusammenhängend ist.

Definition 2.7 (zusammenhängendes Gitter) *Ein Gitter $\Omega_h \subset \Omega \subset \mathbb{R}^2$ heißt zusammenhängend, wenn je zwei Punkte des Gitters durch Geradenstücke im verwendeten Differenzenstern verbunden werden können, die stets innerhalb des Gitters sowie innerhalb Ω liegen.*

Typischerweise liegt ein zusammenhängendes Gitter für hinreichend kleine Schrittweite vor. Nun können wir eine diskrete Version zu Satz 2.1 aufstellen.

Satz 2.8 (diskretes Maximumprinzip)
Gegeben sei eine elliptische Dgl. $Lu = f$ mit $f \leq 0$ in Ω und Dirichlet-Randbedingungen. Sei L_h ein Differenzenoperator in Form eines Fünf-Punkte-Sterns auf einem zusammenhängenden Gitter Ω_h mit negativen Koeffizienten außerhalb des Zentrums und die Summe aller Koeffizienten sei null. Wenn die Werte $\{U_Z : Z \in \Omega_h \cup \partial\Omega_h\}$ das Finite-Differenzen-Schema erfüllen, dann sind entweder alle Werte U_Z konstant oder ein Maximum der Werte U_Z befindet sich nicht in Ω_h sondern auf dem Rand $\partial\Omega_h$.

Beweis:

Wir nehmen an, dass

$$\max_{Z \in \Omega_h} U_Z \geq \max_{R \in \partial\Omega_h} U_R$$

gilt, und zeigen dann, dass U_Z konstant auf $\Omega_h \cup \partial\Omega_h$ ist. Dadurch befindet sich das diskrete Maximum stets auf dem Rand. Sei U_Z das Maximum in Ω_h. Mit den Nachbarwerten gilt die Ungleichung

$$U_Z \geq \max_{\ell=\mathrm{E,W,N,S}} U_\ell.$$

Desweiteren impliziert die Differenzenformel

$$\sum_{\ell=\mathrm{Z,E,S,W,N}} \alpha_\ell U_\ell = f_Z \leq 0.$$

Es folgt

$$\begin{aligned}\sum_{\ell=\mathrm{E,S,W,N}} \alpha_\ell (U_\ell - U_Z) &= \sum_{\ell=\mathrm{Z,E,S,W,N}} \alpha_\ell (U_\ell - U_Z) \\ = \left(\sum_{\ell=\mathrm{Z,E,S,W,N}} \alpha_\ell U_\ell \right) - \left(U_Z \sum_{\ell=\mathrm{Z,E,S,W,N}} \alpha_\ell \right) &= \sum_{\ell=\mathrm{Z,E,S,W,N}} \alpha_\ell U_\ell \leq 0.\end{aligned}$$

Es gilt $\alpha_\ell < 0$ und $U_\ell - U_Z \leq 0$ für alle ℓ in der ursprünglichen Summe. Dadurch sind alle Terme in der ursprünglichen Summe nichtnegativ. Somit muss jeder Term identisch null sein. Aus $\alpha_\ell < 0$ für $\ell \in \{\mathrm{E,S,W,N}\}$ folgt

$$U_Z = U_E = U_W = U_N = U_S,$$

d.h. die Nachbarn von U_Z haben den gleichen Wert. Wir setzen diese Argumentation ausgehend von U_Z bis zum Rand fort. Jeder Nachbar des ursprünglichen U_Z erfüllt die Annahmen und daher haben dessen Nachbarn ebenfalls den gleichen Wert. Dadurch ist U_Z konstant für sowohl alle $Z \in \Omega_h$ als auch alle $Z \in \partial\Omega_h$. Bei dieser Folgerung verwenden wir, dass das Gitter Ω_h zusammenhängend ist, wodurch je zwei Gitterpunkte aus $\Omega_h \cup \partial\Omega_h$ durch Differenzenformeln verbunden sind. □

Bemerkungen:

- Das diskrete Maximumprinzip gilt ebenfalls für Differentialoperatoren aus einem Neun-Punkte-Stern mit entsprechenden Bedingungen an die Koeffizienten. Nur Nahe des Rands sind Fünf-Punkte-Sterne zu verwenden.
- Der Differenzenoperator (2.14) ist konsistent mit Ordnung 2 bezüglich des Differentialoperators (2.13). Jedoch erfüllt die Differenzenformel hier nicht die Bedingung $\alpha_\ell < 0$ für alle Koeffizienten außerhalb des Zentrums. Weitergehende Untersuchungen zeigen, dass in diesem Fall zusätzliche Stabilitätsbedingungen notwendig sind.

Weitere Folgerungen aus dem diskreten Maximumprinzip in Satz 2.8 sind:

- **Diskretes Minimumprinzip:** Falls $Lu = f$ mit $f \geq 0$ gilt, dann ist die diskrete Lösung entweder konstant oder nimmt ein Minimum nicht in Ω_h sondern auf dem Rand $\partial\Omega_h$ an.
- **Diskreter Vergleich:** Falls $L_h U_Z \leq L_h V_Z$ für alle Gitterpunkte $Z \in \Omega_h$ und $U_R \leq V_R$ für alle $R \in \partial\Omega_h$ gilt, dann folgt $U_Z \leq V_Z$ für alle $Z \in \Omega_h$.

Nun können wir nachweisen, dass ein lineares Gleichungssystem aus unserer Finiten-Differenzen-Methode eine eindeutige Lösung besitzt.

Satz 2.9 *Wir betrachten eine elliptische Dgl. $Lu = f$ auf Ω mit Dirichlet-Randbedingungen $u = g$ auf $\partial\Omega$. Sei $A_h U_h = F_h$ ein lineares Gleichungssystem aus einem Finite-Differenzen-Verfahren, welches die Voraussetzungen des diskreten Maximumprinzips erfüllt. Dann ist die Matrix A_h regulär und somit existiert eine eindeutige Lösung des linearen Gleichungssystems.*

Beweis:

Das homogene lineare Gleichungssystem $A_h U_h = 0$ stellt die Diskretisierung der elliptischen Dgl. mit $f \equiv 0$ und $g \equiv 0$ dar. Sei U_h eine Lösung des linearen Gleichungssystems. Das diskrete Maximumprinzip zeigt $U_Z \leq 0$ in jedem $Z \in \Omega_h$, während das diskrete Minimumprinzip $U_Z \geq 0$ in jedem $Z \in \Omega_h$ liefert. Es folgt $U_h = 0$. Somit ist die Matrix nicht singulär. $\square$

Es verbleibt zu zeigen, dass die Finite-Differenzen-Methode konvergent ist. Im folgenden diskutieren wir die Konvergenz nur für Diskretisierungen des Laplace-Opertors, d.h. $Lu = -\Delta u$. Wir verwenden den Fünf-Punkte-Stern, welcher konsistent mit Ordnung mindestens 1 ist und das diskrete Maximumprinzip erfüllt.

Lemma 2.10 *Sei $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$ die Lösung der Poisson-Gl. $-\Delta u = f$ mit Dirichlet-Randbedingungen $u = g$ auf $\partial\Omega$. Es bezeichne L_h den Differenzenopertor zum Fünf-Punkte-Stern auf einem Gitter Ω_h. Dann gilt für den lokalen Fehler und den globalen Fehler aus der Finiten-Differenzen-Methode die Abschätzung*

$$\max_{Z \in \Omega_h} |u(Z) - U_Z| \leq K \max_{Z \in \Omega_h} |\tau(Z)| \tag{2.15}$$

mit einer Konstanten $K > 0$, welche unabhängig von der Schrittweite h ist.

Beweis:

Wir untersuchen die lokalen Fehler und globalen Fehler

$$\tau(Z) = -\Delta u(Z) - L_h u(Z), \qquad \eta(Z) = u(Z) - U_Z \qquad \text{für } Z \in \Omega_h.$$

Mit der Linearität der Operatoren folgt

$$L_h \eta(Z) = L_h u(Z) - L_h U_Z = L_h u(Z) - f(Z) = L_h u(Z) + \Delta u(Z) = -\tau(Z).$$

Der globale Fehler verschwindet auf dem Rand $\partial\Omega_h$, da dort die Näherungslösung identisch mit den vorgegebenen Randwerten ist. Wir untersuchen das Problem

$$L_h \eta = -\tau \quad \text{in } \Omega_h, \qquad \eta = 0 \quad \text{auf } \partial\Omega_h. \tag{2.16}$$

Die Skalierung

$$\tilde{\eta} := \frac{\eta}{\gamma}, \quad \tilde{\tau} := \frac{\tau}{\gamma} \qquad \text{mit } \gamma := \max_{Z \in \Omega_h} |\tau(Z)|$$

erzeugt das Problem

$$L_h \tilde{\eta} = -\tilde{\tau} \qquad \text{mit } -1 \leq \tilde{\tau}(Z) \leq 1 \qquad \text{für alle } Z \in \Omega_h.$$

Es gilt immer noch $\tilde{\eta} = 0$ auf $\partial\Omega_h$. Sei $\Omega \subset \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < R^2\}$. Wir definieren die Hilfsfunktion

$$w(x, y) := \tfrac{1}{4}(R^2 - x^2 - y^2) \geq 0.$$

Da alle dritten Ableitungen von w identisch null sind, verschwinden die lokalen Fehler im Fünf-Punkte-Stern. Es folgt $L_h w = -\Delta w = 1$ in Ω_h.

Weil der Fünf-Punkte-Stern das diskrete Maximumprinzip erfüllt, ergibt der diskrete Vergleich

$$\tilde{\eta} \leq w \leq \tfrac{1}{4}R^2 \qquad \text{für alle } Z \in \Omega_h.$$

Mit $-w$ anstelle von w liefert der diskrete Vergleich

$$-\tfrac{1}{4}R^2 \leq -w \leq \tilde{\eta} \qquad \text{für alle } Z \in \Omega_h.$$

Somit folgt

$$\max_{Z\in\Omega_h} |\eta(Z)| \leq \tfrac{1}{4}R^2 \max_{Z\in\Omega_h} |\tau(Z)|$$

und wir können die Konstante $K = \frac{R^2}{4}$ wählen. □

Im Beweis zu Lemma 2.10 werden sowohl die Konsistenz als auch das diskrete Maximumprinzip angewendet. Das diskrete Maximumprinzip garantiert hier die Stabilität der Finiten-Differenzen-Methode. Jetzt können wir die Konvergenz zeigen.

Satz 2.11 *Sei $u \in C^3(\bar{\Omega})$ die Lösung der Poisson-Gleichung $-\Delta u = f$ mit Dirichlet-Randbedingungen $u = g$ auf $\partial\Omega$. Dann konvergiert die Näherungslösung aus der Diskretisierung mit dem Fünf-Punkte-Stern gegen die exakte Lösung und es gilt*

$$\max_{Z\in\Omega_h} |u(Z) - U_Z| = \mathcal{O}(h).$$

Falls $u \in C^4(\bar{\Omega})$ gilt und alle Schrittweiten identisch sind, dann gilt zudem

$$\max_{Z\in\Omega_h} |u(Z) - U_Z| = \mathcal{O}(h^2).$$

Beweis:

Die Konsistenz des Fünf-Punkte-Sterns liefert $\tau(Z) = \mathcal{O}(h)$ für alle $Z \in \Omega_h$. Die Abschätzung (2.15) aus Lemma 2.10 impliziert

$$\max_{Z\in\Omega_h} |\eta(Z)| \leq \tfrac{1}{4}R^2 Ch \qquad \text{für } 0 < h < h_0$$

mit Konstanten $C, h_0 > 0$. Im Fall von $u \in C^4(\bar{\Omega})$ und identischen Schrittweiten folgt $\tau(Z) = \mathcal{O}(h^2)$. Der globale Fehler ergibt sich nun analog. □

Weitere Untersuchungen zeigen, dass die Konvergenz mit Ordnung $p = 2$ auch für variable Schrittweiten (Shortley-Weller-Stern am Rand) gilt. Zudem liegt die Konvergenz (ohne eine Ordnungsaussage) bereits unter der schwächeren Voraussetzung $u \in C^2(\bar{\Omega})$ vor.

Die Stabilität eines numerischen Verfahrens wird üblicherweise definiert als die Lipschitzstetige Abhängigkeit der Näherungslösung von den Eingabedaten gleichmäßig für alle (hinreichend kleinen) Schrittweiten. Wir demonstrieren diese Eigenschaft durch die Untersuchung zweier Probleme

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{auf } \partial\Omega \end{aligned} \qquad \text{und} \qquad \begin{aligned} -\Delta \tilde{u} &= \tilde{f} \quad \text{in } \Omega \\ \tilde{u} &= 0 \quad \text{auf } \partial\Omega \end{aligned}$$

für $\Omega \subset \mathbb{R}^2$ und $f, \tilde{f} \in C^0(\bar{\Omega})$. Seien U und $\tilde{U}$ die Näherungslösungen aus einer Finiten-Differenzen-Methode auf einem Gitter Ω_h, d.h.

$$L_h U(Z) = f(Z) \qquad \text{und} \qquad L_h \tilde{U}(Z) = \tilde{f}(Z)$$

für $Z \in \Omega_h$. Wegen der Linearität der Differenzenoperatoren liefert eine Subtraktion

$$L_h(U(Z) - \tilde{U}(Z)) = f(Z) - \tilde{f}(Z).$$

Daher erhalten wir die gleiche Struktur (2.16) wie im Beweis zu Lemma 2.10. Die Aussage (2.15) ergibt nun

$$\max_{Z \in \Omega_h} |U(Z) - \tilde{U}(Z)| \leq K \max_{Z \in \Omega_h} |f(Z) - \tilde{f}(Z)| \leq K \max_{(x,y) \in \bar{\Omega}} |f(x,y) - \tilde{f}(x,y)| = K \, \|f - \tilde{f}\|_\infty.$$

Daher hängen die Näherungslösungen Lipschitz-stetig von den Eingabedaten f ab mit einer Lipschitz-Konstanten K, welche unabhängig von der Schrittweite h ist.

Verallgemeinerungen

Wir skizzieren Verallgemeinerungen der vorgestellten Finiten-Differenzen-Methode für weitere Probleme.

- *von-Neumann-Randwertproblem*: Wir betrachten die Poisson-Gleichung $-\Delta u = f$ in $\Omega \subset \mathbb{R}^2$ mit Randbedingungen $\frac{\partial u}{\partial \nu} = g(x,y)$ auf $\partial\Omega$. Sei $\Omega = (0,1) \times (0,1)$. Wir verwenden die Gitterpunkte (2.8) für $i, j = 0, 1, \ldots, M, M+1$. Im Vergleich zum Dirichlet-Randwertproblem

treten jetzt $4M$ zusätzliche Unbekannte auf (die vier Echem im Quadrat werden nicht verwendet). Daher stellen wir $4M$ Gleichungen mit Differenzenformeln auf, welche die Ableitung $\frac{\partial u}{\partial \nu}$ ersetzen:

$$\begin{array}{rcrcll} g(x_i,0) & = & -\frac{\partial u}{\partial y}(x_i,0) & \approx & \frac{1}{h}[u_{i,0}-u_{i,1}] & \text{für } i=1,\ldots,M, \\ g(x_i,1) & = & \frac{\partial u}{\partial y}(x_i,1) & \approx & \frac{1}{h}[u_{i,M+1}-u_{i,M}] & \text{für } i=1,\ldots,M, \\ g(0,y_j) & = & -\frac{\partial u}{\partial x}(0,y_j) & \approx & \frac{1}{h}[u_{0,j}-u_{1,j}] & \text{für } j=1,\ldots,M, \\ g(1,y_j) & = & \frac{\partial u}{\partial x}(1,y_j) & \approx & \frac{1}{h}[u_{M+1,j}-u_{M,j}] & \text{für } j=1,\ldots,M. \end{array}$$

Verfahren für Neumann-Randbedingungen auf beliebigen Gebieten Ω existieren ebenfalls. Lösungen für ein reines Neumann-Randwertproblem sind nicht eindeutig und erfordern eine zusätzliche Bedingung.

- *ortsabhängige Koeffizienten* : Gegeben sei die elliptische Dgl.

$$Lu = a(x,y)\,\frac{\partial^2 u}{\partial x^2} + 2b(x,y)\,\frac{\partial^2 u}{\partial x \partial y} + c(x,y)\,\frac{\partial^2 u}{\partial y^2} = f(x,y)$$

mit ortsabhängigen Koeffizienten $a, b, c : \Omega \to \mathbb{R}$. Geeignete Finite-Differenzen-Verfahren können für diesen Fall konstruiert werden. Für die Konvergenz ist ein gleichmäßig elliptischer Operator, siehe Definition 2.3, vorauszusetzen.

- *rechte Seite enthält die Lösung*: Wir betrachten die (semi-lineare) Dgl. $-\Delta u = f(x,y,u)$ mit einer nichtlinearen Funktion f. Homogene Randbedingungen $u = 0$ auf $\partial\Omega$ seien vorgegeben. Sei Ω das Einheitsquadrat. Der Fünf-Punkte-Stern liefert die Gleichungen

$$\tfrac{1}{h^2}[4u_{i,j} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1}] = f(x_i, y_j, u_{i,j})$$

für $i,j = 1,\ldots,M$. Wir erhalten ein nichtlineares Gleichungssystem für die Unbekannten $u_{i,j}$. Das Newton-Verfahren kann eine zugehörige Näherungslösung erzeugen. Im Spezialfall $f(x,y,u) = b(x,y) + c(x,y)u$ ergibt sich wieder ein lineares Gleichungssystem.

- *dreidimensionales Ortsgebiet*: Der Laplace-Operator in drei Ortskoordinaten lautet $\Delta u = u_{xx} + u_{yy} + u_{zz}$. Wir betrachten ein beschränktes Gebiet $\Omega \subset \mathbb{R}^3$ für die Poisson-Gleichung $-\Delta u = f$. Die Theorie für Finite-Differenzen-Methoden kann in diesem Fall wiederholt werden. Es gelten die gleichen Resultate bezüglich Konsistenz, Stabilität und Konvergenz.

2.3 Sobolev-Räume und Variationsform

In diesem Abschnitt führen wir schwache Lösungen von elliptischen Dgln. ein. Finite-Elemente-Methoden sind konvergente numerische Verfahren für schwache Lösungen, während Finite-Differenzen-Methoden hier versagen.

Klassische Lösungen

Klassische Lösungen (starke Lösungen) einer partiellen Dgl. sind hinreichend glatt in einem gewissen Sinn.

Definition 2.12 (klassische Lösungen) *Sei $\Omega \subset \mathbb{R}^n$ offen, zusammenhängend und beschränkt. Zu einer elliptischen Dgl. $Lu = f$ heißt eine Funktion $u : \bar{\Omega} \to \mathbb{R}$ eine klassische Lösung, wenn gilt*

- *für Dirichlet-Randwertprobleme: $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$,*
- *für Neumann-Randwertprobleme: $u \in C^2(\Omega) \cap C^1(\bar{\Omega})$.*

Als ein Beispiel mit $n = 2$ betrachten wir die Laplace-Gleichung auf einem Dreiviertel des Einheitskreises als Definitionsbereich, d.h.

$$\Omega = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1, x < 0 \text{ oder } y > 0\}.$$

Wir identifizieren $\Omega \subset \mathbb{C}$ über $x = \mathrm{Re}(z)$, $y = \mathrm{Im}(z)$. Nun ist die Funktion $w : \Omega \to \mathbb{C}$, $w(z) = z^{2/3}$ analytisch. (Für $z = r\mathrm{e}^{\mathrm{i}\varphi}$ mit $r > 0$ sowie $0 < \varphi < \frac{3\pi}{2}$ ist dabei $z^{2/3} = r^{2/3}\mathrm{e}^{\mathrm{i}\frac{2}{3}\varphi}$.) Der Imaginärteil $u = \mathrm{Im}(w)$ erfüllt die Gleichungen

$$\begin{aligned} \Delta u &= 0 && \text{in } \Omega, \\ u(\mathrm{e}^{\mathrm{i}\varphi}) &= \sin(\tfrac{2}{3}\varphi) && \text{für } 0 \le \varphi \le \tfrac{3\pi}{2}, \\ u &= 0 && \text{sonst auf } \partial\Omega. \end{aligned}$$

Wir erhalten die Darstellung

$$u(x, y) = \sqrt[3]{x^2 + y^2}\,\sin(\tfrac{2}{3}\arctan(\tfrac{y}{x})) \qquad \text{für } x > 0.$$

Es folgt $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$, wodurch u eine klassische Lösung des Dirichlet-Randwertproblems darstellt. Wegen $w'(z) = \frac{2}{3}z^{-1/3}$ und $w''(z) = -\frac{2}{9}z^{-4/3}$

sind aber sowohl die erste als auch die zweite Ableitung von u unbeschränkt in einer Umgebung von $z = 0$. Es folgt $u \notin C^2(\bar{\Omega})$. Dadurch können wir die Konvergenz einer Finiten-Differenzen-Methode, wie sie in Abschnitt 2.2 konstruiert wurde, nicht garantieren. Ein alternatives numerisches Verfahren ist erforderlich.

Schwache Ableitungen und Sobolev-Räume

Sei $\Omega \subseteq \mathbb{R}^n$ eine nichtleere offene Menge. Wir definieren den Raum der Testfunktionen als

$$C_0^\infty(\Omega) := \{\phi \in C^\infty(\Omega) : \operatorname{supp}(\phi) \subset \Omega,\ \operatorname{supp}(\phi) \text{ ist kompakt}\},$$

wobei $\operatorname{supp}(\phi) = \overline{\{x \in \Omega : \phi(x) \neq 0\}}$ den Träger (support) von ϕ bezeichnet. Falls Ω beschränkt ist, so folgt $\phi(x) = 0$ für $x \in \partial\Omega$. Desweiteren verwenden wir den Hilbert-Raum $L^2(\Omega)$. Dessen Skalarprodukt lautet

$$\langle f, g\rangle_{L^2} = \int_\Omega f(x) \cdot g(x)\ \mathrm{d}x$$

für $f, g \in L^2(\Omega)$. Die induzierte Norm ist

$$\|f\|_{L^2} = \sqrt{\int_\Omega f(x)^2\ \mathrm{d}x}\ .$$

Die Menge $C_0^\infty(\Omega) \subset L^2(\Omega)$ liegt dicht. Ein Multiindex besitzt die Gestalt

$$\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}_0^n, \qquad |\alpha| = \sum_{i=1}^n \alpha_i.$$

Für $u \in C^k(\Omega)$ mit $k = |\alpha|$ lautet ein elementarer Differentialoperator

$$\mathrm{D}^\alpha u = \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}.$$

Für $u \in C^k(\Omega)$ ist $\mathrm{D}^\alpha u$ die übliche (starke) Ableitung.

Definition 2.13 (schwache Ableitung) *Sei $f \in L^2(\Omega)$. Eine Funktion $g \in L^2(\Omega)$ heißt die schwache Ableitung $\mathrm{D}^\alpha f$ von f, wenn*

$$\int_\Omega g(x) \cdot \phi(x)\, \mathrm{d}x = (-1)^{|\alpha|} \int_\Omega f(x) \cdot \mathrm{D}^\alpha \phi(x)\, \mathrm{d}x$$

für alle $\phi \in C_0^\infty(\Omega)$ gilt. Wir schreiben $\mathrm{D}^\alpha f = g$.

Die Bedingung aus dieser Definition können wir schreiben als

$$\langle g, \phi \rangle_{L^2} = (-1)^{|\alpha|} \langle f, \mathrm{D}^\alpha \phi \rangle_{L^2} \qquad \text{für alle } \phi \in C_0^\infty(\Omega).$$

Es kann leicht gezeigt werden, dass die schwache Ableitung einer Funktion eindeutig ist, da der Raum $C_0^\infty(\Omega)$ dicht in $L^2(\Omega)$ liegt.

Beispiel: Im Spezialfall $n = 1$ sei $f \in C^1(a,b) \cap C^0[a,b]$. Es folgt

$$\int_a^b f'(x)\phi(x)\, \mathrm{d}x = [f(x)\phi(x)]_{x=a}^{x=b} - \int_a^b f(x)\phi'(x)\, \mathrm{d}x = -\int_a^b f(x)\phi'(x)\, \mathrm{d}x$$

für jedes $\phi \in C_0^\infty(a,b)$. Somit ist f' die schwache Ableitung von f.

Im allgemeinen Fall $n \geq 1$ liefert das folgende Lemma eine Aussage. Eine Skizze des zugehörigen Beweises ist enthalten in: Werner, Funktionalanalysis, Springer, S. 195.

Lemma 2.14 (partielle Integration) *Sei $\Omega \subseteq \mathbb{R}^n$ nichtleer und offen. Falls $f, g \in C^1(\Omega)$ und der Träger von g kompakt ist mit $\mathrm{supp}(g) \subset \Omega$, dann folgt*

$$\int_\Omega \frac{\partial f}{\partial x_i} \cdot g\, \mathrm{d}x = -\int_\Omega f \cdot \frac{\partial g}{\partial x_i}\, \mathrm{d}x \qquad \textit{für } i = 1, \ldots, n.$$

Lemma 2.14 impliziert, dass aus $f \in C^k(\Omega)$ die Existenz aller schwachen Ableitungen $\mathrm{D}^\alpha f$ mit $|\alpha| \leq k$ folgt, sofern diese Funktionen auch in $L^2(\Omega)$ liegen. In diesem Fall sind dann die üblichen Ableitungen auch die schwachen Ableitungen.

Eine andere Verallgemeinerung der partiellen Integration für beschränkte Gebiete $\Omega \subset \mathbb{R}^n$ mit $n \geq 2$ ist die Green'sche Formel

$$\int_\Omega v \cdot \frac{\partial w}{\partial x_i}\, \mathrm{d}x = \int_{\partial\Omega} v \cdot w \cdot \nu_i\, \mathrm{d}s - \int_\Omega \frac{\partial v}{\partial x_i} \cdot w\, \mathrm{d}x \tag{2.17}$$

für $i = 1, \ldots, n$ und $v, w \in C^1(\bar{\Omega})$. Dabei bezeichnet ν_i die i-te Komponente des äußeren Normalenvektors auf dem Rand $\partial\Omega$ und $\mathrm{d}s$ ist ein zugehöriges Differential. Die Formel (2.17) gilt nicht für beliebige Gebiete Ω. Jedoch sind die erforderlichen Voraussetzungen häufig bei den in der Praxis verwendeten Gebieten erfüllt. Zur Vereinfachung schränken wir uns daher im folgenden auf Gebiete ein, wo die Formel (2.17) gültig ist.

Wir benutzen das Konzept der schwachen Ableitung zur Definition der Sobolev-Räume.

Definition 2.15 (Sobolev-Räume) *Für $m \geq 0$ sei $H^m(\Omega)$ die Menge aller Funktionen $u \in L^2(\Omega)$, zu denen eine schwache Ableitung $\mathrm{D}^\alpha u \in L^2(\Omega)$ für alle $|\alpha| \leq m$ existiert. Der Raum $H^m(\Omega)$ besitzt das Skalarprodukt*

$$\langle u, v \rangle_{H^m} = \sum_{|\alpha| \leq m} \langle \mathrm{D}^\alpha u, \mathrm{D}^\alpha v \rangle_{L^2}.$$

Die Menge $H^m(\Omega)$ heißt ein Sobolev-Raum. $\| \cdot \|_{H^m}$ ist eine Sobolev-Norm.

Daher kann $H^m(\Omega)$ als Verallgemeinerung des Funktionenraums $C^m(\Omega)$, welcher kein Hilbert-Raum ist, interpretiert werden. Die Funktionenräume $(H^m(\Omega), \| \cdot \|_{H^m})$ sind Hilbert-Räume, d.h. sie sind vollständig. Desweiteren gilt $H^m(\Omega) \subset L^2(\Omega)$ für $m \geq 1$ und $H^0(\Omega) = L^2(\Omega)$.

Um Dgln. mit homogenen Dirichlet-Randbedingungen (für beschränktes Ω) zu untersuchen definieren wir die Teilmengen $H_0^m(\Omega) \equiv \overline{C_0^\infty(\Omega)}$ als Abschluss von $C_0^\infty(\Omega) \subset H^m(\Omega)$ bezüglich der Norm $\| \cdot \|_{H^m}$. Genauer gilt

$$H_0^m(\Omega) = \left\{ u \in H^m(\Omega) : \exists (v_i)_{i \in \mathbb{N}} \subset C_0^\infty(\Omega) \text{ mit } \lim_{i \to \infty} \|u - v_i\|_{H^m} = 0 \right\}.$$

Es folgt, dass die Unterräume $(H_0^m, \| \cdot \|_{H^m})$ ebenfalls Hilbert-Räume sind.

Desweiteren erhalten wir eine Halbnorm auf $H^m(\Omega)$ durch

$$|u|_{H^m} = \left(\sum_{|\alpha| = m} \|\mathrm{D}^\alpha u\|_{L^2}^2 \right)^{1/2}.$$

Wenn Ω in einem Würfel mit Kantenlänge s liegt, dann gilt die Normäquivalenz

$$|u|_{H^m} \leq \|u\|_{H^m} \leq (1+s)^m |u|_{H^m} \quad \text{für alle } \; u \in H_0^m(\Omega). \tag{2.18}$$

Die zweite Relation hier wird als *Poincaré-Friedrichs-Ungleichung* bezeichnet.

Die obige Konstruktion kann ebenfalls mit den Funktionenräumen $L^p(\Omega)$ statt $L^2(\Omega)$ mit $1 \leq p \leq \infty$ erfolgen. Es entstehen Sobolev-Räume $W_p^m(\Omega)$, welche für $p \neq 2$ keine Hilbert-Räume sind. Spezialfall ist $W_2^m(\Omega) = H^m(\Omega)$.

