

Numerische Lösung partieller Differentialgleichungen

Peter Bastian

email: `Peter.Bastian@ipvs.uni-stuttgart.de`

29. Mai 2008

`$Id:main.tex4872008-01-2915:03:01Zbastian$`

Universität Stuttgart, Institut für Parallele und Verteilte Systeme,
Universitätsstraße 38, D-70569 Stuttgart

Inhaltsverzeichnis

1	Modellierung mit partiellen Differentialgleichungen	5
1.1	Einleitung	5
1.2	Wiederholung von Begriffen aus der Vektoranalysis	5
1.3	Energieerhaltung	8
1.4	Wärmefluss	9
1.5	Wärmeleitungsgleichung	10
1.6	Weitere Beispiele	11
1.7	Zusammenfassung	15
2	Typeinteilung partieller Differentialgleichungen	17
2.1	Allgemeine Definition	17
2.2	Typeinteilung partieller Differentialgleichungen	17
2.3	Beispiele für verschiedene Typen	19
2.4	Einflussbereich	21
2.5	Zusammenfassung	22
3	Zur Theorie elliptischer partieller Differentialgleichungen	23
3.1	Koordinatentransformation	23
3.2	Fundamentallösung	25
3.3	Grenzen des klassischen Lösungsbegriffes	29
3.4	Separation der Variablen	31
3.5	Mittelwerteigenschaft und Folgen	32
3.6	Lösungsdarstellung mittels Greenscher Funktion	34
3.7	Stabilität	35
3.8	Zusammenfassung	35
4	Differenzenmethode für elliptische Gleichungen	37
4.1	Der eindimensionale Fall	37
4.2	Der n -dimensionale Fall	39
4.3	Neumann Randbedingung	42
4.4	Allgemeine elliptische Gleichung	43
4.5	Zusammenfassung	45
5	M-Matrix-Theorie	47
5.1	Einführende Definitionen	47
5.2	Gerschgorin Kreise und Regularität	48
5.3	Diagonaldominante Matrizen	50
5.4	Zusammenfassung	53
6	Konvergenz des Finite-Differenzen-Verfahrens	55
6.1	Konvergenz	55
6.2	Konsistenz	56
6.3	Stabilität	57
6.4	Diskrete Mittelwerteigenschaft und Maximumprinzip	58

Inhaltsverzeichnis

6.5	Eigenwerte, Eigenvektoren	59
6.6	Zusammenfassung	60
7	Zellenzentrierte Finite Volumen	61
7.1	Problemstellung und Gitterkonstruktion	61
7.2	Finite Volumen	63
7.3	Zellweise Permeabilität	64
7.4	Diskrete Erhaltungseigenschaft	66
7.5	Erweiterung auf unstrukturierte Gitter	67
7.6	Zusammenfassung	68
8	Relaxationsverfahren	69
8.1	Dünn besetzte Matrizen und direkte Lösungsverfahren	69
8.2	Relaxationsverfahren	70
8.3	Matrixschreibweise der Relaxationsverfahren	71
8.4	Konvergenz von linearen Iterationsverfahren	73
8.5	Zusammenfassung	76
9	Abstiegsverfahren	79
9.1	Diagonaldominante Matrizen	79
9.2	Praktische Realisierung; Abbruchkriterium	80
9.3	Abstiegsverfahren	81
9.4	Vorkonditioniertes Gradientenverfahren	83
9.5	Konjugierte Gradienten Verfahren	85
9.6	Zusammenfassung	87
10	Mehrgitterverfahren	89
10.1	Glättungseigenschaft	89
10.2	Prolongation	90
10.3	Restriktion	92
10.4	Grobgitterkorrektur	92
10.5	Zweigitteriteration	93
10.6	Mehrgitterverfahren	93
10.7	Zusammenfassung	95
11	Parabolische partielle Differentialgleichungen	97
11.1	Lösung mittels Fourierreihe	97
11.2	Finite Differenzen für Parabolische Probleme	98
11.3	Fehleranalyse	100
11.4	Numerischer Vergleich der Verfahren	103
11.5	Zusammenfassung	112
12	Finite Differenzen für lineare hyperbolische Gleichungen	113
12.1	Methode der Charakteristiken	113
12.2	Finite Differenzen	115
12.3	Numerischer Vergleich	117

12.4 Numerische Diffusion	126
12.5 Zusammenfassung	127
13 Finite-Volumen-Verfahren für lineare, skalare, hyperbolische Gleichungen	129
13.1 Einführung	129
13.2 Anforderungen an die Flussfunktion	132
13.3 Ein instabiler Fluss	133
13.4 Lax-Friedrich-Verfahren	133
13.5 Upwind-Verfahren	134
13.6 Godunov-Verfahren	135
13.7 Zusammenfassung	135
14 High resolution Schemata für lineare, skalare, hyperbolische Probleme	137
14.1 Verfahren zweiter Ordnung	137
14.2 Höhere Ordnung mit REA	139
14.3 Slope Limiter Verfahren	141
14.4 Numerischer Vergleich	144
14.5 Zusammenfassung	146
15 Nichtlineare Erhaltungsgleichungen	147
15.1 Schwache Lösungen	147
15.2 Bedeutung von FV-Verfahren	152
15.3 Godunov Verfahren im nichtlinearen Fall	153
15.4 Zusammenfassung	154
Literaturverzeichnis	155

Inhaltsverzeichnis

Vorwort

Ziel dieser Vorlesung für Informatiker im Hauptstudium ist es eine kompakte Einführung in die Numerik verschiedener Typen von partiellen Differentialgleichungen zu geben. Bewusst wurde der Wert auf relative einfache (aber hinreichend effektive) Verfahren wie Finite Differenzen und Finite Volumen gelegt um in einer zweistündigen Vorlesungen sowohl elliptische und parabolische als auch hyperbolische Gleichungen behandeln zu können. Jedem Typ wird zunächst ein kurzer Abriss der Theorie vorangestellt um darauf aufbauend die numerischen Verfahren einzuführen.

Erstmals steht im Wintersemester 2007/2008 ein Skript zur Vorlesung und ein Foliensatz zur Verfügung. Für die Erfassung des Textes in \LaTeX und vor allem die hervorragend angefertigten Zeichnungen in \tikZ danke ich Adrian Dempwolff und Jö Fahlke recht herzlich. Alle verbleibenden Fehler gehen natürlich auf mein Konto.

Stuttgart, im Oktober 2007

Peter Bastian

Inhaltsverzeichnis

1 Modellierung mit partiellen Differentialgleichungen

1.1 Einleitung

Partielle Differentialgleichungen sind aus Naturwissenschaft und Technik nicht wegzudenken, siehe [Mar07]. Grundlegend für ihre Herleitung ist dass die gesuchten Größen durch kontinuierliche Funktionen beschrieben werden.

So kann man etwa die Temperatur $T(x, t)$ in einem Körper als Funktion von Raum (x) und Zeit (t) bei vorgegebener Temperatur am Rand bestimmen.

Die Abbildungen 1 und 2 zeigen zwei verschieden geformte Körper. Der Körper in Abbildung 1 kann in guter Näherung durch zwei Koordinaten beschrieben werden.

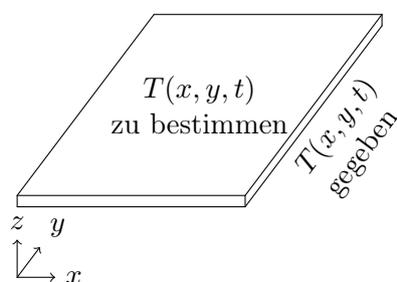


Abbildung 1: „dünne“ Metallplatte, Variation in z klein

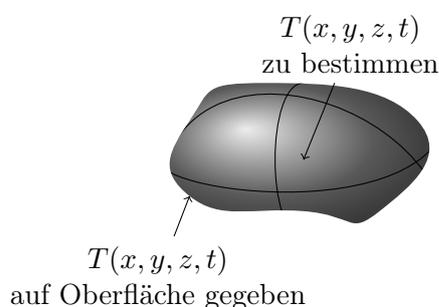


Abbildung 2: beliebig geformter Metallklumpen

1.2 Wiederholung von Begriffen aus der Vektoranalysis¹

Betrachte allgemein Funktionen

$$f : \mathbb{R}^n \rightarrow \mathbb{R}^m$$

n ist die Raumdimension.

$m = 1$: „skalares Feld“, skalare Funktion

$m > 1$: „Vektorfeld“, vektorwertige Funktion

Schreibweisen:

$$f(x, y), \quad f(x, y, z), \quad f(x_1, x_2, \dots, x_n), \quad f(x), \quad x \in \mathbb{R}^n$$
$$f(x) = (f_1(x_1, \dots, x_n), \dots, f_m(x_1, \dots, x_n))^T.$$

¹Mathe für Informatiker II

1 Modellierung mit partiellen Differentialgleichungen

Ob x , f ein Vektor oder ein Skalar ist sollte jeweils aus dem Zusammenhang klar werden. Es wird keine spezielle Auszeichnung wie Unterstreichen oder Fettdruck benutzt.

Sei $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Definiere die partielle Ableitung von f nach x_i :

$$\frac{\partial f}{\partial x_i}(x_1, \dots, x_n) = \lim_{h \rightarrow 0} \frac{f(x_1, \dots, x_i + h, \dots, x_n) - f(x_1, \dots, x_n)}{h}$$

Betrachte die übrigen Variablen x_j , $j \neq i$ als „Parameter“.

Bei Vektorfeldern: Differenzieren der einzelnen Komponenten $\frac{\partial f_j}{\partial x_i}(x)$.

Entsprechend kann man auch höhere Ableitungen bilden. Schreibweise:

$$\frac{\partial^2 f}{\partial x_i^2}(x) = \frac{\partial}{\partial x_i} \left[\frac{\partial}{\partial x_i} f(x) \right] \quad \text{oder} \quad \frac{\partial^2 f}{\partial x_i \partial x_j}(x) = \frac{\partial^2 f}{\partial x_j \partial x_i}(x)$$

„gemischte Ableitung“

Andere, sparsamere Schreibweisen für $\frac{\partial f}{\partial x_i}$ sind $\partial_{x_i} f$ oder f_{x_i} .

Höhere Ableitungen schreibt man dann als $\partial_{x_i} \partial_{x_j} f$ bzw. $f_{x_i x_j}$.

Beispiel 1.1. Betrachte $f(x) = \left(\sum_{j=1}^n x_j^2 \right)^{-\frac{1}{2}} = \|x\|^{-1}$ mit $\|x\| = \left(\sum_{j=1}^n x_j^2 \right)^{\frac{1}{2}}$ „euklidische² Norm“.

$$\begin{aligned} \frac{\partial f}{\partial x_i}(x) &= -\frac{1}{2} \left(\sum_{j=1}^n x_j^2 \right)^{-\frac{3}{2}} \cdot 2x_i = -x_i \left(\sum_{j=1}^n x_j^2 \right)^{-\frac{3}{2}} \\ \frac{\partial^2 f}{\partial x_i^2}(x) &= - \left(\sum_{j=1}^n x_j^2 \right)^{-\frac{3}{2}} + x_i \frac{3}{2} \left(\sum_{j=1}^n x_j^2 \right)^{-\frac{5}{2}} \cdot 2x_i \\ &= 3x_i^2 \left(\sum_{j=1}^n x_j^2 \right)^{-\frac{5}{2}} - \left(\sum_{j=1}^n x_j^2 \right)^{-\frac{3}{2}} \end{aligned}$$

□

Gegeben sei $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $\frac{\partial f}{\partial x_i}(x)$ existiere für alle $i = 1, \dots, n$.

Die vektorwertige Funktion

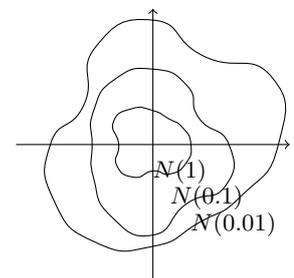
$$g(x) = \left(\frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_n}(x) \right)^T$$

heißt *Gradient* von f .

Kurz schreibt man $g(x) = \nabla f(x)$.

$N(c) = \{x \in \mathbb{R}^n \mid f(x) = c\}$ heißt Niveaulinie (-fläche).

$\nabla f(x)$ steht senkrecht auf $N(f(x))$ und „zeigt“ in Richtung des *größten Anstiegs* von f am Punkt x .



²Euklid von Alexandria, ca. 365-300 v. Chr., griech. Mathematiker.

$$\nabla f(x) = - \underbrace{\left(\sum_{j=1}^n x_j^2 \right)^{-\frac{3}{2}}}_{\text{Länge 1}} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \frac{1}{\|x\|^2} \underbrace{\begin{pmatrix} -x \\ \vdots \\ -x \end{pmatrix}}_{\text{Länge 1}}.$$

Beispiel 1.2 (Fortsetzung von Beispiel 1.1).

Sei $n = 2$. $f(x) = \frac{1}{r}$; r : Abstand vom Ursprung. Niveaulinien sind konzentrische Kreise. □

Oft betrachtet man nicht Funktionen auf ganz \mathbb{R}^n sondern nur einer Teilmenge.

Definition 1.3 (Gebiet). $\Omega \subseteq \mathbb{R}^n$ heißt Gebiet falls Ω offen und zusammenhängend.

offen: Zu $x \in \Omega$ gibt es $B_\epsilon(x) = \{y \in \Omega \mid \|x - y\| < \epsilon\}$ so dass $B_\epsilon(x) \subseteq \Omega$ für ϵ genügend klein.

zusammenhängend: $x, y \in \Omega$, dann gibt es eine stetige Kurve $t(s) : [0, 1] \rightarrow \Omega$ mit $t(0) = x$, $t(1) = y$, $t(s) \in \Omega$.

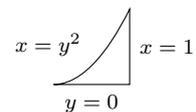
Mit $\bar{\Omega}$ bezeichnet man den Abschluss von Ω , also Ω plus die Grenzwerte aller Folgen, die man mit Elementen aus Ω bilden kann.

$\partial\Omega = \bar{\Omega} \setminus \Omega$ ist dann der Rand von Ω . Oft benötigt man zusätzliche Bedingungen an die Glattheit des Randes.

Schließlich bezeichnet $\nu(x)$ die äußere Einheitsnormale in einem Punkt $x \in \partial\Omega$. □

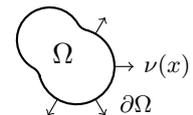
Gegeben sei ein Vektorfeld $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ und $\frac{\partial f_i}{\partial x_i}$ sei wohl definiert.

Ω sei ein Gebiet mit stückweise glattem Rand, das die „Kegelbedingung“ erfüllt.



Dann gilt der Gaußsche³ Integralsatz:

$$\underbrace{\int_{\Omega} \sum_{i=1}^n \frac{\partial f_i(x)}{\partial x_i} dx}_{\text{Volumenintegral}} = \underbrace{\int_{\partial\Omega} f(x) \cdot \nu(x) ds}_{\text{Oberflächenintegral}}$$



Dies entspricht dem Hauptsatz der Differential- und Integralrechnung in mehreren Raumdimensionen, also $\int_a^b f'(x) dx = f(b) - f(a)$.

Für ein Vektorfeld $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ heißt $\sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x)$ die *Divergenz* in x .

Kurz schreibt man

$$\nabla \cdot f(x) = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{pmatrix} \cdot \begin{pmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{pmatrix} = \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x).$$

³Carl Friedrich Gauß, 1777-1855, dt. Mathematiker.

1 Modellierung mit partiellen Differentialgleichungen

Beweis: [Fey70, Abschnitt 3-3] oder [Smi90, Nr. 72].

Der Gaußsche Integralsatz ist ein Spezialfall der partiellen Integration:

$$\int_{\Omega} \nabla \cdot f(x)g(x) \, dx = - \int_{\Omega} f(x) \cdot \nabla g(x) \, dx + \int_{\partial\Omega} f(x) \cdot \nu(x)g(x) \, ds$$

$f(x)$ ist ein Vektorfeld und $g(x)$ eine skalare Funktion.

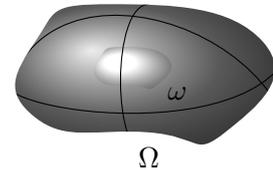
Für $g \equiv 1$ erhält man

$$\int_{\Omega} \nabla \cdot f(x) \, dx = \int_{\partial\Omega} f(x) \cdot \nu(x) \, ds \quad .$$

1.3 Energieerhaltung

Nun zurück zur Modellierung der Temperaturverteilung, siehe auch [Fey70, p. 2-8, p. 3-4].

Sei $\Omega \subseteq \mathbb{R}^3$ unser Körper und $\omega \subseteq \Omega$ ein *beliebiger* Teilbereich des Körpers, sowie $T(x, t)$ die Temperatur in x zur Zeit t (Kontinuums-hypothese).



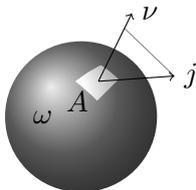
In $\omega \subseteq \Omega$ ist zur Zeit t eine bestimmte Energie $Q_{\omega}(t)$ in Form von Wärme gespeichert. Dabei gilt⁴

$$\underbrace{Q_{\omega}(t)}_{\substack{\text{Wärmeenergie in} \\ \omega}} = \int_{\omega} \underbrace{c}_{\substack{\text{spez. Wärmekapazität} \\ [\frac{\text{J}}{\text{K kg}}]}} \underbrace{\rho(x)}_{\substack{\text{Massen-} \\ \text{dichte} \\ [\frac{\text{kg}}{\text{m}^3}]}} \underbrace{T(x, t)}_{\substack{\text{abs.} \\ \text{Temperatur} \\ [\text{K}]}} \, dx$$

Nun betrachten wir die *zeitliche* Änderung von $Q_{\omega}(t)$ in einem Zeitintervall $[t, t + \Delta t]$ für ein *beliebiges* $\omega \subseteq \Omega$:

$$\underbrace{Q_{\omega}(t + \Delta t) - Q_{\omega}(t)}_{\substack{\text{Änderung der} \\ \text{Wärmeenergie in } \omega}} = \underbrace{\{\text{Zu-/Abfluss von Wärme über Oberfläche } \partial\omega\}}_{\text{positiv: Wärme fließt raus, da } \nu \text{ nach außen}} + \underbrace{\{\text{Zu-/Abfuhr von Wärme über Quellen/Senken in } \omega\}}_{\text{positiv: Wärme fließt zu}}$$

In Formeln:



$$Q_{\omega}(t + \Delta t) - Q_{\omega}(t) = \int_t^{t+\Delta t} \left\{ - \int_{\partial\omega} \underbrace{j(x, t)}_{\substack{\text{Fluss} \\ [\frac{\text{J}}{\text{s m}^2}]}}} \cdot \nu(x) \, ds + \int_{\omega} \underbrace{q(x, t)}_{\substack{\text{Quelle} \\ [\frac{\text{J}}{\text{s m}^3}]}} \, dx \right\} dt$$

⁴Integrale Form von $Q = cmT$ (Wärmeenergie prop. zu T)

Einsetzen der Wärmeenergie, Approximation des Integrals rechts mit Ordnung 1, Gaußscher Integralsatz und Umstellen ergeben:

$$\int_{\omega} \frac{c\rho T(x, t + \Delta t) - c\rho T(x, t)}{\Delta t} dx = - \int_{\omega} \nabla \cdot j(x, t) dx + \int_{\omega} q(x, t) dx.$$

Wir bilden den Limes $\Delta t \rightarrow 0$ und erhalten (alles sei genügend „glatt“):

$$\int_{\omega} \left[\frac{\partial(c\rho T)}{\partial t}(x, t) + \nabla \cdot j(x, t) - q(x, t) \right] dx = 0 \quad .$$

Da der Integrationsbereich ω *beliebig* war folgert man, dass der Integrand punktweise verschwinden muss.

Damit ergibt sich dann

$$\frac{\partial(c\rho T)}{\partial t}(x, t) + \nabla \cdot j(x, t) = q(x, t) \quad \text{für alle } x \in \Omega. \quad (1.1)$$

wichtig!

Diese partielle Differentialgleichung beschreibt die Energieerhaltung in allgemeiner Form. Sie ist sehr typisch für die mathematische Physik und tritt in ähnlicher Form auch für andere Erhaltungsgrößen wie Masse und Impuls auf.

Im stationären Fall gilt $\frac{\partial(c\rho T)}{\partial t} = 0$ und (1.1) reduziert sich auf

$$\nabla \cdot j(x) = q(x) \quad \forall x \in \Omega. \quad (1.2)$$

j und q hängen nun nicht mehr von T ab.

Ist $q \equiv 0$ so erhält man $\nabla \cdot j(x) = 0 \quad \forall x \in \Omega$. Man sagt das Vektorfeld j ist „*divergenzfrei*“. Dies bedeutet, dass keine Zu-/Abfuhr von Wärme in das Gebiet erfolgt!

1.4 Wärmefluss

Bleibt noch die Bestimmung des Wärmeflusses $j(x, t)$. Hier unterscheidet man zwei Fälle:

Konduktion oder auch Wärmeleitung. Temperatur bedeutet kinetische Energie der Atome, diese wird z. B. durch Stöße abgegeben/aufgenommen.

Konvektion Wärmetransport durch mittlere Drift der Atome/Moleküle in einem Fluid.

1 Modellierung mit partiellen Differentialgleichungen

Konduktiver Fluss Bestehen in dem Körper *Temperaturunterschiede* so werden diese ausgeglichen. Wärme fließt von warm nach kalt. Ein mögliches Modell ist

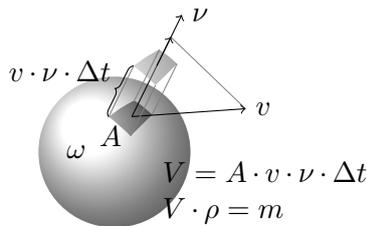
$$j_d(x, t) = \underbrace{\lambda}_{\substack{\text{Wärmeleitfähigkeit} \\ \left[\frac{\text{J}}{\text{s m K}}\right]}} (-\nabla T(x, t)) \quad (1.3)$$

$\nabla T(x, t)$ am Punkt x zeigt in Richtung des größten Anstiegs, $-\nabla T$ in R. d. steilsten Abstiegs. $\left[\frac{\text{J}}{\text{s m K}}\right] \cdot \left[\frac{\text{K}}{\text{m}}\right] \rightarrow \left[\frac{\text{J}}{\text{s m}^2}\right]$ wie gefordert.

(1.3) heißt Fourier'sches⁵ Gesetz. Dies ist eine mehr oder weniger gute Näherung (Modellfehler!).

Konvektiver Fluss Bei der Wärmeleitung findet *kein* Transport von Materie statt. In einem Festkörper „zittern“ die Atome um ihre Ruhelage. Diese Bewegung wird an die Nachbaratome weitergegeben.

Konvektion dagegen bezeichnet den Transport von Wärme durch Bewegung der Atome (Moleküle) eines *Fluids* von einer Stelle zu einer anderen.



$$j_c(x, t) = \underbrace{c}_{\left[\frac{\text{J}}{\text{kg K}}\right]} \underbrace{\rho(x, t)}_{\left[\frac{\text{kg}}{\text{m}^3}\right]} \underbrace{v(x, t)}_{\left[\frac{\text{m}}{\text{s}}\right]} \underbrace{T(x, t)}_{[\text{K}]} \quad (1.4)$$

Konvektiver und konduktiver Fluss addieren sich zum Gesamtfluss

$$j(x, t) = j_c(x, t) + j_d(x, t). \quad (1.5)$$

1.5 Wärmeleitungsgleichung

Einsetzen des Flusses (1.5) in die Energieerhaltungsgleichung (1.1) ergibt die lineare partielle Differentialgleichung

$$\frac{\partial(c\rho T)}{\partial t} + \nabla \cdot \{c\rho v T - \lambda \nabla T\} = q \quad \forall x \in \Omega \text{ und } t \in [t_a, t_b]. \quad (1.6)$$

Die Gleichung (1.6) nennt man Wärmeleitungsgleichung.

Um $T(x, t)$ eindeutig festzulegen benötigt man noch

$$\begin{aligned} \text{Anfangsbedingungen: } & T(x, t_a) = T_o(x) \quad \text{und} \\ \text{Randbedingungen: } & T(x, t) = g(x, t) \quad x \in \partial\Omega, t \in [t_a, t_b]. \end{aligned} \quad (1.7)$$

⁵Jean Baptiste Joseph Fourier, 1768-1830, frz. Mathematiker und Physiker.



Abbildung 3: Illustration der Konvektions-Diffusions-Gleichung an einem herbstlichen Blatt. Chlorophyll wird in der Nähe der Rippen schneller abgebaut als weiter weg im „Inneren“.

Bemerkung 1.4. (1.6) gilt mit obigen Einheiten für $n = 3$. Den Fall $n = 2$ (dünne Platte) bzw. $n = 1$ (langer Stab) erhält man durch $\frac{\partial T}{\partial z} = 0$ bzw. $\frac{\partial T}{\partial y} = \frac{\partial T}{\partial z} = 0$. \square

Allgemein bezeichnet man die Gleichung

$$\frac{\partial C(x, t)}{\partial t} + \nabla \cdot \{uC(x, t) - D\nabla C(x, t)\} = q \quad \forall x \in \Omega \text{ und } t \in [t_a, t_b] \quad (1.8)$$

als Konvektions-Diffusions-Gleichung. Sie beschreibt die Konzentration $C(x, t)$ eines gelösten Stoffes in einer Strömung.

Schön wird dies durch die herbstliche Färbung eines Blattes illustriert.

1.6 Weitere Beispiele

Poisson-Gleichung Sei nun:

$$\frac{\partial T}{\partial t} = 0 \quad (\text{stationär}), \quad v = 0 \quad (\text{keine Konvektion}), \quad \lambda = 1.$$

so reduziert sich (1.6), (1.7) zur sog. Poisson⁶-Gleichung:

$$\begin{aligned} -\nabla \cdot (\nabla T(x)) &= q(x) & x \in \Omega, \\ T(x) &= g(x) & x \in \partial\Omega \end{aligned} \quad (1.9)$$

Es gilt

$$\nabla \cdot (\nabla T(x)) = \nabla \cdot \begin{pmatrix} \frac{\partial T}{\partial x_1}(x) \\ \vdots \\ \frac{\partial T}{\partial x_n}(x) \end{pmatrix} = \sum_{i=1}^n \frac{\partial^2 T}{\partial x_i^2}(x) =: \Delta T(x) \quad .$$

Damit lautet die Poisson-Gleichung auch

$$\begin{aligned} -\Delta T(x) &= q(x) & x \in \Omega, \\ T(x) &= g(x) & x \in \partial\Omega \end{aligned} \quad .$$

Bei $q = 0$ sagt man auch Laplace⁷-Gleichung. Δ heisst Laplace-Operator.

⁶Siméon-Denis Poisson, 1781-1840, fr. Mathematiker und Physiker.

⁷Pierre Simon de Laplace, 1749-1827, frz. Mathematiker und Astronom.

1 Modellierung mit partiellen Differentialgleichungen

Elektrostatik $\vec{E}(x)$ beschreibt das elektrische Feld (Kraft per Einheitsladung) und $\Phi(x)$ das elektrostatische Potential.

Es gelten die Gleichungen

$$\nabla \cdot \vec{E}(x) = \frac{\rho(x)}{\varepsilon_0} \quad \left[\frac{\text{N}}{\text{C}} \right]$$

mit

ρ : Ladungsdichte

ε_0 : el. Feldkonstante $8.854 \cdot 10^{-12} \frac{\text{C}^2}{\text{N m}^2}$

und

$$\vec{E} = -\nabla\Phi$$

d. h. Ladung bewegt sich in Richtung des steilsten Abstieg des Potentials.

Zusammen also:

$$\nabla \cdot (-\nabla\Phi) = -\Delta\Phi = \frac{\rho(x)}{\varepsilon_0}$$

Gravitation Hier ist $\vec{F}(x)$ die Kraftdichte in einer verteilten Masse und $\Phi(x)$ das Gravitationspotential.

Wieder gelten die Gleichungen

$$\nabla \cdot \vec{F}(x) = -4\pi\gamma\rho(x), \quad \vec{F} = -\nabla\Phi$$

wobei nun

ρ : Massendichte,

γ : Gravitationskonstante $6.67428 \cdot 10^{-11} \frac{\text{m}^3}{\text{kg s}^2}$.

Zusammen:

$$\Delta\Phi = 4\pi\gamma\rho(x).$$

Der Faktor 4π gilt in (hier angenommenen) 3 Raumdimensionen (siehe Beispiel unten).

Dies ist die Verallgemeinerung von Newtons⁸ Gravitationsgesetz auf nicht punktförmige Massen.

⁸Sir Isaac Newton, 1643-1727, engl. Physiker und Mathematiker.

Grundwassergleichung Hier ist $u(x, t)$ die Strömungsgeschwindigkeit und $p(x, t)$ der Druck.

Die Erhaltung der Masse wird im kompressiblen Fall beschrieben durch

$$\frac{\partial \varrho(x, t)}{\partial t} + \nabla \cdot \{ \varrho(x, t) u(x, t) \} = f \quad \text{in } \Omega \times \Sigma. \quad (1.10)$$

$\varrho(x, t)$ Massendichte, Σ Zeitintervall.

Das sog. *Darcy*-Gesetz stellt einen Zusammenhang zwischen Geschwindigkeit und Druck her:

$$u(x, t) = -\frac{K(x)}{\mu} (\nabla p(x, t) - \varrho(x, t) G). \quad (1.11)$$

$K(x)$ Permeabilitätstensor, μ dynamische Viskosität, $G = g(0, 0, -1)^T$ Gravitationsvektor.

Im inkompressiblen Fall (ϱ konstant) erhält man

$$-\nabla \cdot \{ K \nabla p \} = \varrho^{-1} \mu f - \nabla \cdot \{ \varrho K G \}. \quad (1.12)$$

Gekoppelter Wärmetransport im porösen Medium Wirkt die Änderung der Temperatur auf die Strömung zurück, so sind Strömungs- und Wärmetransportgleichung gekoppelt zu lösen.

Bei inkompressibler Strömung in einem porösen Medium erhält man:

$$\nabla \cdot u = f, \quad u = -\frac{K}{\mu} (\nabla p - \tilde{\varrho}_w(T(x, t)) G) \quad \text{in } \Omega \times \Sigma \quad (1.13)$$

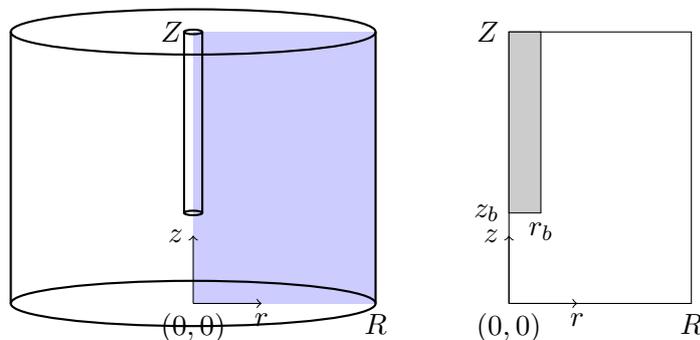
$$\frac{\partial (c_e \varrho_e T)}{\partial t} + \nabla \cdot q + g^- T = g^+, \quad q = c_w \varrho_w u T - \lambda \nabla T \quad \text{in } \Omega \times \Sigma \quad (1.14)$$

für $\Sigma = [t_a, t_b]$, ergänzt um Rand- und Anfangsbedingungen.

$\tilde{\varrho}(T)$ beschreibt die Abhängigkeit der Temperatur von der Dichte. Diese Abhängigkeit wird nur im Auftriebsterm berücksichtigt (Boussinesq-Approximation).

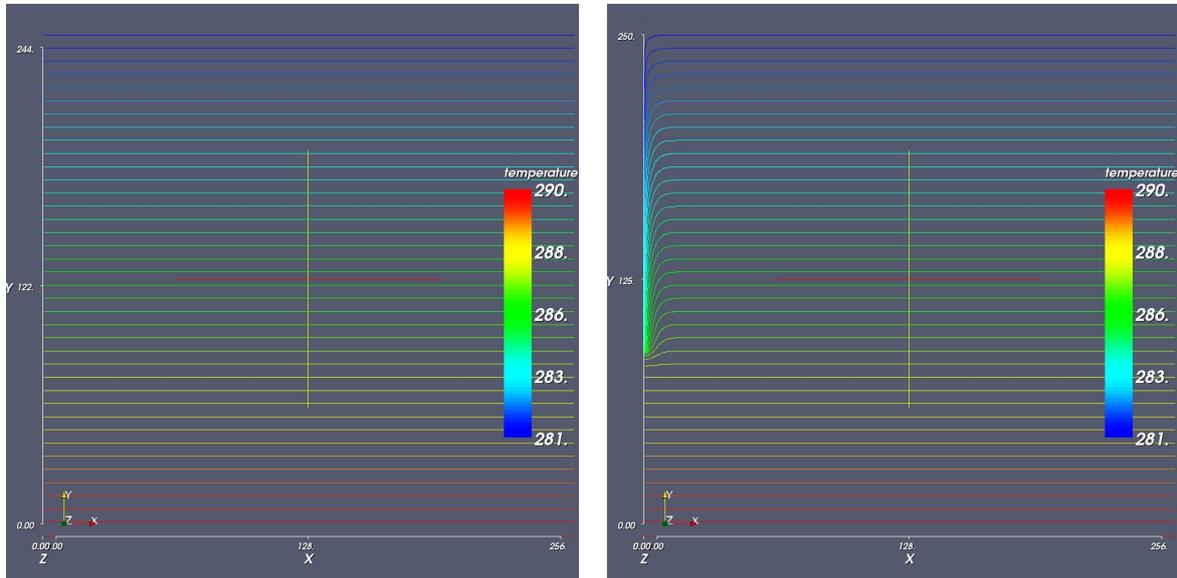
Subskript w bezieht sich auf Wasser, Subskript e auf das wassergesättigte poröse Medium.

Beispiel 1.5 (Geothermie). Das System (1.13),(1.14) beschreibt z. B. eine Geothermieanlage in *offener* Bauweise. In einem Bohrloch fließt Wasser von oben nach unten und nimmt dabei Wärme auf. Das Bohrloch ist nicht gegenüber der Umgebung abgedichtet. In einem isolierten Innenrohr wird das warme Wasser nach oben gepumpt. \square

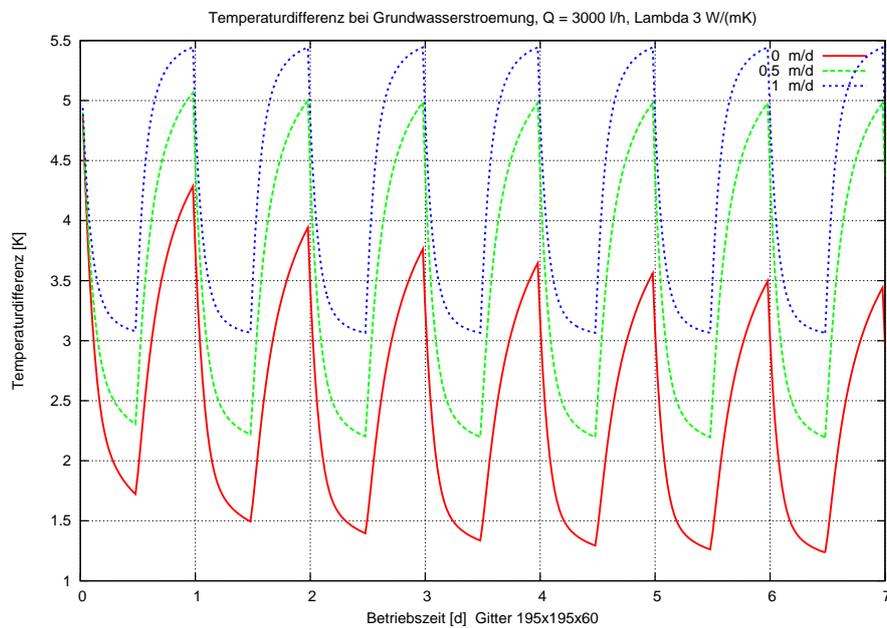


1 Modellierung mit partiellen Differentialgleichungen

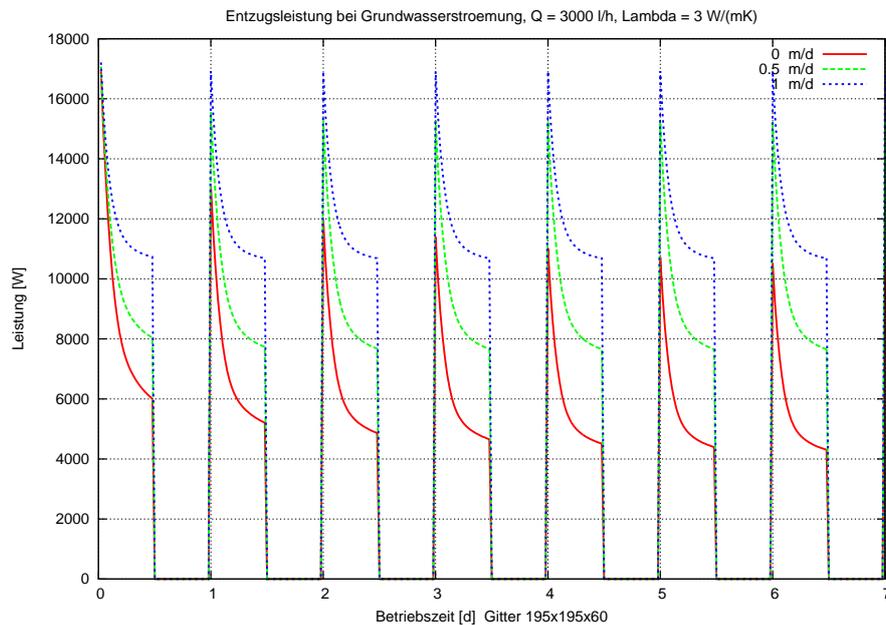
Die Abbildungen zeigen die zylindersymmetrische Geometrie einer solchen Anlage. Hängt Lösung nicht vom Winkel ab, so kann man mit zwei Koordinaten (r, z) auskommen und die Gleichungen entsprechend transformieren.



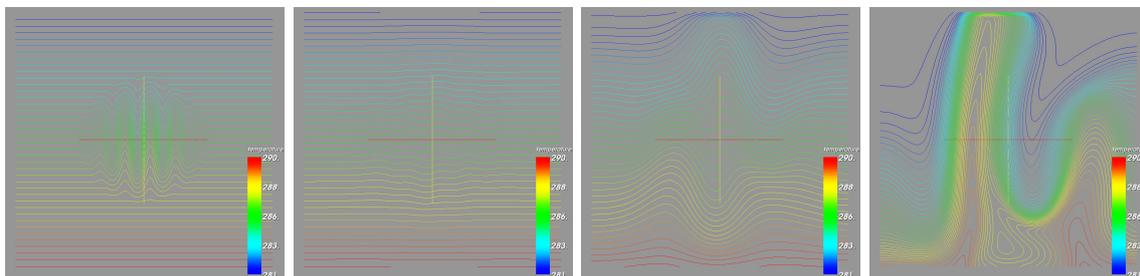
Die Abbildungen zeigen Isolinien der Temperatur [K] nach 238 Tagen Betrieb der Anlage (radialsymmetrischer Fall).



Die Abbildung zeigt die Temperaturdifferenz [K] zwischen Zu- und Ablauf über die Zeit bei 12-Stunden-Betrieb und unter Einfluß einer vorhandenen Grundwasserströmung.



Die Abbildung zeigt die Leistung [W] der Anlage über die Zeit bei 12-Stunden-Betrieb und unter Einfluß einer vorhandenen Grundwasserströmung.



Die Abbildungen zeigen die Entwicklung einer Instabilität der Temperaturschichtung bei einer 50-fachen Überhöhung der Temperaturabhängigkeit der Dichte.

1.7 Zusammenfassung

- Partielle Differentialgleichungen resultieren aus einer kontinuumsmechanischen Beschreibung physikalischer Prozesse.
- Die Wärmeleitungsgleichung wurde hergeleitet. Sie entsteht durch Einsetzen des Flussgesetzes in die Energieerhaltungsgleichung.
- Viele weitere Gleichungen der mathematischen Physik lassen sich sehr ähnlich herleiten.

1 Modellierung mit partiellen Differentialgleichungen

2 Typeinteilung partieller Differentialgleichungen

2.1 Allgemeine Definition

Eine partielle Differentialgleichung (PDGL)

- determiniert eine Funktion $u(x)$ in $n \geq 2$ Variablen $x = (x_1, \dots, x_n)^T$.
- ist eine funktionale Beziehung zwischen partiellen Ableitungen von u an *einem* Punkt.

Also allgemein:

$$F\left(\frac{\partial^m u}{\partial x_1^m}(x), \dots, \frac{\partial^m u}{\partial x_n^m}(x), \frac{\partial^{m-1} u}{\partial x_n^{m-1}}(x), \dots, u(x)\right) = 0 \quad \forall x \in \Omega \quad (2.1)$$

Wichtig:

- u heißt Lösung, wenn u die Gleichung (2.1) an allen Punkten $x \in \Omega$ erfüllt.
- zur eindeutigen Festlegung von u sind noch „zusätzliche Bedingungen“ erforderlich (siehe unten).
- m gibt die Ordnung der Differentialgleichung an.

Eine spezielle Klasse stellen *lineare* partielle DGL dar. Für Raumdimension $n = 2$ und Ordnung $m = 2$ lautet die allgemeine lineare PDGL:

$$\underbrace{a(x, y) \frac{\partial^2 u}{\partial x^2}(x, y) + 2b(x, y) \frac{\partial^2 u}{\partial x \partial y}(x, y) + c(x, y) \frac{\partial^2 u}{\partial y^2}(x, y)}_{\text{Hauptteil}} + d(x, y) \frac{\partial u}{\partial x}(x, y) + e(x, y) \frac{\partial u}{\partial y}(x, y) + f(x, y)u(x, y) + g(x, y) = 0 \quad \text{in } \Omega. \quad (2.2)$$

Die ersten drei Terme stellen den sog. „Hauptteil“ der Gleichung dar.

2.2 Typeinteilung partieller Differentialgleichungen

Definition 2.1 (Typeinteilung). 1. Gleichung (2.2) heißt *elliptisch* im Punkt (x, y) falls

$$\underbrace{a(x, y)c(x, y) - b^2(x, y)}_{\det\begin{pmatrix} a & b \\ b & c \end{pmatrix}} > 0$$

2. Gleichung (2.2) heißt *hyperbolisch* in (x, y) falls $a(x, y)c(x, y) - b^2(x, y) < 0$ und
3. (2.2) heißt *parabolisch* in (x, y) falls $a(x, y)c(x, y) - b^2(x, y) = 0$ und $\text{Rang} \begin{bmatrix} a & b & d \\ b & c & e \end{bmatrix} = 2$ in (x, y) .

□

2 Typeinteilung partieller Differentialgleichungen

Definition 2.2 (Typeinteilung in höheren Raumdimensionen). Die allgemeine lineare PDGL zweiter Ordnung in n Raumdimensionen lautet

$$\underbrace{\sum_{i,j=1}^n a_{ij}(x) \partial_{x_i} \partial_{x_j} u}_{\text{Hauptteil}} + \sum_{i=1}^n a_i(x) \partial_{x_i} u + a_0(x) = 0 \quad \text{in } \Omega.$$

O.B.d.A. kann man $a_{ij} = a_{ji}$ setzen. Mit der Matrix $(A(x))_{ij} = a_{ij}(x)$ ist die Gleichung

1. *elliptisch in x* , falls alle Eigenwerte von $A(x)$ gleiches Vorzeichen besitzen und kein Eigenwert 0 ist.
2. *hyperbolisch in x* , falls kein Eigenwert von $A(x)$ 0 ist, $n - 1$ Eigenwerte gleiches Vorzeichen besitzen und ein Eigenwert das entgegengesetzte Vorzeichen hat.
3. *parabolisch in x* , falls genau ein Eigenwert 0 ist, die übrigen Eigenwerte gleiches Vorzeichen besitzen und $\text{Rang}[A(x), a(x)] = n$.

□

Bemerkung 2.3 (Zur Typeinteilung). 1. Warum diese Einteilung? Theorie und Numerik von PDGL ist nicht einheitlich für alle möglichen PDGLs. Vielmehr sind für die verschiedenen Typen verschiedene Lösungsmethoden notwendig.

2. Obige Typeinteilung ist *vollständig* für die linearen PDGL mit $n = m = 2$. In höheren Raumdimensionen ist die Einteilung nicht mehr vollständig.
3. Der Typ ist invariant unter einer Koordinatentransformation. $\xi = \xi(x, y)$, $\eta = \eta(x, y)$ und $u(x, y) = \tilde{u}(\xi(x, y), \eta(x, y))$, liefert eine neue PDGL für $\tilde{u}(\xi, \eta)$ mit Koeffizienten \tilde{a}, \tilde{b} , etc.. Hat die Gleichung für u in (x, y) den Typ t so auch die für \tilde{u} in $(\xi(x, y), \eta(x, y))$.
4. Der Typ *kann* in verschiedenen Punkten unterschiedlich sein (aber nicht in unseren Anwendungen).
5. Der Typ wird nur vom Hauptteil bestimmt (Einschränkung: parabolisch).
6. Die Definition 2.1(3) (oben) vermeidet pathologische Fälle wie $\frac{\partial^2 u}{\partial x^2} + \frac{\partial u}{\partial x} = 0; u(x, y) = 0$.

Mehr zur Typeinteilung findet man bei [Hac86, S. 14].

□

Definition 2.4. Gleichung (2.2) heißt elliptisch (hyperbolisch, parabolisch) in Ω falls sie für alle $(x, y) \in \Omega$ elliptisch (hyperbolisch, parabolisch) ist.

□

Definition 2.5 (Typeinteilung bei erster Ordnung). Eine Gleichung der Form

$$d(x, y) \frac{\partial u}{\partial x}(x, y) + e(x, y) \frac{\partial u}{\partial y}(x, y) + f(x, y)u(x, y) + g(x, y) = 0$$

heißt hyperbolisch, falls $|d(x, y)| + |e(x, y)| > 0 \quad \forall (x, y) \in \Omega$ (sonst ist es eine gewöhnliche DGL). Für $n \geq 2$ heißt $v(x) \cdot \nabla u(x) + f(x)u(x) + g(x) = 0$ hyperbolisch.

□

Wir behandeln in der Vorlesung vor allem skalare PDGL. Es gibt auch gekoppelte Systeme mehrerer PDGL und entsprechende Typeinteilungen dafür.

2.3 Beispiele für verschiedene Typen

Beispiel 2.6 (Poisson-Gleichung).

$$\frac{\partial^2 u}{\partial x^2}(x, y) + \frac{\partial^2 u}{\partial y^2}(x, y) = f(x, y) \quad \forall (x, y) \in \Omega \quad (2.3)$$

heißt Poisson-Gleichung.

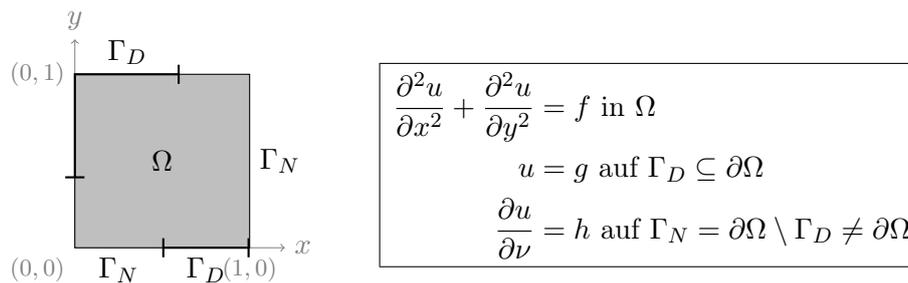
Diese ist der Prototyp für eine *elliptische* PDGL. (2.3) bestimmt die Lösung nicht eindeutig. Mit $u(x, y)$ ist z. B. auch $u(x, y) + c_1 + c_2x + c_3y$ für beliebige c_1, c_2, c_3 eine Lösung. Um u eindeutig festzulegen ist eine *Randwertvorgabe* erforderlich (deswegen sagt man auch „Randwertproblem“).

Hierbei gibt es zwei gebräuchliche Bedingungen:

1. $u(x, y) = g(x, y)$ für $(x, y) \in \Gamma_D \subseteq \partial\Omega$ (Dirichlet⁹),
2. $\frac{\partial u}{\partial \nu}(x, y) = h(x, y)$ für $(x, y) \in \Gamma_N \subset \partial\Omega$ (Neumann¹⁰, Fluß),

und $\Gamma_D \cup \Gamma_N = \partial\Omega$. Wichtig ist auch $\Gamma_N \neq \partial\Omega$, da sonst die Lösung nur bis auf eine Konstante bestimmt ist.

Die vollständige Poisson-Gleichung lautet also



Verallgemeinerung auf n Raumdimensionen:

$$\sum_{i=1}^n \frac{\partial^2 u}{\partial x_i^2} =: \Delta u = f \text{ in } \Omega$$

$$u = g \text{ auf } \Gamma_D \subseteq \partial\Omega$$

$$\nabla u \cdot \nu = h \text{ auf } \Gamma_N = \partial\Omega \setminus \Gamma_D$$

Auch diese Gleichung bezeichnet man als elliptisch. Ist $f \equiv 0$ so spricht man auch von Laplace-Gleichung. □

Beispiel 2.7 (Allgemeine Diffusionsgleichung). Sei $\Omega \subset \mathbb{R}^n$ ein Gebiet und $K : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times n}$ eine Abbildung, die jedem Punkt $x \in \Omega$ eine $n \times n$ Matrix $K(x)$ zuordnet.

Für $K(x)$ fordern wir zusätzlich (für alle $x \in \Omega$)

⁹Peter Gustav Lejeune Dirichlet, 1805-1859, dt. Mathematiker.

¹⁰John von Neumann, 1903-1957, öster.-ungar. Mathematiker.

2 Typeinteilung partieller Differentialgleichungen

1. $K(x) = K^T(x)$ und $\xi^T K(x) \xi > 0 \quad \forall \xi \in \mathbb{R}^n, \xi \neq 0$ (symmetrisch positiv definit),
2. $C(x) := \min \left\{ \xi^T K(x) \xi \mid \|\xi\| = 1 \right\} \geq C_0 > 0$ (uniforme Elliptizität).

Dann ist

$$\begin{aligned} -\nabla \cdot \left\{ K(x) \nabla u(x) \right\} &= f \text{ in } \Omega \\ u &= g \text{ auf } \Gamma_D \subseteq \partial\Omega \\ -\left(K(x) \nabla u(x) \right) \cdot \nu(x) &= h \text{ auf } \Gamma_N = \partial\Omega \setminus \Gamma_D \neq \emptyset \end{aligned} \quad (2.4)$$

die allgemeine Diffusionsgleichung (siehe auch Grundwassergleichung).

