Numerik von Differentialgleichungen

Prof. Dr. Bastian von Harrach

Goethe-Universität Frankfurt am Main
Institut für Mathematik

Sommersemester 2018

http://numerical.solutions

Vorläufige Version, wird während der Vorlesungszeit laufend ergänzt und korrigiert.
Letzte Aktualisierung: 3. Juli 2018

Inhaltsverzeichnis

Kapitel 1

Gewöhnliche Differentialgleichungen

Eine Differentialgleichung ist eine Gleichungen, die eine unbekannte Funktion zusammen mit ihren Ableitungen enthält. In diesem Kapitel beschäftigen wir uns mit der Lösung sogenannter *gewöhnlicher* Differentialgleichungen (engl.: ordinary differential equations, ODE), bei denen die gesuchte Funktion nur von einer reell-wertigen Variablen abhängt, sodass sich alle in der Differentialgleichung vorkommenden Ableitungen auf dieselbe Variable beziehen.

1.1 Einführung und Beispiele

1.1.1 Einfache Beispiele und elementare Begriffe

Beispiel 1.1

(a) Betrachte die Differentialgleichung

$$y'(x) = 0 \quad \textit{(kurz: } y' = 0\textit{)},$$

d.h. gesucht ist eine differenzierbare Funktion

$$y : \mathbb{R} \to \mathbb{R},\ y : x \mapsto y(x) \quad \textit{mit} \quad y'(x) = 0 \quad \forall x \in \mathbb{R}.$$

- Spezielle *Lösungen sind z.B.:*

$$y(x) = 0, \quad y(x) = 1, \quad y(x) = -37, \quad \ldots$$

- *Die* allgemeine *Lösung ist* $y(x) = C$, $C \in \mathbb{R}$, *d.h. man kann zeigen, dass jede solche Funktion die DGL löst und jede Lösung in dieser Form geschrieben werden kann.*

(b) Betrachte die Differentialgleichung

$$y'(x) = ry(x), \quad r \in \mathbb{R} \quad \textit{(kurz: } y' = ry\textit{).}$$

- Spezielle *Lösungen sind z.B.:*

$$y(x) = 0, \quad y(x) = e^{rx}, \quad y(x) = -37e^{rx}, \quad \dots$$

- *Man kann zeigen, dass* $y(x) = Ce^{rx}$, $C \in \mathbb{R}$, *die* allgemeine *Lösung dieser DGL ist.*

(c) Die höchste vorkommende Ableitung bezeichnet man auch als Ordnung *der Differentialgleichung. Die Besipiele in (a) und (b) waren von erster Ordnung. Ein Beispiel für eine Differentialgleichung* höherer Ordnung *ist*

$$y''(x) = -y(x) \quad \textit{(kurz: } y'' = -y\textit{).}$$

- Spezielle *Lösungen sind z.B.:*

$$y(x) = 0, \quad y(x) = \sin(x), \quad y(x) = 37\cos(x), \quad \dots$$

- *Man kann zeigen, dass* $y(x) = C_1 \sin(x) + C_2 \cos(x)$, $C_1, C_2 \in \mathbb{R}$, *die* allgemeine *Lösung dieser DGL ist.*

(d) Betrachte das Differentialgleichungssystem

$$\begin{aligned} y_1''(x) &= -y_1(x) \\ y_2'(x) &= y_3(x) \\ y_3'(x) &= y_3(x) \end{aligned}$$

- *Eine* spezielle *Lösung ist z.B.:*

$$\begin{aligned} y_1(x) &= \sin(x) \\ y_2(x) &= 37 \\ y_3(x) &= 0 \end{aligned}$$

- *Man kann zeigen, dass*

$$\begin{aligned} y_1(x) &= C_1 \sin(x) + C_2 \cos(x), \\ y_2(x) &= C_3 e^x + C_4, \\ y_3(x) &= C_3 e^x, \end{aligned}$$

$C_1, \dots, C_4 \in \mathbb{R}$, *die* allgemeine *Lösung ist.*

(e) Betrachte die Differentialgleichung für eine vektorwertige *Funktion: Gesucht ist (eine differenzierbare Funktion)*

$$y : \mathbb{R} \to \mathbb{R}^2, \quad y : \ x \mapsto y(x) := \begin{pmatrix} y_1(x) \\ y_2(x) \end{pmatrix}.$$

mit $y' = y$, also

$$y'(x) = y(x) \iff \begin{pmatrix} y_1'(x) \\ y_2'(x) \end{pmatrix} = \begin{pmatrix} y_1(x) \\ y_2(x) \end{pmatrix} \iff \begin{cases} y_1'(x) = y_1(x) \\ y_2'(x) = y_2(x) \end{cases}$$

Beispiel 1.2
Betrachte $y : \mathbb{R}^2 \to \mathbb{R}, \quad (x_1, x_2) \mapsto y(x) = y(x_1, x_2)$.

Ein Bespiel für eine nicht gewöhnliche, sondern sogenannte partielle *Differentialgleichung (engl.: Partial Differential Equation, PDE) ist*

$$\frac{\partial^2 y}{\partial x_1^2} + \frac{\partial^2 y}{\partial x_2^2} = 0.$$

Spezielle *Lösungen sind z.B.: $y(x) = 0$, $y(x) = x_1 + x_2$, ...*

1.1.2 Anwendungsbeispiele für gewöhnliche DGL

In praktischen Anwendungen werden mit Differentialgleichungen oft die Änderungen einer messbaren Größe im Laufe der Zeit beschrieben. Je nachdem, um welche messbare Größe es geht, verwenden wir daher in diesem Abschnitt unterschiedliche Bezeichner für die Funktion, und bezeichnen die Zeitvariable mit t. Für Ableitungen bezüglich der Zeit schreiben wir auch $\dot{y}(t)$ statt $y'(x)$.

Stetige Verzinsung Es sei $y(t)$ das Guthaben auf einem Sparkonto zum Zeitpunkt t (gemessen in Jahren). Eine Bank zahle p Zinsen für ein Jahr (z.B. $p = 1\% = 0.01$), also

$$y(t + 1) = y(t)(1 + p).$$

Wir betrachten die Frage, wie sich bei jederzeitiger Verfügbarkeit der faire Wert $y(t)$ des Guthabens entwickelt. Wird das Guthaben nach einem halben Jahr abgehoben, so wäre es nicht fair, dem Kunden dafür $p/2$ Zinsen zu zahlen. Denn dann würde ein Kunde, der nach einem halben Jahr sein Geld abhebt und es sofort wieder für ein weiteres halbes Jahr anlegt, aufgrund der

Zinseszinsen mehr erhalten als ein Kunde, der das Geld nicht zwischendurch abhebt:

$$y(t)(1+\frac{p}{2})(1+\frac{p}{2}) = y(t)\left(1+p+\frac{p^2}{4}\right) > y(t)(1+p).$$

Der faire Zinssatz für ein halbes Jahr wäre q mit $(1+q)^2 = (1+p)$, also

$$y(t+1/2) = y(t) + y(t)q \quad \text{mit } q = \sqrt{1+p} - 1.$$

Entsprechend ist der faire Zinssatz für das Zeitintervall $\Delta t = \frac{1}{n}$ ein q mit $(1+q)^n = 1+p$. Mit $r := \log(1+p)$ ist $\log(1+q) = r\Delta t$ und

$$y(t+\Delta t) = y(t)(1+q) = y(t)e^{\log(1+q)} = y(t)e^{r\Delta t}.$$

Ist der faire Wert $y(t)$ des Guthabens stetig differenzierbar so folgt mit $e^{r\Delta t} = 1 + r\Delta t + O((\Delta t)^2)$, dass $y(t)$ die Differentialgleichung

$$\dot{y}(t) = \lim_{\Delta t \to 0} \frac{y(t+\Delta t) - y(t)}{\Delta t} = ry(t)$$

erfüllt.

Populationsdynamik Es bezeichne $y(t)$ die Anzahl von Individuen einer Population zum Zeitpunkt t. Ein einfaches Populationsmodell (*Malthus-Modell*) besteht darin, anzunehmen, dass die Population wie im obigen Beispiel zur stetigen Verzinsung kontinuierlich mit konstanter Rate $r \in \mathbb{R}$ wächst bzw. fällt. In einem kurzen Zeitintervall Δt, verändert sich die Population also um $\Delta y \approx ry\Delta t$ und die Populationsgröße erfüllt daher die DGL

$$\dot{y} = ry.$$

Es erscheint plausibel, dass eine Population nur bis zu einem gewissen Maximalwert $M > 0$ wachsen kann. Nur für $y << M$ sollte das Wachstum mit Rate r stattfinden, für $y \approx M$ sollte es kein Wachstum mehr geben. Im *Verhulst-Modell* nimmt man daher an, dass die Wachstumsrate $r(1-y/M)$ beträgt, also $\Delta y \approx r(1-y/M)y\Delta t$ und die Populationsgröße also die DGL

$$\dot{y} = r\left(1-\frac{y}{M}\right)y = ry - \frac{r}{M}y^2.$$

erfüllt.

Wir betrachten noch das *Lotka-Volterra-Modell* für Räuber-Beute-Populationen. $y_1(t)$ sei die Populationsgröße der Beutetiere, $y_2(t)$ die der Raubtiere.

Dann sollte die Wachstumsrate der Beutetiere mit steigender Räuberzahl abnehmen und die Wachstumsrate der Räuber mit steigender Beutezahl zunehmen. Wie im Verhulst-Modell erhalten wir

$$\begin{aligned}\dot{y}_1 &= r_1 y_1 - f_1 y_1 y_2,\\ \dot{y}_2 &= -r_2 y_2 + f_2 y_1 y_2\end{aligned}$$

mit Parametern $r_1, r_2, f_1, f_2 > 0$.

Chemische Reaktionen $A(t)$, $B(t)$, $C(t)$, usw. bezeichne die Konzentration von Molekülen A,B,C. Wir nehmen an, dass sich (in einem kurzen Zeitintervall Δt) $kA(t)\Delta t$ Moleküle von A in B umwandeln und schreiben dafür $A \xrightarrow{k} B$. Dann erfüllen die Konzentrationen die DGL

$$\dot{A}(t) = -kA(t), \qquad \dot{B}(t) = kA(t).$$

Betrachten wir noch die Reaktionsvorschrift $A + B \xrightarrow{k} C + 2D$, also das eine chemische Reaktion mit Rate k jeweils ein Molekül A mit einem Molekül B zu einem Molekül C sowie zwei Molekülen D umwandelt. Dann erfüllen die Konzentrationen die DGL

$$\dot{A} = -kAB, \quad \dot{B} = -kAB, \quad \dot{C} = kAB, \quad \dot{D} = 2kAB.$$

Newtonsche Mechanik Es bezeichne

$x(t) = (x_1(t), x_2(t), x_3(t))^T$ die Position eines Körpers zum Zeitpunkt t,

$v(t) = \dot{x}(t) = (\dot{x}_1(t), \dot{x}_2(t), \dot{x}_3(t))^T$ seine Geschwindigkeit und

$a(t) = \dot{v}(t) = \ddot{x}(t) = (\ddot{x}_1(t), \ddot{x}_2(t), \ddot{x}_3(t))^T$ seine Beschleunigung.

Das *Newtonsche Gesetz* besagt, dass die Wirkung einer Kraft $F(t)$ auf einen Körper der Masse m zu einer Beschleunigung führt gemäß

$$F(t) = ma(t) = m\ddot{x}(t).$$

1.1.3 Anfangswertprobleme

Die Lösung einer gewöhnlichen DGL ist üblicherweise nicht eindeutig. Die allgemeine Lösung von $\dot{y}(t) = ry(t)$ ist $y(t) = Ce^{rt}$ mit einem Parameter $C \in \mathbb{R}$. In vielen Anwendungen sind die Parameter eindeutig bestimmt durch die Anfangswerte von y. Bei der stetigen Verzinsung ist z.B. $C = y(0)$ das anfängliche Sparguthaben.

Intuitiv erwarten wir in den anderen Beispielen, dass die Lösung durch folgende Informationen eindeutig bestimmt wird:

- Populationsdynamik: anfängliche Population $y(0)$, bzw., $y_1(0), y_2(0)$.
- Chemische Reaktionen: anfängliche Konzentrationen $A(0)$, $B(0)$, ...
- Newtonsche Gesetze: anfängliche Position $x_1(0), x_2(0), x_3(0)$ und Geschwindigkeit $v_1(0), v_2(0), v_3(0)$.

Eine gewöhnliche DGL zusammen mit Anfangsbedingungen heißt auch *Anfangswertproblem* (AWP).

1.1.4 Elementare Lösungsmethoden

Ähnlich wie Integrale lassen sich manche (aber nicht alle) gewöhnliche Differentialgleichungen analytisch, d.h. in geschlossener Form lösen. Wir geben hier nur Beispiele besonders einfacher Lösungsmethoden an. Es existieren noch einige weitere wichtige analytische Lösungsmethoden, aber im Allgemeinen lassen sich gewöhnliche Differentialgleichungen nur numerisch lösen.

Raten/Wissen der Lösung Für einfache Beispiele kann man die Lösung raten, siehe Beispiel 1.1. Ein so gefundener Lösungskandidat kann durch Einsetzen überprüft werden, womit rigoros bewiesen ist, dass dies tatsächlich eine Lösung ist. Damit ist jedoch noch unklar, ob es noch andere Lösungen gibt.

Für eine große Klasse von Anfangswertproblemen kann die Eindeutigkeit einer Lösung gezeigt werden (siehe Satz 1.5 und 1.13). In dem Fall ist die geratene Lösung die einzige und das Problem durch Raten (genauer: durch das Einsetzen der geratenen Lösung in das AWP) vollständig gelöst.

Separation der Variablen Gewöhnliche DGL der Form

$$y'(x) = g(x)h(y(x))$$

lassen sich formal(!) schreiben als

$$\frac{\mathrm{d}y}{h(y)} = g(x)\,\mathrm{d}x \tag{1.1}$$

und durch Integration lösen

$$\int \frac{1}{h(y)}\,\mathrm{d}y = \int g(x)\,\mathrm{d}x. \tag{1.2}$$

Formal bedeutet dabei, dass dies keine mathematisch rigorosen Umformulierungen sind. Die Ausdrücke $\mathrm{d}x$ und $\mathrm{d}y$ sind nicht definiert!

Man kann diese Methode rigoros formulieren und rechtfertigen (siehe z.B. [Heuser, Satz 8.1]). Aber auch ohne rigorose Rechtfertigung haben formale Methoden einen großen Nutzen (nicht nur im Bereich gewöhnlicher DGL). Oft lässt sich nämlich durch formales Vorgehen ein Lösungskandidat bestimmen und für diesen dann (wie bei einer geratenen Lösung) rigoros überprüfen, ob er tatsächlich eine Lösung ist.

Beispiel 1.3
Betrachte das Verhulst-Modell aus der Populationsdynamik mit $M = r = 1$:

$$\frac{\mathrm{d}y}{\mathrm{d}t} = (1-y)\,y$$

mit Anfangwert $y(0) > 1$.

Obiges formales Vorgehen liefert

$$\int \frac{1}{(1-y)\,y}\,\mathrm{d}y = \int 1\,\mathrm{d}t.$$

Wir erwarten anschaulich, dass die Population von oben gegen ihren Maximalwert $M = 1$ konvergieren diesen aber nicht unterschreiten wird. Wir lösen also die Integrale auf beiden Seiten unter der Vermutung $y > 1$:

$$\begin{aligned}\int 1\,\mathrm{d}t &= t + \mathit{const.}\\ \int \frac{1}{(1-y)\,y}\,\mathrm{d}y &= -\int \frac{1}{y-1}\,\mathrm{d}y + \int \frac{1}{y}\,\mathrm{d}y\\ &= -\ln(y-1) + \ln(y) + \mathit{const.} = \ln\frac{y}{y-1} + \mathit{const.}\end{aligned}$$

Wir erhalten

$$\ln \frac{y}{y-1} = t + const. \quad \Longrightarrow \quad \frac{y}{y-1} = Ce^t$$
$$\Longrightarrow \quad y = \frac{Ce^t}{Ce^t - 1}.$$

C kann aus dem Anfangswert $y(0) = y_0$ bestimmt werden:

$$y_0 = \frac{C}{C-1} \quad \Longrightarrow \quad C = \frac{y_0}{y_0 - 1}.$$

Man prüft leicht nach, dass der so bestimmte Lösungskandidat für $y_0 > 1$ tatsächlich das AWP

$$\frac{\mathrm{d}y}{\mathrm{d}t} = (1-y)\, y, \quad y(0) = y_0$$

löst. Damit ist dann mathematisch rigoros gezeigt, dass dies eine Lösung ist. Man kann zeigen (siehe Satz 1.13), dass das AWP eindeutig lösbar ist, dies also die einzige Lösung ist.

Variation der Konstanten Betrachte $y'(x) = ry(x) + z(x)$.

Allgemeine Lösung der *homogenen* Gleichung $y'(x) = ry(x)$ ist $y(x) = Ce^{rx}$ mit einer Konstante $C \in \mathbb{R}$.

Ansatz: Ersetze zur Lösung der inhomogenen Gleichung die Konstante durch eine Funktion $C(x)$:

$$C'(x)e^{rx} + C(x)re^{rx} = y'(x) = ry(x) + z(x) = rC(x)e^{rx} + z(x)$$
$$\Longleftrightarrow \quad C'(x) = z(x)e^{-rx}$$

Durch Integration erhalten wir $C(x)$ und damit die Lösung $y(x) = C(x)e^{rx}$.

Findet man eine Lösung durch einen Ansatz, so ist damit nur bewiesen, dass dies eine Lösung ist und (wenn die Lösung aus Äquivalenzumformungen aus dem Ansatz hervorgegangen ist), dass es keine andere Lösung der angesetzten Form geben kann. Im Allgemeinen ist damit nicht geklärt, ob es noch andere Lösungen gibt, die nicht dem Ansatz entsprechen. Auch dies folgt aber häufig wieder aus allgemeinen Eindeutigkeitsresulaten wie Satz 1.5 und 1.13.

1.2 Theorie gewöhnlicher DGL

1.2.1 Eine allgemeine Form

Von nun an betrachten wir stets Anfangswertprobleme in der folgenden allgemeinen Form

$$y'(x) = f(x, y(x)), \qquad y(a) = y_0,$$

wobei $x \in [a, b] \subset \mathbb{R}$, $b > a$, $y(x) = (y_1(x), \ldots, y_d(x))^T \in \mathbb{R}^d$ vektorwertig ist, $d \in \mathbb{N}$, und

$$f : \mathbb{R}^{d+1} \to \mathbb{R}^d, \quad f(x, y) = \begin{pmatrix} f_1(x, y_1, \ldots, y_d) \\ \vdots \\ f_d(x, y_1, \ldots, y_d) \end{pmatrix}.$$

Die Differentialgleichung $y' = f(x, y(x))$ ist äquivalent zum Differentialgleichungssystem

$$\begin{aligned} y_1'(x) &= f_1(x, y_1(x), \ldots, y_d(x)), \\ &\vdots \\ y_d'(x) &= f_d(x, y_1(x), \ldots, y_d(x)). \end{aligned}$$

Gleichungen höherer Ordnung (d.h. solche, die höhere Ableitungen von y enthalten) können oft in diese Form transformiert werden, indem y und seine Ableitungen (bis zur zweithöchsten) in einer vektorwertigen Hilfsfunktion $u = (u_1, u_2, \ldots)$ zusammengefasst werden

$$u_1(x) := y(x), \quad u_2(x) := y'(x), \quad u_3(x) = y''(x), \ldots$$

Beispiel 1.4
$y'' = -y$ kann in obige Form transformiert werden durch

$$u = \begin{pmatrix} u_1(x) \\ u_2(x) \end{pmatrix} := \begin{pmatrix} y(x) \\ y'(x) \end{pmatrix}.$$

Damit ist $y'' = -y$ äquivalent zu

$$u' = \begin{pmatrix} u_1'(x) \\ u_2'(x) \end{pmatrix} = \begin{pmatrix} y'(x) \\ y''(x) \end{pmatrix} = \begin{pmatrix} y'(x) \\ -y(x) \end{pmatrix} = \begin{pmatrix} u_2(x) \\ -u_1(x) \end{pmatrix} =: f(x, u(x)).$$

1.2.2 Existenz, Eindeutigkeit und Stabilität

Ein Problem heißt *wohlgestellt* (nach Hadamard) wenn

(a) eine Lösung existiert (*Existenz*),

(b) die Lösung eindeutig ist (*Eindeutigkeit*),

(c) die Lösung stetig von den Eingabeparametern abhängt (*Stabilität*).

Für das Anfangswertproblem

$$y'(x) = f(x, y(x)), \qquad y(a) = y_0.$$

bedeutet Wohlgestelltheit, dass eine eindeutige Lösung $y(x)$ existiert und diese Lösung stetig von den Anfangswerten y_0 (und ggf. weiteren in f vorhandenen Parametern) abhängt. Dies werden wir in diesem Abschnitt untersuchen.

Existenz und Eindeutigkeit In diesem Abschnitt beweisen wir den Existenz- und Eindeutigkeitssatz von Picard-Lindelöf in seiner einfachsten globalen Form (vgl. Satz 1.13 und Bemerkung 1.14 für eine Abschwächung der Voraussetzungen). Hier und in der gesamten Vorlesung bezeichne dabei $\|\cdot\|$ im Raum $\mathbb{R}^d$ stets die Euklidnorm.

Satz 1.5 (Picard-Lindelöf)
Seien $a, b \in \mathbb{R}$, $a < b$, $y_0 \in \mathbb{R}^d$. Für

$$f:\ [a,b] \times \mathbb{R}^d \to \mathbb{R}^d, \quad f:\ (x,y) \mapsto f(x,y)$$

gelte

(a) f ist stetig

(b) f ist (global und bzgl. x gleichmäßig) Lipschitz-stetig in y, d.h. es existiert ein $L > 0$, sodass

$$\|f(x,y) - f(x,z)\| \le L\,\|y - z\| \quad \forall x \in [a,b],\ y, z \in \mathbb{R}^d.$$

Dann existiert genau eine differenzierbare Funktion $y : [a,b] \to \mathbb{R}^d$ mit

$$y'(x) = f(x, y(x)) \quad \forall x \in [a,b] \quad \text{und} \quad y(a) = y_0$$

und diese ist stetig differenzierbar.

Wir werden Satz 1.5 mit Hilfe des Banachschen Fixpunktsatzes und einigen Hilfssätzen beweisen.

Satz 1.6 (Banachscher Fixpunktsatz)
Sei $(X, \|\cdot\|_*)$ *ein Banachraum (d.h. ein vollständiger normierter Vektorraum) und* Φ *eine* kontrahierende Selbstabbildung *auf* X*, d.h. es existiere ein* $q < 1$ *mit*

$$\Phi : X \to X \quad \textit{und} \quad \|\Phi(x) - \Phi(y)\|_* \leq q \, \|x - y\|_* \quad \forall x, y \in X.$$

Dann besitzt Φ *genau einen* Fixpunkt $\hat{x}$*, d.h. genau ein* $\hat{x} \in X$ *mit* $\Phi(\hat{x}) = \hat{x}$*. Für jeden Startwert* $x^{(0)} \in X$ *konvergiert die durch* Fixpunktiteration *definierte Folge*

$$(x^{(k)})_{k \in \mathbb{N}_0}, \qquad x^{(k+1)} := \Phi(x^{(k)}) \quad \forall k \in \mathbb{N}_0$$

(bzgl. der Norm $\|\cdot\|_*$*) gegen* $\hat{x}$*, d.h.*

$$\left\|x^{(k)} - \hat{x}\right\|_* \to 0.$$

Beweis: Wir haben das Resultat für abgeschlossene Teilmengen des $\mathbb{C}^n$ in [NumerikWS1718, Satz 3.1] gezeigt. Wir wiederholen den Beweis, um uns zu überzeugen, dass er auch in allgemeinen (möglicherweise unendlichdimensionalen) Banachräumen gültig bleibt:

(i) Konvergenz der Fixpunktiteration:

Wir zeigen zuerst, dass die durch Fixpunktiteration definierte Folge für jeden Startwert konvergiert. Sei also $x^{(0)} \in X$ und $(x^{(k)})_{k \in \mathbb{N}_0}$ definiert durch $x^{(k+1)} := \Phi(x^{(k)})$ für alle $k \in \mathbb{N}_0$.

Für jedes $k \in \mathbb{N}$ erhalten wir

$$\begin{aligned}\left\|x^{(k+1)} - x^{(k)}\right\|_* &= \left\|\Phi(x^{(k)}) - \Phi(x^{(k-1)})\right\|_* \leq q \left\|x^{(k)} - x^{(k-1)}\right\|_* \\ &\leq \ldots \leq q^k \left\|x^{(1)} - x^{(0)}\right\|_*.\end{aligned}$$

Damit folgt für alle $m, l \in \mathbb{N}$ mit $l > m$

$$\begin{aligned}&\left\|x^{(l)} - x^{(m)}\right\|_* \\ &\leq \left\|x^{(l)} - x^{(l-1)}\right\|_* + \left\|x^{(l-1)} - x^{(l-2)}\right\|_* + \ldots + \left\|x^{(m+1)} - x^{(m)}\right\|_* \\ &\leq \left(q^{l-1} + q^{l-2} + \ldots + q^m\right) \left\|x^{(1)} - x^{(0)}\right\|_* \\ &= q^m \left(q^0 + \ldots + q^{l-m-2} + q^{l-m-1}\right) \left\|x^{(1)} - x^{(0)}\right\|_* \\ &\leq \frac{q^m}{1-q} \left\|x^{(1)} - x^{(0)}\right\|_*.\end{aligned}$$

und damit (beachte $q < 1$)

$$\lim_{l,m\to\infty} \left\|x^{(l)} - x^{(m)}\right\|_* = 0.$$

$(x^{(k)})_{k\in\mathbb{N}_0}$ ist also ein Cauchy-Folge und damit (da X vollständig ist) konvergent. Für jeden Startwert $x^{(0)} \in X$ konvergiert also die durch Fixpunktiteration definierte Folge gegen einen Grenzwert

$$\hat{x} := \lim_{k\to\infty} x^{(k)} \in X.$$

(ii) Existenz und Eindeutigkeit des Fixpunktes:

Nun zeigen wir, dass jeder solche Grenzwert $\hat{x} \in X$ ein Fixpunkt von Φ ist. Aus der Kontraktionseigenschaft folgt insbesondere dass Φ auf X Lipschitz-stetig ist und damit

$$\hat{x} = \lim_{k\to\infty} x^{(k)} = \lim_{k\to\infty} \Phi(x^{(k-1)}) = \Phi\left(\lim_{k\to\infty} x^{(k-1)}\right) = \Phi(\hat{x}).$$

Die Fixpunktiteration konvergiert also für jeden Startwert gegen einen Fixpunkt. Insbesondere ist damit gezeigt, dass Φ (mindestens) einen Fixpunkt $\hat{x}$ besitzt.