Symmetrische Operatoren und Bilinearformen

Gegeben sei eine elliptische Dgl. $Lu = f$ in Ω. Wir setzen im folgenden o.E.d.A. homogene Dirichlet-Randbedingungen auf $\partial\Omega$ voraus. Denn ist $u_0 : \overline{\Omega} \to \mathbb{R}$ hinreichend glatt und $u = u_0$ auf $\partial\Omega$, dann erfüllt die Funktion $w = u - u_0$ die Dgl. $Lw = \tilde{f}$ in Ω mit $\tilde{f} = f - Lu_0$ und $w = 0$ auf $\partial\Omega$. Wir haben somit auf homogene Randbedingungen transformiert.

Wir betrachten zunächst einen allgemeinen linearen Differentialoperator zweiter Ordnung

$$Lu = -\left(\sum_{i,j=1}^{n} a_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j}\right) + \left(\sum_{j=1}^{n} a_j \frac{\partial u}{\partial x_j}\right) + a_0 u. \tag{2.19}$$

Dabei nehmen wir $a_{ij} \in C^2(\bar{\Omega})$, $a_j \in C^1(\bar{\Omega})$, $a_0 \in C^0(\bar{\Omega})$ an. Der zugehörige *adjungierte Operator* L^* ist durch die Bedingung

$$\langle Lu, v \rangle_{L^2} = \langle u, L^* v \rangle_{L^2}$$

für $u, v \in C^2(\bar{\Omega})$ mit $u = 0, v = 0$ auf $\partial\Omega$ definiert. Es folgt

$$L^* v = -\left(\sum_{i,j=1}^{n} \frac{\partial^2 (a_{ij} v)}{\partial x_i \partial x_j}\right) - \left(\sum_{j=1}^{n} \frac{\partial (a_j v)}{\partial x_j}\right) + a_0 v.$$

Ein *symmetrischer* (auch: *selbst-adjungierter*) Operator erfüllt die Bedingung $L = L^*$. Es kann gezeigt werden, dass ein symmetrischer Operator

genau dann vorliegt, wenn er in der Gestalt

$$Lu = -\left(\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right)\right) + a_0 u \tag{2.20}$$

darstellbar ist. Insbesondere ist jeder Operator (2.19) symmetrisch im Fall konstanter Koeffizienten a_{ij} und $a_1 = \cdots = a_n = 0$. Diese Struktur ist bei der Poisson-Gleichung gegeben. Die Green'sche Formel (2.17) liefert

$$\langle Lu, v\rangle = \left(\sum_{i,j=1}^{n} \int_\Omega a_{ij}\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i}\,\mathrm{d}x\right) + \int_\Omega a_0 uv\,\mathrm{d}x. \tag{2.21}$$

Wir bemerken, dass die rechte Seite symmetisch bezüglich u und v ist. Daher gilt $\langle Lu, v\rangle = \langle u, Lv\rangle$. Desweiteren treten nur Ableitungen erster Ordnung in (2.21) auf.

Definition 2.16 *Sei H ein Hilbert-Raum mit der Norm $\|\cdot\|$. Eine Bilinearform $a : H \times H \to \mathbb{R}$ heißt*

i) *symmetrisch, falls $a(u, v) = a(v, u)$ für alle $u, v \in H$ gilt.*

ii) *stetig, falls eine Konstante $C > 0$ existiert mit*

$$|a(u, v)| \le C \cdot \|u\| \cdot \|v\| \qquad \text{für alle } u, v \in H.$$

iii) *positiv, falls $a(u, u) > 0$ für alle $u \in H\backslash\{0\}$ gilt.*

iv) *koerziv (auch: H-elliptisch), falls eine Konstante $\beta > 0$ existiert mit*

$$a(u, u) \ge \beta \cdot \|u\|^2 \qquad \text{für alle } u \in H.$$

Eine koerzive Bilinearform ist insbesondere positiv. Jede symmetrische, positive Bilinearform a induziert eine Norm

$$\|u\|_a = \sqrt{a(u, u)} \qquad \text{für } u \in H, \tag{2.22}$$

welche als *Energienorm* bezeichnet wird. Die Energiernorm ist äquivalent zur Norm des Hilbert-Raums.

Die Gleichung (2.21) motiviert die Definition einer symmetrischen Bilinearform bezüglich eines gleichmäßigen elliptischen Differentialoperators.

Satz 2.17 *Die Bilinearform* $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ *gegeben durch*

$$a(u,v) = \left(\sum_{i,j=1}^{n} \int_\Omega a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} \, \mathrm{d}x \right) + \int_\Omega a_0 uv \, \mathrm{d}x \tag{2.23}$$

mit $a_{ij}, a_0 \in C^0(\bar{\Omega})$, $A = (a_{ij})$ *symmetrisch und positiv definit,* $a_0 \geq 0$ *ist stetig und koerziv, vorausgesetzt der zugehörige Differentialoperator (2.20) ist gleichmäßig elliptisch und* Ω *ist beschränkt.*

Beweis:

Wir definieren $c = \sup\{|a_{ij}(x)| : x \in \Omega,\ 1 \leq i,j \leq n\}$. Es folgt mit der Cauchy-Schwarz'schen Ungleichung

$$\begin{aligned} \left| \sum_{i,j=1}^{n} \int_\Omega a_{ij} u_{x_i} v_{x_j} \, \mathrm{d}x \right| &\leq c \sum_{i,j=1}^{n} \int_\Omega |u_{x_i} v_{x_j}| \, \mathrm{d}x \ \leq \ c \sum_{i,j=1}^{n} \|u_{x_i}\|_{L^2} \cdot \|v_{x_j}\|_{L^2} \\ &= c \Big(\sum_{i=1}^{n} \|u_{x_i}\|_{L^2} \Big) \Big(\sum_{j=1}^{n} \|v_{x_j}\|_{L^2} \Big) \ = \ c\, |u|_{H^1} \, |v|_{H^1}. \end{aligned}$$

Wir setzen $b = \sup\{|a_0(x)| : x \in \Omega\}$. Es folgt

$$\left| \int_\Omega a_0 uv \, \mathrm{d}x \right| \leq b \int_\Omega |uv| \, \mathrm{d}x \leq b \cdot \|u\|_{L^2} \cdot \|v\|_{L^2}.$$

Wir erhalten wegen $\|u\|_{L^2} \leq \|u\|_{H^1}$ und $|u|_{H^1} \leq \|u\|_{H^1}$

$$|a(u,v)| \leq C \cdot \|u\|_{H^1} \cdot \|v\|_{H^1}$$

mit $C = b + c$. Weil der Differentialoperator gleichmäßig elliptisch, siehe (2.6), ist, folgt mit der Monotonie des Integrals

$$\int_\Omega \sum_{i,j=1}^{n} a_{ij} v_{x_i} v_{x_j} \, \mathrm{d}x \geq \alpha \int_\Omega \sum_{i=1}^{n} (v_{x_i})^2 \, \mathrm{d}x$$

für $v \in H_0^1(\Omega)$. Wegen $a_0 \geq 0$ ergibt sich

$$a(v,v) \geq \alpha \sum_{i=1}^{n} \int_\Omega (v_{x_i})^2 \, \mathrm{d}x = \alpha |v|_{H^1}^2$$

für jedes $v \in H_0^1(\Omega)$. Die Äquivalenz (2.18) liefert $a(v,v) \geq \alpha K \|v\|_{H^1}^2$ für $v \in H_0^1(\Omega)$ mit einer Konstanten $K > 0$, die nur von Ω abhängt. Somit ist die Bilinearform a koerziv mit $\beta = \alpha K$. □

Variationsform

Nun betrachten wir eine Dgl. $Lu = f$ mit gleichmäßig elliptischen Differentialoperator L. Sei u eine klassische Lösung und $Lu, f \in L^2(\Omega)$. Es folgt

$$\begin{aligned} Lu - f &= 0 \quad \text{in } \Omega \\ (Lu - f)v &= 0 \quad \text{in } \Omega \\ \langle Lu - f, v\rangle_{L^2} &= 0 \\ \langle Lu, v\rangle_{L^2} - \langle f, v\rangle_{L^2} &= 0 \end{aligned}$$

für jedes $v \in L^2(\Omega)$. We definieren eine lineare Abbildung

$$\ell(v) = \langle f, v\rangle_{L^2} \tag{2.24}$$

für $v \in L^2(\Omega)$ oder einen Unterraum davon. Die Cauchy-Schwarz'sche Ungleichung zeigt

$$|\ell(v)| \leq \|f\|_{L^2}\|v\|_{L^2} \qquad \text{für } v \in L^2(\Omega). \tag{2.25}$$

Dadurch ist ℓ beschränkt (d.h. stetig) auf $L^2(\Omega)$ und für die Operatornorm gilt $\|\ell\| \leq \|f\|_{L^2}$. Daher ist ℓ auch stetig auf $H^1(\Omega)$.

Sei $\ell \in V'$ mit dem Dualraum V'. Somit ist $\ell : V \to \mathbb{R}$ eine beliebige beschränkte lineare Abbildung. Wir verwenden die Notation $\langle \ell, v\rangle = \ell(v)$, welche sich auf die Bilinearform $\langle \cdot, \cdot \rangle : V' \times V \to \mathbb{R}$ bezieht.

Die rechte Seite f der Dgl. liefert eine lineare Abbildung (2.24), während die linke Seite Lu einer Bilinearform (2.23) entspricht. Dies führt auf das Konzept einer schwachen Lösung.

Definition 2.18 (schwache Lösung)
Eine Funktion $u \in H_0^1(\Omega)$ heißt schwache Lösung des elliptischen Problems

$$\begin{aligned} Lu &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{auf } \partial\Omega, \end{aligned}$$

falls mit der zugehörigen Bilinearform (2.23) und der Linearform (2.24) die Bedingung

$$a(u,v) = \langle \ell, v \rangle \qquad \text{für alle } v \in H_0^1(\Omega) \tag{2.26}$$

gilt.

Nun zeigen wir, dass eine klassische Lösung auch eine schwache Lösung des Problems darstellt. Dabei werden wieder gewisse Annahmen für die Koeffizientenfunktionen gemacht.

Satz 2.19 *Sei u eine klassische Lösung des elliptischen Problems*

$$\begin{aligned} -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right) + a_0 u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{auf } \partial\Omega \end{aligned}$$

mit $a_{ij} \in C^1(\Omega) \cap C^0(\bar{\Omega})$ und $a_0, f \in C^0(\bar{\Omega})$ für beliebiges beschränktes Gebiet Ω. Dann stellt u auch eine schwache Lösung des Problems dar, vorausgesetzt alle Ableitungen erster Ordnung befinden sich in $L^2(\Omega)$.

Beweis:

Wir zeigen zuerst $u \in H_0^1(\Omega)$. Da u klassische Lösung ist, gilt $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$ laut Definition 2.12. Insbesondere existieren die ersten Ableitungen, die nach Annahme in $L^2(\Omega)$ sind. Weiter ist $u \in L^2(\Omega)$, da $u \in C^0(\bar{\Omega})$ und Ω beschränkt. Dies zeigt $u \in H^1(\Omega)$. Wegen $u \in H^1(\Omega) \cap C^0(\bar{\Omega})$ und $u = 0$ auf $\partial\Omega$ ist auch $u \in H_0^1(\Omega)$. (Näheres später mit dem Spuroperator in (2.30)).

Es verbleibt (2.26) zu zeigen. Sei zunächst $v \in C_0^\infty(\Omega)$. Wir wenden Lemma 2.14 an mit den beiden Funktionen $v \in C^1(\Omega)$ und $a_{ij}u_{x_j} \in C^1(\Omega)$:

$$\int_\Omega v \frac{\partial}{\partial x_i}\Big(a_{ij}\frac{\partial u}{\partial x_j}\Big)\,\mathrm{d}x = -\int_\Omega a_{ij}\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i}\,\mathrm{d}x.$$

Wir benutzen jeweils die Bilinearform und die Linearform

$$a(u,v) = \int_\Omega \sum_{i,j=1}^n a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} + a_0 uv \, \mathrm{d}x, \qquad \langle \ell, v \rangle = \int_\Omega fv \, \mathrm{d}x.$$

Es folgt

$$\begin{aligned} a(u,v) &= \int_\Omega \Big(- \sum_{i,j=1}^n \frac{\partial}{\partial x_i} \Big(a_{ij} \frac{\partial u}{\partial x_j} \Big) + a_0 u \Big) v \, \mathrm{d}x \\ &= \int_\Omega (Lu)\, v \, \mathrm{d}x \;=\; \int_\Omega fv \, \mathrm{d}x \;=\; \langle \ell, v \rangle \end{aligned}$$

wegen $Lu = f$ auf Ω. Sowohl die Bilinearform a als auch die Linearform ℓ sind stetig auf $H_0^1(\Omega)$, vergleiche den Beweis von Satz 2.17 und (2.25), wobei $a_{ij}, a_0, f \in C^0(\bar{\Omega})$ verwendet werden. Da $C_0^\infty(\Omega) \subseteq H_0^1(\Omega)$ dicht liegt, folgt auch $a(u,v) = \langle \ell, v \rangle$ für alle $v \in H_0^1(\Omega)$. □

Wir zeigen eine wichtige Äquivalenz für unser Problem.

Satz 2.20 *Sei V ein Vektorraum und $a : V \times V \to \mathbb{R}$ eine symmetrische, positive Bilinearform sowie $\ell : V \to \mathbb{R}$ eine lineare Abbildung. Die Funktion*

$$J(v) = \tfrac{1}{2} a(v,v) - \langle \ell, v \rangle \tag{2.27}$$

besitzt ein Minimum in V bei u genau dann, wenn

$$a(u,v) = \langle \ell, v \rangle \quad \textit{für alle} \;\; v \in V \tag{2.28}$$

gilt. Es existiert höchstens ein Minimum.

Beweis:

Eine positive Bilinearform besitzt die Eigenschaft $a(u,u) > 0$ für alle $u \neq 0$. Mit $u, v \in V$ und $t \in \mathbb{R}$ berechnen wir

$$\begin{aligned} J(u+tv) &= \tfrac{1}{2} a(u+tv, u+tv) - \langle \ell, u+tv \rangle \\ &= J(u) + t \left[a(u,v) - \langle \ell, v \rangle \right] + \tfrac{1}{2} t^2 a(v,v). \end{aligned}$$

Falls $u \in V$ die Bedingung (2.28) erfüllt, dann folgt bei $t = 1$

$$J(u + v) = J(u) + \tfrac{1}{2}a(v, v) > J(u)$$

für $v \in V \backslash \{0\}$. Somit ist u eine eindeutige Minimalstelle zu J.

Sei umgekehrt $u \in V$ ein Minimum der Funktion (2.27). Für jedes $v \in V$ gilt

$$\frac{\mathrm{d}}{\mathrm{d}t} J(u + tv)\bigg|_{t=0} = 0.$$

Wegen

$$\frac{\mathrm{d}}{\mathrm{d}t} J(u + tv) = a(u, v) - \langle \ell, v \rangle + ta(v, v)$$

folgt die Bedingung (2.28). □

Satz 2.20 impliziert die Eindeutigkeit einer schwachen Lösung. Falls eine klassische Lösung zusätzliche Bedingungen (wie z.B. $u \in C^2(\bar{\Omega})$) erfüllt, dann stellt diese Funktion auch die eindeutige schwache Lösung dar.

Wir erhalten eine zusätzliche Charakterisierung für die schwache Lösung durch Satz 2.20: Die schwache Lösung eines Dgl.-Problems stellt eine Lösung des Minimierungsproblems

$$J(v) = \tfrac{1}{2}a(v, v) - \langle \ell, v \rangle \quad \longrightarrow \quad \min. \tag{2.29}$$

dar und umgekehrt. Die Aufgabe (2.29) wird *Variationsform* (des Problems) oder *Variationsproblem* genannt.

Satz 2.21 (Lax-Milgram) *Sei H ein Hilbert-Raum und $V \subseteq H$ eine abgeschlossene konvexe Teilmenge. Ist $a : H \times H \to \mathbb{R}$ eine koerzive Bilinearform und $\ell \in H'$ eine Linearform, dann besitzt das Variationsproblem (2.29) eine eindeutige Lösung in V.*

Beweis:

Die Abbildung J ist beschränkt von unten, denn es gilt

$$J(v) \geq \tfrac{1}{2}\beta\|v\|^2 - \|\ell\| \cdot \|v\| = \tfrac{1}{2\beta}(\beta\|v\| - \|\ell\|)^2 - \tfrac{1}{2\beta}\|\ell\|^2 \geq -\tfrac{1}{2\beta}\|\ell\|^2.$$

Wir setzen $c = \inf\{J(v) : v \in V\}$. Sei $(v_n)_{n\in\mathbb{N}} \subset V$ eine Folge mit der Eigenschaft

$$\lim_{n\to\infty} J(v_n) = c.$$

Es folgt

$$\begin{aligned}\beta\|v_n - v_m\|^2 &\le a(v_n - v_m, v_n - v_m)\\ &= 2a(v_n, v_n) + 2a(v_m, v_m) - a(v_n + v_m, v_n + v_m)\\ &= 4J(v_n) + 4J(v_m) - 8J(\tfrac{1}{2}(v_n + v_m))\\ &\le 4J(v_n) + 4J(v_m) - 8c,\end{aligned}$$

weil $\frac{1}{2}(v_n + v_m) \in V$ gilt durch die Konvexität von V. Die obere Schranke konvergiert gegen null für $n, m \to \infty$. Damit folgt $\|v_n - v_m\| \to 0$ für $n, m \to \infty$, d.h. $(v_n)_{n\in\mathbb{N}}$ ist eine Cauchy-Folge. Da V eine abgeschlossene Menge ist, existiert ein Grenzwert $u \in V$. Es folgt

$$J(u) = J\left(\lim_{n\to\infty} v_n\right) = \lim_{n\to\infty} J(v_n) = \inf\{J(v) : v \in V\}$$

wegen der Stetigkeit von J. Somit liegt ein Minimum in u vor.

Die Eindeutigkeit betreffend seien $u_1, u_2 \in V$ zwei Lösungen des Variationsproblems (2.29). Es gilt $J(u_1) = J(u_2) = c$. Wir verwenden u_1 als v_n und u_2 als v_m in obiger Ungleichung. Es folgt

$$\beta\|u_1 - u_2\|^2 \le 4J(u_1) + 4J(u_2) - 8c = 0.$$

Wegen $\beta > 0$ erhalten wir $\|u_1 - u_2\| = 0$ und somit $u_1 = u_2$. □

Wir wenden Satz 2.21 an im Spezialfall $V = H = H_0^1(\Omega)$, weil $H_0^1(\Omega)$ ein Hilbert-Raum ist. Es folgt, dass das Variationsproblem (2.29) eine eindeutige Lösung $u \in H_0^1(\Omega)$ besitzt. Wegen Satz 2.20 ist die Lösung u des Variationsproblems auch eine schwache Lösung des Dgl.-Problems bezüglich Definition 2.18. Wir erhalten direkt als Folgerung.

Satz 2.22 *Sei L ein gleichmäßig elliptischer, symmetrischer Differentialoperator. Dann besitzt die Dgl. $Lu = f$ mit homogenen Dirichlet-Randbedinungen eine eindeutige schwache Lösung in $H_0^1(\Omega)$.*

Bemerkung: Nicht alle beschränkten Gebiete $\Omega \subset \mathbb{R}^n$ sind zulässig, da die Green'sche Formel gültig sein muss. Jedoch ist die Green'sche Formel bei fast allen Gebieten in der Praxis zutreffend.

Im eindimensionalen Fall ($n = 1$) erhalten wir ein homogenes Randwertproblem zu einer gewöhnlichen Dgl. zweiter Ordnung, nämlich

$$-(p(x)u'(x))' + q(x)u(x) = f(x) \quad \text{für } x \in (a, b)$$

mit $u(a) = u(b) = 0$. Unter der Voraussetzung $p(x) \geq p_0 > 0$ und $q(x) \geq 0$ für $x \in (a, b)$ ist der zugehörige Differentialoperator gleichmäßig elliptisch. Ein Variationsproblem kann wie oben konstruiert werden. Eine detaillierte Herleitung in diesem Spezialfall findet sich z.B. in Stoer/Bulirsch: Numerische Mathematik 2, Springer, 2005 (Abschnitt 7.5).

Von-Neumann-Randwertprobleme

Zur Poisson-Gleichung $-\Delta u = f$ auf Ω lauten die von-Neumann-Randbedingungen $\frac{\partial u}{\partial \nu} = g$ auf $\partial\Omega$. Schwache Lösungen werden jetzt im Funktionenraum $H^1(\Omega)$ betrachtet.

Für eine große Klasse von beschränkten Gebieten Ω existiert eine beschränkte lineare Abbildung

$$\gamma : H^1(\Omega) \to L^2(\partial\Omega), \quad \|\gamma(v)\|_{L^2(\partial\Omega)} \leq C \|v\|_{H^1(\Omega)} \tag{2.30}$$

mit einer Konstanten $C > 0$ und der Eigenschaft $\gamma(v) = v|_{\partial\Omega}$ für alle $v \in C^0(\bar{\Omega}) \cap H^1(\Omega)$. Die lineare Abbildung γ heißt *Spuroperator*. Der Spuroperator erlaubt eine alternative Charakterisierung des Hilbert-Raums $H_0^1(\Omega)$. Wir setzen wieder $H_0^1(\Omega) \equiv \overline{C_0^\infty(\Omega)}$, d.h. der Abschluss der Testfunktionen bezüglich der Sovolev-Norm $\|\cdot\|_{H^1}$. Es kann die Darstellung

$$H_0^1(\Omega) = \{u \in H^1(\Omega) : \gamma(u) = 0\}$$

gezeigt werden. Diese Eigenschaft wurde bereits im Beweis von Satz 2.19 angewendet, weil wegen $u \in C^0(\bar{\Omega}) \cap H^1(\Omega)$ und $u = 0$ auf $\partial\Omega$ nun $u \in H_0^1(\Omega)$ gilt.

Eine Variationsform kann auch im Fall der von-Neumann-Randbedingungen hergeleitet werden. Zur Vereinfachung sei $u \in C^2(\bar{\Omega}), v \in C^1(\bar{\Omega})$. Nun liefert

die Green'sche Formel (2.17) für das Skalarprodukt $\langle \Delta u, v \rangle_{L^2}$

$$\int_\Omega v \sum_{i=1}^n \frac{\partial^2 u}{\partial x_i^2} \, \mathrm{d}x = \sum_{i=1}^n \int_\Omega v \, \frac{\partial^2 u}{\partial x_i^2} \, \mathrm{d}x = \sum_{i=1}^n \int_{\partial\Omega} v \, \frac{\partial u}{\partial x_i} \nu_i \, \mathrm{d}s - \int_\Omega \frac{\partial v}{\partial x_i} \frac{\partial u}{\partial x_i} \, \mathrm{d}x.$$

Der zweite Term bewirkt die Definition einer Bilinearform

$$a(u, v) = \int_\Omega \sum_{i=1}^n \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} \, \mathrm{d}x \tag{2.31}$$

wie im Fall von (homogenen) Dirichlet-Randbedingungen. Der erste Term ist nun ungleich null. Es folgt

$$\sum_{i=1}^n \int_{\partial\Omega} v \, \frac{\partial u}{\partial x_i} \nu_i \, \mathrm{d}s = \int_{\partial\Omega} v \left(\sum_{i=1}^n \frac{\partial u}{\partial x_i} \nu_i \right) \, \mathrm{d}s = \int_{\partial\Omega} v (\underbrace{\nabla u \cdot \nu}_{= \frac{\partial u}{\partial \nu}}) \, \mathrm{d}s = \int_{\partial\Omega} v g \, \mathrm{d}s.$$

Daher lautet die zugehörige lineare Abbildung ℓ hier

$$\langle \ell, v \rangle = \int_\Omega f v \, \mathrm{d}x + \int_{\partial\Omega} g v \, \mathrm{d}s \tag{2.32}$$

unter der Annahme $f \in L^2(\Omega)$ und $g \in L^2(\partial\Omega)$. Nun können wir Funktionen $v \in H^1(\Omega)$ betrachten, wobei der Spuroperator (2.30) verwendet wird. Die lineare Abbildung (2.32) modifiziert sich zu

$$\langle \ell, v \rangle = \int_\Omega f v \, \mathrm{d}x + \int_{\partial\Omega} g \gamma(v) \, \mathrm{d}s$$

für $v \in H^1(\Omega)$. Trotzdem wird die Notation (2.32) üblicherweise in der Literatur verwendet.

Die Linearform (2.32) enthält die Information aus sowohl der rechten Seite f als auch der Randwerte g. Numerische Verfahren können zur Lösung des Variationsproblems oder seiner äquivalenten Probleme konstruiert werden. Es sei daran erinnert, dass ein reines von-Neumann-Randwertproblem keine eindeutige Lösung besitzt (ist u eine Lösung, so ist auch $u + c$ eine Lösung für beliebiges $c \in \mathbb{R}$). Daher ist eine zusätzliche Bedingung erforderlich.

Weitere Einzelheiten finden sich in Braess: Finite Elemente. (Kapitel 3)

2.4 Finite-Elemente-Methoden

Nun sezten wir die Theorie aus dem vorhergehenden Abschnitt ein zur Konstruktion eines numerischen Verfahrens zur Bestimmung schwacher Lösungen.

Ritz-Galerkin-Verfahren

Wir betrachten ein homogenes Dirichlet-Randwertproblem zu einem gleichmäßig elliptischen, symmetrischen Differentialoperator (2.20). Es folgt die Existenz und Eindeutigkeit einer schwachen Lösung. Satz 2.20 enthält die zwei Eigenschaften, welche eine schwache Lösung charakterisieren: Variationsform d.h. Minimierung der Funktion (2.27) und die schwache Formulierung (2.28). Numerische Verfahren können jeweils auf einer der beiden Charakterisierungen aufbauen. Sei jetzt H ein beliebiger unendlichdimensionaler Hilbert-Raum. Üblicherweise werden endlichdimensionale Untervektorräume $S_N \subset H$ (N ist die Dimension von S_N) ausgewählt.

Es gibt drei Typen numerischer Verfahren in diesem Zusammenhang:

- *Galerkin-Verfahren*:
  Die schwache Formulierung (2.28) wird verwendet. Eine Näherung der schwachen Lösung soll in S_N bestimmt werden. Die Bedingung (2.28) wird dabei nur für alle $v \in S_N$ gefordert.

- *Petrov-Galerkin-Verfahren* (auch: *Methode der gewichteten Residuen*):
  Die schwache Formulierung (2.28) wird wieder verwendet. Die Näherung $\tilde{u}$ der schwachen Lösung soll in S_N liegen. Die Bedingung (2.28) wird gefordert für $\tilde{u} \in S_N$ und alle $v \in T_N$ mit einem anderen Untervektorraum T_N gleicher Dimension. Der Spezielfall $S_N = T_N$ liefert das Galerkin-Verfahren.

- *Rayleigh-Ritz-Verfahren* (auch: *Ritz-Verfahren*):
  Das Variationsproblem mit der Funktion J aus (2.27) wird betrachtet. Eine Näherung des Minimums in H wird bestimmt als das Minimum von J nur im Teilraum S_N.

Wir diskutieren zuerst das Galerkin-Verfahren. In einem Unterraum S_N wählen wir eine Basis $\{\phi_1, \ldots, \phi_N\}$. Eine Näherung in diesem Unterraum besitzt die Darstellung

$$u_N(x) = \sum_{j=1}^{N} \alpha_j \phi_j(x) \tag{2.33}$$

mit unbekannten Koeffizienten $\alpha_1, \ldots, \alpha_N \in \mathbb{R}$. Ersetzen wir in der Bedingung (2.28) die exakte Lösung $u \in H$ durch eine Näherung $u_N \in S_N$, dann würde sich ergeben

$$a(u_N, v) = \langle \ell, v \rangle \tag{2.34}$$

für alle $v \in H$. Da aber im allgemeinen $u_N \neq u$ ist, kann diese Bedingung nicht für alle $v \in H$ erfüllt werden. Alternativ fordern wir die Bedingung (2.34) nur für alle $v \in S_N$. Unter Verwendung der Basisvektoren ist diese Bedingung äquivalent zu

$$a(u_N, \phi_i) = \langle \ell, \phi_i \rangle \qquad \text{für } \; i = 1, \ldots, N.$$

Einsetzen von (2.33) erzeugt ein lineares Gleichungssystem

$$\sum_{j=1}^{N} \alpha_j a(\phi_j, \phi_i) = \langle \ell, \phi_i \rangle \qquad \text{für } \; i = 1, \ldots, N$$

für die Unbekannten $\alpha_1, \ldots, \alpha_N$. Die Koeffizientenmatrix dieses linearen Gleichungssystems lautet $A = (a(\phi_j, \phi_i)) \in \mathbb{R}^{N \times N}$. Die Symmetrie der Bilinearform a impliziert die Symmetrie der Matrix A. Die Positivität der Bilinearform a ergibt für $\xi = (\xi_1, \ldots, \xi_N)^\top \neq 0$

$$\xi^\top A \, \xi = \sum_{i,j=1}^{N} a(\phi_j, \phi_i) \xi_j \xi_i = a\Big(\sum_{j=1}^{N} \xi_j \phi_j, \sum_{i=1}^{N} \xi_i \phi_i \Big) > 0.$$

Somit ist die Matrix A positiv definit. Folglich existiert eine eindeutige Lösung des linearen Gleichungssystems, welche eine Näherung (2.33) liefert.

Im Petrov-Galerkin-Verfahren fordern wir die Bedingung (2.34) für $u_N \in S_N$ und alle $v \in T_N$ mit einem anderen Teilraum $T_N \subset H$ gleicher Dimension.