In der Praxis ist (2.4) für sehr variables K schwierig zu lösen. □

Beispiel 2.8 (Wellengleichung). Der Prototyp einer hyperbolischen Gleichung zweiter Ordnung ist die Wellengleichung:

$$\frac{\partial^2 u}{\partial x^2}(x, y) - \frac{\partial^2 u}{\partial y^2}(x, y) = 0 \quad \text{in } \Omega \quad (2.5)$$

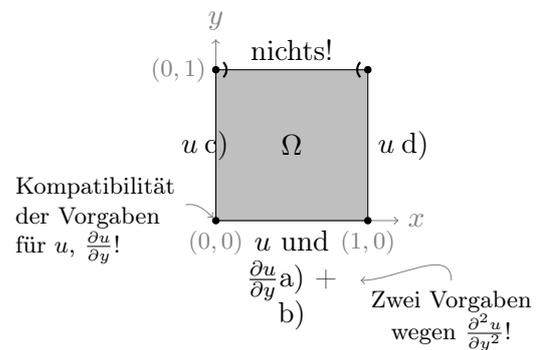
Als Randwertvorgabe kommt für $\Omega = (0, 1)^2$ etwa in Frage:

$x \in [0, 1]$:

- a) $u(x, 0) = u_0(x)$
- b) $\frac{\partial u}{\partial y}(x, 0) = u_1(x)$

$y \in [0, 1]$:

- c) $u(0, y) = g_0(y)$
- d) $u(1, y) = g_1(y)$



Beachte die ausgezeichnete Richtung y , in der Praxis wäre das Zeit! a) + b) heißen deshalb Anfangswerte und c) + d) Randwerte. Vorgaben auf dem ganze Rand sind nicht möglich! □

Beispiel 2.9 (Wärmeleitungsgleichung). Der Prototyp einer parabolischen Gleichung ist die Wärmeleitungsgleichung:

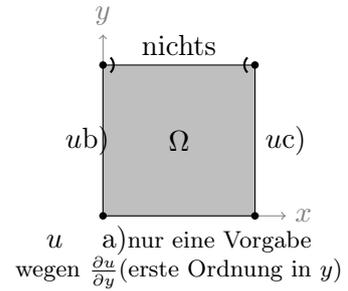
$$\frac{\partial^2 u}{\partial x^2}(x, y) - \frac{\partial u}{\partial y}(x, y) = 0 \quad \text{in } \Omega.$$

Bem.: Das $-$ ist nicht klar, auch $+$ wäre nach Def. 2.1(3) parabolisch \Rightarrow zusätzliche Stabilität bzw. sachgemäß gestellt.

2.4 Einflussbereich

Als Randwertvorgabe in $\Omega = (0, 1)^2$ wählt man für $x \in [0, 1], y \in [0, 1]$:

$$\begin{aligned} u(x, 0) &= u_0(x) \\ u(0, y) &= g_0(y) & u(1, y) &= g_1(y) \end{aligned}$$



□

Beispiel 2.10 (Transportgleichung). Sei $\Omega \subset \mathbb{R}^n, v : \Omega \rightarrow \mathbb{R}^n$ ein gegebenes Vektorfeld. Die Gleichung

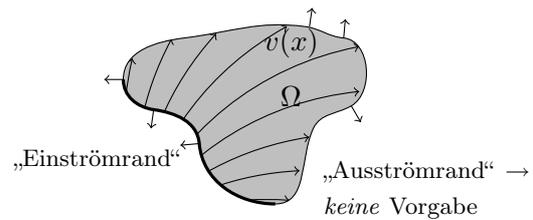
$$\nabla \cdot \{v(x)u(x)\} = f(x) \quad \text{in } \Omega$$

heißt stationäre Transportgleichung und ist hyperbolisch erster Ordnung.

Als Randwertvorgabe kommt in Betracht

$$u(x) = g(x)$$

für $x \in \partial\Omega$ so dass $v(x) \cdot \nu(x) < 0$ (Randvorgabe abhängig von den Daten)



Auch $\frac{\partial u}{\partial t} + \nabla \cdot \{v(x, t)u(x, t)\} = f(x, t)$ ist hyperbolisch 1. Ordnung.

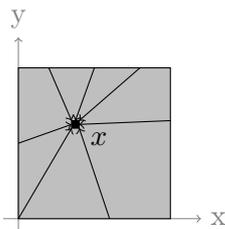
□

2.4 Einflussbereich

Der Typ einer partiellen Differentialgleichung wird auch bei folgender Frage deutlich:

Gegeben $x \in \Omega$. Welche Randwerte/Anfangswerte beeinflussen die Lösung u am Punkt x ?

Elliptisch $u_{xx} + u_{yy} = 0$

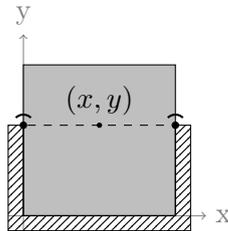


alle Randwerte beeinflussen $u(x)$, d. h. Änderung in $u(y), y \in \partial\Omega \Rightarrow$ Änderung in $u(x)$.

2 Typeinteilung partieller Differentialgleichungen

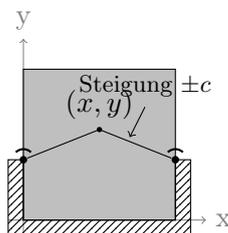
Parabolisch $u_{xx} - u_y = 0$

Bem.: Das $-$ ist wichtig, $+$ ist formal nach Def. 2.1 parabolisch, *aber* nicht sachgemäß gestellt (stabil)
 → s.u.



für (x, y) beeinflussen alle (x', y') mit $y' \leq y$ den Wert in x .
 „unendliche Ausbreitungsgeschwindigkeit“

Hyperbolisch (2. Ordnung) $u_{xx} - u_{yy} = 0$

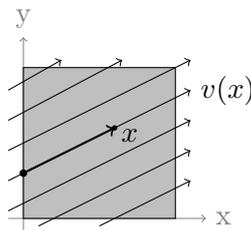


Lösung in (x, y) wird beeinflusst von allen Randpunkten unterhalb des Kegels

$$\{(x', y') \mid y' \leq (x' - x) \cdot c + y \wedge y' \leq (x - x') \cdot c + y\} \cap \partial\Omega$$

„endliche Ausbreitungsgeschwindigkeit“

Hyperbolisch (1. Ordnung) $u_x + u_y = 0$



Genau ein Randpunkt beeinflusst den Wert.

2.5 Zusammenfassung

- Wir beschäftigen uns in dieser Vorlesung vor allem mit linearen partiellen Differentialgleichungen erster und zweiter Ordnung. In zwei Raumdimensionen können diese mittels einer Typeinteilung vollständig klassifiziert werden.
- Die Typen elliptisch, hyperbolisch und parabolisch können auch auf mehr Raumdimensionen erweitert werden, allerdings ist dann nicht jede gegebene PDGL von einem dieser Typen.
- Anhand von Beispielen haben wir mögliche Randvorgaben und den Einflußbereich dieser Randvorgaben für die verschiedenen Typen von Gleichungen diskutiert.

3 Zur Theorie elliptischer partieller Differentialgleichungen

Bevor wir an die numerische Lösung elliptischer Gleichungen gehen wollen wir erst einige analytische Lösungen des Modellproblems und deren Eigenschaften betrachten.

Das elliptische Modellproblem, die Laplace-Gleichung, lautet

$$\Delta u = 0 \quad \text{in } \Omega, \quad u = g \quad \text{auf } \partial\Omega \quad . \quad (3.1)$$

Mit $C^k(\Omega)$ bezeichnen wir alle Funktionen, die k -mal stetig differenzierbar auf Ω sind.

Definition 3.1 (Klassische Lösung). Eine Funktion $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$ heißt klassische Lösung von (3.1). \square

Es gibt auch einen anderen Lösungsbegriff, die sog. schwachen Lösungen, die wir aber in dieser Vorlesung nur am Rande betrachten (bei den hyperbolischen Gleichungen erster Ordnung).

3.1 Koordinatentransformation

Oft ist wegen der Gebietsform eine Transformation der Differentialgleichung in ein anderes Koordinatensystem vorteilhaft.

Hier wollen wir die Polarkoordinaten näher betrachten.

Sei $u(x, y)$ eine klassische Lösung von (3.1), wobei wir damit implizit angenommen haben, dass x, y die kartesischen Koordinaten sind.

Wir führen die Polarkoordinaten (r, φ) ein mit

$$x(r, \varphi) = r \cos \varphi, \quad y(r, \varphi) = r \sin \varphi \quad . \quad (3.2)$$

Für die Jacobimatrix der Transformation gilt:

$$J(r, \varphi) = \begin{pmatrix} \frac{\partial x(r, \varphi)}{\partial r} & \frac{\partial x(r, \varphi)}{\partial \varphi} \\ \frac{\partial y(r, \varphi)}{\partial r} & \frac{\partial y(r, \varphi)}{\partial \varphi} \end{pmatrix} = \begin{pmatrix} \cos \varphi & -r \sin \varphi \\ \sin \varphi & r \cos \varphi \end{pmatrix}$$

Nun führen wir die neue Funktion

$$\hat{u}(r, \varphi) := u(x(r, \varphi), y(r, \varphi))$$

ein und wollen eine partielle Differentialgleichung für \hat{u} in den Koordinaten (r, φ) aufstellen.

Dazu berechnet man die partiellen Ableitungen von \hat{u} nach den neuen Koordinaten bis zur

3 Zur Theorie elliptischer partieller Differentialgleichungen

Ordnung 2. Man erhält nach einiger Rechnerei mit Produkt- und Kettenregel:

$$\begin{pmatrix} \hat{u}_r \\ \hat{u}_\varphi \\ \hat{u}_{rr} \\ \hat{u}_{r\varphi} \\ \hat{u}_{\varphi\varphi} \end{pmatrix} = \begin{pmatrix} \cos \varphi & \sin \varphi & & & \\ -r \sin \varphi & r \cos \varphi & & & \\ & & \cos^2 \varphi & 2 \sin \varphi \cos \varphi & \sin^2 \varphi \\ -\sin \varphi & \cos \varphi & -r \sin \varphi \cos \varphi & r(\cos^2 \varphi - \sin^2 \varphi) & r \sin \varphi \cos \varphi \\ -r \cos \varphi & -r \sin \varphi & r^2 \sin^2 \varphi & -2r^2 \sin \varphi \cos \varphi & r^2 \cos^2 \varphi \end{pmatrix} \begin{pmatrix} u_x \\ u_y \\ u_{xx} \\ u_{xy} \\ u_{yy} \end{pmatrix}$$

Damit rechnet man nach, dass die Laplacegleichung in Polarkoordinaten lautet:

$$\frac{\partial^2 \hat{u}}{\partial r^2} + \frac{1}{r} \frac{\partial \hat{u}}{\partial r} + \frac{1}{r^2} \frac{\partial^2 \hat{u}}{\partial \varphi^2} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \quad (3.3)$$

Beispiel 3.2 (Kreisring).

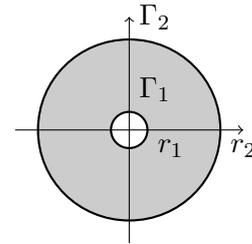
Löse

$$\Delta u = 0$$

in $\Omega = \{(x, y) \mid r_1 < \sqrt{x^2 + y^2} < r_2\}$ mit

$$u(x, y) = u_1 \quad (x, y) \in \Gamma_1,$$

$$u(x, y) = u_2 \quad (x, y) \in \Gamma_2.$$



$\hat{u}(r, \varphi)$ hängt nicht vom Winkel ab:

$$\frac{\partial^2 \hat{u}}{\partial r^2} + \frac{1}{r} \frac{\partial \hat{u}}{\partial r} = 0, \quad \hat{u}(r_1) = u_1, \quad \hat{u}(r_2) = u_2.$$

Die allgemeine Lösung lautet $\hat{u}(r) = a \ln(r) + b$. Die Konstanten a, b bestimmt man mit Hilfe der Randbedingungen. Die spezielle Lösung ist dann:

$$\hat{u}(r) = \frac{u_1 - u_2}{\ln r_1 - \ln r_2} (\ln r - \ln r_1) + u_1 \quad .$$

□

Beispiel 3.3 (Einspringende Ecke).

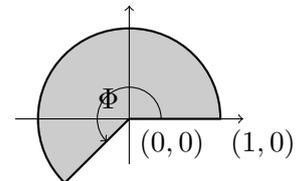
Löse das Modellproblem in

$$\Omega = \{(r, \varphi) \mid 0 < r < 1, 0 < \varphi < \Phi\}$$

mit den Randdaten (in Polarkoordinaten)

$$\hat{u}(r, \varphi) = 0, \quad \varphi \in \{0, \Phi\},$$

$$\hat{u}(r, \varphi) = \sin\left(\frac{\pi}{\Phi} \varphi\right), \quad r = 1.$$



Allgemein löst $\hat{u}(r, \varphi) = r^k \sin(k\varphi)$ die Laplacegleichung in Polarkoordinaten. Die Wahl $k = \pi/\Phi$ erfüllt die Randdaten.

Für die *Ableitung in radialer Richtung* gilt dann

$$\partial_r \hat{u}(r, \varphi) = \frac{\pi}{\Phi} r^{\frac{\pi}{\Phi}-1} \sin\left(\frac{\pi}{\Phi} \varphi\right).$$

Für $\Phi > \pi$ (nichtkonvexes Gebiet!) wird die Ableitung in $(0, 0)$ unendlich! Dies nennt man eine *Singularität*. \square

3.2 Fundamentallösung

Wir untersuchen nun Eigenschaften der Lösung von

$$\Delta u = \sum_{i=1}^n \frac{\partial^2 u}{\partial x_i^2} = 0 \quad (3.4)$$

ohne Randbedingungen.

Definition 3.4 (Harmonische Funktionen). Eine zweimal stetig differenzierbare Funktion u , für die $\Delta u = 0$ gilt, heißt harmonisch. \square

Wir wollen nun versuchen nichttriviale Funktionen (d. h. nicht etwa $u = x^2$) zu finden für die $\Delta u = 0$ gilt.

Da (3.4) invariant unter Rotation ist (sei $\Delta w(\xi, \eta) = 0$; Setze $u(x, y) = w(ax + by, cx + dy)$ mit $\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{pmatrix}$ eine Rotationsmatrix so gilt $\Delta u = 0$), sucht man harmonische Funktionen in radialsymmetrischer Form der Form:

$$u(x) = v(r(x)) \quad \text{mit} \quad r(x) = \left(\sum_{i=1}^n x_i^2 \right)^{\frac{1}{2}}.$$

Berechnen wir die partiellen Ableitungen von u mittels Kettenregel:

$$\frac{\partial r}{\partial x_i} = \frac{1}{2} (x_1^2 + \dots + x_n^2)^{-\frac{1}{2}} \cdot 2x_i = \frac{x_i}{r} \quad (x \neq 0!)$$

also:

$$\frac{\partial u}{\partial x_i}(x) = v'(r(x)) \frac{x_i}{r}, \quad \frac{\partial^2 u}{\partial x_i^2} = v''(r(x)) \left(\frac{x_i}{r}\right)^2 + v'(r(x)) \left(\frac{1}{r} - \frac{x_i^2}{r^3}\right)$$

und damit erhalten wir für Δu :

$$\begin{aligned} \Delta u &= \sum_{i=1}^n \frac{\partial^2 u}{\partial x_i^2} = \sum_{i=1}^n \left[v''(r(x)) \frac{x_i^2}{r^2} + v'(r(x)) \left(\frac{1}{r} - \frac{x_i^2}{r^3} \right) \right] \\ &= v''(r(x)) \frac{r^2}{r^2} + n v'(r(x)) \frac{1}{r} - v'(r(x)) \underbrace{\frac{r^2}{r^3}}_{\frac{1}{r}} \\ &= v''(r(x)) + \frac{n-1}{r(x)} v'(r(x)) \quad . \end{aligned}$$

3 Zur Theorie elliptischer partieller Differentialgleichungen

Nehmen wir an, es sei $v'(r(x)) \neq 0 \forall x$ dann muss für die Funktion $v(r)$ gelten (auflösen):

$$\frac{v''(r)}{v'(r)} = \frac{1-n}{r} .$$

Da dies für alle $r(x)$ gelten muss können wir r nun auch als unabhängige Variable betrachten! Beachte, dass v von der Raumdimension n abhängt.

Wie sieht so eine Funktion v aus? Für die Logarithmusfunktion beobachten wir

$$(\ln v'(r))' = \overbrace{\frac{v''(r)}{v'(r)}}^{\text{Nachdifferenzieren}} \quad \left(\text{wegen } \frac{d}{dz} \ln z = \frac{1}{z} \right) . \quad (3.5)$$

Integration liefert den Zusammenhang

$$\int (\ln v'(r))' dr = \ln v'(r) = \int \frac{1-n}{r} dr = (1-n) \int \frac{1}{r} dr = (1-n) \cdot \ln r$$

und somit hat $v'(r)$ die Gestalt

$$v'(r) = r^{1-n} \quad (\text{wegen } \ln a^b = b \ln a).$$

Wegen $\frac{v''}{v'} = \frac{1-n}{r}$ (3.5) ist mit r^{1-n} auch $b \cdot r^{1-n} + c$ zulässig, siehe auch [Eva98].

Dies motiviert folgende Definition:

Definition 3.5 (Fundamentallösung).

$$\Phi(x) := \begin{cases} -\frac{1}{2\pi} \cdot \ln r(x) & (n=2) \quad (\Rightarrow v'(r) = \frac{1}{r}, \text{ also } v(r) = \ln r) \\ \frac{1}{(n-2)\omega_n} \cdot \frac{1}{r(x)^{n-2}} & (n \geq 3) \quad (\Rightarrow v'(r) = \frac{1}{r^{n-1}} \Rightarrow v = \frac{1}{r^{n-2}}) \end{cases}$$

jeweils so eingerichtet, dass die Singularität $+\infty$ ist!

heißt Fundamentallösung. Dabei ist ω_n die Oberfläche der n -dimensionalen Einheitskugel:

$$\omega_n = \int_{r(x)=1} dx; \quad \omega_3 = 4\pi$$

Nach Konstruktion gilt $\Delta\Phi(x) = 0 \quad \forall x \neq 0$.

Entsprechend ist $\Phi(x-y)$ harmonisch in $\mathbb{R}^n \setminus \{y\}$. □

Damit können wir auch schon praktische Probleme lösen.

Beispiel 3.6 (Elektrostatik). Die Fundamentallösung kann man sich auf ganz \mathbb{R}^n erweitert vorstellen, wenn man zu Distributionen übergeht. Man stellt sich $\Phi(x)$ als Lösung der Gleichung

$$-\Delta\Phi = \delta_0 \quad \text{in ganz } \mathbb{R}^n$$

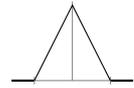
vor, wobei δ_0 die Deltafunktion an der Stelle 0 ist, welche folgende Eigenschaften hat:

$$\delta_0(x) = \begin{cases} 0 & x \neq 0 \\ \infty & x = 0 \end{cases}$$

$$\int_{\mathbb{R}^n} \delta_0(x) \, dx = 1$$

Man kann sich δ_0 als Grenzwert entsprechend skaliertes Pulse vorstellen:

$$(n = 1) \quad \delta_\varepsilon(x) = \begin{cases} 0 & |x| \geq \varepsilon \\ \frac{1}{\varepsilon} - \frac{|x|}{\varepsilon^2} = \frac{1}{\varepsilon} \left(1 - \frac{|x|}{\varepsilon}\right) & |x| < \varepsilon. \end{cases}$$



Es beschreibt dann

$$-\Delta u = \frac{q}{\varepsilon_0} \delta_0(x - y) \quad \text{in } \mathbb{R}^3$$

das elektrostatische Potential einer Punktladung q am Punkt $y \in \mathbb{R}^3$.

$E = -\nabla u$ ist das elektrische Feld dieser Punktladung (ε_0 : elektrische Feldkonstante, $8.854 \cdot 10^{-12} \left[\frac{\text{C}^2}{\text{N m}^2} \right]$, E : $\left[\frac{\text{N}}{\text{C}} \right]$, q : $\left[\frac{\text{C}}{\text{m}^3} \right]$).

$\|E\|$ zeigt für $n = 3$ wegen $u \sim \frac{1}{r}$ ein $\frac{1}{r^2}$ Verhalten (Coulombsches¹¹ Gesetz). Bei einer Dimensionsreduktion muss man also vorsichtig sein. Für $n = 2$ würde $\|E\| \sim \frac{1}{r}$ gelten und dies ist qualitativ falsch!

Als weitere praktische Anwendung betrachten wir die Gravitation.

Beispiel 3.7 (Gravitationsfeld einer Kugel). Und zwar interessieren wir uns für das Gravitationsfeld in und um eine Kugel.

(i) Zunächst das innere: Mit $r(x) = \sqrt{\sum_{i=1}^3 x_i^2}$ setzen wir

$$\Omega_i^R = \{x \in \mathbb{R}^3 \mid r(x) < R\}, \quad \Gamma^R = \partial\Omega_i^R.$$

Das Gravitationsproblem im Inneren einer homogenen Kugel mit Radius R lautet dann

$$\nabla \cdot \vec{F}_i(x) = -4\pi\gamma\rho \quad \text{in } \Omega_i^R, \tag{3.6a}$$

$$\vec{F}_i(x) = -\nabla\Phi_i(x), \tag{3.6b}$$

$$\Phi_i(x) = \Phi_R \quad \text{auf } \Gamma^R. \tag{3.6c}$$

Die Kraft auf eine Punktmasse m am Ort x berechnet sich dann mittels $\vec{F}_i^m(x) = -m\nabla\Phi_i(x)$.

¹¹Charles Augustin de Coulomb, 1736-1808, frz. Physiker

3 Zur Theorie elliptischer partieller Differentialgleichungen

Als Lösung setzen wir an $\Phi_i(x) = ar^2(x) + b$ mit Konstanten $a, b \in \mathbb{R}$.

Zunächst rechnet man

$$\partial_{x_j} \Phi_i(x) = \partial_{x_j} [a(x_1^2 + x_2^2 + x_3^2) + b] = 2ax_j, \quad \partial_{x_j} \partial_{x_j} \Phi_i(x) = 2a$$

und somit

$$-\Delta \Phi_i(x) = -6a = -4\pi\gamma\rho \quad \Rightarrow \quad a = \frac{2\pi\gamma\rho}{3}.$$

Aus der Randbedingung für das Potential erhalten wir

$$\Phi_i(R) = \frac{2\pi\gamma\rho}{3}R^2 + b = \Phi_R \quad \Rightarrow \quad b = \Phi_R - \frac{2\pi\gamma\rho}{3}R^2.$$

also insgesamt

$$\Phi_i(x) = \frac{2\pi\gamma\rho}{3}r^2(x) + \Phi_R - \frac{2\pi\gamma\rho}{3}R^2.$$

Für das Kraftfeld ergibt sich damit

$$\vec{F}_i(x) = -\nabla \Phi_i(x) = -\frac{4\pi\gamma\rho}{3}x = \underbrace{\frac{4\pi\gamma\rho}{3}r(x)}_{\text{Betrag}} \underbrace{\frac{-x}{r(x)}}_{\text{Richtung}}.$$

(ii) Nun zum Kraftfeld im Aussengebiet $\Omega_a^R = \{x \in \mathbb{R}^3 \mid r(x) > R\}$.

Wir wollen lösen

$$\nabla \cdot \vec{F}_a(x) = 0 \quad \text{in } \Omega_a^R, \quad (3.7a)$$

$$\vec{F}_a(x) = -\nabla \Phi_a(x), \quad (3.7b)$$

$$\vec{F}_a(x) \cdot \nu_a(x) = F_R \quad \text{auf } \Gamma^R, \quad (3.7c)$$

$$\Phi_a(x) = 0 \quad \text{für } r(x) \rightarrow \infty. \quad (3.7d)$$

Die Normierungsbedingung für $r \rightarrow \infty$ ersetzt die Randbedingung für das unendlich große Gebiet. ν_a ist die äußere Einheitsnormale am Kugelrand.

Da $1/r(x)$ die Laplacegleichung auch im Kugelaussengebiet löst setzen wir an: $\Phi_a(x) = c/r(x) + d$.

Man rechnet nach, dass $\partial_{x_j} \Phi_a(x) = -\frac{cx_j}{r^3(x)}$.

Wegen $\lim_{r(x) \rightarrow \infty} c/r(x) + d = d$ gilt also $d = 0$.

a erhalten wir aus der inneren Randbedingung:

$$\vec{F}(x \in \Gamma_R) \cdot \nu_a = -\nabla \Phi_a(x) \cdot \nu_a = \frac{c}{R^2} \underbrace{\frac{x}{R} \cdot \frac{-x}{R}}_{=-1} = F_R \quad \Rightarrow \quad c = -F_R R^2.$$

(iii) Es liegt nahe die beiden Lösungen miteinander zu verbinden. Dazu beobachten wir:

1. Das Potential sollte stetig für $x \in \Gamma_R$ sein (sonst wäre ja $\nabla \Phi$ undefiniert).

3.3 Grenzen des klassischen Lösungsbegriffes

2. Die Kraft \vec{F} sollte stetig sein (denn man spürt ja keine plötzliche Änderung wenn man die Hand in die Erde steckt).

Aus der zweiten Forderung schließen wir für die Konstante c im Aussengebiet

$$c = -R^2 \vec{F}_i(r(x) = R) \cdot \nu_a = - \underbrace{\frac{4}{3} R^3 \pi \rho \gamma}_{\text{Masse M}}$$

Damit gilt im Aussengebiet

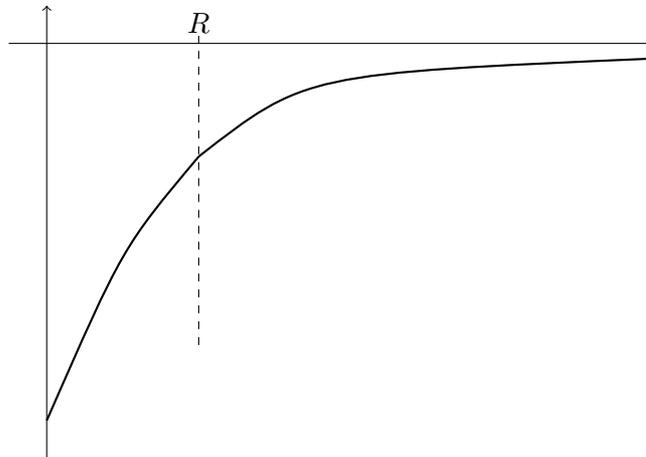
$$\Phi_a(x) = -\frac{M\gamma}{r(x)}, \quad \vec{F}_a(x) = -\nabla\Phi(x) = \frac{M\gamma}{r^2(x)} \frac{-x}{r(x)}.$$

Damit haben wir das Gravitationsgesetz für Punktmassen zurückerhalten.

Umgedreht können wir nun das Potential im Inneren bestimmen indem wir die Stetigkeit ausnutzen (erste Bedingung oben)

$$\Phi_i(x) = \frac{2\pi\gamma\rho}{3}r^2(x) + \Phi_a(R) - \frac{2\pi\gamma\rho}{3}R^2 = \frac{2\pi\gamma\rho}{3}(r^2(x) - R^2) - \frac{M\gamma}{R}.$$

Das Gesamtpotential sieht also etwa so aus:



□

3.3 Grenzen des klassischen Lösungsbegriffes

Können wir das Potential im Innen *und* Aussengebiet einer Kugel auch mit *einem* Problem berechnen?

3 Zur Theorie elliptischer partieller Differentialgleichungen

Es liegt Nahe zu schreiben

$$\nabla \cdot \vec{F}(x) = f(x) \quad \text{in } \Omega, \quad (3.8a)$$

$$\vec{F}(x) = -\nabla\Phi(x), \quad (3.8b)$$

$$\Phi_a(x) = 0 \quad \text{für } r(x) \rightarrow \infty. \quad (3.8c)$$

wobei

$$f(x) = \begin{cases} -4\pi\gamma\rho & \text{falls } r(x) < R, \\ 0 & \text{sonst.} \end{cases}$$

Halt! Das ist aber nun keine klassische Lösung mehr, da $\Phi(x)$ für $r(x) = R$ nicht zweimal stetig differenzierbar ist.

Der sogenannte „schwache Lösungsbegriff“ umgeht diese Schwierigkeit indem er die Gleichung geschickt integriert.

Eine Idee liefert die folgende Betrachtung.

Es sei ω ein Gebiet das in zwei nichtüberlappende Teilgebiete ω_1 und ω_2 zerlegt sei.

Nehmen wir an, Δu sei in ω integrierbar (Bem.: das setzt einen erweiterten Integralbegriff voraus). Dann gilt mit dem Satz von Gauss

$$\int_{\omega} \Delta u \, dx = \int_{\partial\omega} \nabla u \cdot \nu \, ds.$$

Andererseits können wir zerlegen und dann Gauss anwenden:

$$\begin{aligned} \int_{\omega} \Delta u \, dx &= \int_{\omega_1} \Delta u_1 \, dx + \int_{\omega_2} \Delta u_2 \, dx \\ &= \int_{\partial\omega_1} \nabla u_1 \cdot \nu_1 \, ds + \int_{\partial\omega_2} \nabla u_2 \cdot \nu_2 \, ds \\ &= \int_{\partial\omega} \nabla u \cdot \nu \, ds + \int_{\partial\omega_1 \cap \partial\omega_2} (\nabla u_1 - \nabla u_2) \cdot \nu_1 \, ds. \end{aligned}$$

Hierbei ist $u_1 = u|_{\omega_1}$ und $u_2 = u|_{\omega_2}$.

Gleichsetzen beider Ausdrücke liefert die Stetigkeit der Normalenkomponente $\nabla u \cdot \nu$ auf $\partial\omega_1 \cap \partial\omega_2$. Das entspricht der Stetigkeit der Normalenkräfte am Kugelrand.

Somit haben wir (sinnvoll) das Gravitationspotential und Feld im Innen- und Aussengebiet einer Kugel mit Masse M und Radius R bestimmt.

Interessant ist dann noch folgende Beobachtung: Hat man nun N Kugeln mit Radien R_i , Massen M_i und zugehörigem Potential $\Phi_i(x)$ so ist

$$\Phi(x) = \sum_{i=1}^N \Phi_i(x)$$

eine Lösung des Gravitationsproblems mit N (nichtüberlappenden) Kugeln.

Denn:

- $\Delta\Phi = 0$ ausserhalb aller Kugeln,
- $\Delta\Phi = 4\pi\gamma\rho_i$ innerhalb der Kugel i ,
- Φ ist stetig da die Einzelpotential stetig sind,
- $\nabla\Phi$ hat stetige Normalenkomponenten auf den Kugelrändern.

3.4 Separation der Variablen

Eine weitere einfache Methode zur Lösung der Laplacegleichung nennt sich „Separation der Variablen“, siehe [RR93, p. 16].

In zwei Raumdimensionen machen wir den Ansatz $u(x, y) = X(x)Y(y)$. Eingesetzt in $\Delta u = 0$ ergibt sich

$$X''(x)Y(y) + X(x)Y''(y) = 0.$$

Unter der Annahme $u(x, y) = X(x)Y(y) \neq 0$ (erreicht man evtl. durch Verschieben der Lösung) ergibt sich

$$\frac{X''(x)}{X(x)} = -\frac{Y''(y)}{Y(y)} = \lambda \in \mathbb{C},$$

denn: Wäre $X''(x)/X(x)$ nicht konstant wieso sollte es dann genau gleich der völlig unabhängigen Funktion $-Y''(y)/Y(y)$ sein?

Die Funktionen $X(x)$ und $Y(y)$ erhält man durch Lösen der gewöhnlichen Differentialgleichungen

$$X''(x) = \lambda X(x), \quad Y''(y) = -\lambda Y(y).$$

Die allgemeine Lösung dieser Gleichung mit 4 Parametern ist:

$$u(x, y) = A \left(e^{\sqrt{\lambda}x} + B e^{-\sqrt{\lambda}x} \right) \left(e^{\sqrt{-\lambda}y} + C e^{-\sqrt{-\lambda}y} \right).$$

An dieser Stelle muss man nun die Randbedingungen einfließen lassen. Sei $\Omega = (0, 1) \times (0, 1)$. Als Randbedingungen betrachten wir

$$\begin{aligned} u = 0 & \quad \text{für } x = 0, y \in (0, 1), & u = 0 & \quad \text{für } x = 1, y \in (0, 1), \\ u = 0 & \quad \text{für } x \in (0, 1), y = 0, & u = f & \quad \text{für } x \in (0, 1), y = 1. \end{aligned}$$

Setzt man die Parameter $A = \frac{1}{4i}$, $B = -1$, $C = -1$ und $\lambda = -n^2\pi^2$ so ist

$$\frac{1}{2i} (e^{in\pi x} - e^{-in\pi x}) \frac{1}{2} (e^{n\pi y} - e^{-n\pi y}) = \sin n\pi x \sinh n\pi y$$

für alle $n = 1, 2, 3, \dots$ eine Funktion, die die Nullranddaten auf den drei Rändern erfüllt. Auch alle Linearkombinationen erfüllen diese Eigenschaft.

3 Zur Theorie elliptischer partieller Differentialgleichungen

Hat f die Gestalt $f(x) = \sum_{n=1}^N \alpha_n \sin n\pi x$ so lautet die Lösung

$$u(x, y) = \sum_{n=1}^N \frac{\alpha_n}{\sinh n\pi} \sin n\pi x \sinh n\pi y.$$

Was macht man nun bei allgemeinen Funktionen f ?

Fourier hat behauptet, dass „jede“ Funktion (hier mit Nullrändern) in eine möglicherweise unendliche Reihe entwickelt werden kann:

$$f(x) = \sum_{n=1}^{\infty} \alpha_n \sin n\pi x.$$

Allerdings ist die Klasse der so darstellbaren Funktionen nicht $C^0[0, 1]$ sondern die größere Klasse $L_2(0, 1)$ der quadratintegrierbaren Funktionen.

Die so ermittelten Lösungen sind somit keine klassischen Lösungen der Laplacegleichung.

3.5 Mittelwerteigenschaft und Folgen

Harmonische Funktionen haben eine Reihe bemerkenswerter Eigenschaften, die wir in diesem Abschnitt betrachten wollen.

Satz 3.8 (Mittelwerteigenschaft). Sei u harmonisch im Gebiet $U \subset \mathbb{R}^n$ und $B(x, r) \subseteq U$ sei die offene Kugel mit Radius r um den Punkt x . Dann gilt

$$u(x) = \frac{1}{\underbrace{\omega_n r^{n-1}}_{\substack{\text{Oberfläche} \\ \text{der Kugel}}}} \int_{\partial B(x,r)} u \, ds = \frac{1}{\underbrace{\alpha_n r^n}_{\substack{\text{Volumen} \\ \text{der n-dim.} \\ \text{Einheitskugel}}}} \int_{B(x,r)} u(\xi) \, d\xi \quad (3.9)$$

Beweis: [Eva98, S. 25, Theorem 2]. □

Als Folge der Mittelwerteigenschaft gilt das

Satz 3.9 (Starkes Maximumprinzip). Sei $u \in C^2(U) \cap C^0(\bar{U})$ harmonisch im *beschränkten* Gebiet U .

1. Dann ist

$$\max_{\bar{U}} u = \max_{\partial U} u.$$

d. h. das Maximum wird auf dem Rand angenommen.

2. Ist U *zusammenhängend* (sind unsere Gebiete immer) und es gibt $x_0 \in U$ (also im Inneren) so dass

$$u(x_0) = \max_{\bar{U}} u$$

dann muss

u konstant in U sein!

Umkehr: u *nicht* konstant \Rightarrow das Maximum liegt echt am Rand.

Beweis: [Eva98, S. 27, Theorem 4].

Ersetzt man u durch $-u$ (ist auch harmonisch!), so erhält man dieselben Aussagen für \min statt \max . \square

Das Maximumprinzip ist eine wichtige qualitative Eigenschaft der Lösung mit physikalischer Bedeutung: Ohne innere Quellen/Senken wird das Maximum/Minimum der Temperatur im stationären Fall am Rand angenommen.

Auch eine numerische Lösung sollte eine solche Bedingung erfüllen.

Aus dem Maximumprinzip folgt ausserdem sofort die Eindeutigkeit der Lösung.

Satz 3.10 (Eindeutigkeit). Sei $g \in C^0(\partial\Omega)$, $f \in C^0(\Omega)$ und Ω ein beschränktes Gebiet. Dann gibt es höchstens eine Lösung $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$ des Problems

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega \\ u &= g && \text{auf } \partial\Omega. \end{aligned}$$

Beweis: Angenommen u, \tilde{u} wären Lösungen, dann setze $w = u - \tilde{u}$. w ist harmonisch und $w = 0$ auf $\partial\Omega$. Da Maximum und Minimum nicht im Inneren liegen dürfen, bleibt nur $w = \text{const} = 0$. \square

Bemerkenswert ist auch die folgende Aussage:

Satz 3.11 (Glattheit). Erfüllt $u \in C^0(U)$ die Mittelwerteigenschaft ((3.9)) für alle $B(x, r) \subset U$, dann gilt $u \in C^\infty(U)$, d. h. harmonische Funktionen sind *automatisch* unendlich oft differenzierbar!

Beweis: [Eva98, S. 28, Theorem 6] \square

Achtung: Wir wissen also, dass harmonische Funktionen automatisch unendlich oft differenzierbar sind und dass Lösungen des Modellproblems eindeutig sind. Die *Existenz* einer Lösung $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$ des Modellproblems zu zeigen bereitet aber Schwierigkeiten!

3.6 Lösungsdarstellung mittels Greenscher Funktion

Satz 3.12 (Lösungsdarstellung mittels Greenscher Funktion). Ist Ω ein Normalgebiet (partielle Integration ist erlaubt), so erlaubt die Lösung $u \in C^2(\bar{\Omega})$ von

$$-\Delta u = f \quad \text{in } \Omega, \quad u = g \quad \text{auf } \partial\Omega$$

die Darstellung

$$u(x) = \int_{\partial\Omega} g(\xi) \nabla_f G(\xi, x) \cdot \nu(\xi) \, d\xi = \int_{\Omega} G(\xi, x) f(\xi) \, d\xi. \quad (3.10)$$

Dabei heißt $G(\xi, x)$ Greensche¹² Funktion erster Art (d. h. für Dirichlet-Randbedingung).

Diese muss einmal für ein Gebiet Ω gefunden werden und erlaubt dann eine Lösung für verschiedene f, g . G existiert für sehr allgemeine Gebiete, ist nur im Allgemeinen schwer zu finden.

Beweis: [Hac86, Satz 3.2.5]. □

In der Kugel kennt man die Greensche Funktion $B(y, r)$. Eine Anwendung des obigen Satzes ist die

Satz 3.13 (Poissonsche Integralformel). Die Funktion

$$u(x) = \frac{r^2 - \|x - y\|^2}{r\omega_n} \int_{\partial B(y, r)} \frac{g(\xi)}{\|x - \xi\|^n} \, d\xi \quad x \in B(y, r)$$

löst

$$\begin{aligned} \Delta u &= 0 && \text{in } B(y, r) \\ u &= g && \text{auf } \partial B(y, r). \end{aligned}$$

Beweis: [Hac86, Satz 2.3.9]. □

Die Charakterisierung der Lösung über die Greensche Funktion erlaubt auch einen Existenzansatz:

Satz 3.14 (Existenzsatz). Es existiere eine Greensche Funktion erster Art in Ω , es sei $g \in C^0(\partial\Omega)$ und $f \in C^0(\bar{\Omega})$ Hölder-stetig zum Exponenten $\lambda \in (0, 1)$, d.h. $|f(x) - f(y)| \leq C|x - y|^\lambda \forall x, y \in \bar{\Omega}$ (und C unabhängig von x, y).

Dann stellt die Lösungsdarstellung mittels Grenn'scher Funktion (3.10) eine Lösung $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$ (= klassische Lösung) dar.

Beweis: [Hac86, Satz 3.2.13]. □

Die Existenz einer Lösung ist damit auf die Existenz der Greenschen Funktion für das Gebiet zurückgeführt.

Eine wesentlich allgemeinere Lösungstheorie erhält man über sog. „Energimethoden“, die auf einer sog. „schwachen“ Formulierung aufbauen. Dies ist aber einer Vorlesung über die Finite Elemente Methode vorbehalten.

¹²George Green, 1793-1841, engl. Mathematiker und Physiker.

3.7 Stabilität

Existenz und Eindeutigkeit der Lösung ist nicht genug. Zusätzlich benötigt man noch, dass das Problem sachgemäß gestellt ist.

Definition 3.15 (Sachgemäß gestelltes Problem). Eine Aufgabe der abstrakten Form

$$A(x) = y, \quad x \in X, \quad y \in Y$$

heißt sachgemäß gestellt, wenn

1. Zu jedem $y \in Y$ gibt es ein eindeutiges $x \in X$, so dass $A(x) = y$.
2. $x = A^{-1}(y)$ hängt stetig von y ab. (Stabilität)
D. h. man kann zeigen, dass

$$\|A^{-1}(y_1) - A^{-1}(y_2)\|_X \leq C\|y_1 - y_2\|_Y$$

(bei linearem A genügt $\|A^{-1}(y)\| \leq C\|y\|$). $\|\cdot\|_X, \|\cdot\|_Y$ sind Normen auf den Räumen X, Y .

□

Satz 3.16. Laplace- und Poissongleichung sind sachgemäß gestellt. Hier zeigen wir die stetige Abhängigkeit von den Randdaten g .

Sei $\Delta u_1 = \Delta u_2 = 0$ in Ω mit den Randdaten $u_1 = g_1$ auf $\partial\Omega$, $u_2 = g_2$ auf $\partial\Omega$. Dann gilt

$$\Delta w = \Delta(u_1 - u_2) = 0 \text{ in } \Omega, \quad w = g_1 - g_2 \text{ auf } \partial\Omega.$$

Wegen dem Maximumprinzip gilt $\|w\|_\infty \leq \|g_1 - g_2\|_\infty$ mit $C = 1$.

□

Stabilität ist wichtig, damit kleine Fehler (z. B. Rundungsfehler) auch nur kleine Fehler im Ergebnis bewirken.

Bemerkung 3.17. Die Angabe von Rand-/Anfangswertvorgaben wie in Abschnitt 2.3 führt zu sachgemäß gestellten Problemen zu den aufgeführten Typen.

□

3.8 Zusammenfassung

- Mittels Transformation der Laplacegleichung in Polarkoordinaten konnten wir spezielle Lösungen in einem Kreisring und einem Kreissegment konstruieren.
- Fundamentallösungen sind Lösungen der Laplacegleichung im $\mathbb{R}^n \setminus y$. Man kann sie auch als Lösungen zur Deltafunktion als rechter Seite interpretieren.
- Das Maximumprinzip stellt eine wichtige qualitative Eigenschaft von Lösungen der Laplacegleichung dar. Daraus folgt auch Existenz und sachgemäße Gestelltheit des Randwertproblems.

3 Zur Theorie elliptischer partieller Differentialgleichungen

- Mittels Greenscher Funktionen lassen sich Lösungen für verschiedene rechte Seiten und Randdaten darstellen. Allerdings muss man die Greensche Funktion für das vorliegende Gebiet erst mal bestimmen.
- Neben Existenz und- Eindeutigkeit fordert man ausserdem, dass partielle Differentialgleichungen sachgemäß gestellt sind. Dies bedeutet, dass die Lösung stetig von den Daten abhängt.

4 Differenzenmethode für elliptische Gleichungen

4.1 Der eindimensionale Fall

Wir betrachten zunächst die eindimensionale Randwertaufgabe

$$\begin{aligned} -u''(x) &= f(x) & x \in (0, 1) \\ u(0) &= \varphi_0, & u(1) = \varphi_1. \end{aligned}$$

(Im Gegensatz zur Anfangswertaufgabe $u(0) = u_0$, $u'(0) = u_1$).

Aus der Taylorentwicklung für $u(x \pm h)$ erhält man

$$\begin{aligned} u(x+h) &= u(x) + hu'(x) + \frac{h^2}{2}u''(x+\vartheta^+h) & \vartheta^+ \in (0, 1) & (4.1) \\ \Leftrightarrow u'(x) &= \frac{u(x+h) - u(x)}{h} - \frac{h}{2}u''(x+\vartheta^+h) & \vartheta^+ \in (0, 1) \end{aligned}$$

bzw.

$$\begin{aligned} u(x-h) &= u(x) - hu'(x) + \frac{h^2}{2}u''(x-\vartheta^-h) & (4.2) \\ \Leftrightarrow u'(x) &= \frac{u(x) - u(x-h)}{h} + \frac{h}{2}u''(x-\vartheta^-h) \end{aligned}$$

$$\begin{aligned} (\partial^+ u)(x) &:= [u(x+h) - u(x)]/h & \text{heißt Vorwärtsdifferenz und} \\ (\partial^- u)(x) &:= [u(x) - u(x-h)]/h & \text{heißt Rückwärtsdifferenz.} \end{aligned}$$

sowie h Schrittweite.

Entwickelt man bis zur 4. Potenz (bzw. 3. Potenz),

$$\begin{aligned} u(x+h) &= u(x) + hu'(x) + \frac{h^2}{2}u''(x) + \frac{h^3}{6}u'''(x) + \frac{h^4}{24}u''''(x+\vartheta^+h) \\ u(x-h) &= u(x) - hu'(x) + \frac{h^2}{2}u''(x) - \frac{h^3}{6}u'''(x) + \frac{h^4}{24}u''''(x-\vartheta^-h) \end{aligned}$$

so erhält man die Formeln

$$\begin{aligned} u(x+h) - u(x-h) &= 2hu'(x) + \frac{h^3}{6} \{u'''(x+\vartheta^+h) + u'''(x-\vartheta^-h)\} \\ \Leftrightarrow u'(x) &= \frac{u(x+h) - u(x-h)}{2h} - \frac{h^2}{12} \{u'''(x+\vartheta^+h) + u'''(x-\vartheta^-h)\} & (4.3) \end{aligned}$$

Inhaltsverzeichnis

bzw.

$$\begin{aligned}
 u(x+h) + u(x-h) &= 2u(x) + h^2 u''(x) + \frac{h^4}{24} \{u''''(x+\vartheta^+ h) + u''''(x-\vartheta^- h)\} \\
 \iff u''(x) &= \frac{u(x-h) - 2u(x) + u(x+h)}{h^2} - \frac{h^2}{24} \{\dots\} \quad . \quad (4.4)
 \end{aligned}$$

Damit erhalten wir das folgende Lemma.

Lemma 4.1. Es gilt

$$\left. \begin{aligned}
 \frac{1}{h} [u(x+h) - u(x)] &= u'(x) + hR \\
 \frac{1}{h} [u(x) - u(x-h)] &= u'(x) + hR
 \end{aligned} \right\} \quad \text{mit } |R| \leq \frac{1}{2} \|u\|_{C^2(\bar{\Omega})} := \sup_{x \in \bar{\Omega}} u''(x)$$

$$\begin{aligned}
 \frac{1}{2h} [u(x+h) - u(x-h)] &= u'(x) + h^2 R & \text{mit } |R| \leq \frac{1}{6} \|u\|_{C^3(\bar{\Omega})} \\
 \frac{1}{h^2} [u(x-h) - 2u(x) + u(x+h)] &= u''(x) + h^2 R & \text{mit } |R| \leq \frac{1}{12} \|u\|_{C^4(\bar{\Omega})}
 \end{aligned}$$

□

Zur näherungsweisen Lösung des Randwertproblems unterteilen wir $\Omega = (0, 1)$ in N Teilintervalle

$$[x_i, x_{i+1}] \quad i = 0, \dots, N-1 \quad \text{und} \quad x_i = ih \quad \text{mit} \quad h = \frac{1}{N}.$$

und setzen

$$\Omega_h = \{ih \mid i \in \mathbb{N}_0 \quad \text{und} \quad 0 < i < N\} \quad \begin{array}{c} N=8 \\ \frac{0}{x_1} \cdot \frac{\cdot}{x_2} \cdot \frac{\cdot}{x_3} \cdot \frac{\cdot}{x_4} \cdot \frac{\cdot}{x_5} \cdot \frac{\cdot}{x_6} \cdot \frac{\cdot}{x_7} \frac{1}{x_8} \end{array}$$

bzw.

$$\bar{\Omega}_h = \{ih \mid i \in \mathbb{N}_0 \quad \text{und} \quad 0 \leq i \leq N\} \quad \begin{array}{c} \frac{0}{x_0} \cdot \frac{\cdot}{x_1} \cdot \frac{\cdot}{x_2} \cdot \frac{\cdot}{x_3} \cdot \frac{\cdot}{x_4} \cdot \frac{\cdot}{x_5} \cdot \frac{\cdot}{x_6} \cdot \frac{\cdot}{x_7} \frac{1}{x_8} \end{array}$$

Ist $u \in C^4(\bar{\Omega})$ ergibt einsetzen der Differenzenformel in die Differentialgleichung:

$$-\frac{1}{h^2} [u(x-h) - 2u(x) + u(x+h)] = f(x) + O(h^2) \quad \forall x \in \Omega_h.$$

Streichen des Resttermes ergibt $\#\Omega_h = N - 1$ lineare Gleichungen

$$-\frac{1}{h^2} [u_h(x-h) - 2u_h(x) + u_h(x+h)] = f(x) \quad \forall x \in \Omega_h \quad (4.5)$$

für die $\#\bar{\Omega}_h = N + 1$ unbekanntes Werte $u_h(x)$, $x \in \bar{\Omega}_h$. Die restlichen zwei Gleichungen liefert die Randbedingung:

$$u_h(0) = \varphi_0, \quad u_h(1) = \varphi_1. \quad (4.6)$$

$u_h : \bar{\Omega}_h \rightarrow \mathbb{R}$ nennen wir eine Gitterfunktion. Wahlweise fassen wir u_h auch als Vektor $u_h \in \mathbb{R}^{N-1}$

$$u_h = (u_h(h), u_h(2h), \dots, u_h(1-h))^T$$

auf. Hierbei beschränken wir uns auf die unbekanntenen Werte $u_h(x)$, $x \in \Omega_h$ (also exklusive der Randwerte).

Eliminiert man $u(0)$, $u(1)$ in (4.5) mittels (4.6), so erhält man das *lineare Gleichungssystem*

$$\underbrace{\frac{1}{h^2} \begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix}}_{L_h} \underbrace{\begin{bmatrix} u_h(h) \\ u_h(2h) \\ u_h(3h) \\ \vdots \\ u_h(1-2h) \\ u_h(1-h) \end{bmatrix}}_{u_h} = \underbrace{\begin{bmatrix} f(h) + \frac{\varphi_0}{h^2} \\ f(2h) \\ f(3h) \\ \vdots \\ f(1-2h) \\ f(1-h) + \frac{\varphi_1}{h^2} \end{bmatrix}}_{= q_h}$$

mit der Tridiagonalmatrix L_h .