Nun ist nur noch die Eindeutigkeit des Fixpunktes zu zeigen. Seien $\hat{x}, \tilde{x} \in X$ Fixpunkte, also

$$\hat{x} = \Phi(\hat{x}), \quad \text{und} \quad \tilde{x} = \Phi(\tilde{x}).$$

Dann ist

$$\|\hat{x} - \tilde{x}\|_* = \|\Phi(\hat{x}) - \Phi(\tilde{x})\|_* \le q \|\hat{x} - \tilde{x}\|_*$$

und aus $q < 1$ folgt damit $\|\hat{x} - \tilde{x}\|_* = 0$, d.h. $\hat{x} = \tilde{x}$. □

Bemerkung 1.7
Offensichtlich gilt Satz 1.6 mit dem gleichen Beweis bereits auf jedem vollständigen metrischen Raum X, also insbesondere auch falls X eine abgeschlossene Teilmenge eines Banachraums ist.

Wir werden den Banachschen Fixpunktsatz anwenden, indem wir das Anfangswertproblem in eine Fixpunktgleichung im Raum der stetigen Funktionen auf $[a, b]$,

$$C([a, b])^d := \{y : \ [a, b] \to \mathbb{R}^d \text{ stetig}\},$$

umschreiben.

Lemma 1.8
Es gelten die Voraussetzungen von Satz 1.5. Eine differenzierbare Funktion $y: [a,b] \to \mathbb{R}^d$ *erfüllt genau dann das AWP*

$$y'(x) = f(x, y(x)) \quad \forall x \in [a,b] \quad \textit{und} \quad y(a) = y_0,$$

wenn y *stetig ist und die folgende Fixpunktgleichung löst:*

$$y(x) = y_0 + \int_a^x f(t, y(t))\,\mathrm{d}t \quad \forall x \in [a,b].$$

y *ist dann auch stetig differenzierbar.*

Beweis: Ist y differenzierbar und erfüllt das AWP, so ist y' auch stetig und nach dem Hauptsatz der Differential- und Integralrechnung gilt

$$y(x) = y(a) + \int_a^x y'(t)\,\mathrm{d}t = y_0 + \int_a^x f(t, y(t))\,\mathrm{d}t \quad \forall x \in [a,b].$$

Ist umgekehrt $y \in C([a,b])^d$ und erfüllt die Fixpunktgleichung, so ist auch $t \mapsto f(t, y(t))$ stetig und

$$y(x) = y_0 + \int_a^x f(t, y(t))\,\mathrm{d}t$$

ist differenzierbar, $y'(x) = f(x, y(x))$ und $y(a) = y_0$. □

Lemma 1.9
$C([a,b])^d$ *ist bezüglich der* Supremums- *(auch:* Maximumsnorm*)*

$$\|y\|_\infty := \max_{x \in [a,b]} \|y(x)\|$$

ein Banachraum.

Beweis: Übungsaufgabe 1.2. □

Das AWP ist also äquivalent zur Fixpunktgleichung

$$y = y_0 + \int_{x_0}^x f(t, y(t))\,\mathrm{d}t$$

im Banachraum $C([a,b])^d$. Damit die rechte Seite der Fixpunktgleichung eine Kontraktion ist, benötigen wir aber noch eine etwas abgewandelte Norm.

Folgerung 1.10
Für eine Funktion $w \in C([a,b])$ gelte

$$\exists c, C > 0 : \ c \leq w(x) \leq C \quad \forall x \in [a,b].$$

Dann ist die gewichtete Supremumsnorm

$$\|y\|_w := \max_{x \in [a,b]} (w(x) \|y(x)\|)$$

zur Supremumsnorm äquivalent und insbesondere ist $C([a,b])^d$ auch bzgl. $\|\cdot\|_w$ ein Banachraum.

Beweis: Offenbar ist $\|\cdot\|_w$ tatsächlich eine Norm und es gilt

$$c \|y\|_\infty \leq \|y\|_w \leq C \|y\|_\infty \quad \forall y \in C([a,b])^d.$$

Die gewichtete Supremumsnorm ist also zur Supremumsnorm äquivalent. Insbesondere ist jede Cauchy-Folge bzgl. $\|\cdot\|_w$ auch eine bzgl. $\|\cdot\|_\infty$ und der Grenzwert bzgl. $\|\cdot\|_\infty$ ist auch der Grenzwert bzgl. $\|\cdot\|_w$. □

Jetzt können wir Satz 1.5 beweisen:

Beweis von Satz 1.5: Nach unserer Vorarbeit genügt es zu zeigen, dass durch

$$y \mapsto \Phi(y), \quad \Phi(y) : x \mapsto y_0 + \int_a^x f(t, y(t)) \, \mathrm{d}t$$

eine bzgl. einer gewichteten Supremumsnorm kontrahierende Selbstabbildung von $C([a,b])^d$ ist.

Die Selbstabbildungseigenschaft ist klar. Für die Kontraktionseigenschaft betrachten wir für zwei Funktionen $y^{(1)}, y^{(2)} \in C([a,b])^d$ und $x \in [a,b]$

$$\begin{aligned}
&\left\|\Phi(y^{(1)})(x) - \Phi(y^{(2)})(x)\right\| \\
&= \left\| \int_a^x \left(f(t, y^{(1)}(t)) - f(t, y^{(2)}(t)) \right) \mathrm{d}t \right\| \\
&\leq \int_a^x \left\| f(t, y^{(1)}(t)) - f(t, y^{(2)}(t)) \right\| \mathrm{d}t \leq \int_a^x L \left\| y^{(1)}(t) - y^{(2)}(t) \right\| \mathrm{d}t.
\end{aligned}$$

Hieraus folgt

$$\left\|\Phi(y^{(1)}) - \Phi(y^{(2)})\right\|_\infty \leq L(b-a) \left\| y^{(1)} - y^{(2)} \right\|_\infty,$$

sodass Φ nur für kleine L oder nah an a liegendes b eine Kontraktion ist.

Durch Einfügen einer Gewichtsfunktion $w(x)$ (mit den in Folgerung 1.10 genannten Eigenschaften) erhalten wir jedoch

$$\begin{aligned} & w(x) \left\| \Phi(y^{(1)})(x) - \Phi(y^{(2)})(x) \right\| \\ & \leq w(x) \int_a^x L \frac{1}{w(t)} w(t) \left\| y^{(1)}(t) - y^{(2)}(t) \right\| \, \mathrm{d}t \\ & \leq w(x) L \left\| y^{(1)} - y^{(2)} \right\|_w \int_a^x \frac{1}{w(t)} \, \mathrm{d}t \end{aligned}$$

und damit

$$\left\| \Phi(y^{(1)}) - \Phi(y^{(2)}) \right\|_w \leq \left\| y^{(1)} - y^{(2)} \right\|_w L \max_{x \in [a,b]} \left(w(x) \int_a^x \frac{1}{w(t)} \, \mathrm{d}t \right).$$

Φ ist also eine Kontraktion bzgl. $\|\cdot\|_w$ wenn wir eine Gewichtsfunktion w finden mit

$$L \left(w(x) \int_a^x \frac{1}{w(t)} \, \mathrm{d}t \right) < 1 \quad \forall x \in [a,b].$$

Offenbar gilt für jedes $r > 0$ und $x \in [a,b]$, dass

$$e^{-r(x-a)} \int_a^x \frac{1}{e^{-r(t-a)}} \, \mathrm{d}t = e^{-r(x-a)} \frac{1}{r} e^{r(t-a)} \Big|_{t=a}^{t=x} = \frac{1}{r} - e^{-r(x-a)} \frac{1}{r} \leq \frac{1}{r}.$$

Mit $w(x) = e^{-2L(x-a)}$ ist also Φ eine Kontraktion (mit Kontraktionskonstante $1/2$) bzgl. $\|\cdot\|_w$, womit Satz 1.5 bewiesen ist. □

Beispiel 1.11
Auf die Voraussetzung der Lipschitz-Stetigkeit kann nicht verzichtet werden, wie die folgenden Beispiele zeigen:

(a) Betrachte das AWP

$$y' = \sqrt{y}, \quad y(0) = 0.$$

Die rechte Seite $f(x,y) := \sqrt{y}$ ist nicht Lipschitz-stetig. Tatsächlich ist die Lösung des AWPs nicht eindeutig. Zwei Lösungen sind z.B.

$$y(x) = 0 \quad \textit{und} \quad y(x) = \begin{cases} 0 & \textit{für } 0 \leq x \leq 1, \\ \frac{1}{4}(x-1)^2 & \textit{für } x > 1. \end{cases}$$

(b) Betrachte das AWP

$$y' = y^2, \quad y(0) = 1.$$

Die rechte Seite $f(x, y) = y^2$ ist nur lokal aber nicht global Lipschitz stetig. Man kann zeigen, dass nur

$$y(x) = \frac{1}{1-x}$$

das AWP lösen kann. Es existiert daher keine Lösung auf Intervallen $[0, b]$ mit $b > 1$.

Die Annahme globaler Lipschitz-Stetigkeit ist im Allgemein zu restriktiv. Selbst sehr einfache Modelle (wie das Verhulst-Modell aus Abschnitt 1.1.2) erfüllen diese Annahme nicht. Eine sehr viel realistischere Annahme ist es, nur stetige Differenzierbarkeit der rechten Seite der Differentialgleichung zu fordern, womit zumindest noch lokale Lipschitz-Stetigkeit der Lösung folgt.

Lemma 1.12
Seien $a, b \in \mathbb{R}$, $a < b$. Ist

$$f : [a, b] \times \mathbb{R}^d \to \mathbb{R}^d, \quad f : (x, y) \mapsto f(x, y)$$

stetig differenzierbar, so ist f lokal und bzgl. x gleichmäßig Lipschitz-stetig in y, d.h. für jede kompakte Teilmenge $K \subset \mathbb{R}^d$ existiert ein $L > 0$ mit

$$\|f(x, y) - f(x, z)\| \leq L \|y - z\| \quad \forall x \in [a, b],\ y, z \in K.$$

Beweis: Aus dem mehrdimensionalen Mittelwertsatz der Differentialrechnung (siehe z.B. [NumerikWS1718, Lemma 4.1]) folgt, dass

$$f(\eta) - f(\zeta) = \int_0^1 f'(\eta + t(\zeta - \eta))(\zeta - \eta)\,\mathrm{d}t \quad \text{für alle } \eta, \zeta \in [a, b] \times \mathbb{R}^d,$$

wobei $f'(\eta) \in \mathbb{R}^{d \times (d+1)}$ die bezüglich aller $d + 1$ Variablen gebildete Jacobi-Matrix bezeichnet. Insbesondere ist damit auch

$$f(x, y) - f(x, z) = \int_0^1 f_y(x, y + t(z - y))(z - y)\,\mathrm{d}t \quad \text{für alle } y, z \in \mathbb{R}^d,$$

wobei $f_y(x, y) \in \mathbb{R}^{d \times d}$ die bezüglich der y-Variablen gebildete Jacobi-Matrix, also die hinteren d Spalten von $f'(x, y)$ bezeichnet.

Jede kompakte Teilmenge $K \subset \mathbb{R}^d$ ist in einer hinreichend großen Kugel $B_R(0) \subset \mathbb{R}^d$ enthalten und aufgrund der stetigen Differenzierbarkeit existiert ein $L > 0$ mit

$$\|f_y(x, y)\|_{\mathrm{F}} \leq L \quad \text{für alle } x \in [a, b],\ y \in \overline{B_R(0)}.$$

Da für alle $y, z \in K$ die Verbindungslinie von y nach z in $B_R(0)$ verläuft, also $y + t(z-y) \in B_R(0)$ für alle $t \in [0,1]$ gilt, und die Frobenius-Norm $\|\cdot\|_{\mathrm{F}}$ mit der Euklid-Norm $\|\cdot\|$ verträglich ist, folgt damit

$$\|f(x,y) - f(x,z)\| \leq \int_0^1 \|f_y(x, y + t(z-y))(z-y)\| \, \mathrm{d}t \leq L \, \|z-y\|$$

für alle $y, z \in K$. □

Wir übertragen nun die Aussagen des Satzes von Picard-Lindelöf (Satz 1.5) auf die realistischere Annahme lokaler Lipschitz-Stetigkeit. Wir werden zeigen, dass auch unter dieser realistischeren Annahme die Eindeutigkeit der Lösung garantiert ist und eine Lösung zumindest auf einem Teilintervall existiert. Dabei nutzen wir das wichtige Prinzip aus, dass die rechte Seite einer Differentialgleichungen außerhalb einer großen beschränkten Menge typischerweise keine Rolle mehr spielt, da die Lösung innerhalb gewisser Grenzen bleibt oder das durch die DGL beschriebene Modell außerhalb gewisser Grenzen sowieso seine Gültigkeit verliert. Mathematisch können wir dies insofern ausnutzen, dass eine beschränkte Lösung der ursprünglichen DGL auch noch Lösung jeder abgeänderten DGL ist, bei der die rechte Seite nur in solchen Werten abgeändert wurde, die von der Lösung gar nicht erreicht werden. Mit dem gleichen Argument werden wir in späteren Kapiteln numerische Verfahren zunächst unter unrealistisch restriktiven Annahmen untersuchen und diese Ergebnisse dann auch auf realistische Situationen übertragen können.

Satz 1.13
Seien $a, b \in \mathbb{R}$, $a < b$, $y_0 \in \mathbb{R}^d$ und

$$f : \ [a,b] \times \mathbb{R}^d \to \mathbb{R}^d, \quad f : \ (x,y) \mapsto f(x,y)$$

sei stetig in (x,y) und (wie in Lemma 1.12 definiert) lokal und bzgl. x gleichmäßig Lipschitz-stetig in y.

Dann gilt:

(a) Für jedes nicht-leere Teilintervall $[a,\beta] \subseteq [a,b]$ existiert höchstens eine differenzierbare Funktion $y : [a,\beta] \to \mathbb{R}^d$ mit

$$y'(x) = f(x, y(x)) \quad \forall x \in [a,\beta] \quad \text{und} \quad y(a) = y_0$$

und diese ist stetig differenzierbar.

(b) Es existiert ein nicht-leeres Teilintervall $[a,\beta] \subseteq [a,b]$ und eine stetig differenzierbare Funktion $y : [a,\beta] \to \mathbb{R}^d$ mit

$$y'(x) = f(x, y(x)) \quad \forall x \in [a,\beta] \quad \text{und} \quad y(a) = y_0.$$

Beweis: (a) Angenommen $y, z : [a, \beta] \to \mathbb{R}^d$ sind zwei Lösungen des AWP, also

$$\begin{aligned} y'(x) &= f(x, y(x)) \quad \forall x \in [a, \beta] \quad \text{und} \quad y(a) = y_0, \\ z'(x) &= f(x, z(x)) \quad \forall x \in [a, \beta] \quad \text{und} \quad z(a) = y_0. \end{aligned}$$

Da beide Funktionen stetig sind, existiert ein $C > 0$ mit

$$\|y(x)\|^2 \leq C \quad \text{und} \quad \|z(x)\|^2 \leq C \quad \text{für alle } x \in [a, \beta].$$

Wir ersetzen die rechte Seite f des AWP durch eine abgeänderte rechte Seite $\tilde{f}(x, y) = f(x, y)\varphi(\|y\|^2)$, wobei φ eine beliebig oft stetig differenzierbar Funktion ist mit

$$\varphi(r) = \begin{cases} 1 & \text{für } r < C. \\ 0 & \text{für } r > C + 1. \end{cases}$$

Die Existenz einer solchen *Abschneidefunktion* (engl.: *cutoff function*) wird in Übungsaufgabe 2.2 gezeigt. Dort zeigen wir auch, dass die abgeänderte rechte Seite $\tilde{f}(x, y)$ global und bzgl. x gleichmäßig Lipschitz stetig ist.

Sowohl y als auch z lösen nun das abgeänderte Anfangswertproblem

$$\begin{aligned} y'(x) &= f(x, y(x)) = \tilde{f}(x, y(x)) \quad \forall x \in [a, \beta] \quad \text{und} \quad y(a) = y_0, \\ z'(x) &= f(x, z(x)) = \tilde{f}(x, z(x)) \quad \forall x \in [a, \beta] \quad \text{und} \quad z(a) = y_0. \end{aligned}$$

Da das abgeänderte AWP die Voraussetzungen von Satz 1.5 erfüllt, kann jedoch nur eine Lösung existieren, so dass $y(x) = z(x)$ für alle $x \in [a, \beta]$ gelten muss.

(b) Mit einem beliebigen $C > \|y_0\|^2$ definieren wir wie in (a) die abgeänderte rechte Seite $\tilde{f}(x, y)$. Dann existiert nach Satz 1.5 eine stetig differenzierbare Funktion $y : [a, b] \to \mathbb{R}^d$, die das abgeänderte AWP

$$y'(x) = \tilde{f}(x, y(x)) \quad \forall x \in [a, b] \quad \text{und} \quad y(a) = y_0$$

löst. Da y stetig ist und $y(a) = y_0$, existiert ein nicht-leeres Intervall $[a, \beta] \subseteq [a, b]$ mit $\|y\|^2 \leq C$ und auf diesem gilt

$$y'(x) = \tilde{f}(x, y(x)) = f(x, y(x)) \quad \forall x \in [a, \beta] \quad \text{und} \quad y(a) = y_0,$$

so dass das ursprüngliche AWP auf $[a, \beta]$ gelöst wird. □

Bemerkung 1.14
Man kann zeigen, dass für jedes $C > \|y_0\|^2$ das Intervall in Satz 1.13 (b) so groß gewählt werden kann, dass entweder $[a, \beta] = [a, b]$ oder $\|y(\beta)\|^2 = C$. In dem beschränkten Gebiet $[a, b] \times B_{C^2}(0)$ im Graphenraum existiert daher stets solange eine Lösung, bis der Rand dieses Gebietes erreicht wird, entweder bzgl. der x oder der y-Komponente, vgl. die in der Vorlesung gemalten Skizzen.

Es kann dabei jedoch sein (wie in Beispiel 1.11(b)), dass die Lösung eine Singularität besitzt, und daher in der x-Komponente nie über ein gewisses Teilintervall hinauskommt sondern das Gebiet immer in der y-Richtung verlässt.

Bemerkung 1.15
Ein Endwertproblem

$$y'(x) = f(x, y(x)) \quad \forall x \in [a, b] \quad \text{und} \quad y(b) = y_{end}$$

ist offenbar mit der Transformation $z(x) := y(a + b - x)$ äquivalent zu dem Anfangswertproblem

$$z'(x) = \tilde{f}(x, z(x)) \quad \forall x \in [a, b] \quad \text{und} \quad z(a) = y_{end}$$

mit $\tilde{f}(x, z(x)) := -f(a + b - x, z(x))$. Die transformierte rechte Seite $\tilde{f}$ erfüllt die Voraussetzungen von Satz 1.5 bzw. Satz 1.13 genau dann, wenn die ursprüngliche rechte Seite f dies tut. Satz 1.5 und Satz 1.13 gelten also analog auch für Endwertprobleme.

Unter den Voraussetzungen von Satz 1.13 folgt damit insbesondere auch, dass zwei Lösungen einer Differentialgleichung $y_1, y_2 : [a, b] \to \mathbb{R}^d$,

$$y_i'(x) = f(x, y_i(x)) \quad \forall x \in [a, b],\ i = 1, 2,$$

die in einem Punkt $x \in [a, b]$ übereinstimmen, in allen Punkten $x \in [a, b]$ übereinstimmen müssen.

Stabilität Nun untersuchen wir wie sich eine Störung der Anfangswerte auf die Lösung auswirkt.

Satz 1.16
Es gelten die Voraussetzungen von Satz 1.13. y und z seien zwei Lösungen der gleichen DGL, aber mit verschiedenen Anfangswerten, also

$$\begin{aligned} y'(x) &= f(x, y(x)) \quad \text{für alle } x \in [a, b], \qquad y(a) = y_0, \\ z'(x) &= f(x, z(x)) \quad \text{für alle } x \in [a, b], \qquad z(a) = z_0. \end{aligned}$$

$L > 0$ sei die Lipschitz-Konstante aus den Voraussetzungen von Satz 1.13 für eine kompakte Menge, die $y(x)$ und $z(x)$ für alle $x \in [a, b]$ enthält.

Dann gilt

$$\|y(x) - z(x)\| \leq e^{L(x-a)} \|y_0 - z_0\| \quad \forall x \in [a, b].$$

Beweis: Betrachte die Differenz

$$s(x) := \|y(x) - z(x)\|^2 = (y(x) - z(x))^T (y(x) - z(x)).$$

Es ist

$$\begin{aligned} s'(x) &= 2(y'(x) - z'(x))^T (y(x) - z(x)) \\ &= 2(f(x, y(x)) - f(x, z(x)))^T (y(x) - z(x)) \\ &\leq 2 \|f(x, y(x)) - f(x, z(x))\| \, \|y(x) - z(x)\| \\ &\leq 2L \|y(x) - z(x)\|^2 = 2Ls(x). \end{aligned}$$

Im Fall $y_0 = z_0$ gilt nach Satz 1.5 $y(x) = z(x)$ für alle $x \geq a$ und die Behauptung ist bewiesen. Anderenfalls muss nach Bemerkung 1.15 $y(x) \neq z(x)$ für alle $x \in [a, b]$ gelten. Für alle $x \in [a, b]$ gilt dann

$$\frac{d}{dx} \ln s(x) = \frac{s'(x)}{s(x)} \leq 2L,$$

also

$$\ln s(x) = \int_a^x \frac{d}{dt} \ln s(t) \, \mathrm{d}t + \ln s(a) \leq 2L(x - a) + \ln s(a),$$

d.h.

$$\|y(x) - z(x)\|^2 = s(x) \leq e^{2L(x-a)} s(a) = e^{2L(x-a)} \|y_0 - z_0\|^2,$$

womit die Behauptung gezeigt ist. □

1.3 Erste Lösungsmethoden

1.3.1 Das Richtungsfeld

Zur anschaulichen Herleitung einfacher erster Lösungsmethoden betrachten wir ein skalares AWP, in dem wir die Lösung $y : [x_0, \infty) \to \mathbb{R}$ von

$$y'(x) = f(x, y(x)), \qquad y(x_0) = y_0 \in \mathbb{R}$$

suchen.

Wir können uns die DGL $y'(x) = f(x, y(x))$ durch das dazugehörige Richtungsfeld veranschaulichen: Zu jedem Punkt $(x, y) \in \mathbb{R}^2$ zeichnen wir einen Richtungspfeil mit Steigung $f(x, y)$, z.B. den Vektor $(1, f(x, y))^T$ (vgl. die in der Vorlesung gezeichnete Skizze). Eine Funktion löst die DGL genau dann, wenn an jedem Punkt durch den die Funktion geht, die Steigung der Funktion und die Steigung des Richtungspfeils übereinstimmen. Wir können die DGL zeichnerisch lösen, indem wir ausgehend vom Startwert (x, y_0) die Funktion passend zu den Richtungspfeilen zeichnen.

1.3.2 Explizites Euler-Verfahren

Basierend auf dieser zeichnerischen Idee gehen wir nun systematischer vor und entwickeln ein erstes numerisches Lösungsverfahren. Betrachte das allgemeine (vektorwertige) AWP

$$y'(x) = f(x, y(x)), \qquad y(x_0) = y_0 \in \mathbb{R}^d$$

auf dem Intervall $x \in [a, b]$, $a = x_0$. Wir diskretisieren das Intervall durch $n + 1$ Punkte

$$a = x_0 < x_1 < \ldots < x_n = b,$$

also unter Verwendung der *Schrittweite* $h_i := x_{i+1} - x_i$, $i = 0, \ldots, n - 1$. Beginnend mit x_0 (wo wir die Lösung kennen, $y(x_0) = y_0$) berechnen wir nun sukzessiv Approximationen $y_i \approx y(x_i)$.

Die einfachste Möglichkeit ist die *explizite Eulermethode*, bei der wir die Steigung im aktuellen Punkt verwenden, um die Approximation im nächsten Punkt zu berechnen:[1]

$$\begin{aligned}
y_1 &:= y_0 + h_0 f(x_0, y_0), \\
y_2 &:= y_1 + h_1 f(x_1, y_1), \\
&\vdots \\
y_{i+1} &:= y_i + h_i f(x_i, y_i), \\
&\vdots \\
y_n &:= y_{n-1} + h_{n-1} f(x_{n-1}, y_{n-1}).
\end{aligned}$$

[1] Hier und im Folgenden bezeichen wir mit $y_0, y_1, \ldots, y_n \in \mathbb{R}^d$ d-dimensionale Vektoren und nicht die Einträge eines Vektors.