Die Elemente in T_N werden oft als Testfunktionen bezeichnet (gehören jedoch im allgemeinen nicht zur Funktionenmenge C_0^∞). Wir wählen eine Basis $\{\psi_1, \ldots, \psi_N\}$ in T_N. Nun ist die Bedingung (2.34) für alle $v \in T_N$ äquivalent zu

$$a\Big(\sum_{j=1}^{N} \alpha_j \phi_j, \psi_i\Big) = \langle \ell, \psi_i \rangle \quad \text{für } \; i = 1, \ldots, N.$$

Es folgt das lineare Gleichungssystem

$$\sum_{j=1}^{N} \alpha_j a(\phi_j, \psi_i) = \langle \ell, \psi_i \rangle \quad \text{für } \; i = 1, \ldots, N$$

mit der Koeffizientenmatrix $A = (a(\phi_j, \psi_i))$. Diese Matrix ist im allgemeinen nicht symmetrisch. Die Regularität der Matrix hängt von der Wahl der Teilräume ab. Im Spezialfall $S_N = T_N$ und bei Verwendung der gleichen Basis in beiden Teilräumen stimmt dieser Ansatz mit dem Galerkin-Verfahren überein wegen $\phi_i = \psi_i$ für alle i.

Im Rayleigh-Ritz-Verfahren setzen wir eine Näherung (2.33) in die Funktion J aus (2.27) ein. Es folgt ($\alpha = (\alpha_1, \ldots, \alpha_N)^\top$)

$$J\Big(\sum_{j=1}^{N} \alpha_j \phi_j\Big) = \tfrac{1}{2} \sum_{i,j=1}^{N} \alpha_j \alpha_i a(\phi_j, \phi_i) - \sum_{j=1}^{N} \alpha_j \langle \ell, \phi_j \rangle = \tfrac{1}{2} \alpha^\top A \alpha - \alpha^\top b$$

mit der selben Matrix A und rechten Seite b wie im Galerkin-Verfahren. Eine Minimierung von J nur in S_N ergibt die Bedingung

$$\frac{\partial J}{\partial \alpha_k} = 0 \quad \text{für } \; k = 1, \ldots, N.$$

Der Gradient von J als Funktion von α lautet $\nabla J = A\alpha - b$. Es folgt das lineare Gleichungssystem $A\alpha = b$. Daher stimmt diese Methode mit dem Galerkin-Verfahren überein. Unterschiedliche Ansätze können entstehen falls die Bilinearform a nicht symmetrisch oder nicht positiv ist. Bei Problemen, für die das Rayleigh-Ritz-Verfahren und das Galerkin-Verfahren identisch sind, spricht man vom *Ritz-Galerkin-Verfahren*. Die Koeffizientenmatrix A wird auch als *Steifigkeitsmatrix* bezeichnet.

Die Methode der gewichteten Residuen kann auch im Fall glatter Lösungen von Dgln. motiviert werden. Sei $u_N \in S_N \subset C^2(\Omega) \cap C^0(\bar{\Omega})$ eine Näherung zu einer klassischen Lösung. In einem endlichdimensionalen Teilraum S_N wählen wir eine Basis $\phi_1, \ldots, \phi_N$ (aus sogenannten *Ansatzfunktionen*) und betrachten die Näherung (2.33). Es folgt das Residuum $\rho : \Omega \to \mathbb{R}$ mit

$$\rho = Lu_N - f = \Big(\sum_{i=1}^{N} \alpha_i L\phi_i \Big) - f.$$

Wir möchten die Koeffizienten $\alpha_1, \ldots, \alpha_N \in \mathbb{R}$ derart bestimmen, dass das Residuum ρ (in gewisser Weise) klein wird. In der Methode der gewichteten Residuen wird ein Teilraum T_N der Dimension N aus Testfunktionen ausgewählt. Wir fordern, dass das Residuum ρ orthogonal auf dem Teilraum T_N bezüglich des Skalarprodukts von L^2 steht. Dies bedeutet

$$\langle Lu_N - f, v \rangle_{L^2} = 0 \quad \text{für alle } \; v \in T_N.$$

Durch Auswahl einer Basis $\{\psi_1, \ldots, \psi_N\}$ in T_N kann diese Bedingung geschrieben werden als

$$\int_\Omega \rho(x) \cdot \psi_j(x) \, \mathrm{d}x = 0 \quad \text{oder} \quad \langle \rho, \psi_j \rangle_{L^2} = 0 \qquad \text{für } \; j = 1, \ldots, N.$$

Dieser Ausdruck kann als ein gewichtetes Integral des Residuums ρ interpretiert werden, wobei die Funktionen ψ_j die Gewichte darstellen. Es folgt das lineare Gleichungssystem

$$\sum_{i=1}^{N} \alpha_i \langle L\phi_i, \psi_j \rangle_{L^2} = \langle f, \psi_j \rangle_{L^2} \quad \text{für } \; j = 1, \ldots, N$$

mit den unbekannten Koeffizienten als Lösung.

Bemerkung: Ein wesentlicher Vorteil des Ritz-Galerkin-Verfahrens besteht darin, dass die Matrix im linearen Gleichungssystem stets symmetrisch und positiv definit ist. Dies gilt dann auch bei einem elliptischen Problem auf beliebigem Gebiet Ω. Dadurch können iterative Lösungsverfahren effizient eingesetzt werden. In einer Finiten-Differenzen-Methode, siehe Abschnitt 2.2, ist die Matrix im linearen Gleichungssystem symmetrisch und positiv definit im Fall des Einheitsquadrats ($\Omega = (0,1)^2$). Die Matrix wird jedoch bereits unsymmetrisch für andere Gebiete Ω wie den Einheitskreis.

Bezüglich der Stabilität des Ritz-Galerkin-Verfahrens gilt die folgende Aussage.

Satz 2.23 (Stabilität) *Sei H ein Hilbert-Raum, $a : H \times H \to \mathbb{R}$ eine symmetrische, stetige, koerzive Bilinearform und $\ell : H \to \mathbb{R}$ eine stetige Linearform. Dann gilt für die Näherungslösung u_N aus dem Ritz-Galerkin-Verfahren*

$$\|u_N\|_H \leq \tfrac{1}{\beta}\|\ell\| \tag{2.35}$$

unabhängig von der Wahl des Teilraums $S_N \subset H$.

Beweis:

Da ℓ als stetig vorausgesetzt ist, gilt $|\ell(v)| \leq \|\ell\| \cdot \|v\|_H$ für alle $v \in H$ mit der Operatornorm von ℓ. Die Koerzivität liefert

$$0 \leq \beta\|u_N\|_H^2 \leq a(u_N, u_N) = \langle \ell, u_N \rangle \leq \|\ell\| \cdot \|u_N\|_H$$

mit der Konstanten $\beta > 0$. Falls $u_N = 0$, so ist die Ungleichung (2.35) trivial. Falls $u_N \neq 0$, dividieren wir durch $\|u_N\|_H$ und erhalten (2.35). □

Bezüglich der Genauigkeit der Näherung aus dem Ritz-Galerkin-Verfahren gilt der folgende zentrale Satz.

Satz 2.24 (Lemma von Céa) *Sei H ein Hilbert-Raum, $a : H \times H \to \mathbb{R}$ eine symmetrische, stetige, koerzive Bilinearform und $\ell : H \to \mathbb{R}$ eine stetige Linearform. Die Funktion u definiert durch $a(u, v) = \langle \ell, v \rangle$ für alle $v \in H$ und die Näherung u_N aus dem Ritz-Galerkin-Verfahren mit einem endlichdimensionalen Unterraum $S_N \subset H$ erfüllen die Abschätzung*

$$\|u - u_N\|_H \leq \frac{C}{\beta} \inf_{v_N \in S_N} \|u - v_N\|_H. \tag{2.36}$$

Beweis:

Es gilt

$$a(u, v) = \langle \ell, v \rangle \quad \text{für } v \in H, \quad a(u_N, v) = \langle \ell, v \rangle \quad \text{für } v \in S_N \subset H.$$

Subtraktion liefert

$$a(u - u_N, v) = 0 \quad \text{für alle } v \in S_N.$$

Für beliebiges $v_N \in S_N$ erhalten wir

$$\begin{aligned}\beta\|u-u_N\|^2 &\leq a(u-u_N, u-u_N)\\ &= a(u-u_N, u-v_N) + a(u-u_N, v_N-u_N)\\ &= a(u-u_N, u-v_N)\\ &\leq C\cdot\|u-u_N\|\cdot\|u-v_N\|\end{aligned}$$

wegen $v_N - u_N \in S_N$. Für $\|u-u_N\| = 0$ ist die Aussage (2.36) trivial. Anderenfalls liefert eine Division der obigen Ungleichung durch $\|u-u_N\| \neq 0$

$$\|u-u_N\| \leq \frac{C}{\beta}\|u-v_N\|.$$

Da $v_N \in S_N$ beliebig ist, folgt die Ungleichung (2.36). □

Satz 2.24 impliziert bereits die Konvergenz des Ritz-Galerkin-Verfahrens falls gilt

$$\lim_{N\to\infty}\inf_{v_N\in S_N}\|u-v_N\|_H = 0.$$

Daher muss eine Folge $(S_N)_{N\in\mathbb{N}}$ derart gewählt werden, dass der Abstand der Teilräume zur exakten Lösung gegen null geht. Dies ist nun eine Frage der Approximationstheorie, also losgelöst vom ursprünglichen Problem.

Bisher wurde ein abstrakter Hilbert-Raum H betrachtet. Jetzt kehren wir zu einem Randwertproblem einer elliptischen Dgl. auf einem Gebiet Ω zurück. Daher wird der Hilbert-Raum $H_0^1(\Omega)$ verwendet, in dem sich schwache Lösungen befinden. Der endlichdimensionale Teilraum wird nun mit S_h (statt S_N) bezeichnet, wobei $h > 0$ eine Schrittweite bezeichnet, die später noch eingeführt wird. Typischerweise gilt $\dim(S_h) \to \infty$ für $h \to 0$. Eine Näherung in S_h wird als u_h (statt u_N) geschrieben.

Für eine elliptische Dgl. $Lu = f$ auf Ω mit homogenen Dirichlet-Randbedingungen diskutieren wir die Stabilitätseigenschaft aus Satz 2.23. Es gilt $\|\ell\| \leq \|f\|_{L^2}$. Wir betrachten zwei rechte Seiten f_1, f_2 mit den zugehörigen schwachen Lösungen u_1, u_2. Dann ist die Differenz $u_1 - u_2$ eine schwache Lösung zur rechten Seite $f_1 - f_2$. Für die Näherung aus dem Ritz-Galerkin-Verfahren gilt somit

$$\|u_h^1 - u_h^2\|_{H_0^1(\Omega)} \leq \frac{1}{\beta}\|f_1 - f_2\|_{L^2(\Omega)}$$

wegen (2.35) und der Linearität. Diese Abschätzung ist unabhängig von der Wahl des Teilraums S_h, d.h. gleichmäßig bezüglich der (noch einzuführenden) Schrittweite $h > 0$. Somit hängen die Näherungen Lipschitz-stetig von den Eingabedaten ab und die Lipschitz-Konstante ist unabhängig von h.

In einer Finiten-Differenzen-Methode ist die Koeffizientenmatrix im linearen Gleichungssystem typischerweise dünnbesetzt oder sogar eine Bandmatrix. Der Rechenaufwand fällt daher deutlich geringer aus im Vergleich zu einer vollbesetzten Matrix gleicher Dimension. Wir möchten eine dünnbesetzte Matrix oder eine Bandmatrix auch im Ritz-Galerkin-Verfahren erhalten. Wir verwenden Teilräume S_h, welche aus stückweise polynomialen Funktionen bestehen. Jedoch wird die entstehende Steifigkeitsmatrix nur für bestimmte Wahlen der Basisfunktionen dünnbesetzt sein.

Sei $\operatorname{supp}(\phi) = \overline{\{x \in \Omega : \phi(x) \neq 0\}}$ der Träger von ϕ. Die Bilinearform (2.23) besitzt die Eigenschaft

$$a(\phi, \psi) = 0 \quad \text{falls} \;\; \mu(\operatorname{supp}(\phi) \cap \operatorname{supp}(\psi)) = 0$$

mit dem Lebesgue-Maß μ, da die Bilinearform ein Integral über Ω darstellt. Wir werden daher eine Basis konstruieren, wo sich die Träger der Basisfunktionen nur selten überschneiden. Notwendigerweise muss immer noch

$$\bigcup_{j=1}^{N} \operatorname{supp}(\phi_j) = \bar{\Omega}$$

für eine Basis $\{\phi_1, \ldots, \phi_N\}$ gelten. Wir teilen das Gebiet Ω in kleinere Teilbereiche auf, um sowohl die Teilräume S_h zu konstruieren als auch die Basisfunktionen auszuwählen.

Triangulationen

Wir betrachten den zweidimensionalen Fall ($n = 2$). Sei $\Omega \subset \mathbb{R}^2$ ein offenes Polygon. Dadurch kann das Gebiet Ω in Dreiecke aufgeteilt wird.

Definition 2.25 (Triangulation) *Sei $\Omega \subset \mathbb{R}^2$ ein Gebiet mit einem Polygon als Rand. Eine Menge $\mathcal{T} = \{T_1, \ldots, T_Q\}$, wobei jeweils T_j ein nichtleeres abgeschlossenes Dreieck ist, heißt zulässige Triangulation wenn*

(i) $\bar{\Omega} = \bigcup_{j=1}^{Q} T_j$, *und*

(ii) $T_i \cap T_j$ *für* $i \neq j$ *ist entweder leer oder besteht genau aus einer Ecke beider Dreiecke oder einer gesamten Kante beider Dreiecke.*

Aus der Eigenschaft (ii) folgt, dass das Innere der Dreiecke stets paarweise disjunkt ist, d.h. die Dreiecke überlappen sich nicht.

Für ein $T \in \mathcal{T}$ bezeichnen wir den halben Durchmesser des Dreiecks mit

$$h_T = \tfrac{1}{2} \max \{ \|x - y\|_2 : x, y \in T \} .$$

Jedes Dreieck T enthält seinen Innkreis mit Radius ρ_T. Es liegt immer $\rho_T \leq h_T$ vor.

Zu einer Familie $\mathcal{T}_h$ aus Triangulationen mit Schrittweite h und $0 < h < \hat{h}$ (für ein $\hat{h} > 0$) nehmen wir die Bedingung

$$\max\{h_T : T \in \mathcal{T}_h\} \leq h$$

an. Wir sind am Grenzfall $h \to 0$ interessiert.

Definition 2.26 *Eine Familie $\mathcal{T}_h$ aus Triangulationen heißt uniform, wenn es eine Konstante $\kappa > 0$ gibt mit $\rho_T \geq \frac{h}{\kappa}$ für alle $T \in \mathcal{T}_h$. Eine Familie $\mathcal{T}_h$ heißt quasi-uniform, falls $\rho_T \geq \frac{h_T}{\kappa}$ für jedes $T \in \mathcal{T}_h$ gilt.*

Für eine uniforme Triangulationsfamilie ist die Größe der Dreiecke ungefähr gleich für festes h, denn es gilt $\frac{1}{\kappa} h \leq \rho_T \leq h_T \leq h$ für alle $T \in \mathcal{T}_h$. Eine uniforme Familie ist auch quasi-uniform. Bei einer quasi-uniformen Triangulationsfamilie ist garantiert, dass die Winkel in den Dreiecken nicht beliebig klein werden können.

Ist ein beliebiges beschränktes Gebiet $\Omega \subset \mathbb{R}^2$ gegeben, so wird dessen Rand zunächst durch einen Polygonzug approximiert. Dadurch kann eine Triangulation des entstehenden Polygons erfolgen. Für ein Polygon $\Omega \subset \mathbb{R}^2$ könnte auch eine Aufteilung in Vierecke konstruiert werden. Jedoch erlauben Triangulationen eine höhere Flexibilität.

Basisfunktionen

Wir betrachten eine zulässige Triangulation $\mathcal{T}_h$ auf einem offenen Polygon $\Omega \subset \mathbb{R}^2$. Wir definieren endlichdimensionale Funktionenräume S_h, welche aus allen Funktionen $v : \bar{\Omega} \to \mathbb{R}$ bestehen mit den drei Eigenschaften

(i) $v \in C^k(\bar{\Omega})$ für ein $k \geq 0$,

(ii) $v|_{\partial\Omega} = 0$,

(iii) $v|_T$ ist ein Polynom mit Grad (höchstens) $\ell \geq 1$ für jedes $T \in \mathcal{T}_h$.

Somit ist die Menge S_h durch die Wahl der ganzen Zahlen k, ℓ festgelegt, welche unabhängig von der Schrittweite $h > 0$ erfolgt. Also treten stückweise polynomiale Funktionen auf. Wir verwenden den Fall $k = 0$ (global stetige Funktionen) und $\ell = 1$ (stückweise lineare Funktionen). Es gilt

$$v|_T = \alpha_T + \beta_T x + \gamma_T y \qquad \text{für jedes } T \in \mathcal{T}_h$$

mit Koeffizienten $\alpha_T, \beta_T, \gamma_T \in \mathbb{R}$.

Sei $R = \{(x_i, y_i) : i = 1, \ldots, N\}$ die Menge der inneren Knoten, d.h. die Ecken der Dreiecke in Ω. Sei $\partial R = \{(x_i, y_i) : i = N+1, \ldots, N+K\}$ die Menge der Randknoten, d.h. die Ecken der Dreiecke auf $\partial\Omega$. Wir definieren stückweise lineare Basisfunktionen ϕ_i durch

$$\phi_i(x_j, y_j) = \begin{cases} 1 & \text{falls } i = j \\ 0 & \text{falls } i \neq j \end{cases} \tag{2.37}$$

für $i = 1, \ldots, N$ und $j = 1, \ldots, N+K$. Es gilt $\dim(S_h) = N$.

Wir müssen die Bilinearform (2.23) auswerten, welche umgeformt werden kann zu

$$\begin{aligned} a(\phi_i, \phi_j) &= \int_\Omega \sum_{k,\ell=1}^{2} a_{k\ell} \frac{\partial \phi_i}{\partial x_k} \frac{\partial \phi_j}{\partial x_\ell} + a_0 \phi_i \phi_j \, \mathrm{d}x \\ &= \sum_{T \in \mathcal{T}_h} \int_T \sum_{k,\ell=1}^{2} a_{k\ell} \frac{\partial \phi_i}{\partial x_k} \frac{\partial \phi_j}{\partial x_\ell} + a_0 \phi_i \phi_j \, \mathrm{d}x \end{aligned}$$

für $i, j = 1, \ldots, N$. Entsprechend kann die Information aus der rechten Seite ausgewertet werden über

$$\langle \ell, \phi_i \rangle = \int_\Omega f(x)\phi_i(x)\,\mathrm{d}x = \sum_{T \in \mathcal{T}_h} \int_T f(x)\phi_i(x)\,\mathrm{d}x$$

für $i = 1, \ldots, N$.

Im Fall $\Omega \subset \mathbb{R}^n$ geben wir die allgemeine Definition von Finiten Elementen nach Ciarlet an.

Definition 2.27 (Finite Elemente)
Ein Finites Element ist ein Tripel (T, Π, Σ) mit den Eigenschaften:

(i) $T \subset \mathbb{R}^n$ *ist ein Polyhedron (es folgt T beschränkt),*

(ii) $\Pi \subset C^0(T)$ *ist ein Vektorraum endlicher Dimension s,*

(iii) $\Sigma = \{\sigma_1, \ldots, \sigma_s\}$ *ist eine Basis des Dualraums*
$\Pi' = \{f : \Pi \to \mathbb{R}\,,\ f$ *linear*$\}$ *(verallgemeinerte Interpolation).*

Eigenschaft (iii) bedeutet, dass jedes $\pi \in \Pi$ eindeutig durch die reellen Zahlen $\sigma_1(\pi), \ldots, \sigma_s(\pi)$ (Reihenfolge von Bedeutung) bestimmt ist.

Manchmal werden nur die Teilbereiche $T \subset \Omega$ als Finite Elemente bezeichnet. Im zweidimensionalen Fall $\Omega \subset \mathbb{R}^2$ liefert eine Triangulation eine Menge aus Finiten Elementen, wobei T ein Dreieck ist.

Modellproblem

Gegeben sei ein uniformes Gitter im Quadrat $\Omega = (0,1) \times (0,1)$, siehe Abbildung 1. Eine zugehörige Triangulation kann sofort erzeugt werden wie in Abbildung 5 dargestellt. Die Schrittweite h ist hier jedoch nicht die Hälfte des Durchmessers der Dreiecke sondern die Schrittweite des uniformen Gitters. Wir betrachten die Poisson-Gleichung $-\Delta u = f$ mit homogenen Dirichlet-Randbedingungen. Die zugehörige Bilinearform ist in (2.31) gegeben.

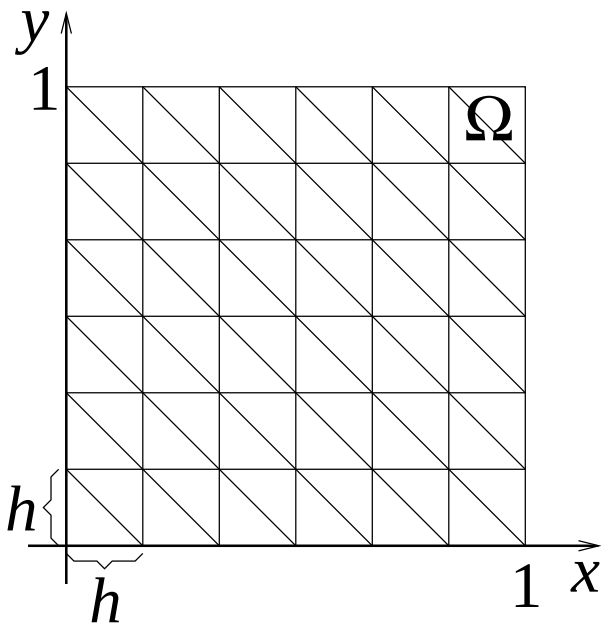


Abbildung 5: Uniformes Gitter mit zugehöriger Triangulation.

Wir verwenden die stückweise linearen Basisfunktionen (2.37). Für ϕ_i sei $Z = (x_i, y_i)$ der zentrale Knoten. Die Nachbarknoten werden bezeichnet wie in Abbildung 6 (links) dargestellt. Die ersten Ableitungen der Basisfunktion sind stückweise konstant, wobei sich die Werte in Abbildung 6 (rechts) finden. Wir berechnen die Steifigkeitsmatrix $A_h = (a(\phi_i, \phi_j))$ aus dem Ritz-Galerkin-Verfahren. Mit (2.31) folgt

$$\begin{aligned} a(\phi_Z, \phi_Z) = \int_\Omega (\nabla \phi_Z)^2 \, \mathrm{d}x\mathrm{d}y &= 2 \int_{\mathrm{I,III,IV}} \left(\frac{\partial \phi_Z}{\partial x}\right)^2 + \left(\frac{\partial \phi_Z}{\partial y}\right)^2 \, \mathrm{d}x\mathrm{d}y \\ &= 2 \int_{\mathrm{I,III}} \left(\frac{\partial \phi_Z}{\partial x}\right)^2 \, \mathrm{d}x\mathrm{d}y + 2 \int_{\mathrm{I,IV}} \left(\frac{\partial \phi_Z}{\partial y}\right)^2 \, \mathrm{d}x\mathrm{d}y \\ &= \frac{2}{h^2} \int_{\mathrm{I,III}} \mathrm{d}x\mathrm{d}y + \frac{2}{h^2} \int_{\mathrm{I,IV}} \mathrm{d}x\mathrm{d}y \;=\; \frac{2}{h^2} \cdot 4 \cdot \frac{h^2}{2} \;=\; 4. \end{aligned}$$

Desweiteren erhalten wir

$$\begin{aligned} a(\phi_Z, \phi_N) &= \int_\Omega (\nabla\phi_Z)\cdot(\nabla\phi_N)\,\mathrm{d}x\mathrm{d}y \;=\; \int_{\mathrm{I,IV}} \frac{\partial\phi_Z}{\partial x}\frac{\partial\phi_N}{\partial x} + \frac{\partial\phi_Z}{\partial y}\frac{\partial\phi_N}{\partial y}\,\mathrm{d}x\mathrm{d}y \\ &= \int_{\mathrm{I,IV}} \left(-\frac{1}{h}\right)\frac{1}{h}\,\mathrm{d}x\mathrm{d}y \;=\; -\frac{1}{h^2}\int_{\mathrm{I,IV}} \mathrm{d}x\mathrm{d}y \\ &= -\frac{1}{h^2}\cdot 2\cdot\frac{h^2}{2} \;=\; -1. \end{aligned}$$

Mit den Symmetrien im Problem ergibt sich auch

$$a(\phi_Z, \phi_N) = a(\phi_Z, \phi_S) = a(\phi_Z, \phi_W) = a(\phi_Z, \phi_E) = -1.$$

Durch Betrachtung der Träger der Basisfunktionen kann leicht bestätigt werden, dass

$$a(\phi_Z, \phi_{NW}) = a(\phi_Z, \phi_{NE}) = a(\phi_Z, \phi_{SW}) = a(\phi_Z, \phi_{SE}) = 0$$

gilt.

Für die rechte Seite benutzen wir die Näherung

$$\langle \ell, \phi_i\rangle = \int_\Omega f(x,y)\phi_i(x,y)\,\mathrm{d}x\mathrm{d}y \approx h^2 f(x_i, y_i),$$

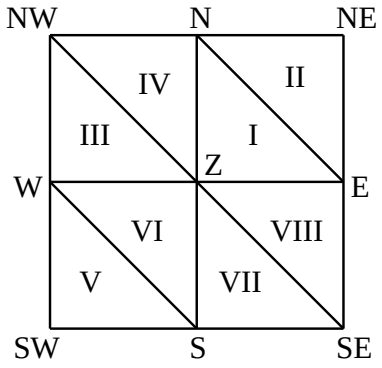
da $f(x_j, y_j)\phi_i(x_j, y_j) = f(x_i, y_i)\delta_{ij}$ für alle j sowie

$$\int_\Omega \phi_i(x,y)\,\mathrm{d}x\mathrm{d}y = h^2$$

gilt. Es folgt der übliche Fünf-Punkte-Stern aus dem Finite-Differenzen-Verfahren, vergleiche (2.9). Allgemein entspricht jede Finite-Differenzen-Mehtode einer gewissen Finite-Elemente-Methode. Jedoch besitzt nicht jede Finite-Elemente-Methode eine äquivalente Finite-Differenzen-Methode. Daher erlauben Techniken mit Finiten Elementen mehr Flexibilität.

Berechnung der Steifigkeitsmatrix

Wir skizzieren die effiziente Berechnung der Steifigkeitsmatrix aus dem Ritz-Galerkin-Verfahren, wobei eine beliebige zulässige Triangulation verwendet



	I	II	III	IV	V	VI	VII	VIII
$\frac{\partial \phi_Z}{\partial x}$	$-h^{-1}$	0	h^{-1}	0	0	h^{-1}	0	$-h^{-1}$
$\frac{\partial \phi_Z}{\partial y}$	$-h^{-1}$	0	0	$-h^{-1}$	0	h^{-1}	h^{-1}	0

Abbildung 6: Triangulation im Modellproblem.

wird, siehe Definition 2.25. Die Struktur der Matrix $A_h = (a(\phi_i, \phi_j)) \in \mathbb{R}^{N \times N}$ legt nahe, eine Schleife über die inneren Knoten (für $i = 1, \ldots, N$) zur Auswertung der Bilinearform zu verwenden (Knoten-orientierte Art). Jedoch erweist sich dieses Vorgehen als ineffizient. Alternativ verläuft die Schleife über die Dreiecke (Element-orientierte Art).

Wir betrachten ein offenes Polygon $\Omega \subset \mathbb{R}^2$ mit einer zulässigen Triangulation $\mathcal{T}_h = \{T_1, \ldots, T_Q\}$. Die Poisson-Gleichung $-\Delta u = f$ mit homogenen Dirichlet-Randbedingungen dient wieder als Modellproblem. Die zugehörige Bilinearform (2.31) ist nun auszuwerten über

$$\begin{aligned} a_{\mu\nu} \;:=\; a(\phi_\mu, \phi_\nu) &= \int_\Omega \frac{\partial \phi_\mu}{\partial x}\frac{\partial \phi_\nu}{\partial x} + \frac{\partial \phi_\mu}{\partial y}\frac{\partial \phi_\nu}{\partial y}\,\mathrm{d}x\mathrm{d}y \\ &= \sum_{q=1}^{Q} \int_{T_q} \frac{\partial \phi_\mu}{\partial x}\frac{\partial \phi_\nu}{\partial x} + \frac{\partial \phi_\mu}{\partial y}\frac{\partial \phi_\nu}{\partial y}\,\mathrm{d}x\mathrm{d}y \end{aligned}$$

für $\mu, \nu = 1, \ldots, N$. Wir setzen

$$a^q_{\mu\nu} = \int_{T_q} \frac{\partial \phi_\mu}{\partial x}\frac{\partial \phi_\nu}{\partial x} + \frac{\partial \phi_\mu}{\partial y}\frac{\partial \phi_\nu}{\partial y}\,\mathrm{d}x\mathrm{d}y. \tag{2.38}$$

Es folgt in der Steifigkeitsmatrix

$$a_{\mu\nu} = \sum_{q=1}^{Q} a^q_{\mu\nu} \quad \text{und} \quad A_h = \sum_{q=1}^{Q} A^q_h \quad \text{mit} \;\; A^q_h := (a^q_{\mu\nu}).$$

Seien i, j, k die Indizes der Ecken im Dreieck T_q. Daher sind nur ϕ_i, ϕ_j, ϕ_k ungleich null in T_q und liefern einen Beitrag im Integral über T_q. Wir erhalten die Struktur

$$A_h^q = \begin{pmatrix} & \vdots & & \vdots & & \vdots & \\ \cdots & a_{ii}^q & \cdots & a_{ij}^q & \cdots & a_{ik}^q & \cdots \\ & \vdots & & \vdots & & \vdots & \\ \cdots & a_{ji}^q & \cdots & a_{jj}^q & \cdots & a_{jk}^q & \cdots \\ & \vdots & & \vdots & & \vdots & \\ \cdots & a_{ki}^q & \cdots & a_{kj}^q & \cdots & a_{kk}^q & \cdots \\ & \vdots & & \vdots & & \vdots & \end{pmatrix} \in \mathbb{R}^{N \times N},$$

wobei (höchstens) neun Einträge ungleich null sind. Diese Matrix kann in kondensierter Form geschrieben werden als

$$\tilde{A}_h^q = \begin{pmatrix} a_{ii}^q & a_{ij}^q & a_{ik}^q \\ a_{ji}^q & a_{jj}^q & a_{jk}^q \\ a_{ki}^q & a_{kj}^q & a_{kk}^q \end{pmatrix} \in \mathbb{R}^{3 \times 3}. \tag{2.39}$$

Zur Berechnung der Integrale (2.38) transformieren wir jedes Dreieck T_q auf das Referenzdreieck $\hat{T} = \{(\xi, \eta) \in \mathbb{R}^2 : 0 \leq \xi, \eta, \ \xi + \eta \leq 1\}$, siehe Abbildung 7. Es folgt die Formel

$$\tilde{A}_h^q = \frac{1}{4|T_q|} E_q E_q^\top \quad \text{mit} \quad E_q = \begin{pmatrix} y_j - y_k & x_k - x_j \\ y_k - y_i & x_i - x_k \\ y_i - y_j & x_j - x_i \end{pmatrix},$$

wobei $|T_q|$ die Fläche des Dreiecks darstellt. Es sei daran erinnert, dass die Indizes i, j, k von q abhängen. Daher ergeben sich nun die Einträge in A_h direkt aus den Koordinaten der Ecken in den Dreiecken.