4.2 Der n -dimensionale Fall

Die Differenzenformel für die zweite Ableitung gilt auch für partielle Ableitungen:

$$\partial_{x_i} \partial_{x_i} u = h^{-2} [u(\dots, x_i - h, \dots) - 2u(\dots, x_i, \dots) + u(\dots, x_i + h, \dots)] + O(h^2) \quad .$$

Lemma 4.2. Sei $u \in C^4(\bar{\Omega})$, dann gilt

$$\Delta u(x) = \frac{1}{h^2} \left\{ \sum_{i=1}^n [u(x - he_i) + u(x + he_i)] - 2nu(x) \right\} + h^2 R$$

mit $|R| \leq \frac{n}{12} \|u\|_{C^4(\bar{\Omega})}$ und $e_i = (0, \dots, \underset{\substack{\uparrow \\ i\text{-te Position}}}{1}, \dots, 0)^T$. (4.7)

□

Man definiert

$$\Delta_h u(x) := \sum_{i=1}^n [u(x - he_i) + u(x + he_i)] - 2nu(x).$$

Speziell für $n = 2$ wählen wir wieder das Gitter

$$\begin{aligned} \Omega_h &= \{(x, y) \in \Omega \mid x/h, y/h \in \mathbb{Z}\} & \text{mit } h = \frac{1}{N} \text{ und} \\ \bar{\Omega}_h &= \{(x, y) \in \bar{\Omega} \mid x/h, y/h \in \mathbb{Z}\}. \end{aligned}$$

Inhaltsverzeichnis

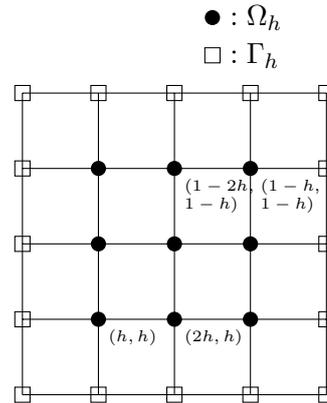
Weiter sei $\Gamma_h := \bar{\Omega}_h - \Omega_h$.

Für die Gitterfunktion $u_h : \bar{\Omega}_h \rightarrow \mathbb{R}$ erhalten wir unter Vernachlässigung des Fehlerterms die $(N - 1)^2$ Bedingungen

$$-\Delta_h u_h(x) = f(x) \quad \forall x \in \Omega_h$$

und die $(N + 1)^2 - (N - 1)^2 = N^2 + 2N + 2 - N^2 + 2N - 2 = 4N$ Bedingungen

$$u_h(x) = g(x) \quad \forall x \in \Gamma_h.$$



Elimination der Randbedingungen liefert wieder ein lineares Gleichungssystem

$$L_h u_h = q_h$$

für den unbekanntem Vektor

$$u_h = (u_h(h, h), u_h(2h, 2), \dots, u_h(1 - 2h, 1 - h), u_h(1 - h, 1 - h))^T \in \mathbb{R}^{(N-1)^2}.$$

Dabei gilt wieder folgende Konvention:

- $u_h : \bar{\Omega}_h \rightarrow \mathbb{R}$ ist eine Gitterfunktion. Da sind die Randwerte mit dabei.
- $u_h \in \mathbb{R}^{(N-1)^2}$ ist ein Vektor. Da sind die Randwerte *nicht* mit dabei.

Nun zur Gestalt des linearen Gleichungssystems.

$$L_h = \frac{1}{h^2} \begin{bmatrix} \begin{array}{ccc|ccc} 4 & -1 & & & & \\ -1 & 4 & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -1 & 4 & -1 & \\ & & & -1 & 4 & \\ & & & & & -1 \end{array} & \begin{array}{ccc} -1 & & \\ & -1 & \\ & & \ddots \\ & & & -1 \end{array} & \\ \hline \begin{array}{ccc} -1 & & \\ & -1 & \\ & & \ddots \\ & & & -1 \end{array} & \begin{array}{ccc|ccc} 4 & -1 & & & & \\ -1 & 4 & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -1 & 4 & -1 & \\ & & & -1 & 4 & \\ & & & & & -1 \end{array} & \begin{array}{ccc} -1 & & \\ & -1 & \\ & & \ddots \\ & & & -1 \end{array} \\ \hline & \begin{array}{ccc} -1 & & \\ & -1 & \\ & & \ddots \\ & & & -1 \end{array} & \begin{array}{ccc|ccc} 4 & -1 & & & & \\ -1 & 4 & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -1 & 4 & -1 & \\ & & & -1 & 4 & \\ & & & & & -1 \end{array} \end{bmatrix} \quad (4.8)$$

Allgemein steht das bei beliebig großen Sternen für

$$\begin{bmatrix} & & \vdots & & \\ & c_{-1,1} & c_{0,1} & c_{1,1} & \\ \dots & c_{-1,0} & c_{0,0} & c_{1,0} & \dots \\ & c_{-1,-1} & c_{0,-1} & c_{1,-1} & \\ & & \vdots & & \end{bmatrix} u_h(x, y) = \sum_{i,j=-\infty}^{\infty} c_{i,j} u_h(x + ih, y + jh) \quad .$$

Idee: Zentriere $c_{0,0}$ über einem Gitterpunkt $x \in \Omega_h$, dann gibt der Stern die Verknüpfung mit den Nachbarn, also eine *Matrixzeile* an.

Das lässt sich auch auf n Raumdimensionen erweitern. Bei $n = 1$ schreibt man etwa

$$-\Delta_h = h^{-2}[-1 \quad 2 \quad -1] \quad .$$

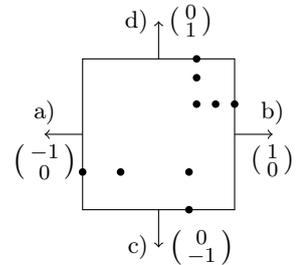
4.3 Neumann Randbedingung

Wir wenden uns nun der Aufgabe mit Neumann Randbedingungen zu:

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega = (0, 1)^2, \\ u &= g && \text{auf } \Gamma_D \subseteq \partial\Omega \\ \frac{\partial u}{\partial \nu} &= \varphi, && \text{auf } \Gamma_N = \partial\Omega \setminus \Gamma_D \quad . \end{aligned}$$

Da Ω das Einheitsquadrat ist, gilt

$$\begin{aligned} \text{a) } \frac{\partial u}{\partial \nu} &= -\frac{\partial u}{\partial x} && \text{für } x = (0, y) \\ \text{b) } \frac{\partial u}{\partial \nu} &= \frac{\partial u}{\partial x} && \text{für } x = (1, y) \\ \text{c) } \frac{\partial u}{\partial \nu} &= -\frac{\partial u}{\partial y} && \text{für } x = (x, 0) \\ \text{d) } \frac{\partial u}{\partial \nu} &= \frac{\partial u}{\partial y} && \text{für } x = (x, 1) \end{aligned}$$



Entsprechend benutzt man in den vier Fällen die $O(h)$ Approximationen

$$\begin{aligned} \text{a) } \frac{1}{h} [u(x, y) - u(x + h, y)] &= \frac{\partial u}{\partial \nu} + O(h) && x = (0, y) \\ \text{b) } \frac{1}{h} [u(x, y) - u(x - h, y)] &= \frac{\partial u}{\partial \nu} + O(h) && x = (1, y) \\ \text{c) } \frac{1}{h} [u(x, y) - u(x, y + h)] &= \frac{\partial u}{\partial \nu} + O(h) && x = (x, 0) \\ \text{d) } \frac{1}{h} [u(x, y) - u(x, y - h)] &= \frac{\partial u}{\partial \nu} + O(h) && x = (x, 1) \end{aligned} \tag{4.9}$$

4.4 Allgemeine elliptische Gleichung

Wie bei der Dirichlet Randbedingung kann (4.9) benutzt werden, um die Randpunkte aus dem LGS zu eliminieren, z. B. am rechten Rand (b):

$$\frac{1}{h} [u(1, y) - u(1 - h, y)] = \varphi(1, y) \quad \Rightarrow \quad u(1, y) = h\varphi(1, y) + u(1 - h, y)$$

Einsetzen in die Gleichung am Punkt $(1 - h, y)$

$$\begin{aligned} \frac{1}{h^2} [-u(1 - 2h, y) - u(1 - h, y - h) + 4u(1 - h, y) - u(1, y) \\ - u(1 - h, y + h)] = f(1 - h, y) \end{aligned}$$

liefert

$$\begin{aligned} \frac{1}{h^2} [-u(1 - 2h, y) - u(1 - h, y - h) + 3u(1 - h, y) - u(1 - h, y + h)] \\ = f(1 - h, y) + h^{-1}\varphi(1, y) \quad . \end{aligned}$$

Der Stern ist also $\begin{bmatrix} & -1 & & \\ -1 & 3 & 0 & \\ & -1 & & \end{bmatrix}$.

4.4 Allgemeine elliptische Gleichung

Um die allgemeine elliptische Gleichung

$$\begin{aligned} a_{xx}(x, y) \frac{\partial^2 u}{\partial x^2} + 2a_{xy}(x, y) \frac{\partial^2 u}{\partial x \partial y} + a_{yy}(x, y) \frac{\partial^2 u}{\partial y^2} \\ + a_x(x, y) \frac{\partial u}{\partial x} + a_y(x, y) \frac{\partial u}{\partial y} + a(x, y)u = f(x, y) \end{aligned}$$

zu diskretisieren, benötigen wir noch Differenzenformeln für die gemischte Ableitung $\frac{\partial^2 u}{\partial x \partial y}$.

Diese wollen wir nun erst einmal mittels Taylorentwicklung herleiten.

Taylor nacheinander in x und dann in y anwenden liefert:

$$\begin{aligned} u(x + h, y + h) = & u(x, y + h) + hu_x(x, y + h) + \frac{h^2}{2}u_{xx}(x, y + h) + \frac{h^3}{6}u_{xxx}(x, y + h) + \frac{h^4}{24}u_{xxxx}(\xi_1) \\ & u(x, y) + hu_y(x, y) + \frac{h^2}{2}u_{yy}(x, y) + \frac{h^3}{6}u_{yyy}(x, y) + \frac{h^4}{24}u_{yyyy}(\xi_2) \\ = & \quad + hu_x(x, y) + h^2u_{xy}(x, y) + \frac{h^3}{6}u_{xyy}(x, y) + \frac{h^4}{24}u_{xyyy}(\xi_3) \\ & \quad + \frac{h^2}{2}u_{xx}(x, y) + \frac{h^3}{2}u_{xxy}(x, y) + \frac{h^4}{4}u_{xxyy}(\xi_4) \\ & \quad + \frac{h^3}{6}u_{xxx}(x, y) + \frac{h^4}{6}u_{xxxxy}(\xi_5) \end{aligned}$$

Inhaltsverzeichnis

entsprechend erhält man in die andere Richtung:

$$\begin{aligned}
 u(x-h, y-h) &= \\
 &= u(x, y-h) - hu_x(x, y-h) + \frac{h^2}{2}u_{xx}(x, y-h) - \frac{h^3}{6}u_{xxx}(x, y-h) + \frac{h^4}{24}u_{xxxx}(\xi_6) \\
 &= \begin{array}{r} u(x, y) \\ -hu_x(x, y) \\ +h^2u_{xx}(x, y) \\ -\frac{h^3}{6}u_{xxx}(x, y) \\ +\frac{h^4}{24}u_{xxxx}(\xi_6) \end{array} \\
 &= \begin{array}{r} u(x, y) \\ -hu_y(x, y) \\ +\frac{h^2}{2}u_{yy}(x, y) \\ -\frac{h^3}{6}u_{yyy}(x, y) \\ +\frac{h^4}{24}u_{yyyy}(\xi_7) \end{array} \\
 &= \begin{array}{r} -hu_x(x, y) \\ +h^2u_{xy}(x, y) \\ -\frac{h^3}{2}u_{xxy}(x, y) \\ +\frac{h^4}{6}u_{xxyy}(\xi_8) \end{array} \\
 &= \begin{array}{r} +\frac{h^2}{2}u_{xx}(x, y) \\ -\frac{h^3}{2}u_{xxy}(x, y) \\ +\frac{h^4}{4}u_{xxyy}(\xi_9) \end{array} \\
 &= \begin{array}{r} -\frac{h^3}{6}u_{xxx}(x, y) \\ +\frac{h^4}{6}u_{xxxxy}(\xi_{10}) \end{array}
 \end{aligned}$$

Bei Addition der Ausdrücke fallen die ungeraden h -Potenzen heraus:

$$\begin{aligned}
 u(x+h, y+h) + u(x-h, y-h) &= \\
 &= 2u(x, y) + h^2[u_{xx}(x, y) + 2u_{xy}(x, y) + u_{yy}(x, y)] + O(h^4) \quad .
 \end{aligned}$$

Nun sind noch die Ableitungen u_{xx} und u_{yy} zu eliminieren was mit den Punkten $x \pm h$ und $y \pm h$ möglich ist:

$$\begin{aligned}
 &u(x+h, y+h) + u(x-h, y-h) \\
 &\quad - [u(x+h, y) + u(x-h, y) + u(x, y+h) + u(x, y-h)] \\
 &= -2u(x, y) + h^2 2u_{xy}(x, y) + O(h^4) \quad .
 \end{aligned}$$

Auflösen nach u_{xy} liefert die Differenzenformel

$$\frac{\partial^2 u}{\partial x \partial y}(x, y) = \frac{1}{2h^2} \begin{bmatrix} 0 & -1 & 1 \\ -1 & 2 & -1 \\ 1 & -1 & 0 \end{bmatrix} u_h(x, y) + O(h^2) \quad . \quad (4.10)$$

Man rechne auch nach, dass ebenfalls folgende Formel gilt:

$$\frac{\partial^2 u}{\partial x \partial y}(x, y) = \frac{1}{2h^2} \begin{bmatrix} -1 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{bmatrix} u_h(x, y) + O(h^2) \quad . \quad (4.11)$$

Dies zeigt nebenbei, dass es im allgemeinen verschiedene Differenzenformeln (gleicher Ordnung) zu einer Ableitung gibt.

Welche man nun wählt, hängt vom Vorzeichen der Koeffizienten ab.

Aufgrund der Elliptizität der Gleichung gilt $a_{xx} \cdot a_{yy} > a_{xy}^2$, also haben a_{xx} und a_{yy} in jedem Fall gleiches Vorzeichen. O. B. d. A. seien a_{xx} und a_{yy} im folgenden als *positiv* angenommen.

Zusätzlich fordern wir noch

$$|a_{xy}(x, y)| \leq \min\{a_{xx}(x, y), a_{yy}(x, y)\} \quad (4.12)$$

(dies folgt sofort aus der Elliptizität falls $a_{xx} = a_{yy}$.)

Dann wähle

- a) falls $a_{xy} \geq 0$ den „rechts oben“-Stern (4.10) und
 b) falls $a_{xy} < 0$ den „links oben“-Stern (4.11).

Mit dieser Wahl und der Bezeichnung

$$a_{xy}^+ = \max\{a_{xy}, 0\} \quad a_{xy}^- = \min\{a_{xy}, 0\}$$

erhält man dann:

$$\begin{aligned} a_{xx} \frac{\partial^2 u}{\partial x^2} + 2a_{xy} \frac{\partial^2 u}{\partial x \partial y} + a_{yy} \frac{\partial^2 u}{\partial y^2} \\ = \frac{1}{h^2} \begin{bmatrix} -a_{xy}^- & a_{yy} - |a_{xy}| & a_{xy}^+ \\ a_{xx} - |a_{xy}| & 2(|a_{xy}| - a_{xx} - a_{yy}) & a_{xx} - |a_{xy}| \\ a_{xy}^+ & a_{yy} - |a_{xy}| & -a_{xy}^- \end{bmatrix} + O(h^2) \end{aligned}$$

Bemerkung 4.3. Mit dieser Konstruktion gilt: Diagonalelement ist negativ und Nebendiagonalelemente sind positiv. Warum dies sehr wichtig ist wird die nun folgende Theorie zeigen. \square

Schließlich noch die Terme der Ordnung kleiner 2:

$$a_x \frac{\partial u}{\partial x} + a_y \frac{\partial u}{\partial y} + au = \frac{1}{2h} \begin{bmatrix} 0 & a_y & 0 \\ -a_x & 0 & a_x \\ 0 & -a_y & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & 0 \end{bmatrix} + O(h^2).$$

Beachte, dass die Vorzeichenbedingung für die zentralen Differenzen erster Ordnung nicht erfüllt sind.

4.5 Zusammenfassung

- Mit Hilfe der Taylorentwicklung lassen sich Differenzenformeln herleiten, die partielle Ableitungen an einem Punkt durch Funktionswerte an Nachbarpunkten ausdrücken.
- Einsetzen in die DGL und ignorieren der Fehlerterme liefert ein lineares Gleichungssystem. Bei strukturierten Gittern und geeigneter Anordnung der Gitterpunkte hat das LGS Bandgestalt.
- Die maximal mögliche (lokale) Fehlerordnung im Falle kompakter 9-Punkte-Sterne (in $2d$) beträgt 2. Höhere Ordnung lässt sich im Prinzip mit Sternen größerer Reichweite erzielen. Dann gibt es aber zusätzliche Schwierigkeiten am Rand.
- Ein kompaktes Verfahren der Ordnung 4 für die Poisson-Gleichung ist die Collatzsche¹³ Mehrstellenformel (siehe Übung).

¹³Lothar Collatz, 1910-1990, dt. Mathematiker.

Inhaltsverzeichnis

- Die hergeleiteten lokalen Fehlerordnungen basieren auf *äquidistanten* Gitterweiten (zumindest in jede der Koordinatenrichtungen) und strukturierten Gittern.
- Eine weitere wichtige Voraussetzung der Herleitung war $u \in C^4(\bar{\Omega})$.

5 M-Matrix-Theorie

Die Konvergenztheorie für das Finite-Differenzen Verfahren basiert auf den Eigenschaften einer speziellen Klasse von Matrizen, den sogenannten M-Matrizen um die es in diesem Abschnitt gehen soll. Eine weitere Anwendung der M-Matrix Theorie sind iterative Lösungsverfahren für lineare Gleichungssysteme.

Dieses Kapitel ist stark an [Hac86, Kap. 4.3–4.5] angelehnt.

5.1 Einführende Definitionen

Sei $A \in \mathbb{R}^{n \times n}$ eine Matrix mit den Elementen $a_{\alpha\beta}$, $\alpha, \beta \in I = \{1, \dots, n\}$ (d. h. wir verwenden in diesem Abschnitt A und I statt L_h und Ω_h). I nennt man eine Indexmenge.

Man schreibt

$$A \geq B \quad \text{falls} \quad a_{\alpha\beta} \geq b_{\alpha\beta} \quad \forall \alpha, \beta \in I.$$

Analog verwendet man $A \geq B$, $A > B$, $A < B$. 0 bezeichne die Nullmatrix.

Definition 5.1 (M-Matrix). A heißt M-Matrix, wenn

- (i) $a_{\alpha\alpha} > 0$ für alle $\alpha \in I$ sowie $a_{\alpha\beta} \leq 0$ für alle $\alpha \neq \beta$
- (ii) A nicht singulär und $A^{-1} \geq 0$.

(i) kann man einfach an der Matrix ablesen („Vorzeichenbedingung“), jedoch braucht man weitere Kriterien, um (ii) *einfach* nachzuweisen. □

Definition 5.2 (Graph von A).

$$G(A) = (I, E), \quad E \subseteq I \times I \\ (\alpha, \beta) \in E \iff a_{\alpha\beta} \neq 0$$

heißt Graph der Matrix A . Ist $(\alpha, \beta) \in E$, so heißt α *direkt verbunden* mit β . α heißt *verbunden* mit β , falls es eine Kette

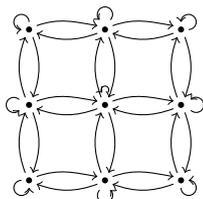
$$\alpha = \alpha_0, \alpha_1, \alpha_2, \dots, \alpha_k = \beta \quad \text{mit} \quad (\alpha_{i-1}, \alpha_i) \in E$$

für $i = 1 \dots k$ gibt. □

Definition 5.3 (Irreduzible Matrix). Eine Matrix A heißt *irreduzibel*, falls jedes $\alpha \in I$ mit jedem $\beta \in I$ verbunden ist. □

Beispiel 5.4. Sei $A = L_h$ zu $h = \frac{1}{4}$.

Graph von A :



- Da $L_h = L_h^T$, sind alle Verbindungen bidirektional
- L_h ist irreduzibel
- $G(A)$ entspricht dem Gitter (da wir nur nächste Nachbarverbindungen haben). □

5 M-Matrix-Theorie

Eine Folgerung der Irreduzibilität ist: Es gibt keine Permutation der Indizes (beschrieben durch eine Permutationsmatrix P), sodass

$$P^T A P = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \begin{matrix} \} I_1 \\ \} I_2 \end{matrix} \quad \text{d.h. } \alpha \in I_2 \text{ ist nicht mit } \beta \in I_1 \text{ verbunden} \\ \text{(das ist schon der Beweis).}$$

Die Invertierbarkeit von L_h folgt aus dem nun folgenden Satz.

5.2 Gerschgorin Kreise und Regularität

Satz 5.5 (Gerschgorin-Kreise). Sei $B(z, r) = \{\xi \in \mathbb{C} \mid |z - \xi| < r\}$ der offene Kreis um z mit Radius r in \mathbb{C} und $\overline{B}(z, r) = \{\xi \in \mathbb{C} \mid |z - \xi| \leq r\}$ der abgeschlossene Kreis.

(a) Alle Eigenwerte von A liegen in

$$\bigcup_{\alpha \in I} \overline{B}(a_{\alpha\alpha}, r_\alpha) \quad \text{mit } r_\alpha = \sum_{\substack{\beta \neq \alpha \\ \beta \in I}} |a_{\alpha\beta}|.$$

(b) Ist A irreduzibel, so liegen die Eigenwerte in

$$\bigcup_{\alpha \in I} \underbrace{B(a_{\alpha\alpha}, r_\alpha)}_{\substack{\text{offener} \\ \text{Kreis!}}} \cup \underbrace{\left(\bigcap_{\alpha \in I} \partial B(a_{\alpha\alpha}, r_\alpha) \right)}_{\text{Schnitte aller Ränder}}.$$

Beweis:

(a) Sei (λ, u) ein Eigenpaar, d. h. $Au = \lambda u$. O. B. d. A. sei $\|u\|_\infty = \max\{|u_\alpha| \mid \alpha \in I\} = 1$ (mit u ist auch $c \cdot u$ ein Eigenvektor für alle $c \in \mathbb{R}, c \neq 0$). Wegen $\|u\|_\infty = 1$ gibt es mindestens ein $\gamma \in I$ mit $|u_\gamma| = 1$. Für dieses γ gilt

$$\sum_{\beta \in I} a_{\gamma\beta} u_\beta = \lambda u_\gamma \quad (\gamma\text{-te Zeile von } Au = \lambda u) \\ \iff (\lambda - a_{\gamma\gamma}) u_\gamma = \sum_{\substack{\beta \neq \gamma \\ \beta \in I}} a_{\gamma\beta} u_\beta$$

Beträge bilden liefert

$$|\lambda - a_{\gamma\gamma}| \underbrace{|u_\gamma|}_{=1} = \left| \sum_{\beta \neq \gamma} a_{\gamma\beta} u_\beta \right| \overset{\text{Dreiecksungleichung}}{\leq} \sum_{\beta \neq \gamma} |a_{\gamma\beta}| \underbrace{|u_\beta|}_{\leq 1} \leq \sum_{\beta \neq \gamma} |a_{\gamma\beta}| = r_\gamma \quad (5.1)$$

also

$$\lambda \in \overline{B}(a_{\gamma\gamma}, r_\gamma)$$

und damit auch $\lambda \in \bigcup_{\alpha \in I} \overline{B}(a_{\alpha\alpha}, r_\alpha)$ (man kennt γ nicht, deshalb nimmt man alle; λ war ja auch beliebig).

(b) A sei nun zusätzlich irreduzibel. Wegen (a) gilt weiterhin

$$\lambda \in \bigcup_{\alpha \in I} \overline{B(a_{\alpha\alpha}, r_\alpha)} = \underbrace{\bigcup_{\alpha \in I} B(a_{\alpha\alpha}, r_\alpha)}_{\Lambda_<} \cup \underbrace{\bigcup_{\alpha \in I} \partial B(a_{\alpha\alpha}, r_\alpha)}_{\Lambda_=}$$

Ist nun $\lambda \notin \Lambda_<$ so gilt klar $\lambda \in \Lambda_=$. Es ist nun zu zeigen, dass aus $\lambda \notin \Lambda_<$ sogar $\lambda \in \bigcap_{\alpha \in I} \partial B(a_{\alpha\alpha}, r_\alpha)$ folgt.

Sei dazu wie oben (λ, u) Eigenpaar und $\|u_\infty\| = 1$. Weiter sei wie oben $|u_\gamma| = 1$ und zusätzlich $a_{\gamma\beta} \neq 0$. (5.1) aus (a) zeigte $|\lambda - a_{\gamma\gamma}| \leq r_\gamma$. Da nun nach Voraussetzung $\lambda \notin \Lambda_<$ muss also $|\lambda - a_{\gamma\gamma}| = r_\gamma$ gelten und damit werden in (5.1) alle \leq zu $=$, also

$$|\lambda - a_{\gamma\gamma}| = \sum_{\beta \neq \gamma} |a_{\gamma\beta}| |u_\beta| = \sum_{\beta \neq \gamma} |a_{\gamma\beta}| = r_\gamma.$$

Dies geht nur, wenn $|u_\beta| = 1$ für alle $a_{\gamma\beta} \neq 0$.

Somit haben wir mindestens ein weiteren $\beta \in I$ gefunden mit $|u_\beta| = 1$. Darauf lässt sich das selbe Argument anwenden und man findet wegen $\lambda \notin \Lambda_<$ einen weiteren Index, nennen wir ihn δ , so dass $|u_\delta| = 1$. Da A irreduzibel ist, erreicht man von γ aus also *alle anderen* Indizes und es gilt

$$|u_\alpha| = 1, \quad |\lambda - a_{\alpha\alpha}| = r_\alpha \quad \forall \alpha \in I.$$

Da $|\lambda - a_{\alpha\alpha}| = r_\alpha$ für *alle* Indizes gleichzeitig haben wir also $\lambda \in \bigcap_{\alpha \in I} \partial B(a_{\alpha\alpha}, r_\alpha)$. \square

Folgerung 5.6 (L_h ist invertierbar). L_h ist irreduzibel, was man sofort am Graphen von A sieht (Gitter, Bidirektionalität).

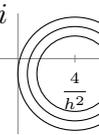
Die Differenzensterne (Zeilen von A) lauten für den 5-Punkte-Stern

innere Knoten:	Randknoten (süd):	
$\frac{1}{h^2} \begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix}$	$\frac{1}{h^2} \begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & 0 & \end{bmatrix}$	$\frac{1}{h^2} \begin{bmatrix} & -1 & \\ -1 & 4 & 0 \\ & 0 & \end{bmatrix}$

Damit gilt für die verschärfte Form der Gerschgorin-Kreise:

$$\begin{aligned} \lambda \in B\left(\frac{4}{h^2}, \frac{4}{h^2}\right) \cup \underbrace{B\left(\frac{4}{h^2}, \frac{3}{h^2}\right) \cup B\left(\frac{4}{h^2}, \frac{2}{h^2}\right)}_{\subset B\left(\frac{4}{h^2}, \frac{4}{h^2}\right)} \\ \cup \underbrace{\left(\partial B\left(\frac{4}{h^2}, \frac{4}{h^2}\right) \cap \partial B\left(\frac{4}{h^2}, \frac{3}{h^2}\right) \cap \partial B\left(\frac{4}{h^2}, \frac{2}{h^2}\right)\right)}_{= \emptyset} \end{aligned}$$

also $\lambda \in B\left(\frac{4}{h^2}, \frac{4}{h^2}\right)$ und damit insbesondere $0 \notin B\left(\frac{4}{h^2}, \frac{4}{h^2}\right)$. \square



5.3 Diagonaldominante Matrizen

Definition 5.7 (Diagonaldominanz). A heißt *diagonaldominant*, falls

$$(i) \quad \sum_{\substack{\beta \neq \alpha \\ \beta \in I}} |a_{\alpha\beta}| < |a_{\alpha\alpha}| \quad \forall \alpha \in I$$

und *irreduzibel diagonaldominant*, falls

(ii) A irreduzibel und

$$(iii) \quad \sum_{\beta \neq \alpha} |a_{\alpha\beta}| \leq |a_{\alpha\alpha}| \quad \forall \alpha \in I$$

(iv) und (i) für *mindestens* einen Index $\alpha \in I$ gilt. □

Bemerkung 5.8. L_h ist irreduzibel diagonaldominant, aber nicht diagonaldominant. □

Definition 5.9 (Spektralradius). Unter dem Spektralradius einer Matrix versteht man den betragsmäßig größten Eigenwert:

$$\rho(A) := \max \{ |\lambda| \mid \lambda \text{ ist Eigenwert von } A \}.$$

□

Um die M-Matrix-Eigenschaft nachzuweisen, brauchen wir folgenden

Satz 5.10. Spalte A auf in $A = D - B$ mit der Diagonalmatrix $d_{\alpha\beta} = \begin{cases} a_{\alpha\alpha} & \beta = \alpha \\ 0 & \beta \neq \alpha \end{cases}$. Ist A diagonaldominant oder irreduzibel diagonaldominant, so gilt $\rho(D^{-1}B) < 1$ (Bemerkung: $D^{-1}B$ ist genau die Iterationsmatrix von Jacobi).

Beweis: Setze $C := D^{-1}B$ mit $c_{\alpha\beta} = \begin{cases} 0 & \beta = \alpha \text{ da } b_{\alpha\alpha} = 0 \\ -\frac{a_{\alpha\beta}}{a_{\alpha\alpha}} & \beta \neq \alpha \end{cases}$

- Sei nun A diagonaldominant, so gilt

$$r_\alpha = \sum_{\beta \neq \alpha} |c_{\alpha\beta}| = \sum_{\beta \neq \alpha} \left| \frac{a_{\alpha\beta}}{a_{\alpha\alpha}} \right| = \frac{1}{|a_{\alpha\alpha}|} \sum_{\beta \neq \alpha} |a_{\alpha\beta}| \stackrel{\text{Diagonaldominant}}{<} \frac{|a_{\alpha\alpha}|}{|a_{\alpha\alpha}|} = 1 \quad (5.2)$$

Nach Gerschgorin Satz 5.5(a) gilt

$$\lambda \in \bigcup_{\alpha \in I} \overline{B(c_{\alpha\alpha}, r_\alpha)} = \bigcup_{\alpha \in I} \overline{B(0, r_\alpha)} \quad \text{mit } r_\alpha < 1, \text{ also } |\lambda| < 1.$$

- Sei A irreduzibel diagonaldominant, so gilt

$$r_\beta \leq 1 \quad \forall \beta \in I \text{ (wie (5.2)) und } r_\alpha < 1 \text{ für mind. ein } \alpha \in I.$$

Nun ist aber Gerschgorin Satz 5.5(b) anwendbar und somit

$$\lambda \in \bigcup_{\beta \in I} \underbrace{B(0, r_\beta)}_{\text{offen}} \cup \left(\bigcap_{\beta \in I} \partial B(0, r_\beta) \right) \subset B(0, 1) \cup \left(\bigcap_{\beta \in I} \partial B(0, r_\beta) \right)$$

Wir unterscheiden zwei Fälle

- I) alle r_β sind gleich, d. h. $r_\beta = r \quad \forall \beta \in I$. Da $r_\alpha < 1$ für ein $\alpha \in I$, muss also $r < 1$ gelten und damit $\bigcap_{\beta \in I} \partial B(0, r_\beta) = \partial B(0, r) \subset B(0, 1)$
- II) die r_β sind nicht alle gleich, dann gilt $\bigcap_{\beta \in I} \partial B(0, r_\beta) = \emptyset$

In jedem Fall also $\lambda \in B(0, 1)$, also $\rho(D^{-1}B) < 1$. □

Schließlich zweigen wir

Lemma 5.11. Sei $A = D - B$ wie in Satz 5.10 und A erfülle die Vorzeichenbedingung (s. 5.1(i)). A ist eine M-Matrix genau dann, wenn $\rho(D^{-1}B) < 1$.

Beweis:

„ \Leftarrow “ Sei $C := D^{-1}B$ und $\rho(C) < 1$.

Dann konvergiert die geometrische Reihe $S := \sum_{\nu=0}^{\infty} C^\nu$ und es gilt $S = (I - C)^{-1}$.

Dies sieht man so ein. Betrachte die Partialsumme

$$(I - C) \sum_{\nu=0}^n C^\nu = \sum_{\nu=0}^n (C^\nu - C^{\nu+1}) = I - C^1 + C^1 - \dots + C^n - C^{n+1} = I - C^{n+1}.$$

Für $\lim_{n \rightarrow \infty}$ gilt nun $C^n \rightarrow 0$, da $\rho(C) < 1$ (schreibe $C = QRQ^T$, Q unitär, R eine untere Dreiecksmatrix, zeige $\|C^\nu\|_\infty \leq K[\rho(C)]^\nu$; Spezialfall: C diagonalisierbar, also $C = QD'Q^T$).

Weiter gilt $D \geq 0$ (da $a_{\alpha\alpha} > 0$) und $B \geq 0$ wegen der Vorzeichenbedingung und damit auch $C = D^{-1}B \geq 0$ und $C^\nu \geq 0$ und $S \geq 0$.

Schließlich ist

$$I = \underbrace{(I - C)^{-1}}_S (I - C) = S \underbrace{(I - D^{-1}B)}_C = SD^{-1} \underbrace{(D - B)}_A = \underbrace{SD^{-1}}_{A^{-1}} A,$$

also ist $A^{-1} = SD^{-1} \geq 0$, da $S \geq 0$ und $D^{-1} \geq 0$.

„ \Rightarrow “ Sei A eine M-Matrix und $D^{-1}Bu = \lambda u$, (λ, u) Eigenpaar mit $u \neq 0$. Definiere $|u|$ für $u \in \mathbb{R}^n$ als $|u|_\alpha := |u_\alpha|$ (Beträge der Komponenten).

Dann gilt:

$$\begin{array}{c}
 \text{Eigenpaar} \qquad \text{gilt wegen } D^{-1}B \geq 0 \\
 \qquad \qquad \qquad \text{(Vorzeichen)} \\
 |\lambda| \underbrace{|u|} = \underbrace{|\lambda u|} \downarrow = |D^{-1}Bu| \leq D^{-1}B \underbrace{|u|} \\
 \text{Vektor!} \quad \text{Vektor!} \qquad \qquad \qquad \text{Vektor!!} \\
 \text{mit Betrag der} \qquad \qquad \qquad \text{mit den Beträgen} \\
 \text{Komponenten}
 \end{array}
 \quad
 \boxed{\begin{array}{l} \text{Dies ist ein} \\ \text{System von} \\ \text{Ungleichun-} \\ \text{gen!} \end{array}}
 \quad (5.3)$$

Nun rechne

$$\begin{aligned}
 \underbrace{|u|} &= A^{-1}A|u| = A^{-1}(D - B)|u| = A^{-1}D(I - D^{-1}B)|u| \\
 \text{Vektor!} & \\
 \text{mit Betrag der} & \\
 \text{Komponenten} & \\
 &= A^{-1}D|u| - A^{-1}DD^{-1}B|u| \qquad \text{aus} \\
 & \qquad \qquad \qquad |\lambda||u| \leq D^{-1}B|u| \quad (5.3) \\
 & \qquad \qquad \qquad \text{folgt} \\
 & \qquad \qquad \qquad -|\lambda||u| \geq -D^{-1}B|u| ! \\
 & \leq \underbrace{A^{-1}D|u| - A^{-1}D|\lambda||u|} \\
 & \qquad \text{komponentenweise,} \\
 & \qquad \text{da } A^{-1}D \geq 0 \\
 & = (1 - |\lambda|)A^{-1}D|u|
 \end{aligned}$$

also $|u| \leq (1 - |\lambda|)A^{-1}D|u|$ als komponentenweises System von Ungleichungen.
 Angenommen $|\lambda| \geq 1$, dann folgt $|u| \leq 0$, was nur für $|u| = 0$ bei nicht singulärem $A^{-1}D \geq 0$ geht. Dies ist ein Widerspruch zur Voraussetzung $u \neq 0$.
 Also folgt $|\lambda| < 1$, also $\varrho(D^{-1}B) < 1$. □

Damit erhalten wir den

Satz 5.12. Erfüllt eine Matrix A die Vorzeichenbedingung 5.1(i) und ist A diagonaldominant oder irreduzibel doagonaldominant, so ist A eine M-Matrix.

Beweis: Folgt sofort aus Lemma 5.11 zusammen mit Satz 5.10. □

Bemerkung 5.13. Damit ist L_h auch eine M-Matrix. □

Es gilt sogar noch die Verschärfung

Satz 5.14. Ist A eine M-Matrix und irreduzibel, so gilt $A^{-1} > 0$. *Beweis:* [Hac86, Satz 4.3.11]. □

Bemerkung 5.15. Dies erklärt die Behauptung „Alle Randwerte beeinflussen die Lösung im Punkt $x \in \Omega$ “ für elliptische Gleichungen im diskreten.

Denn bei $f = 0$ tauchen in q_h nur die Randwerte auf. □

5.4 Zusammenfassung

- In diesem Abschnitt zeigten wir, dass L_h invertierbar, irreduzibel diagonaldominant und M-Matrix ist.
- Dies dient als Vorbereitung auf die im nächsten Abschnitt folgende Konvergenztheorie.

6 Konvergenz des Finite-Differenzen-Verfahrens

In Abschnitt 4 wurde die Poissongleichung

$$-\Delta u = f \quad \text{in } \Omega = (0, 1)^2 \quad (6.1a)$$

$$u = g \quad \text{auf } \partial\Omega \quad (6.1b)$$

mit dem Finite-Differenzen-Verfahren diskretisiert. Dies führte auf eine Differenzengleichung

$$-\Delta_h u_h(x) = f(x) \quad \forall x \in \Omega_h \quad (6.2a)$$

$$u_h(x) = g(x) \quad \forall x \in \Gamma_h \quad (6.2b)$$

für die Gitterfunktion $u_h: \bar{\Omega}_h \rightarrow \mathbb{R}$ und $\bar{\Omega}_h = \{(x, y) \in \bar{\Omega} \mid x/h, y/h \in \mathbb{Z}\}$ und $h = \frac{1}{N}$. Elimination der Randbedingung (6.2b) in (6.2a) lieferte das lineare Gleichungssystem

$$L_h u_h = q_h \quad (6.3)$$

mit invertierbarem L_h (da irreduzibel diagonaldominante M-Matrix). Dabei besteht der Vektor u_h aus den Werten an den Punkten $\Omega_h = \{(x, y) \in \Omega \mid x/h, y/h \in \mathbb{Z}\}$ (innere Punkte).

Wir untersuchen nun die Frage, in welcher Beziehung u und u_h stehen.

Dazu sei $U_h = \{f \mid f: \bar{\Omega}_h \rightarrow \mathbb{R}\}$ der Vektorraum der Gitterfunktionen und

$$R_h: C^0(\bar{\Omega}) \rightarrow U_h$$

definiert durch $\underbrace{(R_h u)}_{\text{Gitterfunktion}}(x) := u(x)$. Weiter ist dann

Gitterfunktion

$$e_h(x) := (R_h u)(x) - u_h(x) \quad \forall x \in \bar{\Omega}_h \quad (6.4)$$

der Fehler in der Finite-Differenzen Lösung u_h .

6.1 Konvergenz

Zu jeder Gitterfunktion $v_h(x) \in U_h$ gehört ein Vektor $v_h \in \mathbb{R}^{\#\Omega_h}$, der durch Weglassen der Randwerte $v_h(x), x \in \Gamma_h$ und Wahl einer Anordnung entsteht.

In diesem Sinne sind $u_h, e_h \in \mathbb{R}^{\#\Omega_h}$ die zu den Gitterfunktionen $u_h(x)$ und $e_h(x)$ gehörenden Vektoren.

Setze nun

$$\eta_h = L_h c_h = L_h \underbrace{(R_h u - u_h)}_{\text{Vektor!}} \stackrel{\text{Linearität}}{\downarrow} = L_h R_h u - L_h u_h.$$

Vektor!

L_h ist eine nichtsinguläre Matrix und daher gilt

$$e_h = L_h^{-1} \eta_h.$$

6 Konvergenz des Finite-Differenzen-Verfahrens

Nun nehmen wir die Maximumnorm auf beiden Seiten

$$\|e_h\|_\infty = \|L_h^{-1}\eta_h\|_\infty.$$

Mit Hilfe der *zugeordneten* Matrixnorm

$$\|A\|_\infty := \sup_{u \neq 0} \frac{\|Au\|_\infty}{\|u\|_\infty} = \underbrace{\max_{\alpha \in I} \left\{ \sum_{\beta \in I} |a_{\alpha\beta}| \right\}}_{\text{„Zeilensummennorm“}}$$

werden wir in diesem Abschnitt zeigen, dass gilt

$$\|e_h\|_\infty \leq \|L_h^{-1}\|_\infty \|\eta_h\|_\infty$$

Stabilität

$\|L_h^{-1}\|_\infty \leq K$
 K unabh. von h

Konsistenz

$\|\eta_h\|_\infty \leq Ch^2\|u\|_{C^4(\bar{\Omega})}$
 C unabh. von h

und damit Konvergenz

$$\|e_h\|_\infty \leq O(h^2).$$

D. h. der Fehler $|u(x) - u_h(x)| \rightarrow 0 \quad \forall x \in \Omega_h$ und $h \rightarrow 0$.

6.2 Konsistenz

Unter Ausnutzung der Fehlerabschätzung für die Differenzenformeln aus Abschnitt 4 erhält man:

Lemma 6.1 (Konsistenz). Mit den Bezeichnungen von oben gilt

$$\|\eta_h\|_\infty = \|L_h e_h\|_\infty = \|L_h R_h u - L_h u_h\|_\infty \leq Ch^2 \|u\|_{C^4(\bar{\Omega})} \quad (6.5)$$

Der Vektor $\eta_h = L_h e_h$ heißt *lokaler Abschneidefehler* und e_h ist die Lösung des LGS $L_h e_h = \eta_h$.

Beweis: Wir zeigen das konkret für zwei Raumdimensionen. Betrachte zunächst einen randfernen inneren Punkt $(x, y) \in \Omega_h$, so dass $(x \pm h, y \pm h) \notin \Gamma_h$.

Dann gilt für die (x, y) entsprechende Zeile in $L_h e_h$ unter Zuhilfenahme von Lemma 4.2:

$$\frac{1}{h^2} \underbrace{[-u(x-h, y) - u(x, y-h) + 4u(x, y) - u(x+h, y) - u(x, y+h)]}_{\substack{[L_h R_h u]_{(x,y)} = -\Delta_h u(x, y) = \Delta u(x, y) + h^2 R \\ \text{mit } |R| \leq \frac{1}{6} \|u\|_{C^4(\bar{\Omega})}}}$$

$$- \underbrace{f(x, y)}_{\text{Element von } L_h u_h = q_h} = h^2 R \quad \text{mit } |R| \leq \frac{1}{6} \|u\|_{C^4(\bar{\Omega})}.$$

Für einen randnahen Punkt, z. B. $x = (x, h)$ erhält man entsprechend:

$$\underbrace{\frac{1}{h^2} [-u(x-h, h) + 4u(x, h) - u(x+h, h) - u(x, 2h)] - \frac{g(x, 0)}{h^2}}_{\substack{[L_h R_h u]_{(x, h)} \\ = -\Delta_h(x, h) = \Delta u(x, h) + h^2 R}} + \frac{g(x, 0)}{h^2} - \underbrace{\left(f(x, y) + \frac{g(x, 0)}{h^2} \right)}_{(q_h)_{(x, h)}} = h^2 R \quad \text{mit } |R| \leq \frac{1}{6} \|u\|_{C^4(\Omega)}.$$

Da $\|\cdot\|_\infty$ die Maximum-Norm, gilt (6.5) mit $C = \frac{1}{6}$. □

6.3 Stabilität

Wir brauchen eine Abschätzung für $\|L_h^{-1}\|_\infty$. Hier geht wieder entscheidend ein, dass L_h eine M-Matrix ist.

Zunächst beweist man den

Satz 6.2. Sei A eine M-Matrix und es gebe einen Vektor w mit $Aw \geq \mathbb{1}$ (d. h. komponentenweise $\sum a_{\alpha\beta} w_\beta \geq 1 \quad \forall \alpha$). Dann gilt

$$\|A^{-1}\|_\infty \leq \|w\|_\infty.$$

Beweis: Wie in Lemma 5.11 sei $|u|$ der Vektor mit den Komponenten $|u_\alpha|$. Für jedes u ist dann

$$\begin{array}{ccccc} |u| & \leq & \|u\|_\infty \mathbb{1} & \leq & \|u\|_\infty Aw \\ & \uparrow & & \uparrow & \\ & |u_\alpha| & \leq & \|u\|_\infty & \text{Vor.} \end{array}$$

Wegen $A^{-1} \geq 0$ (M-Matrix) gilt

$$\underbrace{|A^{-1}u|}_{\text{Vektor}} \leq A^{-1}|u| \leq A^{-1}\|u\|_\infty Aw = \|u\|_\infty w \quad (\text{Vektorungleichung!})$$

mit Komp. $|(A^{-1}u)_\alpha|$

Wegen $\|\underbrace{A^{-1}u}_{\text{Vektor}}\|_\infty = \|A^{-1}u\|_\infty \leq \|u\|_\infty \|w\|_\infty$ gilt unter Ausnutzung der Definition $\|A^{-1}\|_\infty$

$$\|A^{-1}\|_\infty = \sup_{u \neq 0} \frac{\|A^{-1}u\|_\infty}{\|u\|_\infty} \leq \frac{\|u\|_\infty \|w\|_\infty}{\|u\|_\infty} = \|w\|_\infty \quad \square$$

6 Konvergenz des Finite-Differenzen-Verfahrens

Damit gilt dann der

Satz 6.3 (Stabilität). Ist L_h die Matrix aus der Diskretisierung von $-\Delta$ mit dem Fünfpunktestern, so gilt

$$\|L_h^{-1}\|_\infty \leq \frac{1}{8}$$

Beweis: Wähle $w_h = R_h w$ mit $w(x, y) = x(1-x)/2$. Dann gilt für (x, y) randfern:

$$\begin{aligned} (L_h w_h)_{(x,y)} &= \frac{1}{h^2} \left[-\frac{(x-h)(1-x+h)}{2} - \frac{x(1-x)}{2} + 4\frac{x(1-x)}{2} \right. \\ &\quad \left. - \frac{(x+h)(1-x-h)}{2} - \frac{x(1-x)}{2} \right] \\ &= \frac{1}{2h^2} [x(1-x) - x(1-x) + h(1-x) - xh + h^2 + x(1-x) - x(1-x) \\ &\quad + xh - h(1-x) + h^2] = 1 \end{aligned}$$

Dies gilt auch für $x = h, x = 1-h$, da dort $x-h$ bzw. $x+h = 0$, also $w = 0$. Für $y = h, y = 1-h$ bleibt $\frac{1}{2h^2} \left[2h^2 + \frac{x(1-x)}{2} \right] = 1 + \frac{x(1-x)}{4h^2} > 1$ und damit also $L_h w \geq \mathbb{1}$.

Nach Satz 6.2 gilt dann $\|L_h^{-1}\|_\infty \leq \|w\|_\infty$ und $\|w\|_\infty \leq w(\frac{1}{2}, y) = \frac{1}{8}$.

Damit ist die Konvergenz des Finite-Differenzen-Verfahrens gezeigt. \square

Bemerkung 6.4.

1. Der Konvergenzbeweis

$$\|e_h\| \leq Ch^2 \|u\|_{C^4(\bar{\Omega})}$$

erfordert $u \in C^4(\bar{\Omega})$. Dies ist eine sehr starke Forderung, da eine klassische Lösung ja nur $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$ erfüllt. Es zeigt sich über den Umweg des Finite-Elemente-Verfahrens, dass die Diskretisierung auch bei $u \notin C^4(\bar{\Omega})$ anwendbar bleibt. Dies lässt sich eben auf dem hier gezeigten Weg aber nicht beweisen.

2. Die Konvergenzordnung h^2 basiert auf äquidistanten Gittern in jeder Richtung. Bei nicht äquidistanten Gittern gilt nur h^1 .
3. Es gibt Verfahren (Finite Elemente, bestimmte Finite Volumen), die auch auf nicht äquidistanten Gittern optimale Konvergenzordnung erreichen.
4. Der Konvergenzbeweis lässt sich auf allgemeinere PDGL's z. B. mit nichtkonstantem Koeffizient sowie gemischten und niedrigeren Ableitungen übertragen. Wesentlich ist, dass L_h eine M-Matrix bleibt. \square

6.4 Diskrete Mittelwerteigenschaft und Maximumprinzip

Die Lösung der Laplace-Gleichung erfüllt ein Maximumprinzip. Wir wollen in diesem Abschnitt zeigen, dass die M-Matrix Eigenschaft ein Maximumprinzip für die diskrete Lösung liefert (diskretes Maximumprinzip).

Für $f(x, y) = 0$ (Laplace-Gleichung) liefert Auflösen der Differenzgleichung (6.2a)

$$u_h(x, y) = \frac{1}{4} [u_h(x - h, y) + u_h(x, y - h) + u_h(x + h, y) + u_h(x, y + h)] \quad \forall (x, y) \in \Omega_h, \quad (6.6)$$

also eine diskrete Mittelwerteigenschaft.

Damit folgt wie im kontinuierlichen Fall ein Maximumprinzip:

Satz 6.5. Sei $u_h(x)$ eine nichtkonstante Lösung der diskreten Laplacegleichung. Die Extrema $\max\{u_h(x) \mid x \in \bar{\Omega}_h\}$ und $\min\{u_h(x) \mid x \in \bar{\Omega}_h\}$ werden nicht auf Ω_h , sondern auf Γ_h angenommen.

Beweis: Wäre u_h maximal in $(x, y) \in \Omega_h$, so muss wegen (6.6) auch $u_h(x \pm h, y \pm h) = u(x, y)$ gelten. Wegen der Irreduzibilität von L_h (alle Punkte sind verbunden) folgt $u_h = \text{const}$ im $\bar{\Omega}_h$ zur Voraussetzung. \square

Die Beziehung (6.6) lässt sich auf allgemeinere Diskretisierungen übertragen. Für $a_{\alpha\alpha} \neq 0, \alpha \in I$ gilt

$$\forall \alpha : \sum_{\beta \in I} a_{\alpha\beta} u_\beta = \overset{\text{keine Quellen/Senken}}{\downarrow} 0 \iff u_\alpha = \frac{1}{a_{\alpha\alpha}} \sum_{\beta \neq \alpha} -a_{\alpha\beta} u_\beta = \sum_{\beta \neq \alpha} -\frac{a_{\alpha\beta}}{a_{\alpha\alpha}} u_\beta$$

Ist A M-Matrix, so gilt

$$-\frac{a_{\alpha\beta}}{a_{\alpha\alpha}} \geq 0 \quad (a_{\alpha\beta} \leq 0, a_{\alpha\alpha} > 0).$$

Ist A irreduzibel diagonaldominant, so gilt

$$\sum_{\beta \neq \alpha} -\frac{a_{\alpha\beta}}{a_{\alpha\alpha}} \leq 1.$$

Diskretisierungen mit irreduzibel diagonaldominanter M-Matrix erfüllen also ein diskretes Maximumprinzip.