Dies entspricht dem zeichnerischen Lösen der DGL im Richtungsfeld durch eine stückweise lineare Funktion, bei der die Steigung der Liniensegmente im *linken Punkt* mit dem Richtungsfeld übereinstimmt.

Das gleiche Verfahren erhalten wir auch durch Diskretisierung der DGL mit finiten Differenzen (Vorwärtsdifferenzenquotient):

$$\frac{y_{i+1} - y_i}{x_{i+1} - x_i} \approx \frac{y(x_{i+1}) - y(x_i)}{x_{i+1} - x_i} \approx y'(x_i) = f(x_i, y(x_i)) \approx f(x_i, y_i).$$

Das explizite Eulerverfahren heißt auch Vorwärts-Euler-Verfahren (engl.: foward Euler).

1.3.3 Implizites Euler-Verfahren

Bei der zeichnerischen Lösung der DGL im Richtungsfeld durch eine stückweise lineare Funktion, könnten wir auch versuchen die Funktion so zu wählen, dass die Steigung der Liniensegmente im *rechten Punkt* mit dem Richtungsfeld übereinstimmt.

Dann haben wir keine explizite Formel für die Berechnung von y_{i+1} aus y_i. Stattdessen ist y_{i+1} *implizit* gegeben als Lösung von

$$y_{i+1} = y_i + h_i f(x_{i+1}, y_{i+1}).$$

Bei diesem *impliziten Euler-Verfahren* müssen wir also für jedes $y_i \in \mathbb{R}^d$, $i = 2, \ldots, n$, ein (üblicherweise nicht-lineares) d-dimensionales Gleichungssytem lösen.

Wieder erhalten wir das Verfahren auch durch Diskretisierung der DGL mit finiten Differenzen, diesmal mit dem Rückwärtsdifferenzenquotienten:

$$\frac{y_{i+1} - y_i}{x_{i+1} - x_i} \approx \frac{y(x_{i+1}) - y(x_i)}{x_{i+1} - x_i} \approx y'(x_{i+1}) = f(x_{i+1}, y(x_{i+1})) \approx f(x_{i+1}, y_{i+1}).$$

Enstprechend heißt das implizite Euler-Verfahren auch Rückwärts-Euler-Verfahren (engl.: *backward* Euler).

Bemerkung 1.17
Wie wir noch sehen werden, besitzen implizite Verfahren für manche (die sogenannten steifen*) Differentialgleichungen so große Vorteile, dass sie den erhöhten Aufwand wert sind.*

1.3.4 Weitere explizite und implizite Methoden

Die anschauliche Idee, stückweise lineare Funktionen ins Richtungsfeld zu zeichnen, führt auf viele weitere explizite und implizite Methoden:

Implizite Mittelpunktsregel Zuerst zeichnen wir die Liniensegmente so, dass die Steigung der Segmente in ihrem *Mittelpunkt* mit dem Richtungsfeld übereinstimmt.

Wir bestimmen also erst $y_{i+1/2}$ durch Lösung der Gleichung

$$y_{i+1/2} = y_i + \frac{h_i}{2} f\left(\frac{x_{i+1} + x_i}{2}, y_{i+1/2}\right)$$

und setzen dann

$$y_{i+1} := y_i + h_i f\left(\frac{x_{i+1} + x_i}{2}, y_{i+1/2}\right).$$

Dies ist die sogenannte *implizite Mittelpunktsregel*, die auch als Kombination eines halben impliziten Eulerschrittes mit einem halben expliziten Eulerschritt interpretiert (und implementiert) werden kann.

Verfahren von Runge Wir können auch versuchen, eine explizite Formel für $y_{i+1/2}$ zu finden. Dafür setzen wir

$$y_{i+1/2} := y_i + \frac{h_i}{2} f(x_i, y_i)$$

und wählen dann

$$y_{i+1} := y_i + h_i f\left(\frac{x_{i+1} + x_i}{2}, y_{i+1/2}\right).$$

Dies ist das *Verfahren von Runge*. Beachte, dass dies nicht nur die Kombination zweier Halbschritte des expliziten Eulerverfahrens ist.

Crank-Nicolson-Methode Die nächste Methode erhalten wir durch die Forderung, dass die Steigung der Liniensegmente mit dem Mittelwert der Steigungen im Richtungsfeld im linken und rechten Randpunkt des Segments übereinstimmen soll:

$$y_{i+1} = y_i + h_i \frac{f(x_i, y_i) + f(x_{i+1}, y_{i+1})}{2}.$$

Dies ist die *Crank-Nicolson-Methode*.

Verfahren von Heun Eine explizite Alternative zur Crank-Nicolson Methode erhalten wir, indem wir y_{i+1} auf der rechten Seite der Crank-Nicolson-Formel durch einen Schritt mit dem expliziten Euler-Verfahren approximieren:

$$\begin{aligned} \eta &:= y_i + h_i f(x_i, y_i) \\ y_{i+1} &:= y_i + h_i \frac{f(x_i, y_i) + f(x_{i+1}, \eta)}{2}. \end{aligned}$$

1.4 Einschrittmethoden höherer Ordnung

1.4.1 Eine Generalvoraussetzung

In diesem Abschnitt sei $[a, b] \subset \mathbb{R}$, $b > a$ stets ein festes Intervall. Wir werden in dieser Vorlesung nur Differentialgleichungen

$$y'(x) = f(x, y(x))$$

mit unendlich oft differenzierbarer rechter Seite f betrachten. Insbesondere ist f also nach Lemma 1.12 lokal Lipschitz stetig.

Zur Untersuchung numerischer Lösungsmethoden gehen werden wir *zunächst* die folgende (unrealistisch restriktive!) Generalvoraussetzung annehmen und erst danach zeigen, wie sich die Ergebnisse auf den unbeschränkten Fall übertragen lassen. Wie sagen, dass die rechte Seite f die **Generalvoraussetzung** erfüllt, falls f unendlich oft differenzierbar ist und f sowie alle partiellen Ableitungen (beliebiger Ordnung und beliebiger Kombination von x- und y_i-Ableitungen) beschränkt sind, d.h. für jeden Multiindex $\alpha = (\alpha_x, \alpha_{y_1}, \ldots, \alpha_{y_d}) \in \mathbb{N}_0^{d+1}$ existiert eine Konstante $C_\alpha > 0$, so dass

$$\left| \frac{\partial^{|\alpha|}}{\partial^{\alpha_x} x \partial^{\alpha_{y_1}} y_1 \cdots \partial^{\alpha_{y_d}} y_1} f(x, y) \right| \leq C_\alpha \quad \text{für alle } (x, y) \in [a, b] \times \mathbb{R}^d.$$

Wie in Lemma 1.12 folgt daraus insbesondere, dass ein $L > 0$ existiert mit

$$\|f_y(x, y)\|_{\mathrm{F}} \leq L \quad \forall (x, y) \in [a, b] \times \mathbb{R}^d,$$

und damit für alle $x \in [a, b]$ und $y, z \in \mathbb{R}^d$ gilt, dass

$$\|f(x, z) - f(x, y)\| = \left\| \int_0^1 f_y(x, y + t(z - y))(z - y) \, \mathrm{d}t \right\| \leq L \|z - y\|.$$

Die Generalvoraussetung impliziert also globale Lipschitz-stetig und nach Abschnitt 1.2.2 ist somit für jede Wahl der Anfangswerte $x_0 \in [a, b]$ und $y_0 \in \mathbb{R}^d$ die eindeutige Existenz von Lösungen und ihre stetige Abhängigkeit von den Anfangswerten garantiert.

Außerdem kann man zeigen, dass dann jede Lösung und alle ihre Ableitungen eines AWP mit rechter Seite f beschränkt sind. Präzise ausgedrückt: Zu jeder rechten Seite f, die die Generalvoraussetung erfüllt, existieren Konstanten $C_k > 0$ ($k \in \mathbb{N}$), sodass für jede Wahl der Anfangswerte $x_0 \in [a, b]$ und $y_0 \in \mathbb{R}^d$ die zugehörige Lösung von

$$y'(x) = f(x, y(x)), \qquad y(x_0) = y_0 \in \mathbb{R}^d$$

erfüllt, dass

$$\sup_{x \in [x_0, b]} \left\| y^{(k)}(x) \right\| \leq C_k.$$

Diese Generalvoraussetzung ist in der Praxis üblicherweise nicht erfüllt und bereits bei einfachsten Beispielen (wie $y' = ry$) verletzt. Sie erleichtert uns jedoch die Konvergenzanalyse und wir werden im Abschnitt 1.4.6 zeigen, dass alle unter der Generalvoraussetzung erhaltenen Ergebnisse, auch für den praxisrelevanteren allgemeineren Fall gelten, dass die rechte Seite unendlich oft differenzierbar ist und eine Lösung auf dem betrachteten Intervall existiert.

1.4.2 Konsistenz und Konvergenz

Zur Lösung des AWP

$$y'(x) = f(x, y(x)) \quad \forall x \in [a, b], \qquad y(a) = y_0 \in \mathbb{R}^d,$$

diskretisieren wir das Intervall in $n + 1$ Punkte

$$a = x_0 < x_1 < \ldots < x_n = b.$$

Beginnend mit x_0 (wo wir die Lösung $y(x_0) = y_0$ kennen) versuchen wir sukzessive Approximationen $y_i \approx y(x_i) \in \mathbb{R}^d$ zu bestimmen.

Einschrittmethoden: Nur der vorherige Punkt y_i (und x_i, x_{i+1}, $h_i := x_{i+1} - x_i$) wird zur Berechnung von y_{i+1} verwendet.

Mehrschrittmethoden: Mehrere vorherige Punkte y_i, y_{i-1}, $\ldots y_{i-m}$ (und x_{i+1}, x_i, $\ldots$, x_{i-m}) werden zur Berechnung von y_{i+1} verwendet. (Eine $m + 1$-Schritt Methode benötigt *Startprozedur* zur Berechnung der ersten m Werte y_1,$\ldots$, y_m.)

Die Methoden in Abschnitt 1.3 sind allesamt Einschrittmethoden.

Definition 1.18
Eine Einschrittmethode heißt konsistent*, falls für jede (unsere Generalvoraussetzung erfüllende) rechte Seite f gilt, dass*

$$\lim_{h\to 0} \sup_{x_i\in[a,b], y_i\in\mathbb{R}^d} \frac{|y_{i+1} - y(x_i + h)|}{h} = 0$$

wobei y die Lösung des AWP

$$y'(x) = f(x, y(x)), \qquad y(x_i) = y_i$$

ist, und y_{i+1} durch Anwendung der Methode auf y_i mit Schrittweite h erzeugt wurde.

Die Methode besitzt Konsistenzordnung $p \in \mathbb{N}$*, falls für jede (unsere Generalvoraussetzung erfüllende) rechte Seite f*

$$\limsup_{h\to 0} \sup_{x_i\in[a,b], y_i\in\mathbb{R}^d} \frac{|y_{i+1} - y(x_i + h)|}{h^{p+1}} < \infty,$$

d.h. Konstanten $C > 0$, $h_0 > 0$ existieren, sodass

$$\sup_{x_i\in[a,b], y_i\in\mathbb{R}^d} |y_{i+1} - y(x_i + h)| \le Ch^{p+1} \quad \forall 0 < h < h_0.$$

Eine Methode der Konsistenzordnung p macht in jedem Intervall $[x_i, x_{i+1}]$ einen *lokalen Fehler* der Größenordnung h^{p+1}. Da es $n = \frac{b-a}{h}$ solche Intervalle gibt, erwarten wir dass der *globale Fehler* in der Größenordnung h^p liegen wird. Der folgende Satz zeigt, dass dies tatsächlich der Fall ist.

Satz 1.19
Sei $y : [a, b] \to \mathbb{R}^d$ die Lösung des AWP

$$y'(x) = f(x, y(x)) \quad \forall x \in [a, b], \qquad y(a) = y_0 \in \mathbb{R}^d$$

Wir betrachten die Anwendung einer Einschrittmethode mit einer Diskretisierung mit Höchstschrittweite $h > 0$,

$$a = x_0 < x_1 < \ldots < x_n = b, \qquad x_i - x_{i-1} \le h \quad (i = 1, \ldots, n),$$

für die ein $c > 0$ existiert[2] *mit $nh \le c(b - a)$.*

[2]Dies ist erfüllt, wenn $h_{\min} \le x_i - x_{i-1} \le h$ für alle $i = 1, \ldots, n$ und der Quotient $h/h_{\min}$ für $h \to 0$ durch c beschränkt bleibt. Für äquidistante Gitter gilt offenbar $c = 1$.

(a) Ist die Methode konsistent, so gilt

$$\max_{i=0,\dots,n} \|y_i - y(x_i)\| \to 0 \quad \text{für } h \to 0.$$

(b) Besitzt die Methode Konsistenzordnung p, so gilt

$$\max_{i=0,\dots,n} \|y_i - y(x_i)\| \leq \frac{e^{cL(b-a)} - 1}{L} Ch^p \quad \forall 0 < h < h_0,$$

wobei $C, h_0 > 0$ die Konstanten aus der Definition der Konsistenzordnung und L die Lipschitz-Konstante aus der Generalvoraussetzung ist.

Beweis: Wir beginnen mit (b). Für die Höchstschrittweite gelte $h < h_0$. Da wir mit dem korrekten Startwert $y_0 = y(x_0)$ beginnen, gilt für den Fehler nach dem ersten Schritt

$$\|y_1 - y(x_1)\| \leq Ch^{p+1}.$$

Im nächsten Schritt, bei dem wir y_2 aus y_1 berechnen, gibt es zwei Fehlerquellen:

(i) Die Berechnung von y_2 aus y_1 mit dem Einschrittverfahren entspricht der Anwendung des Verfahrens auf das (wegen unserer Generalvoraussetzung eindeutig lösbare) AWP

$$z'(x) = f(x, z(x)), \qquad z(x_1) = y_1 \in \mathbb{R}^d \tag{1.3}$$

Dabei macht das Verfahren den Fehler

$$\|z(x_2) - y_2\| \leq Ch^{p+1}.$$

(ii) Da y_1 nur eine Approximation an $y(x_1)$ ist, stimmt die Lösung $z(x)$ von (1.3) nicht mit $y(x)$ überein. Nach Satz 1.16 gilt aber

$$\|y(x_2) - z(x_2)\| \leq e^{L(x_2 - x_1)} \|y(x_1) - y_1\|.$$

Insgesamt erhalten wir also

$$\begin{aligned} \|y_2 - y(x_2)\| &\leq \|y_2 - z(x_2)\| + \|z(x_2) - y(x_2)\| \\ &\leq Ch^{p+1} + e^{Lh} \|y_1 - y(x_1)\| \\ &\leq (1 + e^{Lh}) Ch^{p+1}. \end{aligned}$$

Mit trivialer Induktion erhalten wir für alle $i = 1, \ldots, n$:

$$\begin{aligned}\|y_i - y(x_i)\| &\leq Ch^{p+1} + e^{Lh} \|y_{i-1} - y(x_{i-1})\| \\ &\leq Ch^{p+1} + e^{Lh} \left(Ch^{p+1} + e^{Lh} \|y_{i-2} - y(x_{i-2})\|\right) \\ &\leq \ldots \leq \left(1 + e^{Lh} + e^{2Lh} + e^{(i-1)Lh}\right) Ch^{p+1} \\ &\leq \sum_{j=0}^{n-1} \left(e^{Lh}\right)^j Ch^{p+1} = \frac{(e^{Lh})^n - 1}{e^{Lh} - 1} Ch^{p+1}.\end{aligned}$$

und mit $e^{Lh} \geq 1 + Lh$ folgt

$$\max_{i=0,\ldots,n} \|y_i - y(x_i)\| \leq \frac{e^{nhL} - 1}{Lh} Ch^{p+1} \leq \frac{e^{cL(b-a)} - 1}{L} Ch^p.$$

Der Beweis von (a) geht analog. □

Bemerkung 1.20
Im Folgenden verwenden wir im Zusammenhang mit Anfangswertproblemen die Landau-Notation $O(h^p)$, $o(h)$, etc., mit der Konvention, dass die darin vorkommenden Konstanten von der rechten Seite f, nicht jedoch von dem Anfangswert $x_0 \in [a, b]$, $y_0 \in \mathbb{R}^d$ abhängen dürfen. Mit dieser Konvention ist ein Einschrittverfahren

- *konsistent, falls aus $y_i = y(x_i)$ folgt, dass*

$$y(x_{i+1}) - y_{i+1} = o(h).$$

- *konsistent mit Ordnung p, falls aus $y_i = y(x_i)$ folgt, dass*

$$y(x_{i+1}) - y_{i+1} = O(h^{p+1}).$$

Beispiel 1.21
(a) ***Explizites Euler-Verfahren:*** *Für $y(x_i) = y_i$ erhalten wir durch Taylorentwicklung*

$$\|y(x_{i+1}) - y(x_i) - hy'(x_i)\| \leq \frac{h^2}{2} \max_{\xi \in [x_i, x_{i+1}]} \|y''(\xi)\| \leq \frac{h^2}{2} C_2$$

mit einer (aufgrund unserer Generalvoraussetzung nur von der rechten Seite f abhängigen) Konstante C_2. Es ist also mit obiger Konvention

$$\begin{aligned}y(x_{i+1}) &= y(x_i) + hy'(x_i) + O(h^2) \\ &= y(x_i) + hf(x_i, y_i) + O(h^2) = y_{i+1} + O(h^2)\end{aligned}$$

und das explizite Euler-Verfahren besitzt somit Konsistenzordnung 1.

(b) ***Implizites Euler-Verfahren:*** *Für $y(x_i) = y_i$ erhalten wir wiederum durch Taylorentwicklung*

$$\begin{aligned} y(x_i) &= y(x_{i+1}) - hy'(x_{i+1}) + O(h^2) \\ &= y(x_{i+1}) - hf(x_{i+1}, y(x_{i+1})) + O(h^2). \end{aligned}$$

Zusammen mit $y_{i+1} = y_i + hf(x_{i+1}, y_{i+1})$ erhalten wir

$$\begin{aligned} \|y_{i+1} - y(x_{i+1})\| &= \|y_i + hf(x_{i+1}, y_{i+1}) - y(x_{i+1})\| \\ &= h\,\|f(x_{i+1}, y_{i+1}) - f(x_{i+1}, y(x_{i+1}))\| + O(h^2) \\ &\leq hL\,\|y_{i+1} - y(x_{i+1})\| + O(h^2). \end{aligned}$$

Für hinreichend kleine h gilt also

$$\|y_{i+1} - y(x_{i+1})\| = \frac{1}{1 - hL} O(h^2) = O(h^2).$$

Das implizite Euler-Verfahren besitzt also Konsistenzordnung 1.

1.4.3 Runge-Kutta-Methoden

Wir betrachten nun einen allgemeinen Ansatz um Einschrittmethoden hoher Konsistenzordnung zu konstruieren. Im ersten Schritt (mit $h = x_1 - x_0$) soll gelten

$$y_1 \approx y(x_1) = y_0 + \int_{x_0}^{x_1} y'(t)\,\mathrm{d}t = y_0 + \int_{x_0}^{x_1} f(t, y(t))\,\mathrm{d}t.$$

Durch Approximation des Integrals auf der rechten Seite durch eine Quadraturformel erhalten wir

$$\int_{x_0}^{x_1} f(t, y(t))\,\mathrm{d}t \approx h \sum_{j=1}^{s} b_j f(x_0 + c_j h, \eta_j).$$

Die Quadraturformel sollte zumindest konstante Funktionen exakt integrieren, deshalb fordern wir

$$\sum_{j=1}^{s} b_j = 1.$$

Die c_j heißen *Knoten*, b_j *Gewichte* und s heißt *Stufenzahl* des Verfahrens.

Dieser Ansatz verallgemeinert die Ideen aus Abschnitt 1.3.4, indem nun ein gewichtetes Mittel aus s unterschiedlichen Steigungen im Richtungsfeld an

den Punkten $(x_0 + c_j h, \eta_j)$, $j = 1, \ldots, s$ verwendet wird. Für das explizite Eulerverfahren ist $s = 1$, $c_1 = 0$ und $\eta_1 = y_0$.

b_j und c_j sollten aus den Knoten und Gewichten eines möglichst guten Quadraturverfahrens bestimmt werden. Zur Wahl der η_j fordern wir, dass

$$\eta_j \approx y(x_0 + c_j h) = y_0 + \int_{x_0}^{x_0 + c_j h} y'(t)\,\mathrm{d}t = y_0 + \int_{x_0}^{x_0 + c_j h} f(t, y(t))\,\mathrm{d}t.$$

Wir wenden wiederum ein Quadraturverfahren für die Integrale auf der rechten Seite an und verwenden dabei die gleichen Quadraturpunkte wie für das erste Integral, d.h. für $j = 1, \ldots, s$ verwenden wir

$$\int_{x_0}^{x_0 + c_j h} f(t, y(t))\,\mathrm{d}t \approx h \sum_{l=1}^{s} a_{jl} f(x_0 + c_l h, \eta_l).$$

Wiederum sollten die Quadraturformeln zumindest konstante Funktionen exakt integrieren, deshalb fordern wir (für alle $j = 1, \ldots, s$)

$$\sum_{l=1}^{s} a_{jl} = c_j.$$

So erhalten wir die allgemeinen *Runge-Kutta-Methoden*:

Gegeben $a_{jl}, b_j, c_j \in \mathbb{R}$, $j = 1, \ldots, s$, $l = 1, \ldots, s$ mit

$$\sum_{j=1}^{s} b_j = 1 \quad \text{und} \quad \sum_{l=1}^{s} a_{jl} = c_j \quad \text{für alle } j = 1, \ldots, s.$$

- Bestimme $\eta_j \in \mathbb{R}^d$, $j = 1, \ldots, s$ aus

$$\eta_j = y_i + h_i \sum_{l=1}^{s} a_{jl} f(x_i + c_l h_i, \eta_l), \quad j = 1, \ldots, s. \tag{1.4}$$

- Setze

$$y_{i+1} := y_i + h_i \sum_{j=1}^{s} b_j f(x_i + c_j h_i, \eta_j).$$

Runge-Kutta-Methoden können explizit oder implizit sein. Ist

$$a_{jl} = 0 \quad \text{für } l \geq j$$

dann ist $\eta_1 = y_i$, η_2 kann aus η_1 berechnet werden, usw. (explizite Runge-Kutta-Methoden). Ansonsten ist (1.4) ein System aus sd Gleichungen für die sd unbekannten Einträge der d-dimensionalen Vektoren $\eta_j \in \mathbb{R}^d$, $j = 1, \ldots, s$ (implizite Runge-Kutta-Methoden).

Eine äquivalente Formulierung erhalten wir durch

$$k_j := f(x_i + c_j h_i, \eta_j).$$

Gegeben $a_{jl}, b_j, c_j \in \mathbb{R}$, $j = 1, \ldots, s$, $l = 1, \ldots, s$ mit

$$\sum_{j=1}^{s} b_j = 1 \quad \text{und} \quad \sum_{l=1}^{s} a_{jl} = c_j \quad \text{für alle } j = 1, \ldots, s.$$

- Bestimme $k_j \in \mathbb{R}^d$, $j = 1, \ldots, s$ aus

$$k_j = f(x_i + c_j h, y_i + h_i \sum_{l=1}^{s} a_{jl} k_l), \quad j = 1, \ldots, s.$$

- Setze

$$y_{i+1} := y_i + h_i \sum_{j=1}^{s} b_j k_j.$$

Die Koeffizienten $A = (a_{jl}) \in \mathbb{R}^{s \times s}$, $b = (b_j) \in \mathbb{R}^s$ und $c = (c_j) \in \mathbb{R}^s$ einer Runge-Kutta-Methode lassen sich im sogenannten *Butcher Tableau* zusammenfassen:

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array} \qquad \begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \ldots & a_{1s} \\ c_2 & a_{21} & a_{22} & \ldots & a_{2s} \\ \vdots & \vdots & \vdots & & \vdots \\ c_s & a_{s1} & a_{s2} & \ldots & a_{ss} \\ \hline & b_1 & b_2 & \ldots & b_s \end{array}$$

Mit dieser Notation erhalten wir für das explizite und implizite Euler-Verfahren:

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array} \qquad \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

1.4.4 Wohldefiniertheit impliziter Methoden

Die Vorteile impliziter Runge-Kutta-Methoden werden wir erst in Abschnitt 1.5 kennenlernen. Wir zeigen aber an dieser Stelle schon, dass das nicht-

lineare Gleichungssystem (1.4) für hinreichend kleine Schrittweiten eindeutig lösbar ist.

Satz 1.22
Zu jeder rechten Seite f (die die Generalvoraussetzung erfüllt) und jedem Runge-Kutta-Verfahren (A, b, c) existiert eine Schrittweite $h_0 > 0$, sodass für jedes $x \in [a, b]$, $y \in \mathbb{R}^d$ und $0 < h \leq h_0$ das Gleichungsystem für die η_j

$$\eta_j = y + h \sum_{l=1}^{s} a_{jl} f(x + c_l h, \eta_l), \quad j = 1, \ldots, s, \tag{1.5}$$

eindeutig lösbar ist.