Falls eine Ecke in T_q, sagen wir mit Index i, nicht zu den inneren Knoten gehört sondern zu den Randknoten, dann ist die zugehörige Basisfunktion ϕ_i nicht definiert. Dadurch werden die erste Zeile und die erste Spalte in (2.39) weggelassen. Entsprechend werden zwei Zeilen und zwei Spalten entfernt, falls zwei Ecken auf dem Rand liegen. Diese Technik steht in Einklang mit den homogenen Dirichlet-Randbedingungen.

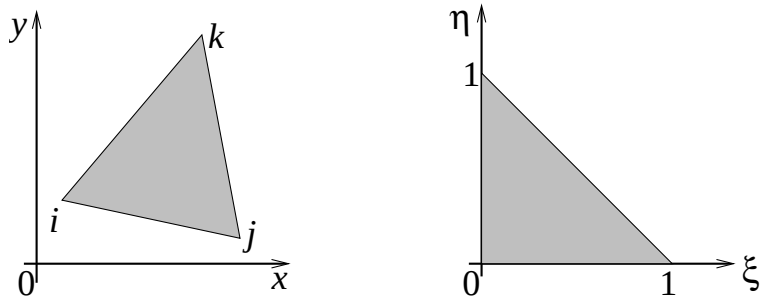


Abbildung 7: Transformation eines beliebigen Dreiecks auf Referenzdreieck.

Die Steifigkeitsmatrix $A_h \in \mathbb{R}^{N\times N}$ enthält insgesamt N^2 Einträge. Weil A_h die Summe der A_h^q für $q = 1, \ldots, Q$ ist, erhalten wir eine grobe Abschätzung der Nicht-Null-Einträge in A_h: höchstens 9Q Einträge sind ungleich null.

Approximationen höherer Ordnung

Bisher haben wir stückweise lineare Polynome auf zugehörigen Triangulationen $\mathcal{T}_h = \{T_1, \ldots, T_Q\}$ eines Polygons Ω verwendet. Stückweise Polynome höheren Grades können ebenfalls konstruiert werden. Sei $\mathcal{P}_\ell$ die Menge aller Polynome mit Grad kleinergleich ℓ, also

$$\mathcal{P}_\ell = \left\{ p(x,y) = \sum_{i,j\geq 0, i+j\leq \ell} c_{ij} x^i y^j \right\}.$$

Es gilt $\dim(\mathcal{P}_\ell) = \frac{1}{2}(\ell+1)(\ell+2)$, was auch der Anzahl der Koeffizienten c_{ij} entspricht.

Auf jedem Dreieck $T \in \mathcal{T}_h$ wählen wir $\frac{(\ell+1)(\ell+2)}{2}$ Punkte $z_s = (x_s, y_s)$ aus zur Durchführung einer Interpolation. Abbildung 8 zeigt die Auswahl der Punkte im Referenzdreieck $\hat{T}$. Es folgt der eindeutige Interpolationsoperator

$$I_T : C^0(T) \to \mathcal{P}_\ell, \quad (I_T u)(z_s) = u(z_s) \ \text{ für } \ s = 1, \ldots, \tfrac{1}{2}(\ell+1)(\ell+2).$$

Wir bilden einen globalen Interpolationsoperator durch

$$I_h : C^0(\bar{\Omega}) \to C^0(\bar{\Omega}), \quad I_h|_T = I_T. \tag{2.40}$$

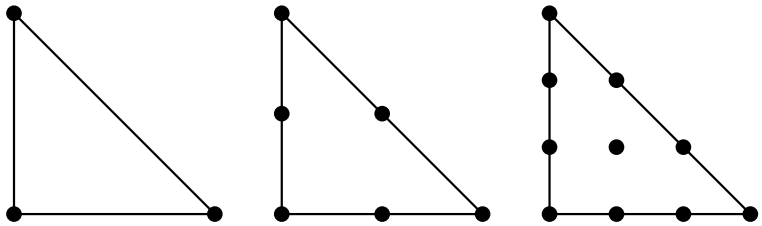


Abbildung 8: Knoten für lineare (links), quadratische (mitte) und kubische (rechts) polynomiale Interpolation im Referenzdreieck.

Somit ist $I_h u$ eine stückweise Funktion aus Polynomen mit Grad höchstens ℓ. Zudem ist die Funktion $I_h u$ global stetig. Die Einschränkung eines Polynoms $I_T u$ auf die Kante eines Dreiecks T liefert ein Polynom in einer Veränderlichen mit Grad höchstens ℓ. Da jede Kante $\ell + 1$ Knoten enthält, stimmen die eindimensionalen Polynome zweier benachbarter Dreiecke überein wegen der Eindeutigkeit der Polynominterpolation.

Wir möchten die Sobolev-Räume $H^m(\Omega)$ anwenden. Das Lemma von Sobolev bei $\Omega \subset \mathbb{R}^n$ liefert $H^m(\Omega) \subset C^k(\Omega)$ für $m > k + \frac{n}{2}$. Mit $n = 2$ und $k = 0$ folgt $H^m(\Omega) \subset C^0(\Omega)$ für $m \geq 2$, d.h. jedes $u \in H^m(\Omega)$ besitzt einen stetigen Repräsentanten. Somit können wir den Interpolationsoperator erweitern zu einem Operator $I_h : H_0^m(\Omega) \to C^0(\bar{\Omega})$ unter der Voraussetzung $m \geq 2$. Wir fordern $(I_h u)(z_s) = 0$ für einen Knoten $z_s \in \partial\Omega$ wegen der homogenen Dirichlet-Randbedingungen.

Wenn der Grad ℓ in den stückweise polynomialen Funktionen hinreichend hoch ist, dann können globale Interpolanten $I_h u \in C^k(\bar{\Omega})$ für $k \geq 1$ definiert werden. Jedoch wird die Konstruktion dann deutlich komplizierter. Die Wahl der Interpolation hängt zusammen mit der Festsetzung der endlichdimensionalen Teilräume S_h im Ritz-Galerkin-Verfahren.

Bemerkung: Auf einer Triangulation können wir bereits Funktionen $\tilde{u}$ erhalten, die global $\tilde{u} \in C^k(\bar{\Omega})$ für ein beliebiges $k \geq 0$ erfüllen, vorausgesetzt der Grad ℓ der lokalen Polynome ist hinreichend hoch. Daher benötigen wir keine komplizierteren Teilbereiche von Ω als Finite Elemente um Approximationen höherer Ordnung zu erzielen.

Konvergenz der Finiten-Elemente-Methode

Wir betrachten eine Finite-Elemente-Methode zu einem allgemeinem Problem $Lu = f$ mit gleichmäßig elliptischem Differentialoperator und homogenen Dirichlet-Randbedinungen in einem offenen Polygon $\Omega \subset \mathbb{R}^2$. Sei eine zulässige Triangulation $\mathcal{T}_h = \{T_1, \ldots, T_Q\}$ gegeben. Die Konvergenz des Verfahrens folgt aus Satz 2.24, wobei wir die Approximationen aus dem Interpolationsschema diskutieren müssen.

Eine Finite-Elemente-Methode auf Grundlage einer Triangulation $\mathcal{T}_h$ verwendet einen Teilraum

$$S_h = \left\{ v \in C^0(\bar{\Omega}) : v|_{\partial\Omega} = 0,\, v|_T \in \mathcal{P}_\ell \text{ für jedes } T \in \mathcal{T}_h \right\}. \tag{2.41}$$

Wir verwenden den globalen Interpolationsoperator

$$I_h : H_0^m(\Omega) \to S_h \subset C^0(\bar{\Omega}), \quad I_h u|_T \in \mathcal{P}_\ell \text{ für jedes } T \in \mathcal{T}_h$$

unter der Annahme $m \geq 2$, vergleiche (2.40).

Wir definieren für $m \geq 0$ die Norm

$$\|v\|_{m,h} = \sqrt{\sum_{T \in \mathcal{T}_h} \|v\|^2_{H^m(T)}}$$

für Funktionen $v : \Omega \to \mathbb{R}$, welche $v|_T \in H^m(T)$ für jedes $T \in \mathcal{T}_h$ erfüllen. Die Funktionen $v_h \in S_h$ aus (2.41) besitzen diese Eigenschaft. Es sei bemerkt, dass $v \in C^k(\bar{\Omega})$ nur $v \in H^{k+1}(\Omega)$ impliziert, sogar für stückweise polynomiale Funktionen v. Es gilt

$$\|v\|_{m,h} = \|v\|_{H^m} \qquad \text{für } v \in H^m(\Omega).$$

Jedoch erfüllt der Teilraum (2.41) nur $S_h \subset H_0^1(\Omega)$.

Der nachfolgende Satz gilt für allgemeine Funktionen, d.h. sie sind nicht notwendigerweise die Lösungen zu Dgln. Eine quasi-uniforme Triangulationsfamilie aus Definition 2.26 wird vorausgesetzt.

Satz 2.28 *Sei $t \geq 2$ und $(\mathcal{T}_h)_{h>0}$ eine Familie aus quasi-uniformen Triangulationen auf Ω. Die zugehörige Interpolation mittels stückweisen Polynomen vom Grad $t-1$ besitzt die Fehlerabschätzung*

$$\|u - I_h u\|_{m,h} \leq c \cdot h^{t-m} \cdot |u|_{H^t} \quad \textit{für } u \in H^t(\Omega)$$

und $0 \leq m \leq t$. Die Konstante $c \geq 0$ hängt von Ω, der Konstanten κ aus der quasi-uniformen Triangulationsfamilie und der ganzen Zahl t ab.

Zum Beweis siehe D. Braess: Finite Elemente.

Da schwache Lösungen einer elliptischen Dgl. in $H_0^1(\Omega)$ liegen, wenden wir nur den Fall $m = 1$ an. Dabei ist $I_h u \in H_0^1(\Omega)$ garantiert. Wir nehmen an, dass die schwache Lösung $u \in H_0^t(\Omega) \subset H_0^1(\Omega)$ mit einem $t \geq 2$ erfüllt. Nun erhalten wir die Konvergenz mittels Satz 2.24. Wegen $I_h u \in S_h$ gilt

$$\inf_{v_h \in S_h} \|u - v_h\|_{H^1} \leq \|u - I_h u\|_{H^1} \leq c \cdot h^{t-1} \cdot |u|_{H^t}.$$

Wir folgern die Konvergenz der Ordnung $p \geq 1$ für die Näherungen $u_h \in S_h$ aus dem Ritz-Galerkin-Verfahren in der Norm von $H^1(\Omega)$, da gilt

$$\|u - u_h\|_{H^1} \leq K_p \cdot h^p \cdot |u|_{H^{p+1}} \quad \text{für } u \in H^{p+1}(\Omega) \tag{2.42}$$

mit der Konstanten $K_p = \frac{Cc}{\beta}$ abhängig von p. Für $t = 2$ verwenden wir stückweise lineare Polynome. Für $t > 2$ erreichen wir höhere Ordnung nur durch Einsatz von Polynomen höheren Grads in jedem Dreieck. Die Näherung u_h bleibt jedoch global nur stetig.

Wegen (2.42) benötigen wir mindestens $u \in H^2(\Omega)$ um Konvergenz der Ordnung $p \geq 1$ zu garantieren. Es kann gezeigt werden, dass eine schwache Lösung auch $u \in H^2(\Omega)$ aufweist, vorausgesetzt für die rechte Seite gilt $f \in L^2(\Omega)$ und das Gebiet Ω besitzt gewisse elementare Eigenschaften.

Satz 2.28 liefert auch eine Abschätzung in der Norm des Hilbert-Raums $L^2(\Omega)$, d.h. im Fall $m = 0$. Wir erwarten eine Konvergenz der Ordnung t in der L^2-Norm. Leider kann Satz 2.24 jetzt nicht angewendet werden, da die Bilinearform nicht stetig bezüglich der L^2-Norm ist. Trotzdem ergibt die Technik von Aubin und Nitsche eine Abschätzung

$$\|u - u_h\|_{L^2} \leq \tilde{K}_p \cdot h^{p+1} \cdot |u|_{H^{p+1}} \quad \text{für } u \in H^{p+1}(\Omega)$$

mit der Konstanten $\tilde{K}_p > 0$ abhängig von p.

Für manche Probleme kann eine gleichmäßige Konvergenz der Gestalt

$$\sup_{x \in \Omega} |u(x) - u_h(x)| \leq c \cdot h \cdot \|f\|_{L^2}$$

mit einer Konstanten $c > 0$ gezeigt werden. Solche Abschätzungen entsprechen dem Banach-Raum $L^\infty(\Omega)$.

Wir haben die Konvergenz in einer Sobolev-Norm sowie in der L^2-Norm diskutiert. Weitere Abschätzungen können bezüglich der Energienorm (2.22) hergeleitet werden.

Kapitel 3

Parabolische Differentialgleichungen

Nun betrachten wir parabolische Differentialgleichungen, welche zeitabhängige Probleme beschreiben. Die Wärmeleitungsgleichung stellt das Musterbeispiel für diesen Typ von partiellen Differentialgleichungen dar. Numerische Verfahren für Anfangs-Randwert-Probleme zu parabolischen Differentialgleichungen werden hergeleitet und untersucht.

3.1 Anfangs-Randwert-Probleme

Zeitabhängige parabolische Differentialgleichungen besitzen häufig die Gestalt

$$\frac{\partial u}{\partial t} + Lu = f(x_1, \ldots, x_n)$$

mit der Lösung $u : D \times [t_0, t_{\text{end}}] \to \mathbb{R}$, wobei $D \subseteq \mathbb{R}^n$ ein Ortsgebiet ist. Der lineare Differentialoperator L enthält partielle Ableitungen zweiter Ordnung im Ortsgebiet (d.h. nicht in der Zeit) und weist häufig einen elliptischen Typ auf. In diesem Kapitel schränken wir uns auf eine Raumdimension ($n = 1$) ein.

Die Wärmeleitungsgleichung lautet

$$\frac{\partial v}{\partial t} = \lambda(x) \frac{\partial^2 v}{\partial x^2} \tag{3.1}$$

mit Koeffizientenfunktion $\lambda : D \to \mathbb{R}$ ($D \subseteq \mathbb{R}$) und $\lambda(x) > 0$ für alle x.

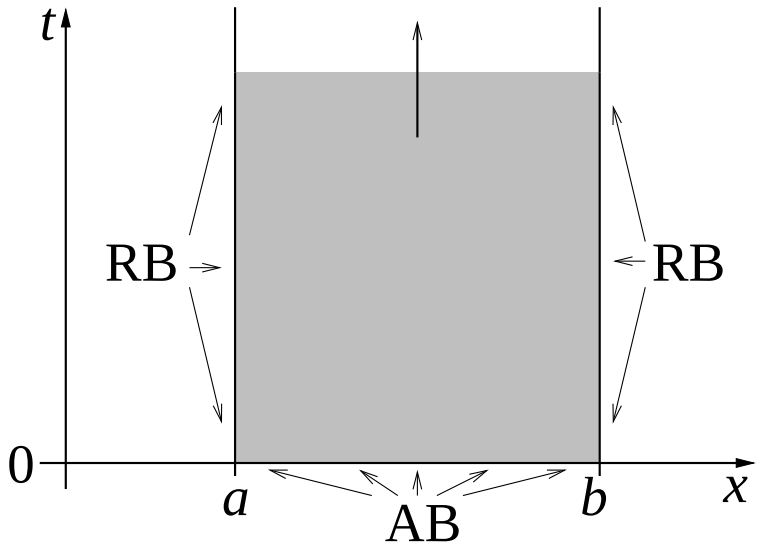


Abbildung 9: Anfangs-Randwert-Problem.

O.E.d.A. sei $\lambda(x) \equiv 1$, d.h.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}. \tag{3.2}$$

Für eine Lösung u von (3.2) erhalten wir eine Lösung von (3.1) für konstanten Parameter λ über die Transformation $v(x,t) = u(x, \lambda t)$.

Wir wählen ein endliches Intervall $[a,b]$ im Ort ($a < b$). Randbedingungen (RB) werden bei $x = a$ und $x = b$ festgelegt. Anfangsbedingungen (AB) werden in der Form

$$u(x,t_0) = u_0(x) \quad \text{für} \;\; x \in [a,b] \tag{3.3}$$

gestellt, wobei $u_0 : [a,b] \to \mathbb{R}$ eine vorgegebene Funktion ist. O.E.d.A. setzen wir $t_0 = 0$. Abbildung 9 enthält eine Skizze des Anfangs-Randwert-Problems.

Wir unterscheiden drei Typen bei Randwertproblemen:

(i) Randbedingungen vom *Dirichlet-Typ* lauten

$$u(a,t) = \alpha(t), \quad u(b,t) = \beta(t) \quad \text{für alle} \;\; t \geq 0 \tag{3.4}$$

mit vorgegebenen Funktionen $\alpha, \beta : [0,\infty) \to \mathbb{R}$.

(ii) Randbedingungen vom *von-Neumann-Typ* fordern

$$\left.\frac{\partial u}{\partial x}\right|_{x=a} = \alpha(t), \quad \left.\frac{\partial u}{\partial x}\right|_{x=b} = \beta(t) \quad \text{für alle } t \geq 0 \tag{3.5}$$

mit vorgegebenen Funktionen $\alpha, \beta : [0, \infty) \to \mathbb{R}$.

(iii) Randbedingungen vom *Robin-Typ* sind ein gemischtes Problem der Typen (i) und (ii), d.h.

$$\gamma_a(t)u(a,t) + \delta_a(t) \left.\frac{\partial u}{\partial x}\right|_{x=a} = \alpha(t), \quad \gamma_b(t)u(b,t) + \delta_b(t) \left.\frac{\partial u}{\partial x}\right|_{x=b} = \beta(t)$$

für alle $t \geq 0$ mit vorgegebenen Funktionen $\alpha, \beta, \gamma_a, \gamma_b, \delta_a, \delta_b$.

Die Anfangswerte (3.3) müssen mit den Randbedingungen kompatibel sein. Beispielsweise ist $u_0(a) = \alpha(0)$ und $u_0(b) = \beta(0)$ erforderlich im Fall von Dirichlet-Randbedingungen.

Sei u eine Lösung der Wärmeleitungsgleichung (3.2) mit homogenen Dirichlet-Randbedingungen ($\alpha, \beta \equiv 0$). Die Funktion

$$\hat{u}(x) = \tfrac{b-x}{b-a}\,\alpha + \tfrac{x-a}{b-a}\,\beta$$

erfüllt die inhomogenen Randbedingungen (3.4) für Konstanten $\alpha, \beta \neq 0$. Es folgt, dass $v := u + \hat{u}$ eine Lösung von (3.2) ist, welche die inhomogenen Dirichlet-Randbedingungen erfüllt. O.E.d.A. betrachten wir daher homogene Dirichtlet-Randbedingungen.

Wir lösen die Wärmeleitungsgleichung (3.2) mit homogenen Dirichlet-Randbedingungen analytisch für $a = 0, b = 1$. Ein Separationsansatz führt auf

$$u(x,t) = \phi(t)\psi(x).$$

Einsetzen dieser Gleichung in (3.2) ergibt

$$\phi'(t)\psi(x) = \phi(t)\psi''(x) \quad \Leftrightarrow \quad \frac{\phi'(t)}{\phi(t)} = \frac{\psi''(x)}{\psi(x)} = \kappa.$$

Dabei stellt $\kappa \in \mathbb{R}$ die Separationskonstante dar.

Zu den beiden gewöhnlichen Differentialgleichungen

$$\phi'(t) = \kappa\phi(t), \qquad \psi''(x) = \kappa\psi(x)$$

folgen die allgemeinen Lösungen

$$\phi(t) = C\mathrm{e}^{\kappa t}, \qquad \psi(x) = A\mathrm{e}^{\sqrt{\kappa}x} + B\mathrm{e}^{-\sqrt{\kappa}x}.$$

mit Konstanten $C \in \mathbb{R}$, $A, B \in \mathbb{C}$. Wir erhalten die allgemeine Lösung

$$u(x,t) = \mathrm{e}^{\kappa t}\left[\tilde{A}\mathrm{e}^{\sqrt{\kappa}x} + \tilde{B}\mathrm{e}^{-\sqrt{\kappa}x}\right]$$

mit beliebigen Konstanten $\tilde{A}, \tilde{B} \in \mathbb{C}$. Die homogenen Randbedingungen sind genau dann erfüllt, wenn

$$\kappa = -k^2\pi^2 \quad \text{für} \;\; k = 1, 2, 3, \ldots \; .$$

Es folgt eine Familie aus Lösungen

$$v_k(x,t) = \tilde{A}_k\mathrm{e}^{-k^2\pi^2 t}\sin(k\pi x)$$

für $k \in \mathbb{N}$ mit neuen Konstanten $\tilde{A}_k \in \mathbb{R}$. Wir verwenden diese Lösungen zur Konstruktion einer einzelnen Lösung, welche die vorgegebenen Anfangsbedingungen (3.3) erfüllt. Sei $u_0 \in C^0([0,1]) \cap C^1((0,1))$. Es gilt $u_0(0) = u_0(1) = 0$ wegen der homogenen Randbedingungen. Wir können die Funktion u_0 zu einer ungeraden Funktion $\hat{u} : [-1,1] \to \mathbb{R}$ erweitern durch die Festsetzung $\hat{u}(x) = u_0(x)$ für $x \geq 0$ und $\hat{u}(x) = -u_0(-x)$ für $x < 0$. Eine periodische Fortsetzung auf ganz $\mathbb{R}$ ist dann stetig und stückweise stetig differenzierbar. Unter diesen Annahmen konvergiert die Fourier-Reihe

$$u_0(x) = \sum_{k=1}^{\infty} a_k \sin(k\pi x)$$

gleichmäßig. Es folgt $\tilde{A}_k = a_k$. Somit lautet die Lösung des Anfangs-Randwert-Problems

$$u(x,t) = \sum_{k=1}^{\infty} a_k\mathrm{e}^{-k^2\pi^2 t}\sin(k\pi x). \tag{3.6}$$

Um die Formel (3.6) auszuwerten muss die Reihe abgeschnitten werden und die enthaltenen Fourier-Koeffizienten a_k sind numerisch zu berechnen.

Die Formel (3.6) charakterisiert die Kondition unseres Anfangs-Randwert-Problems. Seien $u_0, \tilde{u}_0$ zwei Anfangsbedingungen mit zugehörigen Fourier-Koeffizienten a_k bzw. $\tilde{a}_k$. Unter den gemachten Annahmen gilt auch $u_0, \tilde{u}_0 \in L^2((0,1))$. Die entstehenden Lösungen besitzen als Differenz

$$u(x,t) - \tilde{u}(x,t) = \sum_{k=1}^{\infty} (a_k - \tilde{a}_k) \mathrm{e}^{-k^2\pi^2 t} \sin(k\pi x).$$

Damit erhalten wir

$$|u(x,t) - \tilde{u}(x,t)| \leq \sum_{k=1}^{\infty} |a_k - \tilde{a}_k| \cdot \mathrm{e}^{-k^2\pi^2 t}.$$

Die Cauchy-Schwarzsche Ungleichung zum Hilbert-Raum ℓ^2 zusammen mit der Parseval-Identität liefern

$$\begin{aligned} \sum_{k=1}^{\infty} |a_k - \tilde{a}_k| \cdot \mathrm{e}^{-k^2\pi^2 t} &\leq \sqrt{\sum_{k=1}^{\infty} |a_k - \tilde{a}_k|^2} \sqrt{\sum_{k=1}^{\infty} |\mathrm{e}^{-k^2\pi^2 t}|^2} \\ &= \|u_0 - \tilde{u}_0\|_{L^2((0,1))} \sqrt{\sum_{k=1}^{\infty} \mathrm{e}^{-2k^2\pi^2 t}}. \end{aligned}$$

Dabei haben wir Erweiterungen $\hat{u}, \hat{\tilde{u}}$ zu $u_0, \tilde{u}_0$ verwendet, welche die Periode 2 besitzen, sowie $\|\hat{u} - \hat{\tilde{u}}\|^2_{L^2((-1,1))} = 2\|u_0 - \tilde{u}_0\|^2_{L^2((0,1))}$ wegen der Symmetrie.

Wir wenden die Formel zum Grenzwert einer geometrischen Reihe an

$$\sum_{k=1}^{\infty} \mathrm{e}^{-2k^2\pi^2 t} = \sum_{k=1}^{\infty} \left(\mathrm{e}^{-2\pi^2 t}\right)^{k^2} < \frac{1}{1 - \mathrm{e}^{-2\pi^2 t}} - 1 = \frac{\mathrm{e}^{-2\pi^2 t}}{1 - \mathrm{e}^{-2\pi^2 t}}.$$

Es folgt

$$|u(x,t) - \tilde{u}(x,t)| \leq \|u_0 - \tilde{u}_0\|_{L^2((0,1))} \frac{\mathrm{e}^{-\pi^2 t}}{\sqrt{1 - \mathrm{e}^{-2\pi^2 t}}}$$

für alle $x \in [0,1]$ und $t > 0$. Dadurch werden Unterschiede in den Anfangswerten mit der Zeit exponentiell gedämpft. Die Kondition dieses Anfangs-Randwert-Problems is dadurch exzellent. Umgekehrt sind Anfangs-Randwert-Probleme rückwärts in der Zeit (von $t_0 = 0$ zu einen $t_{\mathrm{end}} < 0$) extrem

schlecht konditioniert, da auch kleinste Unterschiede exponentiell verstärkt werden.

Für die Wärmeleitungsgleichung (3.1) hängt die Kondition des Anfangs-Randwert-Problems von der Konstante $\lambda \in \mathbb{R}\backslash\{0\}$ ab gemäß:

	vorwärt in Zeit	rückwärts in Zeit
$\lambda > 0:$	gut konditioniert	schlecht konditioniert
$\lambda < 0:$	schlecht konditioniert	gut konditioniert

Anfangswertprobleme rückwärts in der Zeit werden auch Endwertprobleme genannt (die Werte bei der frühreren Zeit $t_{\text{end}} < 0$ sind unbekannt, während der Zustand zur Endzeit $t_0 = 0$ vorgegeben ist).

Wir erhalten eine Lösung des Anfangswertproblems der Wärmeleitungsgleichung (3.2) auch im Fall $x \in (-\infty, +\infty)$, wo kein Rand auftritt. Es folgt

$$u(x,t) = \frac{1}{2\sqrt{\pi t}} \int_{-\infty}^{+\infty} \mathrm{e}^{-\frac{(x-y)^2}{4t}} u_0(y)\, \mathrm{d}y. \tag{3.7}$$

Die Integrale existieren z.B. für eine beschränkte messbare Funktion u_0 oder $u_0 \in L^2(\mathbb{R})$. Anderenfalls müssen Integrabilitätsbedingungen gestellt werden. Die Formel (3.7) kann nicht für $t = 0$ ausgewertet werden. Stattdessen sind die Anfangsbedingungen erfüllt im Sinne von

$$\lim_{t \to 0+} u(x,t) = u_0(x) \qquad \text{für jedes } \ x \in \mathbb{R}.$$

Zudem ist diese Konvergenz gleichmäßig auf jeder kompakten Ortsmenge $D \subset \mathbb{R}$.

Sei u_0 stetig, $u_0 \geq 0$ und $u_0 \not\equiv 0$. Selbst falls u_0 einen kompakten Träger besitzt, folgt $u(x,t) > 0$ für alle $x \in \mathbb{R}$ und jedes $t > 0$. Daher findet der Informationstransport mit unendlicher Geschwindigkeit statt. Dies gilt auch bei einem Anfangs-Randwert-Problem mit einem endlichen Ortsbereich $x \in [a,b]$.

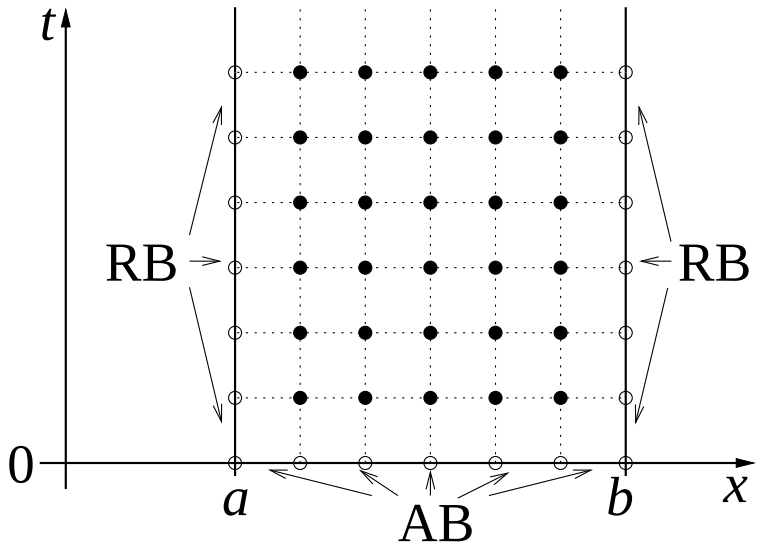


Abbildung 10: Gitter in Finiter-Differenzen-Methode.