6.5 Eigenwerte, Eigenvektoren

Sei L_h wieder die Matrix aus dem Fünfpunktstern für $-\Delta u = 0$ in $\Omega = (0, 1)^2$. Dann hat L_h die $(N - q)^2$ Eigenvektoren $R_h u^{\nu\mu}$ mit

$$u^{\nu\mu}(x, y) = \sin(\nu\pi x) \sin(\mu\pi y) \quad \text{für } 1 \leq \nu, \mu \leq N - 1.$$

Die dazugehörigen Eigenwerte sind

$$\lambda^{\nu\mu} = \frac{4}{h^2} \left(\sin^2 \left(\frac{\nu\pi h}{2} \right) + \sin^2 \left(\frac{\mu\pi h}{2} \right) \right).$$

6 Konvergenz des Finite-Differenzen-Verfahrens

Beweis: Zeige nur 1d: $u^\nu = \sin(\nu\pi x)$.

$$\begin{aligned}
 -\Delta_h \cdot u^\nu &= \frac{1}{h^2} [-\sin(\nu\pi(x-h)) + 2\sin(\nu\pi x) - \sin(\nu\pi(x+h))] \\
 &= \frac{1}{h^2} \left\{ -[\sin(\nu\pi x)\cos(\nu\pi h) - \cos(\nu\pi x)\sin(\nu\pi h)] + 2\sin(\nu\pi x) \right. \\
 &\quad \left. - [\sin(\nu\pi x)\cos(\nu\pi h) + \cos(\nu\pi x)\sin(\nu\pi h)] \right\} \\
 &= \frac{1}{h^2} 2\sin(\nu\pi x) \underbrace{[1 - \cos(\nu\pi h)]}_{1 - \cos(\nu\pi h) = 2\sin^2\left(\frac{\nu\pi h}{2}\right)} \\
 &= \frac{4}{h^2} \sin^2\left(\frac{\nu\pi h}{2}\right) \sin(\nu\pi x) = \frac{4}{h^2} \sin^2\left(\frac{\nu\pi h}{2}\right) u^\nu
 \end{aligned}$$

$$\begin{aligned}
 &\sin(a \pm b) \\
 &\sin(a)\cos(\pm b) \pm \cos(a)\sin(\pm b)
 \end{aligned}$$

Multiplikation mit $\sin(\mu\pi y)$ liefert die x -Richtung, und die y -Richtung geht analog.

Wie sehen außerdem:

- Alle Eigenwerte sind reell und positiv $\rightarrow L_h$ ist symmetrisch und positiv definit (dies kann man auch algebraisch zeigen, ohne die Eigenwerte zu kennen. Symmetrie ist ohnehin klar.
- Es ist

$$\left. \begin{aligned}
 \lambda_{max} &= \max_{\nu\mu} \lambda^{\nu\mu} \leq \frac{4}{h^2} \overset{\sin x \in I}{\downarrow} (1 + 1) = \frac{8}{h^2} \\
 \lambda_{min} &= \min_{\nu\mu} \lambda^{\nu\mu} \geq \frac{4}{h^2} \left(2\sin^2\left(\frac{\pi h}{2}\right) \right) \\
 &\geq \frac{8}{h^2} \frac{\pi^2 h^2}{4} \cdot C > 0
 \end{aligned} \right\} \begin{aligned}
 \kappa(L_h) &= \frac{\lambda_{max}}{\lambda_{min}} = O(h^{-2}) \\
 &\rightarrow \infty \text{ für } h \rightarrow 0.
 \end{aligned}$$

6.6 Zusammenfassung

- In diesem Abschnitt haben wir die Konvergenz des Finite-Differenzen Verfahrens bewiesen. Wesentlich war dabei die Aufspaltung in (lokale) Konsistenz, die aus der Taylorreihenentwicklung folgt, sowie Stabilität, die wesentlich auf der M-Matrixeigenschaft fusst.
- Schließlich kann man aus der M-Matrixeigenschaft auch ein diskretes Maximumprinzip folgern.
- Obwohl wir uns hier im wesentlichen auf den Fünfpunktstern aus der Diskretisierung des Laplaceoperators in zwei Raumdimensionen beschränkt haben lässt sich das auf allgemeine M-Matrizen und n Raumdimensionen verallgemeinern.

7 Zellenzentrierte Finite Volumen

In der Praxis (z. B. bei der Strömungssimulation in porösen Medien) trifft man häufig auf Gleichungen mit ortsvariablen Koeffizienten. Wollen wir das Finite-Differenzen Verfahren auf

$$-\nabla \cdot \{k(x)\nabla u\} = f \text{ in } \Omega, \quad u = g \text{ auf } \partial\Omega,$$

mit einer skalaren Koeffizientenfunktion $k(x)$ anwenden, so können wir die Differenzenformeln für u zunächst nicht direkt einsetzen.

Die eben aufgeführte Gleichung ist in sog. *Erhaltungsform* wie sie üblicherweise in der Modellierung auftritt. Durch Anwendung der Produktregel erhalten wir

$$-\nabla \cdot \{k(x)\nabla u\} = -k(x) \sum_{i=1}^d \partial_{x_i} \partial_{x_i} u - \sum_{i=1}^d (\partial_{x_i} k(x)) (\partial_{x_i} u),$$

also eine Gleichung mit Termen erster und zweiter Ordnung auf die nun das Finite-Differenzen Verfahren angewendet werden kann.

Diese Vorgehen erfordert die Differenzierbarkeit der Koeffizientenfunktion $k(x)$. Oft hat man in der Praxis aber nur ein stückweise konstantes k .

In diesem Abschnitt wollen wir ein Verfahren welches auf stückweise konstante Diffusionskoeffizienten anwendbar ist und welches auch einige weitere Vorteile besitzt.

7.1 Problemstellung und Gitterkonstruktion

Wir schreiben die Diffusionsgleichung mit isotropem aber ortsabhängigem Diffusionskoeffizienten in der Form

$$\begin{aligned} \nabla \cdot \sigma &= f && \text{in } \Omega \\ \sigma &= -k(x)\nabla u \\ u &= g && \text{auf } \Gamma_D \subseteq \partial\Omega \\ \sigma \cdot \nu &= \varphi && \text{auf } \Gamma_N = \partial\Omega - \Gamma_D \end{aligned} \tag{7.1}$$

mit $k(x) \geq k_0 > 0$.

Wie bei den Finiten Differenzen sei nun $\Omega = (0, 1)^d$ mit einem äquidistanten Gitter der Schrittweite $h = \frac{1}{N}$, $N \in \mathbb{N}$ überzogen.

Inhaltsverzeichnis

Wir betrachten nun die inneren Gitterpunkte

$$\Omega_h = \left\{ (x_1, \dots, x_d) \in \Omega \mid \frac{x_i}{h} - \frac{1}{2} \in \mathbb{Z} \right\}$$

sowie die Gitterpunkte am Rand

$$\Gamma_h = \left\{ (x_1, \dots, x_d) \in \partial\Omega \mid \exists i \in \{1, \dots, d\} : \frac{x_i}{h} - \frac{1}{2} \in \mathbb{Z} \right\}$$

und setzen

$$\bar{\Omega}_h = \Omega_h \cup \Gamma_h.$$

Wir definieren die Indexmenge

$$I_h = \{1, \dots, \#\bar{\Omega}_h\}$$

und numerieren damit die Gitterpunkte

$$x : I_h \rightarrow \bar{\Omega}_h, \quad x_i := x(i) \quad \forall i \in I_h.$$

Die Indexmenge kann partitioniert werden in innere Punkte und Randpunkte:

$$I_h = I_h^\Omega \cup I_h^\Gamma, \quad I_h^\Omega = \{i \in I_h \mid x_i \in \Omega\}, \quad I_h^\Gamma = \{i \in I_h \mid x_i \in \partial\Omega\}.$$

Jedem Gitterpunkte wird nun eine Zelle C_i zugeordnet und zwar getrennt für Rand- und innere Gitterpunkte:

$$\begin{aligned} i \in I_h^\Omega : & \quad C_i = \{x \in \Omega \mid \|x - x_i\| < \|x - x_j\| \quad \forall j \in I_h^\Omega\} \subset \Omega, \\ i \in I_h^\Gamma : & \quad C_i = \{x \in \partial\Omega \mid \|x - x_i\| < \|x - x_j\| \quad \forall j \in I_h^\Gamma\} \subset \partial\Omega. \end{aligned}$$

Die Zellen zerlegen also sowohl Ω als auch $\partial\Omega$. Für jede innere Zelle C_i , $i \in I_h^\Omega$ bezeichne

$$\gamma_{ij} = \begin{cases} \partial C_i \cap \partial C_j & \text{falls } j \in I_h^\Omega \\ \partial C_i \cap C_j & \text{falls } j \in I_h^\Gamma \end{cases}$$

den Schnitt zweier Zellen. Für unser strukturiertes Gitter gilt

$$\int_{C_i} 1 \, dx = h^d \quad \text{und} \quad \int_{\gamma_{ij}} 1 \, dx = h^{d-1}.$$

Damit können wir die Nachbarzellen definieren. Zu jeder innere Zelle $i \in I_h^\Omega$ sei

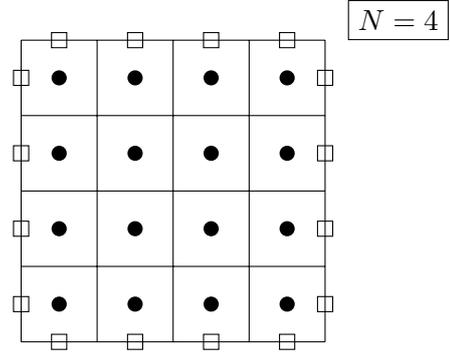
$$N_i = \{i \in I_h^\Omega \mid \gamma_{ij} \text{ hat Dimension } d-1\}$$

die Menge der inneren Nachbarn und

$$B_i = \{i \in I_h^\Gamma \mid \gamma_{ij} \text{ hat Dimension } d-1\}$$

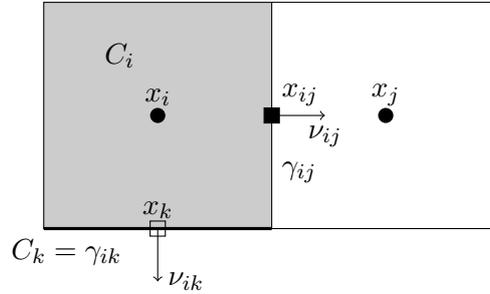
die Menge der Randnachbarn. Die Randnachbarn können wir noch partitionieren in

$$B_i^D = \{j \in B_i \mid x_j \in \Gamma_D\} \quad \text{und} \quad B_i^N = \{j \in B_i \mid x_j \in \Gamma_N\}.$$



Schließlich bezeichne ν_{ij} noch für jede innere Zelle die Einheitsnormale auf γ_{ij} , die von Zelle C_i nach aussen zeigt.

Alle Definitionen sind nochmal in folgender Zeichnung zusammengefasst:



7.2 Finite Volumen

Damit kommen wir nun endlich zum Finite-Volumen Verfahren, genauer zu seiner „zellenzentrierten“ Variante.

Wir integrieren die erste Gleichung in (7.1) über jede innere Zelle:

$$\int_{C_i} \nabla \cdot \sigma \, dx = \int_{C_i} f \, dx \quad \forall i \in I_h^\Omega. \tag{7.2}$$

Für die linke Seite rechnen wir:

$$\begin{aligned} \int_{C_i} \nabla \cdot \sigma \, dx &\stackrel{\text{Gauß}}{=} \int_{\partial C_i} \sigma \cdot \nu \, ds \\ &= \sum_{j \in N_i} \int_{\gamma_{ij}} \sigma \cdot \nu_{ij} \, ds + \sum_{j \in B_i} \int_{\gamma_{ij}} \sigma \cdot \nu_{ij} \, ds \\ &\stackrel{\text{Mittelpunktregel}}{=} \sum_{j \in N_i} h^{d-1} \sigma \cdot \nu_{ij} + \sum_{j \in B_i} h^{d-1} \sigma \cdot \nu_{ij} + \text{Fehler} \\ &\stackrel{\text{Differenzen}}{=} \sum_{j \in N_i} h^{d-1} \left[-k(x_{ij}) \frac{u(x_j) - u(x_i)}{h} \right] \\ &\quad + \sum_{j \in B_i^D} h^{d-1} \left[-k(x_j) \frac{u(x_j) - u(x_i)}{h/2} \right] + \sum_{j \in B_i^N} h^{d-1} \varphi(x_j) + \text{Fehler}. \end{aligned}$$

Für die rechte Seite wendet man auch die Mittelpunktsregel an:

$$\int_{C_i} f \, dx = h^d f(x_i) + \text{Fehler}.$$

Inhaltsverzeichnis

Unter Ignorieren der Fehlerterme und Kürzen entsprechender h -Potenzen erhalten wir folgende diskrete Gleichung für die unbekannte Gitterfunktion $u_h : \bar{\Omega}_h \rightarrow \mathbb{R}$:

$$\left(\sum_{j \in N_i} k(x_{ij}) + \sum_{j \in B_i^D} 2k(x_j) \right) u_h(x_i) - \sum_{j \in N_i} k(x_{ij}) u_h(x_j) - \sum_{j \in B_i^D} 2k(x_j) u_h(x_j) = h^2 f(x_i) - h \sum_{j \in B_i^N} \varphi(x_j) \quad i \in I_h^\Omega. \quad (7.3)$$

Dies entspricht wieder einem linearen Gleichungssystem für die unbekanntesten Werte diesmal in den Zellmittelpunkten (deswegen zellenzentrierte Finite Volumen).

Die Matrix ist bei mindestens einem Dirichletrandpunkt irreduzibel diagonaldominant und erfüllt das Vorzeichenmuster. Mithin ist die entstehende Matrix eine M-Matrix.

Leider ist die Konvergenzanalyse des Verfahrens nicht so einfach wie bei Finiten Differenzen. Eine simple Analyse des Konsistenzfehlers in obiger Herleitung (Mittelpunktsregel, zentrale Differenz für erste Ableitung) führt nur auf $O(h)$ in randfernen Zellen und gar nur $O(1)$ in randnahen Zellen. Trotzdem ist die Konvergenzordnung des Verfahrens $O(h^2)$ auf äquidistanten Gittern. Dies erfordert allerdings andere Beweistechniken als wir sie hier kennengelernt haben.

Im Fall $k(x) \equiv 1$ und zwei Raumdimensionen führt das Verfahren bis auf den Faktor h^{-2} (der sich nun auf der rechten Seite befindet) auf den bekannten Stern

$$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}.$$

Am Rand ergeben sich allerdings andere Sterne. An den Rändern links unten und unten ergibt sich für Dirichlet Randbedingungen

$$\begin{bmatrix} 0 & -1 & 0 \\ 0 & 6 & -1 \\ 0 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & 0 & 0 \end{bmatrix}.$$

und für Neumann Randbedingungen

$$\begin{bmatrix} 0 & -1 & 0 \\ 0 & 2 & -1 \\ 0 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & -1 & 0 \\ -1 & 3 & -1 \\ 0 & 0 & 0 \end{bmatrix}.$$

7.3 Zellweise Permeabilität

Wir betrachten nun den Fall, dass die Funktion $k(x)$ stückweise konstant auf jeder Zelle C_i ist.

Im Finite-Volumen-Verfahren ist $k(x_{ij})$ genau auf den Zellrändern, also den Unstetigkeitsstellen auszuwerten. Welchen Wert von k soll man dann nehmen?

Betrachte die 1D-Situation:

$$\begin{aligned} \frac{d\sigma}{dx} &= 0 && \text{Länge } l \\ &&& \downarrow \\ &&& \text{in } \Omega = (0, l) \\ \sigma &= -k(x) \frac{du}{dx} \\ u(0) &= u_0 && \boxed{\text{--- } k(x) \text{ ---}} \\ u(l) &= u_l && \begin{array}{c} u_0 \qquad \qquad \qquad u_l \end{array} \end{aligned}$$

Aus $\frac{d\sigma}{dx} = 0$ in Ω folgt $\sigma(x) = \Sigma \in \mathbb{R}$ (das gilt nur in einer Raumdimension).

Also gilt für die zweite Gleichung

$$\begin{aligned} \Sigma = -k(x) \frac{du}{dx} &\iff \frac{du}{dx} = -\frac{\Sigma}{k(x)} \\ &\iff \int_0^l \frac{du}{dx} dx = [u(x)]_0^l = p_l - p_0 = -\Sigma \int_0^l \frac{1}{k(x)} dx \\ &\iff \underbrace{\Sigma}_{\text{Fluss}} = - \underbrace{\frac{l}{\int_0^l \frac{1}{k(x)} dx}}_{\substack{\text{effektive} \\ \text{Permeabilität} \\ \text{harmonisches Mittel}}} \cdot \underbrace{\frac{p_l - p_0}{l}}_{\text{Gradient}} \end{aligned}$$

Speziell für den Fall

$$k(x) = \begin{cases} k_l & x \leq \frac{l}{2} \\ k_r & x > \frac{l}{2} \end{cases} \quad \begin{array}{c} \boxed{k_l \quad \quad \quad k_r} \\ 0 \qquad \qquad \qquad l \end{array}$$

ergibt sich

$$k_{eff} = \frac{l}{\int_0^l k^{-1}(x) dx} = \frac{l}{\frac{l}{2} \cdot \frac{1}{k_l} + \frac{l}{2} \cdot \frac{1}{k_r}} = \frac{2}{\frac{1}{k_l} + \frac{1}{k_r}}$$

(harmonisches Mittel).

Deshalb wählt man bei zellweise konstanten Permeabilitäten

$$k(x_{ij}) = \frac{2}{\frac{1}{k(x_i)} + \frac{1}{k(x_{ij})}}$$

Zu den Mittelwerten: Die drei geläufigsten Mittelwerte von n Zahlen sind das harmonische, geometrische und arithmetische Mittel:

$$\begin{aligned} \bar{k}_{\text{harm}}(k_1, \dots, k_n) &= \frac{n}{\sum_{i=1}^n \frac{1}{k_i}}, \\ \bar{k}_{\text{geom}}(k_1, \dots, k_n) &= \sqrt[n]{k_1 \cdot k_2 \cdot \dots \cdot k_n} \\ \bar{k}_{\text{arith}}(k_1, \dots, k_n) &= \frac{k_1 + k_2 + \dots + k_n}{n}. \end{aligned}$$

Es gilt

$$\bar{k}_{\text{harm}}(k_1, \dots, k_n) \leq \bar{k}_{\text{geom}}(k_1, \dots, k_n) \leq \bar{k}_{\text{arith}}(k_1, \dots, k_n).$$

. Das arithmetische Mittel entspricht der Parallelschaltung von Widerständen und das harmonische Mittel der Reihenschaltung von Widerständen (und ist deshalb hier das geeignete).

7.4 Diskrete Erhaltungseigenschaft

Die diskrete Lösung $u_h(x)$ erfüllt nach Konstruktion ein diskretes, lokales Erhaltungsprinzip:

$$\int_{C_i} \nabla \cdot \sigma_h \, dx = \overbrace{\int_{\partial C_i} \sigma_h \cdot \nu \, ds}^{\text{Fluss über Rand}} = \overbrace{\int_{C_i} f \, dx}^{\text{Zufuhr über } Q/S \text{ in } C_i} \quad i \in I_h^\Omega$$

$$\parallel$$

$$\sum_{j \in N_i} \int_{\gamma_{ij}} \sigma_h \cdot \nu_{ij} \, ds$$

Bei der Herleitung der Erhaltungsgleichung haben wir so eine Bilanz für beliebige $\omega \subseteq \Omega$ erhalten und daraus nach Anwendung des Gaußschen Integralsatzes auf die partielle Differentialgleichung geschlossen. Bei Finite-Volumen-Verfahren macht man diesen letzten Schritt wieder rückgängig und fordert die Bilanz der physikalischen Größe (z. B. Masse oder Energie) auf einer diskreten Menge von Volumina, den Zellen C_i .

In diesem Sinne gilt also für die diskrete Lösung auch ein Erhaltungsprinzip, man spricht von diskreter Erhaltung. Praktisch bedeutet dies, dass in einem Finite-Volumen-Verfahren (bis auf Rundungsfehler) keine Masse, Energie, Impuls, . . . , „verloren“ gehen kann. Bei Finite-Differenzen-Verfahren gilt dies nur bis auf den Diskretisierungsfehler.

Die weitere Bedeutung des diskreten Erhaltungsprinzips wird sich uns erst bei den hyperbolischen Differentialgleichungen (erster Ordnung) erschließen.

Summiert man über mehrere zusammenhängende Zellen so heben sich alle internen Flüsse heraus. Betrachte zwei Zellen $i, j, j \in N_i$, dann gilt:

$$\int_{C_i} \nabla \cdot \sigma_h \, dx + \int_{C_j} \nabla \cdot \sigma_h \, dx$$

$$= \dots \int_{\gamma_{ij}} \sigma_h \cdot \nu_{ij} \, ds + \dots + \dots \int_{\gamma_{ji}} \sigma_h \cdot \nu_{ji} \, ds + \dots$$

$$= \int_{\partial(C_i \cup C_j)} \sigma_h \cdot \nu$$

da $\nu_{ij} = -\nu_{ji}$ und σ_h in gleicher Weise auf beiden Seiten bestimmt wird. Dies lässt sich rekursiv auch auf eine größere Menge von Zellen erweitern.

7.5 Erweiterung auf unstrukturierte Gitter

Das beschriebene Verfahren lässt sich unmittelbar auf unstrukturierte Gitter verallgemeinern. Dazu sei

$$T_h = \{t_1, \dots, t_N\}$$

ein *konformes* Dreiecksgitter, d. h. :

1. Der Schnitt von zwei verschiedenen Dreiecken $t_i \cap t_j$ ist entweder leer, ein gemeinsamer Knoten oder eine gemeinsame Kante.

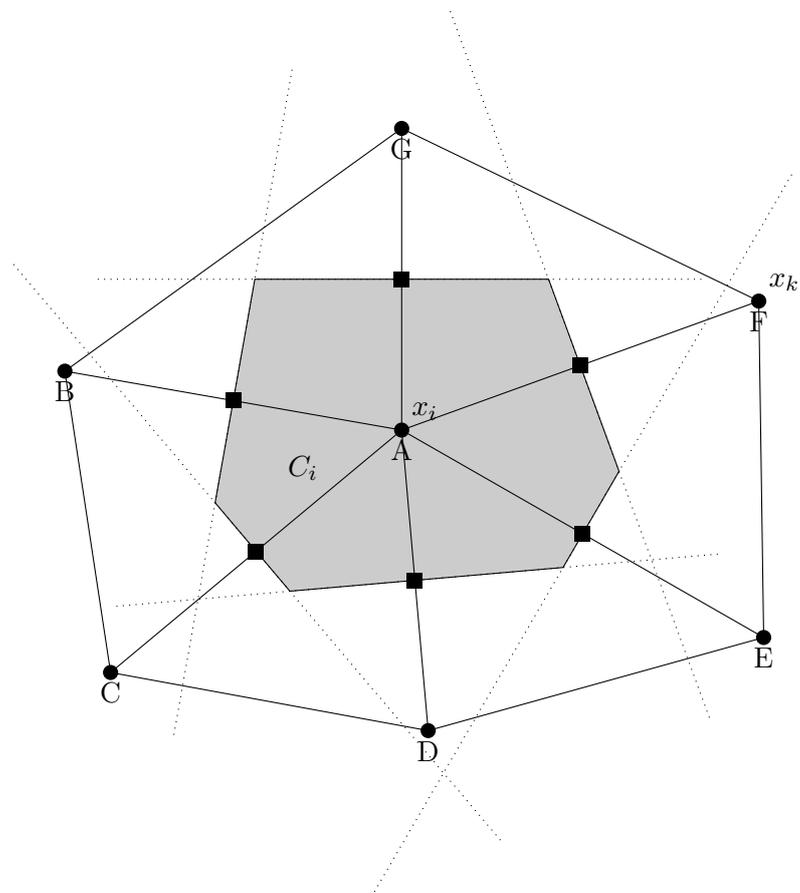
Und der *Delaunay*¹⁴-*Eigenschaft*:

2. Alle Innenwinkel der Dreiecke sind höchstens 90 Grad.

Die Zellen können dann auf die oben beschriebene Weise definiert werden und werden in diesem Zusammenhang Voronoi¹⁵-Diagramm genannt. Hier eine Zeichnung:

¹⁴Boris Nikolajewitsch Delone, 1890-1980, russ. Mathematiker.

¹⁵Georgi Feodosjewitsch Woronoi, 1868-1908, russischer Mathematiker ukrainischer Herkunft.



In diesem Zusammenhang wird das Verfahren auch Boxmethode oder Methode der integralen Finiten Differenzen genannt. Auch viereckförmige Zellen sind unter gewissen Einschränkungen an die Innenwinkel möglich.

7.6 Zusammenfassung

- In diesem Abschnitt haben wir ein Diskretisierungsverfahren kennengelernt, das sich für die Diffusionsgleichung mit ortsvariablem, insbesondere stückweise konstantem Diffusionskoeffizienten eignet.
- Es lässt sich einfach auf gewisse unstrukturierte Gitter verallgemeinern, was die Lösung der Differentialgleichung in komplizierteren Gebieten ermöglicht.
- Eine Schwierigkeit bei dem vorgestellten Verfahren ist die Erweiterung auf allgemeine Diffusionstensoren $K(x)$. Eine Lösung sind sog. „multipoint flux approximations“.

8 Relaxationsverfahren

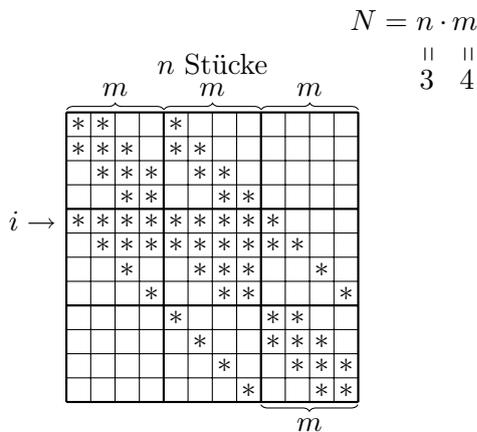
8.1 Dünn besetzte Matrizen und direkte Lösungsverfahren

Wir untersuchen die Anwendung der Gauß-Elimination auf

$$Ax = b, \quad A \in \mathbb{R}^{N \times N} \text{ regulär}, \quad x, b \in \mathbb{R}^N,$$

wobei $A = L_h$ unsere Matrix aus dem FD-Verfahren sein soll.

Bei lexikographischer Anordnung ist A eine Bandmatrix. Dabei entstehe A aus einem Gitter mit n Zeilen zu je m Gitterpunkten. Die Matrix hat also $N = nm$ Zeilen und die Bandbreite beträgt m .



- A lässt sich ohne Pivotisierung eliminieren, da s. p. d.
- Es werden im Laufe der Elimination neue Nichtnullelemente erzeugt \rightarrow „Fill in“.
- Diese entstehen nur innerhalb der äußersten Diagonalen.
- Das Bild zeigt die Situation wenn i die Pivotzeile ist.

Die Anzahl der Nichtnullelemente von A ist $O(N) = O(n \cdot m)$ (denn nm ist die Anzahl der Zeilen in A mal eine konstante Zahl von Einträgen). Daraus werden nach der Elimination $O(n \cdot m \cdot m) = O(nm^2)$ Matrixeinträge, da jede der nm Einträge dann bis zu $2m$ Einträge hat. Falls $n = m$ (quadratisches Gitter) sind dies also $O(n^3)$ entsprechend $O(N^{\frac{3}{2}})$ (N ist die Anzahl der Unbekannten).

Der Aufwand A in Fließkommaoperationen für die Elimination ist

$$A \leq \sum_{i=1}^{nm} \underset{\substack{\# \text{ zu eliminierende} \\ \text{Elemente bis zur Diagonale} \\ \text{in Zeile } i}}{m} \cdot \underset{\substack{\text{untere Schranke für} \\ \text{Elimination eines Elements} \\ \text{(eigentlich } 2m \dots m)}}{m} = nmm^2 = nm^3.$$

Für $n = m$ (quadratisches Gitter) ist der Aufwand für die Elimination also $O(n^4)$ entsprechend $O(N^2)$ mit N der Zahl der Unbekannten. Immerhin ist das besser als $O(N^3)$ für die LU -Zerlegung vollbesetzter Matrizen.

Diese Betrachtung gilt nur für zwei Raumdimensionen. In drei Raumdimensionen findet man den Aufwand $O(N^{8/3})$ für den Bandlöser.

Bemerkung 8.1. Es geht auch in $O(N^{\frac{3}{2}})$ in zwei Raumdimensionen wenn man die Nummerierung optimal wählt. Dies ist die sog. *Nested Dissection*. Siehe auch die Grundlagenvorlesung „Modellbildung und Simulation“. □

8 Relaxationsverfahren

Iterative Lösungsverfahren konstruieren ausgehend von $x^{(0)} \in \mathbb{R}^N$ dem sog. „Startwert“, eine Folge von „Iterierten“

$$x^{(0)}, x^{(1)}, \dots, x^{(k)}, \dots \quad \text{mit} \quad \lim_{k \rightarrow \infty} x^{(k)} \rightarrow x.$$

Der Schritt von $x^{(k)}$ nach $x^{(k+1)}$ hat dabei typischerweise Aufwand $O(N)$, dafür ist die entscheidende Frage nach wievielen Schritten man die Iteration abbricht, man also den „Iterationsfehler“ $\|x - x^{(k)}\|$ akzeptiert.

8.2 Relaxationsverfahren

Relaxationsverfahren sind eine einfache Klasse von Iterationsverfahren.

In $Ax = b$ isolieren wir die i -te Gleichung

$$\sum_{j=1}^N a_{ij}x_j = b_i$$

und lösen nach x_i auf.

$$x_i = \frac{1}{a_{ii}} \left(b_i - \sum_{j \neq i} a_{ij}x_j \right) \quad (8.1)$$

Voraussetzung: offensichtlich $a_{ii} \neq 0$.

Hiermit ergeben sich verschiedene Verfahren:

gedämpftes Jacobi- oder Gesamtschrittverfahren:

$$x_i^{(k+1)} = \underbrace{(1 - \omega)x_i^{(k)}}_{\text{alter Wert}} + \underbrace{\frac{\omega}{a_{ii}} \left(b_i - \sum_{i \neq j} a_{ij} \overbrace{x_j^{(k)}}^{\substack{\text{nur Werte aus der} \\ \text{letzten Iteration}}} \right)}_{\text{„neuer“ Wert nach (8.1)}} \quad i = 1, \dots, N$$

und $\omega \in (0, 1]$. $\omega = 1$: Jacobi-Verfahren.

SOR-Verfahren (successive overrelaxation)

$$x_i^{(k+1)} = (1 - \omega)x_i^{(k)} + \frac{\omega}{a_{ii}} \left(b_i - \sum_{j < i} a_{ij} \overbrace{x_j^{(k+1)}}^{\text{möglichst neue Werte}} - \sum_{j > i} a_{ij}x_j^{(k)} \right) \quad i = 1, \dots, N$$

und $\omega \in (0, 2)$. $\omega = 1$: Gauß-Seidel-Verfahren

$\omega < 1$: Unterrelaxation

$\omega > 1$: Überrelaxation

8.3 Matrixschreibweise der Relaxationsverfahren

Eine kompaktere Schreibweise und der Zugang zu Analyse der Iterationsverfahren ergibt sich mittels Matrixdarstellung.

Dazu zerlege

$$A = \begin{array}{c} L \\ \text{strikte untere} \end{array} + \begin{array}{c} D \\ \text{Diagonale} \end{array} + \begin{array}{c} U \\ \text{strikte obere} \end{array}$$

$$l_{ij} = \begin{cases} a_{ij} & i > j \\ 0 & \text{sonst} \end{cases} \quad d_{ij} = \begin{cases} a_{ij} & i = j \\ 0 & \text{sonst} \end{cases} \quad u_{ij} = \begin{cases} a_{ij} & i < j \\ 0 & \text{sonst} \end{cases}$$

Für das gedämpfte Jacobi-Verfahren erhalten wir:

$$x_i^{(k+1)} = (1 - \omega)x_i^{(k)} + \frac{\omega}{a_{ij}} \underbrace{\left(b_i - \sum_{j \neq i} a_{ij}x_j^{(k)} \right)}_{i = 1, \dots, n}$$

$$\begin{aligned} \Leftrightarrow x^{(k+1)} &= (1 - \omega)x^{(k)} + \omega D^{-1} \left(b - (L + U)x^{(k)} \right) \\ &= x^{(k)} - \omega D^{-1} D x^{(k)} + \omega D^{-1} \left(b - (L + U)x^{(k)} \right) \\ &= x^{(k)} + \omega D^{-1} \underbrace{\left(b - A x^{(k)} \right)}_{=: d^{(k)} \text{ „Defekt“}} \end{aligned}$$

und so die SOR-Iteration:

$$x_i^{(k+1)} = (1 - \omega)x_i^{(k)} + \frac{\omega}{a_{ii}} \left(b_i - \sum_{j < i} a_{ij}x_j^{(k+1)} - \sum_{j > i} a_{ij}x_j^{(k)} \right) \quad i = 1, \dots, N$$

$$\Leftrightarrow \omega \sum_{j < i} a_{ij}x_j^{(k+1)} + a_{ii}x_i^{(k+1)} = a_{ii}(1 - \omega)x_i^{(k)} + \omega \left(b_i - \sum_{j > i} a_{ij}x_j^{(k)} \right) \quad i = 1, \dots, N$$

$$\Leftrightarrow \omega L x^{(k+1)} + D x^{(k+1)} = (1 - \omega) D x^{(k)} + \omega \left(b - U x^{(k)} \right)$$

8 Relaxationsverfahren

$$\begin{aligned}
 \Leftrightarrow x^{(k+1)} &= (1 - \omega)(\omega L + D)^{-1} D x^{(k)} + \omega(\omega L + D)^{-1} (b - U x^{(k)}) \\
 &= (1 - \omega)(\omega L + D)^{-1} D x^{(k)} \\
 &\quad + \underbrace{\omega(\omega L + D)^{-1} (L + D) x^{(k)} - \omega(\omega L + D)^{-1} (L + D) x^{(k)}}_{=0} \\
 &\quad + \omega(\omega L + D)^{-1} (b - U x^{(k)}) \\
 &= (\omega L + D)^{-1} \left[\underbrace{(1 - \omega)D + \omega(L + D)}_{D - \omega D + \omega L + \omega D} \right] x^{(k)} \\
 &\quad + \underbrace{\omega(\omega L + D)^{-1}}_{(L + \frac{1}{\omega} D)^{-1}} (b - A x^{(k)}) \\
 &= x^{(k)} + \left(L + \frac{1}{\omega} D \right)^{-1} (b - A x^{(k)})
 \end{aligned}$$

Folgende Herleitung zeigt, dass die Formulierung kein Zufall ist.

Definiere den Fehler

$$e^{(k)} := \begin{array}{ccc} & x & \\ \nearrow & & \nwarrow \\ \text{Lösung von } Ax = b & x - x^{(k)} & \text{Iterierte nach Schritt } k. \end{array} \quad (8.2)$$

Es gilt

$$A e^{(k)} = A (x - x^{(k)}) = Ax - A x^{(k)} = b - A x^{(k)} = d^{(k)} \quad (\text{Defektgleichung}). \quad (8.3)$$

Gegeben $x^{(k)}$, so könnte man durch

1. Lösen von $A e^{(k)} = d^{(k)} \Rightarrow e^{(k)} = A^{-1} d^{(k)} = A^{-1} (b - A x^{(k)})$
2. und korrigieren mittels (8.2) $x = x^{(k)} + e^{(k)} = x^{(k)} + A^{-1} (b - A x^{(k)})$

die Lösung x berechnen. Nun ist natürlich Lösen von (8.3) nicht einfacher als $Ax = b$ selbst.

Idee: Ersetze A in (8.3) durch $M \in \mathbb{R}^{N \times N}$, sodass

1. $M \approx A$
2. M einfach invertierbar.

Dies liefert die *Iterationsvorschrift*

$$x^{(k+1)} = x^{(k)} + M^{-1} (b - A x^{(k)})$$

Bemerkung 8.2. Falls A und M regulär sind, ist x der einzige Fixpunkt dieser Iteration. \square

Wir erkennen:

1. $M = \frac{1}{\omega}D$ liefert das gedämpfte Jacobi-Verfahren.
2. $M = L + \frac{1}{\omega}D$ liefert die SOR-Iteration.

Weitere Verfahren sind

1. $M = \frac{1}{\omega}(L + D)$ gedämpftes Gauß-Seidel-Verfahren.
2. $M = \frac{1}{\omega}I$ gedämpfte Richardson-Iteration.
 \uparrow
Einheitsmatrix!

8.4 Konvergenz von linearen Iterationsverfahren

Für die Fortpflanzung des Fehlers gilt:

$$\begin{aligned}
 x^{(k+1)} &= x^{(k)} + M^{-1} (b - Ax^{(k)}) \\
 \iff \underbrace{x - x^{(k+1)}}_{e^{(k+1)}} &= \underbrace{x - x^{(k)}}_{e^{(k)}} - M^{-1} \left(\underbrace{b - Ax^{(k)}}_{Ax} \right) = x - x^{(k)} - M^{-1} (Ax - Ax^{(k)}) \\
 e^{(k+1)} &= \underbrace{(I - M^{-1}A)}_{=:S} e^{(k)}
 \end{aligned}$$

S heißt Iterationsmatrix.

Wegen $e^{(k+1)} = Se^{(k)}$ heißen die Iterationsverfahren auch „linear“. Offensichtlich ist

$$e^{(k)} = S^k e^{(0)}$$

und damit $e^{(k)} \rightarrow 0$, falls $S^{(k)} \rightarrow 0$ für $k \rightarrow \infty$.

Satz 8.3. Ein Iterationsverfahren der Form $x^{(k+1)} = x^{(k)} + M^{-1} (b - Ax^{(k)})$ konvergiert unabhängig vom Startwert genau dann, wenn $\varrho(S) < 1$.

Beweisskizze:

„ \Leftarrow “ Für eine zugeordnete Matrixnorm gilt $\|e^{(k)}\| \leq \|S\|^m \|e^{(0)}\|$. Man kann zeigen, dass, falls $\varrho(S) < 1$, es immer eine Matrixnorm $\|\cdot\|_\varepsilon$ gibt mit $\|S\|_\varepsilon \leq \varrho(S) + \varepsilon < 1$ und somit $\|S\|_\varepsilon^m \rightarrow 0$, also $\|e^{(k)}\|_\varepsilon \rightarrow 0$.

Der Beweis ist konstruktiv: siehe [Ran06, Satz 6.1].

8 Relaxationsverfahren

„ \Rightarrow “ Verfahren konvergent unabhängig vom Startwert.

Sei $w \neq 0$ ein Eigenvektor von S zum betragsgrößten Eigenwert λ . Wähle den Startwert $x^{(0)} = x - w \Rightarrow e^{(0)} = x - x^{(0)} = w$.

Also gilt $e^{(k)} = S^k w = \lambda^k w$.

Aus $e^{(k)} \rightarrow 0$ (Voraussetzung) und $w \neq 0$ folgt $\lambda^k \rightarrow 0$, was $|\lambda| < 1$ impliziert. \square

Aussagen über das Spektrum von S gelingen z. B. im symmetrisch positiv definiten Fall.

Besonders einfach ist die Richardson-Iteration zu analysieren.

Satz 8.4. Sei A symmetrisch und positiv definit. Dann konvergiert die gedämpfte Richardson-Iteration, wenn $\omega \leq \lambda_{\max}(A)$. $\lambda_{\max}(A)$ und $\lambda_{\min}(A)$ bezeichnen den größten bzw. kleinsten Eigenwert von A .

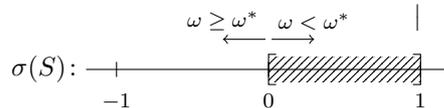
Beweis: A s. p. d.: $\sigma(A) = \{\lambda_{\min}(A) = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_{N-1} \leq \lambda_N = \lambda_{\max}(A)\}$ mit $0 < \lambda_i \in \mathbb{R}$.

Iterationsmatrix $S_{Richardson} = I - \omega A$

Somit ist $\sigma(S_{Richardson}) = \{\mu_i \in \mathbb{R} \mid \mu_i = 1 - \omega \lambda_i, \lambda_i \in \sigma(A)\}$.

Für $\omega^* = \frac{1}{\lambda_{\max}(A)}$ gilt

$$\begin{aligned} \lambda_{\min}(S_{Richardson, \omega^*}) &= 1 - \frac{\lambda_{\max}(A)}{\lambda_{\max}(A)} = 0 \\ \varrho(S_{Richardson, \omega^*}) &= \lambda_{\max}(S_{Richardson, \omega^*}) = 1 - \underbrace{\frac{\lambda_{\min}(A)}{\lambda_{\max}(A)}}_{< 1} < 1 \end{aligned}$$



\square

Bemerkung 8.5. Mit der Konditionszahl $\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} = O(h^{-2})$ erhalten wir

$$\varrho(S_{Richardson, \omega^*}) = 1 - \frac{\lambda_{\min}(A)}{\lambda_{\max}(A)} = 1 - \frac{1}{\kappa(A)}.$$

d.h. der Spektralradius nähert sich quadratisch der 1. \square

Bemerkung 8.6. In der Praxis benötigt man für den Einsatz der Richardson-Iteration eine Schätzung für den größten Eigenwert. Die erhält man leicht mittels des Satzes von Gerschgorin:

$$\lambda_{\max}(A) \leq \max_{i=1, \dots, N} \left(|a_{ii}| + \sum_{j \neq i} |a_{ij}| \right).$$

\square

Bemerkung 8.7 (Aufwandsabschätzung). Wieviele Iterationen benötigt man, um den Anfangsfehler um den Faktor ε zu reduzieren?

Sei $w \neq 0$ ein Eigenvektor zum betragsgrößten Eigenwert λ von S , also

$$Sw = \underbrace{\left(1 - \frac{1}{\kappa(A)}\right)}_{\lambda} w.$$

Weiter nehmen wir an, dass $e^{(0)} = w$. Somit ist

$$\begin{aligned} e^{(k)} &= S^k e^{(0)} = \lambda^k e^{(0)} \\ \Leftrightarrow \|e^{(k)}\| &= |\lambda|^k \|e^{(0)}\| \\ \Leftrightarrow \frac{\|e^{(k)}\|}{\|e^{(0)}\|} &= |\lambda|^k = \varepsilon \\ \Leftrightarrow k \underbrace{\log|\lambda|}_{<0 \text{ wegen } |\lambda| < 1} &= \log \varepsilon \\ \Leftrightarrow k = \frac{\log \varepsilon}{\log|\lambda|} &= \frac{\log \varepsilon}{\log\left(1 - \underbrace{\frac{1}{\kappa(A)}}_{\text{klein}}\right)} \approx \frac{\overbrace{\log \varepsilon}^{\varepsilon < 1, \text{ also negativ}}}{\underbrace{-\frac{1}{\kappa(A)}}_{\text{Taylorntw.}}} = \kappa(A) |\log \varepsilon| \\ & \stackrel{\kappa(A) = Ch^{-2}}{=} Ch^{-2} |\log \varepsilon| \end{aligned}$$

Für $d = 2$ und $h = \frac{1}{\sqrt{N}}$ gilt also $k = O(N)$. Aufwand für eine Iteration ist ebenfalls $O(N)$, also ist der Gesamtaufwand $O(N^2)$ in $d = 2$.

Für $d = 3$ ist $h = \frac{1}{N^{\frac{1}{3}}}$ $\Rightarrow k = O(N^{\frac{2}{3}})$

\Rightarrow Gesamtaufwand $O(N^{\frac{5}{3}})$, besser als Band-Gauß! Hier verbessert sich also die Komplexität mit steigender Raumdimension im Gegensatz zu den direkten Lösern. Man überlege sich woran das liegt. □

Nun wollen wir noch die Konvergenz des Jacobi-Verfahrens untersuchen.

Satz 8.8. Seien A und $2D - A$ (D wie immer die Diagonale von A) symmetrisch positiv definit. Dann konvergiert die Jacobi-Iteration.

Beweis:

- Bezeichne $\langle x, y \rangle = \sum_{i=1}^N x_i y_i$ das euklidische Skalarprodukt.
- Für jedes symmetrisch positiv definite C gilt

$$\lambda_{\max}(C) = \sup_{x \neq 0} \frac{\langle x, Cx \rangle}{\langle x, x \rangle}, \quad \lambda_{\min} = \inf_{x \neq 0} \frac{\langle x, Cx \rangle}{\langle x, x \rangle}.$$

Die Ausdrücke werden jeweils von den Eigenwerten zu den größten/kleinsten Eigenwerten maximiert/minimiert (Raleigh-Quotienten).

8 Relaxationsverfahren

- $S_{Jac} = I - D^{-1}A$ ist nicht symmetrisch, aber es gilt mit $D^{\frac{1}{2}} = \text{diag}(\sqrt{a_{ii}})$ ($a_{ii} > 0$):

$$\sigma(S_{Jac}) = \sigma\left(D^{\frac{1}{2}}S_{Jac}D^{-\frac{1}{2}}\right) = \sigma\left(I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}}\right)$$

- Die Vor. $2D - A$ positiv definit bedeutet

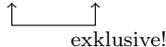
$$\langle x, (2D - A)x \rangle > 0 \iff \boxed{2\langle x, Dx \rangle > \langle x, Ax \rangle}$$

- Nun rechne für den größten Eigenwert der Iterationsmatrix

$$\begin{aligned} \lambda_{\max}\left(I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}}\right) &= \sup_{x \neq 0} \frac{\langle x, \left(I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}}\right)x \rangle}{\langle x, x \rangle} \\ &= 1 - \inf_{x \neq 0} \frac{\langle x, D^{-\frac{1}{2}}AD^{-\frac{1}{2}}x \rangle}{\langle x, x \rangle} \\ &= 1 - \inf_{x=D^{\frac{1}{2}}y \neq 0} \underbrace{\frac{\langle y, Ay \rangle}{\langle y, Dy \rangle}}_{> 0, \text{ da } A \text{ s. p. d. und } D \text{ s. p. d.}} < 1 \end{aligned}$$

- Und für den kleinsten Eigenwert der Iterationsmatrix

$$\begin{aligned} \lambda_{\min}\left(I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}}\right) &= 1 - \sup_{y \neq 0} \underbrace{\frac{\langle y, Ay \rangle}{\langle y, Dy \rangle}}_{< \sup_{y \neq 0} \frac{2\langle y, Dy \rangle}{\langle y, Dy \rangle} = 2} \\ &> 1 - 2 = -1 \end{aligned}$$

also $\sigma(S_{Jac}) \subseteq (-1, 1)$


□

Bemerkung 8.9. Man kann auch zeigen, dass die geeignet gedämpfte Jacobi-Iteration für alle symmetrisch positiv definiten Matrizen konvergiert.

8.5 Zusammenfassung

- Für die Lösung der bei Finite-Differenzen und Finite-Volumen Verfahren auftretenden linearen Gleichungssysteme eignen sich direkte Verfahren wegen des unvermeidlichen Fill-in nur bedingt.
- Besser geeignet sind iterative Verfahren, allerdings ist hier entscheidend ob und wie schnell das Verfahren konvergiert.

8.5 Zusammenfassung

- Wir haben einfache lineare Iterationsverfahren kennengelernt und die Konvergenz des Richardson- und Jacobiverfahrens für symmetrisch positiv definite Matrizen bewiesen.
- Hinsichtlich des Aufwandes zeigt sich, dass selbst diese einfachen Verfahren den direkten Verfahren bei dreidimensionalen Problemen schon überlegen sind.

8 *Relaxationsverfahren*

9 Abstiegsverfahren

Einen weiteren Zugang zu Konvergenzbeweisen für lineare Iterationsverfahren liefern diagonal-dominante Matrizen.

9.1 Diagonaldominante Matrizen

Aufgrund der Fehlergleichung

$$e^{(k+1)} = Se^{(k)}$$

gilt für beliebige verträgliche Normen

$$\|e^{(k+1)}\| \leq \|S\| \|e^{(k)}\|.$$

Kann man $\|S\| < 1$ zeigen, so konvergiert das zugehörige Iterationsverfahren (geometrische Reihe).

Für diagonaldominante Matrizen ist die Maximumnorm besonders geeignet.

Wegen $S_{Jac} = I - D^{-1}A = D^{-1}(\underbrace{-L - R}_{=:B})$ haben wir die Konvergenz des Jacobi-Verfahrens bereits gezeigt.

Wie zeigen nun die Konvergenz des Gauß-Seidel Verfahrens.

Satz 9.1. Sei A diagonaldominant oder irreduzibel diagonaldominant. Dann konvergiert das Gauß-Seidel Verfahren.

Beweis:

$$\begin{aligned} S = I - (L + D)^{-1}A &\iff (L + D)S = L + D - A = -U \\ &\iff DS = -(U + LS) \iff \boxed{S = -D^{-1}(U + LS)} \end{aligned}$$

in Komponenten gilt also für $(Sx)_i$ die Rekursionsformel:

$$(Sx)_i = (-D^{-1}(U + LS)x)_i = -\frac{1}{a_{ii}} \left(\sum_{j>i} a_{ij}x_j + \sum_{j<i} a_{ij}(Sx)_j \right)$$

also für den Betrag:

$$|(Sx)_i| \leq \frac{1}{|a_{ii}|} \left(\sum_{j>i} |a_{ij}| |x_j| + \sum_{j<i} |a_{ij}| |(Sx)_j| \right)$$

diagonaldominanter Fall: $\sum_{j \neq i} |a_{ij}| < |a_{ii}|$:

Per Induktion zeige $|(Sx)_i| < \|x\|_\infty \quad i = 1, \dots, N$.

Der Fall $i = 1$.

$$|(Sx)_i| \leq \frac{1}{|a_{ii}|} \sum_{\substack{j>i \\ j<i \text{ ist leer}}} |a_{ij}| |x_j| \leq \|x\|_\infty \underbrace{\frac{1}{|a_{ii}|} \sum_{j>i} |a_{ij}|}_{<1} < \|x\|_\infty$$

9 Abstiegsverfahren

Der Fall $i - 1 \rightarrow i$.

$$\begin{aligned} |(Sx)_i| &\leq \frac{1}{|a_{ii}|} \left(\sum_{j>i} |a_{ij}| \underbrace{|x_j|}_{\leq \|x\|_\infty} + \sum_{j<i} |a_{ij}| \underbrace{|(Sx)_j|}_{\leq \|x\|_\infty} \right) \\ &\leq \|x\|_\infty \underbrace{\frac{1}{|a_{ii}|} \sum_{j \neq i} |a_{ij}|}_{< 1} < \|x\|_\infty \end{aligned}$$

damit gilt $\|S\|_\infty = \sup_{x \neq 0} \frac{\|Sx\|_\infty}{\|x\|_\infty} < 1$.

irreduzibel diagonaldominanter Fall: $\sum_{j \neq i} |a_{ij}| \leq |a_{ii}|$

oberer Beweis zeigt dann zunächst nur $\|Sx\|_\infty \leq \|x\|_\infty$, also $\varrho(S) \leq 1$.