Beweis: Wir schreiben das Gleichungssystem (1.5) als Fixpunktgleichung

$$\eta = \Phi(\eta)$$

mit

$$\eta := \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_s \end{pmatrix} \in \mathbb{R}^{ds}, \quad \Phi(\eta) := \begin{pmatrix} \Phi_1(\eta) \\ \vdots \\ \Phi_s(\eta) \end{pmatrix} \in \mathbb{R}^{ds}$$

und

$$\Phi_j(\eta) := y + h \sum_{l=1}^{s} a_{jl} f(x + c_l h, \eta_l) \in \mathbb{R}^d.$$

Es ist

$$\begin{aligned} \Phi'(\eta) &= \begin{pmatrix} \Phi_1'(\eta) \\ \vdots \\ \Phi_s'(\eta) \end{pmatrix} \in \mathbb{R}^{ds \times ds} \\ \Phi_j'(\eta) &= \begin{pmatrix} \frac{\partial \Phi_j}{\partial \eta_1} & \ldots & \frac{\partial \Phi_j}{\partial \eta_s} \end{pmatrix} \in \mathbb{R}^{d \times ds} \\ \frac{\partial \Phi_j}{\partial \eta_l} &= h a_{jl} f_y(x + c_l h, \eta_l) \in \mathbb{R}^{d \times d}. \end{aligned}$$

Aufgrund unserer Generalvoraussetung und der Äquivalenz aller Normen auf dem $\mathbb{R}^{d \times d}$ existiert ein $C > 0$, so dass für alle $j, l = 1, \ldots, d$ jeder Eintrag der Matrix $\frac{\partial \Phi_j}{\partial \eta_l}$ durch Ch beschränkt ist. Damit ist jeder Eintrag von $\Phi'(\eta)$ durch Ch beschränkt und (wegen der Äquivalenz aller Normen auf dem $\mathbb{R}^{ds \times ds}$) existiert ein $C' > 0$ mit

$$\|\Phi'(\eta)\| \leq C' h \quad \text{für alle } h > 0,\ \eta \in \mathbb{R}^{ds}.$$

Für hinreichend kleines h_0 gilt daher

$$\|\Phi'(\eta)\| \leq \frac{1}{2} \quad \text{für alle } 0 < h \leq h_0,\ \eta \in \mathbb{R}^{ds}$$

und Φ ist eine Kontraktion, da

$$\begin{aligned}\left\|\Phi(\eta^{(2)}) - \Phi(\eta^{(1)})\right\| &= \left\|\int_0^1 \Phi'(\eta^{(1)} + t(\eta^{(2)} - \eta^{(1)}))(\eta^{(2)} - \eta^{(1)})\,\mathrm{d}t\right\| \\ &\leq \frac{1}{2}\left\|\eta^{(2)} - \eta^{(1)}\right\|.\end{aligned}$$

Aus dem Banachschen Fixpunktsatzes (Satz 1.6) folgt daher für $0 < h \leq h_0$ die eindeutige Lösbarkeit der Fixpunktgleichung $\Phi(\eta) = \eta$ und damit die eindeutige Lösbarkeit des Gleichungssystems (1.5). □

1.4.5 Runge-Kutta-Ordnungsbedingungen

Im letzten Abschnitt haben wir gesehen, dass jede Wahl der Runge-Kutta-Koeffizienten (A, b, c) für hinreichend kleine Schrittweiten auf lösbare implizite (oder sogar explizite) Gleichungssysteme führt, das Verfahren also für jede Wahl der Koeffizienten durchführbar ist. Jetzt wenden wir uns der Frage zu, wie die Koeffizienten gewählt werden müssen, um ein Verfahren möglichst hoher Ordnung zu erhalten.

Satz 1.23
Seien $A = (a_{ij})_{i,j=1,\dots,s}$, $b = (b_i)_{i=1,\dots,s}$ und $c = (c_i)_{i=1,\dots,s}$ die Koeffizienten eines Runge-Kutta-Verfahrens.

(a) Aus

$$\sum_{j=1}^{s} b_j = 1 \quad \textit{und} \quad \sum_{k=1}^{s} a_{jk} = c_j \quad (j = 1, \dots, s)$$

folgt, dass das Verfahren (mindestens) Konsistenzordnung 1 besitzt.

(b) Das Verfahren hat genau dann mindestens Konsistenzordnung 2, wenn zusätzlich gilt, dass

$$\sum_{j=1}^{s} b_j c_j = \frac{1}{2}.$$

(c) Das Verfahren hat genau dann mindestens Konsistenzordnung 3, wenn zusätzlich gilt, dass

$$\sum_{j=1}^{s} b_j c_j^2 = \frac{1}{3} \quad \textit{und} \quad \sum_{j=1}^{s} b_j \sum_{k=1}^{s} a_{jk} c_k = \frac{1}{6}.$$

Beweis: Wegen Übungsaufgabe 4.2 genügt es, die Behauptung für *autonome* Differentialgleichungen

$$y' = f(y), \quad f : \mathbb{R}^d \to \mathbb{R}^d$$

zu beweisen.

Sei y eine Lösung der DGL, $x_i \in [a, b]$, $y_i = y(x_i)$, $h = x_{i+1} - x_i$ und y_{i+1} die durch das Runge-Kutta-Verfahren erzielte Näherung. Nach Übungsaufgabe 3.3 gilt

$$y(x_{i+1}) = y_i + hf + 1/2h^2 f'f + 1/6h^3(f^T f''f + (f')^2 f) + O(h^4)$$

wobei wir das Argument (y_i) bei f und seinen Ableitungen zur Vereinfachung der Schreibweise weglassen.

Genauso entwickeln wir die Näherung y_{i+1}. Dabei verwenden wir immer wieder

$$\eta_j = y_i + h \sum_{k=1}^{s} a_{jk} f(\eta_k), \qquad \sum_{k=1}^{s} a_{jk} = c_j. \tag{1.6}$$

Zuerst erhalten wir damit

$$\eta_j = y_i + O(h), \quad \forall j = 1, \ldots, s,$$

und daher

$$f(\eta_j) = f(y_i) + f'(y_i)(\eta_j - y_i) = f + O(h) \quad \forall j = 1, \ldots, s,$$

da f' nach unserer Generalvoraussetzung beschränkt ist.

Nochmalige Anwendung von (1.6) führt zu

$$\eta_j - y_i = h \sum_{k=1}^{s} a_{jk} f(\eta_k) = h \sum_{k=1}^{s} a_{jk} f(y_i) + O(h^2) = hc_j f(y_i) + O(h^2).$$

Ein weiteres Mal verwenden wir (1.6) und kombinieren es mit

$$f(\eta_k) = f(y_i) + f'(y_i)(\eta_k - y_i) + O(h^2).$$

So erhalten wir

$$\begin{aligned}
\eta_j &= y_i + h\sum_{k=1}^{s} a_{jk} f(\eta_k) \\
&= y_i + h\sum_{k=1}^{s} a_{jk}(f(y_i) + f'(y_i)(\eta_k - y_i) + O(h^2)) \\
&= y_i + h\sum_{k=1}^{s} a_{jk}(f(y_i) + f'(y_i)\left(hc_k f(y_i) + O(h^2)\right) + O(h^2)) \\
&= y_i + hc_j f(y_i) + h^2 f'(y_i) f(y_i) \sum_{k=1}^{s} a_{jk} c_k + O(h^3).
\end{aligned}$$

Mit

$$y_{i+1} := y_i + h\sum_{j=1}^{s} b_j f(\eta_j), \quad \sum_{j=1}^{s} b_j = 1$$

folgt (wobei wir das Argument (y_i) bei f und seinen Ableitungen weglassen)

$$\begin{aligned}
y_{i+1} &= y_i + h\sum_{j=1}^{s} b_j f(\eta_j) \\
&= y_i + h\sum_{j=1}^{s} b_j \left(f + f'(\eta_j - y_i) + \frac{1}{2}(\eta_j - y_i)^T f''(\eta_j - y_i) + O(h^3) \right)
\end{aligned}$$

und damit

$$\begin{aligned}
y_{i+1} &= y_i + hf + hf' \sum_{j=1}^{s} b_j(\eta_j - y_i) + \frac{1}{2}h\sum_{j=1}^{s} b_j(\eta_j - y_i)^T f''(\eta_j - y_i) + O(h^4) \\
&= y_i + hf + hf' \sum_{j=1}^{s} b_j \left(hc_j f + h^2 f' f \sum_{k=1}^{s} a_{jk} c_k + O(h^3) \right) \\
&\quad + \frac{1}{2}h\sum_{j=1}^{s} b_j \left(hc_j f + O(h^2)\right)^T f'' \left(hc_j f + O(h^2)\right) + O(h^4) \\
&= y_i + hf + h^2 f' f \sum_{j=1}^{s} b_j c_j + h^3 (f')^2 f \sum_{j=1}^{s} b_j \sum_{k=1}^{s} a_{jk} c_k \\
&\quad + \frac{1}{2}h^3 f^T f'' f \sum_{j=1}^{s} b_j c_j^2 + O(h^4).
\end{aligned}$$

Die Behauptung folgt nun aus dem Vergleich der Entwicklungen von $y(x_{i+1})$ und y_{i+1}. □

Bemerkung 1.24

(a) Mit Satz 1.23 lässt sich die Ordnung der Verfahren in 1.3.4 bestimmen.

(b) Mit einem systematischeren symbolischen *Ansatz können Methoden beliebig hoher Ordnung konstruiert werden.*

(c) Man kann zeigen, dass die Ordnung eines s-stufigen Runge-Kutta-Verfahrens höchstens $2s$ ist. Explizite Runge-Kutta-Verfahren können höchstens Ordnung s haben (siehe Abschnitt 1.5.5).

(d) Die in der Praxis wohl am häufigsten verwendete explizite Runge-Kutta-Methode ist eine Methode 5. Ordnung von Dormand und Prince*, die durch folgendes Tableau gegeben ist:*

0						
$\frac{1}{5}$	$\frac{1}{5}$					
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$				
$\frac{4}{5}$	$\frac{44}{45}$	$-\frac{56}{15}$	$\frac{32}{9}$			
$\frac{8}{9}$	$\frac{19372}{6561}$	$-\frac{25360}{2187}$	$\frac{64448}{6561}$	$-\frac{212}{729}$		
1	$\frac{9017}{3168}$	$-\frac{355}{33}$	$\frac{46732}{5247}$	$\frac{49}{176}$	$-\frac{5103}{18656}$	
	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$

Dieses Verfahren ist (in Kombination mit einer zur adaptiven Schrittweitensteuerung verwendeten Methode 4. Ordnung) unter dem Namen `dopri5` *oder* `ode45` *in vielen Programmpaketen der Standardlöser für Anfangswertprobleme.*

1.4.6 Wohldefiniertheit und Konvergenz ohne Generalvoraussetzung

Wir haben in den letzten Abschnitten die Wohldefiniertheit und Konvergenz von Einschrittmethoden nur unter der für praktische Anwendungen meist zu restriktiven Generalvoraussetzung untersucht, dass die rechte Seite des Anfangswertproblems und ihre partiellen Ableitungen jeder Ordnung beschränkt sind. In Beispiel 1.11 und Bemerkung 1.14 haben wir gesehen, dass es für rechte Seiten mit unbeschränkter Ableitung vorkommen kann, dass das Problem nur auf einem Teil des betrachteten Intervalls lösbar ist. Ist die Lösbarkeit jedoch auf dem gesamten Intervall sichergestellt, so übertragen sich die Konvergenzresultate aus den letzten Abschnitten. Wir fassen dies im folgenden Satz zusammen.

Satz 1.25
Seien $a, b \in \mathbb{R}$, $b > a$ und $y_0 \in \mathbb{R}^d$. Auf das Anfangswertproblem

$$y'(x) = f(x, y(x)) \quad \text{für } x \in [a, b], \quad y(a) = y_0 \in \mathbb{R}^d,$$

mit beliebig oft stetig differenzierbarer rechter Seite

$$f : \ [a, b] \times \mathbb{R}^d \to \mathbb{R}^d$$

werde ein Runge-Kutta-Verfahren auf dem Gitter

$$a = x_0 < x_1 < \ldots < x_n = b$$

mit Höchstschrittweite $h = \max_{i=1,\ldots,n}(x_i - x_{i-1})$ angewendet, wobei wie in Satz 1.19 ein $c > 0$ existiere mit mit $nh \leq c(b - a)$.

Falls eine Lösung $y : [a, b] \to \mathbb{R}^d$ dieses AWP existiert[3]*, so gilt:*

(a) Ist das RKV explizit und besitzt Konsistenzordnung $p \in \mathbb{N}$, so erfüllen die Iterierten

$$\max_{i=0,\ldots,n} \|y_i - y(x_i)\| = O(h^p).$$

(b) Ist das RKV implizit und besitzt Konsistenzordnung $p \in \mathbb{N}$, so existiert für hinreichend kleine $h > 0$ eine Lösung der impliziten Gleichungen (1.4), sodass für die zugehörigen Iterierten gilt:

$$\max_{i=0,\ldots,n} \|y_i - y(x_i)\| = O(h^p).$$

Beweis: Wir gehen analog zum Beweis von Satz 1.13 vor. Sei $y : [a, b] \to \mathbb{R}^d$ die Lösung des Anfangswertproblems. Sei

$$C := 1 + \max_{x \in [a,b]} \|y(x)\|^2 .$$

Wir ersetzen die rechte Seite f des AWP durch eine abgeänderte rechte Seite $\tilde{f}(x, \eta) = f(x, \eta)\varphi(\|\eta\|^2)$ mit der Abschneidefunktion φ aus Übungsaufgabe 2.2. Dann ist $\tilde{f}$ weiterhin beliebig oft stetig differenzierbar, und wegen

$$\tilde{f}(x, \eta) = \begin{cases} f(x, \eta) & \text{für } \|\eta\|^2 \leq C, \\ 0 & \text{für } \|\eta\|^2 > C + 1, \end{cases}$$

erfüllt $\tilde{f}$ die Generalvoraussetzung.

[3] also insbesondere dann, wenn f global und bzgl. x gleichmäßig Lipschitz stetig ist oder sogar die Generalvoraussetzung aus Abschnitt 1.4.1 erfüllt

Dann ist y auch die (nach Satz 1.5 eindeutige) Lösung des Anfangswertproblems

$$y'(x) = \tilde{f}(x, y(x)) \quad \text{für } x \in [a, b], \quad y(a) = y_0 \in \mathbb{R}^d. \tag{1.7}$$

Bei Anwendung des RKV auf (1.7) sind für hinreichend kleine Schrittweiten die Iterierten $\tilde{y}_0, \ldots, \tilde{y}_n$ wohldefiniert und

$$\max_{i=0,\ldots,n} \|\tilde{y}_i - y(x_i)\| = O(h^p).$$

Insbesondere ist, für hinreichend kleine $h > 0$, $\|\tilde{y}_i\|^2 \leq C$. Außerdem folgt wie im Beginn vom Beweis von Satz 1.23 aus der Beschränktheit von $\tilde{f}$, dass für alle Zwischenwerte $\tilde{\eta}_j$ $(j = 1, \ldots, s)$ des i-ten Schrittes des auf das abgeänderte AWP angewandten Verfahrens gilt:

$$\tilde{\eta}_j = \tilde{y}_i + h_i \sum_{l=1}^{s} a_{jl} \tilde{f}(x + c_l h_i, \tilde{\eta}_l) = \tilde{y}_i + O(h).$$

Für hinreichend kleine $h > 0$ gilt daher auch $\|\tilde{\eta}_j\|^2 \leq C$ in jedem Schritt des Verfahrens und es folgt, dass

$$\tilde{\eta}_j = \tilde{y}_i + h_i \sum_{l=1}^{s} a_{jl} \tilde{f}(x + c_l h_i, \tilde{\eta}_l) = \tilde{y}_i + h_i \sum_{l=1}^{s} a_{jl} f(x + c_l h_i, \tilde{\eta}_l)$$

$$\tilde{y}_{i+1} = \tilde{y}_i + h_i \sum_{j=1}^{s} b_j \tilde{f}(x_i + c_j h_i, \tilde{\eta}_j) = \tilde{y}_i + h_i \sum_{j=1}^{s} b_j f(x_i + c_j h_i, \tilde{\eta}_j)$$

Die $\tilde{\eta}_j$ lösen also auch das Gleichungssystem zum nicht abgeänderten AWP und die damit bestimmten Iterierten y_i zum nicht abgeänderten AWP stimmen mit den Iterierten $\tilde{y}_i$ zum abgeänderten AWP überein. □

Bemerkung Durch Satz 1.25 ist für implizite Verfahren nur garantiert, dass *eine* Lösung $\eta_1, \ldots, \eta_s$ zu konvergenten Iterierenden führt. Die Lösung ist jedoch im Allgemeinen nicht eindeutig, wie das folgende Beispiel zeigt.

Das sklare AWP

$$y'(x) = y^2(x), \quad y(0) = 0,$$

besitzt offenbar die (nach Satz 1.13 eindeutige) Lösung $y(x) = 0$. Bei Anwendung des impliziten Euler-Verfahrens ergibt sich im ersten Schritt mit Schrittweite $h > 0$

$$y_1 = h y_1^2,$$

was für jedes $h > 0$ die zwei Lösungen $y_1 = 0$ und $y_1 = 1/h$ besitzt. Die impliziten Gleichungen sind also nicht eindeutig lösbar und es existieren Lösungen der impliziten Gleichung, für die die zugehören Iterierten nicht gegen die Lösung des AWP konvergieren.

Wie in Satz 1.25 kann man jedoch zeigen, dass sich bei Lösung der impliziten Gleichungen mit einer Fixpunktiteration wie im Beweis vom Satz 1.22 mit Startwerten $(eta_1, \ldots, \eta_s) = (y_i, \ldots, y_i)$ eine zu einem konvergenten Verfahren führende Lösung ergibt, da sich die selben Fixpunktiterierten wie bei Anwendung auf ein modifiziertes AWP ergeben.

1.5 Numerik steifer Differentialgleichungen

1.5.1 Steife Differentialgleichungen

Bei dem Pendel aus Übungsaufgabe 3.5 waren implizite Verfahren (trotz gleicher Konsistenzordnung) den expliziten weit überlegen. Differentialgleichungen, in denen dieser Effekt auftritt, werden *steif* genannt. Steif ist dabei kein mathematisch präzise definierter Begriff, sondern wird anschaulich für solche Differentialgleichungen verwendet, bei denen naheliegende Standardverfahren (z.B. explizite Runge-Kutta-Verfahren) keine (oder nur für extrem kleine Schrittweiten) zufriedenstellenden Ergebnisse liefern.

Betrachten wir das einfache Beispiel

$$y'(x) = \lambda y, \quad y(0) = 1, \quad \lambda < 0.$$

Offenbar ist die Lösung $y(x) = e^{\lambda x}$. Aufgrund der Annahme $\lambda < 0$ konvergiert die Lösung mit exponentieller Geschwindigkeit gegen Null.

Durch Anwendung des expliziten Euler-Verfahrens mit Schrittweite h auf dieses AWP erhalten wir

$$\begin{aligned} y_1 &= y_0 + h\lambda y_0 = (1 + h\lambda), \\ y_2 &= y_1 + h\lambda y_1 = (1 + h\lambda) y_1 = (1 + h\lambda)^2, \\ &\vdots \\ y_n &= (1 + h\lambda)^n. \end{aligned}$$

Da $\lambda < 0$ folgt für $n \to \infty$

$$\begin{cases} y_n > 0 \text{ und } y_n \to 0 & \text{für } 1 + h\lambda \geq 0, \\ y_n \text{ alterniert im Vorzeichen, } y_n \to 0 & \text{für } 0 > 1 + h\lambda > -1, \\ y_n \text{ alterniert zwischen } +1 \text{ und } -1 & \text{für } 1 + h\lambda = -1, \\ y_n \text{ alterniert im Vorzeichen, } |y_n| \to \infty & \text{für } 1 + h\lambda < -1. \end{cases}$$

Nur für $1 + h\lambda > -1$ (d.h. $h < -2/\lambda$) zeigen die Approximationen also das korrekte Langzeitverhalten und konvergieren gegen Null, und für $h > -1/\lambda$ oszillieren die Approximationen. Für dieses AWP mit stark negativem λ liefert die explizite Euler Methode also nur für extrem kleine Schrittweiten brauchbare Ergebnisse.

Betrachten wir zum Vergleich die implizite Euler-Methode:

$$\begin{aligned} y_1 &= y_0 + h\lambda y_1 \quad \Longrightarrow y_1 = (1 - h\lambda)^{-1}, \\ y_2 &= y_1 + h\lambda y_2 \quad \Longrightarrow y_2 = (1 - h\lambda)^{-1} y_1 = (1 - h\lambda)^{-2}, \\ &\vdots \\ y_n &= (1 - h\lambda)^{-n}. \end{aligned}$$

Für jede Schrittweite h ist $1 - h\lambda > 1$. y_n bleibt also stets positiv und konvergiert gegen Null für $n \to \infty$.

Implizites und explizites Euler-Verfahren besitzen die gleiche Konsistenzordnung. Für $h \to 0$ konvergieren sie gleich schnell gegen die wahre Lösung. Jedoch gibt es zwei Eigenschaften der wahren Lösung, Positivität und Abfallverhalten, die nur die Iterierten des impliziten Euler-Verfahrens für alle Schrittweiten zeigen. Die Iterierten des expliziten Euler-Verfahrens haben diese Eigenschaften erst für extrem kleine Schrittweiten.

Das betrachtete AWP ist also *steif* in dem Sinne, dass die wahre Lösung gewisse Eigenschaften besitzt, die so wichtig sind, dass man nur solche numerischen Approximationen akzeptieren wird, die diese Eigenschaften auch besitzen.

1.5.2 Die Testgleichung

Wir motivieren nun heuristisch, dass sich das im letzten Abschnitt beobachtete Verhalten auch in allgemeinen Anfangswertproblemen wiederfinden lässt.

Betrachten wir das allgemeine AWP

$$y'(x) = f(x, y) \quad \forall x \in [a, b], \qquad y(a) = y_0 \in \mathbb{R}^d.$$

Gemäß Übungsaufgabe 4.2 können wir es in eine autonome Gleichung umformen. Außerdem können wir durch Verschiebung annehmen, dass $x_0 = 0$.

$$y'(x) = f(y), \quad y(0) = y_0 \in \mathbb{R}^d$$

Für kleine Änderungen in x wird sich die Lösung nur wenig verändern. Wir erwarten also, dass sich y *lokal* wie die Lösung der linearisierten Gleichung

$$y'(x) = f(y) \approx f(y_0) + f'(y_0)(y - y_0), \quad y(0) = y_0 \in \mathbb{R}^d,$$

verhält.

Wir nehmen noch an, dass sich die Shifts $f(y_0)$ und y_0 durch geeignete Transformationen eliminieren lassen. Lokal lässt sich das AWP dann durch die Lösung der Gleichung

$$y'(x) = My$$

mit einer Matrix $M \in \mathbb{R}^{d\times d}$ approximieren. Ist M diagonalisierbar mit Eigenwerten $\lambda_1, \ldots, \lambda_d$, dann ist dies äquivalent zu d skalaren Testgleichungen

$$y_j' = \lambda_j y_j, \quad j = 1, \ldots, d.$$

Die Eigenwerte λ_j werden im Allgemeinen komplex sein. Offenbar gelten aber alle Ergebnisse dieses Kapitels auch genauso für komplexwertige Gleichungen.

Insgesamt scheint es also erstrebenswert, Methoden zu konstruieren, die nicht nur möglichst schnell konvergieren, sondern auch qualitativ richtiges Verhalten zeigen für die komplexe skalare Testgleichung

$$y' = \lambda y, \quad \lambda \in \mathbb{C}.$$

Aufgrund der Linearität der Gleichung können wir dabei den Anfangswert auf $y(0) = 1$ setzen.

1.5.3 Die Stabilitätsfunktion

Nach Abschnitt 1.5.1 gilt für die Iterierten des expliziten und impliziten Euler-Verfahrens bei Anwendung auf die Testgleichung (mit $\lambda < 0$)

$$\begin{aligned} y_i^{(\text{expl})} &= (1 + h\lambda)^i y_0 = \left(R^{(\text{expl})}(h\lambda)\right)^i y_0, \\ y_i^{(\text{impl})} &= (1 - h\lambda)^{-i} y_0 = \left(R^{(\text{impl})}(h\lambda)\right)^i y_0, \end{aligned}$$

wobei

$$R^{(\text{expl})}(\zeta) := (1 + \zeta) \quad \text{und} \quad R^{(\text{impl})}(\zeta) = (1 - \zeta)^{-1}.$$

Offenbar gilt das auch für $\lambda \in \mathbb{C}$. Wie gut die Verfahren für die Testgleichung funktionieren, lässt sich also vollständig mit der Funktion $R(\zeta)$ beschreiben. Gleiches gilt für allgemeine Runge Kutta Methoden.

Definition 1.26
Seien

$$A = (a_{ij})_{i,j=1,\dots,s} \in \mathbb{R}^{s\times s}, \quad b = (b_i)_{i=1,\dots,s} \in \mathbb{R}^s, \text{ und } c = (c_i)_{i=1,\dots,s} \in \mathbb{R}^s$$

die Koeffizienten einer Runge-Kutta-Methode. Sei $\mathbb{1} := (1,\dots,1)^T \in \mathbb{R}^s$ *und* $I \in \mathbb{R}^{s\times s}$ *sei die Einheitsmatrix. Für* $\zeta \in \mathbb{C}$ *definieren wir*

$$R(\zeta) := 1 + \zeta b^T (I - \zeta A)^{-1} \mathbb{1} \in \mathbb{C}$$

falls $I - \zeta A$ *invertierbar ist. Ansonsten schreiben wir formal* $R(\zeta) = \infty$. *(Offenbar ist dies genau dann der Fall, wenn* $\frac{1}{\zeta}$ *ein Eigenwert von* A *ist, also für höchstens* s *komplexe Zahlen).*

Satz 1.27
Betrachte die Anwendung des Runge-Kutta-Verfahrens mit Koeffizienten $A \in \mathbb{R}^{s\times s}$, $b, c \in \mathbb{R}^s$ *auf die Testgleichung*

$$y'(x) = \lambda y(x), \quad y_0 = 1,$$

mit $\lambda \in \mathbb{C}$ *und Schrittweite* $h > 0$.