3.2 Finite-Differenzen-Methoden

Wir möchten ein Finite-Differenzen-Verfahren verwenden, um ein Anfangs-Randwert-Problem zur Wärmeleitungsgleichung (3.2) aus Abschnitt 3.1 numerisch zu lösen. Ein Gitter wird im (x,t)-Bereich für $x \in [a,b]$ und $t \in [0,T]$ konstruiert, siehe Abbildung 10. O.E.d.A. sei $[a,b] = [0,1]$. Wir setzen die Gitterpunkte fest

$$\begin{aligned} x_j &= jh \quad \text{für} \quad j = 0,1,\ldots,M-1,M, \quad h = \tfrac{1}{M}, \\ t_n &= nk \quad \text{für} \quad n = 0,1,\ldots,N-1,N, \quad k = \tfrac{T}{N}. \end{aligned}$$

Die zugehörigen Schrittweiten sind $h = \Delta x$ und $k = \Delta t$ in Ort bzw. Zeit. Seien $u_j^n = u(x_j,t_n)$ die Werte der exakten Lösung und U_j^n die entsprechenden Näherungen auf dem Gitter. Jetzt ersetzen wir die partiellen Ableitungen in der Wärmeleitungsgleichung (3.2) durch Differenzenformeln.

Klassisches Explizites Verfahren

Wir ersetzen die Zeitableitung durch den gewöhnlichen Differenzenquotienten erster Ordnung (als Vorwärtsdifferenz) und die Ortsableitung durch den

symmetrischen Differenzenquotienten zweiter Ordnung, d.h.

$$u_t(x_j,t_n) = \tfrac{1}{k}(u(x_j,t_{n+1}) - u(x_j,t_n)) - \tfrac{k}{2}u_{tt}(x_j,t_n+\vartheta k)$$

$$\begin{aligned} u_{xx}(x_j,t_n) = {} & \tfrac{1}{h^2}(u(x_{j-1},t_n) - 2u(x_j,t_n) + u(x_{j+1},t_n)) \\ & - \tfrac{h^2}{12}u_{xxxx}(x_j+\theta h,t_n) \end{aligned}$$

mit Zwischenwerten $\vartheta \in (0,1)$, $\theta \in (-1,1)$. Dabei setzen wir u hinreichend glatt voraus. Die Wärmeleitungsgleichung zeigt $u_t(x_j,t_n) = u_{xx}(x_j,t_n)$. Es folgt

$$\tfrac{1}{k}(u_j^{n+1} - u_j^n) + \mathcal{O}(k) = \tfrac{1}{h^2}(u_{j-1}^n - 2u_j^n + u_{j+1}^n) + \mathcal{O}(h^2).$$

Das entstehende Finite-Differenzen-Verfahren lautet

$$\begin{aligned} \tfrac{1}{k}(U_j^{n+1} - U_j^n) &= \tfrac{1}{h^2}(U_{j-1}^n - 2U_j^n + U_{j+1}^n), \\ U_j^{n+1} &= U_j^n + \tfrac{k}{h^2}(U_{j-1}^n - 2U_j^n + U_{j+1}^n). \end{aligned}$$

Wir definieren das Verhältnis $r = \frac{k}{h^2}$. Die Formel des Verfahrens resultiert zu

$$U_j^{n+1} = rU_{j-1}^n + (1-2r)U_j^n + rU_{j+1}^n \tag{3.8}$$

für $j = 1,\ldots,M-1$. Dies ist ein explizites Einschritt-Verfahren. Die Zeitschichten können sukzessive berechnet werden. Die Anfangswerte ergeben sich aus (3.3) über

$$U_j^0 = u_0(x_j) \quad \text{für } \ j = 0,1,\ldots,M.$$

In nachfolgenden Schichten müssen die Randbedingungen einbezogen werden. Dirichlet-Randbedingungen liefern

$$U_0^n = \alpha(t_n), \quad U_M^n = \beta(t_n) \quad \text{für jedes } \ n.$$

Von-Neumann-Randbedingungen werden später diskutiert.

Der lokale Diskretisierungsfehler lautet

$$\tau(h,k) = \tfrac{k}{2}u_{tt}(x_j,t_n+\vartheta k) - \tfrac{h^2}{12}u_{xxxx}(x_j+\theta h,t_n).$$

Wir nehmen an, dass u_{tt} und u_{xxxx} existieren und stetig auf $[0,1]\times[0,T]$ sind. Sei

$$C_1 = \max_{x\in[0,1],t\in[0,T]} |u_{tt}(x,t)|, \quad C_2 = \max_{x\in[0,1],t\in[0,T]} |u_{xxxx}(x,t)|.$$

Es folgt

$$|\tau(h,k)| \leq (k+h^2)(\max\{\tfrac{1}{2}C_1, \tfrac{1}{12}C_2\}) = C(k+h^2) \tag{3.9}$$

gleichmäßig für alle Gitterpunkte (x_j, t_n) in $[0,1] \times [0,T]$. Dadurch ist das Finite-Differenzen-Verfahren konsistent. Für $k, h \to 0$ konvergiert der lokale Diskretisierungsfehler gleichmäßig gegen null. Wir erhalten Konsistenz von Ordnung 1 in der Zeit und Ordnung 2 im Ort.

Klassisches Implizites Verfahren

Wir verwenden die selben Differenzenquotienten im Gitterpunkt (x_j, t_{n+1}) und diskretisieren (als Rückwärtsdifferenz)

$$u_t(x_j, t_{n+1}) = \tfrac{1}{k}(u(x_j, t_{n+1}) - u(x_j, t_n)) + \tfrac{k}{2}u_{tt}(x_j, t_n + \vartheta k)$$

$$u_{xx}(x_j, t_{n+1}) = \tfrac{1}{h^2}(u(x_{j-1}, t_{n+1}) - 2u(x_j, t_{n+1}) + u(x_{j+1}, t_{n+1})) + \tfrac{h^2}{12}u_{xxxx}(x_j + \theta h, t_{n+1})$$

mit Zwischenwerten $\vartheta \in (-1,0)$, $\theta \in (-1,1)$. Die Wärmeleitungsgleichung liefert $u_t(x_j, t_{n+1}) = u_{xx}(x_j, t_{n+1})$. Wir erhalten

$$\tfrac{1}{k}(U_j^{n+1} - U_j^n) = \tfrac{1}{h^2}(U_{j-1}^{n+1} - 2U_j^{n+1} + U_{j+1}^{n+1}).$$

Es folgt die Methode (wieder mit $r = \frac{k}{h^2}$)

$$-rU_{j-1}^{n+1} + (1+2r)U_j^{n+1} - rU_{j+1}^{n+1} = U_j^n \tag{3.10}$$

für $j = 1, \ldots, M-1$. Dies entspricht einem impliziten Einschritt-Verfahren. In jedem Zeitschritt muss ein lineares Gleichungssystem gelöst werden, um die Näherungen sukzessive zu berechnen. Die Koeffizientenmatrix dieses Gleichungssystems lautet

$$B = r \begin{pmatrix} 2+\frac{1}{r} & -1 & & & \\ -1 & 2+\frac{1}{r} & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2+\frac{1}{r} & -1 \\ & & & -1 & 2+\frac{1}{r} \end{pmatrix} \in \mathbb{R}^{(M-1)\times(M-1)}. \tag{3.11}$$

Diese Matrix ist symmetrisch, tridiagonal und strikt diagonaldominant. Eine LR-Zerlegung kann ohne Pivotsuche erfolgen, wobei der Rechenaufwand proportional zu M ist.

Wir fassen die Näherungen in einem Vektor

$$U^n = (U_1^n, U_2^n, \ldots, U_{M-2}^n, U_{M-1}^n)^\top \in \mathbb{R}^{M-1}$$

zusammen. Inhomogene Dirichlet-Randbedingungen müssen in der rechten Seite einbezogen werden durch

$$b^n = \left(rU_0^{n+1}, 0, \ldots, 0, rU_M^{n+1}\right)^\top \in \mathbb{R}^{M-1}.$$

Es folgt ein lineares Gleichungssystem $BU^{n+1} = U^n + b^n$. Bei homogenen Dirichlet-Randbedingungen erhalten wir einfach $BU^{n+1} = U^n$.

Leapfrog-Verfahren

Wir möchten nun Methoden höherer Ordnung erhalten. Wir verwenden den symmetrischen Differenzenquotienten zweiter Ordnung für die (erste) Zeitableitung, d.h.

$$u_t(x_j, t_n) = \tfrac{1}{2k}(u(x_j, t_{n+1}) - u(x_j, t_{n-1})) + \mathcal{O}(k^2)$$

$$u_{xx}(x_j, t_n) = \tfrac{1}{h^2}(u(x_{j-1}, t_n) - 2u(x_j, t_n) + u(x_{j+1}, t_n)) + \mathcal{O}(h^2).$$

Wegen $u_t(x_j, t_n) = u_{xx}(x_j, t_n)$ folgt die Verfahrensvorschrift

$$U_j^{n+1} = U_j^{n-1} + \tfrac{2k}{h^2}(U_{j-1}^n - 2U_j^n + U_{j+1}^n)$$

und mit $r = \frac{k}{h^2}$

$$U_j^{n+1} = U_j^{n-1} + 2r(U_{j-1}^n + U_{j+1}^n) - 4rU_j^n \tag{3.12}$$

für $j = 1, \ldots, M-1$. Wir erhalten ein explizites Zweischritt-Verfahren, genannt das Leapfrog-Verfahren. Diese Methode ist konsistent von Ordnung 2 sowohl in Zeit als auch Ort. Jedoch ist die Leapfrog-Methode instabil für alle $r > 0$ wie wir im nächten Abschnitt sehen werden. Daher ist dieses Verfahren nutzlos in der Praxis.

Crank-Nicolson-Verfahren

Wir erreichen ein Einschritt-Verfahren zweiter Ordnung in sowohl Zeit als auch Ort durch die folgende Konstruktion mit $t_{n+\frac{1}{2}} = t_n + \frac{k}{2}$

$$u_t(x_j, t_{n+\frac{1}{2}}) = \tfrac{1}{k}(u(x_j, t_{n+1}) - u(x_j, t_n)) + \mathcal{O}(k^2)$$

$$\begin{aligned} u_{xx}(x_j, t_{n+\frac{1}{2}}) &= \tfrac{1}{2}(u_{xx}(x_j, t_n) + u_{xx}(x_j, t_{n+1})) + \mathcal{O}(k^2) \\ &= \tfrac{1}{2h^2}(u(x_{j-1}, t_n) - 2u(x_j, t_n) + u(x_{j+1}, t_n)) + \mathcal{O}(h^2) \\ &\quad + \tfrac{1}{2h^2}(u(x_{j-1}, t_{n+1}) - 2u(x_j, t_{n+1}) + u(x_{j+1}, t_{n+1})) + \mathcal{O}(h^2) \\ &\quad + \mathcal{O}(k^2). \end{aligned}$$

Hier erfolgt eine Mittelwertbildung über die symmetrischen Differenzenquotienten im Ort. Die Wärmeleitungsgleichung $u_t(x_j, t_{n+\frac{1}{2}}) = u_{xx}(x_j, t_{n+\frac{1}{2}})$ liefert mit $r = \frac{k}{h^2}$

$$-rU_{j-1}^{n+1} + 2(1+r)U_j^{n+1} - rU_{j+1}^{n+1} = rU_{j-1}^n + 2(1-r)U_j^n + rU_{j+1}^n \qquad (3.13)$$

für $j = 1, \ldots, M-1$. Dies stellt ein implizites Einschritt-Verfahren dar, genannt das Crank-Nicolson-Verfahren. In jedem Zeitschritt ist ein lineares Gleichungssystem mit der Koeffizientenmatrix

$$B = r \begin{pmatrix} -2(1+\frac{1}{r}) & 1 & & & & \\ 1 & -2(1+\frac{1}{r}) & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & 1 & -2(1+\frac{1}{r}) & 1 & \\ & & & 1 & -2(1+\frac{1}{r}) & \end{pmatrix}$$

zu lösen. Diese Matrix ist wieder symmetrisch, tridiagonal und strikt diagonaldominant.

Obwohl die Crank-Nicolson-Methode konsistent von Ordnung 2 in Zeit und Ort ist, liegt der gleiche Rechenaufwand vor wie im klassischen impliziten Verfahren (3.10), welches nur konsistent von Ordnung 1 in der Zeit ist.

Von-Neumann Randbedingungen

Für von-Neumann Randbedingungen (3.5) sind die Werte U_j^n für $j = 0, M$ zunächst unbekannt. Daher fügen wir zwei Gleichungen in jedem Zeitschritt der Finiten-Differenzen-Methode hinzu. Beispielsweise verwenden wir den gewöhnlichen Differenzenquotienten erster Odnung, um die ersten Ableitungen in (3.5) zu diskretisieren. Es folgt

$$\begin{aligned} \alpha(t_n) &= \frac{\partial u}{\partial x}(x_0, t_n) &&= \frac{1}{h}(u(x_1, t_n) - u(x_0, t_n)) + \mathcal{O}(h) \\ \beta(t_n) &= \frac{\partial u}{\partial x}(x_M, t_n) &&= \frac{1}{h}(u(x_M, t_n) - u(x_{M-1}, t_n)) + \mathcal{O}(h) \end{aligned}$$

für jedes n. Wir erhalten die beiden Gleichungen

$$U_0^n = U_1^n - \alpha(t_n)h, \qquad U_M^n = U_{M-1}^n + \beta(t_n)h,$$

die verwendet werden können um die Unbekannten U_0^n, U_M^n in jeder Zeitschicht zu eliminieren. Daraufhin kann die Finite-Differenzen-Methode eingesetzt werden wie bereits besprochen.

Quellterme

Die Finiten-Differenzen-Methoden können direkt verallgemeinert werden für Wärmeleitungsgleichungen inklusive eines Quellterms f, also

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x, t, u).$$

Beispielsweise modifiziert sich das klassische explizite Verfahren (3.8) zu

$$U_j^{n+1} = rU_{j-1}^n + (1 - 2r)U_j^n + rU_{j+1}^n + kf(jh, nk, U_j^n)$$

für $j = 1, \ldots, M-1$. Es ist nur eine Funktionsauswertung von f notwendig um den einzelnen Näherungswert zu berechnen. Für das klassische implizite Verfahren (3.10) erhalten wir

$$-rU_{j-1}^{n+1} + (1 + 2r)U_j^{n+1} - rU_{j+1}^{n+1} - kf(jh, (n+1)k, U_j^{n+1}) = U_j^n$$

für $j = 1, \ldots, M-1$. Falls der Quellterm f nichtlinear u abhängt, dann muss ein nichtlineares Gleichungssystem gelöst werden um die Näherungen in einer Zeitschicht zu bestimmen. Anderenfalls entsteht wieder ein lineares Gleichungssystem.

3.3 Stabilitätsanalyse

Nun untersuchen wir die Stabilität der Finite-Differenzen-Methoden. Wir stellen die Verstärkung (oder Dämpfung) von Fehlern in den Anfangswerten fest.

Direkte Abschätzung

Seien $u_j^n = u(x_j, t_n)$ die Werte der exakten Lösung zur Wärmeleitungsgleichung (3.2) und U_j^n die Näherungen aus einer Finiten-Differenzen-Methode. Wir definieren die globalen Fehler

$$z_j^n = u_j^n - U_j^n.$$

Unter der Annahme, dass die Randwerte exakt gegeben sind, gilt $z_j^n = 0$ für $j = 0, M$. Beim klassischen expliziten Verfahren (3.8) erhalten wir

$$z_j^{n+1} = r z_{j-1}^n + (1-2r) z_j^n + r z_{j+1}^n + \mathcal{O}(k^2 + kh^2).$$

Wir setzen $r \leq \frac{1}{2}$ voraus. Es folgt die Abschätzung

$$|z_j^{n+1}| \leq r|z_{j-1}^n| + (1-2r)|z_j^n| + r|z_{j+1}^n| + C(k^2 + kh^2)$$

mit einer Konstanten $C \geq 0$, siehe (3.9). Wir definieren

$$\|z^n\|_\infty = \max_{j=0,\ldots,M} |z_j^n|,$$

was dem maximalen Fehler in jedem Zeitschritt entspricht. Es folgt

$$\|z^{n+1}\|_\infty \leq r\|z^n\|_\infty + (1-2r)\|z^n\|_\infty + r\|z^n\|_\infty + C(k^2 + kh^2)$$

und daher

$$\|z^{n+1}\|_\infty \leq \|z^n\|_\infty + C(k^2 + kh^2).$$

Wir erhalten sukzessive wegen $nk \leq T$

$$\|z^n\|_\infty \leq \|z^0\|_\infty + nC(k^2 + kh^2) \leq \|z^0\|_\infty + CT(k + h^2)$$

für alle $n = 1, \ldots, N$. Falls $k, h \to 0$ und $\|z^0\|_\infty \to 0$ gilt, dann konvergiert der globale Fehler gegen null unter der Voraussetzung $r \leq \frac{1}{2}$.

Matrix-Stabilitäts-Konzept

Wir untersuchen jetzt das klassische implizite Verfahren (3.10). Seien

$$U^n = (U_1^n, \ldots, U_{M-1}^n)^\top, \quad z^n = (z_1^n, \ldots, z_{M-1}^n)^\top$$

die Näherungen und die zugehörigen globalen Fehler. Wir setzen wieder exakte Randwerte in der Finiten-Differenzen-Methode voraus. Der globale Fehler erfüllt das lineare Gleichungssystem

$$Bz^{n+1} = z^n + k\tau^n$$

mit der Koeffizientenmatrix (3.11) und dem lokalen Diskretisierungsfehler

$$\tau^n = (\tau_1^n, \ldots, \tau_{M-1}^n)^\top.$$

Es gilt $\tau_j^n = \mathcal{O}(k + h^2)$. Wir erhalten sukzessize

$$z^n = (B^{-1})^n z^0 + k \sum_{i=1}^{n} (B^{-1})^i \tau^{n-i}. \tag{3.14}$$

Es gilt $B = I + r\hat{B}$ mit der Tridiagonalmatrix

$$\hat{B} = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{(M-1)\times(M-1)}.$$

Seien λ_i für $i = 1, \ldots, M-1$ die Eigenwerte von $\hat{B}$. Der Kreisesatz von Gerschgorin zeigt $0 \leq \lambda_i \leq 4$. Die inverse Matrix lautet

$$B^{-1} = \left(I + r\hat{B}\right)^{-1}.$$

Seien μ_i die Eigenwerte der Matrix B^{-1}. Es folgt

$$\mu_i = \frac{1}{1 + r\lambda_i}$$

für $i = 1, \ldots, M-1$. Dadurch erhalten wir $0 < \mu_i \leq 1$ für alle i und alle $r > 0$. Weil die Matrix B^{-1} symmetrisch ist, folgt $\|B^{-1}\|_2 = \rho(B^{-1}) \leq 1$ (ρ: Spektralradius).

Die Formel (3.14) führt auf eine Abschätzung in der Euklidischen Norm, wobei wir $\|(B^{-1})^n\|_2 \leq \|B^{-1}\|_2^n$ und (3.9) verwenden:

$$\begin{aligned}
\|z^n\|_2 &\leq \|B^{-1}\|_2^n \cdot \|z^0\|_2 + k\sum_{i=1}^{n} \|B^{-1}\|_2^i \cdot \|\tau^{n-i}\|_2 \\
&\leq \|z^0\|_2 + k\sum_{i=1}^{n} \|\tau^{n-i}\|_2 \;\leq\; \|z^0\|_2 + nMC(k^2 + kh^2) \\
&\leq \|z^0\|_2 + CT(\tfrac{k}{h} + h) \;=\; \|z^0\|_2 + CT(r+1)h.
\end{aligned}$$

Somit ist das Verfahren konvergent für konstantes $r > 0$ und $h \to 0$. Dabei nehmen wir $\|z^0\|_2 = \mathcal{O}(h)$ an. Es sei daran erinnert, dass gilt

$$\|z^0\|_2 \leq M \max_{j=0,\ldots,M} |z_j^0| = \tfrac{1}{h} \max_{j=0,\ldots,M} |z_j^0|.$$

Die obige Herleitung impliziert die Konvergenz mit Ordnung 1 im Ort und Ordnung $\frac{1}{2}$ in der Zeit für konstantes Verhältnis r. Trotzdem können Abschätzungen in anderen Normen erhalten werden, welche die Konvergenz von Ordnung 2 im Ort und Ordnung 1 in der Zeit bestätigen.

Der Fall $C = 0$ (wähle z.B. $u(x,t) \equiv 0$) führt auf $\|z^n\|_2 \leq \|z^0\|_2$ für alle n, was zeigt, dass die Fehler in den Anfangswerten nicht mit der Zeit verstärkt werden. Es folgt die Stabilität des klassischen impliziten Verfahrens, weil diese Abschätzung unabhängig von der Wahl der Schrittweiten k und h ist.

Die Stabilität allein kann auch wie folgt erhalten werden. Seien Anfangswerte U^0, V^0 gegeben. Im klassischen impliziten Verfahren ergeben sich die zugehörigen Näherungen aus $BU^{n+1} = U^n$ und $BV^{n+1} = V^n$ bei homogenen Randbedingungen. Weiter setzen wir $Z^n = U^n - V^n$, wodurch $BZ^{n+1} = Z^n$ folgt. Wir erhalten

$$Z^n = (B^{-1})^n Z^0 \quad \Rightarrow \quad \|Z^n\|_2 \leq \|(B^{-1})^n\|_2 \cdot \|Z^0\|_2 \leq \|Z^0\|_2.$$

Diese Abschätzung ist unabhängig von den verwendeten Schrittweiten, welche auch die Dimension der Vektoren festlegt. Die Ungleichung impliziert die Stabilität des Finite-Differenzen-Verfahrens.

Von-Neumann Stabilität

Sei $\lambda \in \mathbb{R}$ eine beliebige Konstante. Die Funktion

$$u(x,t) = \mathrm{e}^{\alpha t}\mathrm{e}^{\mathrm{i}\lambda x} \tag{3.15}$$

erfüllt die Wärmeleitungsgleichung (3.2) unter der Voraussetzung $\alpha = -\lambda^2$. Insbesondere gilt dann $\alpha \leq 0$ für alle λ, welches die Stabilität von Anfangswertproblemen der Dgl. zeigt. Für Anfangswerte $u_0 \equiv 0$ ist die Lösung von (3.2) einfach $u \equiv 0$. Für gestörte Anfangswerte $\tilde{u}_0(x) = \mathrm{e}^{\mathrm{i}\lambda x}$ konvergiert die Lösung (3.15) für $t \to \infty$ gegen die ursprüngliche Lösung $u \equiv 0$ falls $\alpha < 0$. Wir betrachten jetzt reine Anfangswertprobleme mit $x \in (-\infty, +\infty)$, d.h. es tritt kein Rand auf.

Gegeben sei ein Gitter $(x_j, t_n) = (jh, nk)$ mit $j \in \mathbb{Z}$, $n \in \mathbb{N}_0$. Für die Näherungen aus einer Finite-Differenzen-Methode machen wir den Ansatz

$$U_j^n = \mathrm{e}^{\alpha t_n}\mathrm{e}^{\mathrm{i}\lambda x_j} = \mathrm{e}^{\alpha nk}\mathrm{e}^{\mathrm{i}\lambda jh} \tag{3.16}$$

mit $\lambda \in \mathbb{R}$ und $\alpha \in \mathbb{C}$. Bei $t_0 = 0$ stellen die Anfangswerte

$$U_j^0 = \mathrm{e}^{\mathrm{i}\lambda jh}$$

eine harmonische Schwingung dar, wobei die Frequenz durch die Konstante $\lambda \in \mathbb{R}$ festgelegt wird. Wir interpretieren diese Anfangswerte wieder als eine Störung der Anfangswerte $u_0(x) \equiv 0$.

Für gegebenes $\lambda \in \mathbb{R}$ bestimmt sich das zugehörige $\alpha \in \mathbb{C}$ in (3.16) aus der Finiten-Differenzen-Methode. Wir unterscheiden die Fälle:

- $\mathrm{Re}(\alpha) > 0$ ($\Leftrightarrow |\mathrm{e}^{\alpha k}| > 1$): Die Anfangsstörung U_j^0 wird mit zunehmender Zeit $t > 0$ verstärkt.
- $\mathrm{Re}(\alpha) < 0$ ($\Leftrightarrow |\mathrm{e}^{\alpha k}| < 1$): Die Anfangsstörung U_j^0 wird mit zunehmender Zeit $t > 0$ gedämpft.
- $\mathrm{Re}(\alpha) = 0$ ($\Leftrightarrow |\mathrm{e}^{\alpha k}| = 1$): Die Größenordnung der Anfangsstörung U_j^0 bleibt konstant mit der Zeit.

Das Wachstum der Störungen in Abhängigkeit des Koeffizienten α motiviert die nächste Dfinition.

Definition 3.1 (von-Neumann Stabilität)
Eine Finite-Differenzen-Methode für festes k, h heißt stabil bezüglich des Konzepts nach von-Neumann, wenn $Re(\alpha) \leq 0$ gilt für jedes $\lambda \in \mathbb{R}$. Die Methode ist instabil, wenn $Re(\alpha) > 0$ auftritt für (mindestens) ein $\lambda \in \mathbb{R}$.

Ist die Methode stabil bezüglich des Konzepts nach von-Neumann, dann werden Fehler in den Anfangswerten mit der Zeit nicht verstärkt. Um die von-Neumann Stabilität zu untersuchen reicht es aus, den Term $|e^{\alpha k}|$ zu betrachten.

Für das klassische explizite Verfahren (3.8) führt der Ansatz (3.16) auf

$$e^{\alpha(n+1)k}e^{i\lambda jh} = re^{\alpha nk}e^{i\lambda(j-1)h} + (1-2r)e^{\alpha nk}e^{i\lambda jh} + re^{\alpha nk}e^{i\lambda(j+1)h}.$$

Division durch $e^{\alpha nk}e^{i\lambda jh}$ ergibt

$$\begin{aligned} e^{\alpha k} &= re^{i\lambda(-h)} + 1 - 2r + re^{i\lambda h} \;=\; 1 - 2r + 2r\cos(\lambda h) \\ &= 1 + 2r(\cos(\lambda h) - 1) \;\in\; [1-4r, 1]. \end{aligned}$$

Für $r \leq \frac{1}{2}$ gilt $-1 \leq e^{\alpha k} \leq 1$ für alle $\lambda \in \mathbb{R}$. Dadurch ist das Verfahren stabil für $r \leq \frac{1}{2}$. Für jedes $r > \frac{1}{2}$ existiert eine Konstante $\lambda \in \mathbb{R}$ mit $|e^{\alpha k}| > 1$. Somit ist das Verfahren instabil für $r > \frac{1}{2}$. Desweiteren gilt $0 \leq e^{\alpha k} \leq 1$ für $r \leq \frac{1}{4}$. Das Konzept nach von-Neumann ist in Übereinstimmung mit der direkten Abschätzung beim klassischen expliziten Verfahren unter der Voraussetzung $r \leq \frac{1}{2}$. Zudem wird hier die Instabilität bei $r > \frac{1}{2}$ nachgewiesen, welches aus der direkten Abschätzung nicht folgt.

Beim klassischen impliziten Verfahren (3.10) verwenden wir wieder den Ansatz (3.16) und erhalten

$$-re^{\alpha(n+1)k}e^{i\lambda(j-1)h} + (1+2r)e^{\alpha(n+1)k}e^{i\lambda jh} - re^{\alpha(n+1)k}e^{i\lambda(j+1)h} = e^{\alpha nk}e^{i\lambda jh}.$$

Division durch $e^{\alpha nk}e^{i\lambda jh}$ führt auf

$$-re^{\alpha k}e^{i\lambda(-h)} + (1+2r)e^{\alpha k} - re^{\alpha k}e^{i\lambda h} = e^{\alpha k}\left(1 + 2r(1-\cos(\lambda h))\right) = 1$$

und daher (mit $1-\cos(\gamma) = 2\sin^2(\frac{\gamma}{2})$)

$$\mathrm{e}^{\alpha k} = \frac{1}{1 + 4r\sin^2\left(\frac{\lambda h}{2}\right)} \in [0,1].$$

Es folgt die Stabilität des klassischen impliziten Verfahrens für alle $r > 0$. Dieses Kriterium ist in Übereinstimmung mit dem Matrix-Stabilitäts-Konzept angewendet auf das klassische implizite Verfahren.