Wir führen nun $\varrho = 1$ zu einem Widerspruch mit den Annahmen:

Sei also $\varrho = 1$, dann gibt es einen normierten Eigenvektor $\|Sx\|_\infty = \|x\|_\infty = 1$ und ein i , sodass

$$\begin{aligned} 1 = |(Sx)_i| &\leq \frac{1}{|a_{ii}|} \left(\sum_{j>i} |a_{ij}| |x_j| + \sum_{j<i} |a_{ij}| \underbrace{|(Sx)_j|}_{= x_j, \text{ da } x \text{ Eigenvektor zu } \lambda = 1} \right) \leq \|x\|_\infty = 1 \\ \Rightarrow \sum_{j>i} |a_{ij}| |x_j| + \sum_{j<i} |a_{ij}| |x_j| &= |a_{ii}|. \end{aligned} \tag{9.1}$$

Diese Gleichung kann nur erfüllt sein, wenn

$$\begin{aligned} |x_j| &= 1 && \text{wenn } a_{ij} \neq 0 \text{ und } j > i \\ |(Sx)_j| &= 1 && \text{wenn } a_{ij} \neq 0 \text{ und } j < i \end{aligned}$$

und $\sum_{j \neq i} |a_{ij}| = |a_{ii}|$ (d. h. in der Diagonaldominanz wird gleich angenommen).

Da A irreduzibel ist kann man von i aus alle Komponenten erreichen und somit gilt Gleichheit (9.1) für alle Zeilen. Dies ist aber ein Widerspruch zu $\sum_{j=1} |a_{ij}| < |a_{ii}|$ für mindestens ein i . \square

9.2 Praktische Realisierung; Abbruchkriterium

Wann bricht man die Iteration ab?

Wegen $Ae^{(k)} = b - Ax^{(k)} = d^{(k)}$ erscheint der Defekt als geeignete Größe:

$$e^{(k)} = 0 \iff d^{(k)} = 0,$$

weil A invertierbar.

Aber $\|e^{(k)}\| \leq \|A^{-1}\| \|d^{(k)}\|$, insofern muss kleiner Defekt nicht unbedingt kleiner Fehler bedeuten wenn $\|A^{-1}\| \gg 1$.

Idee: Verwende relatives Fehlerkriterium:

Akzeptiere $x^{(k)}$, falls

$$\|d^{(k)}\| \leq \varepsilon \|d^{(0)}\|$$

mit vorgegebenem ε , z. B. gekoppelt an Diskretisierungsfehler.

Algorithmus:

```

geg.  $x, b;$            //  $x = x^{(0)}$  zu Beginn
 $d = b - Ax;$ 
 $d_0 = \|d\|;$        // z. B. euklidische Norm
 $d_k = d\Phi;$ 
while( $d_k > \varepsilon d_0$ ) {
  Löse  $Mv = d;$     // nur hier kommt das verwendete
                    // Iterationsverfahren ins Spiel
   $x = x + v;$ 
   $d = d - Av;$      // wegen  $d = b - A(x + v) = b - Ax - Av$ 
   $d_k = \|d\|;$ 
}

```

9.3 Abstiegsverfahren

Satz 9.2. Ist A symmetrisch positiv definit, dann nimmt das Funktional

$$F(x) = \frac{1}{2}x^T Ax - b^T x$$

sein eindeutiges Minimum in $x^* = A^{-1}b$, der Lösung des linearen Gleichungssystems $Ax^* = b$, an.

Beweis: Für beliebiges x setze $x = x^* + v$ und rechne

$$\begin{aligned}
 F(x) &= F(x^* + v) = \frac{1}{2}(x^* + v)^T A(x^* + v) - b^T(x^* + v) \\
 &= \frac{1}{2} \left[x^{*\top} Ax^* + \underbrace{x^{*\top} Av + v^T Ax^*}_{2v^T Ax^*} + v^T Av \right] - b^T x^* - b^T v \\
 &= \underbrace{\frac{1}{2} x^{*\top} Ax^* - b^T x^*}_{F(x^*)} + v^T \underbrace{(Ax^* - b)}_{=0} + \frac{1}{2} v^T Av \\
 &= F(x^*) + \frac{1}{2} \underbrace{v^T Av}_{> 0 \text{ für } v \neq 0} > F(x^*)
 \end{aligned}$$

Eindeutigkeit: Sei F in x^* und $x' \neq x^*$ minimal, so gilt für $x' = x^* + v$, $v \neq 0$: $F(x') = F(x^*) + \frac{1}{2}v^T Av > F(x^*)$, also ζ zu F minimal in x' . \square

9 Abstiegsverfahren

Gegeben ein $x^{(k)}$, eine Näherungslösung von x^* , so kann man $x^{(k)}$ in Richtung einer ebenfalls gegebenen *Suchrichtung* $p^{(k)} \in \mathbb{R}^N$ verbessern, indem man

$$F\left(x^{(k)} + \alpha p^{(k)}\right) \rightarrow \min$$

löst.

Das optimale α lässt sich wie folgt berechnen. Zunächst reche aus:

$$F\left(x^{(k)} + \alpha p^{(k)}\right) = F\left(x^{(k)}\right) + \alpha \left(p^{(k)}\right)^T \left(Ax^{(k)} - b\right) + \frac{\alpha^2}{2} \left(p^{(k)}\right)^T Ap^{(k)}$$

(setze $x^* = x^{(k)}$ und $v = \alpha p^{(k)}$ in obiger Rechnung).

Eine notwendige Bedingung für ein Minimum ist das Verschwinden der Ableitung

$$\frac{d}{dx} F\left(x^{(k)} + \alpha p^{(k)}\right) = \left(p^{(k)}\right)^T \left(Ax^{(k)} - b\right) + \alpha p^{(k)T} Ap^{(k)} \stackrel{!}{=} 0$$

$$\Leftrightarrow \alpha^{(k)} = \frac{\left(p^{(k)}\right)^T \left(b - Ax^{(k)}\right)}{\underbrace{\left(p^{(k)}\right)^T Ap^{(k)}}_{\substack{\text{Defekt!} \\ \downarrow}}}$$

$\neq 0$ wg. pos. definit und $p \neq 0$

Die zweite Ableitung ist gerade

$$\frac{d^2}{dx^2} F\left(x^{(k)} + \alpha p^{(k)}\right) = p^{(k)T} Ap^{(k)}$$

also positiv und es liegt tatsächlich ein Minimum vor.

Wähle nun die spezielle Suchrichtung $p^{(k)} = -\nabla F\left(x^{(k)}\right)$ (negative Gradientenrichtung). Ausgeschrieben gilt

$$F(x) = \frac{1}{2} \underbrace{\sum_{i=1}^N \sum_{j=1}^N x_i a_{ij} x_j}_{\text{}} - \sum_{i=1}^N b_i x_i$$

und damit

$$\frac{\partial F}{\partial x_m} = \frac{1}{2} \left\{ 2a_{mm}x_m + \sum_{j \neq m} a_{mj}x_j + \sum_{i \neq m} x_i a_{im} \right\} = (Ax - b)_m.$$

also gilt $p^{(k)} = -\nabla F\left(x^{(k)}\right) = b - Ax^{(k)}$ und es ergibt sich der folgende Algorithmus, das sogenannte „Gradientenverfahren“:

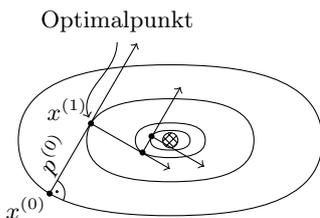
```

geg.  $x, b$ ;
 $d = b - Ax$ ;
 $d_0 = \|d\|$ ;
 $d_k = d_0$ ;
while( $d_k > \varepsilon d_0$ ) {
     $q = Ad$ ;
     $\alpha = \frac{d^T d}{d^T q}$ ;
     $x = x + \alpha d$ ;
     $d = d - \alpha q$ ;
}
    
```

Problem:

$$A = \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow F(x) = x_1^2 + \varepsilon x_2^2$$

Höhenlinien von F sind Ellipsen um den Ursprung



- „Zickzackkurve = langsame Konvergenz“
- Man zeigt:

$$\|e^{(k)}\|_A \leq \frac{\kappa(A) + 1}{\kappa(A) - 1} \|e^{(k-1)}\|_A.$$

- „Energienorm“ $\|x\|_A = \sqrt{x^T A x}$.
- „Energieskalarprodukt“

$$\langle x, y \rangle_A = x^T A y.$$

Wir lernen zwei Möglichkeiten zur Verbesserung des Gradientenverfahrens kennen, die auch kombiniert werden können.

9.4 Vorkonditioniertes Gradientenverfahren

Idee: Wende Gradientenverfahren auf das transformierte System

$$M^{-1}Ax = M^{-1}b \tag{9.2}$$

an.

Problem: $M^{-1}A$ ist im allgemeinen nicht symmetrisch, selbst wenn M^{-1} und A symmetrisch sind.

Ist M symmetrisch positiv definit, so gibt es aber ein T mit $M = TDT^{-1}$ und $D = \text{diag}(d_{ii})$, $d_{ii} > 0$. Man definiert dann formal

$$M^{\frac{1}{2}} = TD^{\frac{1}{2}}T^{-1} \quad \text{mit} \quad (D^{\frac{1}{2}})_{ii} = \sqrt{d_{ii}}$$

denn damit gilt $M^{\frac{1}{2}}M^{\frac{1}{2}} = M$.

9 Abstiegsverfahren

Damit ist $M^{-1}A$ ähnlich zu $M^{\frac{1}{2}}M^{-1}AM^{-\frac{1}{2}} = M^{-\frac{1}{2}}AM^{-\frac{1}{2}}$, also

$$\sigma(M^{-1}A) = \sigma(M^{-\frac{1}{2}}AM^{-\frac{1}{2}}).$$

Nun multipliziere (9.2) von links mit $M^{\frac{1}{2}}$ und setze $\hat{x} = M^{\frac{1}{2}}x$

$$\underbrace{M^{-\frac{1}{2}}AM^{-\frac{1}{2}}}_{\hat{A}} \underbrace{M^{\frac{1}{2}}x}_{\hat{x}} = \underbrace{M^{-\frac{1}{2}}b}_{\hat{b}}$$

$$\iff \hat{A}\hat{x} = \hat{b}.$$

Das transformierte System $\hat{A}\hat{x} = \hat{b}$ ist symmetrisch positiv definit und hat die Eigenwerte von $M^{-1}A$.

Das Gradientenverfahren ist formal anwendbar:

$$\begin{aligned} &\text{geg. } \hat{x}, \hat{b}; \\ &\hat{d} = \hat{b} - \hat{A}\hat{x}; \\ &\text{while}(\dots) \{ \\ &\quad \hat{q} = \hat{A}\hat{d}; \\ &\quad \hat{\alpha} = \frac{\hat{d}^T \hat{d}}{\hat{d}^T \hat{q}}; \\ &\quad \hat{x} = \hat{x} + \hat{\alpha}\hat{d}; \\ &\quad \hat{d} = \hat{d} - \hat{\alpha}\hat{q}; \\ &\} \end{aligned}$$

Allerdings möchte man das Verfahren in dieser Weise nicht praktisch durchführen, da \hat{A} im allgemeinen nicht mehr dünn besetzt ist.

Die Idee ist nun die Transformation *in jedem einzelnen Schritt* des Gradientenverfahrens zu berücksichtigen aber nur die untransformierten Größen zu speichern.

Gegeben seien also x, b

$$\begin{aligned} \text{Beachte: } x &= M^{-\frac{1}{2}}\hat{x} \quad \text{und} \quad \hat{b} = M^{-\frac{1}{2}}b \quad \text{sowie} \quad \hat{A} = M^{-\frac{1}{2}}AM^{-\frac{1}{2}} \\ \hat{x} &= M^{\frac{1}{2}}x \quad \quad \quad b = M^{\frac{1}{2}}\hat{b} \\ \hat{d} &= \hat{b} - \hat{A}\hat{x} = M^{-\frac{1}{2}}b - \left(M^{-\frac{1}{2}}AM^{-\frac{1}{2}}\right) \left(M^{\frac{1}{2}}x\right) = M^{-\frac{1}{2}} \underbrace{(b - Ax)}_d \end{aligned}$$

berechne $d = b - Ax$;

while(...) {

$$\hat{q} = \hat{A}\hat{d} = \left(M^{-\frac{1}{2}}AM^{-\frac{1}{2}}\right) \left(M^{-\frac{1}{2}}d\right) = M^{-\frac{1}{2}} \underbrace{A \underbrace{M^{-1}d}_v}_q;$$

berechne nur $q = Av$; $v = M^{-1}d$;

$$\hat{\alpha} = \frac{\hat{d}^T \hat{d}}{\hat{d}^T \hat{q}} = \frac{\left(M^{-\frac{1}{2}}d\right)^T \left(M^{-\frac{1}{2}}d\right)}{\left(M^{-\frac{1}{2}}d\right)^T \left(M^{-\frac{1}{2}}q\right)} = \frac{d^T (M^{-1}d)}{d^T (M^{-1}q)} = \frac{d^T v}{q^T v}$$

α und $\hat{\alpha}$
sind identisch.

$$\hat{x} = \hat{x} + \hat{\alpha}\hat{d} = M^{\frac{1}{2}}x + \hat{\alpha} \cdot M^{-\frac{1}{2}}d = M^{\frac{1}{2}} \left(x + \hat{\alpha} \underbrace{M^{-1}d}_v\right)$$

$$= M^{\frac{1}{2}} \underbrace{(x + \hat{\alpha}v)}$$

speichere nur das x und seine updates.

\hat{x} taucht nirgendwo anders auf!

$$\hat{d} = \hat{d} - \hat{\alpha}\hat{q} = M^{-\frac{1}{2}}d - \hat{\alpha}M^{-\frac{1}{2}}q = M^{-\frac{1}{2}} \underbrace{(d - \hat{\alpha}q)}$$

speichere nur d und seine updates.

}

Das sogenannte vorkonditionierte Gradientenverfahren lässt sich damit folgendermaßen algorithmisch realisieren:

```

geg.  $x, b$ ;
 $d = b - Ax$ ;
 $d_0 = \|d\|$ ;
 $d_k = d_0$ ;
while( $d_k > \varepsilon d_0$ ) {
    Löse  $v = M^{-1}d$ ; //  $M$  ist sym. pos. definit
     $q = Av$ ;
     $\alpha = \frac{d^T v}{q^T v}$ ;
     $x = x + \alpha v$ ;
     $d = d - \alpha q$ ;
     $d_k = \|d\|$ ;
}

```

9.5 Konjugierte Gradienten Verfahren

Die Vorkonditionierung behebt nicht das Problem der langsamen Konvergenz des Gradientenverfahrens, d.h. das Verfahren ist nicht schneller als die Basisiteration $\varrho(I - M^{-1}A)$.

Die liegt daran, dass das Gradientenverfahren die Optimalität bezüglich einer Suchrichtung wieder verliert.

9 Abstiegsverfahren

Sei $p^{(k)}$ eine Suchrichtung. Die Iterierte $x^{(k)}$ ist in Richtung $p^{(k)}$ bereits optimal, falls

$$\alpha = \frac{(d^{(k)})^T p^{(k)}}{(p^{(k)})^T Ap^{(k)}} = 0, \quad \text{was nur für } (d^{(k)})^T p^{(k)} = 0 \quad \text{möglich ist.}$$

Nach einem Schritt des Gradientenverfahrens gilt zwar

$$\underbrace{(d^{(k)})^T}_{\substack{\text{aktueller} \\ \text{Defekt}}} \underbrace{d^{(k-1)}}_{\substack{\text{letzte} \\ \text{Suchrichtung}}} = 0$$

(nachrechnen!), aber leider im allgemeinen bereits

$$\underbrace{(d^{(k)})^T}_{\substack{\text{aktueller} \\ \text{Defekt}}} \underbrace{d^{(k-2)}}_{\substack{\text{vorletzte} \\ \text{Suchrichtung}}} \neq 0,$$

d. h. die Optimalität bezüglich aller Suchrichtungen geht verloren.

Seien $p^{(k)}$ die im Laufe eines Abstiegsverfahrens verwendeten Richtungen (nicht notwendigerweise die Gradientenrichtung).

Minimierung in Richtung $p^{(k)}$ im Schritt k liefert den neuen Defekt

$$d^{(k+1)} = d^{(k)} - \alpha^{(k)} Ap^{(k)}.$$

Damit gilt dann natürlich

$$\begin{aligned} (d^{(k+1)})^T p^{(k)} &= (d^{(k)} - \alpha^{(k)} Ap^{(k)})^T p^{(k)} \\ &= (d^{(k)})^T p^{(k)} - \frac{(d^{(k)})^T p^{(k)}}{(p^{(k)})^T Ap^{(k)}} (p^{(k)})^T Ap^{(k)} = 0 \end{aligned}$$

Damit $x^{(k+1)}$ *zusätzlich* noch optimal bezüglich aller alten Richtungen $p^{(0)}, \dots, p^{(k-1)}$ ist, muss gelten

$$(d^{(k+1)})^T p^{(l)} = 0 \quad \forall 0 \leq l < k$$

also

$$\begin{aligned} (d^{(k)} - \alpha^{(k)} Ap^{(k)})^T p^{(l)} &= \underbrace{(d^{(k)})^T p^{(l)}}_{=0} - \alpha^{(k)} \underbrace{(Ap^{(k)})^T p^{(l)}}_{=0} = 0 \\ &\quad \text{per Induktion im Schritt } \quad \quad \quad \text{im Schritt } (k) \\ &\quad (k-1) \text{ für } d^{(k)} \text{ sichergestellt} \quad \quad \quad \text{sicherzustellen} \end{aligned}$$

Wählt man also die $p^{(k)}$ so, dass $(p^{(k)})^T Ap^{(l)} \neq 0 \quad \forall 0 \leq l < k$, so bleibt die Optimalität bezüglich aller alten Richtungen erhalten.

Dies nennt man „Verfahren der konjugierten Richtungen“.

Die Verbindung mit Gradientenverfahren liefert das „Verfahren der konjugierten Gradienten“. Für dieses Verfahren zeigt man die Konvergenzrate

$$\|e^{(k)}\|_A \leq \frac{\sqrt{\kappa(A)} + 1}{\sqrt{\kappa(A)} - 1} \|e^{(k-1)}\|_A.$$

Auch hier kann wieder die Technik der Vorkonditionierung eingesetzt werden.

Zum Schluss noch eine Tabelle mit ein paar Zahlen.

$\Omega = (0, 1)^2$, $-\Delta u = 0$, $u = g$ auf $\partial\Omega$, Fünfpunktstern, Abbruch: $\|T^m\| \leq 10^{-4} \|T^0\|$

	$h = \frac{1}{32}$	$h = \frac{1}{64}$	$h = \frac{1}{128}$
Jacobi	1203	3967 $T=1.5 s$	12444 $T=36.2 s$
SSOR(1) (sym.GS)	303 $\varrho_{GS} = \varrho_{Jac}^2$ SGS = 2 · GS	994 $T=0.5 s$	3113 $T=10.2 s$
ILU(0)	178 selber Aufwand wie SGS	583 $T=0.3 s$	1824 $T=5.79 s$
Grad + Jac	1268	4240	13579 $T=46.255 s$
CG + Jacobi	60	116 $T=0.05 s$	224 $T=0.8 s$
CG + SSOR(1)	22	42	79 $T=0.31 s$
CG + ILU(0)	19	36 $T=0.023 s$	68 $T=0.26 s$

ILU: „Incomplete Lower Upper“-Zerlegung.

9.6 Zusammenfassung

- In diesem Abschnitt haben wir Abstiegsverfahren eingeführt.
- Details zum Verfahren der konjugierten Gradienten findet man in der Grundlagenvorlesung.
- Abstiegsverfahren können durch Vorkonditionierung verbessert werden

9 Abstiegsverfahren

10 Mehrgitterverfahren

Die Konvergenzrate aller bisher behandelten Iterationsverfahren hängt von h , der Gitterweite, ab.

Nun führen wir ein Verfahren ein welches eine h -unabhängige Konvergenzrate besitzt. Dies bedeutet, dass bei vorgegebener Genauigkeit eine feste Anzahl von Iterationen unabhängig von der Zahl der Unbekannten benötigt wird.

10.1 Glättungseigenschaft

Die gedämpfte Richardson Iteration besitzt die Iterationsmatrix

$$S = I - \omega A.$$

Für eine symmetrisch positiv definite Matrix A mit Eigenvektoren $\lambda_1 \leq \dots \leq \lambda_N$ lauten die Eigenwerte der Iterationsmatrix

$$\mu_i = 1 - \omega \lambda_i$$

bzw. bei $\omega = \frac{1}{\lambda_N}$

$$\mu_i = 1 - \frac{\lambda_i}{\lambda_N}.$$

Zusätzlich gilt: Jeder Eigenvektor e_i von A ist auch Eigenvektor von S .

Für den Fünfpunktstern sind $(\lambda^{\nu\mu}, e^{\nu\mu})$ explizit bekannt (Abschnitt 6.5).

$$e_{ij}^{\nu\mu} = \sin(\nu\pi ih) (\mu\pi jh)$$

$$\lambda^{\nu\mu} = \frac{4}{h^2} \left(\sin^2 \left(\frac{\nu\pi h}{2} \right) + \sin^2 \left(\frac{\mu\pi h}{2} \right) \right) \quad 1 \leq \nu\mu \leq n$$

$$\mu^{\nu\mu} = 1 - \frac{1}{8} \lambda^{\nu\mu} = 1 - \frac{1}{2} \left(\underbrace{\sin^2 \left(\frac{\nu\pi}{n} \right) + \sin^2 \left(\frac{\mu\pi}{n} \right)}_{\substack{\leq \frac{1}{2} \quad \text{für } \nu \leq \frac{n}{2} \\ > \frac{1}{2} \quad \text{für } \nu > \frac{n}{2}}} \right),$$

wobei $n = \sqrt{N}$.

Wir sehen:

- Niederfrequente Fehler werden schlecht gedämpft:

$$\mu^{\nu\mu} > \frac{1}{2} \quad \text{für } 1 \leq \nu, \mu < \frac{n}{2} \quad \begin{array}{l} \text{worst case} \\ \xi^{1,1} \rightarrow 1 \text{ für } h \rightarrow 0 \end{array}$$

10 Mehrgitterverfahren

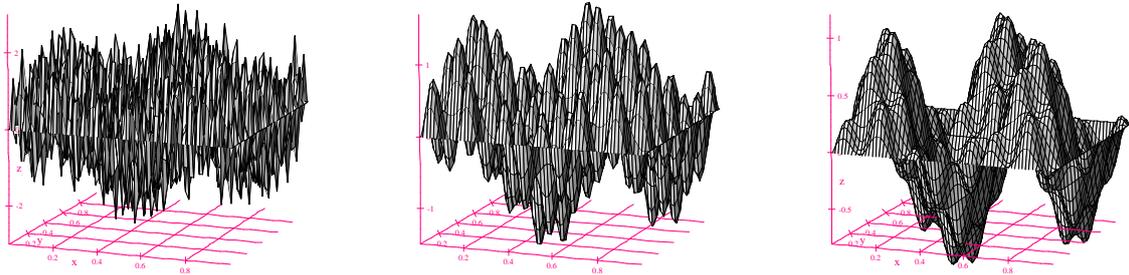


Abbildung 4: Beispiel für die Fehlerentwicklung. Anfangsfehler links, nach einer Iteration mitte und nach fünf Iterationen rechts.

- Hochfrequente Fehler werden durch die (gedämpfte) Richardson-Iteration gut reduziert:

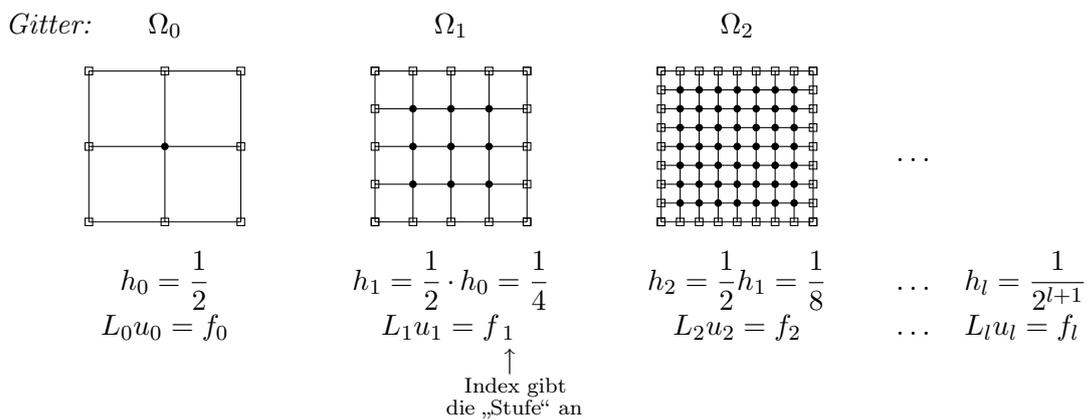
$$\mu^{\nu\mu} \leq \frac{3}{4} \quad \text{für} \quad \frac{n}{2} \leq \nu, \mu < n \quad \text{best case} \quad \xi^{n-1, n-1} = O(h^2)$$

Abbildung 10.1 zeigt anschaulich die Reduktion der hochfrequenten Fehler nach einer und fünf Iterationen mit der Richardson-Iteration.

Mehrgitterverfahren nutzen die Diskretisierung des kontinuierlichen Problems

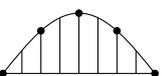
$$\begin{aligned} -\Delta u &= f && \text{in } \Omega = (0, 1)^2 \\ u &= g && \text{auf } \partial\Omega \end{aligned}$$

auf unterschiedlich feinen Gittern:



Für jede Stufe $l = 0, 1, \dots$ erhalten wir Gitterfunktionen aus der Menge $U_l = \{f \mid f: \Omega_l \rightarrow \mathbb{R}\}$.

10.2 Prolongation



Idee: Niederfrequente Fehler (also $1 \leq \nu, \mu < \frac{n}{2}$) können auf einem gröberen Gitter sehr gut

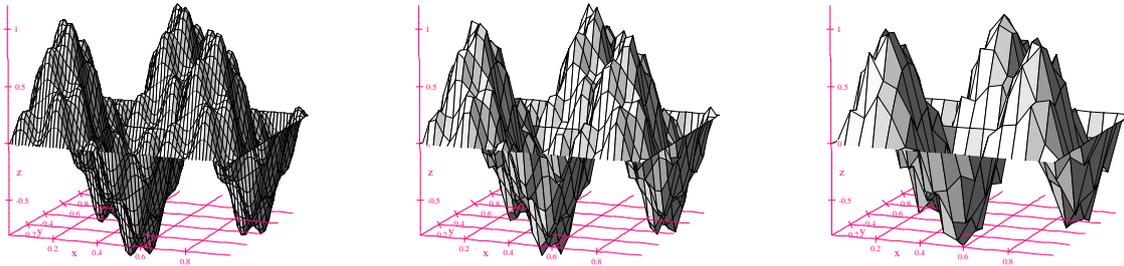


Abbildung 5: Darstellung einer niederfrequenten Gitterfunktion auf größeren Gittern.

dargestellt werden. Abbildung 10.2 zeigt dies anschaulich für den Fehler nach fünf Iterationen aus Abbildung 10.1.

Formal können wir das folgendermaßen schreiben:

$$\begin{array}{ccccccc}
 & & e_l & = & P_l & e_{l-1} & + & \text{Fehler} \\
 & \nearrow & & & \nearrow & & & \nearrow \\
 \text{Fehler} & & & & \text{Prolongations-} & & & \text{klein, falls} \\
 \text{auf Stufe } l & & & & \text{matrix} & & \text{Darstellung des} & e_l \text{ niederfrequent} \\
 \text{niederfrequent} & & & & & & \text{Fehlers auf} & \\
 & & & & & & \text{Stufe } l-1 &
 \end{array}$$

Dabei ist

$$P_l: U_{l-1} \rightarrow U_l$$

eine Rechteckmatrix, genannt „Prolongationsmatrix“.

$$\begin{array}{c} \boxed{} \end{array} = \begin{array}{c} \boxed{} \end{array} \begin{array}{c} \boxed{} \end{array}$$

Für die Prolongation gibt es – je nach Diskretisierungsverfahren – verschiedene Möglichkeiten.

Eine Möglichkeit definiert die Prolongation durch bilineare Interpolation:

$$(P_l e_{l-1})(x_{ij}) \begin{cases} e_{l-1}(x_{\frac{i}{2}, \frac{j}{2}}) & i, j \text{ gerade} \\ \frac{1}{2}(e_l(x_{\frac{i-1}{2}, \frac{j}{2}}) + e_l(x_{\frac{i+1}{2}, \frac{j}{2}})) & i \text{ ungerade} \\ & j \text{ gerade} \\ \frac{1}{2}(e_l(x_{\frac{i}{2}, \frac{j-1}{2}}) + e_l(x_{\frac{i}{2}, \frac{j+1}{2}})) & i \text{ gerade} \\ & j \text{ ungerade} \\ \frac{1}{4}(e_l(x_{\frac{i-1}{2}, \frac{j-1}{2}}) + e_l(x_{\frac{i+1}{2}, \frac{j-1}{2}}) \\ + e_l(x_{\frac{i-1}{2}, \frac{j+1}{2}}) + e_l(x_{\frac{i+1}{2}, \frac{j+1}{2}})) & i, j \text{ ungerade} \end{cases}$$

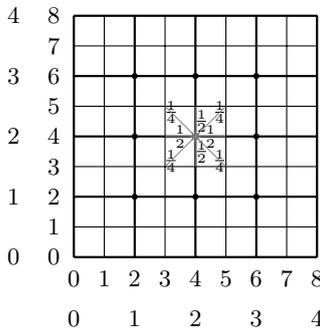
Bemerkung 10.1. Am Dirichlet Rand sind entsprechende Auswertungen von $e_l(x \in \partial\Omega)$ zu ignorieren da am Dirichletrand der Fehler Null ist. Dies entspricht dem Nullsetzen der Werte. Die Punkte sind ja gar nicht in Ω_l enthalten (siehe Definition von U_l oben). square

10.3 Restriktion

Die Restriktion ist eine Abbildung, die Feingitterfunktionen in Grobgitterfunktionen überführt.

Man setzt

$$R_l: U_l \rightarrow U_{l-1} \quad \text{mit } R_l = \frac{1}{4}P_l^T$$



$$(R_l d_l)(x_{ij}) = \frac{1}{4} \left[d_l(x_{2i,2j}) + \frac{1}{2} (d_l(x_{2i-1,2j}) + d_l(x_{2i+1,2j}) + d_l(x_{2i,2j-1}) + d_l(x_{2i,2j+1})) + \frac{1}{4} (d_l(x_{2i-1,2j-1}) + d_l(x_{2i+1,2j-1}) + d_l(x_{2i-1,2j+1}) + d_l(x_{2i+1,2j+1})) \right]$$

Die hiermit definierte Restriktion ist für die Aufstellung der Matrizen L_l jeder Stufe mit dem Finite Differenzen Verfahren aus (4.8) geeignet.

10.4 Grobgitterkorrektur

Lineare Iterationsverfahren basieren auf einer genäherten Lösung der Fehlergleichung

$$L e^{(k)} = d^{(k)} \rightsquigarrow M v^{(k)} = d^{(k)}.$$

Nun entsteht der Defekt auf der Stufe J eines J mal verfeinerten Gitters: $d_J^{(k)} = f_J^{(k)} - L_J u_J^{(k)}$ und es ist eine Korrektur $v_J^{(k)}$ gesucht.

Berechne diese in folgender Weise:

- (i) $d_J^{(k)} = f_J^{(k)} - L_J u_J^{(k)}$ Feingitterdefekt
- (ii) $d_{J-1}^{(k)} = R_J d_J^{(k)}$ Restriktion des Defektes
- (iii) Löse $L_{J-1} v_{J-1}^{(k)} = d_{J-1}^{(k)}$ Löse Grobgitterproblem
- (iv) $v_J^{(k)} = P_J v_{J-1}^{(k)}$ Prolongation der Korrektur.

Dies ergibt die sogenannte Grobgitterkorrektur

$$u_J^{(k+1)} = u_J^{(k)} + \underbrace{P_J L_{J-1}^{-1} R_J}_{G_J} (f_J - L_J u_J^{(k)}) \tag{10.1}$$

10 Mehrgitterverfahren

```

multigrid( $l, u_l, f_l$ )
{
  if( $l == 0$ ) {
    Löse  $L_0 u_0 = f_0$ ;
  } else {
    „Vorglättung“ for ( $i = 1; i \leq \nu_1; i = i + 1$ )  $u_l = u_l + M_l^{-1} (f_l - L_l u_l)$ ;
                   $d_{l-1} = R_l (f_l - L_l u_l)$ ;
                   $v_{l-1} = 0$ ; // Startwert für Korrektur
                  for( $i = 1; i \leq \gamma; i = i + 1$ )
                    multigrid( $l - 1, v_{l-1} d_{l-1}$ );
                   $u_l = u_l + P_l v_l$ ;
    „Nachglättung“ for( $i = 1; i \leq \nu_2; i = i + 1$ )  $u_l = u_l + M_l^{-1} (f_l - L_l u_l)$ ;
  }
}

```

Bei $\gamma = 1$ spricht man vom V-Zyklus, bei $\gamma = 2$ vom W-Zyklus (größere Werte von γ sind nicht in Gebrauch).

Auch die Konvergenz des Mehrgitterverfahrens ist unabhängig von h , und das sogar schon für den V-Zyklus mit einem Glättungsschritt, also $\gamma = 1$, $\nu_1 = 1$ und $\nu_2 = 0$.

Der Aufwand für einen V-Zyklus in zwei Raumdimensionen beträgt

$$\begin{aligned}
 A(N) &\leq \underbrace{CN}_{\text{feinste Stufe}} + C \frac{N}{4} + C \frac{N}{16} + C \frac{N}{64} \dots \\
 &\qquad\qquad\qquad \swarrow \qquad \nwarrow \\
 &\qquad\qquad\qquad \text{Vergrößerungs-} \\
 &\qquad\qquad\qquad \text{faktor in } 2d
 \end{aligned}$$

$$= CN \underbrace{\left(1 + \frac{1}{4} + \frac{1}{16} + \dots \right)}_{\text{Reihe konvergiert}} \leq \frac{4}{3} CN$$

Der Aufwand ist somit unwesentlich größer als für ein Eingitterverfahren (d. h. für den Glätter auf der feinsten Stufe).

Entsprechend lassen sich auch Komplexitätsabschätzungen für allgemeines γ und Raumdimension d geben.

Mehrgitterverfahren gehören zu den schnellsten bekannten Verfahren zur Lösung diskretisierter Poissongleichung. Es kann auf unstrukturierte Gitter und kompliziertere Probleme (z. B. variable Diffusionskoeffizienten) verallgemeinert werden. Die Elliptizität der Gleichung ist allerdings eine wesentliche Voraussetzung für die Glättungseigenschaft.

Die hier beschriebene Variante nennt man *geometrisches* Mehrgitterverfahren. Hier setzt man die Existenz einer Gitterhierarchie voraus auf denen man die Gleichung diskretisieren kann. In der Praxis ist das oft ein Problem (z. B. bei komplizierten Geometrien), dann kann man ein *algebraisches* Mehrgitterverfahren einsetzen bei dem die Hierarchie von Gleichungssystemen direkt aus einem gegebenen Gleichungssystem erzeugt wird.

10.7 Zusammenfassung

- Mehrgitterverfahren kombinieren ein Relaxationsverfahren mit einer Grobgitterkorrektur.
- Voraussetzung hierfür ist eine Reduktion hochfrequenter Fehlerkomponenten durch das Relaxationsverfahren.
- Mehrgitterverfahren besitzen optimale Komplexität $O(N)$ für die Lösung der diskretisierten Poissongleichung.

10 Mehrgitterverfahren

11 Parabolische partielle Differentialgleichungen

11.1 Lösung mittels Fourierreihe

Wir betrachten das Problem: Finde $u(x, t)$ so, dass

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0 \quad \text{in } (0, 1) \times (0, \infty) \quad (11.1a)$$

$$u(x, 0) = f(x) \quad \text{für } t = 0 \quad (\text{Anfangsbedingung}), \quad (11.1b)$$

$$\left. \begin{array}{l} u(0, t) = 0 \\ u(1, t) = 0 \end{array} \right\} \quad (\text{Randbedingung}). \quad (11.1c)$$

Eine Lösungsmöglichkeit ist mittels Separation der Variablen. Setze an

$$u(x, t) = X(x) \cdot T(t).$$

Damit ist

$$\frac{\partial u}{\partial t} = XT' \quad \text{und} \quad \frac{\partial^2 u}{\partial x^2} = X''T.$$

Einsetzen in die PDE liefert

$$XT' - X''T = 0 \iff \frac{T'(t)}{T(t)} = \frac{X''(x)}{X(x)}, \quad (11.2)$$

vorausgesetzt, dass $u(x, t) = X(x) \cdot T(t) \neq 0$.

Linke Seite in (11.2) ist unabhängig von x , rechte Seite ist unabhängig von t . Damit beide Seiten gleich sind für *alle* x, t , kommt nur

$$\frac{T'(t)}{T(t)} = \frac{X''(x)}{X(x)} = \lambda \quad (\text{const})$$

in Frage. Daraus erhält man dann

$$\begin{array}{ll} T'(t) = \lambda T(t) & \Rightarrow T(t) = c_1 e^{\lambda t} \\ X''(x) = \lambda X(x) & \Rightarrow X(x) = c_2 e^{\sqrt{\lambda}x} + c_3 e^{-\sqrt{\lambda}x} \end{array} \quad \text{das gleiche } \lambda!$$

Also

$$u(x, t) = e^{\lambda t} \left(A e^{\sqrt{\lambda}x} + B e^{-\sqrt{\lambda}x} \right).$$

Um die Randbedingungen (11.1c) zu erfüllen, setzen wir

$$\left. \begin{array}{l} A = a + ib \\ B = a - ib \end{array} \right\} \Rightarrow A + B = 2a \stackrel{!}{=} 0 \quad A - B = i2b$$

und $\lambda = -n^2\pi^2$, $n \in \mathbb{N}$ ($\rightsquigarrow \sin n\pi x$).

11 Parabolische partielle Differentialgleichungen

Dies gibt

$$\begin{aligned}
 Ae^{\sqrt{\lambda}x} + Be^{-\sqrt{\lambda}x} &= Ae^{\underbrace{\sqrt{-n^2\pi^2}}_{in\pi}x} + Be^{-in\pi x} \\
 &= A(\cos n\pi x + i \sin n\pi x) + B\left(\underbrace{\cos(-n\pi x)}_{=\cos n\pi x} + i \underbrace{\sin(-n\pi x)}_{=-\sin n\pi x}\right) \\
 &= (A + B) \cos n\pi x + i(A - B) \sin n\pi x \\
 &= \underbrace{2a \cos n\pi x}_0 - \underbrace{2b \sin n\pi x} \\
 &\quad \text{erfüllt die RB}
 \end{aligned}$$

Lösungen, die die Randbedingung erfüllen, haben also die Gestalt (mit einem neuen A):

$$u(x, t) = Ae^{-n^2\pi^2 t} \sin n\pi x.$$

Um die Anfangsbedingungen (11.1b) zu erfüllen, entwickeln wir f in eine Fourierreihe

$$f(x) = \sum_{n=1}^{\infty} A_n \sin n\pi x$$

und damit erfüllt

$$u(x, t) = \sum_{n=1}^{\infty} A_n e^{-n^2\pi^2 t} \sin n\pi x$$

Die Gleichung (11.1a).

Bemerkung 11.1. Das so definierte $u(x, t)$ ist keine klassische Lösung in $C^2(\Omega \times (0, \infty))$. Für jedes t ist $u(\cdot, t)$ eine Funktion in $L^2(\Omega)$, da es Grenzwert einer Fourierreihe ist.

Bemerkung 11.2. „Hochfrequente“ Anteile (großes n) in der Anfangsbedingung werden sehr viel schneller gedämpft als niederfrequente wegen des n^2 -Terms in der e -Funktion. Dies bezeichnet man als „Glättungseigenschaft“ parabolischer Probleme.

11.2 Finite Differenzen für Parabolische Probleme

Wir beschränken uns auf die räumlich eindimensionale Aufgabe

$$\begin{aligned}
 \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} &= f & \text{in } \Omega \times T, & \quad \Omega = (0, 1), T = (0, T_{end}) \\
 u &= g & \text{auf } \partial\Omega & \\
 u &= u_0 & \text{für } t = 0. &
 \end{aligned} \tag{11.3}$$

Diskretisierung erfolgt mittels der sog. Linienmethode, d. h. man diskretisiert erst im Ort und dann in der Zeit.

11.2 Finite Differenzen für Parabolische Probleme

Ortsdiskretisierung: Finite Differenzen mit Gitter $x_i = i \cdot h$, $h = \frac{1}{N}$, $i = 0, \dots, N$.

Taylorentwicklung für $\frac{\partial^2 u}{\partial x^2}$ am Punkt (x_i, t) liefert:

$$\frac{\partial u(x_i, t)}{\partial t} = \frac{1}{h^2} \underbrace{[u(x_{i-1}, t) - 2u(x_i, t) + u(x_{i+1}, t)]}_{F(x_i, t)} + f(x_i, t) + O(h^2) \quad i = 1, \dots, N-1 \quad (11.4)$$

Dies ist ein gekoppeltes System gewöhnlicher Differentialgleichungen für die $N-1$ unbekannt Funktionen „ $u_i(t) = u(x_i, t)$ “.

Für die Zeitdiskretisierung wählen wir nun das Gitter $t^k = k \cdot \tau$, $\tau = \frac{T_{end}}{K}$, $k = 0, \dots, K$.

Einschritt- θ -Verfahren: Numerische Integration liefert:

$$\begin{aligned} \frac{\partial u(x_i, t)}{\partial t} &= F(x_i, t) \quad i = 1, \dots, N-1 \\ \Rightarrow \int_{t^k}^{t^{k+1}} \frac{\partial u(x_i, t)}{\partial t} dt &= \int_{t^k}^{t^{k+1}} F(x_i, t) dt \\ \Leftrightarrow u(x_i, t^{k+1}) - u(x_i, t^k) &= \tau \left[(1-\theta)F(x_i, t^k) + \theta \cdot H(x_i, t^{k+1}) \right] + O(\tau^p) \\ \text{mit } p &= \begin{cases} 3 & \theta = \frac{1}{2} \quad \text{„Trapezregel“} \\ 2 & 0 \leq \theta \leq 1, \theta \neq \frac{1}{2} \end{cases} \end{aligned}$$

Etwas umstellen (setze F ein, bringe alle $u(\cdot, t^{k+1})$ nach links) liefert dann

$$\begin{aligned} &\frac{\tau}{h^2} \\ &\parallel \\ &-\theta \gamma u(x_{i-1}, t^{k+1}) + (1 + 2\theta \gamma)u(x_i, t^{k+1}) - \theta \gamma u(x_{i+1}, t^{k+1}) = \\ &= (1-\theta)\gamma u(x_{i-1}, t^k) + (1 - 2(1-\theta)\gamma)u(x_i, t^k) + (1-\theta)\gamma u(x_{i+1}, t^k) \quad i = 1, \dots, N-1 \quad (11.5) \\ &\quad + \tau \left[(1-\theta)f(x_i, t^k) + \theta f(x_i, t^{k+1}) \right] + O(\tau h^2 + \tau^p). \\ &\quad \quad \quad \uparrow \\ &\quad \quad \quad \text{wegen } \tau \cdot F! \end{aligned}$$

mit der Abkürzung $\frac{\tau}{h^2} = \gamma$.

Für jeden diskreten Zeitpunkt t^k betrachten wir die Gitterfunktion $u_h^k: \bar{\Omega}_h \rightarrow \mathbb{R}$.

Die Gleichung für die Gitterfunktion erhält man durch Weglassen des Fehlerterms in (11.5) und Einsetzen der Rand- und Anfangsbedingungen:

$$\begin{aligned} &-\theta \gamma u_h^{k+1}(x_{i-1}) + (1 + 2\theta \gamma)u_h^{k+1}(x_i) - \theta \gamma u_h^{k+1}(x_{i+1}) \\ &= (1-\theta)\gamma u_h^k(x_{i-1}) + (1 - 2(1-\theta)\gamma)u_h^k(x_i) + (1-\theta)\gamma u_h^k(x_{i+1}) \\ &\quad + \tau \left[(1-\theta)f(x_i, t^k) + \theta f(x_i, t^{k+1}) \right] \quad i = 1, \dots, N-1, \quad k \geq 0 \quad (11.6a) \end{aligned}$$

11 Parabolische partielle Differentialgleichungen

$$u_h^{k+1}(x_i) = g(x_i, t^{k+1}) \quad i = 0, \dots, N, \quad k \geq 0 \quad (11.6b)$$

$$u_h^0(x_i) = u_0(x_i) \quad i = 0, \dots, N. \quad (11.6c)$$

Bemerkung 11.3. Dieses System hat folgende Eigenschaften.

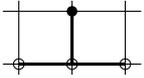
1. (11.6a)/(11.6b) ist eine Rekursionsgleichung für die Gitterfunktion zu den Zeitpunkten t^k .
2. In jedem Zeitschritt ist ein lineares Gleichungssystem

$$L_h u_h^{k+1} = M_h u_h^k + \tau f_h^k$$

zu lösen.

3. L_h ist diagonal im Fall $\theta = 0$, bzw. tridiagonal sonst.

Beispiel 11.4 (Bezeichnung der Standardverfahren).



$\theta = 0$ bezeichnet man als explizites Eulerverfahren.

Wegen $L_h = I$ berechnen sich die Werte von u_h^{k+1} ohne Lösen eines Gleichungssystems aus denen von u_h^k .



$\theta = 1$ ist das implizite Eulerverfahren.

L_h ist tridiagonal (räumlich 1D).



$\theta = \frac{1}{2}$ heißt Crank-Nicolson-Verfahren. Dies entspricht der Trapezregel für gewöhnliche Differentialgleichungen.

Hier ist auch ein Gleichungssystem zu lösen, aber das Verfahren ist genauer in der Zeit (siehe unten).

11.3 Fehleranalyse

Zur Fehleranalyse geht man ähnlich vor wie im elliptischen Fall.

Wir definieren die Fehlerfunktion zur Zeit t^k :

$$e_h^k = \underbrace{R_h}_{\text{Restriktionsoperator}} \underbrace{u(\cdot, t^k)}_{\text{exakte Lösung von (11.3) zur Zeit } t^k} - \underbrace{u_h^k}_{\text{durch das FD-Verfahren erzeugte Lösung}}$$

Für u_h^{k+1} gilt die Gleichung

$$L_h u_h^{k+1} = M_h u_h^k + \tau f_h^k.$$

Wir definieren z_h^{k+1} durch die Gleichung

$$L_h z_h^{k+1} = M_h \underbrace{R_h u(\cdot, t^k)}_{\substack{\text{exakte Werte} \\ \text{im letzten Zeitschritt}}} + \tau f_h^k$$

Für den Fehler in z_h^{k+1} (also nach einem Schritt mit exakten Werten) gilt:

$$\begin{aligned}
 L_h \left(R_h u(\cdot, t^{k+1}) - z_h^{k+1} \right) &= L_h R_h u(\cdot, t^{k+1}) - L_h z_h^{k+1} \\
 &= \underbrace{L_h R_h u(\cdot, t^{k+1}) - M_h R_h u(\cdot, t^k) - \tau f_h^k}_{\substack{\text{das ist die exakte Lösung, eingesetzt in die} \\ \text{Differenzgleichung (11.6) Dies ist aber (11.5) bis} \\ \text{auf den Fehlerterm!}}} \\
 &=: \eta_h^k \quad \text{„lokaler Abschneidefehler“}
 \end{aligned} \tag{11.7}$$

Aus (11.5) folgt

$$\| \eta_h^k \|_\infty = O(\tau h^2 + \tau^p) \quad \text{mit} \quad p = \begin{cases} 2 & 0 \leq \theta \leq 1, \theta \neq \frac{1}{2} \\ 3 & \theta = \frac{1}{2}. \end{cases}$$

Nun zurück zum globalen Fehler. Anwendung von L_h auf e_h liefert:

$$\begin{aligned}
 L_h e_h^{k+1} &= L_h \left(R_h u(\cdot, t^{k+1}) - u_h^{k+1} \right) \\
 &= \underbrace{L_h R_h u(\cdot, t^{k+1})}_{(11.7)} - \underbrace{L_h u_h^{k+1}}_{\text{Bem. 11.3}} \\
 &= \underbrace{M_h R_h u(\cdot, t^k) + \tau f_h^k + \eta_h^k}_{\substack{\text{aus (11.7) und Bem. 11.3}}} - \underbrace{M_h u_h^k + \tau f_h^k}_{\text{aus Bem. 11.3}} \\
 &= M_h \underbrace{\left(R_h u(\cdot, t^k) - u_h^k \right)}_{e_h^k} + \eta_h^k
 \end{aligned}$$

also:

$$\boxed{L_h e_h^{k+1} = M_h e_h^k + \eta_h^k} \quad \text{Rekursionsgleichung für den Fehler}$$

und hieraus folgt nach Auflösen

$$e_h^{k+1} = L_h^{-1} M_h e_h^k + L_h^{-1} \eta_h^k$$

Nun bilden wir wieder Normen und zwar wie gewohnt die Maximumsnorm:

$$\| e_h^{k+1} \|_\infty \leq \| L_h^{-1} M_h \|_\infty \| e_h^k \|_\infty + \| L_h^{-1} \|_\infty \| \eta_h^k \|_\infty$$

also

z. B. Rundungsfehler

↓

$$\begin{aligned}
 \| e_h^1 \|_\infty &\leq \| L_h^{-1} M_h \|_\infty \| e_h^0 \|_\infty + \| L_h^{-1} \|_\infty \| \eta_h^0 \|_\infty \\
 \| e_h^2 \|_\infty &= \| L_h^{-1} M_h \|_\infty \left(\| L_h^{-1} M_h \|_\infty \| e_h^0 \|_\infty + \| L_h^{-1} \|_\infty \| \eta_h^0 \|_\infty \right) + \| L_h^{-1} \|_\infty \| \eta_h^1 \|_\infty \\
 &= \| L_h^{-1} M_h \|_\infty^2 \| e_h^0 \|_\infty + \| L_h^{-1} M_h \|_\infty \| L_h^{-1} \|_\infty \| \eta_h^0 \|_\infty + \| L_h^{-1} \|_\infty \| \eta_h^1 \|_\infty \\
 \| e_h^k \|_\infty &= \| L_h^{-1} M_h \|_\infty^k \| e_h^0 \|_\infty + \sum_{i=0}^{k-1} \| L_h^{-1} M_h \|_\infty^{k-1-i} \| L_h^{-1} \|_\infty \| \eta_h^i \|_\infty
 \end{aligned} \tag{11.8}$$

$\theta = \frac{1}{2}$ (Crank-Nicolson)

stabil in $\|\cdot\|_\infty$ für $\tau \leq h^2$, uneingeschränkt stabil in der $\|\cdot\|_2$ -Norm.