Ist die Matrix $I - h\lambda A \in \mathbb{C}^{s\times s}$ *invertierbar, so ist die Runge-Kutta-Methode anwendbar (d.h. die möglicherweise impliziten Gleichungen lösbar) und ihre Iterierten sind gegeben durch*

$$y_i = (R(h\lambda))^i.$$

Beweis: Anwendung der Runge-Kutta-Methode liefert das (möglicherweise implizite) Gleichungssystem

$$\eta_j = y_i + h \sum_{l=1}^{s} a_{jl} \lambda \eta_l, \quad j = 1,\dots,s.$$

Mit $\eta := (\eta_1,\dots,\eta_s)^T \in \mathbb{C}^s$ ist das äquivalent zu

$$\eta = y_i \mathbb{1} + h\lambda A \eta \quad \Longleftrightarrow \quad (I - h\lambda A)\eta = y_i \mathbb{1}.$$

Ist $I - h\lambda A$ invertierbar, so existiert eine eindeutige Lösung η und wir erhalten

$$\begin{aligned} y_i &:= y_{i-1} + h \sum_{j=1}^{s} b_j \lambda \eta_j = y_{i-1} + h\lambda b^T \eta \\ &= y_{i-1} + h\lambda b^T (I - h\lambda A)^{-1} y_{i-1} \mathbb{1} = (1 + h\lambda b^T (I - h\lambda A)^{-1} \mathbb{1}) y_{i-1} \\ &= (1 + \zeta b^T (I - \zeta A)^{-1} \mathbb{1})^i y_0 = (R(\zeta))^i, \quad \zeta := h\lambda \qquad \square \end{aligned}$$

Beispiel 1.28
(a) Die Stabilitätsfunktion des expliziten Eulerverfahrens ist $R(\zeta) := 1 + \zeta$.

(b) Die Stabilitätsfunktion des impliziten Eulerverfahrens ist $R(\zeta) = (1 - \zeta)^{-1}$.

(c) Das Butcher Tableau für die implizite Mittelpunktsformel aus Abschnitt 1.3.4 ist (vgl. Übungsaufgabe 4.1):

$$\begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array}$$

Die Stabilitätsfunktion ist also

$$R(\zeta) = 1 + \zeta b^T (I - \zeta A)^{-1} \mathbb{1} = 1 + \zeta 1 (1 - \zeta\, 1/2)^{-1} 1 = \frac{1 + \zeta/2}{1 - \zeta/2}.$$

1.5.4 Stabilität

Die exakte Lösung der Testgleichung $y' = \lambda y$, $y(0) = 1$, ist

$$y(x) = e^{\lambda x}.$$

Es gilt also

$$|y(x)| \begin{cases} \to \infty & \text{für } x \to \infty, \quad \text{wenn } \mathrm{Re}(\lambda) > 0, \\ \to 0 & \text{für } x \to \infty, \quad \text{wenn } \mathrm{Re}(\lambda) < 0, \\ = 1 & \text{für alle } x \geq 0, \quad \text{wenn } \mathrm{Re}(\lambda) = 0, \end{cases}$$

und außerdem ist, für alle $x > 0$,

$$|y(x)| \to 0, \quad \text{wenn } \mathrm{Re}(\lambda) \to -\infty.$$

Dies motiviert die folgende Definition.

Definition und Satz 1.29
$y_i \approx y(hi)$ seien die Approximationen einer Einschrittmethode auf die Testgleichung $y' = \lambda y$, $y(0) = 1$, mit äquidistanten Gitterpunkten $x_i = hi$. Die Methode heißt

- A-stabil, *falls für* $\mathrm{Re}(\lambda) \leq 0$ *stets gilt, dass*

$$|y_{i+1}| \leq |y_i| \quad \textit{für alle } i \textit{ und alle Schrittweiten } h,$$

- Isometrie-erhaltend, *wenn für* $\mathrm{Re}(\lambda) = 0$ *stets gilt, dass*

$$|y_{i+1}| = |y_i| \qquad \textit{für alle } i \textit{ und alle Schrittweiten } h,$$

- L-stabil, *wenn sie A-stabil ist und (für alle* $h > 0$*)*

$$|y_1| \to 0 \quad \textit{für } |\lambda| \to \infty.$$

Eine Runge-Kutta-Methode ist genau dann

- *A-stabil, wenn* $|R(\zeta)| \leq 1$ *für alle* $\zeta \in \mathbb{C}$ *mit* $\mathrm{Re}(\zeta) \leq 0$,
- *Isometrie-erhaltend, wenn* $|R(\zeta)| = 1$ *für alle* $\zeta \in \mathbb{C}$ *mit* $\mathrm{Re}(\zeta) = 0$,
- *L-stabil, wenn A-stabil und* $|R(\zeta)| \to 0$ *für* $|\zeta| \to \infty$.

Beweis: Die Äquivalenzen folgen aus $y_i = (R(h\lambda))^i$. □

Aus der Cramerschen Regel folgt, dass die Stabilitätsfunktion eines Runge-Kutta-Verfahrens stets eine rationale Funktion ist, so dass (bei der Definition der L-Stabilität) das Verhalten für $|\zeta| \to \infty$ mit dem für $\mathrm{Re}(\zeta) \to -\infty$ übereinstimmt.

Beispiel 1.30

(a) Explizites Euler-Verfahren:

$$|R(\mathrm{i})| = |1 + \mathrm{i}| = \sqrt{2} > 1.$$

Das explizite Euler-Verfahren ist also weder A-stabil noch Isometrie-erhaltend.

(b) Implizites Euler-Verfahren:

Für alle $\zeta \in \mathbb{C}$ *mit* $\mathrm{Re}(\zeta) \leq 0$ *ist*

$$\begin{aligned} |R(\zeta)| &= \frac{1}{|1-\zeta|} = \frac{1}{\sqrt{\mathrm{Re}(1-\zeta)^2 + \mathrm{Im}(1-\zeta)^2}} \\ &= \frac{1}{\sqrt{(1-\mathrm{Re}(\zeta))^2 + \mathrm{Im}(\zeta)^2}} \leq 1. \end{aligned}$$

Das implizite Euler-Verfahren ist also A-stabil. Außerdem ist $|R(\zeta)| \to 0$ *für* $|\zeta| \to \infty$*, das Verfahren ist also auch L-stabil.*

Es ist jedoch $R(\mathrm{i}) = 1/|1 - \mathrm{i}| = 1/\sqrt{2} < 1$*, das Verfahren ist also nicht Isometrie-erhaltend.*

(c) Die implizite Mittelpunktsformel ist A-stabil (jedoch nicht L-stabil) und Isometrie-erhaltend (siehe Übungsaufgabe 7.2).

Betrachten wir noch einmal die Testgleichung mit $\mathrm{Re}(\lambda) < 0$. A-Stabilität bedeutet, dass die Approximationen das korrekte qualitative Verhalten $|y_{i+1}| \leq |y_i|$ für jede Schrittweite $h > 0$ zeigen. Auch wenn eine Methode nicht A-stabil ist, kann sie dennoch dieses korrekte Verhalten zeigen, wenn nur die Schrittweite klein genug gewählt ist (sodass $|R(h\lambda)| \leq 1$). Dies motiviert die folgende Definition.

Definition 1.31
Zu einer Runge-Kutta-Methode mit Stabilitätsfunktion $R(\zeta)$ definieren wir das Stabilitätsgebiet *durch*

$$\mathcal{S} := \{\zeta \in \mathbb{C} \ : \ |R(\zeta)| \leq 1\} \subseteq \mathbb{C}.$$

Beispielsweise besteht das Stabilitätsgebiet des expliziten Euler-Verfahrens aus allen $\zeta \in \mathbb{C}$ mit

$$1 \geq |R(\zeta)|^2 = |1 + \zeta|^2 = (1 + \mathrm{Re}(\zeta))^2 + \mathrm{Im}(\zeta)^2,$$

d.h. dem abgeschlossenen Kreis mit Radius 1 um $z = -1$ in der komplexen Ebene.

1.5.5 Nachteile expliziter Verfahren

In unseren Beispielen waren nur implizite Verfahren A-stabil oder Isometrie-erhaltend. Tatsächlich können explizite Verfahren diese Eigenschaften nicht besitzen, wie wir in diesem Abschnitt zeigen.

Satz 1.32
Die Stabilitätsfunktion einer expliziten Runge-Kutta-Methode mit s Stufen ist ein Polynom der Ordnung s.

Beweis: Seien $A \in \mathbb{R}^{s\times s}$, $b \in \mathbb{R}^s$, $c \in \mathbb{R}^s$ die Koeffizienten der Methode. Da die Methode explizit ist, ist A eine strikte untere Dreiecksmatrix. Man zeigt leicht, dass in höheren Potenzen von A immer mehr Diagonalen durch Nullen aufgefüllt werden, und schließlich $A^s = 0$ gilt:

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 \\ * & 0 & 0 & 0 & \dots & 0 \\ * & * & 0 & 0 & \dots & 0 \\ * & * & * & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ * & * & * & * & \dots & 0 \end{pmatrix}, \quad A^2 = \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ * & 0 & 0 & 0 & \dots & 0 \\ * & * & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ * & * & * & * & \dots & 0 \end{pmatrix},$$

$$A^3 = \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ * & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ * & * & * & * & \dots & 0 \end{pmatrix}, \quad A^s = \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 \end{pmatrix}.$$

Aus $A^s = 0$ folgt dass

$$(I - \zeta A)(I + \zeta A + \dots \zeta^{s-1} A^{s-1}) = I$$

also $(I - \zeta A)^{-1} = I + \zeta A + \dots \zeta^{s-1} A^{s-1}$.

$R(\zeta) := 1 + \zeta b^T (I - \zeta A)^{-1} \mathbb{1}$ ist also ein Polynom der Ordnung s. □

Satz 1.33

Explizite Runge-Kutta-Methoden sind weder A-stabil noch Isometrie-erhaltend.

Beweis: Für jedes Polynom $R(\zeta)$ gilt $|R(\zeta)| \to \infty$ für $|\zeta| \to \infty$. □

Außerdem erhalten wir noch die schon in Bemerkung 1.24 angesprochene Höchstgrenze in der Ordnung expliziter Verfahren:

Satz 1.34

Die Konsistenzordnung einer expliziten Runge-Kutta-Methode mit s Stufen ist höchstens s.

Beweis: Nach Satz 1.32 ist die Stabilitätsfunktion ein Polynom der Ordnung s, also

$$R(\zeta) = r_0 + r_1 \zeta + \dots + r_s \zeta^s, \quad r_0, \dots, r_s \in \mathbb{R}.$$

Betrachte die Anwendung der Methode auf die Testgleichung mit $\lambda = 1$, also $y' = y$, $y(0) = 1$:

$$y_1 = R(h) = r_0 + r_1 h + \dots + r_s h^s.$$

$$y(x_1) = e^h = 1 + h + \frac{1}{2} h^2 + \dots + \frac{1}{s!} h^s + \frac{1}{(s+1)!} h^{s+1} + O(h^{s+2})$$

Höchstens die ersten s Terme der Entwicklungen können übereinstimmen, sodass der lokale Fehler einer expliziten Methode höchstens $O(h^{s+1})$, also die Ordnung höchstens s sein kann.[4] □

[4] Streng genommen haben wir in dieser Vorlesung nur für AWP, die die Generalvoraussetzung erfüllen, die Konsistenzordnung über den lokalen Fehler definiert, und die Testgleichung erfüllt die Beschränktheitsbedingung aus der Generalvoraussetzung nicht. Da die Testgleichung aber offensichtlich lösbar ist, gilt mit dem Abschneideargument aus Satz 1.25 der Zusammenhang zwischen Konsistenzordnung und lokalem Fehler auch für die Testgleichung.

1.6 Linear implizite Methoden

Wir haben gesehen, dass steife Differentialgleichungen implizite Methoden erfordern. Im Allgemeinen erfordert die Anwendung eines impliziten Runge-Kutta-Verfahrens aber die Lösung von s gekoppelten d-dimensionalen nicht-linearen Gleichungen

$$k_j = f(x_i + c_j h, y_i + h \sum_{l=1}^{s} a_{jl} k_l), \quad j = 1, \ldots, s,$$

nach den sd unbekannten Einträgen der k_j, $j = 1, \ldots, s$. Ziel dieses Abschnitts ist die Herleitung von einfacheren und weniger Rechenaufwand erfordernden, aber dennoch stabilen Methoden.

Wir beschränken uns dabei auf autonome AWP

$$y' = f(y), \quad y(x_0) = y_0$$

(nach Übungsaufgabe 4.2 kann jedes AWP in diese Form gebracht werden).

Die erste Vereinfachung ist, dass wir eine Runge-Kutta-Methode verwenden, für die A eine linke untere Dreiecksmatrix ist, also $a_{jl} = 0$ für $l > j$. Dann können die Gleichungen für die k_j,

$$k_j = f(y_i + h \sum_{l=1}^{j-1} a_{jl} k_l + a_{jj} h k_j), \quad j = 1, \ldots, s,$$

beginnend mit k_1 eine nach der anderen gelöst werden. Statt eines sd-dimensionalen nicht-linearen Gleichungssystems müssen wir so nur s mal ein d-dimensionales nicht-lineares Gleichungssystem lösen. Diese bringen wir auf Nullstellenform, also gegeben $k_1, \ldots, k_{j-1}$ ist k_j so zu bestimmen, dass

$$0 = F_j(k_j) := k_j - f(y_i + h \sum_{l=1}^{j-1} a_{jl} k_l + a_{jj} h k_j).$$

Anwendung des Newton-Verfahrens ergibt ausgehend von einer Startnäherung $k_j^{(0)}$ die Iterationen

$$k_j^{(n+1)} := k_j^{(n)} - F_j'(k_j^{(n)})^{-1} F_j(k_j^{(n)}),$$

wobei

$$F_j'(k_j) = I - f'(y_i + h \sum_{l=1}^{j-1} a_{jl} k_l + a_{jj} h k_j) a_{jj} h.$$

Als weitere Vereinfachung ersetzen wir für alle j die wahre Jacobi-Matrix $F_j'(k_j)$ durch

$$F_j'(k_j) \approx I - a_{jj}hJ, \quad J := f'(y_i).$$

Außerdem führen wir nur einen einzelnen Newton-Schritt durch, d.h. für alle $j = 1, \ldots, s$ setzen wir

$$\begin{aligned} k_j &:= k_j^{(0)} - (I - a_{jj}hJ)^{-1} F_j(k_j^{(0)}) \\ &= k_j^{(0)} - (I - a_{jj}hJ)^{-1} \left(k_j^{(0)} - f(y_i + h\sum_{l=1}^{j-1} a_{jl}k_l + a_{jj}hk_j^{(0)}) \right) \\ &= (I - a_{jj}hJ)^{-1} \left(f(y_i + h\sum_{l=1}^{j-1} a_{jl}k_l + a_{jj}hk_j^{(0)}) - a_{jj}hJk_j^{(0)} \right). \end{aligned}$$

Es bleibt noch die Wahl der Startwerte $k_j^{(0)}$ zu klären. Hierzu verwenden wir eine lineare Kombination der bereits berechneten k_l, $l = 1, \ldots, j-1$:

$$k_j^{(0)} := \sum_{l=1}^{j-1} d_{jl}/a_{jj}k_l$$

mit noch zu bestimmenden Koeffizienten d_{jl}. Insgesamt erhalten wir so die *linear impliziten* (auch: *Rosenbrock-*) Runge Kutta Methoden.

Gegeben $a_{jl}, d_{jl}, b_j, c_j \in \mathbb{R}$, $j = 1, \ldots, s$, $l = 1, \ldots, s$.

- Setze $J := f'(y_i)$ und bestimme k_j, $j = 1, \ldots, s$ nacheinander aus

$$k_j := (I - a_{jj}hJ)^{-1} \left(f(y_i + h\sum_{l=1}^{j-1}(a_{jl} + d_{jl})k_l) - hJ\sum_{l=1}^{j-1} d_{jl}k_l \right).$$

- Setze

$$y_{i+1} := y_i + h\sum_{j=1}^{s} b_j k_j.$$

Bemerkung 1.35
Für alle $k \in \mathbb{R}^d$ ist

$$\|(I - a_{jj}hJ)k\| \geq \|k\| - \|a_{jj}hJk\| \geq (1 - |a_{jj}|h\,\|J\|)\,\|k\|.$$

Für $1-|a_{jj}|h\,\|J\|>0$ *(also* $h<\frac{1}{|a_{jj}\|J\|}$*) ist* $I-a_{jj}hJ\in\mathbb{R}^{d\times d}$ *deshalb injektiv und damit auch bijektiv. Außerdem gilt in dem Fall für alle* $k\in\mathbb{R}^d$*, dass*

$$(1-|a_{jj}|h\,\|J\|)\left\|(I-a_{jj}hJ)^{-1}k\right\|\le\left\|(I-a_{jj}hJ)(I-a_{jj}hJ)^{-1}k\right\|=\|k\|$$

und damit $\|(I-a_{jj}hJ)^{-1}\|\le\frac{1}{1-|a_{jj}|h\|J\|}$.

Satz 1.36
Seien (A,b,c) *die Koeffizienten einer Runge-Kutta-Methode, wobei* A *eine linke untere Dreiecksmatrix und* $a_{jj}\neq 0$ *sei. Dann hat die dazugehörige linear implizite Runge-Kutta-Methode im folgenden Sinne die gleichen Stabilitätseigenschaften wie die ursprüngliche Methode:*

Ist $R(\zeta)$ *die Stabilitätsfunktion der ursprünglichen Methode, dann ergeben sich für jede Wahl der* d_{jl} *bei Anwendung der linear impliziten Methode auf die Testgleichung*

$$y'=\lambda y,\quad y(0)=1$$

die Approximationen

$$y_i=(R(h\lambda))^i,$$

wenn $I-h\lambda A$ *invertierbar ist (also* $\frac{1}{h\lambda}\neq a_{jj}$ *für alle* j*).*

Beweis: Wir wenden die linear implizite Methode auf die Testgleichung an

$$y'=\lambda y=:f(y),\quad y(0)=1.$$

Für alle y ist $J=f'(y)=\lambda$ und damit

$$\begin{aligned}k_j&:=(1-a_{jj}h\lambda)^{-1}\left(\lambda(y_i+h\sum_{l=1}^{j-1}(a_{jl}+d_{jl})k_l)-h\lambda\sum_{l=1}^{j-1}d_{jl}k_l\right)\\&=(1-a_{jj}h\lambda)^{-1}\left(\lambda y_i+h\lambda\sum_{l=1}^{j-1}a_{jl}k_l\right).\end{aligned}$$

Mit $k:=(k_1,\dots,k_s)^T$ ist das äquivalent zu

$$\begin{pmatrix}1-a_{11}h\lambda & 0 & \dots & 0\\ -a_{21}h\lambda & 1-a_{22}h\lambda & \dots & 0\\ \vdots & \vdots & \ddots & \\ -a_{s1}h\lambda & -a_{s2}h\lambda & \dots & 1-a_{ss}h\lambda\end{pmatrix}\begin{pmatrix}k_1\\k_2\\\vdots\\k_s\end{pmatrix}=\begin{pmatrix}\lambda y_i\\\lambda y_i\\\vdots\\\lambda y_i\end{pmatrix}.$$

Wenn $I-h\lambda A$ invertierbar ist, dann ist also

$$k=\lambda y_i(I-h\lambda A)^{-1}\mathbb{1}$$

und damit

$$y_{i+1}=y_i+hb^Tk=(1+h\lambda b^T(I-h\lambda A)^{-1}\mathbb{1})y_i=R(h\lambda)y_i.\quad\square$$

Beispiel 1.37

(a) **Linear-implizites Euler-Verfahren**

Für das implizite Euler-Verfahren

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

ist A eine linke untere Dreiecksmatrix und keine d-Koeffizienten nötig. Das dazugehörige linear-implizite Euler-Verfahren *lautet*

$$y_{i+1} := y_i + hk, \quad mit \quad k := (I - hf'(y_i))^{-1} f(y_i).$$

(b) **Linear-implizites Mittelpunktsverfahren**

Genauso erhalten wir das linear-implizite Mittelpunktsverfahren*:*

$$y_{i+1} := y_i + hk, \quad mit \quad k := (I - h/2f'(y_i))^{-1} f(y_i).$$

(c) `ode23s`

Das wohl am häufigsten verwendete linear-implizite Verfahren besteht aus der folgenden Kombination einer zweistufigen (y) und einer dreistufigen (ŷ) Methode:

$$\begin{aligned} k_1 &:= (I - ahJ)^{-1} f(y_i) \\ k_2 &:= (I - ahJ)^{-1} \left(f(y_i + h/2\, k_1) - ahJk_1\right) \\ k_3 &:= (I - ahJ)^{-1} \left(f(y_i + hk_2) - d_{31}hJk_1 - d_{32}hJk_2\right) \\ y_{i+1} &:= y_i + hk_2 \\ \hat{y}_{i+1} &:= y_i + \frac{h}{6}(k_1 + 4k_2 + k_3), \end{aligned}$$

mit

$$J := f'(y_i), \quad a := \frac{1}{2+\sqrt{2}}, \quad d_{31} := -\frac{4+\sqrt{2}}{2+\sqrt{2}}, \quad d_{32} := \frac{6+\sqrt{2}}{2+\sqrt{2}}.$$

y und ŷ werden wie in Übungsaufgabe 5.4 zur adaptiven Schrittweitensteuerung kombiniert. Das Verfahren ist in Matlab unter dem Namen `ode23s` *eines der zur Lösung steifer DGL empfohlenen Verfahren.*

Lemma 1.38

Das linear implizite Euler-Verfahren ist L-stabil, das linear implizite Mittelpunktsverfahren ist A-stabil und Isometrie-erhaltend.

Beweis: Dies folgt aus Satz 1.36 und den Stabilitätseigenschaften des impliziten Eulerverfahrens und des impliziten Mittelpunktsverfahrens. □

Eine linear implizite Methode besitzt nach Satz 1.36 die gleichen Stabilitätseigenschaften wie die ursprüngliche Methode, aber (je nach Wahl der d_{jl}) kann sich die Konsistenzordnung unterscheiden. Wie in Satz 1.23, lassen sich Ordnungsbedingungen für die Koeffizienten von linear impliziten Verfahren herleiten. Wir zeigen nur exemplarisch am Beispiel `ode23s` die Berechnung der Ordnung eines linear impliziten Verfahrens unter der Generalvoraussetzung aus Abschnitt 1.4.1. Analog zu Abschnitt 1.4.6 folgt daraus auch die Konvergenz auch für den allgemeinen Fall einer unendlich oft differenzierbaren rechten Seite f, falls die Lösbarkeit auf dem kompletten betrachteten Intervall sichergestellt ist.

Satz 1.39
Die in Beispiel 1.37 beschriebene zweistufige Methode zur Berechnung von y in `ode23s` *besitzt Konsistenzordnung 2.*

Beweis: Für jedes $k \in \mathbb{R}^d$ ist

$$\|(I - ahJ)k\| \geq \|k\| - ah\,\|J\|\,\|k\|$$

und $J = f'(y_i)$ ist aufgrund unserer Generalvoraussetzung unabhängig vom Anfangswert x_i, y_i beschränkt.

Für hinreichend kleine $h > 0$ ist die Matrix $I - ahJ$ also invertierbar und es gilt (mit unserer Konvention bzgl. der Landau-Notation aus Abschnitt 1.4.2)

$$\left\|(I - ahJ)^{-1}\right\| \leq \frac{1}{1 - ah\,\|J\|} = \frac{1}{1 + O(h)} = O(1).$$

Wir gehen nun wie im Beweis von Satz 1.23 vor. Nach Übungsaufgabe 3.3 gilt für die Lösung von $y' = f(y)$, $y(x_i) = y_i$

$$y(x_{i+1}) = y_i + hf + 1/2h^2 f'f + O(h^3)$$

wobei wir wieder das Argument (y_i) von f und f' weglassen.

Wir wollen dies mit

$$y_{i+1} = y_i + hk_2,$$

vergleichen und müssen dazu also k_2 bis zur Ordnung $O(h^2)$ entwickeln. Dazu benötigen wie die Entwicklung von k_1. Aus

$$k_1 = (I - ahJ)^{-1} f \quad \text{und} \quad \left\|(I - ahJ)^{-1}\right\| = O(1)$$

folgt $k_1 = O(1)$. Wir verwenden die Definition von k_1 noch einmal und erhalten

$$k_1 = f + ahJk_1 = f + O(h).$$

Für k_2 folgt zuerst

$$k_2 = (I - ahJ)^{-1} \left(f(y_0 + h/2\, k_1) - ahJk_1\right) = O(1)$$

und dann durch nochmalige Anwendung der Definition von k_2

$$\begin{aligned} k_2 &= f(y_0 + h/2\, k_1) - ahJk_1 + ahJk_2 \\ &= f + h/2 f' k_1 + O(h^2) - ahJk_1 + ahJk_2 = f + O(h). \end{aligned}$$

Noch ein weiteres Mal verwenden wir die Definition von k_2 und erhalten zusammen mit $k_1 = f + O(h)$, dass

$$\begin{aligned} k_2 &= f + h/2\, f' k_1 + O(h^2) - ahJk_1 + ahJk_2 \\ &= f + h/2\, f' f - ahJf + ahJf + O(h^2) = f + h/2 f' f + O(h^2). \end{aligned}$$

Insgesamt ist also

$$y_{i+1} = y_i + hk_2 = y_i + hf + h^2/2\, f' f + O(h^3) = y(x_i) + O(h^3),$$

die Methode besitzt also Konsistenzordnung 2. □

1.7 Mehrschrittverfahren

Wir beschreiben nun noch kurz die wesentliche Idee der Mehrschrittverfahren. Dabei beschränken wir uns in diesem Abschnitt auf äquidistant gewählte Gitterpunkte

$$x_i = x_0 + ih, \quad h > 0.$$

In einem m-Schritt Verfahren verwenden wir die letzten m Approximationen

$$y_{i-m+1} \approx y(x_{i-m+1}), \ldots, y_i \approx y(x_i)$$

zur Bestimmung der nächsten Approximation $y_{i+1} \approx y(x_{i+1})$. Für die Bestimmung der dafür nötigen ersten Werte $y_1, \ldots, y_{m-1}$ können dabei Einschrittverfahren oder Mehrschrittverfahren mit weniger Schritten verwendet werden.