Das Leapfrog-Verfahren (3.12) liefert die Gleichung

$$\mathrm{e}^{\alpha(n+1)k}\mathrm{e}^{\mathrm{i}\lambda jh} = \mathrm{e}^{\alpha(n-1)k}\mathrm{e}^{\mathrm{i}\lambda jh} + 2r(\mathrm{e}^{\alpha nk}\mathrm{e}^{\mathrm{i}\lambda(j-1)h} + \mathrm{e}^{\alpha nk}\mathrm{e}^{\mathrm{i}\lambda(j+1)h}) - 4r\mathrm{e}^{\alpha nk}\mathrm{e}^{\mathrm{i}\lambda jh}.$$

Division durch $\mathrm{e}^{\alpha nk}\mathrm{e}^{\mathrm{i}\lambda jh}$ ergibt

$$\mathrm{e}^{\alpha k} = \mathrm{e}^{-\alpha k} + 2r(\mathrm{e}^{\mathrm{i}\lambda(-h)} + \mathrm{e}^{\mathrm{i}\lambda h}) - 4r = \mathrm{e}^{-\alpha k} + 4r(\cos(\lambda h) - 1).$$

Es folgt die quadratische Gleichung (mit $1-\cos(\gamma) = 2\sin^2(\frac{\gamma}{2})$)

$$\left(\mathrm{e}^{\alpha k}\right)^2 + 8r\sin^2\left(\tfrac{\lambda h}{2}\right)\mathrm{e}^{\alpha k} - 1 = 0. \tag{3.17}$$

Sei $\xi = \mathrm{e}^{\alpha k}$ und $b = 8r\sin^2\left(\frac{\lambda h}{2}\right)$. Die Nullstellen aus der quadratischen Gleichung lauten

$$\xi_{1/2} = \tfrac{1}{2}\left[-b \pm \sqrt{b^2+4}\right] \in \mathbb{R}.$$

Dadurch gilt $\xi_2 < 0 < \xi_1$, insbesondere $\xi_1 \neq \xi_2$. Wir erhalten

$$\xi_1 \cdot \xi_2 = \tfrac{1}{4}\left((-b)^2 - \sqrt{b^2+4}^2\right) = -1.$$

Es folgt

$$|\xi_1| = \frac{1}{|\xi_2|}, \quad |\xi_2| = \frac{1}{|\xi_1|}.$$

Falls $|\xi_1| < 1$, dann $|\xi_2| > 1$ und umgekehrt. Der Fall $\xi_1 = 1$, $\xi_2 = -1$ kann nur für $\sin(\frac{\lambda h}{2}) = 0$ auftreten, siehe (3.17). Dadurch liegt $|\mathrm{e}^{\alpha k}| > 1$ bei einer Nullstelle vor. Somit ist das Leapfrog-Verfahren instabil für alle $r > 0$.

Desweiteren kann gezeigt werden, dass das Crank-Nicolson-Verfahren (3.13) stabil ist für alle $r > 0$. Diese Aussage ist in Übereinstimmung mit dem Matrix-Stabilitäts-Konzept angewendet auf das Crank-Nicolson-Verfahren.

Bemerkung: Die Stabilität gemäß des Konzepts nach von-Neumann ist notwendig und hinreichend für die Konvergenz einer konsistenten Finiten-Differenzen-Methode.

Wärmeleitungsgleichung mit Koeffizient

Für die Wärmeleitungsgleichung $v_t = \lambda v_{xx}$ mit einer Konstanten $\lambda > 0$ liefert die lineare Transformation $v(x,t) = u(x, \lambda t)$ die standardisierte Wärmeleitungsgleichung $u_t = u_{xx}$, welche bisher diskutiert wurde. Wir betrachten eine Finite-Differenzen-Methode. Sei $r = \frac{k}{h^2}$. Die Stabilitätsforderung kann eine Restiktion $r \leq c$ mit einer Konstanten $c > 0$ im Fall $u_t = u_{xx}$ bewirken. Dann folgt die Bedingung $r \leq \frac{c}{\lambda}$ im Fall $v_t = \lambda v_{xx}$. Somit ergibt sich eine nachteilhafte Restriktion an die Zeitschrittweite ($k \leq \frac{c}{\lambda} h^2$) bei hohen Konstanten λ.

3.4 Semidiskretisierung

Das Konzept der Semidiskretisierung besteht darin, nur eine partielle Ableitung in der Dgl. durch eine Differenzenformel zu ersetzen. Es folgt ein System aus gew. Dgln. Dementsprechend kann das System gew. Dgln. mit dafür üblichen numerischen Verfahren gelöst werden.

Linienmethode

Wir betrachten die Wärmeleitungsgleichung

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x,t,u) \tag{3.18}$$

inklusive eines Quellterms f. Ein zugehöriges Anfangs-Randwert-Problem sei gegeben mit dem Ortsbereich $x \in [0,1]$ und Dirichlet-Randbedingungen (3.4). Wir verwenden eine Diskretisierung im Ort mit den Gitterpunkten $x_j = jh$ für $j = 0, 1, \ldots, M$ und der Schrittweite $h = \frac{1}{M}$. Im Definitionsbereich werden die Mengen $\{(x_j, t) \in \mathbb{R}^2 : t \geq 0\}$ auch als Linien bezeichnet. Wir bestimmen als Näherungen die zeitabhängigen Funktionen $U_j(t) \approx u(x_j, t)$ für $j = 1, \ldots, M-1$. Abbildung 11 verdeutlicht diese Konstruktion.

Nun ersetzen wir in (3.18) die Ableitung im Ort durch einen symmetrischen

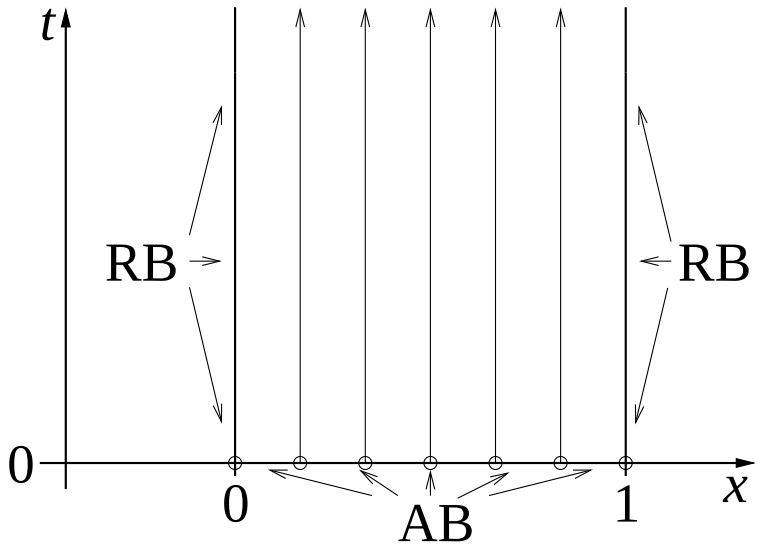


Abbildung 11: Linienmethode.

Differenzenquotienten zweiter Ordnung. Es folgt

$$\frac{\partial u}{\partial t}(x_j,t) = \frac{1}{h^2}\big[u(x_{j-1},t) - 2u(x_j,t) + u(x_{j+1},t)\big] + f(x_j,t,u(x_j,t)) + \mathcal{O}(h^2)$$

für $j = 1,\dots,M-1$. Wir schreiben diese Gleichungen als ein System aus gew. Dgln.

$$U_j'(t) = \frac{1}{h^2}\Big[U_{j-1}(t) - 2U_j(t) + U_{j+1}(t)\Big] + f(x_j,t,U_j(t)) \tag{3.19}$$

für $j = 1,\dots,M-1$. Für eine kompakte Notation setzen wir

$$B = \frac{1}{h^2}\begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{pmatrix} \in \mathbb{R}^{(M-1)\times(M-1)},$$

$$U(t) = \begin{pmatrix} U_1(t) \\ \vdots \\ U_{M-1}(t) \end{pmatrix},\ F(t,U) = \begin{pmatrix} f(x_1,t,U_1) \\ \vdots \\ f(x_{M-1},t,U_{M-1}) \end{pmatrix},\ b(t) = \begin{pmatrix} \alpha(t)/h^2 \\ 0 \\ \vdots \\ 0 \\ \beta(t)/h^2 \end{pmatrix}.$$

Nun besitzt das System aus gew. Dgln. die Form

$$U'(t) = BU(t) + F(t, U(t)) + b(t). \tag{3.20}$$

Die Anfangswerte folgen aus (3.3) über

$$U(0) = (U_1(0), \ldots, U_{M-1}(0))^\top = (u_0(x_1), \ldots, u_0(x_{M-1}))^\top. \tag{3.21}$$

Falls kein Quellterm in (3.18) auftritt, dann folgt $F \equiv 0$ und das Differentialgleichungssystem (3.20) ist linear. Die Eigenwerte μ_ℓ der Matrix B können ausgerechnet werden und es gilt

$$\mu_\ell = -\tfrac{4}{h^2} \sin^2\left(\tfrac{\pi}{2}\ell h\right) < 0 \qquad \text{für } \ell = 1, 2, \ldots, M-1.$$

Der größte und kleinste Eigenwert ist

$$\mu_{\max} \approx -\pi^2 \quad (\ell = 1) \quad \text{und} \quad \mu_{\min} \approx -\frac{4}{h^2} \quad (\ell = M-1).$$

Für kleines h erhalten wir $\mu_{\min} \ll \mu_{\max} < 0$. Dadurch ist das System gew. Dgln. (3.20) steif. Somit sind implizite Verfahren erforderlich um zugehörige Anfangswertprobleme zu lösen.

Nun können Software-Pakete zur numerischen Lösung des Anfangswertproblems (3.20), (3.21) eingesetzt werden. Das explizite Euler-Verfahren und das implizite Euler-Verfahren ergeben gerade das klassische explizite Verfahren (3.8) bzw. das klassische implizite Verfahren (3.10). Ausgereiftere numerische Verfahren können angewendet werden wie Runge-Kutta-Methoden oder Mehrschritt-Verfahren.

Seien $\tilde{U}_j(\tau_i)$ die Näherungen zur exakten Lösung $U_j(t)$ des Anfangswertproblems (3.20), (3.21), welche mit einer Methode für gew. Dgln. der Konvergenzordnung p berechnet werden. Der Fehler kann abgeschätzt werden durch

$$\left|\tilde{U}_j(\tau_i) - u(x_j, \tau_i)\right| \le \left|\tilde{U}_j(\tau_i) - U_j(\tau_i)\right| + \left|U_j(\tau_i) - u(x_j, \tau_i)\right|.$$

Da die Ortsdiskretisierung konsistent von Ordnung 2 ist, erwarten wir als Fehler

$$\left|\tilde{U}_j(\tau_i) - u(x_j, \tau_i)\right| \le C(\Delta t)^p + D(\Delta x)^2 \tag{3.22}$$

mit $\Delta x = h$ und $\tau_{i+1} - \tau_i \leq \Delta t$ für alle i. Der Fehler besteht aus zwei Anteilen: der Fehler aus der Ortsdiskretisierung und der Fehler aus der anschließenden Zeitdiskretisierung. Leider hängt die Konstante C aus der Zeitdiskretisierung vom Differentialgleichungssystem (3.20) ab und damit auch von der Schrittweite h im Ort, d.h. $C = C(h)$. Insbesondere ergibt sich die Dimension $M - 1$ des System (3.20) aus der Schrittweite $h = \frac{1}{M}$. Die Konvergenz kann daher nicht direkt nachgewiesen werden, da üblicherweise die Eigenschaft $C = \mathcal{O}(\frac{1}{h})$ vorliegt. Die beiden Terme in der Abschätzung (3.22) sind nicht unabhängig voneinander.

Von-Neumann-Randbedingungen (3.5) können in der Linienmethode einbezogen werden in gleicher Weise wie bei den Finite-Differenzen-Methoden, siehe Abschnitt 3.2.

Rothe-Methode

Wir betrachten wieder die Wärmeleitungsgleichung (3.18) mit einem Quellterm. Seien Dirichlet-Randbegingungen (3.4) vorgegeben. In der Rothe-Methode wird nun zuerst die Zeitableitung diskretisiert in den Zeitpunkten $t_n = kn$ für $n = 0, 1, \ldots, N$ mit $k = \frac{T}{N}$. Es folgt (mit Rückwärtsdifferenzen)

$$\tfrac{1}{k}\left[u(x, t_{n+1}) - u(x, t_n)\right] = \frac{\partial^2 u}{\partial x^2}(x, t_{n+1}) + f(x, t_{n+1}, u(x, t_{n+1}))$$

für $n = 0, 1, \ldots, N - 1$. Abbildung 12 skizziert diese Semidiskretisierung.

Wir setzen als Näherungen $z_n(x) \approx u(x, t_n)$ für $n = 1, \ldots, N$. Es ergibt sich ein Zwei-Punkt-Randwertproblem einer gew. Dgl. zweiter Ordnung

$$\begin{aligned} z''_{n+1}(x) &= \tfrac{1}{k}\left[z_{n+1}(x) - z_n(x)\right] - f(x, t_{n+1}, z_{n+1}(x)), \\ z_{n+1}(0) &= \alpha(t_{n+1}), \qquad z_{n+1}(1) = \beta(t_{n+1}). \end{aligned} \tag{3.23}$$

Die Anfangsbedingungen (3.3) liefern $z_0(x) = u_0(x)$ für $x \in [0, 1]$. Dadurch können wir die unbekannten Funktionen z_n sukzessive berechnen. Dabei werden numerische Methoden für Randwertprobleme mit gew. Dgln. angewendet. Oft wird dazu das äquivalente Differentialgleichungssystem erster

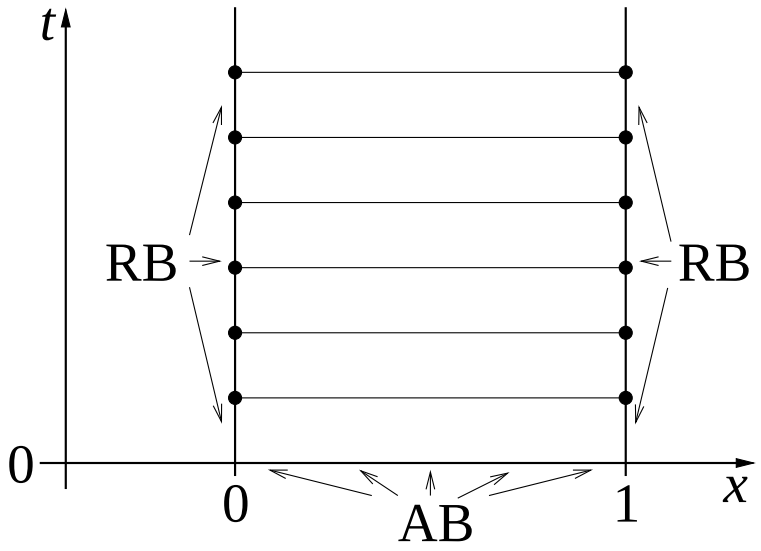


Abbildung 12: Rothe-Methode.

Ordnung zu (3.23) verwendet. Mit $v_j = z_j$ und $w_j = z_j'$ lautet das Zwei-Punkt-Randwertproblem dann

$$\begin{aligned} v_{n+1}'(x) &= w_{n+1}(x), \\ w_{n+1}'(x) &= \tfrac{1}{k}\left[v_{n+1}(x) - v_n(x)\right] - f(x, t_{n+1}, v_{n+1}(x)), \\ v_{n+1}(0) &= \alpha(t_{n+1}), \\ v_{n+1}(1) &= \beta(t_{n+1}). \end{aligned}$$

Üblicherweise erzeugt ein numerisches Verfahren für gew. Dgln. hier Näherungen $z_n(x_j)$ in den Gitterpunkten $0 < x_1 < \cdots < x_R < 1$. Daher muss ein Interpolationsverfahren eingesetzt werden, um die rechte Seite der Dgln. (3.23) für beliebiges $x \in [0, 1]$ auszuwerten.

Seien $\tilde{z}_n(x_j)$ die Näherungen aus dem Lösungsverfahren für die gew. Dgln. Wir erhalten wieder eine Abschätzung des Fehler

$$|\tilde{z}_n(x_j) - u(x_j, t_n)| \leq |\tilde{z}_n(x_j) - z_n(x_j)| + |z_n(x_j) - u(x_j, t_n)|.$$

Falls die numerische Methode zur Lösung der gew. Dgln. konvergent von Ordnung q ist, dann erwarten wir den Fehler

$$|\tilde{z}_n(x_j) - u(x_j, t_n)| \leq C\Delta t + D(\Delta x)^q \tag{3.24}$$

mit $x_{j+1} - x_j \leq \Delta x$ für alle j. Jetzt sind die beiden Terme in der Abschätzung (3.24) unabhängig voneinander, da die gew. Dgln. (3.23) qualitativ gleich sind für jedes $k = \Delta t$ (nur die rechte Seite ist geringfügig verschieden). Dadurch gewinnen wir gute Konvergenzeigenschaften in der Rothe-Methode. Desweiteren ist es leicht adaptive Verfahren bezüglich sowohl der Zeitschrittweite Δt als auch der Ortsschrittweite Δx in der Rothe-Methode anzuwenden. Im Gegensatz dazu bedeutet eine Änderung der Ortsschrittweite Δx eine Änderung der Dimension des Differentialgleichungssystems in der Linienmethode.

In der Linienmethode ist ein Anfangswertproblem eines (relativ großen) steifen Systems aus gew. Dgln. zu lösen. In der Rothe-Method ist ein Randwertproblem einer einzelnen gew. Dgl. zweiter Ordnung (oder eines Systems erster Ordnung mit zwei Gleichungen) sukzessive zu lösen. Jedoch ist der Rechenaufwand für ein Randwertproblem deutlich höher als bei einem Anfangswertproblem (etwa 20-mal mehr bei gleicher Dimension).

Mehrdimensionales Ortsgebiet

Wir skizzieren die Anwendung der Linienmethode im Fall der Wärmeleitungsgleichung

$$\frac{\partial u}{\partial t} = \Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \tag{3.25}$$

in zwei Raumdimensionen. Das Ortsgebiet sei $\Omega = (0,1)^2$ mit homogenen Dirichlet-Randbedingungen auf $\partial\Omega$. Wir verwenden eine Diskretisierung im Ort entsprechend der Finite-Differenzen-Methode aus Abschnitt 2.2. Sei $x_i = ih$ und $y_j = jh$ für $i, j = 0, 1, \ldots, M+1$ Mit der Schrittweite $h = \frac{1}{M+1}$. Die Näherungen sind $U_{i,j}(t) \approx u(x_i, y_j, t)$. Es folgt ein System gew. Dgln.

$$U'_{i,j}(t) = \tfrac{1}{h^2} \left[U_{i-1,j}(t) + U_{i+1,j}(t) + U_{i,j-1}(t) + U_{i,j+1}(t) - 4U_{i,j}(t) \right]$$

für $i, j = 1, \ldots, M$. Die homogenen Randbedingungen liefern

$$U_{i,j}(t) = 0 \qquad \text{für } \; i = 0, M+1 \; \text{ oder } \; j = 0, M+1$$

und alle $t \geq 0$. Die Anfangsbedingungen $u_0 : \Omega \to \mathbb{R}$ bilden

$$U_{i,j}(0) = u_0(x_i, y_j) \qquad \text{für } \; i, j = 1, \ldots, M.$$

Wieder entsteht ein System aus gew. Dgln. in der Linienmethode. Der Fall von drei Raumdimensionen kann in analoger Weise behandelt werden.

Ebenso können wir eine Finite-Elemente-Methode zur Ortsdiskretisierung anwenden, siehe Abschnitt 2.4. Entsprechend dem Ritz-Galerkin-Verfahren lauten die Näherungen

$$u_h(x,y,t) = \sum_{j=1}^{N} \alpha_j(t)\phi_j(x,y)$$

mit den zeitabhängigen Koeffizienten α_j und den ortsabhängigen Basisfunktionen ϕ_j. Wir erhalten die Gleichungen

$$\sum_{i=1}^{N} \alpha_i'(t)\langle\phi_i,\phi_j\rangle_{L^2(\Omega)} = -\sum_{i=1}^{N} \alpha_i(t)a(\phi_i,\phi_j) \quad \text{für } j = 1,\ldots,N$$

mit der Bilinearform a. Es entsteht ein implizites System gew. Dgln.

$$M\alpha'(t) = A\alpha(t) \tag{3.26}$$

für die unbekannten Koeffizienten $\alpha = (\alpha_1,\ldots,\alpha_N)^\top$. In den konstanten Matrizen $M = (m_{ij})$, $A = (a_{ij})$ lauten die Einträge

$$m_{ij} = \langle\phi_i,\phi_j\rangle_{L^2(\Omega)}, \qquad a_{ij} = -a(\phi_i,\phi_j).$$

Insbesondere sind beide Matrizen symmetrisch. Zudem ist M positiv definit und daher regulär. Wieder können wir numerische Methoden für Systeme aus gew. Dgln. zur Lösung der Anfangswertprobleme zu (3.26) einsetzen.

Methoden vom Rothe-Typ können ebenfalls konstruiert werden. Zur part. Dgl. (3.25) auf $\Omega = (0,1)^2$ diskretisieren wir die Zeitableitung einfach durch den Differenzenquotienten erster Ordnung, d.h.

$$\tfrac{1}{k}\left[u(x,y,t_{n+1}) - u(x,y,t_n)\right] + \mathcal{O}(k) = \Delta u|_{t=t_{n+1}}.$$

Sei $z_n(x,y) \approx u(x,y,t_n)$. Es folgt das Schema

$$\Delta z_{n+1} = \tfrac{1}{k}\left[z_{n+1} - z_n\right] \tag{3.27}$$

mit gegebenem z_n und unbekanntem z_{n+1}. Die Semidiskretisierung (3.27) entspricht einer Poisson-Gleichung mit lösungsabhängiger rechter Seite. Die entstehenden Randwertprobleme können mit verfügbaren Algorithmen numerisch gelöst werden.

Kapitel 4

Hyperbolische Differentialgleichungen

Wir besprechen numerische Verfahren für hyperbolische Dgln. zweiter Ordnung. Das Musterbeispiel hierzu ist die Wellengleichung. Die Geschwindigkeit des Informationstransports ist endlich in hyperbolischen Modellen.

4.1 Wellengleichung

In einer Raumdimension lautet die Wellengleichung

$$\frac{\partial^2 u}{\partial t^2} = c^2 \, \frac{\partial^2 u}{\partial x^2} \tag{4.1}$$

mit der Wellengeschwindigkeit $c \neq 0$ (o.E.d.A. $c > 0$). Mit einer beliebigen Funktion $\Phi : \mathbb{R} \to \mathbb{R}$ und $\Phi \in C^2$ sind die Funktionen

$$u(x,t) = \Phi(x + ct) \qquad \text{und} \qquad u(x,t) = \Phi(x - ct)$$

beide Lösungen von (4.1).

Ein reines Anfangswertproblem wird *Cauchy-Problem* genannt und lautet

$$u(x,0) = u_0(x), \qquad \tfrac{\partial u}{\partial t}(x,0) = u_1(x) \tag{4.2}$$

mit vorgegebenen Funktionen $u_0, u_1 : \mathbb{R} \to \mathbb{R}$ beim Zeitpunkt (o.E.d.A.) $t_0 = 0$. Wir nehmen $u_0 \in C^2$ und $u_1 \in C^1$ an. Die Lösung des Cauchy-

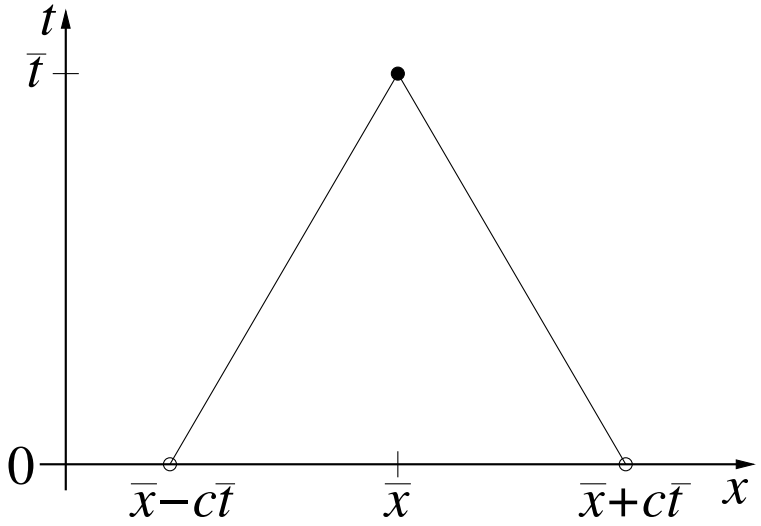


Abbildung 13: Analytischer Abhängigkeitsbereich und Informationstransport bei der Wellengleichung $u_{tt} = c^2 u_{xx}$ in einer Raumdimension.

Problems (4.1), (4.2) ist gegeben durch, siehe (1.3),

$$u(x,t) = \frac{1}{2}\left(u_0(x+ct) + u_0(x-ct) + \frac{1}{c}\int_{x-ct}^{x+ct} u_1(s)\,\mathrm{d}s \right). \tag{4.3}$$

Dies kann leicht durch Differentiation verifiziert werden. Wir bemerken eine endliche Geschwindigkeit des Informationstransports. Die Lösung u an einer Stelle $(\bar{x}, \bar{t})$ hängt nur von Anfangswerten im Intervall $x \in [\bar{x} - c\bar{t}, \bar{x} + c\bar{t}]$ zur Zeit $t_0 = 0$ ab, siehe Abbildung 13. Daher wird vom analytischen Abhängigkeitsbereich der Lösung gesprochen.

In drei Raumdimensionen lautet die Wärmeleitungsgleichung

$$\frac{\partial^2 u}{\partial t^2} = c^2 \Delta u = c^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right) \tag{4.4}$$

mit Wellengeschwindigkeit $c > 0$. Sei $r = (x, y, z)$. Spezielle Lösungen zu (4.4) sind gegeben durch

$$u(x,y,z,t) = \mathrm{e}^{\mathrm{i}(r\cdot k - \omega t)} = \mathrm{e}^{\mathrm{i}(k_x x + k_y y + k_z z - \omega t)}$$

mit der Frequenz $\omega > 0$ und dem Wellenvektor $k = (k_x, k_y, k_z)$ unter der Voraussetzung

$$\omega^2 = c^2(k_x^2 + k_y^2 + k_z^2) \quad \Rightarrow \quad \omega = c\|k\|_2.$$

Ein Cauchy-Problem ist wieder durch (4.2) festgesetzt, jetzt mit Funktionen $u_0, u_1 : \mathbb{R}^3 \to \mathbb{R}$.

Satz 4.1 *Die Lösung des Cauchy-Problems (4.4), (4.2) lautet*

$$u(x,t) = \frac{1}{4\pi c^2 t^2} \iint_{S(x,t)} u_0(y) + t u_1(y) + (y-x)^\top \nabla u_0(y) \, \mathrm{d}\sigma_y \qquad (4.5)$$

für $x \in \mathbb{R}^3$ und $t > 0$ mit der Kugeloberfläche
$S(x,t) = \{y \in \mathbb{R}^3 \ : \ \|y - x\|_2 = ct\}$.

Beweis:

Wir definieren die Sphärenmittel

$$w(x,\theta,t) = \frac{1}{4\pi} \iint_{\widehat{S}} u(x + \theta z, t) \, \mathrm{d}\sigma_z$$

mit der Einheitssphäre $\widehat{S} = \{z \in \mathbb{R}^3 \ : \ \|z\|_2 = 1\}$. Für jede stetige Funktion u gilt

$$\lim_{\theta \to 0} w(x,\theta,t) = u(x,t).$$

Wir zeigen, dass die Sphärenmittel die Wellengleichung $(\theta w)_{tt} = c^2 (\theta w)_{rr}$ zum eindimensionalen Fall erfüllen. Die Funktion w ist eine Lösung der partiellen Dgl.

$$\Delta_x w = \frac{1}{4\pi} \iint_{\widehat{S}} \Delta_x u(x + \theta z, t) \, \mathrm{d}\sigma_z = \iint_{\widehat{S}} \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2}(x + \theta z, t) \, \mathrm{d}\sigma_z = \frac{1}{c^2} w_{tt}.$$

Die Formel von Darboux für Sphärenmittel liefert $(\theta^2 w_\theta)_\theta = \Delta_x (\theta^2 w)$. Es folgt

$$\theta w_{tt} = \theta c^2 \Delta_x w = \tfrac{1}{\theta} c^2 (\Delta_x (\theta^2 w)) = \tfrac{1}{\theta} c^2 (\theta^2 w_\theta)_\theta = c^2 (\theta w)_{\theta\theta}.$$

Für die eindimensionale Wellengleichung haben wir die Lösung (4.3), d.h.

$$\theta w(x,\theta,t) = \tfrac{1}{2} \left((\theta + ct) w(x, \theta + ct, 0) + (\theta - ct) w(x, \theta - ct, 0) + \frac{1}{c} \int_{\theta - ct}^{\theta + ct} s w_t(x, s, 0) \, \mathrm{d}s \right).$$

Mit der Symmetrie $w(x, \theta - ct, 0) = w(x, ct - \theta, 0)$ folgt

$$w(x,\theta,t) = \tfrac{1}{2\theta} \left[(ct + \theta) w(x, \theta + ct, 0) - (ct - \theta) w(x, ct - \theta, 0) \right] + \tfrac{1}{2c\theta} \int_{\theta - ct}^{\theta + ct} s w_t(x, s, 0) \, \mathrm{d}s.$$

Für eine beliebige Funktion $f \in C^1$ ergibt der symmetrische Differenzenquotient

$$\lim_{\theta \to 0} \tfrac{1}{2\theta} \left[f(ct + \theta) - f(ct - \theta) \right] = f'(ct) = \tfrac{1}{c} \cdot \tfrac{\mathrm{d}}{\mathrm{d}t} f(ct).$$

Es folgt weiter

$$\lim_{\theta\to 0} \tfrac{1}{2\theta}\left[(ct+\theta)w(x,\theta+ct,0)-(ct-\theta)w(x,ct-\theta,0)\right] = \tfrac{1}{c}\cdot\tfrac{\mathrm{d}}{\mathrm{d}t}\left[ct\cdot w(x,ct,0)\right] =: A.$$

Die Funktion w und daher auch w_t sind symmetrisch bezüglich θ. Es gilt damit

$$\begin{aligned}\int_{\theta-ct}^{\theta+ct} s w_t(x,s,0)\,\mathrm{d}s &= \int_{ct-\theta}^{ct+\theta} s w_t(x,s,0)\,\mathrm{d}s + \int_{\theta-ct}^{ct-\theta} s w_t(x,s,0)\,\mathrm{d}s\\ &= \int_{ct-\theta}^{ct+\theta} s w_t(x,s,0)\,\mathrm{d}s.\end{aligned}$$

Wir erhalten

$$\lim_{\theta\to 0}\frac{1}{2c\theta}\int_{ct-\theta}^{ct+\theta} s w_t(x,s,0)\,\mathrm{d}s = t w_t(x,ct,0) =: B.$$

Es folgt

$$u(x,t) = A + B = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{t}{4\pi}\iint_{\widehat{S}} u_0(x+ctz)\,\mathrm{d}\sigma_z\right) + \frac{t}{4\pi}\iint_{\widehat{S}} u_1(x+ctz)\,\mathrm{d}\sigma_z.$$

Mit der Produktregel der Differentiation berechnen wir

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{t}{4\pi}\iint_{\widehat{S}} u_0(x+ctz)\,\mathrm{d}\sigma_z\right)\\ =&\frac{1}{4\pi}\iint_{\widehat{S}} u_0(x+ctz)\,\mathrm{d}\sigma_z + \frac{t}{4\pi}\iint_{\widehat{S}} (cz)^\top \nabla u_0(x+ctz)\,\mathrm{d}\sigma_z.\end{aligned}$$

Die Substitution

$$y = x + ctz, \qquad \mathrm{d}\sigma_y = (ct)^2\mathrm{d}\sigma_z$$

ergibt die Formel (4.5). Die Kugeloberfläche $\{y : \|y-x\|_2 = ct\}$ besitzt die Fläche $4\pi(ct)^2$. Es folgt $y - x = \mathcal{O}(t)$. Für $t \approx 0$ impliziert die Formel (4.5)

$$u(x,t) \approx u_0(x) + t u_1(x).$$

Somit sind die Anfangsbedingungen erfüllt. □

Wir bemerken erneut die endliche Geschwindigkeit c für den Informationstransport. Für eine Stelle $(\bar{x},\bar{t})$ mit $\bar{x}\in\mathbb{R}^3$ und $\bar{t} > 0$ hängt die Lösung $u(\bar{x},\bar{t})$ nur von den Anfangswerten auf der Menge $\{x\in\mathbb{R}^3 : \|x-\bar{x}\|_2 = c\bar{t}\}$ ab.