Maximumprinzip gilt nur bei Einhaltung der Zeitschrittbedingung.

Konvergenzordnung $O(h^2 + \tau^2)$

$\theta = 0$ (expl. Euler)

stabil in $\|\cdot\|_\infty$ und $\|\cdot\|_2$ nur für $\tau \leq \frac{h^2}{2}$

Konvergenzordnung $O(h^2 + \tau)$

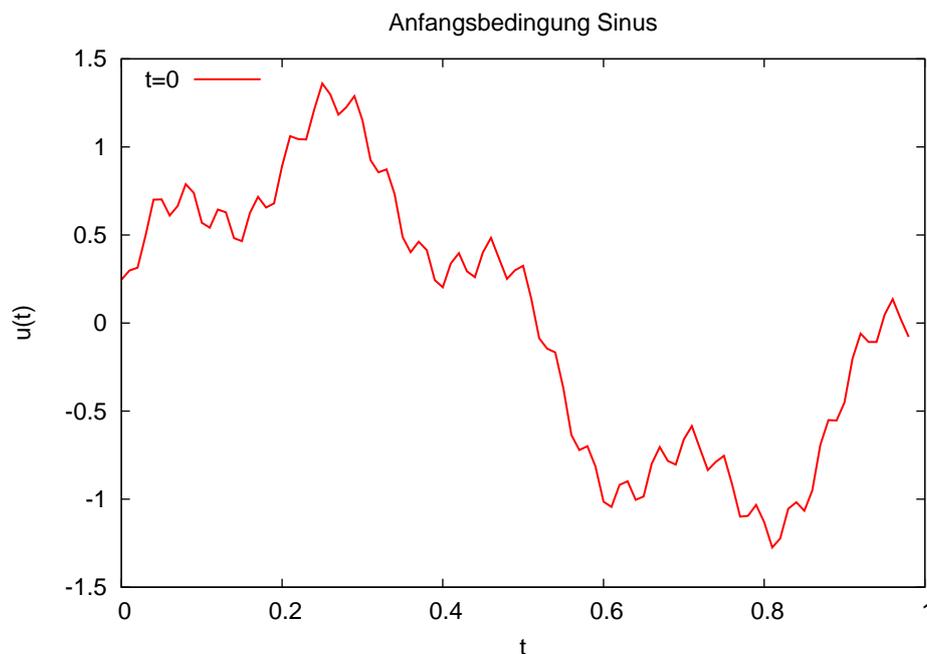
Bemerkung 11.5. Wegen der Glättungseigenschaft parabolischer Gleichungen wird die Lösung sich am Anfang schnell in der Zeit ändern. Mit fortschreitender Zeit sind (bei geeigneten Randbedingungen und rechter Seite) nur noch die örtlich langwelligen Anteile vorhanden, die sich entsprechend langsamer in der Zeit ändern.

Dementsprechend möchte man zu Beginn der Simulation kleine Zeitschritte und mit fortschreitender Zeit immer größere Zeitschritte verwenden. Die Einhaltung der Bedingung $\tau \leq Ch^2$ beim expliziten Euler (und auch bei Crank-Nicolson wenn man auf das Maximumprinzip Wert legt) verhindert dies. Das explizite Zeitschrittverfahren ist also ungeeignet.

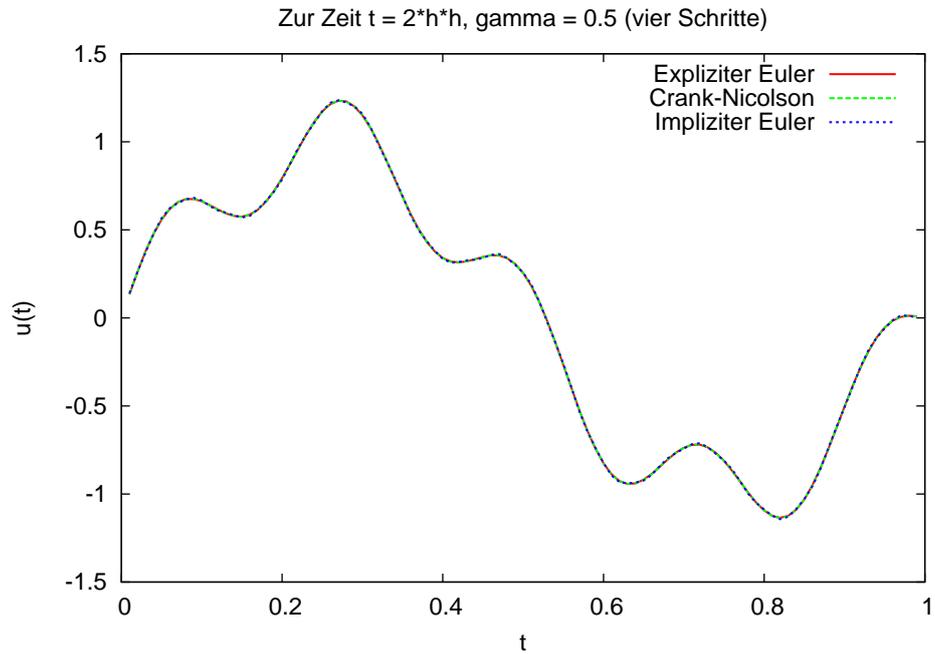
Die ortsdiskretisierte parabolische Gleichung (11.4) stellt ein steifes System gewöhnlicher Differentialgleichungen dar. Größter zu kleinster Eigenwert wächst wie $O(h^{-2})$ (dies ist ein diskretisierter elliptischer Operator). Daher ist auch verständlich, dass A-stabile Zeitdiskretisierungsverfahren erforderlich sind. \square

11.4 Numerischer Vergleich der Verfahren

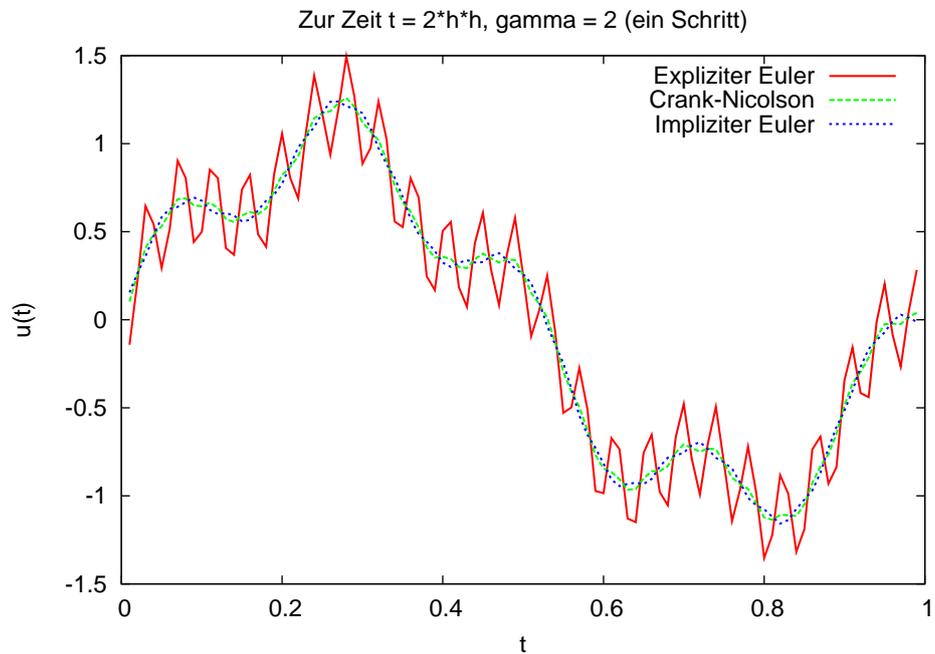
Wir lösen (11.1) mit $\Delta t = \gamma \cdot h^2$ und der Anfangsbedingung:



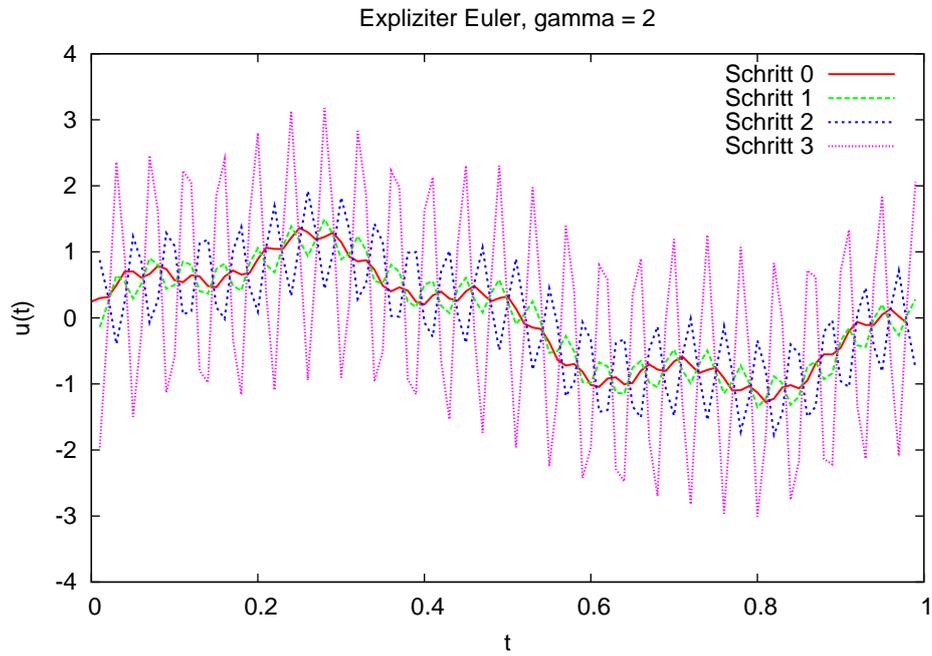
11 Parabolische partielle Differentialgleichungen



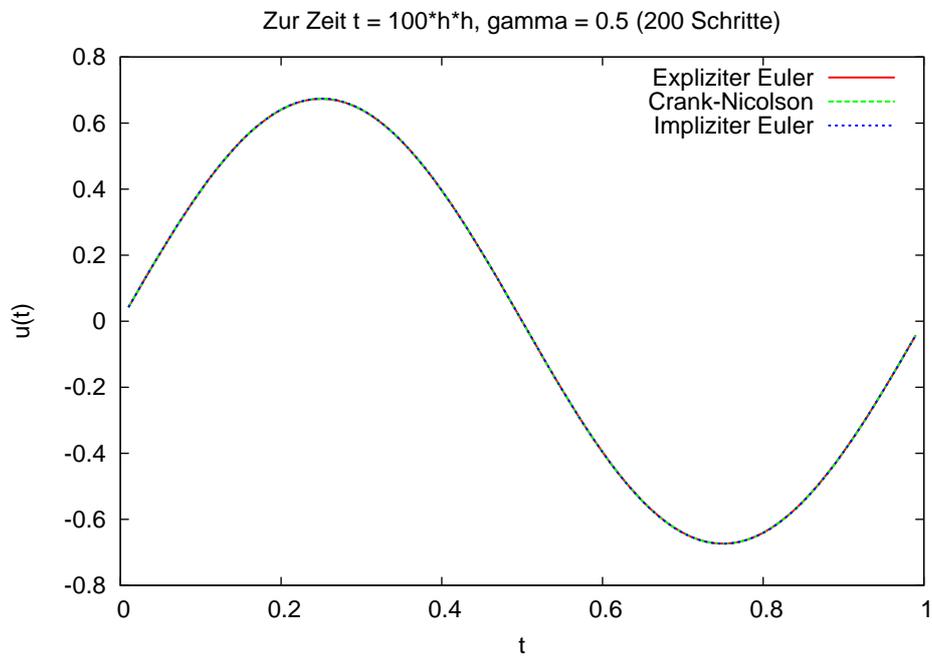
Zunächst mit $\gamma = 1/2$ und den drei Verfahren $\theta = 0, 1/2, 1$ nach 4 Schritten.



Jetzt mit $\gamma = 2$ und einem Schritt. Stabilität verletzt für $\theta = 0, 1/2$.

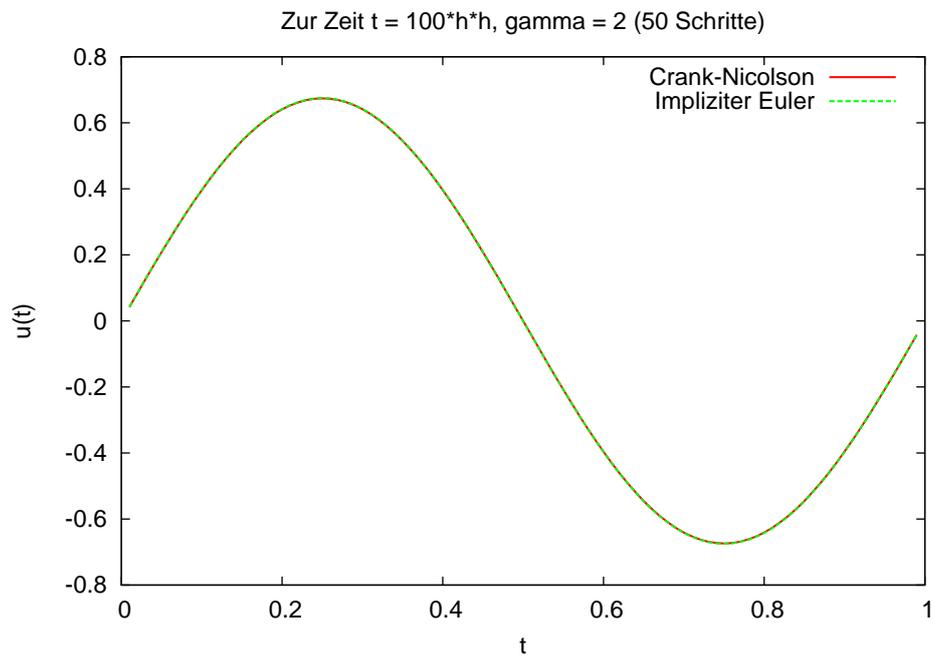


Der explizite Euler bei $\gamma = 2$: instabil.

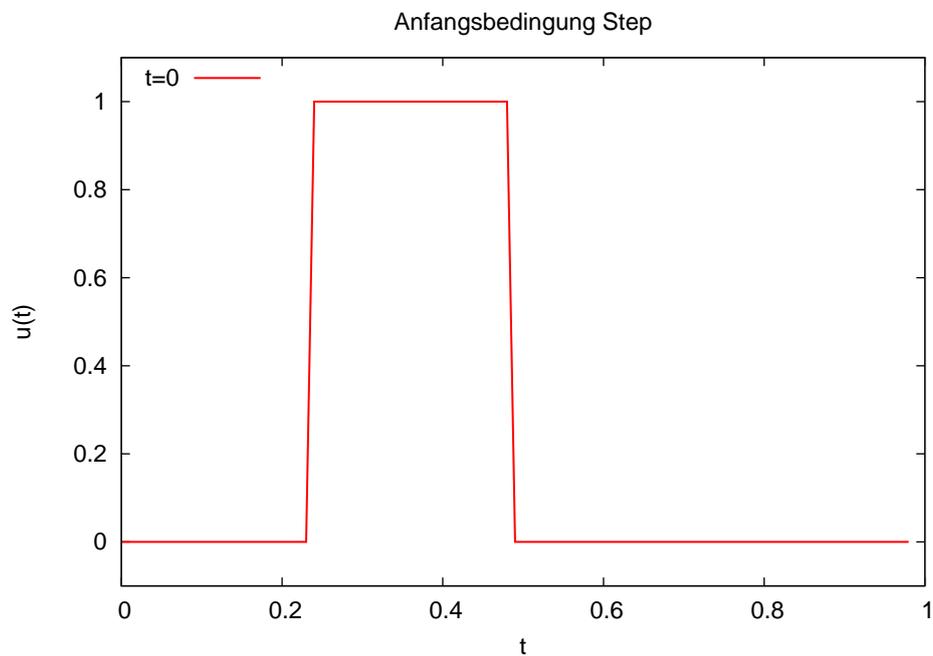


$\gamma = 1/2$ und 200 Schritte.

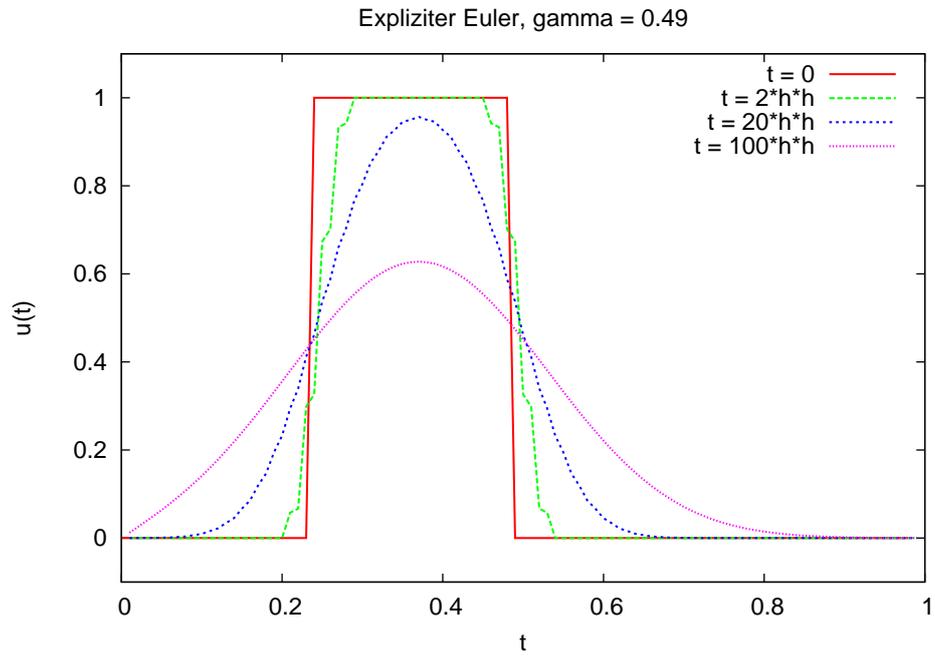
11 Parabolische partielle Differentialgleichungen



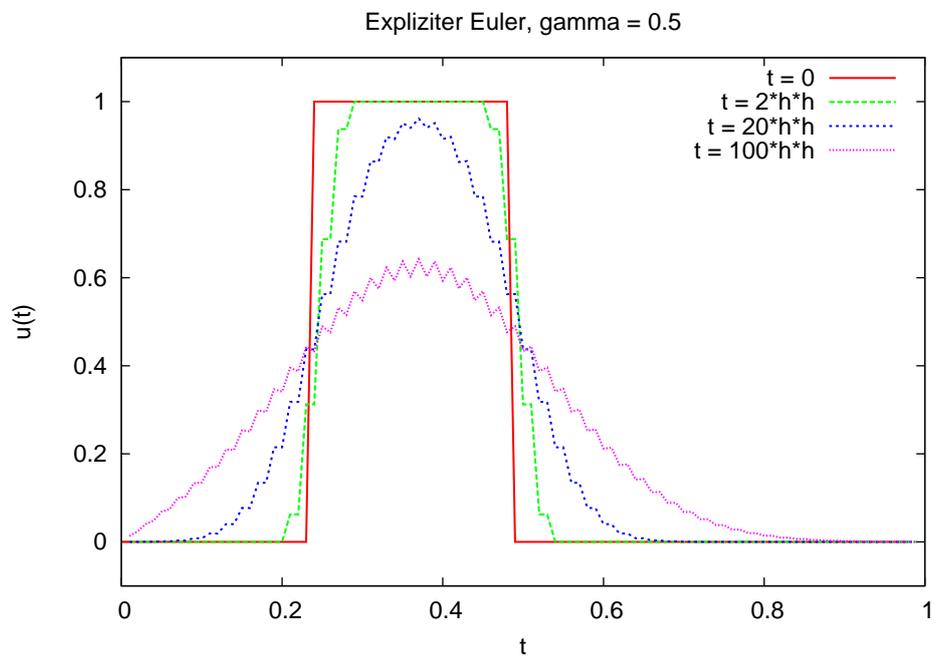
$\gamma = 2$ und 50 Schritte: $\theta = 1/2$ doch stabil?



Nun testen wir diese Anfangsbedingung.

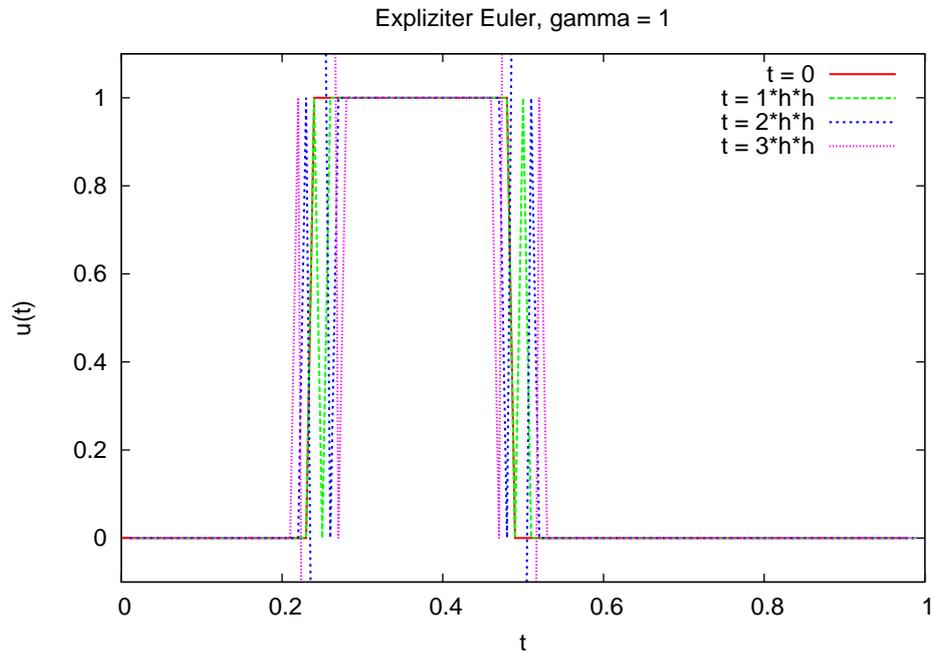


Expliziter Euler bei $\gamma = 49/100$: Stabil.

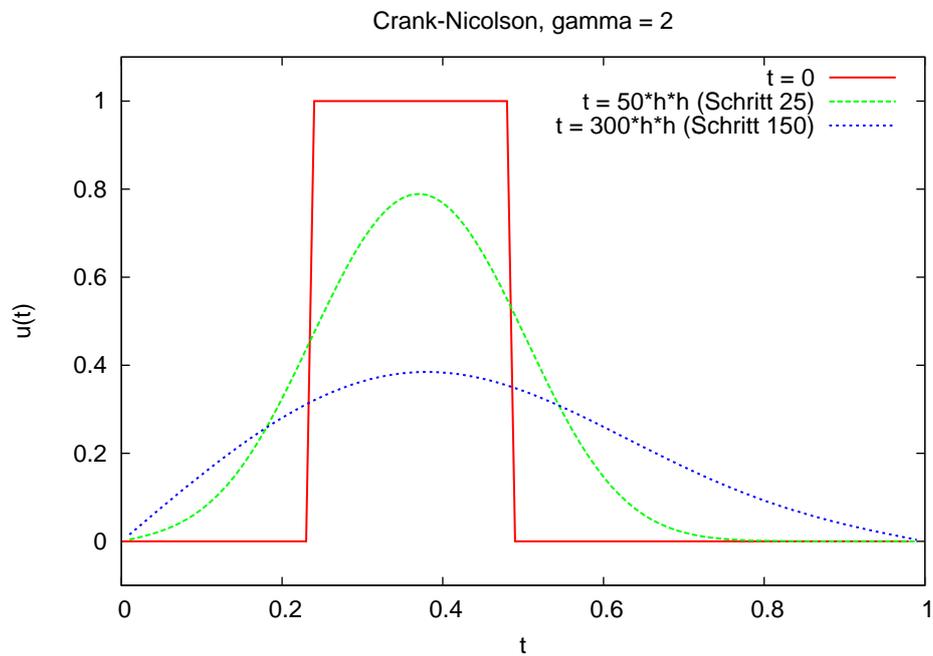


Expliziter Euler bei $\gamma = 1/2$: Das ist die Grenze.

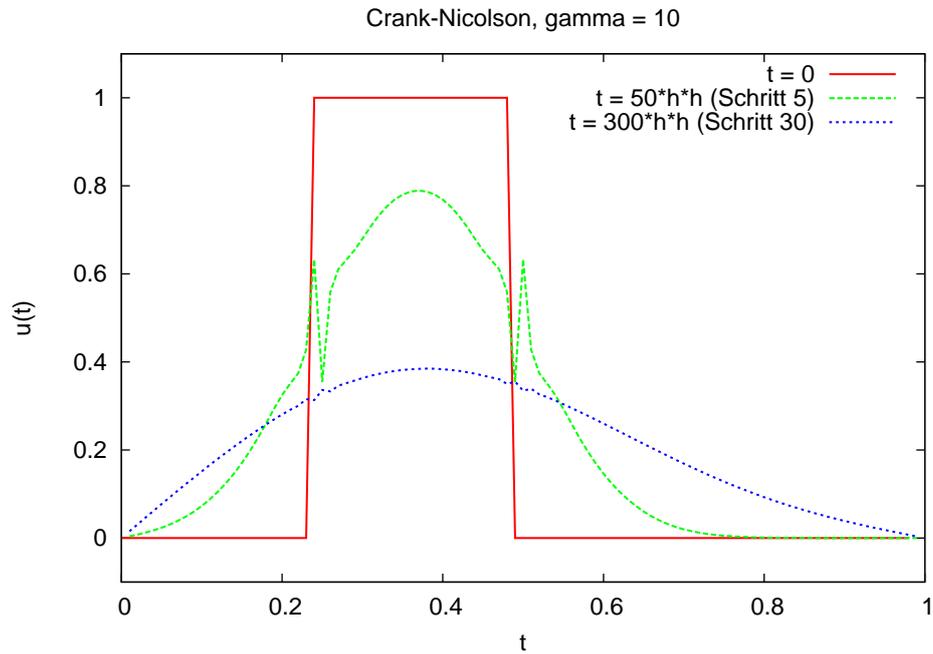
11 Parabolische partielle Differentialgleichungen



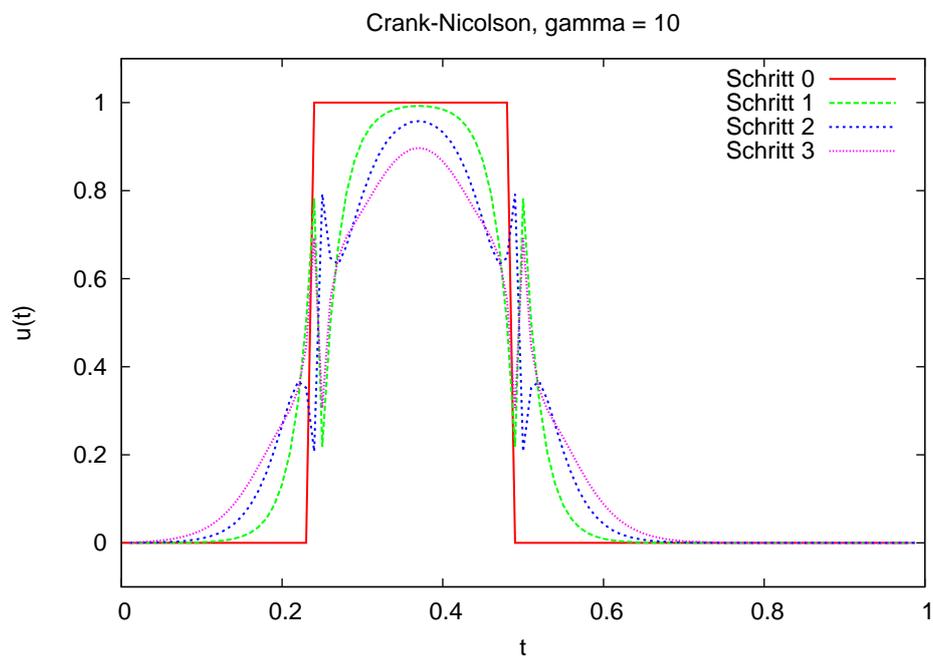
Expliziter Euler für $\gamma = 1$: instabil.



Crank-Nicolson für $\gamma = 2$: Sieht gut aus.

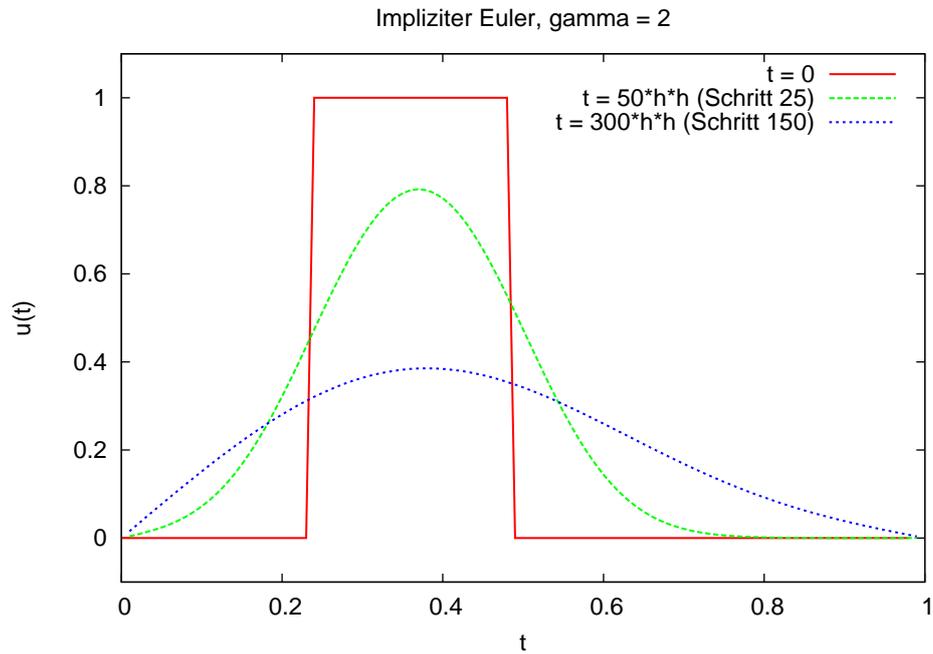


... auch noch für $\gamma = 10$ (bis auf die seltsamen Zacken).

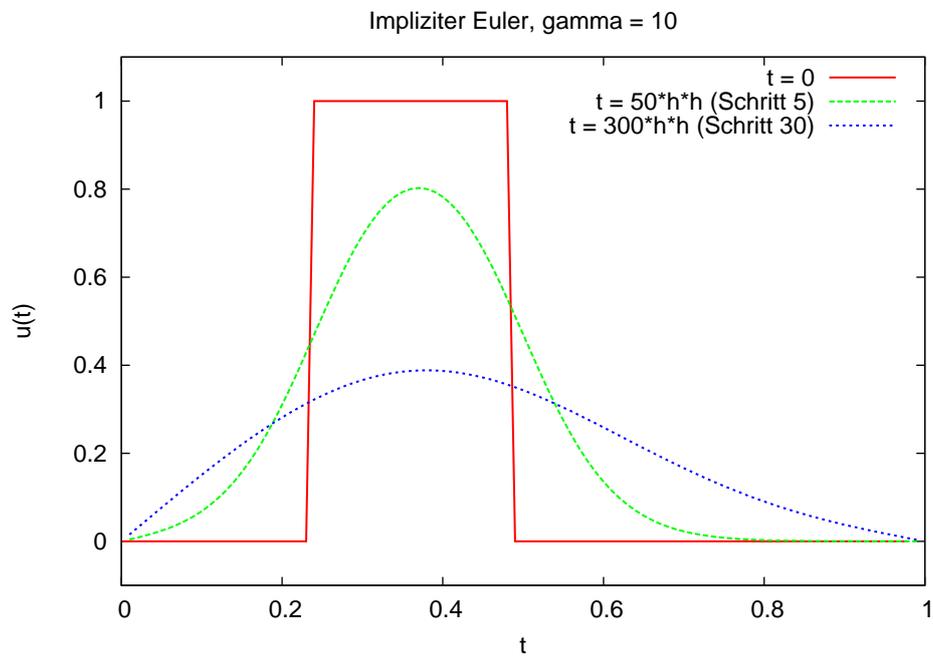


Crank-Nicolson für $\gamma = 10$: In den ersten Zeitschritten unphysikalisches Verhalten um die Sprungstelle.

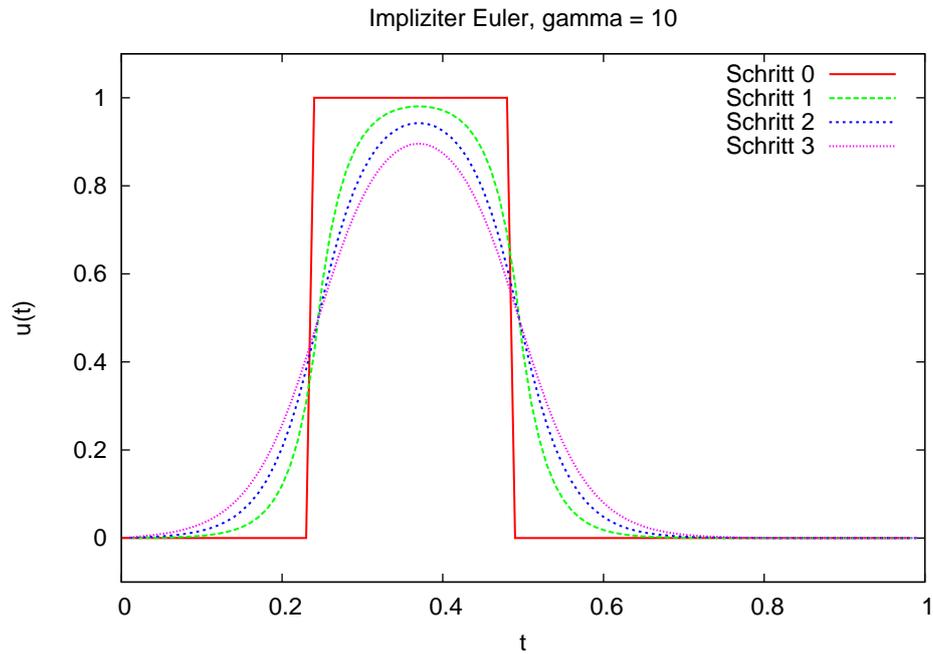
11 Parabolische partielle Differentialgleichungen



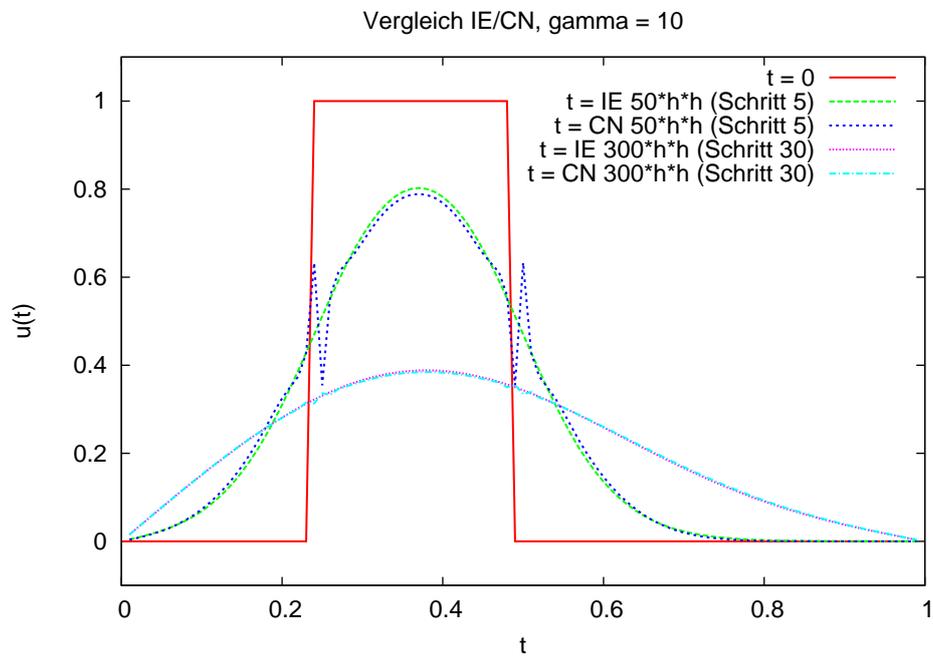
Impliziter Euler für $\gamma = 2$: Stabil.



... und für $\gamma = 10$: Auch stabil.



Auch in den ersten Schritten zeigt der implizite Euler kein unphysikalisches Verhalten.



Aber: Crank-Nicolson hat asymptotisch eine bessere Konvergenzrate.

11.5 Zusammenfassung

- Parabolische Gleichungen besitzen Lösungen, die mit fortschreitender Zeit immer glatter werden.
- Bei der Linienmethode diskretisiert man zunächst im Ort, das dann entstehende System gewöhnlicher DGL wird dann in der Zeit diskretisiert.
- Zur Zeitdiskretisierung setzt man A-stabile und damit implizite Verfahren ein, da diese für steife Systeme geeignet sind. Explizite Verfahren erfordern bei hinreichend kleiner Ortsschrittweite extrem kleine Zeitschritte.

12 Finite Differenzen für lineare hyperbolische Gleichungen

Wir betrachten die mehrdimensionale, hyperbolische, lineare Transportgleichung

$$\begin{aligned} \frac{\partial u}{\partial t} + \nabla \cdot (vu) &= f && \text{in } \Omega \times T \\ u &= g && \text{auf } \Gamma_{in} = \{(x, t) \in \partial\Omega \times T \mid v(x) \cdot n(x) < 0\} \\ &&& \uparrow \\ &&& \text{äußere Normale} \\ u &= u_0 && \text{für } t = 0 \end{aligned} \quad (12.1)$$

für ein gegebenes Geschwindigkeitsfeld $v: \Omega \times T \rightarrow \mathbb{R}^d$.

Oft beschränken wir uns auf den räumlich eindimensionalen Fall mit konstanter Geschwindigkeit $a > 0$:

$$\begin{aligned} \frac{\partial u}{\partial t} + \frac{\partial (au)}{\partial x} &= 0 && \text{in } (0, 1) \times (0, \infty) \\ u(0, t) &= g(t) \\ u(x, 0) &= u_0(x) \end{aligned} \quad (12.2)$$

12.1 Methode der Charakteristiken

Unter Annahme von $\nabla \cdot v = 0$ (Quell-/Senken-freies Fließfeld) und $f = 0$ folgt aus (12.1):

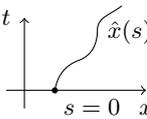
$$\begin{aligned} \frac{\partial u}{\partial t} + v \cdot \nabla u + \underbrace{(\nabla \cdot v)}_{=0} u &= 0 \\ \iff \frac{\partial u}{\partial t} + v \cdot \nabla u &= 0 \end{aligned}$$

Dies nennt man die „nichtkonservative“ Form der hyperbolischen Gleichung.

Sei $(\hat{x}(s), \hat{t}(s))$ eine Kurve in $\Omega \times T$ parametrisiert mit s .

Berechne Ableitung von u in Richtung der Kurve

$$\frac{d}{ds} \left[u(\hat{x}(s), \hat{t}(s)) \right] = \sum_{i=1}^d \frac{\partial u}{\partial x_i} \Big|_{(\hat{x}(s), \hat{t}(s))} \cdot \frac{\partial \hat{x}_i}{\partial s} \Big|_s + \frac{\partial u}{\partial t} \Big|_{(\hat{x}(s), \hat{t}(s))} \cdot \frac{\partial \hat{t}}{\partial s} \Big|_s \quad (12.3)$$



Bis jetzt war die Kurve beliebig. Nun wählen wir eine spezielle:

$$\begin{aligned} \frac{d\hat{t}}{ds} \Big|_s &= 1, && \hat{t}(0) = t_0 \\ \frac{d\hat{x}_i}{ds} \Big|_s &= v_i(\hat{x}(s), \hat{t}(s)), && \hat{x}_i(0) = x_{0,i} \end{aligned} \quad (12.4)$$

Dies funktioniert auch für unstetige Anfangsbedingungen!

$$u_0(x) = \begin{cases} 1 & x \geq 0 \\ 0 & \text{sonst} \end{cases}$$

Die Sprungstelle pflanzt sich mit Geschwindigkeit a nach rechts fort.

$$u(x, t) = \begin{cases} 1 & x \geq a \cdot t \\ 0 & \text{sonst} \end{cases}$$

Allgemein (mit Rand, mehrdimensional): Definiere den „Trackingoperator“ $\Phi(x, t, t') \in \bar{\Omega}$ über:
D. h. $\Phi(x, t, t')$ verfolgt den Punkt (x, t) bis zur Zeit t' und liefert dann die Position.

Löse (12.4) für $t_0 = t$, $x_0 = x$.

Setze $\Phi(x, t, t') = \hat{x}(s^*)$, sodass $\hat{t}(s^*) = t'$.

$$u(x, t) = \begin{cases} u_0(\Phi(x, t, 0)) & \text{falls } \Phi(x, t, 0) \in \bar{\Omega} \\ g(t^*) & \text{für } \Phi(x, t, t^*) \in \partial\Omega \end{cases}$$

12.2 Finite Differenzen

Selber Ansatz wie bei parabolischen Gleichungen: Linienmethode.

Gleich die vordiskreten Versionen:

Zweite Ordnung im Ort (zentrale Differenz), Einschritt- θ -Verfahren in der Zeit:

$$\begin{aligned} \frac{u_h^{k+1}(x_i) - u_h^k}{\tau} + \frac{(1-\theta)a}{2h} [u_h^k(x_{i+1}) - u_h^k(x_{i-1})] \\ + \frac{\theta a}{2h} [u_h^{k+1}(x_{i+1}) - u_h^{k+1}(x_{i-1})] = 0 \quad k \geq 0, i = 1, \dots, N-1 \end{aligned}$$

$$\begin{aligned} \iff -\frac{\tau\theta a}{2h} u_h^{k+1}(x_{i-1}) + u_h^{k+1}(x_i) + \frac{\tau\theta a}{2h} u_h^{k+1}(x_{i+1}) = \\ = \frac{\tau(1-\theta)a}{2h} u_h^k(x_{i-1}) + u_h^k(x_i) - \frac{\tau(1-\theta)a}{2h} u_h^k(x_{i+1}) \end{aligned}$$

Selbe Struktur wie im parabolischen Fall $L_h u_h^{k+1} = M_h u_h^k$.

- L_h ist *keine* M-Matrix (pos. Vorzeichen) falls $\theta > 0$.
- L_h ist nicht symmetrisch.

12 Finite Differenzen für lineare hyperbolische Gleichungen

- L_h diagonaldominant für $2 \cdot \frac{\tau\theta a}{2h} < 1$

$$\begin{array}{ll} \text{also } \theta = 0 & \text{und } \tau, h, a \text{ beliebig, klar: } L_h = I \\ \theta \neq 0 & \text{und } \tau < \frac{h}{\theta a}. \end{array}$$

- Bemerkung: Wie behandelt man den rechten Rand bei $a > 0$? Wie haben bei der Herleitung der Differenzenmethode den Rand noch nicht berücksichtigt.

$\theta = 0$, expliziter Fall

$$\begin{aligned} u_h^{k+1} &= M_h u_h^k \quad \text{mit } M_h = \text{tridiag} \left(-\frac{\tau a}{2h}, 1, \frac{\tau a}{2h} \right) \\ &\Rightarrow \|M_h\|_\infty = 1 + \frac{\tau|a|}{h} > 1 \text{ für alle } \tau, h \\ &\Rightarrow \text{Verfahren ist uneingeschränkt instabil in der Maximumnorm} \end{aligned}$$

$\theta = 1$, voll impliziter Fall

$$L_h u_h^{k+1} = u_h^k \quad \text{mit } L_h = \text{tridiag} \left(\frac{\tau a}{2h}, \overset{\text{keine M-Matrix mehr!}}{1}, -\frac{\tau a}{2h} \right)$$

Numerische Resultate unten zeigen, dass Verfahren stabil für $\frac{\tau}{h} \geq C(a)$, also τ *groß genug* (!), allerdings nicht in der Maximumnorm.

mit einseitiger Differenz im Ort (welche?).

$$\begin{aligned} &\frac{u_h^{k+1}(x_i) - u_h^k(x_i)}{\tau} + \frac{(1-\theta)a}{h} [u_h^k(x_i) - u_h^k(x_{i-1})] + \frac{\theta a}{h} [u_h^{k+1}(x_i) - u_h^{k+1}(x_{i-1})] = 0 \\ \Leftrightarrow &-\frac{\tau\theta a}{h} u_h^{k+1}(x_{i-1}) + \left(1 + \frac{\tau\theta a}{h}\right) u_h^{k+1}(x_i) = \frac{\tau(1-\theta)a}{h} u_h^k(x_{i-1}) + \left(1 - \frac{\tau(1-\theta)a}{h}\right) u_h^k(x_i) \end{aligned}$$

wieder $L_h u_h^{k+1} = M_h u_h^k$

L_h ist eine M-Matrix, falls $a \geq 0$. Im Fall $a < 0$ wählt man die andere einseitige Differenz

$$\frac{\partial u}{\partial x}(x_i, t) = \frac{u(x_{i+1}, t) - u(x_i, t)}{h} + O(h)$$

under erhält eine M-Matrix!

D. h. die Wahl der Differenz hängt vom Vorzeichen von a ab.

- L_h ist unsymmetrisch, aber bi-diagonal.
- Randbedingung am rechten Rand entfällt nun wie im kontinuierlichen Problem!

$\theta = 0$, expliziter Fall

$$u_h^{k+1} = M_h u_h^k \text{ mit } M_h = \text{bidiag} \left(\frac{\tau a}{h}, 1 - \frac{\tau a}{h} \right)$$

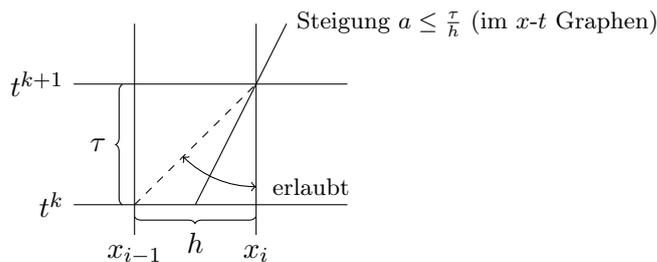
Diagonale
↓

$$\|M_h\|_\infty = \left| \frac{\tau a}{h} \right| + \left| 1 - \frac{\tau a}{h} \right| = 1, \text{ falls } 0 \leq \frac{\tau a}{h} \leq 1$$

$\frac{\tau a}{h} \geq 0$ klar wegen $a > 0$, ansonsten wählt man wieder die andere Differenz!

$\frac{\tau a}{h} \leq 1$ heißt CFL-Bedingung nach Courant, Friedrich, Levy (1928)

anschaulich:



$\theta = 1$, impliziter Fall

$$L_h u_h^{k+1} = u_h^k \quad L_h = \text{bidiag} \left(-\frac{\tau a}{h}, 1 + \frac{\tau a}{h} \right)$$

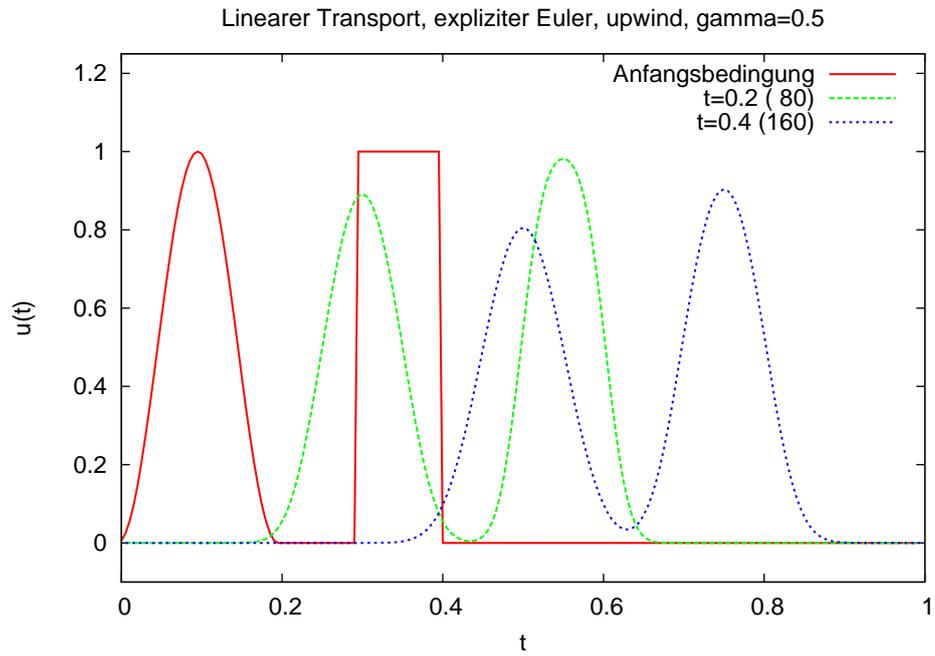
$$\|L_h\|_\infty \leq 1 \text{ für alle } \frac{\tau}{h} \text{ nach Satz 6.2, denn } L_h \mathbb{1} \geq \mathbb{1}$$

Verfahren ist uneingeschränkt stabil!