1.7.1 Adams-Bashforth Methoden

Zur Bestimmung von $y_{i+1} \approx y(x_{i+1})$ aus $y_{i-m+1}, \ldots, y_i$ verwenden wir zuerst wie bei der Herleitung der Runge Kutta Methoden

$$y_{i+1} - y_i \approx y(x_{i+1}) - y(x_i) = \int_{x_i}^{x_{i+1}} y'(x)\,\mathrm{d}x = \int_{x_i}^{x_{i+1}} f(x, y(x))\,\mathrm{d}x.$$

Die Funktion

$$x \mapsto f(x, y(x))$$

ist (zumindest näherungsweise) an den Stellen

$$f_j := f(x_j, y_j) \approx f(x_j, y(x_j)), \quad j = i - m + 1, \ldots, i$$

bekannt. Es liegt daher nahe, die unbekannte Funktion $x \mapsto f(x, y(x))$ durch ihr Interpolationspolynom $f(x, y(x)) \approx p(x), p \in \Pi_{m-1}$ durch die Stützstellen (x_j, f_j), $j = i - m + 1, \ldots, i$ zu ersetzen. Mit Hilfe der (aus der Numerik I bekannten) Lagrange-Grundpolynome

$$l_k(x) = \prod_{\substack{l=i-m+1,\ldots,i \\ l \neq k}} \frac{x - x_l}{x_k - x_l}, \quad k = i - m + 1, \ldots, i$$

können wir das Interpolationspolynom schreiben als

$$p(x) = \sum_{k=i-m+1}^{i} f_k l_k(x).$$

So erhalten wir

$$\begin{aligned}
y_{i+1} - y_i &\approx \int_{x_i}^{x_{i+1}} p(x)\,\mathrm{d}x = \sum_{k=i-m+1}^{i} f_k \int_{x_i}^{x_{i+1}} l_k(x)\,\mathrm{d}x \\
&= h \sum_{k=i-m+1}^{i} f_k \int_0^1 l_k(x_i + th)\,\mathrm{d}t \\
&= h \sum_{k=i-m+1}^{i} f_k \int_0^1 \prod_{\substack{l=i-m+1,\ldots,i \\ l \neq k}} \frac{x_i + th - x_l}{x_k - x_l}\,\mathrm{d}t \\
&= h \sum_{k=i-m+1}^{i} f_k \int_0^1 \prod_{\substack{l=i-m+1,\ldots,i \\ l \neq k}} \frac{i - l + t}{k - l}\,\mathrm{d}t.
\end{aligned}$$

Mit der Umnummerierung $k = i - m + j$ und $l = i - m + j'$ können wir das schreiben als

$$y_{i+1} - y_i = h \sum_{j=1}^{m} f_{i-m+j} \underbrace{\int_0^1 \prod_{\substack{j'=1,\dots,m \\ j' \neq j}} \frac{m - j' + t}{j - j'} \, \mathrm{d}t}_{=:\beta_j} = h \sum_{j=1}^{m} \beta_j f_{i-m+j},$$

mit (von h und i unabhängigen!) Konstanten $\beta_j \in \mathbb{R}$.

Die so erhaltenen Methoden heißen *explizite Adams Methoden* oder *Adams-Bashforth Methoden.*

Beispiel 1.40
Für den Spezialfall $m = 1$ ergibt sich die explizite Euler-Methode. Für $m = 2$ ist

$$\beta_1 := \int_0^1 \frac{2 - 2 + t}{1 - 2} \, \mathrm{d}t = -\int_0^1 t \, \mathrm{d}t = -\frac{1}{2}$$
$$\beta_2 := \int_0^1 \frac{2 - 1 + t}{-1 + 2} \, \mathrm{d}t = \int_0^1 (t + 1) \, \mathrm{d}t = \frac{3}{2},$$

also $y_{i+1} := y_i + h(\frac{3}{2} f_i - \frac{1}{2} f_{i-1})$.

1.7.2 Weitere auf Integration basierende Methoden

Analog lassen sich implizite Adams Methoden (*Adams-Moulton-Methoden*) aufstellen, indem das Interpolationspolynom durch die Stützstellen (x_j, f_j) für $j = i-m+1, \dots, i+1$, also inklusive der noch zu bestimmenden Stützstelle gewählt wird. Dies führt auf Formeln der Form

$$y_{i+1} = y_i + h \sum_{j=1}^{m+1} \beta_j f_{i-m+j} = y_i + h \sum_{j=1}^{m+1} \beta_j f(x_{i-m+j}, y_{i-m+j}). \tag{1.8}$$

Eine verbreitete Methode diese impliziten Gleichungen zu lösen, ist es zuerst eine Näherung an y_{i+1} (und damit an f_{i+1}) durch die explizite Adams Methode zu bestimmen. Diese Näherung wird dann als Startwert für eine Fixpunktiteration der Gleichung (1.8) verwendet (üblich sind ein oder zwei Iterationsschritte). Dieses Vorgehen heißt *Predictor-Corrector-Verfahren.*

Das Integrationsintervall bei der Herleitung der Methoden könnte auch vor x_i liegende Bereiche umfassen, z.B.

$$y_{i+1} - y_{i-1} \approx y(x_{i+1}) - y(x_{i-1}) = \int_{x_{i-1}}^{x_{i+1}} y'(x) \, \mathrm{d}x = \int_{x_{i-1}}^{x_{i+1}} f(x, y(x)) \, \mathrm{d}x.$$

Analag zum Adams-Verfahren können wir f durch sein Interpolationspolynom (mit oder ohne Verwendung der unbekannten Stützstelle x_{i+1}, f_{i+1}) annähern und erhalten (implizite bzw. explizite) Formeln der Form

$$y_{i+1} = y_{i-1} + h \sum_{j=1}^{m+1} \beta_j f_{i-m+j} \quad \text{bzw.} \quad y_{i+1} := y_{i-1} + h \sum_{j=1}^{m} \beta_j f_{i-m+j}.$$

Diese Formeln heißen *Nyström-Methoden* (die explizite Variante) oder *Milne-Simpson-Methoden* (die implizite Variante).

1.7.3 Auf Differentiation basierende Methoden

Die bisher betrachteten Mehrschrittverfahren beruhten auf der Idee die Funktion $x \mapsto f(x, y(x)) = y'(x)$ durch ein Interpolationspolynom zu approximieren und dieses zu integrieren. Stattdessen können wir auch die (Approximationen an die) Funktionswerte $y_{i-m+1}, \ldots, y_{i+1}$ durch ein Polynom interpolieren. Wie bei der Herleitung der Adams-Bashforth Methode lässt sich das Interpolationspolynom $q \in \Pi_m$ schreiben als

$$q(x) = \sum_{k=i+1-m}^{i+1} y_k l_k(x), \quad l_k(x) = \prod_{\substack{l=i+1-m,\ldots,i+1 \\ l \neq k}} \frac{x - x_l}{x_k - x_l}$$

also

$$\begin{aligned} q(x_i + th) &= \sum_{k=i+1-m}^{i+1} y_k \prod_{\substack{l=i+1-m,\ldots,i+1 \\ l \neq k}} \frac{x_i + th - x_l}{x_k - x_l} \\ &= \sum_{k=i+1-m}^{i+1} y_k \prod_{\substack{l=i+1-m,\ldots,i+1 \\ l \neq k}} \frac{i - l + t}{k - l} \\ &= \sum_{j=1}^{m+1} y_{i-m+j} \prod_{\substack{j'=1,\ldots,m+1 \\ j' \neq j}} \frac{m - j' + t}{j - j'}. \end{aligned}$$

Wir können nun versuchen, den unbekannten Wert y_{i+1} so zu bestimmen, dass das Interpolationspolynom q im aktuellen Gitterpunkt x_i die Differentialgleichung erfüllt, also

$$q'(x_i) = f(x_i, y_i).$$

Wegen

$$\begin{aligned} q'(x_i) &= \frac{1}{h}\frac{\partial}{\partial t} q(x_i + th)|_{t=0} \\ &= \frac{1}{h}\sum_{j=1}^{m+1} y_{i-m+j} \underbrace{\left(\frac{\partial}{\partial t} \prod_{\substack{j'=1,\dots,m+1 \\ j' \neq j}} \frac{m - j' + t}{j - j'} \right) \Big|_{t=0}}_{=:\alpha_j} \end{aligned}$$

führt dies auf Formeln der Form

$$\sum_{j=1}^{m+1} \alpha_j y_{i-m+j} = h f(x_i, y_i),$$

die sich (für $\alpha_{m+1} \neq 0$) explizit nach y_{i+1} auflösen lassen.

Genauso führt die Forderung, dass das Interpolationspolynom q im nächsten Gitterpunkt x_{i+1} die Differentialgleichung erfüllt, mittels

$$\begin{aligned} q'(x_{i+1}) &= \frac{1}{h}\frac{\partial}{\partial t} q(x_{i+1} + th)|_{t=0} \\ &= \frac{1}{h}\sum_{j=1}^{m+1} y_{i-m+j} \underbrace{\left(\frac{\partial}{\partial t} \prod_{\substack{j'=1,\dots,m+1 \\ j' \neq j}} \frac{m + 1 - j' + t}{j - j'} \right) \Big|_{t=0}}_{=:\alpha_j} \end{aligned}$$

auf implizite Methoden der Form

$$\sum_{j=1}^{m+1} \alpha_j y_{i-m+j} = h f(x_{i+1}, y_{i+1}).$$

Die so entstandenen impliziten Formeln heißen auch *BDF-Methoden* (Backward differentiation formula).

Beispiel 1.41
Für die implizite BDF-Methode mit $m = 1$ ergibt sich

$$\begin{aligned} \alpha_1 &= \frac{\partial}{\partial t} \frac{1 + 1 - 2 + t}{1 - 2} \Big|_{t=0} = -1, \\ \alpha_2 &= \frac{\partial}{\partial t} \frac{1 + 1 - 1 + t}{2 - 1} \Big|_{t=0} = 1, \end{aligned}$$

also

$$-1 y_i + 1 y_{i+1} = h f(x_{i+1}, y_{i+1}),$$

und damit gerade die implizite Euler-Methode.

1.7.4 Konvergenz linearer Mehrschrittmethoden

Alle bisher kennengelernten Mehrschrittmethoden können wir in der allgemeinen Form

$$\sum_{j=1}^{m+1} \alpha_j y_{i-m+j} = h \sum_{j=1}^{m+1} \beta_j f_{i-m+j}$$

mit Konstanten $\alpha_1, \ldots, \alpha_{m+1}, \beta_1, \ldots, \beta_{m+1} \in \mathbb{R}$ schreiben.

Wir geben in dieser Vorlesung nur eine ganz kurze Zusammenfassung der Theorie dieser Methoden. Eine ausführlichere Darstellung findet sich z.B. in [HairerNorsettWanner].

Analog zu Einschrittmethoden definiert man auch bei Mehrschrittverfahren die *Konsistenzordnung* durch Betrachtung des *lokalen Fehlers*

$$\|y_{i+1} - y(x_{i+1})\|,$$

der sich ergibt, wenn die Methode auf m exakte Werte

$$y_{i-m+1} = y(x_{i-m+1}), \ldots, y_i = y(x_i)$$

angewendet wird.

Das Konvergenzresultat für Einschrittverfahren in Satz 1.19 lässt sich jedoch nicht unmittelbar auf Mehrschrittverfahren übertragen. Im Gegensatz zu Einschrittverfahren folgt aus der Konsistenz eines Mehrschrittverfahrens nicht automatisch Konvergenz, sondern dies erfordert eine zusätzliche Stabilitätseigenschaft. Die in Abschnitt 1.7.1 und 1.7.2 vorgestellten Adam-Varianten erfüllen diese zusätzlichen Eigenschaften, die BDF-Formeln jedoch nur bis $m \leq 6$.

1.8 Eindimensionale Randwertprobleme

1.8.1 Motivation: Diffusionsprozesse

Neben Anfangswertproblemen treten in der Praxis auch *Randwertprobleme* für gewöhnliche Differentialgleichungen auf. Die Theorie und Numerik dieser Probleme ist eng mit der für partielle Differentialgleichungen verwandt, da (wie in der folgenden Motivation) Randwertprobleme für gewöhnliche Differentialgleichungen oft als eindimensionale stationäre Spezialfälle von Randwertproblemen für partielle Differentialgleichungen (PDGL) auftreten. Die

folgende Modellierung von Diffusionsprozessen folgt dem sehr lesenswerten Buch [FulfordBroadbridge].

Wir betrachten ein Rohr mit Querschnitt A, das von einer Lösung durchflossen wird. Wir bezeichnen mit

x: die Position innerhalb des Rohres, etwa $x \in [0, 1]$,

$C(x, t)$: die Konzentration des gelösten Stoffes am Ort x zur Zeit t,

$J(x, t)$: die Flussdichte der Lösung, d.h. welche Masse des Stoffes einen Einheitsquerschnitt pro Zeiteinheit durchquert.

Wir betrachten den Rohrabschnitt zwischen x und $x+\delta x$. Dabei nehmen wir an, dass δx so klein ist, dass die Konzentration in diesem Abschnitt räumlich konstant ist. Die Gesamtmasse innerhalb des Abschnitts ist also

$$A\delta x C(x, t).$$

Nun nehmen wir an, dass δt so klein ist, dass der Fluss im Zeitabschnitt zwischen t und $t+\delta t$ zeitlich konstant ist. Aufgrund des Flusses wird sich im betrachteten Abschnitt des Rohres die Gesamtmasse in diesem Zeitabschnitt ändern um

$$J(x, t)A\delta t - J(x + \delta x, t)A\delta t,$$

vgl. die in der Vorlesung gemalten Skizzen.

Wenn es keine anderen die Masse ändernden Phänomene gibt, so gilt also

$$A\delta x C(x, t + \delta t) = A\delta x C(x, t) + J(x, t)A\delta t - J(x + \delta x, t)A\delta t$$

also

$$\frac{C(x, t + \delta t) - C(x, t)}{\delta t} = -\frac{J(x + \delta x, t) - J(x, t)}{\delta x}$$

und mit $\delta x \to 0$, $\delta t \to 0$ erhalten wir die *Bilanzgleichung* (auch: *Kontinuitätsgleichung*)

$$\frac{\partial C(x, t)}{\partial t} = -\frac{\partial J(x, t)}{\partial x}.$$

Das einfachste Model für Diffusion ist *Fick's Gesetz*, das besagt, dass die Flussdichte proportional ist zum Konzentrationsgefälle

$$J(x, t) = -D(x)\frac{\partial C(x, t)}{\partial x}$$

($D(x)$ heißt Diffusionskonstante).

So erhalten wir eine partielle Differentialgleichung, die sogenannte *Diffusionsgleichung* oder auch *Wärmeleitungsgleichung*

$$\frac{\partial C(x,t)}{\partial t} = \frac{\partial}{\partial x}\left(D(x)\frac{\partial C(x,t)}{\partial x}\right).$$

Konvektion und Absorption können ähnlich modelliert werden. Wenn die Flüssigkeit sich mit der Geschwindigkeit $v(x,t)$ bewegt, dann muss der Term $v(x,t)C(x,t)$ zum Fluss addiert werden. Wenn pro Zeiteinheit und Raumeinheit die Masse $M(x,t)$ hinzugegeben wird oder $a(x,t)C(x,t)$ z.B. aufgrund einer chemischen Reaktion verbraucht wird, so müssen diese Änderungen in der Massenbilanz berücksichtigt werden. Insgesamt erhalten wir

$$\begin{aligned}\frac{\partial C}{\partial t}(x,t) = \frac{\partial}{\partial x}\left(D(x)\frac{\partial}{\partial x}C(x,t)\right) - \frac{\partial}{\partial x}(v(x,t)C(x,t))\\ - a(x,t)C(x,t) + M(x,t).\end{aligned}$$

Es erscheint natürlich, dass diese partielle Differentialgleichung Anfangsbedingungen $C(x,0)$ für alle $x \in [0,1]$ und Randbedingungen für $x = 0$ und $x = 1$ benötigt. Als Randbedingungen können wir z.B. die Konzentration $C(0,t)$ und $C(1,t)$ für alle $t > 0$ (Dirichlet-Randbedingungen) oder den Fluss $-D(0)\frac{\partial C(0,t)}{\partial x}$ und $-D(1)\frac{\partial C(1,t)}{\partial x}$ (Neumann-Randbedingungen) vorschreiben.

Sind alle Koeffizienten der Gleichung von der Zeit unabhängig, so stellt sich oft mit der Zeit ein Gleichgewichtszustand ein, d.h. die Konzentration ändert sich nicht mehr. Für diesen muss also gelten

$$\frac{\partial}{\partial x}\left(D(x)\frac{\partial}{\partial x}C(x)\right) - \frac{\partial}{\partial x}(v(x)C(x)) - a(x)C(x) + M(x) = 0.$$

Dies ist wieder eine *gewöhnliche Differentialgleichung*, für die wir jedoch (Dirichlet- oder Neumann-)Randwerte anstelle von Anfangswerten kennen.

1.8.2 Differenzenverfahren

Motiviert durch den letzten Abschnitt betrachten wir nun die leicht vereinfachte Diffusionsgleichung

$$L[u] := -u''(x) + b(x)u'(x) + c(x)u(x) = f(x) \quad x \in (0,1)$$

und zwar zuerst mit *homogenen* Dirichlet-Randbedingungen $u(0) = u(1) = 0$.

Es ist naheliegend, dass Randwertproblem zu lösen, indem wir die Funktion u diskretisieren durch ein äquidistantes Gitter $x_i = ih$, $i = 0, \dots n+1$, $h := 1/(n+1)$. Es bezeichne

$$U := (u(x_1), \dots, u(x_n))^T \quad \text{und} \quad F := (f(x_1), \dots, f(x_n))^T$$

die Auswertungen von u und f auf diesem Gitter.

Wir ersetzen die Ableitungen durch *zentrale finite Differenzen*

$$u'(x) \approx D_h[u](x) := \frac{u(x+h) - u(x-h)}{2h}$$
$$u''(x) \approx D_h^2[u](x) := \frac{u(x+h) - 2u(x) + u(x-h)}{h^2}$$

(wobei wir am Rand $u(0) = 0 = u(1)$ verwenden).

Aus der Gleichung $L[u] = f$ ergibt sich so das LGS

$$L_h U_h = F$$

mit einer Matrix $L_h \in \mathbb{R}^{n \times n}$. Durch Lösung des LGS erhalten wir einen Vektor

$$U_h := (u_1, \dots, u_n)^T \in \mathbb{R}^n$$

von Approximationen an $(u(x_1), \dots, u(x_n))^T$.

Finite Differenzen für ein einfaches Beispiel Mit diesem Ansatz ergibt sich für das einfache Beispiel $-u'' = f$

$$\underbrace{\begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{pmatrix}}_{=:F} = - \begin{pmatrix} u''(x_1) \\ u''(x_2) \\ \vdots \\ u''(x_n) \end{pmatrix} \approx \underbrace{h^{-2} \begin{pmatrix} 2 & -1 & & 0 \\ -1 & 2 & -1 & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{pmatrix}}_{=:L_h} \begin{pmatrix} u(x_1) \\ u(x_2) \\ \vdots \\ u(x_n) \end{pmatrix}.$$

Wir können daher erwarten, dass wir durch Lösung von $F = L_h U_h$ einen Vektor $U_h = (u_1, \dots, u_n)^T$ aus Approximationen an $(u(x_1), \dots, u(x_n))^T$ erhalten.

Finite Differenzen für die Diffusionsgleichung Genauso diskretisieren wir

$$-u''(x) + b(x)u'(x) + c(x)u(x) = f(x), \quad u(0) = u(1) = 0$$

und erhalten

$$\begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{pmatrix} \approx h^{-2} \begin{pmatrix} d_1 & s_1 & & 0 \\ r_2 & d_2 & s_2 & \\ & \ddots & \ddots & s_{n-1} \\ 0 & & r_n & d_n \end{pmatrix} \begin{pmatrix} u(x_1) \\ u(x_2) \\ \vdots \\ u(x_n) \end{pmatrix}$$

mit

$$\begin{aligned} d_i &= 2 + h^2 c(x_i), \\ r_i &= -1 - hb(x_i)/2, \\ s_i &= -1 + hb(x_i)/2. \end{aligned}$$

Wiederum ergibt sich ein LGS $F \approx L_h U$, und wir können erwarten, dass die Lösung $U_h := L_h^{-1} F$ die wahren Lösungswerte in U approximiert.

Inhomogene Dirichlet-Bedingungen Im Falle inhomogener Dirichlet-Bedingungen $u(0) = \alpha \in \mathbb{R}$, $u(1) = \beta \in \mathbb{R}$ ergibt sich

$$\begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{pmatrix} \approx h^{-2} \begin{pmatrix} d_1 & s_1 & & 0 \\ r_2 & d_2 & s_2 & \\ & \ddots & \ddots & s_{n-1} \\ 0 & & r_n & d_n \end{pmatrix} \begin{pmatrix} u(x_1) \\ u(x_2) \\ \vdots \\ u(x_n) \end{pmatrix} + h^{-2} \begin{pmatrix} r_1 \alpha \\ 0 \\ \vdots \\ s_n \beta \end{pmatrix}.$$

Wir erhalten das LGS $F \approx L_h U + B_h$ und können erwarten, dass

$$U_h := L_h^{-1}(F - B_h) \approx U.$$

Neumann-Randbedingungen Neumann-Randbedingungen $u'(x_0) = \alpha$, $u'(x_{n+1}) = \beta$ können behandelt werden, indem wir die unbekannten Auswertungen von u an den Randwerten x_0 und x_{n+1} zu den Vektoren hinzufügen. So erhalten wir zunächst das unterbestimmte Gleichungssystem

$$\begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{pmatrix} \approx h^{-2} \underbrace{\begin{pmatrix} r_1 & d_1 & s_1 & & & 0 \\ & r_2 & d_2 & s_2 & & \\ & & \ddots & \ddots & s_{n-1} & 0 \\ 0 & & & r_n & d_n & s_n \end{pmatrix}}_{\in \mathbb{R}^{n \times (n+2)}} \begin{pmatrix} u(x_0) \\ u(x_1) \\ u(x_2) \\ \vdots \\ u(x_n) \\ u(x_{n+1}) \end{pmatrix}.$$

Analog ergeben sich durch Verwendung zentraler Finiter Differenzen in x_0 und x_{n+1} die Gleichungen

$$\begin{aligned} f(x_0) &\approx h^{-2}\left(r_0 u(x_{-1}) + d_0 u(x_0) + s_0 u(x_1)\right), \\ f(x_{n+1}) &\approx h^{-2}\left(r_{n+1} u(x_n) + d_{n+1} u(x_{n+1}) + s_{n+2} u(x_{n+2})\right), \end{aligned}$$

wobei $x_{-1} = x_0 - h$ und $x_{n+2} = x_{n+1} + h$. Aus den Neumann-Randbedingungen können wir Näherungen an $u(x_{-1})$ und $u(x_{n+2})$ berechnen,

$$\begin{aligned} u(x_{-1}) &\approx u(x_0) - hu'(x_0) = u(x_0) - \alpha h, \\ u(x_{n+2}) &\approx u(x_{n+1}) + hu'(x_{n+1}) = u(x_{n+1}) + \beta h. \end{aligned}$$

Damit ist

$$\begin{aligned} \underbrace{\begin{pmatrix} f(x_0) \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{pmatrix}}_{=:F} &\approx h^{-2} \begin{pmatrix} d_0 & s_0 & & & \\ r_1 & d_1 & s_1 & & \\ & \ddots & \ddots & \ddots & \\ & & r_n & d_n & s_n \\ & & & r_{n+1} & d_{n+1} \end{pmatrix} \begin{pmatrix} u(x_0) \\ u(x_1) \\ \vdots \\ u(x_{n+1}) \end{pmatrix} \\ &\quad + h^{-2} \begin{pmatrix} r_0(u(x_0) - \alpha h) \\ 0 \\ \vdots \\ s_{n+1}(u(x_{n+1}) + \beta h) \end{pmatrix} \\ &= \underbrace{h^{-2} \begin{pmatrix} d_0 + r_0 & s_0 & & & \\ r_1 & d_1 & s_1 & & \\ & \ddots & \ddots & \ddots & \\ & & r_n & d_n & s_n \\ & & & r_{n+1} & d_{n+1} + s_{n+1} \end{pmatrix}}_{=:L_h} \underbrace{\begin{pmatrix} u(x_0) \\ u(x_1) \\ \vdots \\ u(x_{n+1}) \end{pmatrix}}_{=:U} \\ &\quad + \underbrace{h^{-1} \begin{pmatrix} -r_0\alpha \\ 0 \\ \vdots \\ s_{n+1}\beta \end{pmatrix}}_{=:B_h} \end{aligned}$$

Wiederum ergibt sich, dass der Vektor $U \in \mathbb{R}^{n+1}$ der Auswertungen von u in den (um die Randpunkte erweiterten) Gitterpunkten $x_0, \ldots, x_{n+1}$ annähernd ein LGS $F \approx L_h U + B_h$ löst und wir erwarten, dass $U_h := L_h^{-1}(F - B_h) \approx U$.