4.2 Finite-Differenzen-Methoden

Zuerst besprechen wir Finite-Differenzen-Methoden im Fall einer Raumdimension. Später werden diese Verfahren auf den mehrdimensionalen Fall verallgemeinert.

Eine Raumdimension

Wir wenden eine Finite-Differenzen-Methode an auf die Wellengleichung mit Quellterm in einer Raumdimension

$$\frac{\partial^2 u}{\partial t^2} = c^2 \, \frac{\partial^2 u}{\partial x^2} + f(x,t,u). \tag{4.6}$$

Die Schrittweiten $k, h > 0$ werden in der Zeit bzw. im Ort verwendet. Die Gitterpunkte seien $x_j = jh$ und $t_n = nk$. Üblicherweise werden die partiellen Ableitungen ersetzt durch die Differenzenquotienten

$$\frac{\partial^2 u}{\partial x^2}(x,t) = \frac{u(x+h,t) - 2u(x,t) + u(x-h,t)}{h^2} - \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(x+\vartheta h, t)$$

$$\frac{\partial^2 u}{\partial t^2}(x,t) = \frac{u(x,t+k) - 2u(x,t) + u(x,t-k)}{k^2} - \frac{k^2}{12}\frac{\partial^4 u}{\partial t^4}(x, t+\eta k)$$

mit $-1 < \vartheta, \eta < 1$. Es folgt die Finite-Differenzen-Methode

$$\tfrac{1}{k^2}\left[U_j^{n+1} - 2U_j^n + U_j^{n-1}\right] = c^2 \tfrac{1}{h^2}\left[U_{j-1}^n - 2U_j^n + U_{j+1}^n\right] + f(x_j, t_n, U_j^n)$$

für die Näherungen $U_j^n \approx u(x_j, t_n)$, oder äquivalent

$$\begin{aligned} U_j^{n+1} = -U_j^{n-1} + 2\left(1 - c^2\tfrac{k^2}{h^2}\right) U_j^n + c^2 \tfrac{k^2}{h^2}\left[U_{j-1}^n + U_{j+1}^n\right] \\ + k^2 f(x_j, t_n, U_j^n). \end{aligned} \tag{4.7}$$

Daher erhalten wir ein Zweischritt-Verfahren. Die Diskretisierung entspricht einem Fünf-Punkte-Stern. Der lokale Diskretisierungsfehler der Finiten-Differenzen-Methode lautet

$$\tau(k,h) = \tfrac{k^2}{12}\tfrac{\partial^4 u}{\partial t^4}(x, t+\eta k) - c^2 \tfrac{h^2}{12}\tfrac{\partial^4 u}{\partial x^4}(x+\vartheta h, t).$$

Für $u \in C^4$ folgt die Konsistenz aus der gleichmäßigen Abschätzung

$$|\tau(k,h)| \leq k^2 \tfrac{1}{12} \max_{x\in[a,b],t\in[0,T]} \left|\tfrac{\partial^4 u}{\partial t^4}\right| + h^2 \tfrac{c^2}{12} \max_{x\in[a,b],t\in[0,T]} \left|\tfrac{\partial^4 u}{\partial x^4}\right|$$

für $x \in (a,b)$ und $t \in (0,T)$ bei beliebigen $a, b \in \mathbb{R}$ $(a < b)$ und $T > 0$.

Wir betrachten das Cauchy-Problem (4.2). Zur Anwendung der Finite-Differenzen-Methode (4.7) benötigen wir die Anfangswerte U_j^0 und U_j^1 für jedes j. Die vorgegebenen Anfangswerte liefern

$$U_j^0 = u_0(x_j), \qquad U_j^1 = u_0(x_j) + k u_1(x_j).$$

Jedoch ist diese Diskretisierung nur konsistent von Ordnung 1. Um eine Methode mit insgesamt der Ordnung 2 zu erhalten, benutzen wir die Diskretisierung

$$\tfrac{1}{2k}[u(x_j,t_1) - u(x_j,t_{-1})] = u_t(x_j,t_0) + \mathcal{O}(k^2)$$

mit der Hilfszeitschicht $t_{-1} = -k$. Es folgt

$$U_j^1 = U_j^{-1} + 2k u_1(x_j).$$

Die Finite-Differenzen-Methode (4.7) ergibt für $n = 0$

$$U_j^1 = -U_j^{-1} + 2(1 - c^2\tfrac{k^2}{h^2})U_j^0 + c^2\tfrac{k^2}{h^2}\left[U_{j-1}^0 + U_{j+1}^0\right] + k^2 f(x_j, 0, U_j^0)$$

und daher

$$\begin{aligned} U_j^1 = {} & -U_j^{-1} + 2(1 - c^2\tfrac{k^2}{h^2})u_0(x_j) + c^2\tfrac{k^2}{h^2}\left[u_0(x_{j-1}) + u_0(x_{j+1})\right] \\ & + k^2 f(x_j, 0, u_0(x_j)). \end{aligned}$$

Es folgen die Näherungen

$$\begin{aligned} U_j^1 = {} & k u_1(x_j) + (1 - c^2\tfrac{k^2}{h^2})u_0(x_j) + c^2\tfrac{k^2}{2h^2}\left[u_0(x_{j-1}) + u_0(x_{j+1})\right] \\ & + \tfrac{k^2}{2} f(x_j, 0, u_0(x_j)), \end{aligned}$$

wobei nun alle Terme auf der rechten Seite bekannt sind.

Wir diskutieren das Cauchy-Problem (4.1), (4.2), d.h. ohne Randbedingungen zu betrachten. Die Finite-Differenzen-Methode (4.7) wird verwendet

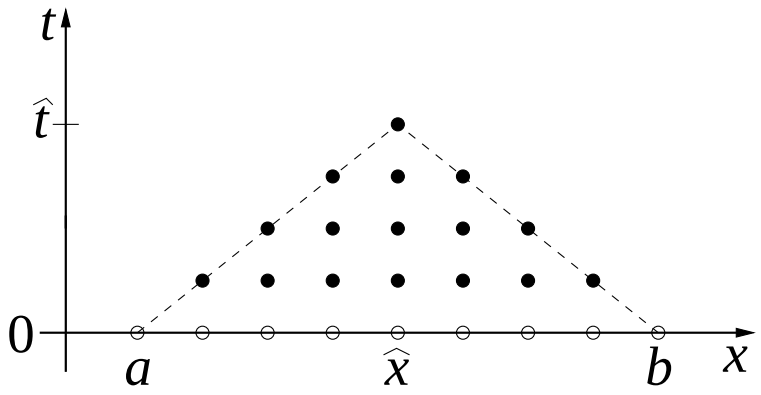


Abbildung 14: Gitter in Finite-Differenzen-Methode bei reinem Anfangswertproblem.

(mit $f \equiv 0$). Sei $r = \frac{k}{h}$ konstant. Wir wählen ein endliches Intervall $x \in [a, b]$ und $h = \frac{b-a}{2M}$ mit einer ganzen Zahl M. Sei $x_j = a + jh$ für $j = 0, 1, \ldots, 2M$. Wenn R Gitterpunkte in der Zeitschicht t_n gegeben sind, dann können damit nur $R-2$ Gitterpunkte in der neuen Zeitschicht t_{n+1} bestimmt werden. Dieses Vorgehen wird in Abbildung 14 skizziert. Es folgt, dass bis zu M Zeitschritte durchgeführt werden können. Wir erhalten eine Näherung im Endpunkt $(\hat{x}, \hat{t})$ mit

$$\hat{x} = \tfrac{a+b}{2}, \qquad \hat{t} = Mk = Mrh = r\tfrac{b-a}{2},$$

dessen Lage unabhängig von M ist unter der Annahme eines Konstanten r. Das Intervall $[a, b]$ stellt einen numerischen Abhängigkeitsbereich für die Methode (Abhängigkeit von Anfangswerten) dar.

Gemäß (4.3) hängt die exakte Lösung $u(\hat{x}, \hat{t})$ von den Anfangswerten in $x \in [\hat{x} - c\hat{t}, \hat{x} + c\hat{t}]$ ab, siehe Abbildung 13. Daher kann das Verfahren nur konvergent sein, wenn

$$\mathcal{D}(\hat{x}, \hat{t}) := [\hat{x} - c\hat{t}, \hat{x} + c\hat{t}] \subseteq [a, b] =: \mathcal{D}_0(\hat{x}, \hat{t}).$$

Anderenfalls könnten wir nämlich die Anfangswerte für $x \notin [a, b]$ derart verändern, dass $u(\hat{x}, \hat{t})$ sich ändert, während die numerische Lösung gleich bleibt. In diesem Kontext stellen $\mathcal{D}$ und $\mathcal{D}_0$ den analytischen Abhängigkeitsbereich bzw. den numerischen Abhängigkeitsbereich dar. Als notwendige Bedingung für Konvergenz folgt

$$c\hat{t} \leq \tfrac{b-a}{2} \qquad \Rightarrow \qquad r \leq \tfrac{1}{c}.$$

Falls die Ortsschrittweite h gegeben ist, dann erhalten wir an die Zeitschrittweite k eine Restriktion wegen $r = \frac{k}{h}$ konstant. Jedoch ist diese Einschränkung nicht so stark wie bei expliziten Finite-Differenzen-Methoden für parabolische Probleme, wo $r = \frac{k}{h^2}$ gilt.

Desweiteren können Randbedingungen verwendet werden bei $x = a$ und / oder $x = b$ für $t \geq 0$, siehe (3.4) und (3.5).

Wir analysieren die Stabilität der Finiten-Differenzen-Methode mit dem Kriterium nach von-Neumann. Typische Lösungen zu $u_{tt} = c^2 u_{xx}$ lauten

$$u(x,t) = \mathrm{e}^{\mathrm{i}(\lambda x - \omega t)} = \mathrm{e}^{-\mathrm{i}\omega t}\mathrm{e}^{\mathrm{i}\lambda x}$$

mit $\lambda, \omega \in \mathbb{R}$ und $\omega = c|\lambda|$. Wir erhalten die Konstante $\alpha = -\mathrm{i}\omega$ in dieser exakten Lösung. Es folgt $\mathrm{Re}(\alpha) = 0$ und somit $|\mathrm{e}^{\alpha t}| = 1$. Störungen in den Anfangswerten werden weder verstärkt noch gedämpft, sie werden in der Zeit weitertransportiert. Im Gegensatz dazu ist bei der Wärmeleitungsgleichung $u_t = u_{xx}$ die Lösung $u(x,t) = \mathrm{e}^{\alpha t}\mathrm{e}^{\mathrm{i}\lambda x}$ mit $\alpha = -\lambda^2 \leq 0$. Es folgt $|\mathrm{e}^{\alpha t}| < 1$ für $t > 0$ und alle $\lambda \neq 0$.

Wir verwenden den Ansatz $U_j^n = \mathrm{e}^{\alpha nk}\mathrm{e}^{\mathrm{i}\lambda jh}$ in der Finiten-Differenzen-Methode (4.7) ohne Quellterm ($f \equiv 0$). Es folgt

$$\begin{aligned}\mathrm{e}^{\alpha(n+1)k}\mathrm{e}^{\mathrm{i}\lambda jh} = &-\mathrm{e}^{\alpha(n-1)k}\mathrm{e}^{\mathrm{i}\lambda jh} + 2\left(1 - c^2r^2\right)\mathrm{e}^{\alpha nk}\mathrm{e}^{\mathrm{i}\lambda jh} \\ &+ c^2r^2\left[\mathrm{e}^{\alpha nk}\mathrm{e}^{\mathrm{i}\lambda(j-1)h} + \mathrm{e}^{\alpha nk}\mathrm{e}^{\mathrm{i}\lambda(j+1)h}\right].\end{aligned}$$

Division durch $\mathrm{e}^{\alpha nk}\mathrm{e}^{\mathrm{i}\lambda jh}$ ergibt

$$\mathrm{e}^{\alpha k} = -\mathrm{e}^{-\alpha k} + 2\left(1 - c^2r^2\right) + c^2r^2\left[\mathrm{e}^{\mathrm{i}\lambda(-h)} + \mathrm{e}^{\mathrm{i}\lambda h}\right].$$

Mit $\xi = \mathrm{e}^{\alpha k}$ folgt die quadratische Gleichung

$$\xi^2 + \left(4r^2c^2\sin^2\left(\tfrac{\lambda h}{2}\right) - 2\right)\xi + 1 = 0.$$

Wir benutzen die Abkürzung $b = 4r^2c^2\sin^2\left(\frac{\lambda h}{2}\right) - 2$. Offensichtlich gilt $b \in [-2, 4r^2c^2 - 2]$. Die Lösungen der quadratischen Gleichung lauten

$$\xi_{1/2} = \tfrac{1}{2}\left[-b \pm \sqrt{b^2 - 4}\right].$$

Wir setzen die notwendige Bedingung $r \leq \frac{1}{c}$ voraus. Es folgt $b \in [-2, 2]$ wegen $r^2c^2 \leq 1$. Dadurch ergeben sich die komplexen Lösungen

$$\xi_{1/2} = \tfrac{1}{2}\left[-b \pm \mathrm{i}\sqrt{4 - b^2}\right]$$

mit $4 - b^2 \geq 0$. Es folgt

$$|\xi_{1/2}|^2 = \tfrac{1}{4}\left[(-b)^2 + \sqrt{4 - b^2}^2\right] = 1.$$

Wegen $|\xi_1| = 1$ und $|\xi_2| = 1$ ist die Finite-Differenzen-Methode (4.7) stabil bezüglich des Kriteriums nach von-Neumann unter der Voraussetzung $r \leq \frac{1}{c}$. Zudem stimmt die Größenordnung der Terme $\xi = \mathrm{e}^{\alpha k}$ mit der Struktur der exakten Lösungen von $u_{tt} = c^2 u_{xx}$ überein. Da die Methode konsistent ist, folgt die Konvergenz mit Ordnung 2 falls $r \leq \frac{1}{c}$. Das Verfahren (4.7) ist nicht konvergent falls $r > \frac{1}{c}$.

Desweiteren ist die Geschwindigkeit des Informationstransports endlich in einem expliziten Verfahren, bei parabolischen sowie hyperbolischen Problemen. Im Gegensatz dazu ist die Geschwindigkeit des Informationstransports unbegrenzt bei impliziten Verfahren, bei parabolischen sowie hyperbolischen Problemen. Daher passen explizite Methoden besser zur Struktur hyperbolischer Dgln., während implizite Methoden geeigneter für parabolische Dgln. sind.

Mehrere Raumdimensionen

Wir betrachten das Cauchy-Problem (4.2) der Wellengleichung mit Quellterm f in drei Raumdimensionen

$$\frac{\partial^2 u}{\partial t^2} = c^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}\right) + f(x, y, z, t, u). \tag{4.8}$$

Die Zeitableitung wird wieder durch den symmetrischen Differenzenquotienten zweiter Ordnung mit Schrittweite k diskretisiert. Wir benutzen identische Schrittweiten h zur Diskretisierung der Ortsableitungen. Jedoch dürfen die Differenzenformeln auch verschieden sein. Seien $x_j = jh$, $y_p = ph$,

$z_q = qh$, $t_n = nk$ die Gitterpunkte und $U^n_{j,p,q} \approx u(x_j, y_p, z_q, t_n)$ die zugehörigen Näherungen. Wir verwenden Diskretisierungen der Gestalt

$$\begin{aligned}
\frac{\partial^2 u}{\partial x^2}(x_j, y_p, z_q, t_n) &\approx \frac{1}{h^2} \sum_{\nu=-N}^{N} w^x_\nu U^n_{j+\nu,p,q} \\
\frac{\partial^2 u}{\partial y^2}(x_j, y_p, z_q, t_n) &\approx \frac{1}{h^2} \sum_{\nu=-N}^{N} w^y_\nu U^n_{j,p+\nu,q} \\
\frac{\partial^2 u}{\partial z^2}(x_j, y_p, z_q, t_n) &\approx \frac{1}{h^2} \sum_{\nu=-N}^{N} w^z_\nu U^n_{j,p,q+\nu}
\end{aligned}$$

mit den Koeffizienten $w^x_\nu, w^y_\nu, w^z_\nu \in \mathbb{R}$. Die symmetrische Differenzenformel zweiter Ordnung besitzt die Koeffizienten $w_0 = -2$, $w_1 = w_{-1} = 1$. Die symmetrische Differenzenformel vierter Ordnung führt auf die Koeffizienten $w_0 = -\frac{30}{12}$, $w_1 = w_{-1} = \frac{16}{12}$, $w_2 = w_{-2} = -\frac{1}{12}$.

Die resultierende Finite-Differenzen-Methode lautet

$$\begin{aligned}
&U^{n+1}_{j,p,q} - 2U^n_{j,p,q} + U^{n-1}_{j,p,q} \\
&= c^2 r^2 \Big(\sum_{\nu=-N}^{N} w^x_\nu U^n_{j+\nu,p,q} + \sum_{\nu=-N}^{N} w^y_\nu U^n_{j,p+\nu,q} + \sum_{\nu=-N}^{N} w^z_\nu U^n_{j,p,q+\nu} \Big) \\
&\quad + k^2 f(x_j, y_p, z_q, t_n, U^n_{j,p,q})
\end{aligned}$$

mit $r = \frac{k}{h}$-

Wir analysieren die Stabilität mit dem Kriterium nach von-Neumann im Fall der Wellengleichung (4.8) ohne Quellterm ($f \equiv 0$). Der Ansatz

$$U^n_{j,p,q} = \mathrm{e}^{\alpha n k} \mathrm{e}^{\mathrm{i}(\lambda_x j h + \lambda_y p h + \lambda_z q h)}$$

mit beliebigen Konstanten $\lambda_x, \lambda_y, \lambda_z \in \mathbb{R}$ wird in die Formel der Finite-Differenzen-Methode eingesetzt. Eine Division durch $U^n_{j,p,q}$ zeigt mit der Variable $\xi = \mathrm{e}^{\alpha k}$

$$\xi - 2 + \xi^{-1} = c^2 r^2 \sum_{\nu=-N}^{N} w^x_\nu \mathrm{e}^{\mathrm{i}\lambda_x \nu h} + w^y_\nu \mathrm{e}^{\mathrm{i}\lambda_y \nu h} + w^z_\nu \mathrm{e}^{\mathrm{i}\lambda_z \nu h}.$$

Es folgt die quadratische Gleichung $\xi^2 + b\xi + 1 = 0$ mit $b = -2 - c^2 r^2 A$ und

$$A := \sum_{\nu=-N}^{N} w_\nu^x \mathrm{e}^{\mathrm{i}\lambda_x \nu h} + w_\nu^y \mathrm{e}^{\mathrm{i}\lambda_y \nu h} + w_\nu^z \mathrm{e}^{\mathrm{i}\lambda_z \nu h}.$$

Im Folgenden nehmen wir $A \in \mathbb{R}$ und $A \leq 0$ an, welches bei symmetrischen Differenzenformeln gegeben ist. Die Lösungen ξ_1, ξ_2 der quadratischen Gleichung erfüllen $|\xi_{1/2}| = 1$ falls $b^2 - 4 \leq 0$. Es folgt die Forderung

$$(-2 - c^2 r^2 A)^2 \leq 4 \quad \Leftrightarrow \quad 4A + c^2 r^2 A^2 \leq 0 \quad \Leftrightarrow \quad 4 + c^2 r^2 A \geq 0$$

unter der zusätzlichen Annahme $A < 0$. Mit $A = -|A|$ ergibt sich

$$r \leq \frac{2}{c\sqrt{|A|}}.$$

Im Fall $A = 0$ entsteht keine Restriktion (r beliebig). Wir bestimmen eine obere Schranke für $|A|$. Die Dreiecksungleichung liefert sukzessive

$$|A| \leq \sum_{\nu=-N}^{N} |w_\nu^x| + |w_\nu^y| + |w_\nu^z| =: B. \tag{4.9}$$

Daher ist für die Stabilität bezüglich des Kriteriums nach von-Neumann hinreichend

$$r \leq \frac{2}{c\sqrt{B}} \leq \frac{2}{c\sqrt{|A|}}.$$

Wir erhalten den eindimensionalen und zweidimensionalen Fall der Wellengleichung, indem einfach die Koeffizientnen w_ν^y oder w_ν^z weggelassen werden. Unter der Annahme $c = 1$ folgen die in der Tabelle dargestellten Restriktionen $\frac{2}{\sqrt{B}}$ an die Schrittweiten bei den symmetrischen Differenzenformeln der Ordnung 2 und Ordnung 4 (im Ort):

	eindim.	zweidim.	dreidim.
Ordnung 2	1	$\frac{1}{\sqrt{2}} \approx 0.707$	$\frac{1}{\sqrt{3}} \approx 0.577$
Ordnung 4	$\frac{\sqrt{3}}{2} \approx 0.866$	$\sqrt{\frac{3}{8}} \approx 0.612$	$\frac{1}{2} = 0.5$

Eine weitere Untersuchung zeigt, dass diese Schranken für r auch notwendig für die Stabilität nach von-Neumann sind. Es folgt eine Restriktion an die Wahl der Zeitschrittweite k bei gegebener Ortsschrittweite h.

4.3 Charakteristikenverfahren

Nun führen wir charakteristische Kurven einer allgemeinen partiellen Dgl. zweiter Ordnung ein. Wir konstruieren damit eine numerische Methode für hyperbolische Dgln.

Motivation

Wir betrachten eine semi-lineare partielle Dgl. zweiter Ordnung

$$Au_{xx} + 2Bu_{xy} + Cu_{yy} = f(x, y, u, u_x, u_y) \tag{4.10}$$

mit Lösung $u : \mathbb{R}^2 \to \mathbb{R}$ und kostanten Koeffizienten $A, B, C \in \mathbb{R}$. Entsprechend der Klassifikation für Dgln. (4.10) aus Abschnitt 1.2 gilt

$$\begin{array}{ll} \text{elliptisch} & \text{für } AC - B^2 > 0, \\ \text{parabolisch} & \text{für } AC - B^2 = 0, \\ \text{hyperbolisch} & \text{für } AC - B^2 < 0. \end{array}$$

Wir suchen eine Koordinatentransformation $\xi = \xi(x, y)$, $\eta = \eta(x, y)$ mit $w(\xi, \eta) = u(x, y)$ durch die in der transformierten Gleichung

$$A^* w_{\xi\xi} + 2B^* w_{\xi\eta} + C^* w_{\eta\eta} = \tilde{f}(\xi, \eta, w, w_\xi, w_\eta)$$

dann $A^* = C^* = 0$ gilt. Dementsprechend nehmen wir $A \neq 0$ oder $C \neq 0$ in (4.10) an. O.E.d.A. sei $A \neq 0$. Die Transformation ist bijektiv genau dann, wenn

$$\det \begin{pmatrix} \xi_x & \xi_y \\ \eta_x & \eta_y \end{pmatrix} = \xi_x \eta_y - \xi_y \eta_x \neq 0. \tag{4.11}$$

Wir erhalten

$$\begin{aligned} u_x &= w_\xi \xi_x + w_\eta \eta_x \\ u_{xx} &= (w_{\xi\xi}\xi_x + w_{\xi\eta}\eta_x)\xi_x + w_\xi \xi_{xx} + (w_{\eta\xi}\xi_x + w_{\eta\eta}\eta_x)\eta_x + w_\eta \eta_{xx} \\ &= w_{\xi\xi}\xi_x^2 + 2w_{\eta\xi}\xi_x\eta_x + w_{\eta\eta}\eta_x^2 + w_\xi \xi_{xx} + w_\eta \eta_{xx} \\ &\text{etc.} \end{aligned}$$

Es ergibt sich das transformierte System

$$\begin{aligned}&\left(A\xi_x^2 + 2B\xi_x\xi_y + C\xi_y^2\right) w_{\xi\xi}\\ +\,2\left(A\xi_x\eta_x + B(\xi_x\eta_y + \xi_y\eta_x) + C\xi_y\eta_y\right) &w_{\xi\eta}\\ +\left(A\eta_x^2 + 2B\eta_x\eta_y + C\eta_y^2\right) &w_{\eta\eta} = \tilde{f}(\xi,\eta,w,w_\xi,w_\eta).\end{aligned}$$

Wir möchten erreichen, dass

$$\begin{aligned}A^* &:= A\xi_x^2 + 2B\xi_x\xi_y + C\xi_y^2 &&= 0,\\ C^* &:= A\eta_x^2 + 2B\eta_x\eta_y + C\eta_y^2 &&= 0.\end{aligned}$$

Wir erhalten zwei quadratische Gleichungen jeweils für $\frac{\xi_x}{\xi_y}$ und $\frac{\eta_x}{\eta_y}$. Jedoch sind die beiden quadratischen Gleichungen identisch. Um eine bijektive Koordinatentransformation zu erhalten, benötigen wir $\frac{\xi_x}{\xi_y} \neq \frac{\eta_x}{\eta_y}$ wegen (4.11), d.h. zwei verschiedene Lösungen der quadratischen Gleichung. Für $A \neq 0$ sind die Lösungen der quadratischen Gleichung $A\mu^2 + 2B\mu + C = 0$ dann

$$\mu_{1/2} = \frac{-B \pm \sqrt{B^2 - AC}}{A}.$$

Die Bedingung $B^2 - AC > 0$ ist äquivalent zur Existenz zweier verschiedener reeller Lösungen $\mu_1, \mu_2 \in \mathbb{R}$. Dies gilt nur bei hyperbolischen Dgln. Wir erhalten die Koordinatentransformation

$$2B^* w_{\xi\eta} = \tilde{f}(\xi,\eta,w,w_\xi,w_\eta).$$

Die beteiligten Koeffizienten erfüllen

$$B^* = A\xi_x\eta_x + B(\xi_x\eta_y + \xi_y\eta_x) + C\xi_y\eta_y = \cdots = -\tfrac{2}{A}(B^2 - AC)\xi_y\eta_y \neq 0$$

für $\xi_y, \eta_y \neq 0$.

Da A, B, C konstant sind, liefern die Gleichungen $\xi_x = \mu_1\xi_y$ und $\eta_x = \mu_2\eta_y$ die Transformation

$$\xi = \mu_1 x + y, \qquad \eta = \mu_2 x + y.$$

Für ξ = konst. oder η = konst. erhalten wir Geraden im Definitionsbereich (x, y). Diese Geraden sind die charakteristischen Kurven. Wegen $\mu_1 \neq \mu_2$ liegen zwei Familien aus charakteristischen Kurven vor.