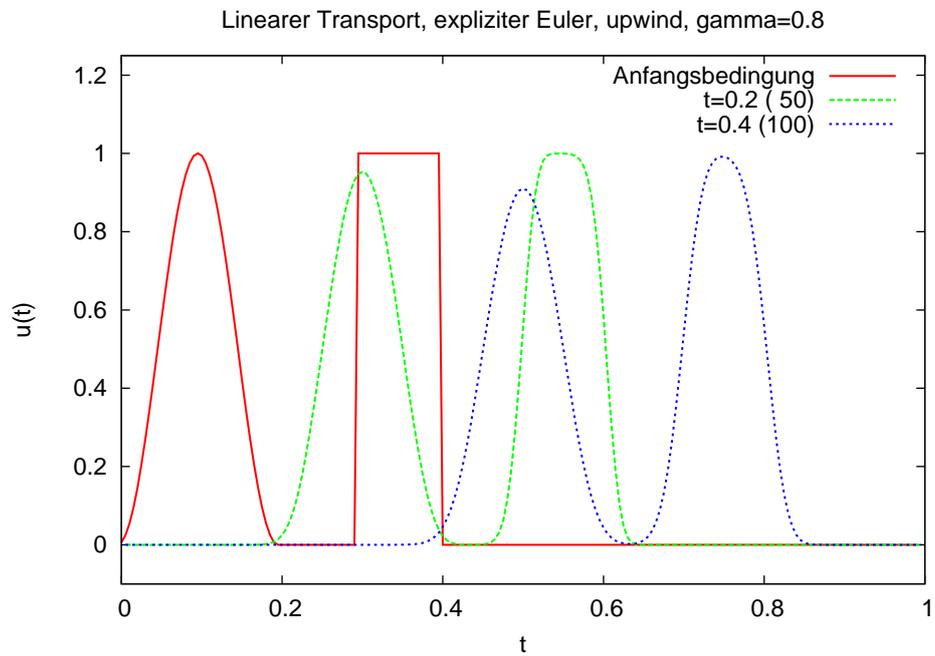
12.3 Numerischer Vergleich

Modellproblem, $a = 1$, $h = 1/200$.

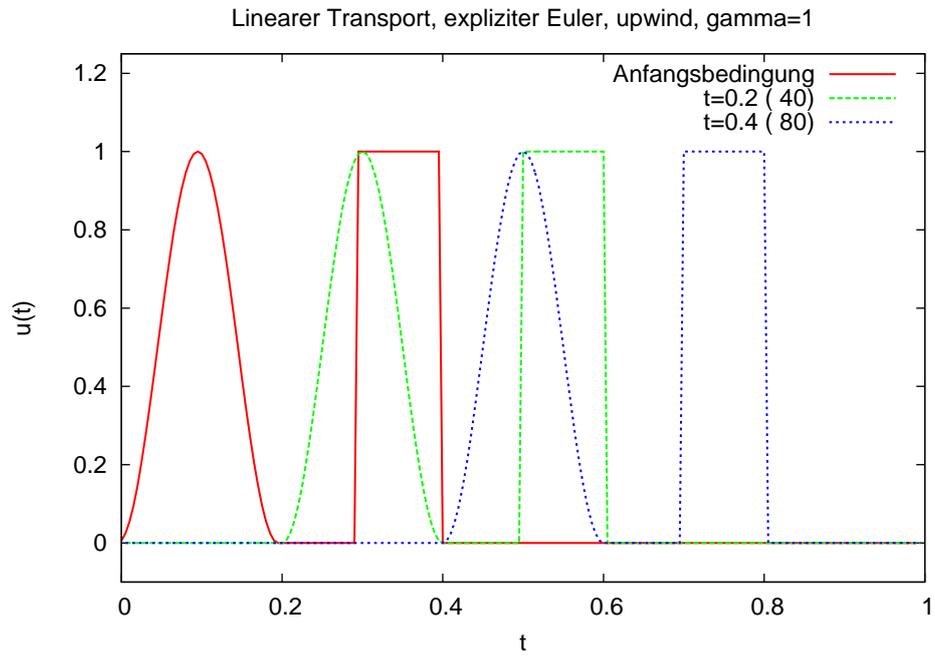
12 Finite Differenzen für lineare hyperbolische Gleichungen



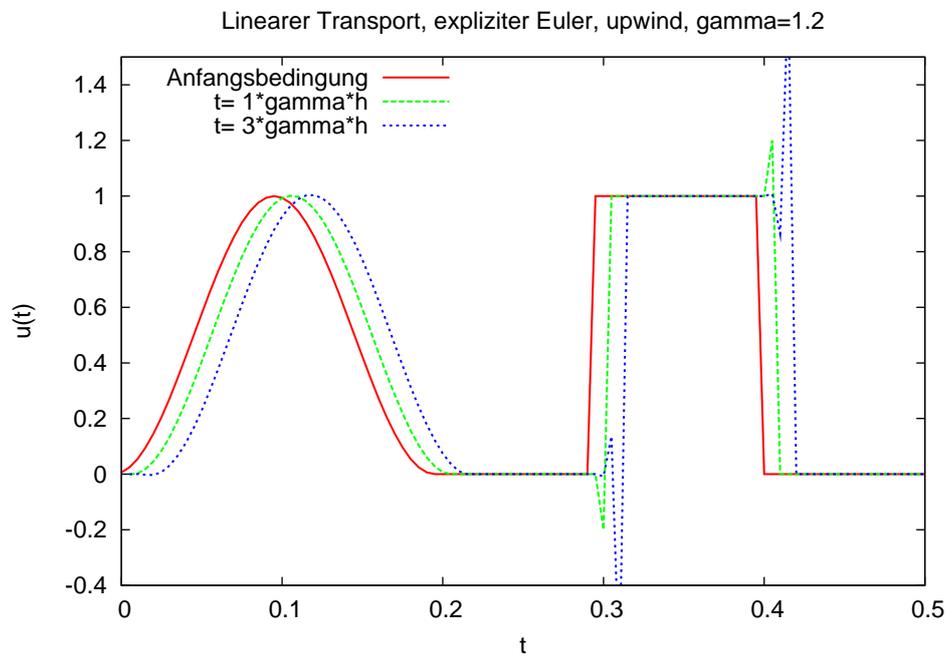
Expliziter Euler, upwind bei $\gamma = 1/2$.



Expliziter Euler, upwind bei $\gamma = 4/5$.

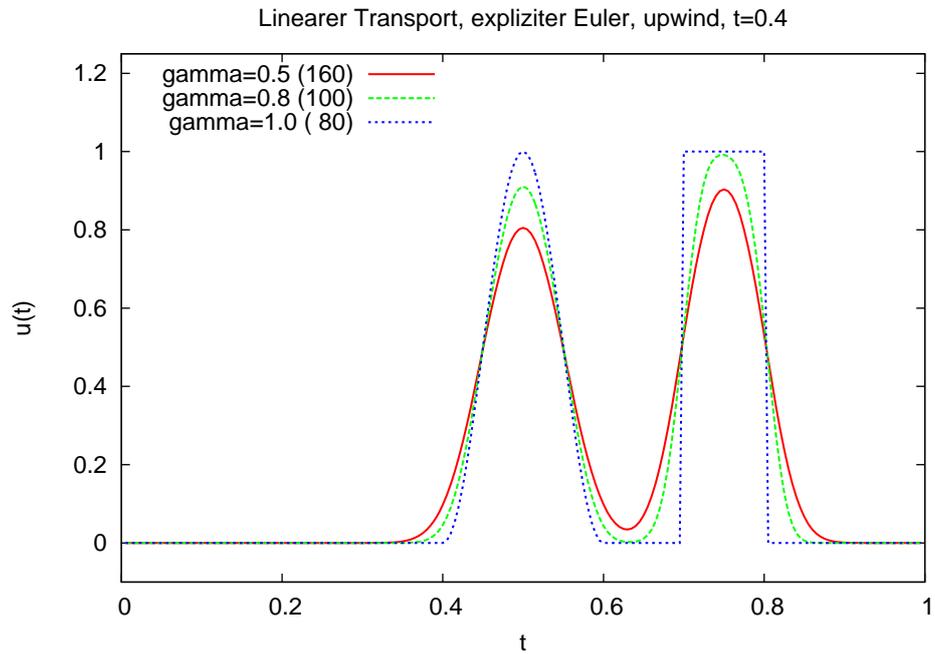


Expliziter Euler, upwind bei $\gamma = 1$.

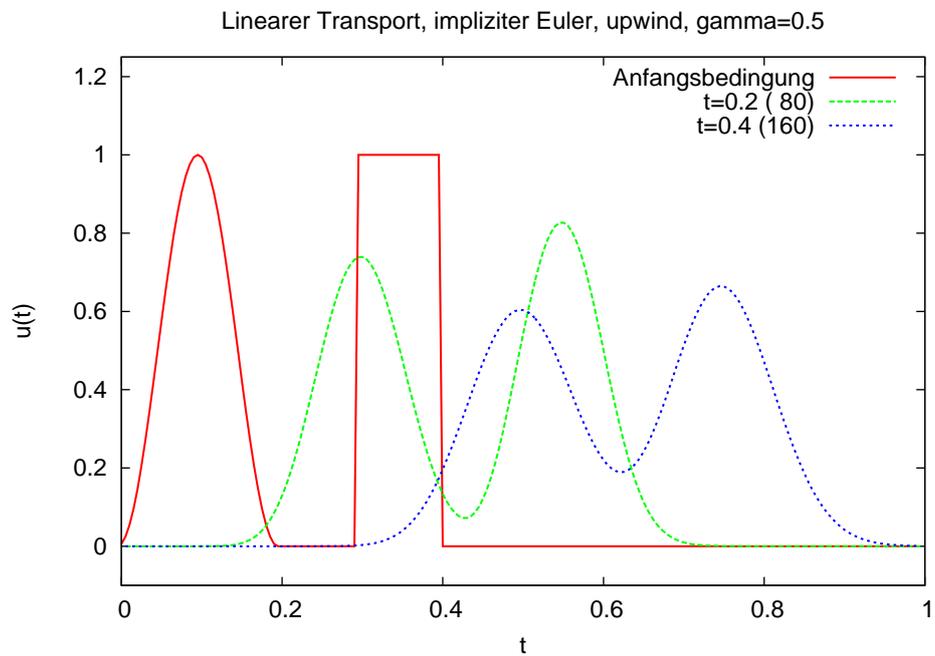


Expliziter Euler, upwind bei $\gamma = 1.2$: Courantbedingung ist scharf.

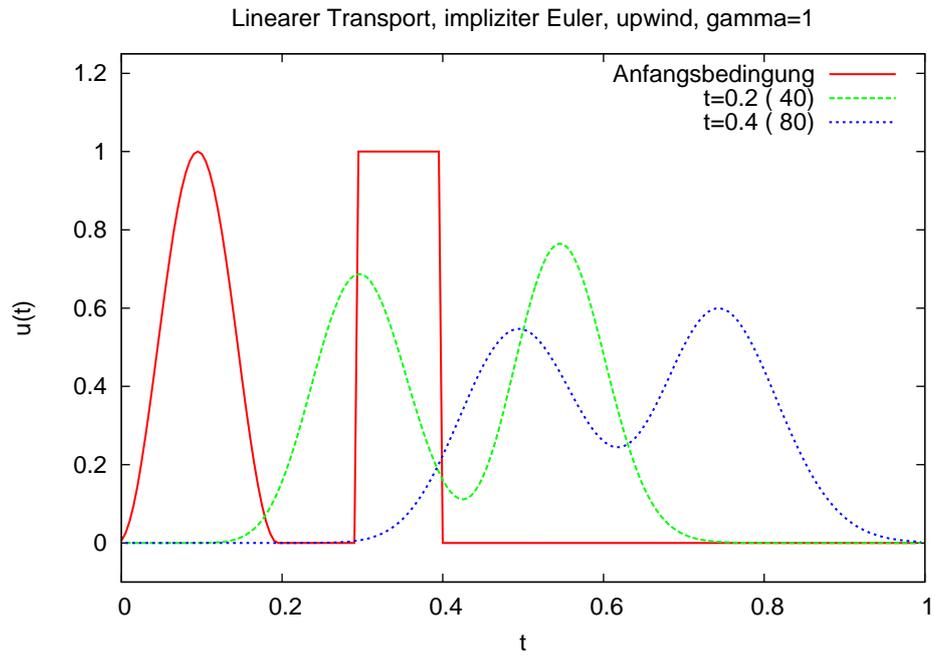
12 Finite Differenzen für lineare hyperbolische Gleichungen



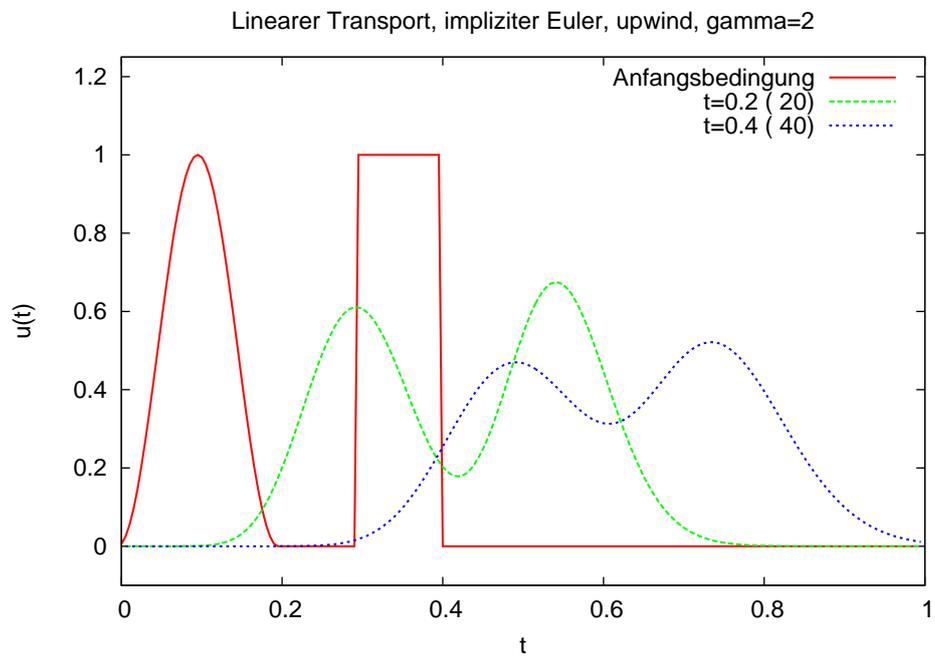
Expliziter Euler, upwind: Stabil für $\gamma \leq 1$, wird besser mit steigendem γ .



Impliziter Euler mit upwind bei $\gamma = 0.5$.

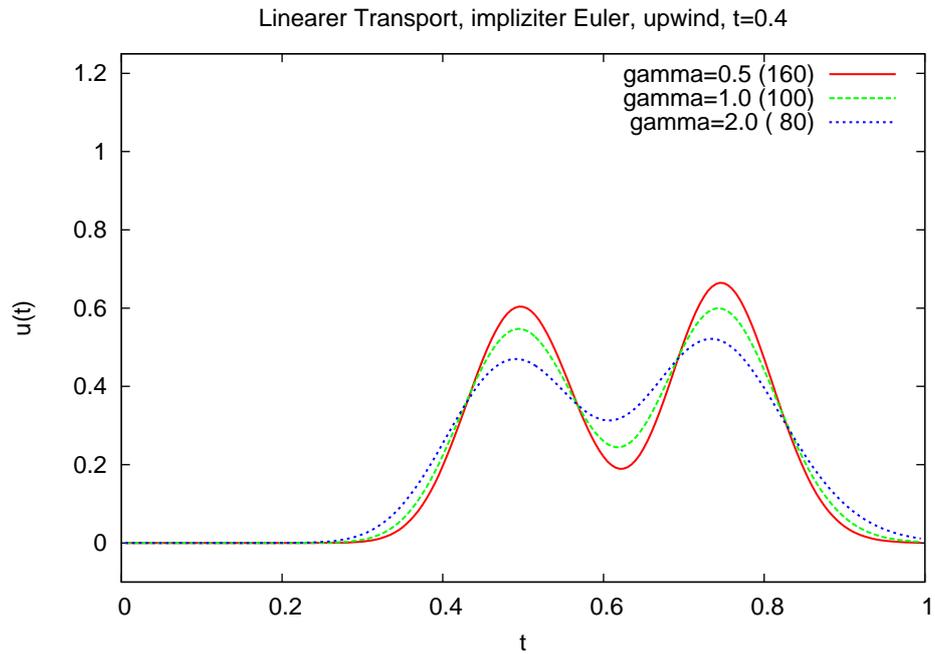


Impliziter Euler mit upwind bei $\gamma = 1$.

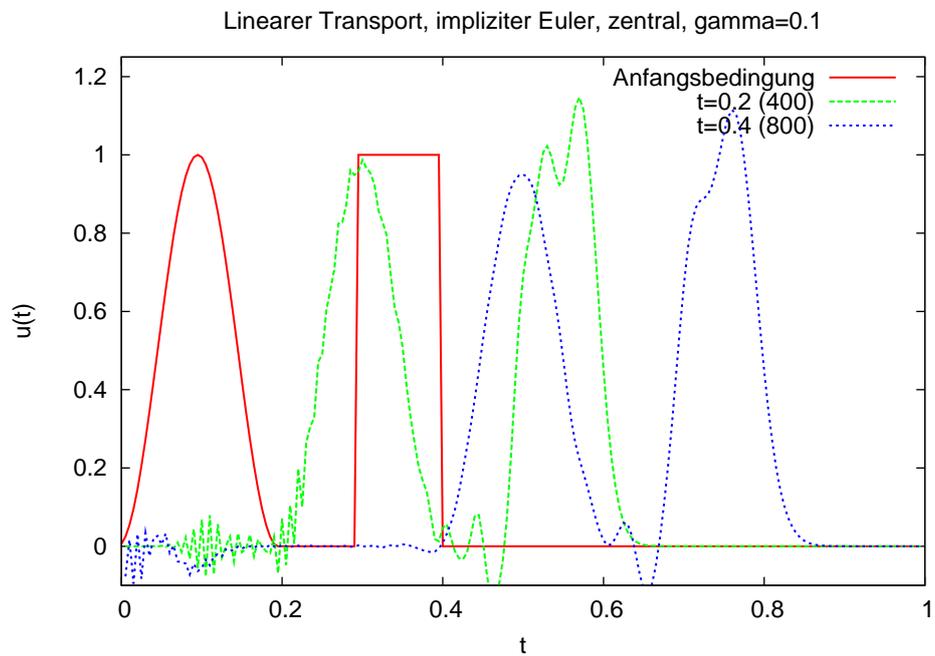


Impliziter Euler mit upwind bei $\gamma = 2$.

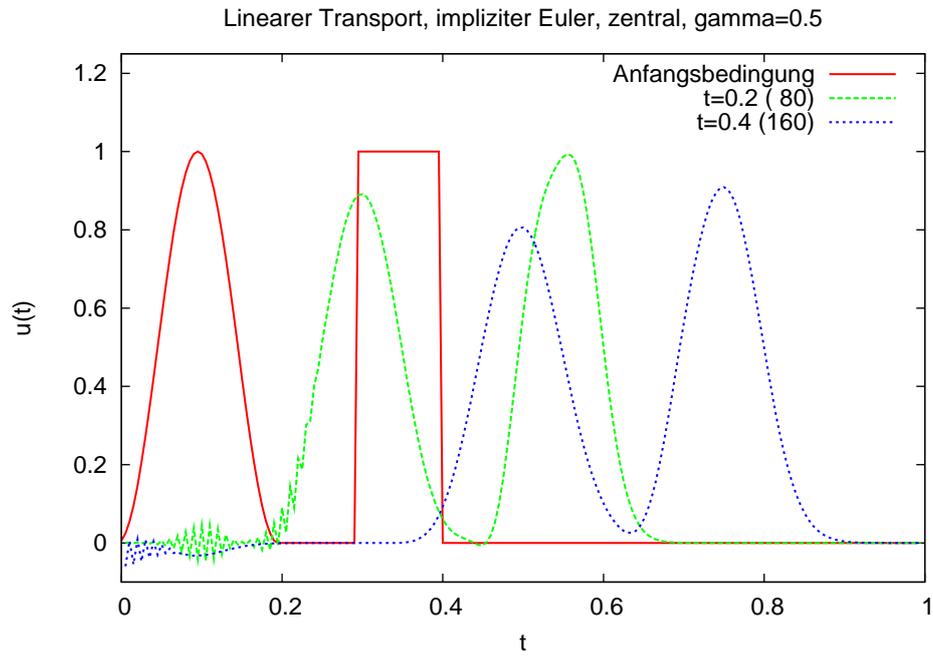
12 Finite Differenzen für lineare hyperbolische Gleichungen



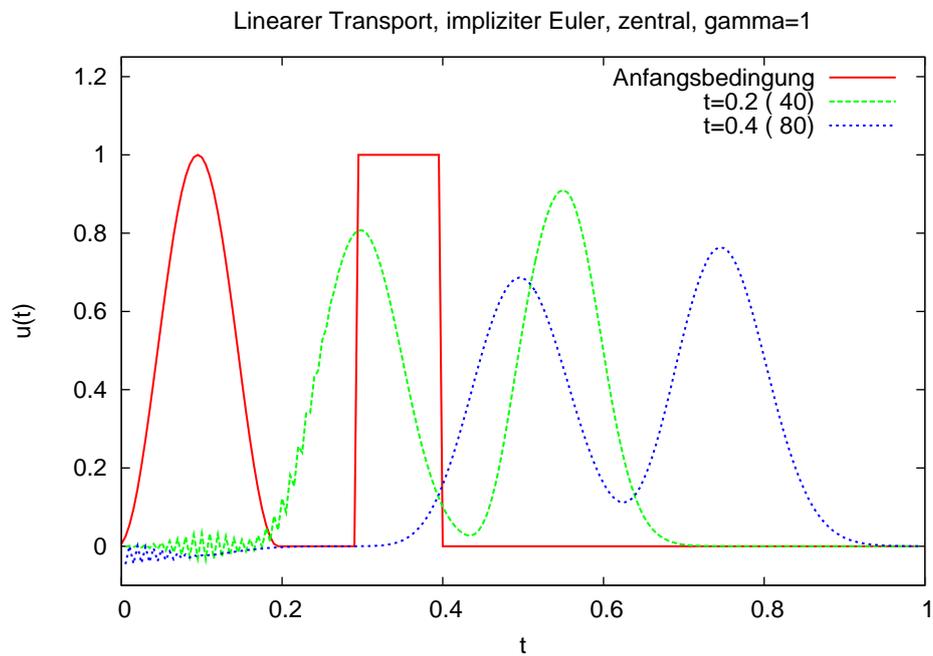
Impliziter Euler mit upwind: Stabil für alle γ aber diffusiv. Wird schlechter mit steigendem γ



Impliziter Euler mit zentraler Differenz bei $\gamma = 0.1$.

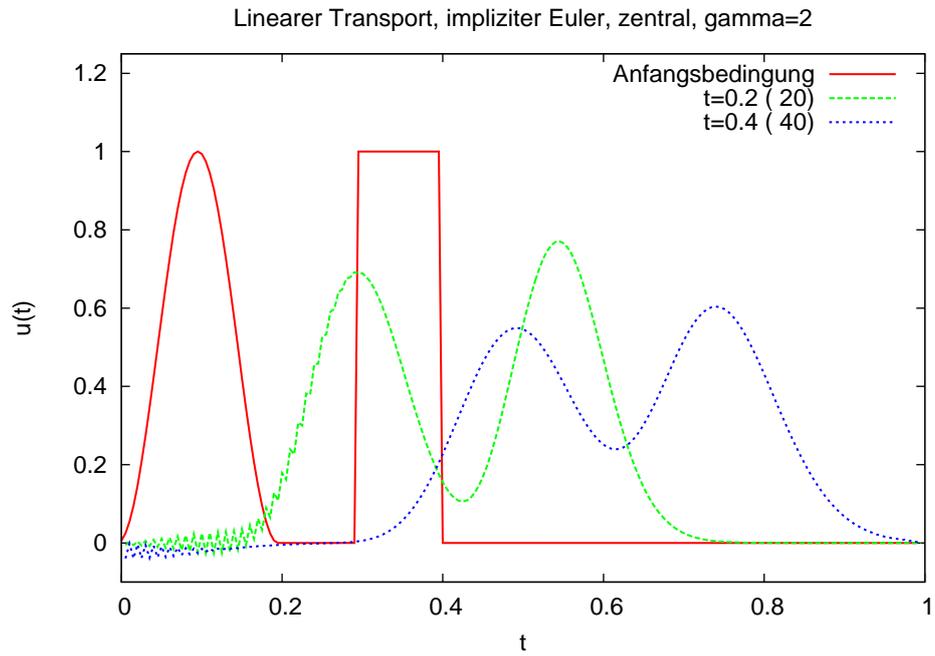


Impliziter Euler mit zentraler Differenz bei $\gamma = 0.5$.

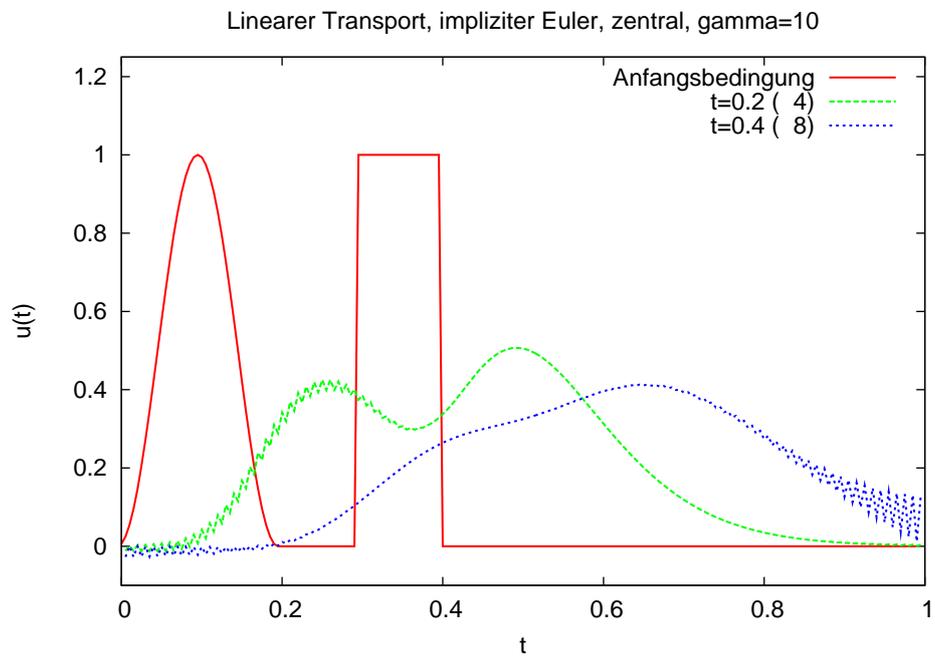


Impliziter Euler mit zentraler Differenz bei $\gamma = 1$.

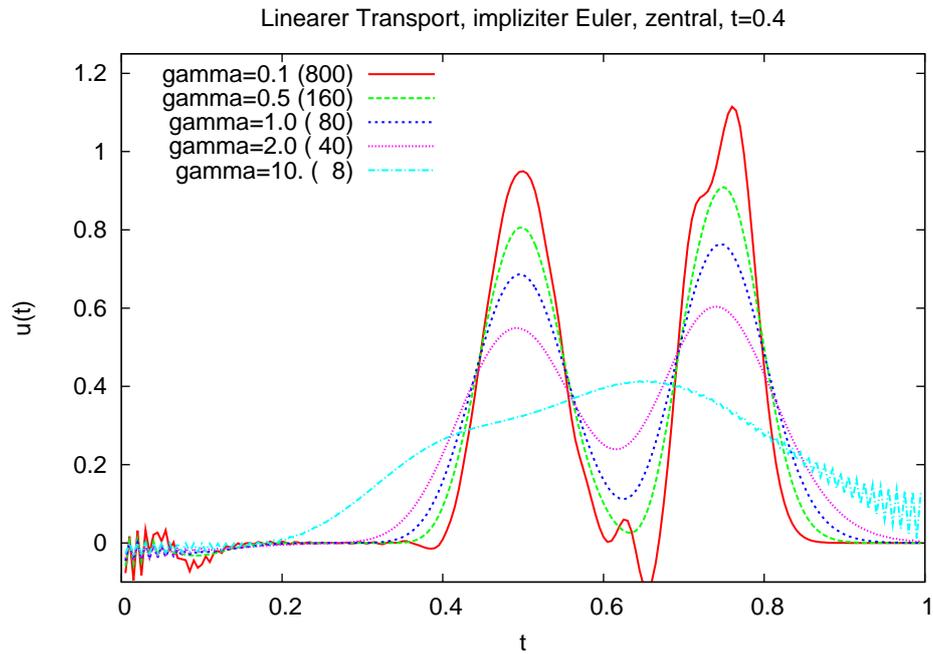
12 Finite Differenzen für lineare hyperbolische Gleichungen



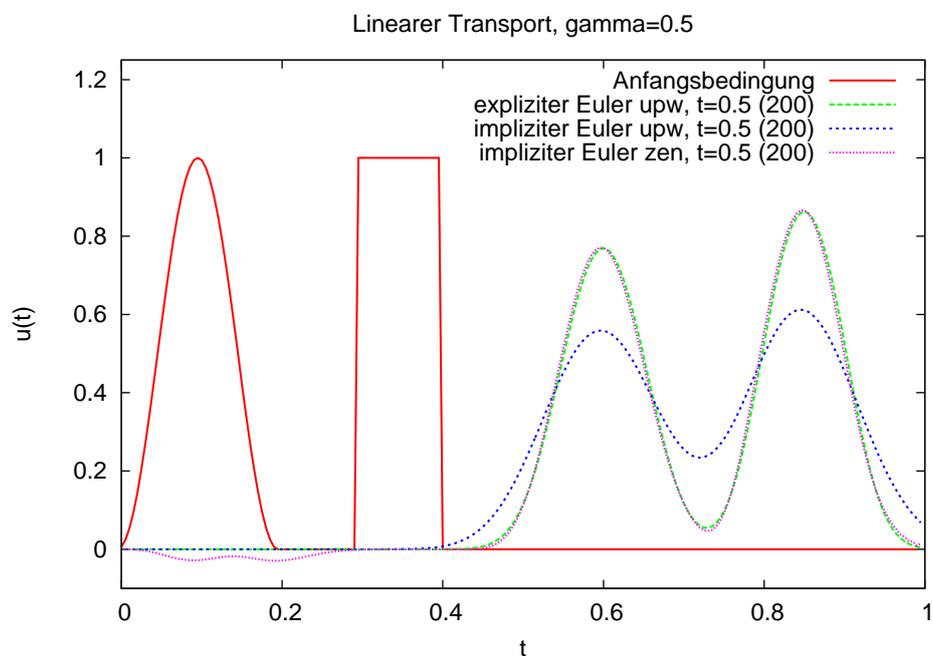
Impliziter Euler mit zentraler Differenz bei $\gamma = 2$.



Impliziter Euler mit zentraler Differenz bei $\gamma = 10$.



Impliziter Euler mit zentraler Differenz: Diffusiv, Oszillationen werden weniger mit steigendem γ .



Vergleich aller Verfahren bei $\gamma = 0.5$: Expliziter Euler mit Upwind ist die Methode der Wahl.

12.4 Numerische Diffusion

Einen Hinweis darauf, warum die einseitigen Differenzen gut funktionieren, liefert die Interpretation mit der „effektiven Gleichung“.

Wie analysieren die einseitige Differenz mit implizitem Euler:

Taylor liefert:

$$\frac{\partial u}{\partial t} : \quad \frac{u(x, t + \tau) - u(x, t)}{\tau} = \frac{\partial u}{\partial t} \Big|_{(x, t+\tau)} - \frac{\tau}{2} \frac{\partial^2 u}{\partial t^2} \Big|_{(x, t+\tau)} + O(\tau^2)$$

\downarrow
 ex
 \uparrow
 entw. Punkt

$$\frac{\partial u}{\partial x} : \quad \frac{u(x, t + \tau) - u(x - h, t + \tau)}{h} = \frac{\partial u}{\partial x} \Big|_{(x, t+\tau)} - \frac{h}{2} \frac{\partial^2 u}{\partial x^2} \Big|_{(x, t+\tau)} + O(h^2)$$

Außerdem gilt für genügend glattes u :

$$\frac{\partial u}{\partial t} + a \cdot \frac{\partial u}{\partial x} = 0 \quad \left\{ \begin{array}{l} \Rightarrow \frac{\partial^2 u}{\partial t^2} + a \cdot \frac{\partial^2 u}{\partial x \partial t} = 0 \\ \Rightarrow \frac{\partial^2 u}{\partial t \partial x} + a \cdot \frac{\partial^2 u}{\partial x^2} = 0 \end{array} \right\} \quad \frac{\partial^2 u}{\partial t^2} - a^2 \frac{\partial^2 u}{\partial x^2} = 0$$

also

$$\boxed{\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2}}$$

Für die exakte Lösung eingesetzt in die Differenzengleichung erhalten wir:

$$\begin{aligned} \frac{u(x, t + \tau) - u(x, t)}{\tau} + a \frac{u(x, t + \tau) - u(x - h, t + \tau)}{h} &= \\ &= \left(\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} \right) \Big|_{(x, t+\tau)} - \left(\frac{\tau}{2} \frac{\partial^2 u}{\partial t^2} + \frac{ah}{2} \frac{\partial^2 u}{\partial x^2} \right) \Big|_{(x, t+\tau)} + O(h^2 + \tau^2) \\ &= \left(\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} \right) \Big|_{(x, t+\tau)} - \frac{a^2 \tau + ah}{2} \frac{\partial^2 u}{\partial x^2} \Big|_{(x, t+\tau)} + O(h^2 + \tau^2) \end{aligned}$$

Dies bedeutet:

- Der führende Term des Diskretisierungsfehlers wirkt wie ein Diffusionsterm. Beachte, dass das Vorzeichen stimmt.

- Man kann das diskrete Verfahren auch als die Diskretisierung der Konvektions-Diffusionsgleichung

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} - \frac{a^2 \tau + ah}{2} \frac{\partial^2 u}{\partial x^2} = 0$$

mit zweiter Ordnung (!) auffassen. Der Diffusionskoeffizient ist ortsabhängig.

- Die zentrale Differenz kann durch „künstliches“ Hinzufügen eines Diffusionstermes stabilisiert werden.
- Der Diskretisierungsfehler des impliziten Euler *in der Zeit* kann wegen $\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2}$ als Diffusionsterm im Ort interpretiert werden. Dies erklärt die Stabilisierung der zentralen Differenz für τ genügend groß (!).
- Das upwind-Verfahren verschmiert steile Fronten in der Lösung.
⇒ Dies nennt man das Phänomen der „numerischen“ Diffusion.

12.5 Zusammenfassung

- Hyperbolische Gleichungen erster Ordnung erlauben (im Gegensatz zu elliptischen und parabolischen Gleichung) unstetige Lösungen wenn man sie mit der Methode der Charakteristiken löst. Dies sind natürlich keine klassischen Lösungen.
- Die Linienmethode ist anwendbar, aber in der Kombination mit finiten Differenzen im Ort nicht anwendbar. Bei diskretisierung erster Ordnung erhält man stabile Verfahren mit numerischer Diffusion, ein Verfahren zweiter Ordnung haben wir bis jetzt nicht kennengelernt.

12 Finite Differenzen für lineare hyperbolische Gleichungen

13 Finite-Volumen-Verfahren für lineare, skalare, hyperbolische Gleichungen

Wir halten uns im wesentlichen an [Lev02, Kap. 4].

13.1 Einführung

Im folgenden behandeln wir die Gleichung

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0 \quad \text{in } (0, 1) \times (0, \infty) \quad (13.1)$$

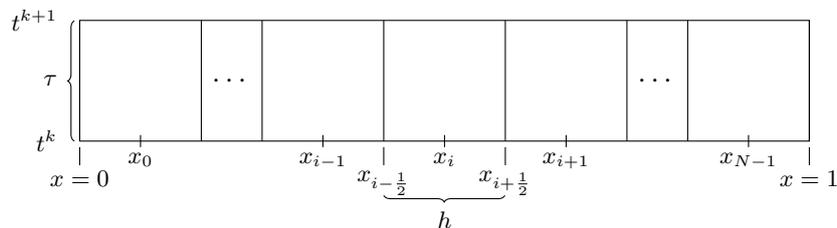
mit geeignetem Rand und Anfangsbedingungen.

Unterschied zu vorher: Einführung der Flussfunktion $f: \mathbb{R} \rightarrow \mathbb{R}$.

Für unser Modellproblem gilt $f(u) = a \cdot u$ mit $\mathbb{R} \rightarrow a > 0$ und mehr wollen wir in diesem Abschnitt auch gar nicht betrachten.

Wie erinnern uns an die Physik hinter Gl. (13.1): Sie beschreibt die Erhaltung einer Größe (z. B. Energie, Masse, Impuls). f beschreibt den Fluss der Erhaltungsgröße an einer Stelle.

Zur Einführung der Finite-Volumen-Verfahren (FV) benötigen wir wieder ein Gitter, diesmal in Raum und Zeit:



also:

$$\begin{aligned} t^k &= k \cdot \tau \\ x_i &= i \cdot h + \frac{h}{2} \\ x_{i \pm \frac{1}{2}} &= i \cdot h + h \left(\frac{1}{2} \pm \frac{1}{2} \right) \end{aligned}$$

Das Intervall $\omega_i = (x_{i-1/2}, x_{i+1/2})$ heißt Zelle oder Kontrollvolumen (engl.: cell, control volume). Äquidistanz ist nicht unbedingt notwendig. Durch Integration von (13.1) über eine Zelle ω_i erhalten

13 Finite-Volumen-Verfahren für lineare, skalare, hyperbolische Gleichungen

wir die integrale Form:

$$\int_{\omega_i} \frac{\partial u}{\partial t} dx + \int_{\omega_i} \frac{\partial f(u)}{\partial x} dx = 0$$

$$\Leftrightarrow \underbrace{\frac{d}{dt} \int_{\omega_i} u(x, t) dx}_{\text{„Masse, Energie in } \omega_i \text{ zur Zeit } t\text{“}} + f(u(x_{i+\frac{1}{2}}, t)) - f(u(x_{i-\frac{1}{2}}, t)) = 0 \quad (13.2)$$

vertauschen

Die (klassische) Lösung von (13.1) erfüllt (13.2) für beliebige Intervalle ω (partielle Integration).

Um ein voll diskretes Verfahren zu erhalten, integrieren wir auch noch in der Zeit über das Intervall (t^k, t^{k+1}) :

$$\underbrace{\frac{1}{h} \int_{\omega_i} u(x, t^{k+1}) dx}_{\text{Zellmittelwert zu Zeit } t^{k+1}} = \underbrace{\frac{1}{h} \int_{\omega_i} u(x, t^k) dx}_{\text{Zellmittelwert zu Zeit } t^k} - \frac{\tau}{h} \left[\underbrace{\frac{1}{\tau} \int_{t^k}^{t^{k+1}} f(u(x_{i+\frac{1}{2}}, t)) dt}_{\text{mittlerer Fluss über Grenze } x_{i+\frac{1}{2}} \text{ im Zeitintervall } (t^k, t^{k+1})} - \underbrace{\frac{1}{\tau} \int_{t^k}^{t^{k+1}} f(u(x_{i-\frac{1}{2}}, t)) dt}_{\text{entsprechend an } x_{i-\frac{1}{2}}} \right] \quad (13.3)$$

Diese Gleichung beschreibt die Evolution der Zellmittelwerte in *exakter* Weise.

Finite-Volumen-Verfahren nutzen die *Zellmittelwerte*.

$$U_i^k = \frac{1}{h} \int_{\omega_i} u(x, t^k) dx + \text{Fehler}$$

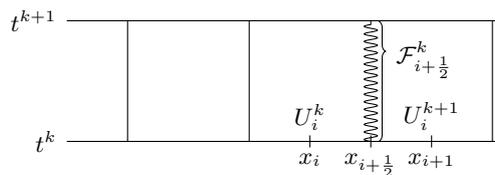
als Unbekannte. Die Approximation besteht darin, dass die mittleren Flüsse, z. B.

$\frac{1}{\tau} \int_{t^k}^{t^{k+1}} f(u(x_{i+\frac{1}{2}}, t)) dt$ nur näherungsweise berechnet werden (können).

Betrachtet man nur explizite Verfahren, so liegt es nahe

$$\frac{1}{\tau} \int_{t^k}^{t^{k+1}} f(u(x_{i+\frac{1}{2}}, t)) dt = \underbrace{\mathcal{F}(U_i^k, U_{i+1}^k)}_{=F_{i+\frac{1}{2}}^k} + \text{Fehler} \quad (13.4)$$

zu nutzen, also:



\mathcal{F} wird als numerische Flussfunktion bezeichnet.

Das voll diskrete Verfahren erhält man wie immer durch Ignorieren der Fehlerterme. (13.3) und (13.4) zusammen geben:

$$U_i^{k+1} = U_i^k - \frac{\tau}{h} (\mathcal{F}(U_i^k, U_{i+1}^k) - \mathcal{F}(U_i^k, U_{i-1}^k)) \quad (13.5)$$

Man sieht: Wie im expliziten Finite-Differenzen-Verfahren hängt U_i^{k+1} nur von den drei Werten $U_{i-1}^k, U_i^k, U_{i+1}^k$ ab.

Beispiel Umstellen von (13.5) und die Wahl $\mathcal{F}(Q, Q') = a \cdot Q$ (bei $a > 0$) liefert:

$$\frac{U_i^{k+1} - U_i^k}{\tau} + a \cdot \frac{U_i^k - U_{i-1}^k}{h} = 0,$$

was nichts anderes als das explizite upwind-Verfahren ist. In diesem Fall sind also FD- und FV-Verfahren äquivalent.

Finite-Volumen-Verfahren sind global konservativ:

Gesamt -masse
-energie zur Zeit $t^{k+1} =$

$$= \sum_{i=0}^{N-1} h \cdot U_i^{k+1} = \sum_{i=0}^{N-1} h \left(U_i^k - \frac{\tau}{h} (\mathcal{F}(U_i^k, U_{i+1}^k) - \mathcal{F}(U_{i-1}^k, U_i^k)) \right)$$

über alle Zellen \nearrow Zellmittelwert!
 Masse, Energie in Zelle i

$$= \sum_{i=0}^{N-1} h U_i^k \left(-\tau \mathcal{F}(U_N^k, U_{N-1}^k) - \mathcal{F}(U_{-1}^k, U_0^k) \right)$$

Masse zur Zeit t^k $\underbrace{\hspace{10em}}$ Dies sind spezielle Flüsse, die über die Randbedingungen definiert sind!
 Alle internen Flüsse heben sich weg.

FV-Verfahren geben die Erhaltungsgröße im Gesamtgebiet *exakt* wieder. Dies ist bei FD-Verfahren im allgemeinen (d. h. bei nichtäquidistanten Gittern, variablen Koeffizienten, Nichtlinearitäten) *nicht* der Fall.

Diffusionsgleichung Wir betrachten nochmal kurz die Wärmeleitungsgleichung:

$$\frac{\partial u}{\partial t} - \frac{\partial}{\partial x} \left(\beta \frac{\partial u}{\partial x} \right) = 0 \iff \frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0 \text{ mit } f(u) = -\beta \frac{\partial u}{\partial x}$$

13 Finite-Volumen-Verfahren für lineare, skalare, hyperbolische Gleichungen

Bei der Wahl der Flussfunktion

$$\mathcal{F}(U_i^k, U_{i+1}^k) = -\beta \frac{U_{i+1}^k - U_i^k}{h} \quad \left(\approx \frac{1}{\tau} \int_{t^k}^{t^{k+1}} -\beta \frac{\partial u}{\partial x} \Big|_{x_{i+\frac{1}{2}}} dt \right)$$

ergibt sich nach Einsetzen in (13.5)

$$\begin{aligned} U_{i+1}^{k+1} &= U_i^k - \frac{\tau}{h} \left(-\beta \frac{U_{i+1}^k - U_i^k}{h} + \beta \frac{U_i^k - U_{i-1}^k}{h} \right) \\ &= U_i^k + \tau \frac{\beta}{h^2} (U_{i-1}^k - 2U_i^k + U_{i+1}^k) \end{aligned} \quad (13.6)$$

also das explizite FD Differenzen Verfahren.

Implizite Verfahren sind auch möglich durch $\mathcal{F}(U_i^k, U_{i+1}^k, U_i^{k+1}, U_{i+1}^{k+1})$.

13.2 Anforderungen an die Flussfunktion

Analyse der FD-Verfahren führte auf Konsistenz (lokaler Abschneidefehler, lokale Approximation) und Stabilität (Fehlerfortpflanzung). Dies ist bei FV-Verfahren genauso.

Um Konsistenz sicherzustellen benötigt man zwei Bedingungen an die Flussfunktion:

1. $\mathcal{F}(Q, Q) = f(Q)$ Ist u konstant in x und t , dann sollte die Flussauswertung für jedes Q ! t , dann sollte die Flussauswertung exakt sein.

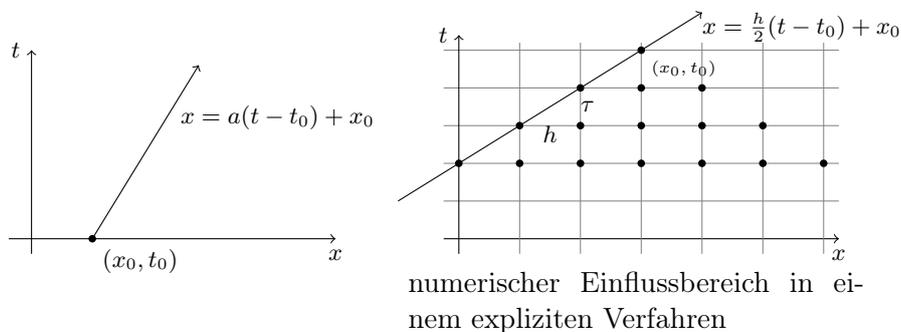
2. Stetigkeit der Flussfunktion:

$$|\mathcal{F}(Q_i, Q_{i+1}) - f(\bar{Q})| \leq L \max(|Q_i - \bar{Q}|, |Q_{i+1} - \bar{Q}|).$$

Konvergieren $Q_i, Q_{i+1} \rightarrow \bar{Q}$, sp soll auch der numerische Fluss gegen den richtigen Wert konvergieren.

Für die Stabilität von expliziten Verfahren ist die CFL-Bedingung eine notwendige (aber nicht hinreichende, wie das uneingeschränkt instabile Verfahren zeigt) Voraussetzung.

In hyperbolischen Gleichungen breitet sich Information mit endlicher Geschwindigkeit aus. Dies zeigt direkt das Verfahren der Charakteristiken.



Forderung: Charakteristik muss in numerischem Einflussbereich enthalten sein, d. h.

$$|a| \leq \frac{h}{\tau} \iff \boxed{\left| \frac{a\tau}{h} \right| \leq 1}$$

$\nu = \left| \frac{a \cdot \tau}{h} \right|$ heißt *Courantzahl*.

13.3 Ein instabiler Fluss

Zur Erinnerung:

$$\mathcal{F}(U_i^k, U_{i+1}^k) \approx \frac{1}{\tau} \int_{t^k}^{t^{k+1}} f(u(x_{i+\frac{1}{2}}, t)) dt$$

Also wäre (Trapezregel)

$$\mathcal{F}(U_i^k, U_{i+1}^k) = \frac{1}{2} [f(U_i^k) + f(U_{i+1}^k)] \tag{13.7}$$

eine naheliegende Wahl.

\mathcal{F} ist konsistent (erfüllt 1, 2 von oben).

Für $f(u) = au$ erhalten wir das Verfahren

$$\begin{aligned} U_i^{k+1} &= U_i^k - \frac{\tau}{h} \left(\frac{1}{2} [aU_i^k + aU_{i+1}^k] - \frac{1}{2} [aU_{i-1}^k + aU_i^k] \right) \\ &= U_i^k - a\tau \underbrace{\frac{1}{2h} (U_{i+1}^k - U_{i-1}^k)}_{\text{zentrale Differenz}} \end{aligned}$$

Dieses Verfahren haben wir in (12.2) als uneingeschränkt instabil erkannt.

13.4 Lax-Friedrich-Verfahren

Das Verfahren ist definiert durch die Flussfunktion

$$\mathcal{F}_{LF}(U_i^k, U_{i+1}^k) = \underbrace{\frac{1}{2} [f(U_i^k) + f(U_{i+1}^k)]}_{\text{wie oben}} - \underbrace{\frac{h}{2\tau} (U_{i+1}^k - U_i^k)}_{\substack{= -\beta \frac{U_{i+1}^k - U_i^k}{h} \text{ mit } \beta = \frac{h^2}{2\tau} \\ \text{Diffusionsterm}}}$$

Hier wird zu dem instabilen Fluss ein stabilisierender Fluss aus einem Diffusionsterm addiert. Man löst also eigentlich die Gleichung (mit einem Verfahren zweiter Ordnung)

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} - \beta \frac{\partial^2 u}{\partial x^2} = 0.$$

Wegen $\beta = \frac{h^2}{2\tau}$ gilt $\beta \rightarrow 0$ für $h \rightarrow 0$ bei $\frac{h}{\tau}$ fest (Courantbedingung). Aber: Die Methode hat dadurch einen Konsistenzfehler $O(h)$ statt $O(h^2)$.

Bemerkung: Das Verfahren ist asymptotisch stabil für $h \rightarrow 0$.

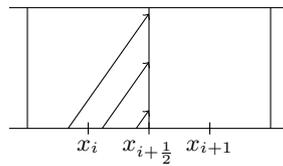
Ab welchem h z. B. ein Maximumprinzip erfüllt ist, hängt aber von a ab!

Das Lax-Friedrich-Verfahren addiert mehr Diffusion als eigentlich nötig.

13.5 Upwind-Verfahren

Idee: Nutze Wissen über Charakteristiken und Informationsausbreitung in der numerischen Flussfunktion.

Sei $a > 0$. Die Form der Charakteristik



legt nahe, dass $\mathcal{F}(U_i^k, U_{i+1}^k)$ nur von U_i abhängen sollte. Wir setzen also:

$$\mathcal{F}(U_i^k, U_{i+1}^k) = f(U_i^k) \stackrel{\text{in unserem Modellproblem}}{=} a \cdot U_i^k.$$

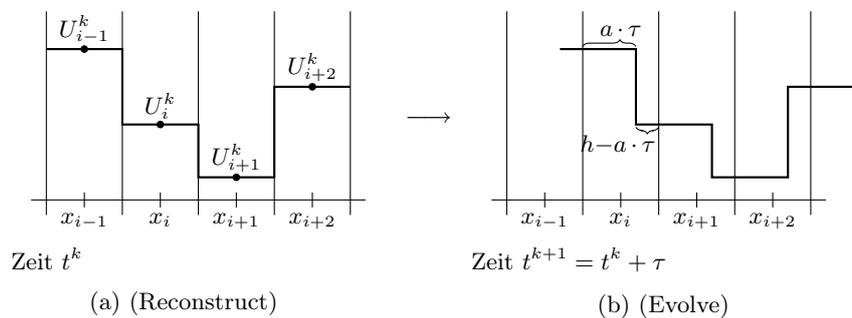
Wir erhalten für diese Flussfunktion das Verfahren:

$$U_i^{k+1} = U_i^k - \frac{\tau}{h} a \left(U_i^k - U_{i-1}^k \right). \quad (13.8)$$

Das FV-Verfahren erlaubt eine weitere Interpretation.

Die Werte U_i^k stellen *Zellmittelwerte* dar.