1.8.3 Konsistenz, Stabilität und Konvergenz

Wir betrachten in diesem Abschnitt nur das spezielle Randwertproblem

$$L[u] := -u''(x) + b(x)u'(x) + c(x)u(x) = f(x) \quad x \in (0,1)$$

mit homogenen Dirichlet-Randbedingungen und die dazugehörige Diskretisierung

$$L_h U_h = F$$

über einem äquidistanten Gitter $x_i = ih$, $i = 0, \dots n+1$, $h := 1/(n+1)$, aus dem letzten Abschnitt. Außerdem nehmen wir an, dass $b \in C^2([0,1])$, $c \in C([0,1])$ sowie $c > 0$ gilt[5], und dass das Randwertproblem eine eindeutige Lösung $u \in C^4([0,1])$ besitzt.

Zuerst charakterisieren wir, wie gut die wahren Lösungswerte

$$U := (u(x_1), \dots, u(x_n))^T$$

die diskretisierte Gleichung lösen.

Lemma 1.42
Es existiert ein $C > 0$, sodass

$$\|L_h U - F\|_\infty \leq Ch^2.$$

Man sagt auch, das Differenzenverfahren hat Konsistenzordnung *2.*

Beweis: Der i-te Eintrag ($i = 1, \dots, n$, $u(x_0) = 0 = u(x_{n+1})$) von $L_h U - F$ ist

$$-D_h^2[u](x_i) + b(x_i)D_h[u](x_i) + c(x_i)u(x_i) - f(x_i).$$

Da u die DGL $-u''(x_i) + b(x_i)u'(x_i) + c(x_i)u(x_i) - f(x_i) = 0$ löst, genügt es zu zeigen, dass

$$D_h[u](x_i) = u'(x_i) + O(h^2) \quad \text{und} \quad D_h^2[u](x_i) = u''(x_i) + O(h^2).$$

In der Tat erhalten wir durch Taylorentwichlung

$$u(x_i + h) = u(x_i) + hu'(x_i) + \frac{1}{2}h^2u''(x_i) + \frac{1}{3!}h^3u'''(x_i) + O(h^4)$$
$$u(x_i - h) = u(x_i) - hu'(x_i) + \frac{1}{2}h^2u''(x_i) - \frac{1}{3!}h^3u'''(x_i) + O(h^4)$$

[5]Da die stetige Funktion c auf dem Kompaktum $[0,1]$ ihr Minimum annimmt gilt damit sogar $c(x) \geq c_0 := \min_{x\in[0,1]} c(x) > 0$.

und damit

$$\begin{aligned} D_h[u](x_i) &= \frac{u(x_i+h)-u(x_i-h)}{2h} = \frac{2hu'(x_i)+O(h^3)}{2h} = u'(x_i)+O(h^2) \\ D_h^2[u](x_i) &= \frac{u(x_i+h)-2u(x_i)+u(x_i-h)}{h^2} = \frac{h^2u''(x_i)+O(h^4)}{h^2} \\ &= u''(x_i)+O(h^2), \end{aligned}$$

womit die Behauptung gezeigt ist. □

Aus Konsistenz (im Sinne von Lemma 1.42) folgt mit dem folgenden einfachen Argument Konvergenz

$$\|U-U_h\|_\infty = \left\|L_h^{-1}L_h(U-U_h)\right\|_\infty \le \left\|L_h^{-1}\right\|_\infty \|L_hU-F\|_\infty ,$$

wenn wir zeigen können, dass L_h invertierbar ist **und** $\left\|L_h^{-1}\right\|_\infty$ **(gleichmäßig in h) beschränkt ist**. Die zweite Eigenschaft heißt auch *Stabilität* des Differenzenverfahrens. Um die Stabilität zu zeigen, konstruieren wir eine Lösung w eines Randwertproblems, für die zugehörigen Auswertungen W

$$L_hW \ge \mathbb{1}$$

erfüllen. Zusammen mit einer noch zu zeigenden eintragsweisen Nichtnegativität von L_h^{-1} und einer daraus folgenden Monotonieeigenschaft folgt dann

$$\left\|L_h^{-1}\right\|_\infty = \left\|L_h^{-1}\mathbb{1}\right\|_\infty \le \left\|L_h^{-1}L_hW\right\|_\infty \le \max_{x\in[0,1]} w(x).$$

Bemerkung 1.43
Eine komponentenweise nicht-negative Matrix $M=(m_{ij})_{i,j=1}^n$ hat die Monotonieeigenschaft

$$x \le y \quad \Longrightarrow \quad Mx \le My,$$

wobei die Ungleichheitszeichen für die Vektoren $x, y, Mx, My \in \mathbb{R}^n$ komponentenweise zu verstehen sind.

Lemma 1.44
(a) Ist $A \in \mathbb{R}^{n\times n}$ eine strikt diagonaldominante Matrix,

$$a_{ii} > \sum_{\substack{j=1\\ j\neq i}}^{N} |a_{ij}|, \quad i=1,\dots,n,$$

mit positiven Diagonalelementen und nicht-positiven Nichtdiagonalelementen, dann ist A invertierbar und A^{-1} ist komponentenweise nichtnegativ.

(b) Ist $A \in \mathbb{R}^{n\times n}$ eine invertierbare, diagonaldominante Matrix,

$$a_{ii} \geq \sum_{\substack{j=1\\ j\neq i}}^{N} |a_{ij}|, \quad i = 1, \ldots, n,$$

mit positiven Diagonalelementen und nicht-positiven Nichtdiagonalelementen, dann ist A^{-1} komponentenweise nicht-negativ.

Insbesondere gilt gemäß Bemerkung 1.43 in beiden Fällen komponentenweise

$$Au \leq Av \quad \Longrightarrow \quad u \leq v.$$

Beweis: (a) Wir zerlegen $A = D - N$ in seinen Diagonal- und Nichtdiagonalanteil. Nach Voraussetzung ist $D \geq 0$ und $N \geq 0$. Für $R = D^{-1}N$ gilt offenbar

$$A = D(I - R), \quad R \geq 0, \quad \|R\|_\infty = \left\|D^{-1}N\right\|_\infty < 1.$$

Mit Hilfe der Neumannschen Reihe (siehe z.B. [NumerikWS1718, Lemma 4.16]) folgt, dass $I - R$ invertierbar ist und $(I - R)^{-1} = \sum_{k=0}^{\infty} R^k$. Damit ist auch A invertierbar und

$$A^{-1} = (I - R)^{-1}D^{-1} = \sum_{k=0}^{\infty} R^k D^{-1}.$$

Die Einträge von A^{-1} sind also Grenzwerte von Summen und Produkten nicht-negativer Zahlen und damit nicht-negativ.

(b) Für (nicht notwendigerweise strikt) diagonaldominantes und invertierbares A (mit positiven Diagonalelementen und nicht-positiven Nichtdiagonalelementen) erhalten wir aus Teil (a), dass $(A+\epsilon I)$ invertierbar ist, und dass $(A + \epsilon I)^{-1}$ komponentenweise nicht-negativ ist. Da A nach Voraussetzung invertierbar ist, folgt mit der Stetigkeit der Matrixinversen (siehe z.B. [NumerikWS1718, Lemma 4.17]), dass (für $\epsilon \to 0$) $(A+\epsilon I)^{-1} \to A^{-1}$ konvergiert. Die Einträge von A^{-1} sind also Grenzwerte nicht-negativer Zahlen und damit nicht-negativ. □

Lemma 1.45
Es existieren $h_0 > 0$ und $C > 0$ mit

$$\left\|L_h^{-1}\right\|_\infty \leq C \quad \text{für alle } 0 < h < h_0.$$

Beweis: Nach Übungsaufgabe 9.1 existiert eine Lösung $w \in C^4[0,1]$ des Randwertproblems

$$-w''(x) + b(x)w'(x) = 1 \quad x \in (0,1), \qquad w(0) = 0 = w(1).$$

Da w stetig ist, besitzt w sein globales Minimum in $[0,1]$. Da in jedem inneren Minimum $w'(x) = 0 \leq w''(x)$ gilt und damit die DGL nicht erfüllt sein kann, muss das Minimum auf dem Rand liegen und es folgt

$$w(x) \geq 0 \quad \forall x \in [0,1].$$

w erfüllt

$$L[w] = -w''(x) + b(x)w'(x) + c(x)w(x) = 1 + c(x)w(x).$$

Nach Lemma 1.42 existiert deshalb ein $C' > 0$, sodass für hinreichend kleine $h > 0$ mit den Bezeichnungen

$$\begin{aligned} W &= (w(x_1), \ldots, w(x_n))^T \quad \text{und} \\ G &= (1 + c(x_1)w(x_1), \ldots, 1 + c(x_n)w(x_n))^T \end{aligned}$$

gilt, dass

$$\|L_h W - G\|_\infty \leq C' h^2,$$

und damit insbesondere

$$L_h W \geq G - C' h^2 \mathbb{1}.$$

Da c und w nicht-negativ sind, ist $G \geq \mathbb{1}$ und es folgt

$$L_h W \geq \mathbb{1} - C' h^2 \mathbb{1}.$$

Für hinreichend kleine $h > 0$ ist $1 - C'h^2 > \frac{1}{2}$ und die Matrix L_h erfüllt die Voraussetzungen von Lemma 1.44 (ist also strikt diagonaldominant mit postiven Diagonal- und nichtpositiven Nebendiagonalelementen). Mit Bemerkung 1.43 folgt dann für hinreichend kleine $h > 0$

$$L_h W \geq \frac{1}{2}\mathbb{1} \quad \Longrightarrow \quad W \geq \frac{1}{2} L_h^{-1} \mathbb{1}$$

und damit

$$\left\|L_h^{-1}\right\|_\infty = \left\|L_h^{-1}\mathbb{1}\right\|_\infty \leq 2\|W\|_\infty \leq 2 \max_{x \in [0,1]} w(x),$$

womit die Behauptung gezeigt ist. □

Folgerung 1.46
Es existieren $h_0 > 0$ und $C > 0$ mit

$$\|U - U_h\|_\infty \leq Ch^2 \quad \text{für alle } 0 < h < h_0.$$

Beweis: Mit

$$\|U - U_h\|_\infty = \left\|L_h^{-1} L_h (U - U_h)\right\| \leq \left\|L_h^{-1}\right\| \|L_h U - F\|$$

folgt die Behauptung aus Lemma 1.42 und Lemma 1.45. □

Kapitel 2

Partielle Differentialgleichungen

2.1 Motivation und Klassifikation

2.1.1 Mehrdimensionale Diffusion

Analog zu Abschnitt 1.8.1 können wir Diffusionsprozesse auch im Mehrdimensionalen modellieren. $x = (x_1, \ldots, x_n)^T$ und die Flussdichte $J(x,t)$ sind nun n-dimensionale Vektoren. Die j-te Komponente von $J(x,t)$ bezeichne dabei den Anteil des Flusses in die j-te Koordinatenrichtung. Ersetzen wir in Abschnitt 1.8.1 den Rohrabschnitt durch einen n-dimensionalen Würfel, so erhalten wir für die Änderung der Massenkonzentration $C(x,t)$ aufgrund eines Flusses $J(x,t)$ mit $\delta x_j \to 0$, $\delta t \to 0$ die *Bilanzgleichung*

$$\begin{aligned}\frac{\partial C(x,t)}{\partial t} &= -\frac{\partial J_1(x,t)}{\partial x_1} - \frac{\partial J_2(x,t)}{\partial x_2} - \cdots - \frac{\partial J_n(x,t)}{\partial x_n} \\ &= -\mathrm{div}(J(x,t)) = -\nabla \cdot J(x,t),\end{aligned}$$

wobei wir in der gesamten Vorlesung die Konvention verwenden, dass sich die (meist kurz mit dem Nabla-Operator geschriebenen) Operatoren Divergenz, Gradient und Rotation stets nur auf die räumlichen Koordinaten beziehen.

Fick's Gesetz lautet entsprechend

$$J_1(x,t) = -D(x,t)\frac{\partial C(x,t)}{\partial x_1}, \quad \ldots \quad J_n(x,t) = -D(x,t)\frac{\partial C(x,t)}{\partial x_n},$$

also

$$J(x,t) = -D(x,t)\nabla C(x,t).$$

D sei dabei weiterhin ein Skalar. (Der Fall *anisotroper*, d.h. richtungabshängiger, Diffusion lässt sich weitgehend analog mit einem Matrix-wertigem D behandeln.)

Konvektion, Absorption und Quellterme lasen sich wie im Eindimensionalen in die Gleichung integrieren (wobei die Geschwindigkeit $v(x,t)$ nun ein Vektor sei, dessen Einträge die Geschwindigkeit in die jeweilige Richtung darstellen):

$$\begin{aligned}\frac{\partial C}{\partial t}(x,t) = \nabla \cdot (D(x,t)\nabla C(x,t)) \\ - \nabla \cdot (v(x,t)C(x,t)) - a(x,t)C(x,t) + M(x,t).\end{aligned}$$

2.1.2 Typen von Differentialgleichungen

Die Diffusionsmotivation enthält bereits drei der vier wichtigsten speziellen partiellen Differentialgleichungen:

(a) Treten nur Diffusionsphänomene auf, so erhalten wir

$$\frac{\partial C}{\partial t}(x,t) = \nabla \cdot (D(x,t)\nabla C(x,t)) \,.$$

Diese Gleichung und insbesondere ihr Spezialfall (bei dem wir die gesuchte Funktion mit $u(x,t)$ und ihre zeitliche Ableitung mit $u_t(x,t)$ bezeichnen)

$$u_t = \Delta u.$$

heißt *Wärmeleitungsgleichung* (engl.: heat equation). Sie ist das Musterbeispiel einer sogenannten *parabolischen Gleichung*, bei der sich eine zu Beginn gegebene Konzentrations- (oder Temperatur-)verteilung mit der Zeit immer gleichmäßiger verteilt.

Intuitiv erscheint es sinnvoll, die Gleichung mit Anfangsbedingungen $u(x,t)|_{t=0}$, $x \in \Omega \subset \mathbb{R}^n$ und Randbedingungen $u(x,t)|_{x\in\partial\Omega}$ zu kombinieren.

(b) Wie in Abschnitt 1.8.1 erwarten wir intuitiv, dass (wenn alle Parameter, Randvorgaben und Quellen zeitlich konstant sind) sich eine Temperatur- oder Konzentrationsverteilung immer mehr einem Gleichgewichtszustand annähert. In diesem würde die zeitliche Ableitung verschwinden und wir erhalten

$$0 = \nabla \cdot (D(x)\nabla C(x)) \,.$$

Diese Gleichung und insbesondere ihr Spezialfall (bei dem wir die gesuchte Funktion wieder mit $u(x,t)$ bezeichnen)

$$\Delta u = 0$$

heißt *Laplace-Gleichung*. Die Variante, bei der noch äußere Quellen vorhanden sind, also

$$-\Delta u = f,$$

heißt auch *Poisson-Gleichung*. Dies sind die Musterbeispiele sogenannter *elliptischer Gleichung*, die den Gleichgewichtszustand eines Diffusionsprozesses beschreiben.

Intuitiv erscheint es sinnvoll, die Gleichung mit Randbedingungen

$$u(x,t)|_{x\in\partial\Omega}$$

zu kombinieren.

(c) Wird der Stoff lediglich mit der Geschwindigkeit $v(x,t)$ transportiert (nur Konvektion, keine Diffusion), so erhalten wir

$$\frac{\partial C}{\partial t}(x,t) = -\nabla\cdot(v(x,t)C(x,t)).$$

Diese Gleichung und insbesondere ihr Spezialfall konstanter Geschwindigkeit (bei dem wir die gesuchte Funktion wieder mit $u(x,t)$ bezeichnen)

$$u_t = -v\cdot\nabla u$$

heißt *Transport-Gleichung*. Sie ist das Musterbeispiel einer sogenannten *hyperbolischen Gleichung*, bei der die Masse lediglich transportiert wird.

Intuitiv erscheint es sinnvoll, die Gleichung mit Anfangsbedingungen $u(x,t)|_{t=0}$, $x\in\Omega\subset\mathbb{R}^n$ und Randbedingungen $u(x,t)|_{x\in\Gamma}$, $\Gamma\subseteq\partial\Omega$ auf dem ganzen Rand oder zumindest einem *einfallenden Teil des Randes* zu kombinieren.

(d) Um die vierte wichtige spezielle PDGL zu motivieren, stellen wir uns vor, dass $u(x,t)$ die Auslenkung einer Gitarrenseite beschreibe, vgl. die in der Vorlesung gemalten Skizzen. Ähnlich wie bei einem Diffusionsprozess zieht eine starke Auslenkung an einer Stelle (etwa nach oben) die danebenliegenden weniger ausgelenkten Punkte mit nach oben. Dies geschieht jedoch nicht durch Herüberwandern von Teilchen wie bei einem Diffusionprozess, sondern durch elastische Kräfte mit denen nebenliegende Punkte *beschleunigt* werden. Die Beschleunigung ist die zweite

Ableitung der Auslenkung und so ergibt sich ähnlich wie bei der Diffusionsgleichung die sogenannte *Wellengleichung*

$$u_{tt} = \Delta u.$$

Diese Gleichung wird ebenfalls als *hyperbolische Gleichung* bezeichnet, die Auslenkung scheint sich wie durch einen Transportprozess auszubreiten.

Bemerkung 2.1
Wir betrachten eine lineare partielle Differentialgleichung 2. Ordnung der Form

$$\sum_{i,j=1}^{n} a_{ij}(x)\frac{\partial^2}{\partial x_i \partial x_j}u(x) + \sum_{i=1}^{n} b_i(x)\frac{\partial}{\partial x_i}u(x) + c(x)u(x) = f(x),$$

wobei (o.B.d.A) $a_{ij}(x) = a_{ji}(x)$. Die Abbildung

$$L: \; u \mapsto \sum_{i,j=1}^{n} a_{ij}(x)\frac{\partial^2}{\partial x_i \partial x_j}u(x) + \sum_{i=1}^{n} b_i(x)\frac{\partial}{\partial x_i}u(x) + c(x)u(x)$$

bezeichnet man auch als Differentialoperator*. Der Term mit den höchsten Ableitungen $\sum_{i,j=1}^{n} a_{ij}(x)\frac{\partial^2}{\partial x_i \partial x_j}u(x)$ wird als* Hauptteil *bezeichnet. (Beachte, dass für eine rigorose mathematische Definition der Abbildung noch der Ausgangs- und Zielraum festgelegt werden muss.)*

Die zum Hauptteil gehörige symmetrische Matrix $A(x) = (a_{ij}(x))_{i,j=1}^{n} \in \mathbb{R}^{n\times n}$ bestimmt den Typ der Differentialgleichung. Die Gleichung heißt

- elliptisch in x*, falls alle Eigenwerte von A positiv oder alle negativ sind.*
- hyperbolisch in x*, falls genau $n-1$ Eigenwerte positiv sind und einer negativ ist, oder $n-1$ Eigenwerte negativ sind und einer positiv ist.*
- parabolisch in x*, falls ein Eigenwert Null ist und die anderen $n-1$ Eigenwerte entweder alle positiv oder alle negativ sind.*

2.2 Finite Differenzen für elliptische Differentialgleichungen

Wir beginnen mit der numerischen Lösung von Gleichungen, die Gleichgewichtszustände beschreiben und betrachten dazu exemplarisch

$$-\Delta u(x) = f(x) \tag{2.1}$$

in einer beschränkten offenen Menge $\Omega \subseteq \mathbb{R}^n$. Dabei sei $f \in C(\Omega)$ eine stetige Funktion. Entsprechend der Modellierung aus dem letzten Abschnitt können wir uns f als die Verteilung angelegter Wärmequellen vorstellen und die Lösung u beschreibt dann die sich im Gleichgewicht einstellende Temperatur.

Es ist anschaulich klar, dass für die Gleichgewichtsverteilung der Temperatur auch der Rand des Gebiets $\partial\Omega$ eine Rolle spielen wird, etwa wenn dieser Rand immer auf einer konstanten Temperatur gehalten wird. Tatsächlich werden wir sehen, dass u durch (2.1) und Vorgabe von $u|_{\partial\Omega}$ eindeutig bestimmt ist.

Damit die Gleichung (2.2) und eventuelle Randwerte überhaupt einen Sinn ergeben, betrachten wir als Lösungskandidaten nur Funktionen $u \in C^2(\Omega) \cap C(\overline{\Omega})$ (sogenannte *klassische Lösungen*). Lösungen der Laplace-Gleichung $\Delta u = 0$ heißen auch *harmonische* Funktionen.

2.2.1 Das Maximumsprinzip

Satz 2.2 (Maximumsprinzip)
Es sei $f \in C(\Omega)$ punktweise nicht-positiv und die Funktion $u \in C^2(\Omega) \cap C(\overline{\Omega})$ erfülle

$$-\Delta u(x) = f(x) \leq 0 \quad \forall x \in \Omega. \tag{2.2}$$

Dann nimmt u sein Maximum auf dem Rand $\partial\Omega$ an (d.h. mindestens ein globales Maximum von u liegt auf $\partial\Omega$).

Beweis: (i) Betrachte zunächst den Fall $f(x) < 0$ für alle $x \in \Omega$.

Angenommen es gibt ein inneres Maximum, also $y \in \Omega$ mit

$$u(y) \geq u(x) \quad \forall x \in \overline{\Omega}.$$

Dann ist y insbesondere ein Maximum in jeder Koordinatenrichtung, also

$$\frac{\partial^2}{\partial x_j^2} u(y) \leq 0 \quad j = 1, \ldots, n$$

und damit $-\Delta u \geq 0$, was $-\Delta u = f < 0$ widerspricht. u kann also kein innereres Maximum haben. Da u als stetige Funktion auf dem Kompaktum $\overline{\Omega}$ aber mindestens ein Maximum besitzt, muss ein Maximum auf dem Rand liegen.

(ii) Nun sei $f(x) \leq 0$ für alle $x \in \Omega$. Angenommen es liegt kein Maximum auf dem Rand, dann gibt es ein inneres Maximum $y \in \Omega$ mit

$$u(y) \geq u(x) \quad \forall x \in \Omega \quad \text{und} \quad u(y) > u(x) \quad \forall x \in \partial\Omega.$$

Mit diesem y definieren wir die Funktion

$$h(x) := \|x - y\|^2 = \sum_{j=1}^{n} (x_i - y_i)^2.$$

Da Ω beschränkt ist, ist $\partial\Omega$ kompakt. h ist also auf $\partial\Omega$ beschränkt und es gilt $h(y) = 0$. Für hinreichend kleines $\delta > 0$ nimmt deshalb

$$w(x) := u(x) + \delta h(x)$$

sein Maximum nicht auf dem Rand an. Aus $\Delta h(x) = 2n$ folgt aber

$$-\Delta w(x) = f - 2n\delta < 0,$$

und wir erhalten den Widerspruch aus der in Teil (i) gezeigten Aussage. □

Satz 2.3
Sei $f \in C(\Omega)$.

*(a) Ist $f \geq 0$ und $u \in C^2(\Omega) \cap C(\overline{\Omega})$ eine Lösung von $-\Delta u = f \geq 0$ in Ω, so nimmt u sein Minimum auf dem Rand $\partial\Omega$ an (*Minimumsprinzip*).*

(b) Gilt für $u, v \in C^2(\Omega) \cap C(\overline{\Omega})$

$$-\Delta u \leq -\Delta v \text{ in } \Omega \quad \text{und} \quad u \leq v \text{ auf } \partial\Omega$$

so gilt $u \leq v$ in ganz Ω.

(c) Die Nullfunktion ist die einzige Lösung $u \in C^2(\Omega) \cap C(\overline{\Omega})$ von

$$-\Delta u = 0, \qquad u|_{\partial\Omega} = 0.$$

Eine Lösung $u \in C^2(\Omega) \cap C(\overline{\Omega})$ von

$$-\Delta u = f$$

ist also (wenn sie existiert) eindeutig durch f und $u|_{\partial\Omega}$ bestimmt.

(d) Erfüllen $u_1, u_2 \in C^2(\Omega) \cap C(\overline{\Omega})$

$$-\Delta u_1 = f = -\Delta u_2$$

so ist

$$\|u_1 - u_2\|_\infty := \max_{x \in \overline{\Omega}} |u_1(x) - u_2(x)| = \max_{x \in \partial\Omega} |u_1(x) - u_2(x)|.$$

Die Lösungen des Dirichlet-Problems hängen also (so sie denn existieren) stetig von den vorgegebenen Dirichlet-Randdaten ab.

(e) Es existiert ein (von der Menge Ω abhängiges) $C > 0$, sodass für alle $u \in C^2(\Omega) \cap C(\overline{\Omega})$

$$\|u\|_\infty = \max_{x\in\overline{\Omega}} |u(x)| \leq \max_{x\in\partial\Omega} |u(x)| + C \sup_{x\in\Omega} |\Delta u|.$$

In diesem Sinne hängt eine Lösung von $\Delta u = f$ (so sie denn existiert) also auch stetig von der rechten Seite ab.