Charakteristische Kurven

Wir betrachten die semi-lineare partielle Dgl.

$$A(x,y)u_{xx} + 2B(x,y)u_{xy} + C(x,y)u_{yy} = f(x,y,u,u_x,u_y) \tag{4.12}$$

mit nichtkonstanten Koeffizienten A, B, C. Wir möchten ein korrekt gestelltes Anfangswertproblem festlegen. Im Definitionsbereich sei eine Kurve

$$\mathcal{K} = \{(x(\tau), y(\tau)) : \tau \in [\tau_0, \tau_{\text{end}}]\}$$

gegeben mit $x, y \in C^1$ und $\dot{x}(\tau)^2 + \dot{y}(\tau)^2 > 0$ für alle τ. In einem Cauchy-Problem werden Anfangswerte auf dieser Kurve festgesetzt durch

$$u(x(\tau), y(\tau)) = u_0(\tau), \qquad \tfrac{\partial u}{\partial n}\big|_{x=x(\tau),y=y(\tau)} = u_1(\tau) \tag{4.13}$$

mit vorgegebenen Funktionen $u_0, u_1 : [\tau_0, \tau_{\text{end}}] \to \mathbb{R}$. Dabei ist $n = (n_1, n_2)$ ein Normalenvektor zur Kurve $\mathcal{K}$ mit $\|n\|_2 = 1$. Sei $u_0 \in C^1$. Die Ableitung von u in Tangentialrichtung $s = (s_1, s_2)$ ist gegeben durch

$$\tfrac{\partial u}{\partial s}\big|_{x=x(\tau),y=y(\tau)} = u_x(x(\tau), y(\tau))\dot{x}(\tau) + u_y(x(\tau), y(\tau))\dot{y}(\tau) = \dot{u}_0(\tau).$$

Da s und n linear unabhängig sind, werden im Cauchy-Problem alle ersten Ableitungen u_x, u_y entlang der Kurven $\mathcal{K}$ festgelegt. Eine weitere Differentiation liefert die zweiten Ableitungen

$$\dot{u}_x = \tfrac{\mathrm{d}}{\mathrm{d}\tau}u_x = u_{xx}\dot{x} + u_{xy}\dot{y}, \qquad \dot{u}_y = \tfrac{\mathrm{d}}{\mathrm{d}\tau}u_y = u_{yx}\dot{x} + u_{yy}\dot{y}. \tag{4.14}$$

Es gilt $u_{xy} = u_{yx}$ für $u \in C^2$. Sei

$$\tilde{f}(\tau) = f(x(\tau), y(\tau), u(\tau), u_x(\tau), u_y(\tau)).$$

We schreiben die Gleichungen (4.14) zusammen mit der Dgl. (4.12) als lineares Gleichungssystem

$$\begin{pmatrix} A & 2B & C \\ \dot{x} & \dot{y} & 0 \\ 0 & \dot{x} & \dot{y} \end{pmatrix} \begin{pmatrix} u_{xx} \\ u_{xy} \\ u_{yy} \end{pmatrix} = \begin{pmatrix} \tilde{f} \\ \dot{u}_x \\ \dot{u}_y \end{pmatrix}. \tag{4.15}$$

Wir möchten, dass die Vorgaben u, u_x, u_y auf $\mathcal{K}$ eine eindeutige Lösung der Dgl. (4.12) liefern. Es gilt, dass jedes Cauchy-Problem (4.13) eine eindeutige Lösung besitzt genau dann, wenn das lineare Gleichungssystem (4.15)

eindeutig lösbar ist. Äquivalent fordern wir, dass die Determinante der Koeffizientenmatrix in (4.15) ungleich null ist, also

$$A\dot{y}^2 - 2B\dot{x}\dot{y} + C\dot{x}^2 \neq 0.$$

Wir nehmen $\dot{x} \neq 0$ der Einfachheit wegen an. Mit $y' = \frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\dot{y}}{\dot{x}}$ führt die Annahme des Gegenteils auf die quadratische Gleichung

$$A(y')^2 - 2By' + C = 0.$$

Es ergibt sich die Definition von charakteristischen Kurven.

Definition 4.2 (Charakteristiken) *Die charakteristischen Kurven (oder: Charakteristiken) einer partiellen Dgl. (4.12) zweiter Ordnung sind die reellwertigen Lösungen $y(x)$ der gewöhnlichen Dgl.*

$$y'(x) = \frac{B(x,y) \pm \sqrt{B(x,y)^2 - A(x,y)C(x,y)}}{A(x,y)} \tag{4.16}$$

unter der Annahme $A(x,y) \neq 0$.

Es folgt, dass die Existenz und Eindeutigkeit einer Lösung der partiellen Dgl. (4.12) im Cauchy-Problem (4.13) nicht erfüllt ist, falls die Anfangskurve $\mathcal{K}$ tangential zu einer charkteristischen Kurve in einem Punkt liegt. Umgekehrt existiert eine eindeutige Lösung, falls die Anfangskurve $\mathcal{K}$ niemals tangential zu einer charakteristischen Kurve verläuft. Die gewöhnliche Dgl. (4.16) beschreibt eine Familie aus charakteristischen Kurven.

Bei einer elliptischen Dgl. gilt $B^2 - AC < 0$. Folglich existieren keine charakteristischen Kurven. Eine eindeutige Lösung des Cauchy-Problems existiert für beliebige Amfangskurven $\mathcal{K}$. Jedoch sind Anfangswertprobleme bei elliptischen Dgln. nicht korrekt gestellt, da die Lösungen nicht stetig von den Anfangsdaten abhängen.

Bei einer parabolischen Dgl. gilt $B^2 - AC = 0$ und daher $y' = \frac{B}{A}$. Eine Familie aus charakteristischen Kurven existiert. Jedoch werden oft keine Cauchy-Probleme der Gestalt (4.13) betrachtet. Beispielsweise werden beim reinen

Anfangswertproblem zur Wärmeleitungsgleichung nur die Anfangswerte u bei $t = 0$ vorgegeben und nicht die Ableitung u_t in Normalenrichtung.

Bei einer hyperbolischen Dgl. gilt $B^2 - AC > 0$. Dadurch existieren zwei Familien aus charakteristischen Kurven. Die Anfangskurve $\mathcal{K}$ darf nie tangential zu einer dieser Charakteristiken liegen. Beispielsweise ist diese Bedingung erfüllt bei der Wellengleichung $u_{tt} = c^2 u_{xx}$ und den Anfangswerten u, u_t bei $t = 0$. Charakteristische Kurven sind nur bei hyperbolischen Dgln. von Interesse, da das Cauchy-Problem (4.13) irrelevant bei elliptischen Problemen oder parabolischen Problemen ist.

Definition 4.2 bleibt auch im quasi-linearen Fall $A = A(x, y, u, u_x, u_y)$, $B = B(x, y, u, u_x, u_y)$, $C = C(x, y, u, u_x, u_y)$ bestehen. Jedoch hängen die charakteristischen Kurven dann von der a priori unbekannten Lösung ab.

Numerisches Verfahren

Wir betrachten ein Cauchy-Problem (4.13) bei einer hyperbolischen Dgl. (4.12). Es existieren zwei Familien aus charakteristischen Kuven, siehe Definition 4.2. Der Informationstransport vollzieht sich entlang der charakteristischen Kurven. Wir können diese Eigenschaft verwenden, um ein numerisches Verfahren zu konstruieren.

Entlang einer charakteristischen Kurve $(x(\tau), y(\tau))$ besitzt das lineare Gleichungssystem (4.15) keine eindeutige Lösung, da die Koeffizientenmatrix singulär ist. Insbesondere gilt

$$\text{Rang} \begin{pmatrix} A & 2B & C \\ \dot{x} & \dot{y} & 0 \\ 0 & \dot{x} & \dot{y} \end{pmatrix} = 2$$

für $\dot{x} \neq 0$. Trotzdem nehmen wir an, dass eine eindeutige Lösung eines Cauchy-Problems existiert, wobei die Anfangskurve $\mathcal{K}$ nicht tangential zu einer charakteristischen Kurve verläuft. Es folgt, dass das lineare Gleichungssystem (4.15) auch entlang einer charakteristischen Kurve eine Lösung be-

sitzt. Dies führt auf

$$\text{Rang} \begin{pmatrix} A & 2B & C & \tilde{f} \\ \dot{x} & \dot{y} & 0 & \dot{u}_x \\ 0 & \dot{x} & \dot{y} & \dot{u}_y \end{pmatrix} = 2.$$

Wenn wir aus den vier Spaltenvektoren drei auswählen, dann muss die zugehörige Determinante null sein. Eine bestimmte Wahl ergibt

$$\det \begin{pmatrix} A & C & \tilde{f} \\ \dot{x} & 0 & \dot{u}_x \\ 0 & \dot{y} & \dot{u}_y \end{pmatrix} = 0,$$

was äquivalent ist zu

$$A\dot{u}_x\dot{y} + C\dot{u}_y\dot{x} - \tilde{f}\dot{x}\dot{y} = 0. \tag{4.17}$$

Eine andere äquivalente Formulierung lautet

$$A\frac{\dot{u}_x}{\dot{x}} + C\frac{\dot{u}_y}{\dot{y}} = \tilde{f} \qquad \text{für } \dot{y}, \dot{x} \neq 0. \tag{4.18}$$

Daher erhalten wir eine Information über die Änderung von u_x, u_y entlang der charakteristischen Kurven. Als Abkürzungen führen wir ein

$$\alpha = \frac{B + \sqrt{B^2 - AC}}{A} \quad \text{und} \quad \beta = \frac{B - \sqrt{B^2 - AC}}{A}, \tag{4.19}$$

wobei α und β von x, y abhängen. Es gilt $\alpha \neq \beta$ bei hyperbolischen Dgln. Die gewöhnliche Dgl. (4.16) impliziert $\dot{y} = \alpha\dot{x}$ und $\dot{y} = \beta\dot{x}$. Die beiden Familien aus charakteristischen Kurven können wir schreiben als

$$\mathcal{K}_\alpha = \{(x(\tau), y(\tau)) : \dot{y} = \alpha\dot{x}\} \quad \text{und} \quad \mathcal{K}_\beta = \{(x(\tau), y(\tau)) : \dot{y} = \beta\dot{x}\}.$$

Die Gleichung (4.17) liefert

$$\begin{aligned} A\alpha\dot{u}_x + C\dot{u}_y &= \tilde{f}\dot{y} = \alpha\tilde{f}\dot{x}, \\ A\beta\dot{u}_x + C\dot{u}_y &= \tilde{f}\dot{y} = \beta\tilde{f}\dot{x}. \end{aligned} \tag{4.20}$$

Diese beiden Gleichungen können wir verwenden, um u_x und u_y zu bestimmen.

Beispiel: Zur hyperbolischen Dgl. $u_{xx}-u_{yy}=2(y^2-x^2)$ lösen wir das Anfangswertproblem $u(0,y)=y^2$, $u_x(0,y)=0$ für $y\in\mathbb{R}$ analytisch mittels der charakteristischen Kurven.

Es gilt hier $A=1$, $B=0$, $C=-1$. Es folgt $y'=\pm 1$ in (4.16), d.h. $\alpha=1$, $\beta=-1$. Die charakteristischen Kurven können geschrieben werden als

$$\mathcal{K}_\alpha : y = C_\alpha + x, \qquad \mathcal{K}_\beta : y = C_\beta - x$$

mit Konstanten $C_\alpha, C_\beta \in \mathbb{R}$. Die Gleichung (4.18) liefert

$$\begin{aligned}
\frac{\dot{u}_x}{\dot{x}}(x, C_\alpha + x) - \frac{\dot{u}_y}{\dot{y}}(x, C_\alpha + x) &= 2((C_\alpha + x)^2 - x^2) &= 4C_\alpha x + 2C_\alpha^2,\\
\frac{\dot{u}_x}{\dot{x}}(x, C_\beta - x) - \frac{\dot{u}_y}{\dot{y}}(x, C_\beta - x) &= 2((C_\beta - x)^2 - x^2) &= -4C_\beta x + 2C_\beta^2.
\end{aligned}$$

Desweiteren gilt

$$\frac{\dot{u}_x}{\dot{x}} = \frac{\mathrm{d}u_x}{\mathrm{d}x}, \quad \frac{\dot{u}_y}{\dot{y}} = \frac{\mathrm{d}u_y}{\mathrm{d}y} = \frac{\mathrm{d}u_y}{\mathrm{d}x}\cdot\frac{\mathrm{d}x}{\mathrm{d}y} = \frac{\mathrm{d}u_y}{\mathrm{d}x}\cdot\frac{1}{y'}.$$

Integration bezüglich x ergibt

$$\begin{aligned}
u_x(x, C_\alpha + x) - u_y(x, C_\alpha + x) &= 2C_\alpha x^2 + 2C_\alpha^2 x + C_\alpha^*,\\
u_x(x, C_\beta - x) + u_y(x, C_\beta - x) &= -2C_\beta x^2 + 2C_\beta^2 x + C_\beta^*.
\end{aligned}$$

mit Konstanten $C_\alpha^*, C_\beta^* \in \mathbb{R}$. Die Anfangswerte liefern $u_x(0,y)=0$, $u_y(0,y)=2y$, was die Konstanten identifiziert als $C_\alpha^*=-2C_\alpha$, $C_\beta^*=2C_\beta$. In einem beliebigen Punkt $(\bar{x},\bar{y})$ schneiden sich zwei charakteristische Kuven $\bar{y}=C_\alpha+\bar{x}$ aus $\mathcal{K}_\alpha$ und $\bar{y}=C_\beta-\bar{x}$ aus $\mathcal{K}_\beta$. Es folgt

$$C_\alpha = \bar{y} - \bar{x} \quad \text{und} \quad C_\beta = \bar{y} + \bar{x}.$$

Wir erhalten die Gleichungen

$$\begin{aligned}
u_x(\bar{x},\bar{y}) - u_y(\bar{x},\bar{y}) &= 2(\bar{y}-\bar{x})\bar{x}^2 + 2(\bar{y}-\bar{x})^2\bar{x} - 2(\bar{y}-\bar{x}),\\
u_x(\bar{x},\bar{y}) + u_y(\bar{x},\bar{y}) &= -2(\bar{y}+\bar{x})\bar{x}^2 + 2(\bar{y}+\bar{x})^2\bar{x} + 2(\bar{y}+\bar{x}).
\end{aligned}$$

Wir lösen dieses lineare Gleichungssystem für u_x, u_y direkt und bekommen (mit Umbenennung $\bar{x},\bar{y}$ zu x,y)

$$u_x(x,y) = 2x(1+y^2), \qquad u_y(x,y) = 2y(1+x^2).$$

Wir erhalten die Lösung u durch Integration

$$u(x,y) = u(x_0,y_0) + \int_{\mathcal{J}} u_x\,\mathrm{d}x + u_y\,\mathrm{d}y$$

entlang einer beliebigen Kurve $\mathcal{J}$, welche (x_0, y_0) und (x, y) verbindet. Wir verwenden

$$\begin{aligned} u(x,y) = u(0,y) + \int_0^x u_x(s,y)\,\mathrm{d}s \;&=\; y^2 + \int_0^x 2s(1+y^2)\,\mathrm{d}s \\ &= y^2 + x^2(1+y^2). \end{aligned}$$

Dabei haben wir eine spezielle Kurve $\mathcal{J}$ gewählt, die auf eine einfache Rechnung führt. Hier könnte auch eine Kurve $\mathcal{K}_\alpha$ oder $\mathcal{K}_\beta$ verwendet werden. Es ist leicht zu verifizieren, dass die Funktion u das Cauchy-Problem der Dgl. erfüllt.

Jetzt konstruieren wir ein numerisches Verfahren um automatisch Näherungen zu berechnen. Wir betrachten eine Kurve $\mathcal{K}$, welche ein Cauchy-Problem (4.13) spezifiziert. Wir nehmen an, dass die Kurve $\mathcal{K}$ niemals tangential zu einer charakteristischen Kurve der hyperbolischen Dgl. (4.12) liegt. Auf der Anfangskurve wählen wir die Punkte $P_1, \ldots, P_n$. Dann seien $x(P_j), y(P_j)$ die Koordinaten dieser Punkte. Die Werte $u(P_j)$, $u_x(P_j)$, $u_y(P_j)$ für $j = 1, \ldots, n$ folgen aus den Anfangsvorgaben.

Im Falle konstanter Koeffizienten A, B, C besteht jede Charakteristikenfamilie aus einer Schar von Geraden, siehe Abbildung 15. Im Falle nichtkonstanter Koeffizienten A, B, C sind die Charakteristiken allgemeine Kurven, siehe Abbildung 16. Sei $\mathcal{K}_\alpha^{(j)}$ die charakteristische Kurve aus der ersten Familie und $\mathcal{K}_\beta^{(j)}$ die charakteristische Kurve aus der zweiten Familie, welche beide durch den Punkt P_j verlaufen. Der Schnittpunkt von $\mathcal{K}_\alpha^{(j)}$ durch P_j und $\mathcal{K}_\beta^{(j+1)}$ durch P_{j+1} ergibt einen neuen Punkt Q_j für $j = 1, \ldots, n-1$. Wir beschreiben wie die Daten $x(Q_1)$, $y(Q_1)$, $u(Q_1)$, $u_x(Q_1)$, $u_y(Q_1)$ berechnet werden können aus den entsprechenden Daten in P_1 und P_2. Sukzessive können dann die Daten in anderen Punkten aus dem charakterischen Gitter bestimmt werden.

Wir diskretisieren die gewöhnlichen Dgln. $\dot{y} = \alpha\dot{x}$ und $\dot{y} = \beta\dot{x}$ aus den beiden Charakteristikenfamilien, siehe (4.19). Der Hauptsatz der Differential-

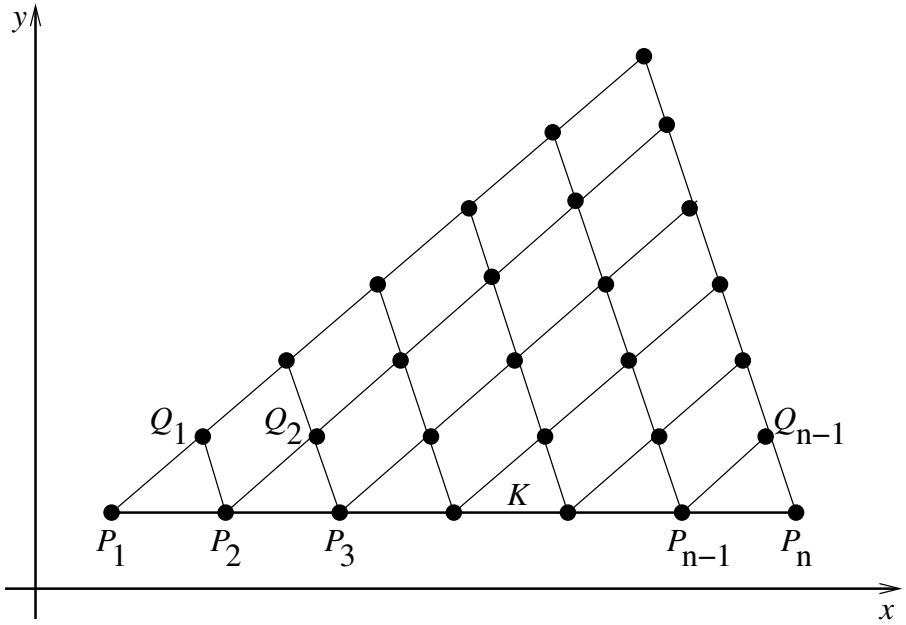


Abbildung 15: Gitter in Charakteristikenverfahren für Dgl. mit konstanten Koeffizienten.

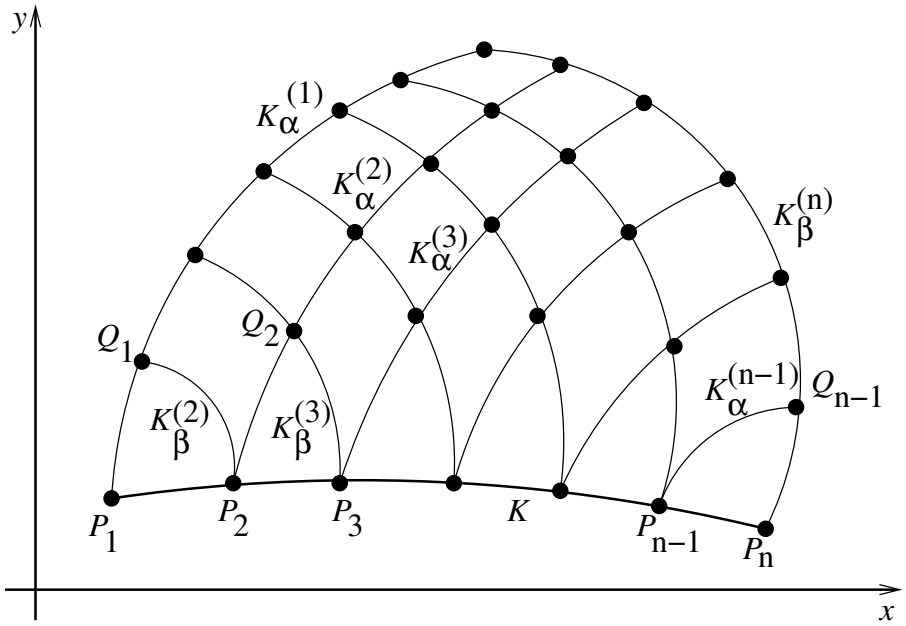


Abbildung 16: Gitter in Charakteristikenverfahren für Dgl. mit nichtkonstanten Koeffizienten.

Integralrechnung und eine Substitution liefern

$$\begin{aligned} y(Q_1) - y(P_1) &= \int_{\tau(P_1)}^{\tau(Q_1)} \dot{y} \, \mathrm{d}\tau = \int_{\tau(P_1)}^{\tau(Q_1)} \alpha \, \dot{x} \, \mathrm{d}\tau = \int_{x(P_1)}^{x(Q_1)} \alpha \, \mathrm{d}x, \\ y(Q_1) - y(P_2) &= \int_{\tau(P_2)}^{\tau(Q_1)} \dot{y} \, \mathrm{d}\tau = \int_{\tau(P_2)}^{\tau(Q_1)} \beta \, \dot{x} \, \mathrm{d}\tau = \int_{x(P_2)}^{x(Q_1)} \beta \, \mathrm{d}x. \end{aligned} \tag{4.21}$$

Wir approximieren die obigen Integrale durch die (linksseitige) Rechteckregel. Es folgt

$$\begin{aligned} y(Q_1) - y(P_1) &= \alpha(P_1)(x(Q_1) - x(P_1)), \\ y(Q_1) - y(P_2) &= \beta(P_2)(x(Q_1) - x(P_2)). \end{aligned} \tag{4.22}$$

Wegen $\alpha = \alpha(A, B, C)$ gilt

$$\alpha(P_1) = \alpha(A(x(P_1), y(P_1)), B(x(P_1), y(P_1)), C(x(P_1), y(P_1)))$$

oder in kürzerer Form $\alpha(P_1) = \alpha(x(P_1), y(P_1))$ und entsprechend für β. Dadurch können wir direkt $\alpha(P_1), \beta(P_2)$ auswerten. Wir erhalten ein lineares Gleichungssystem (4.22) für $x(Q_1), y(Q_1)$, welches direkt gelöst werden kann

$$\begin{aligned} x(Q_1) &= \frac{y(P_2) - y(P_1) + \alpha(P_1)x(P_1) - \beta(P_2)x(P_2)}{\alpha(P_1) - \beta(P_2)}, \\ y(Q_1) &= \frac{\alpha(P_1)y(P_2) - \beta(P_2)y(P_1) + \alpha(P_1)\beta(P_2)(x(P_1) - x(P_2))}{\alpha(P_1) - \beta(P_2)}. \end{aligned}$$

Es gilt $\alpha(P_j) \neq \beta(P_j)$ für alle j. Somit ist $\alpha(P_1) \neq \beta(P_2)$ erfüllt für P_2 hinreichend nahe bei P_1 wegen der Stetigkeit der Funktionen. Wir erhalten schließlich $u(Q_1)$ durch eine Näherung erster Ordnung aus einer Taylor-Entwicklung

$$u(Q_1) = u(P_1) + u_x(P_1)(x(Q_1) - x(P_1)) + u_y(P_1)(y(Q_1) - y(P_1)).$$

Um dieses Verfahren in den anderen Gitterpunkten fortzusetzen, benötigen wir auch Näherungen für $u_x(Q_1), u_y(Q_1)$. Die Gleichungen (4.20) erlauben eine Bestimmung von u_x, u_y. Wir verwenden wieder eine Diskretisierung

ähnlich der Rechteckregel für Integrale

$$\begin{aligned} &A(P_1)\alpha(P_1)(u_x(Q_1) - u_x(P_1)) + C(P_1)(u_y(Q_1) - u_y(P_1)) \\ &= f(P_1)(y(Q_1) - y(P_1)), \\ &A(P_2)\beta(P_2)(u_x(Q_1) - u_x(P_2)) + C(P_2)(u_y(Q_1) - u_y(P_2)) \\ &= f(P_2)(y(Q_1) - y(P_2)), \end{aligned} \tag{4.23}$$

wobei

$$f(P_j) = f(x(P_j), y(P_j), u(P_j), u_x(P_j), u_y(P_j)).$$

Die Gleichungen (4.23) bilden ein lineares Gleichungssystem für die Unbekannten $u_x(Q_1)$ und $u_y(Q_1)$, wobei die Koeffizientenmatrix

$$G = \begin{pmatrix} A(P_1)\alpha(P_1) & C(P_1) \\ A(P_2)\beta(P_2) & C(P_2) \end{pmatrix}$$

auftritt. Wir erhalten die Determinante

$$\det G = A(P_1)C(P_2)\alpha(P_1) - A(P_2)C(P_1)\beta(P_2).$$

Daher ist $\det G \neq 0$ garantiert für P_1, P_2 hinreichend nahe beieiander (unter der Annahme $A, C \neq 0$). Daher erhalten wir $u_x(Q_1), u_y(Q_1)$ aus dem linearen Gleichungssystem (4.23).

Im quasi-linearen Fall $A = A(x, y, u, u_x, u_y)$, $B = \cdots$, $C = \cdots$ kann dieses Verfahren mit den gleichen Formeln angewendet werden, weil die Daten

$$A(P_j) = A(x(P_j), y(P_j), u(P_j), u_x(P_j), u_y(P_j)), \quad \text{etc.}$$

bekannt sind.

Nun möchten wir ein Charakteristikenverfahren konstruieren, welches konsistent mit Ordnung 2 ist. Dadurch werden die gewöhnlichen Dgln. mit der Trapezregel diskretisiert, d.h. einer impliziten Methode. Für die beiden Charakteristikenfamilien gegeben durch $\dot{y} = \alpha\dot{x}$ und $\dot{y} = \beta\dot{x}$ approximieren wir die Integrale (4.21) mit der Trapezregel. Es folgt

$$\begin{aligned} y(Q_1) - y(P_1) &= \tfrac{1}{2}(\alpha(P_1) + \alpha(Q_1))(x(Q_1) - x(P_1)), \\ y(Q_1) - y(P_2) &= \tfrac{1}{2}(\beta(P_2) + \beta(Q_1))(x(Q_1) - x(P_2)). \end{aligned} \tag{4.24}$$

Da $\alpha(Q_1) = \alpha(x(Q_1), y(Q_1))$ und $\beta(Q_1) = \beta(x(Q_1), y(Q_1))$ gilt, erhalten wir ein nichtlineares Gleichungssystem (4.24) für die Unbekannten $x(Q_1), y(Q_1)$. Das Newton-Verfahren generiert eine Näherungslösung. Wenn eine Näherungslösung $x(Q_1), y(Q_1)$ zum System (4.24) vorliegt, dann können die Terme $\alpha(Q_1), \beta(Q_1)$ ausgewertet werden.

Wir verwenden wieder die Gleichungen (4.20) zur Bestimmung von u_x, u_y. Eine Diskretisierung zweiter Ordnung liefert

$$\begin{aligned}
&(A(P_1)\alpha(P_1) + A(Q_1)\alpha(Q_1))(u_x(Q_1) - u_x(P_1)) \\
&+ (C(P_1) + C(Q_1))(u_y(Q_1) - u_y(P_1)) \\
&= (f(P_1) + f(Q_1))(y(Q_1) - y(P_1)), \\
&(A(P_2)\beta(P_2) + A(Q_1)\beta(Q_1))(u_x(Q_1) - u_x(P_2)) \\
&+ (C(P_2) + C(Q_1))(u_y(Q_1) - u_y(P_2)) \\
&= (f(P_2) + f(Q_1))(y(Q_1) - y(P_2)).
\end{aligned} \tag{4.25}$$

Für $f = f(x, y)$ ergibt sich die Auswertung $f(Q_1)$ aus der Näherungslösung $x(Q_1), y(Q_1)$ des nichtlinearen Gleichungssystems (4.24). Dementsprechend ist auch $\alpha(Q_1), \beta(Q_1)$ durch $x(Q_1), y(Q_1)$ bekannt. Wir erhalten erneut ein lineares Gleichungssystem (4.25) für die Unbekannten $u_x(Q_1), u_y(Q_1)$. Es kann gezeigt werden, dass die Koeffizientenmatrix regulär ist falls P_1, P_2 hinreichend nahe beieinanderliegen. Eine Formel für die unbekannten Ableitungswerte $u_x(Q_1), u_y(Q_1)$ kann aus der Cramerschen Regel erhalten werden. Allgemein gilt

$$\dot{u} = u_x \dot{x} + u_y \dot{y}.$$

Schließlich erfüllt die Lösung die Gleichung

$$u(Q_1) = u(P_1) + \int_{\mathcal{K}_\alpha^{(1)}} \dot{u} \, \mathrm{d}\tau = u(P_1) + \int_{\mathcal{K}_\alpha^{(1)}} u_x \, \mathrm{d}x + u_y \, \mathrm{d}y$$

unter Verwendung der charakteristischen Kurve $\mathcal{K}_\alpha^{(1)}$ von P_1 nach Q_1. Die Trapezregel produziert die Näherung

$$u(Q_1) = u(P_1) + \tfrac{1}{2}(u_x(P_1) + u_x(Q_1))(x(Q_1) - x(P_1)) \tag{4.26}$$

$$+ \tfrac{1}{2}(u_y(P_1) + u_y(Q_1))(y(Q_1) - y(P_1)), \tag{4.27}$$

welche konsistent von Ordnung 2 ist.

Im semi-linearen Fall $f = f(x, y, u, u_x, u_y)$ lösen wir die Gleichungen (4.25) zusammen mit (4.26) für die Unbekannten $u(Q_1), u_x(Q_1), u_y(Q_1)$, welches im allgemeinen ein nichtlineares Gleichungssystem darstellt.

Im quasi-linearen Fall $A = A(x, y, u, u_x, u_y)$, $B = \cdots$, $C = \cdots$, lösen wir ein nichtlineares Gleichungssystem bestehend aus (4.24), (4.25), (4.26) für die fünf Unbekannten $x(Q_1), y(Q_1), u(Q_1), u_x(Q_1), u_y(Q_1)$.

Im linearen Fall (auch mit nichtkonstanten Koeffizienten) können alle Gitterpunkte a priori berechnet werden ohne Kenntnis von u, u_x, u_y. Jedoch stellt dies keinen wesentlichen Vorteil dar. Im quasi-linearen Fall müssen die Gitterpunkte zusammen mit u, u_x, u_y sukzessive berechnet werden.