Wir können uns diese auch als stückweise konstante Funktion vorstellen (Bild a):



Nach der Methode der Charakteristik wird diese Funktion in dem Zeitintervall τ um $a \cdot \tau$ nach rechts bewegt. Courant $\frac{a \cdot \tau}{h} \leq 1 \iff a \cdot \tau < h$ bedeutet, dass maximal eine Bewegung um eine Gitterzelle erlaubt ist (Bild b).

Die Zellmittelwerte zur Zeit t^{k+1} ergeben sich nun als Mittelwerte über diese unstetige Funktion in jeder Zelle:

$$\begin{aligned}
 U_i^{k+1} &= \frac{a \cdot \tau}{h} U_{i-1}^k + \frac{h - a \cdot \tau}{h} U_i^k = \frac{a \cdot \tau}{h} U_{i-1}^k + \left(1 - \frac{a \cdot \tau}{h}\right) U_i^k \\
 &\quad \uparrow \qquad \qquad \qquad \uparrow \\
 &\quad \text{Konvexkombination, da } \frac{a \cdot \tau}{h} \leq 1 \\
 &\quad \text{!} \Rightarrow \text{Maximumprinzip.} \\
 &= U_i^k - \frac{\tau}{h} a \left(U_i^k - U_{i-1}^k \right)
 \end{aligned}$$

Das ist identisch zu (13.8)!

Für $a < 0$ erhält man ein entsprechendes Verfahren.

Für ein beliebiges a setzt man

$$\begin{aligned}
 \mathcal{F}(U_i^k, U_{i+1}^k) &= \max(a, 0) \cdot U_i^k + \min(a, 0) \cdot U_{i+1}^k \\
 &= \begin{cases} aU_i^k & a \geq 0 \\ aU_{i-1}^k & a < 0 \end{cases}
 \end{aligned}$$

13.6 Godunov-Verfahren

Das oben beschriebene Verfahren lässt sich verallgemeinern zum sogenannten REA (*Reconstruct, Evolve, Average*) Verfahren:

1. Rekonstruiere eine *stückweise polynomiale* Funktion aus den Zellmittelwerten. Im einfachsten Fall stückweise konstant.
2. Löse die hyperbolische Gleichung mit diesem Anfangswert exakt, um eine Lösung zum Zeitpunkt $t + \tau$ zu erhalten.
3. Berechne aus dieser Lösung neue Zellmittelwerte.

Dieses Verfahren lässt sich vor allem auf kompliziertere Gleichungen verallgemeinern und wurde von Godunov 1957 erstmalig für die (nichtlinearen) Euler-Gleichungen der Gasdynamik vorgeschlagen. Es ist auch der Ausgangspunkt für Methoden höherer Ordnung, die das Phänomen der numerischen Diffusion vermindern.

13.7 Zusammenfassung

- Finite-Volumen-Verfahren basieren auf einer Integration der partiellen Differentialgleichung über Kontrollvolumen und entsprechende Approximation der Flüsse über Kontrollvolumengrenzen.
- Beim Lax-Friedrich-Verfahren stabilisiert man den instabilen Fluss der sich aus dem Mittelwert ergibt durch einen künstlichen Diffusionsterm.

13 Finite-Volumen-Verfahren für lineare, skalare, hyperbolische Gleichungen

- Das Godunov-Verfahren nutzt eine einseitige Auswertung und entspricht den Upwind-Verfahren bei Finiten Differenzen.

14 High resolution Schemata für lineare, skalare, hyperbolische Probleme

Wir betrachten wieder

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = 0 \quad \text{in } \Omega \times T \quad (14.1)$$

Da wir Randbedingungen erst einmal gar nicht diskutieren, lassen wir sie gleich weg. Der Parameter a ist in \mathbb{R} , falls $a > 0$ wird dies explizit erwähnt.

14.1 Verfahren zweiter Ordnung

Die bisher „erfolgreichen“ Verfahren (explizit, implizit upwind, Lax-Friedrich) waren alle nur erster Ordnung.

Lax-Wendroff: ist ein erstes Verfahren zweiter Ordnung. Dieses leiten wir zunächst als Finite-Differenzen-Methode her und schreiben es dann in ein Finite-Volumen-Verfahren um.

Herleitung über Taylor:

Aus

$$\left. \frac{\partial u}{\partial t} \right|_{(x,t)} = \frac{u(x, t + \tau) - u(x, t)}{\tau} - \frac{\tau}{2} \left. \frac{\partial^2 u}{\partial t^2} \right|_{(x,t)} + O(\tau^2)$$

und

$$\left. \frac{\partial u}{\partial x} \right|_{(x,t)} = \frac{u(x + h, t) - u(x - h, t)}{2h} + O(h^2)$$

folgt

$$\begin{aligned} \left. \frac{\partial u}{\partial t} \right|_{(x,t)} + a \left. \frac{\partial u}{\partial x} \right|_{(x,t)} &= \frac{u(x, t + \tau) - u(x, t)}{\tau} + \\ &+ a \frac{u(x + h, t) - u(x - h, t)}{2h} - \frac{\tau}{2} \left. \frac{\partial^2 u}{\partial t^2} \right|_{(x,t)} + O(a h^2 + \tau^2). \end{aligned} \quad (14.2)$$

Nun nutze

$$\left. \frac{\partial^2 u}{\partial t^2} \right|_{(x,t)} = a^2 \left. \frac{\partial^2 u}{\partial x^2} \right|_{(x,t)} = a^2 \frac{u(x + h, t) - 2u(x, t) + u(x - h, t)}{h^2} + O(h^2)$$

(für genügend glattes u).

Einsetzen dieser Approximation für $\frac{\partial^2 u}{\partial t^2}$ liefert

$$\begin{aligned} \frac{u(x, t + \tau) - u(x, t)}{\tau} + a \frac{u(x + h, t) - u(x - h, t)}{2h} + \\ + \frac{\tau a^2}{2} \frac{u(x + h, t) - 2u(x, t) + u(x - h, t)}{h^2} &= \left. \frac{\partial u}{\partial t} \right|_{(x,t)} + a \left. \frac{\partial u}{\partial x} \right|_{(x,t)} + O(h^2 + \tau^2) \end{aligned}$$

14 High resolution Schemata für lineare, skalare, hyperbolische Probleme

und damit nach Umstellen das Verfahren von Lax-Wendroff:

$$U_i^{k+1} = U_i^k - \frac{a\tau}{2h} (U_{i+1}^k - U_{i-1}^k) + \underbrace{\frac{a^2\tau^2}{2h^2} (U_{i+1}^k - 2U_i^k + U_{i-1}^k)}_{\substack{\text{stabilisierender Diffusionsterm} \\ \text{der exakt richtigen Größe!}}} \quad (14.3)$$

Durch Koeffizientenvergleich kann man das (für äquidistante Gitter) in ein FV-Verfahren umschreiben:

$$U_i^{k+1} = U_i^k - \frac{\tau}{h} [\mathcal{F}(U_i, U_{i+1}) - \mathcal{F}(U_{i-1}, U_i)]$$

mit

$$\mathcal{F}(U_i^k, U_{i+1}^k) = \frac{a}{2} (U_i^k + U_{i+1}^k) - \underbrace{\frac{1}{2} \frac{\tau}{h} a^2 (U_{i+1}^k - U_i^k)}_{\substack{\text{„diffusiver Fluss“ mit} \\ \beta = \frac{1}{2} \tau a^2}} \quad (14.4)$$

Lax-Wendroff ist eine zentrale 3-Punkt-Formel. Mittels der einseitigen Differenzen

$$\begin{aligned} \left. \frac{\partial u}{\partial x} \right|_{(x,t)} &= \frac{3u(x,t) - 4u(x-h,t) + u(x-2h,t)}{2h} + O(h^2) \\ \left. \frac{\partial^2 u}{\partial x^2} \right|_{(x,t)} &= \frac{u(x,t) - 2u(x-h,t) + u(x-2h,t)}{h^2} + O(h) \end{aligned}$$

\uparrow
 das genügt, damit
 in (14.2) der Fehler
 $O(\tau h)$ wird und $\tau \sim h$

erhält man das **Beam-Warming-Verfahren**:

$$U_i^{k+1} = U_i^k - \frac{a\tau}{2h} (3U_i^k - 4U_{i-1}^k + U_{i-2}^k) + \frac{a^2\tau^2}{2h^2} (U_i^k - 2U_{i-1}^k + U_{i-2}^k) \quad (14.5)$$

Die entsprechende Flussfunktion für das FV-Verfahren lautet:

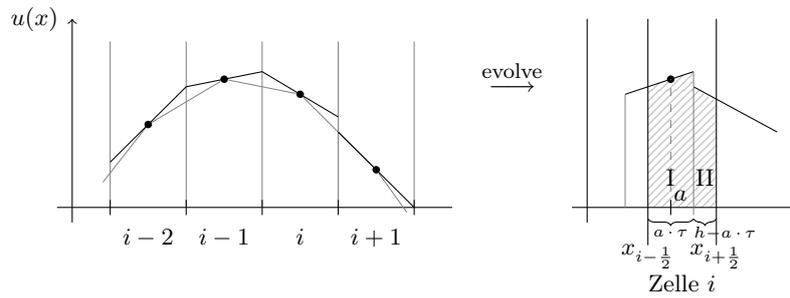
$$\mathcal{F}_{i+\frac{1}{2}}(\underbrace{U_{i-1}^k, U_i^k, U_{i+1}^k}_{\text{zus. Argument}}) = aU_i^k + \frac{a}{2} \left(1 - \frac{\tau}{h} a\right) (U_i^k - U_{i-1}^k). \quad (14.6)$$

Der Lax-Wendroff-Fluss ist laut (14.4) der instabile zentrale Fluss plus ein stabilisierender Diffusionsterm.

Es geht aber auch anders:

$$\mathcal{F}(U_i^k, U_{i+1}^k) = \underbrace{\max(a, 0)}_{a^+} \cdot U_i^k + \underbrace{\min(a, 0)}_{a^-} \cdot U_{i+1}^k = a^+ U_i^k + a^- U_{i+1}^k$$

14 High resolution Schemata für lineare, skalare, hyperbolische Probleme



In Zelle i : Setze

$$\tilde{u}_i^k(x) = U_i^k + \sigma_i^k(x - x_i)$$

Beachte:

$$\frac{1}{h} \cdot \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} U_i^k + \sigma_i^k(x - x_i) dx = U_i^k$$

also: Steigung σ_i^k beeinflusst den Mittelwert nicht \Rightarrow Verfahren ist *konservativ*.

Die Wahl der Steigungen σ_i diskutieren wir unten. Zunächst tun wir so als wüssten wir σ_i .

Evolve und Averaging liefert für $t + \tau$ unter Courant $\frac{|a|\tau}{h} \leq 1$ und $a > 0$:

$$\begin{aligned} \underbrace{h \cdot U_i^{k+1}}_{\text{Fläche}} &= a \cdot \tau \cdot \tilde{u}_{i-1}^k \left(\underbrace{\left(x_{i-\frac{1}{2}} + \frac{a \cdot \tau}{2} \right) - a \cdot \tau}_{\substack{\text{Auswertungspunkt } a \\ \text{evolution} \\ \text{des Profils}}} \right) + (h - a \cdot \tau) \tilde{u}_i^k \left(\underbrace{\left(x_{i+\frac{1}{2}} - \frac{h - a \cdot \tau}{2} \right) - a \cdot \tau}_{\substack{x_i + \frac{h}{2} - \frac{h}{2} - \frac{a \cdot \tau}{2} \\ = x_i - \frac{a \cdot \tau}{2}}} \right) \\ &= x_{i-1} + \frac{h}{2} - \frac{a \cdot \tau}{2} = x_{i-1} + \frac{1}{2}(h - a \cdot \tau) \\ &= a \cdot \tau \cdot \left(U_{i-1}^k + \sigma_{i-1}^k \left(x_{i-1} + \frac{1}{2}(h - a \cdot \tau) - x_{i-1} \right) \right) \\ &\quad + (h - a \cdot \tau) \left(U_i^k - \sigma_i^k \left(x_i - \frac{a \cdot \tau}{2} - x_i \right) \right) \\ &= a \cdot \tau U_{i-1}^k + (h - a \cdot \tau) U_i^k + \frac{a \cdot \tau}{2} (h - a \cdot \tau) \sigma_{i-1}^k - (h - a \cdot \tau) \frac{a \cdot \tau}{2} \sigma_i^k \end{aligned}$$

teilen durch h

$$\Leftrightarrow \boxed{U_i^{k+1} = U_i^k - a \frac{\tau}{h} (U_i^k - U_{i-1}^k) - \frac{a \cdot \tau}{2} \left(1 - \frac{a \cdot \tau}{h} \right) (\sigma_i^k - \sigma_{i-1}^k)} \quad (14.8)$$

upwind + correction abhängig von Steigungen

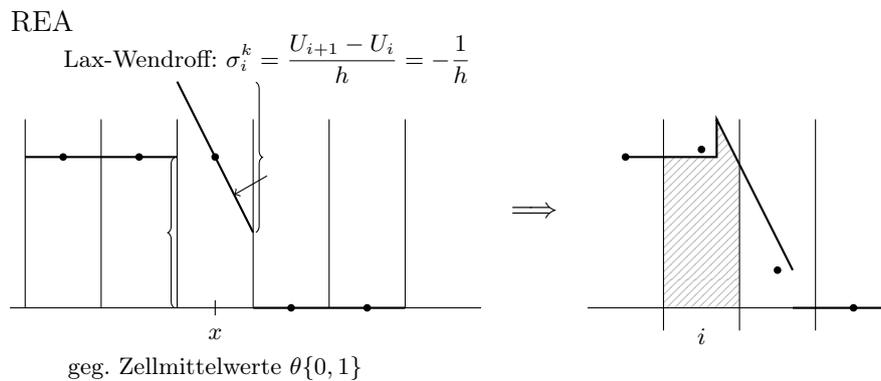


Wie wählt man die σ_i ? Drei offensichtliche Möglichkeiten sind

$$\begin{aligned}
 \text{zentral: } \sigma_i^k &= \frac{U_{i+1}^k - U_{i-1}^k}{2h} && \text{(Fromm)} \\
 \text{upwind: } \sigma_i^k &= \frac{U_i^k - U_{i-1}^k}{h} && \text{(Beam-Warming!) } \sigma_{i-1} \text{ braucht } U_{i-2}! \\
 \text{downwind: } \sigma_i^k &= \frac{U_{i+1}^k - U_i^k}{h} && \text{(Lax-Wendroff)}
 \end{aligned}
 \tag{14.9}$$

(Übung: Durchrechnen und Koeffizientenvergleich).

Damit kann man erklären, wie die Oszillationen im Lax-Wendroff-Verfahren entstehen:



Ein Negativresultat:

Satz 14.1 (Godunov, 1959). Alle monotoneerhaltenden, *linearen* Verfahren sind höchstens von erster Ordnung genau.

Siehe [Lev02] □

Monotoneerhaltend = führt keine neuen Minima/Maxima ein.

Godunov sagt: Es gibt kein lineares Verfahren (d. h. $U^{k+1} = M_{h,\tau} U^k$), das zweiter Ordnung und monotoneerhaltend ist.

14.3 Slope Limiter Verfahren

Wie umgeht man Godunov? Durch *nichtlineare* Verfahren! (trotz linearen Problems)

Idee: Wähle σ_i^k *lösungsabhängig*.

Da man die σ_i^k begrenzen muss, spricht man von *Slope Limitern*.

Ein Weg, die Oszillationen zu messen, ist die *totale Variation* TV:

$$TV(U^k) := \sum_{i=-\infty}^{\infty} |U_i^k - U_{i-1}^k|$$

hier: unendliches Gebiet.

im unendlichen Fall: Reihe konvergiert, d. h. $U_i \rightarrow const$ für $i \rightarrow \pm 1$ notwendig.

Definition 14.2. Ein Verfahren heißt total variation non increasing (TVNI), falls in jedem Schritt gilt

$$TV(U^{k+1}) \leq TV(U^k).$$

□

Satz 14.3 (Harten 1983). Ein TVNI-Schema erzeugt keine neuen Extrema in der Lösung. Oder: War U^k monoton, so ist auch U^{k+1} monoton.

Beweis: gegeben U^k , nehme an $U_i^k \leq U_{i+1}^k$ (geht auch in andere Richtung). Offensichtlich:

$$TV(U^k) = \sum_{i=-\infty}^{\infty} \underbrace{|U_i - U_{i-1}|}_{\geq 0} = \sum_{i=-\infty}^{\infty} U_i - U_{i-1} = U_{\infty}^k - U_{-\infty}^k$$

↑
Teleskop

Habe nun U^{k+1} in U_j^k ein lokales Minimum, so gilt:

$$\begin{aligned} TV(U^{k+1}) &= \underbrace{\sum_{i=-\infty}^j |U_i^{k+1} - U_{i-1}^{k+1}|}_{<0} + \underbrace{|U_{j+1} - U_j|}_{<0} + \sum_{i=j+2}^{\infty} |U_i - U_{i-1}| \\ &= U_j^{k+1} - U_{-\infty}^{k+1} + U_j^{k+1} - U_{j+1}^{k+1} + U_{\infty}^{k+1} - U_{j+1}^{k+1} \\ &= \underbrace{U_{\infty}^{k+1} - U_{-\infty}^{k+1}}_{= TV(U^k)} + 2 \underbrace{(U_j^{k+1} - U_{j+1}^{k+1})}_{>0 \text{ nach Ann.}} \leq TV(U^k) \quad \not\Leftarrow \quad \square \end{aligned}$$

die können sich in einem Schritt nicht ändern! Courant!

Es ist also sinnvoll nach Verfahren zu suchen, die die totale Variation nicht erhöhen.

Im REA-Verfahren wird die totale Variation nur durch die Rekonstruktion bestimmt. Evolve und Average vergrößern die TV nicht (ohne Beweis).

Eine mögliche Wahl für die Steigung ist:

$$\sigma_i^k = \minmod \left(\underbrace{\frac{U_{i+1}^k - U_i^k}{h}}_{\substack{\text{downwind} \\ \text{slope} \\ \text{(LW-d)}}, \underbrace{\frac{U_i^k - U_{i-1}^k}{h}}_{\substack{\text{upwind} \\ \text{slope} \\ \text{(BW)}}} \right)$$

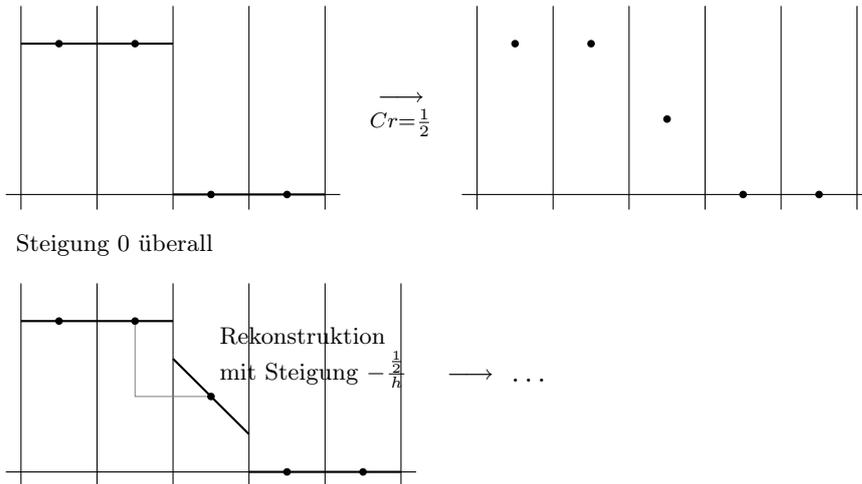
mit

$$\minmod(a, b) = \begin{cases} a & \text{falls } |a| < |b| \text{ und } a \cdot b > 0 \\ b & \text{falls } |b| < |a| \text{ und } a \cdot b > 0 \\ 0 & \text{falls } a \cdot b < 0 \text{ (d. h. verschiedene Vorzeichen)} \end{cases}$$

Idee: Nehme die kleinere Steigung (Variation klein halten) bzw. 0, falls ein lokales Extremum vorliegt.

Was passiert an einer Diskontinuität?

Annahme: $Cr = \frac{1}{2}$



Beobachtung: Die rekonstruierte Steigung könnte eigentlich Faktor 2 größer sein, ohne Monotonie zu verletzen.

Tatsächlich erfüllt auch noch folgender Limiter mit Namen „Superbee“ die TVNI-Eigenschaft:

$$\sigma_i^k = \text{maxmod} \left(\sigma_i^{(1)}, \sigma_i^{(2)} \right), \quad \text{maxmod}(a, b) = \begin{cases} a & \text{falls } |a| \geq |b| \\ b & \text{falls } |b| \geq |a| \end{cases}$$

mit

$$\sigma_i^{(1)} = \text{minmod} \left(\frac{U_{i+1}^k - U_i^k}{h}, 2 \frac{U_i^k - U_{i-1}^k}{h} \right)$$

$$\sigma_i^{(2)} = \text{minmod} \left(2 \frac{U_{i+1}^k - U_i^k}{h}, \frac{U_i^k - U_{i-1}^k}{h} \right)$$

Bemerkung: bei verschiedenen Vorzeichen gilt $\sigma_i^{(1)} = \sigma_i^{(2)} = 0 \Rightarrow \sigma_i^k = 0$

Beispiel:

$$\sigma_i^{(1)} = \text{minmod}(0.8, 2 \cdot 0.2) = 0.4$$

$$\sigma_i^{(2)} = \text{minmod}(2 \cdot 0.8, 0.2) = 0.2$$

$$\sigma_i = \text{maxmod}(0.2, 0.4) = 0.4$$

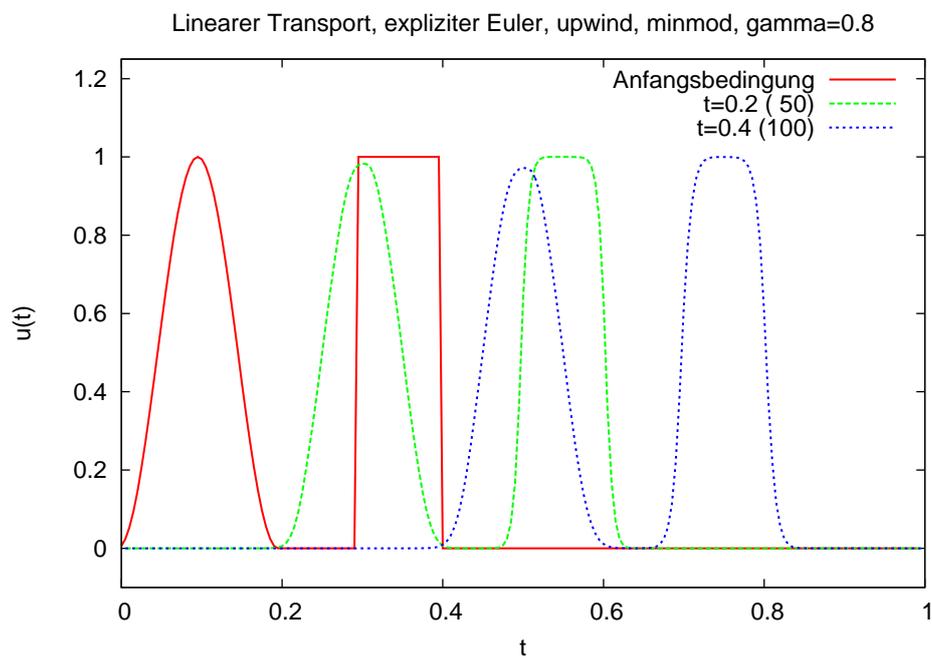
Sind die Steigungen sehr verschieden, wird zweimal die kleinere geliefert (sind beide Steigungen $\frac{1}{2}$, bleibt es aber bei $\frac{1}{2}$).

14 High resolution Schemata für lineare, skalare, hyperbolische Probleme

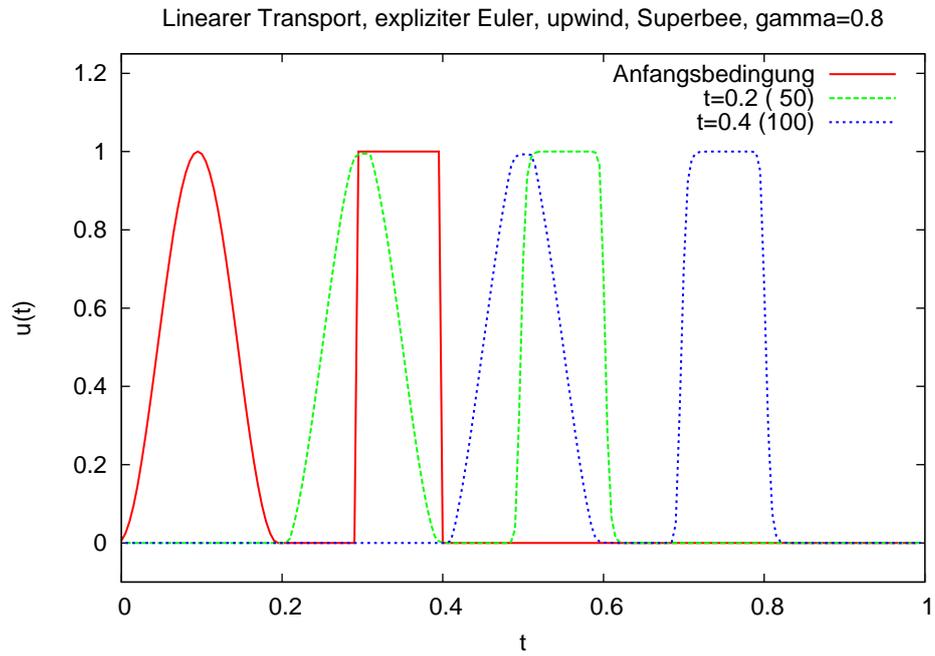
- Es gibt viele verschiedene Limiter
- Die Kriterien an eine TVNI-Limiterfunktion sind genau bekannt werden in diesem Skript aber nicht behandelt, siehe etwa [Lev02] und die Literatur dort.

14.4 Numerischer Vergleich

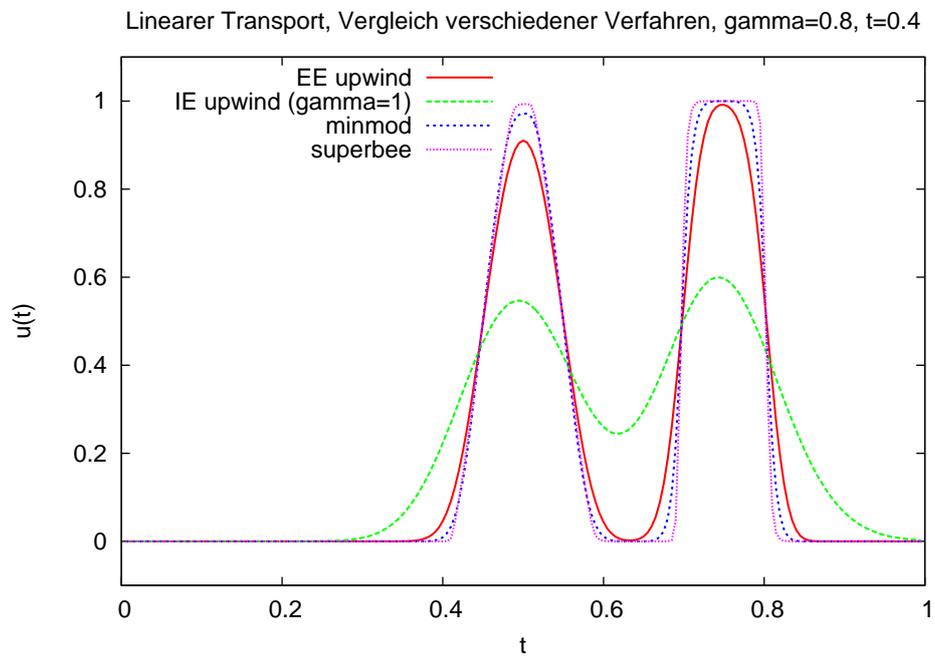
Wieder das Modellproblem, $a = 1$, $h = 1/200$.



Minmod bei $\gamma = 0.8$.



Superbee bei $\gamma = 0.8$.



Vergleich verschiedener Verfahren bei $\gamma = 0.8$.

14.5 Zusammenfassung

- Mit Lax-Wendroff haben wir ein erstes Verfahren zweiter Ordnung für das linear hyperbolische Modellproblem kennengelernt. Allerdings ergibt dieses Verfahren nichtphysikalische Lösungen bei unstetigen (oder sehr steilen) Anfangsbedingungen.
- Der Satz von Godunov zeigt, dass alle monotonen linearen Verfahren höchstens erste Ordnung genau sind.
- Dies führt zu den Slope-Limiter-Verfahren, die lösungsabhängig zwischen erster und zweiter Ordnung umschalten (und somit aufgrund der Nichtlinearität den Satz von Godunov umgehen).

15 Nichtlineare Erhaltungsgleichungen

Wir betrachten die *nichtlineare* hyperbolische Gleichung

$$\begin{aligned} \frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} &= 0 \quad \text{in } \mathbb{R} \times \mathbb{R}^+ \\ u(x, 0) &= \quad x \in \mathbb{R} \end{aligned} \quad (15.1)$$

Ein Beispiel für die Nichtlinearität ist

$$f(u) = \frac{u^2}{2}.$$

Anwendung: Verkehrssimulation $\left(u : \text{Dichte} : \frac{\text{Autos}}{\text{Autolänge}} \in [0, 1]\right)$
 Mehrphasenströmung.

15.1 Schwache Lösungen

(15.1) macht keinen Sinn für unstetige Lösungen (sondern nur für genügend glatte).

Durch Multiplikation mit einer Testfunktion Φ und partielle Integration erhält man:

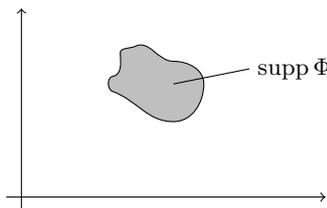
u ist eine schwache Lösung von (15.1), falls

$$\int_0^\infty \int_{-\infty}^\infty \left[u(x, t) \cdot \frac{\partial \Phi}{\partial t}(x, t) + f(u(x, t)) \cdot \frac{\partial \Phi}{\partial x}(x, t) \right] dx dt + \int_{-\infty}^\infty u_0(x, t) \Phi(x, 0) dx = 0 \quad (15.2)$$

für alle $\Phi \in C_0^1(\mathbb{R} \times \mathbb{R}^+) =$

$$= \{ \Phi \in C^1(\mathbb{R} \times \mathbb{R}^+) \mid \exists r > 0 \text{ s. t. } \text{supp } \Phi \subset B_r(0) \cap (\mathbb{R} \times \mathbb{R}^+) \}$$

Man zeigt: u glatte Lösung von (15.1) $\Rightarrow u$ ist auch Lösung von (15.2).
 Die Umkehrung gilt nicht unbedingt.



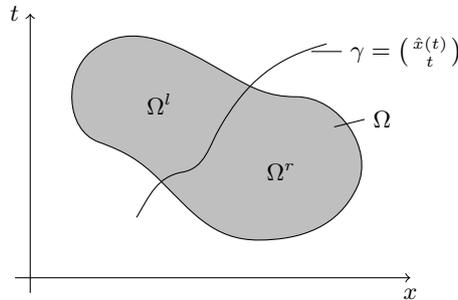
Rankine-Hugoniot-Bedingung

ist eine notwendige Bedingung, die eine schwache Lösung an einer Sprungstelle erfüllen muss.

Kontraktion:

15 Nichtlineare Erhaltungsgleichungen

- $\Omega \subset \mathbb{R} \times \mathbb{R}^+$ offenes Teilgebiet.
- $\gamma: (\hat{x}(t), t)$ eine Kurve, die Ω in zwei Teile Ω^l, Ω^r zerlegt.



Für einen Punkt $(x, t) \in \gamma$ definiert man den Sprung

$$[u](x, t) = \lim_{(x', t') \rightarrow (x, t) \text{ in } \Omega^l} u(x', t') - \lim_{(x', t') \rightarrow (x, t) \text{ in } \Omega^r} u(x', t')$$

Satz 15.1. Sei u eine schwache Lösung von (15.1) im Sinne von (15.2) mit den folgenden zusätzlichen Bedingungen:

1. u ist eine klassische Lösung in Ω^l und in Ω^r
2. u ist unstetig entlang der Kurve γ , d. h. $[u](\hat{x}(t), t) \neq 0 \forall (x, t) \in \gamma$
3. Der Sprung ist stetig entlang γ , d. h. $[u](\hat{x}(t), t)$ ist eine stetige Funktion in t .

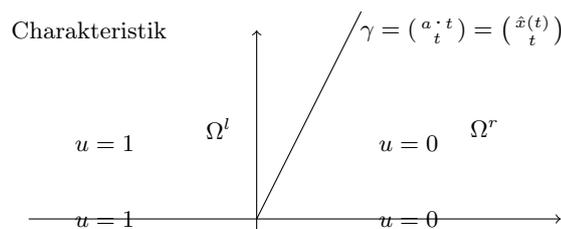
Dann gilt

$$\frac{d\hat{x}}{dt}(t) \cdot [u](\hat{x}(t), t) = [f(u)](\hat{x}(t), t) \quad \forall (\hat{x}(t), t) \tag{15.3}$$

Zur Illustration: Wir erinnern uns an die Methode der Charakteristiken

linearer Fall:

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = 0, \quad u_0(x) = \begin{cases} 1 & x \leq 0 \\ 0 & \text{sonst} \end{cases}$$



- $x'(t) = a$ ist die Geschwindigkeit, mit der sich die Unstetigkeit bewegt!
- Andererseits gilt:

$$\frac{[f(u)](\hat{x}(t), t)}{[u](\hat{x}(t), t)} = \frac{a \cdot 1 - a \cdot 0}{1 - 0} = a$$

$f(u) = a \cdot u$
↙

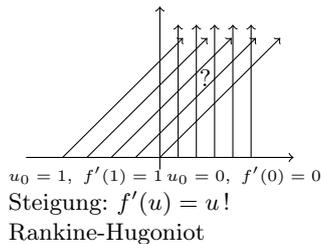
nichtlinearer Fall:

$$f(u) = \frac{u^2}{2}; \quad \frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0; \quad u_0(x) = \begin{cases} 1 & x \leq 0 \\ 0 & \text{sonst} \end{cases}$$

nichtkonservative Form $f'(u) = u$

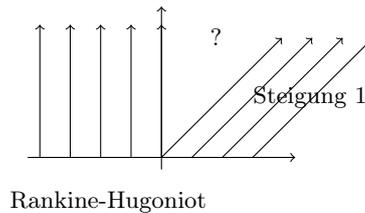
$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = \frac{\partial u}{\partial t} + f'(u(x, t)) \frac{\partial u}{\partial x} = \frac{\partial u}{\partial t} + u(x, t) \cdot \frac{\partial u}{\partial x} = 0$$

Methode der Charakteristik: (Man stelle sich einen ε -Übergang vor!)

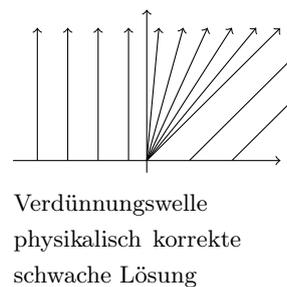
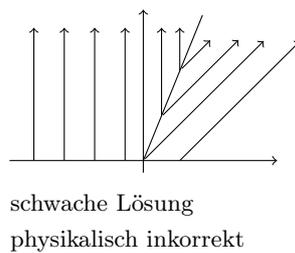
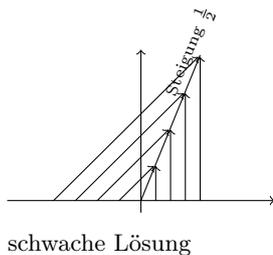


$$\hat{x}'(t) = \frac{[f(u)]}{[u]} = \frac{\frac{1}{2} - \frac{0^2}{2}}{1 - 0} = \frac{1}{2} = \text{Schockgeschw.}$$

Warum ist Charakteristiken-Verf. ok?
1.) ε -Übergang
2.) zwei getrennte Gebiete aneinander kleben.



$$\hat{x}'(t) = \frac{[f(u)]}{[u]} \Big|_{\gamma} = \frac{\frac{0^2}{2} - \frac{1^2}{2}}{0 - 1} = \frac{-\frac{1}{2}}{-1} = \frac{1}{2}$$



Merke:

1. Schwache Lösungen sind nicht eindeutig.
2. Rankine-Hugoniot ist eine notwendige Bedingung, die Auskunft über die Schockgeschwindigkeit gibt.

15 Nichtlineare Erhaltungsgleichungen

3. Man benötigt noch zusätzliche Bedingungen, um *die* physikalisch sinnvolle schwache Lösung auszuwählen.

Eine Möglichkeit:

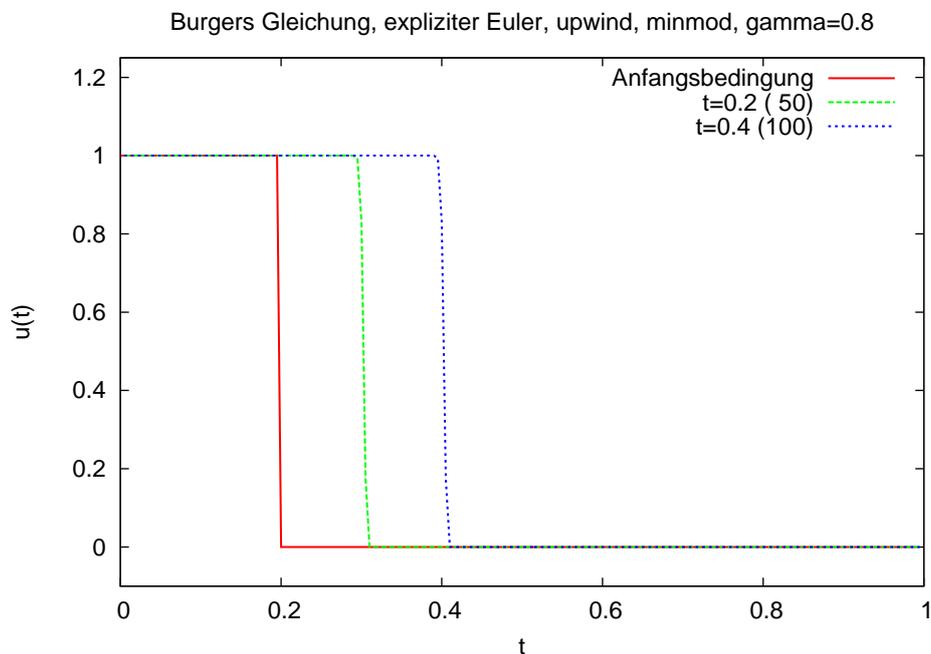
Löse

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = \underbrace{\varepsilon \frac{\partial^2 u}{\partial x^2}}_{\text{Diffusionsterm}}$$

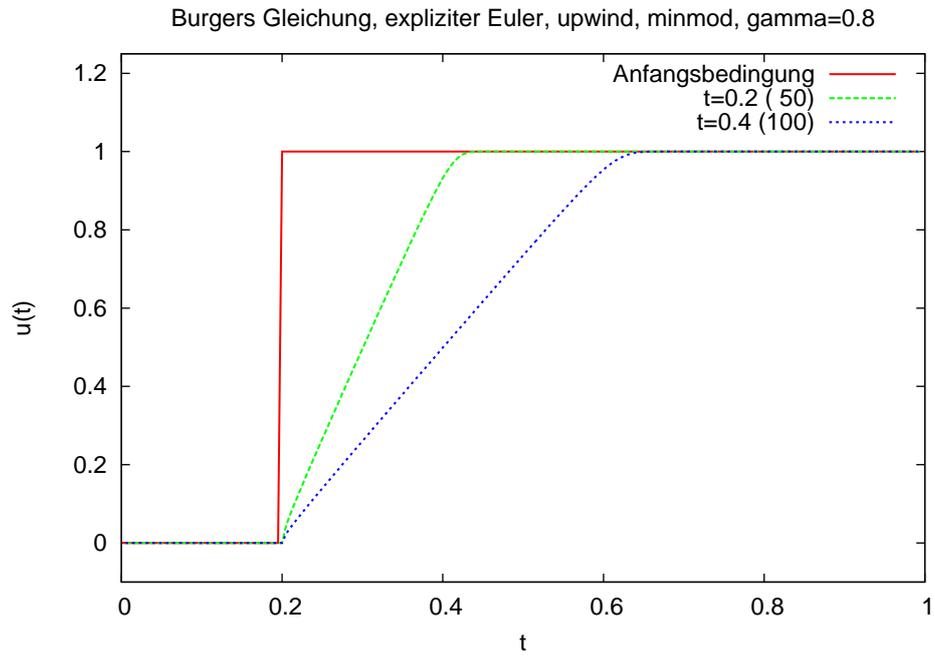
und betrachte $\varepsilon \rightarrow 0$. „vanishing viscosity“ solution.

Bemerkung: Oft liefert das „upwinding“ den entsprechenden Diffusionsterm.

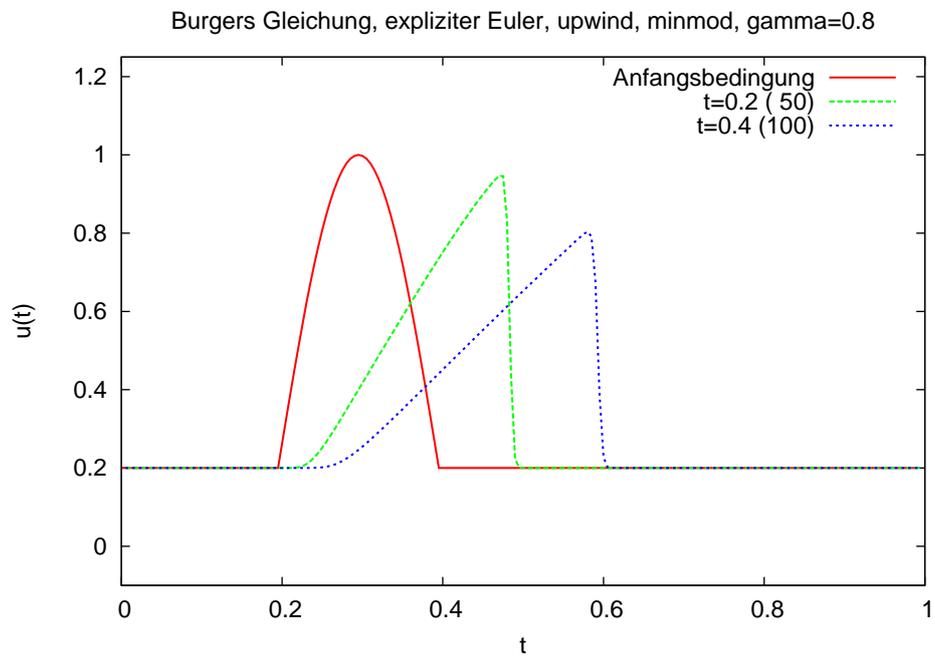
Beispiel Wir lösen die Burger's Gleichung, d. h. die Flussfunktion $f(u) = u^2$.



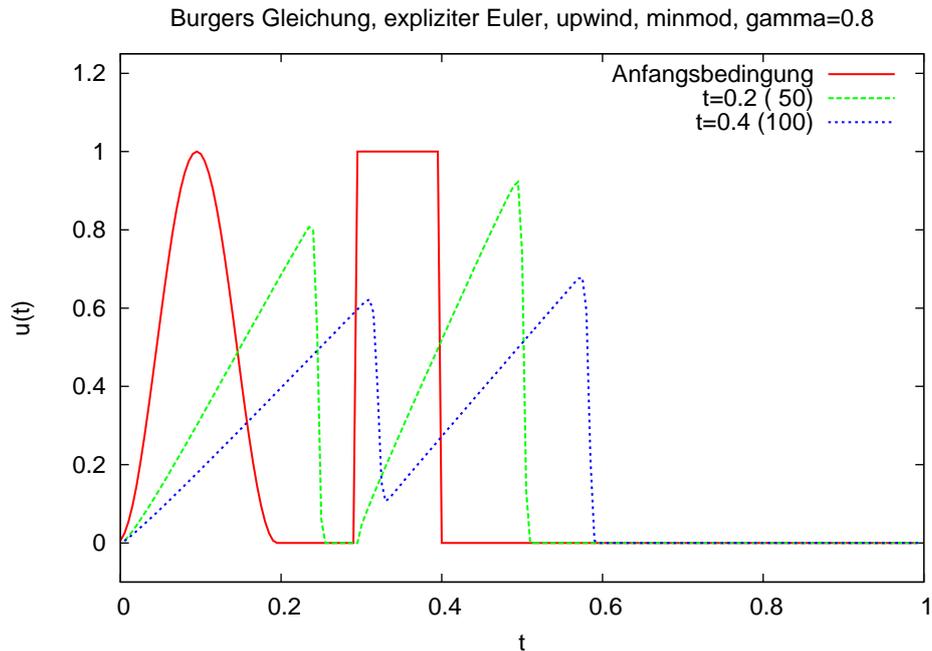
Sprung von oben nach unten: Schock.



Sprung von unten nach oben: Verdünnungswelle.



Ein sinusförmiger Puls: Ein Schock entwickelt sich.



Zwei verschiedene Pulse.

15.2 Bedeutung von FV-Verfahren

Wir betrachten folgendes FD-Verfahren zur Lösung der Burger's Gleichung.

Für genügend glattes u gilt

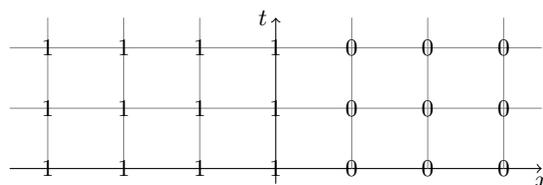
$$\frac{\partial u}{\partial t} + \frac{\partial \left(\frac{u^2}{2} \right)}{\partial x} = \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = 0; \quad u_0(x) = \begin{cases} 1 & x \leq 0 \\ 0 & \text{sonst} \end{cases}$$

FD:

$$\frac{U_i^{k+1} - U_i^k}{\tau} + U_i^k \frac{U_i^k - U_{i-1}^k}{h} = 0 \quad \text{upwind, da } U_i^k \geq 0$$

$$\iff U_i^{k+1} = U_i^k - \frac{\tau}{h} U_i^k (U_i^k - U_{i-1}^k)$$

Dies ergibt für die gegebenen Startwerte



- Falsche Schockgeschwindigkeit (0 statt $\frac{1}{2}$)
- und das gilt unabhängig von τ und $h \rightarrow$ keine Konvergenz
- In Abhängigkeit der Startwerte ergeben sich andere (falsche) Schockgeschwindigkeiten.
- Naive Ansätze können im nichtlinearen Fall böse Überraschungen liefern.

Satz 15.2 (Lax-Wendroff). Gegeben sei

- eine Sequenz von Gittern mit $\tau_l, h_l \rightarrow 0$ mit $l \rightarrow \infty$
- eine konservative Methode (=FV-Ansatz) mit konsistenter Flussfunktion.

Konvergiert dann die numerische Lösung für $l \rightarrow \infty$ gegen ein u , dann ist dieses eine schwache Lösung der Erhaltungsgleichung! □

Mit FV-Verfahren kann einem das nicht passieren!

Allerdings: Dies muss nicht die physikalisch korrekte Lösung sein.

15.3 Godunov Verfahren im nichtlinearen Fall

FV-Verfahren:

$$U_i^{k+1} = U_i^k - \frac{\tau}{h} \left(\mathcal{F}(U_i^k, U_{i+1}^k) - \mathcal{F}(U_{i-1}^k, U_i^k) \right)$$

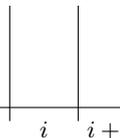
mit der Flussfunktion

$$\mathcal{F}(U_i^k, U_{i+1}^k) = \begin{cases} f(U_i^k) & s = \frac{f(U_i^k) - f(U_{i+1}^k)}{U_i^k - U_{i+1}^k} \geq 0 \\ f(U_{i+1}^k) & \text{sonst} \end{cases} \quad \begin{matrix} \text{Burger's} \\ s = \frac{1}{2} \frac{(U_i^k)^2 - (U_{i+1}^k)^2}{U_i^k - U_{i+1}^k} = \frac{1}{2} (U_i^k + U_{i+1}^k) \end{matrix}$$

Dies ist konsistent mit dem linearen Fall:

$$s = \frac{a \cdot U_i^k - a \cdot U_{i+1}^k}{U_i^k - U_{i+1}^k} = a.$$

Höhere Ordnung: Wieder mit linearer Rekonstruktion.



15.4 Zusammenfassung

- Hyperbolische Gleichungen können unstetige Lösungen haben, wie wir schon mit der Methode der Charakteristiken gesehen haben. Bei nichtlinearer Flussfunktion können sich selbst aus stetigen Anfangsbedingungen zu späterer Zeit unstetige Lösungen entwickeln.
- Da diese nicht mehr vom klassischen Lösungsbegriff erfasst werden entwickelt man eine sogenannte schwache Formulierung als Erweiterung.
- Wieder sind Finite-Volumen-Verfahren sehr gut geeignet, da man hier zeigen kann, dass diese gegen eine schwache Lösung konvergieren (falls sie überhaupt konvergieren).

Lehrbücher Numerik

- [GR06] GROSSMANN, C. und H.-G. ROOS: *Numerische Behandlung partieller Differentialgleichungen*. Teubner, 3. Auflage, 2006.
- [Hac86] HACKBUSCH, W.: *Theorie und Numerik elliptischer Differentialgleichungen*. Teubner, 1986. http://www.mis.mpg.de/scicomp/articleshackbusch_d.html.
- [Hac91] HACKBUSCH, W.: *Iterative Lösung großer schwachbesetzter Gleichungssysteme*. Teubner, 1991.
- [Lev02] LEVEQUE, R. J.: *Finite Volume Methods for Hyperbolic Problems*. Cambridge University Press, 2002.
- [Ran06] RANNACHER, R.: *Einführung in die Numerische Mathematik II (Numerik partieller Differentialgleichungen)*. <http://numerik.iwr.uni-heidelberg.de/~lehre/notes>, 2006.

Sonstige Literatur

- [Arn04] ARNOLD, V. I.: *Lectures on Partial Differential Equations*. Springer, 2004.
- [Eva98] EVANS, L. C.: *Partial Differential Equations*. American Mathematical Society, 1998.
- [Fey70] FEYNMAN, R. P.: *Feynman Lectures on Physics*, Band II. Addison-Wesley, 1970.
- [Mar07] MARKOWICH, P. A.: *Applied Partial Differential Equations*. Springer, 2007.
- [Ran06] RANNACHER, R.: *Einführung in die Numerische Mathematik (Numerik 0)*. <http://numerik.iwr.uni-heidelberg.de/~lehre/notes>, 2006.
- [RR93] RENARDY, M. und R. C. ROGERS: *An Introduction to Partial Differential Equations*, Band 13 der Reihe *Texts in Applied Mathematics*. Springer, 1993.
- [Smi90] SMIRNOW, W. I.: *Lehrbuch der höheren Mathematik*, Band II. Harri Deutsch, 17. Auflage, 1990.