(f) Sind $c, f \in C(\Omega)$, $c \geq 0$, $f \leq 0$ und $u \in C^2(\Omega) \cap C(\overline{\Omega})$ erfüllt

$$-\Delta u + cu = f \leq 0,$$

so gilt

$$\max_{x\in\overline{\Omega}} u(x) \leq \max\{0, \max_{x\in\partial\Omega} u(x)\}$$

Beweis: (a) folgt aus Anwendung des Maximumprinzips auf $-u$.

(b) folgt aus Anwendung des Maximumprinzips auf $u - v$.

(c) folgt aus Anwendung des Maximumsprinzips und des Minimumsprinzips.

(d) folgt aus Anwendung des Maximums- und Minimumprinzips auf $u_1 - u_2$.

(e) Für $\sup_{x\in\Omega} |\Delta u(x)| = \infty$ ist die Aussage erfüllt. Sei also $\Delta u(x)$ beschränkt.

Wähle $R > 0$ so groß, dass $\Omega \subseteq B_R(0) := \{x \in \mathbb{R}^n \mid \|x\| < R\}$. Für

$$w(x) := R^2 - \frac{1}{n} \|x\|^2$$

gilt $-\Delta w(x) = 2$ und $w(x) \geq 0$ für alle $x \in \overline{\Omega}$.

Definiere außerdem

$$v(x) := \max_{z\in\partial\Omega} |u(z)| + \frac{w(x)}{2} \sup_{z\in\Omega} |\Delta u(z)|.$$

Dann ist

$$-\Delta v(x) = \sup_{z\in\Omega} |\Delta u(z)| \geq -\Delta u(x) \ \forall x \in \Omega \quad \text{und} \quad v|_{\partial\Omega} \geq u|_{\partial\Omega}$$

also nach (b) $u \leq v$ auf Ω. Genauso folgt $u \geq -v$ auf Ω und damit

$$\begin{aligned} |u(x)| \leq |v(x)| = v(x) &= \max_{z\in\partial\Omega} |u(z)| + \frac{w(x)}{2} \sup_{z\in\Omega} |\Delta u(z)| \\ &\leq \max_{z\in\partial\Omega} |u(z)| + \frac{R^2}{2} \sup_{z\in\Omega} |\Delta u(z)|, \end{aligned}$$

also folgt die Behauptung mit $C := \frac{R^2}{2}$.

(f) Wir definieren die (möglicherweise leere) Menge

$$O := \{x \in \Omega : \ u(x) > 0\}.$$

O ist das Urbild des offenen Intervalls $(0, \infty)$ unter der stetigen Abbildung $u : \ \Omega \to \mathbb{R}$ und damit offen in der Relativtopologie von Ω. Da Ω offen in $\mathbb{R}^n$ ist, ist auch O offen in $\mathbb{R}^n$.

Zur Anwendung des Maximumsprinzips charakterisieren wir noch ∂O. Da O offen ist, gilt für jedes $x \in \partial O$, dass $x \notin O$ und damit entweder $x \notin \Omega$ oder $u(x) \leq 0$. Im ersten Fall $x \notin \Omega$ muss $x \in \partial\Omega$ gelten, da aus $O \subseteq \Omega$ folgt dass

$$\partial O \subseteq \overline{O} \subseteq \overline{\Omega} = \Omega \cup \partial\Omega.$$

Insgesamt ist also

$$\partial O \subseteq \partial\Omega \cup \{x \in \Omega : \ u(x) \leq 0\}.$$

Man kann leicht zeigen, dass sogar $\partial O \subseteq \partial\Omega \cup \{x \in \Omega : \ u(x) = 0\}$ gilt, aber das benötigen wir im Folgenden nicht.

Nun können wir die Behauptung

$$\max_{x \in \overline{\Omega}} u(x) \leq \max\{0, \max_{x \in \partial\Omega} u(x)\} \tag{2.3}$$

beweisen. Ist $O = \emptyset$, so ist $\max_{x \in \Omega} u(x) \leq 0$ und wegen der Stetigkeit von u gilt dies auch auf $\overline{\Omega}$, so dass (2.3) folgt. Ansonsten gilt dass

$$-\Delta u(x) = f(x) - c(x)u(x) \leq 0 \quad \text{ für alle } x \in O \neq \emptyset,$$

und aus der Konstruktion von O sowie dem Maximumsprinzip Satz 2.2 folgt, dass ein Punkt $\hat{x} \in \partial O$ existiert mit

$$\max_{x \in \overline{\Omega}} u(x) = \max_{x \in \overline{O}} u(x) \leq u(\hat{x}).$$

Mit der obigen Charakterisierung folgt, dass entweder $\hat{x} \in \partial\Omega$ oder $u(\hat{x}) \leq 0$ gilt. In beiden Fällen folgt (2.3). □

2.2.2 Finite Differenzen

Wie in Abschnitt 1.8.2 betrachten wir finite Differenzen und definieren für eine Funktion $u(x)$

$$\begin{aligned}
D^+_{h,i}[u](x) &:= \frac{u(x+he_i)-u(x)}{h},\\
D^-_{h,i}[u](x) &:= \frac{u(x)-u(x-he_i)}{h},\\
D_{h,i}[u](x) &:= \frac{u(x+he_i)-u(x-he_i)}{2h},\\
D^2_{h,i}[u](x) &:= D^+_{h,i}D^-_{h,i}[u](x) = D^-_{h,i}D^+_{h,i}[u](x)\\
&= \frac{u(x+he_i)-2u(x)+u(x-he_i)}{h^2}.
\end{aligned}$$

Wir verwenden für mehrfache partielle Ableitungen auch die folgende *Multiindex-Notation*. Für $\alpha := (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}_0^n$ ist

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

und $|\alpha| = \alpha_1 + \ldots + \alpha_n$, $\alpha! = \alpha_1! \cdots \alpha_n!$. Für $\alpha, \beta \in \mathbb{N}_0^n$, $\beta \le \alpha$ ist außerdem $\binom{\alpha}{\beta} = \frac{\alpha!}{\beta!(\alpha-\beta)!}$ wobei

$$\alpha \le \beta \quad :\Longleftrightarrow \quad \alpha_j \le \beta_j \quad \text{für alle } j = 1, \ldots, n.$$

Mit dieser Notation definieren wir für $u \in C^k(\overline{\Omega})$ die Halbnorm

$$|u|_{C^k(\overline{\Omega})} := \max_{|\alpha|=k} \max_{x\in\overline{\Omega}} |D^\alpha u(x)|$$

und die Norm

$$\|u\|_{C^k(\overline{\Omega})} := \max_{|\alpha|\le k} \max_{x\in\overline{\Omega}} |D^\alpha u(x)|.$$

Man rechnet leicht nach, dass letzteres tatsächlich eine Norm ist, die $C^k(\overline{\Omega})$ zu einem Banachraum (also einem vollständigen normierten Vektorraum) macht.

Lemma 2.4
Sei $x \in \Omega$, $j \in \{1, \ldots, n\}$ und $h > 0$ hinreichend klein, sodass

$$x \pm the_j \in \Omega \quad \textit{für alle } t \in [-1, 1].$$

(a) Für $u \in C^2(\overline{\Omega})$ *ist* $\left| D^+_{h,i}[u](x) - \dfrac{\partial}{\partial x_i} u(x) \right| \leq \dfrac{h}{2} |u|_{C^2(\overline{\Omega})}$.

(b) Für $u \in C^2(\overline{\Omega})$ *ist* $\left| D^-_{h,i}[u](x) - \dfrac{\partial}{\partial x_i} u(x) \right| \leq \dfrac{h}{2} |u|_{C^2(\overline{\Omega})}$.

(c) Für $u \in C^3(\overline{\Omega})$ *ist* $\left| D_{h,i}[u](x) - \dfrac{\partial}{\partial x_i} u(x) \right| \leq \dfrac{h^2}{6} |u|_{C^3(\overline{\Omega})}$.

(d) Für $u \in C^4(\overline{\Omega})$ *ist* $\left| D^2_{h,i}[u](x) - \dfrac{\partial^2}{\partial x_i^2} u(x) \right| \leq \dfrac{h^2}{12} |u|_{C^4(\overline{\Omega})}$.

Beweis: Das folgt wie im Beweis von Lemma 1.42 durch Taylorentwicklung. □

Beispiel 2.5
Wir betrachten die Gleichung

$$-\Delta u = f \quad \text{in } \Omega \quad \text{und} \quad u|_{\partial\Omega} = 0,$$

zu gegebenen Quelltermen $f : \Omega \to \mathbb{R}$, *wobei* $\Omega = (0,1)^2 \subset \mathbb{R}^2$ *ein zweidimensionales Quadrat mit Seitenlänge* 1 *sei.*

Wir diskretisieren Ω *durch ein Punktegitter mit der Schrittweite*

$$h = 1/(k+1), \quad k \in \mathbb{N},$$

und ordnen die inneren Punkte entsprechend der in der Vorlesung gemalten Skizze an

$$x^{(i+(j-1)k)} := (ih, jh) \in \Omega, \quad i, j = 1, \ldots, k.$$

Wir setzen noch $f_h = (f^{(i)})_{i=1,\ldots,k^2} \in \mathbb{R}^{k^2}$, $f^{(i)} = f(x^{(i)})$ *und versuchen einen Vektor von Approximationen*

$$u_h = (u^{(i)})_{i=1,\ldots,k^2} \in \mathbb{R}^{k^2}, \quad u^{(i)} = u(x^{(i)})$$

zu finden. Für jeden inneren Gitterpunkt $x^{(i)} \in \Omega$ *erhalten wir durch die obigen Differenzenverfahren 2. Ordnung die lineare Gleichung*

$$\begin{aligned} f^{(i)} = f(x^{(i)}) &= -\Delta u|_{x=x^{(i)}} \approx -D^2_{h,1} u(x)|_{x=x^{(i)}} - D^2_{h,2} u(x)|_{x=x^{(i)}} \\ &= -\frac{1}{h^2} \big(u(x^{(i)} + he_1) - 2u(x^{(i)}) + u(x^{(i)} - he_1) \\ &\qquad + u(x^{(i)} + he_2) - 2u(x^{(i)}) + u(x^{(i)} - he_2) \big) \\ &\approx \frac{1}{h^2} \left(4u^{(i)} - u^{(i-1)} - u^{(i+1)} - u^{(i-k))} - u^{(i+k)} \right), \end{aligned}$$

falls alle benachbarten Punkte ebenfalls innere Punkte sind. Ist ein benachbarter Punkt ein Randpunkt, so erhalten wir einen analogen Ausdruck, bei dem der zum Randpunkt gehörige Summand (wegen $u|_{\partial\Omega} = 0$) fehlt.

Insgesamt erhalten wir so ein lineares Gleichungssystem

$$A_h u_h = f_h$$

wobei die Matrix A_h die folgende Blocktridiagonalgestalt besitzt

$$A_h = \frac{1}{h^2} \begin{pmatrix} C & -I & & & \\ -I & C & -I & & \\ & -I & \ddots & \ddots & \\ & & \ddots & \ddots & -I \\ & & & -I & C \end{pmatrix} \in \mathbb{R}^{k^2 \times k^2}$$

mit

$$C := \begin{pmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 4 \end{pmatrix} \in \mathbb{R}^{k \times k},$$

und der $k \times k$-Einheitsmatrix I.

Bemerkung 2.6
Für zwei Matrizen $K = (k_{i,j}) \in \mathbb{R}^{l \times m}$ und $L \in \mathbb{R}^{r \times s}$ ist das Kronecker Produkt *definiert durch*

$$K \otimes L = \begin{pmatrix} k_{1,1}L & k_{1,2}L & \dots & k_{1,m}L \\ k_{2,1}L & k_{2,2}L & \dots & k_{2,m}L \\ \vdots & \vdots & & \vdots \\ k_{l,1}L & k_{l,2}L & \dots & k_{l,m}L \end{pmatrix} \in \mathbb{R}^{lr \times ms}.$$

Damit lässt sich die Matrix A_h aus 2.5 schreiben als

$$A_h = \frac{1}{h^2} (I \otimes T + T \otimes I)$$

mit

$$T := \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{k \times k}.$$

Inhomogene Dirichlet-Probleme

$$-\Delta u = f, \qquad u|_{\partial\Omega} = g$$

lassen sich analog behandeln, indem entweder der Effekt der Randpunkte auf die Differenzenverfahren der randnahen Punkte in der rechten Seite berücksichtigt wird, oder indem die Randpunkte als Unbekannte hinzugenommen werden und für jeden Randpunkt $x^{(j)}$ die Gleichung

$$u^{(j)} = g(x^{(j)})$$

aufgenommen wird.

2.2.3 Allgemeinere Fälle und ein diskretes Maximumsprinzip

Allgemeinere Gebiete $\Omega \subset \mathbb{R}^n$ lassen sich analog behandeln, indem wir über Sie ein Gitter mit Maschenweite h legen

$$\Omega_h := \{x = hk, \quad k \in \mathbb{Z}^n\} \cap \Omega$$

und zusätzlich diejenigen Randpunkte dazunehmen, die eine der Gitterlinien schneiden

$$\Gamma_h := \left\{x = (x_1, \ldots, x_n)^T \in \partial\Omega, \quad \exists i \in \{1, \ldots, n\} \ : \ x_i = hk, \ k \in \mathbb{Z}\right\}$$

(vgl. das in der Vorlesung gemalte Bild). Für die *randnahen* Punkte (also denen die Nachbarn in Γ_h besitzen) lassen sich finite Differenzen aufstellen, die in unterschiedliche Richtungen verschiedene Schrittweiten besitzen:

Lemma 2.7 (Shortley-Weller-Approximation)
Sei $x \in \Omega \subseteq \mathbb{R}^n$, $n = 1$ oder $n = 2$. Sei $u \in C^3(\overline{\Omega})$. Dann existiert $C > 0$, sodass für $h > 0$ und $h_O, h_W, h_N, h_S \leq h$ (die hinreichend klein seien, sodass die Auswertungen von u definiert sind)

(a) für $n = 1$

$$\left|\frac{2}{h_O(h_O + h_W)}u_O - \frac{2}{h_O h_W}u_Z + \frac{2}{h_W(h_O + h_W)}u_W - u''(x)\right| \leq C|u|_{C^3(\overline{\Omega})}h,$$

wobei $u_Z := u(x)$, $u_W := u(x - h_W)$ und $u_O := u(x + h_O)$.

(b) für $n = 2$

$$\left| \frac{2}{h_O(h_O+h_W)}u_O + \frac{2}{h_W(h_O+h_W)}u_W + \frac{2}{h_S(h_S+h_N)}u_S + \frac{2}{h_N(h_S+h_N)}u_N - \left(\frac{2}{h_O h_W} + \frac{2}{h_S h_N}\right) u_Z - \Delta u(x) \right| \leq C|u|_{C^3(\overline{\Omega})} h,$$

wobei $u_Z := u(x)$, $u_W := u(x - h_W e_1)$, $u_O := u(x + h_O e_1)$, $u_N := u(x + h_N e_2)$ und $u_S := u(x - h_S e_2)$.

Beweis: Übungsaufgabe 11.3. □

Offensichtlich lassen sich auch allgemeinere Gleichungen (mit veränderlichen Diffusionskoeffizienten, Absorptions- und Konvektionstermen) analog behandeln und führen wiederum auf lineare Gleichungssysteme der Form

$$A_h u_h = f_h.$$

Die so entstehenden Diskretisierungsmatrizen A_h sind üblicherweise von der Gestalt, dass die Diagonaleinträge positiv sind, die Nebendiagonaleinträge negativ sind, und zeilenweise der Diagonaleintrag die Zeile dominiert. Wir werden sehen, dass solche Matrizen ein diskretes Maximumsprinzip erfüllen. Zur Motivation betrachten wir zunächst eine Gleichung der Gestalt

$$4u_Z - u_W - u_O - u_S - u_N = 0 \tag{2.4}$$

die einen postiven und ansonsten nur negative Koeffizienten enthält, und bei der die Summe aller Koeffizienten Null ist. Dies kann so interpretiert werden kann, dass

$$u_Z = \frac{1}{4}(u_W + u_O + u_S + u_N),$$

d.h. der Eintrag mit dem positiven Koeffizienten ist der Mittelwert der Einträge mit den negativen Koeffizienten ist. u_Z kann daher nicht größer als $\max\{u_W, u_O, u_S, u_N\}$ sein, und $u_Z = \max\{u_W, u_O, u_S, u_N\}$ ist nur möglich im Falle $u_Z = u_W = u_O = u_S = u_N$. Dieses Maximumsargument bleibt offenbar auch gültig, wenn die rechte Seite (2.4) nicht-positiv ist. Unter der Zusatzannahme, dass $u_Z \geq 0$ ist, bleibt das Maximumsargument auch gültig, wenn die Summe aller Koeffizienten größer Null ist.

Lemma 2.8 (Sternlemma)
Sei $L \geq 0$ und α_l, x_l, $l = 0, \ldots, L$ erfüllen

$$\alpha_l < 0 \quad \forall l > 0, \qquad \sum_{l=0}^{L} \alpha_l \geq 0, \qquad \sum_{l=0}^{L} \alpha_l x_l \leq 0 \quad \text{und} \quad x_0 \geq 0.$$

Ist $x_0 \geq \max_{l=1,\ldots,L} x_l$, so gilt $x_0 = x_1 = \ldots = x_L$.

Beweis: Für $L = 0$ ist die Aussage trivial. Ansonsten ist

$$\sum_{l=1}^{L} \underbrace{\alpha_l}_{<0} \underbrace{(x_l - x_0)}_{\leq 0} = \sum_{l=0}^{L} \alpha_l (x_l - x_0) = \underbrace{\sum_{l=0}^{L} \alpha_l x_l}_{\leq 0} - x_0 \underbrace{\sum_{l=0}^{L} \alpha_l}_{\geq 0} \leq 0$$

und es folgt $x_l = x_0$ für alle $l = 1, \ldots, L$. □

Offenbar kann für $\sum_{l=0}^{L} \alpha_l = 0$ auf die Voraussetzung $x_0 \geq 0$ in Lemma 2.8 verzichtet werden.

Satz 2.9
Sei $A_h \in \mathbb{R}^{N \times N}$ eine (nicht-notwendigerweise strikt) diagonaldominante Matrix mit positiven Diagonal- und negativen Nebendiagonaleinträgen, also

$$a_{ii}^h \geq \sum_{i \neq j} |a_{ij}^h| \quad \forall i = 1, \ldots, N, \qquad a_{ij}^h \leq 0 \quad \forall i \neq j.$$

Sei $f_h \in \mathbb{R}^N$ und $u_h \in \mathbb{R}^N$ sei eine Lösung von $A_h u_h = f_h$. Außerdem sei $f_h \leq 0$, also komponentenweise nicht-positiv.

Betrachte die i-te Zeile des LGS $A_h u_h = f_h$,

$$\sum_{j=1}^{N} a_{ij}^h u_j^h = f_i^h \leq 0.$$

Ist $u_i^h \geq \max\{0, \max_{j:\ j \neq i, a_{ij} \neq 0} u_j^h\}$, so ist $u_i^h = u_j^h$ für alle j mit $a_{ij} \neq 0$.

Beweis: Wir setzen $\alpha_0 = a_{ii}^h$, $x_0 = u_i^h$. Entsprechend seien α_l und x_l, $l = 1, \ldots, k$ die anderen von Null verschiedenen Einträge a_{ij}^h, $j \neq i$ und dazugehörigen u_j^h. Dann folgt die Behauptung aus Lemma 2.8. □

Bemerkung 2.10
(a) Im Kontext unserer Finite-Differenzen-Diskretisierungen kann Theorem 2.9 als diskretes Maximumsprinzip *(in Analogie zu Theorem 2.3(f)) interpretiert werden. Die diskrete Lösung u_h kann nicht in einem Gitterpunkt einen nicht-negativen Wert annehmen, der strikt maximal ist unter allen im Rahmen der Differenzenquotienten betrachten Nachbarwerten. Mehr noch: nimmt die diskrete Lösung einen in diesem Sinne maximalen Wert an, so besitzen alle (in diesem Sinne) benachbarten Werte den gleichen maximalen Wert.*

(b) Mit dem diskreten Maximumsprinzip lässt sich oft die Lösbarkeit der diskretisierten Gleichung beweisen. Betrachten wir exemplarisch die Matrix A_h aus der Diskretisierung des Poisson-Problems auf dem Einheitsquadrat. A_h ist genau dann invertierbar, wenn es injektiv ist, also nur der Nullvektor das homogene LGS $A_h u_h = 0$ löst. Sei also u_h eine solche Lösung. Dann besitzt u_h einen maximalen Eintrag u_j^h. O.B.d.A. sei $u_h^j \geq 0$, ansonsten betrachten wir $-u_h$. Da wir jeden anderen Gitterpunkt über einen Weg aus (bei den Differenzenquotienten vorkommenden) Nachbarwerten erreichen können (das Gitter ist diskret zusammenhängend*), folgt aus dem diskreten Maximumsprinzip, dass alle Einträge von u_j^h übereinstimmen. Aus der ersten Zeile von A_h folgt dann, dass $u_h = 0$ gelten muss.*

(c) Auch für die Monotonieeigenschaft aus Theorem 2.3(b) existiert ein diskretes Analogon. In Lemma 1.44 hatten wir gezeigt, dass für jede invertierbare, diagonaldominante Matrix $A \in \mathbb{R}^{N\times N}$ mit positiven Diagonalelementen und nicht-positiven Nichtdiagonalelementen gilt, dass

$$Au \leq Av \quad \Longrightarrow \quad u \leq v.$$

2.2.4 Konsistenz, Stabilität und Konvergenz

Sei $A_h u_h = f_h$ die entsprechend den letzten Abschnitten erstellte Diskretisierung des Poisson-Problems

$$-\Delta u = f, \quad u|_{\partial\Omega} = g.$$

Der Einfachheit halber bezeichnen wir mit $u \in \mathbb{R}^N$ auch den Vektor der Auswertungen der Funktion $u(x)$ auf den Gitterpunkten der Diskretisierung.

Bereits in Abschnitt 1.8.3 haben wir den folgenden Zusammenhang zwischen Konsistenz, Stabilität und Konvergenz kennengelernt:

$$\|u - u_h\|_\infty = \left\|A_h^{-1} A_h (u - u_h)\right\|_\infty \leq \left\|A_h^{-1}\right\|_\infty \|A_h u - f_h\|_\infty .$$

Erfüllt also (für $h \to 0$) die wahre Lösung (genauer: ihre Auswertungen) immer besser die diskretisierte Gleichung (Konsistenz), d.h.

$$\lim_{h\to 0} \|A_h u - f_h\|_\infty = 0,$$

und bleibt $\left\|A_h^{-1}\right\|_\infty$ beschränkt (Stabilität), so nähert die Lösung der diskreten Gleichung die wahre Lösung immer besser an.

Die Konsistenz von finiten Differenzenverfahren erhalten wir sofort aus Abschätzungen, wie sie in Lemma 2.4 und Lemma 2.7 vorkommen, die ihrerseits leicht aus Anwendungen der Taylor-Formel folgen. Konsistente Differenzenverfahren lassen sich deshalb in der Praxis meist sehr leicht für eine gegebene partielle Differentialgleichung konstruieren.

Der Beweis von Stabilität ist ungleich schwerer. Wir zeigen im Rahmen dieser Vorlesung nur exemplarisch am Poisson-Problem, wie sich (ähnlich wie in Abschnitt 1.8.3) mit Hilfe einer speziellen Lösung und der im letzten Abschnitt gezeigten Monotonieeigenschaft die Stabilität zeigen lässt.

Satz 2.11
Sei $\Omega \subseteq \mathbb{R}^2$ beschränkt. Dann existiert ein $C > 0$, sodass für die entsprechend den letzten Abschnitten (mit den Differenzenquotienten aus Lemma 2.4 und Lemma 2.7) erstellte Diskretisierungsmatrix A_h zum Poisson-Problem

$$-\Delta u = f, \quad u|_{\partial\Omega} = g$$

gilt $\left\|A_h^{-1}\right\|_\infty < C$ für alle $h > 0$.

Beweis: Sei $R > 0$ hinreichend groß, sodass $\Omega \subseteq B_R(0)$ und betrachte die Funktion $w(x) = \frac{1}{4}\left(R^2 - \|x\|^2\right)$. Offenbar gilt $-\Delta w = 1$.

Wir bezeichnen den Vektor der Auswertungen von $w(x)$ in den Gitterpunkten mit W. Für jeden nicht-randnahen Gitterpunkt x_j gilt, dass

$$(A_h W)_j = -D^2_{h,1}[w](x_j) - D^2_{h,2}[w](x_j) = -\Delta w(x_j) = 1,$$

wobei wir ausgenutzt haben, dass nach Lemma 2.4(d) wegen $|w|_{C^4(\overline{\Omega})} = 0$ die verwendeten Differenzenquotienten für die Funktion w exakt sind.

Nach Lemma 2.7 ist auch die Anwendung der Shortley-Weller-Differenzenquotienten in den randnahen Punkten wegen $|w|_{C^3(\overline{\Omega})} = 0$ exakt. Jedoch bleiben bei der Anwendung von A_h auf W in den randnahen Gitterpunkten die Randpunkte unberücksichtigt. Ist etwa x_j ein Punkt, an den links (und an keiner anderen Seite) ein Randpunkt x_W angrenzt, so ist

$$(A_h W)_j - \frac{2}{h_W(h_O + h_W)} w(x_W) = -\Delta w(x_j) = 1.$$

Da $w(x_W) \geq 0$ ist, gilt also auch in diesem Fall (und analog in allen anderen randnahen Punkten) $(A_h W)_j \geq 1$, sodass insgesamt

$$A_h W \geq \mathbb{1}$$

folgt. Damit ist

$$\left\|A_h^{-1}\right\|_\infty = \left\|A_h^{-1}\mathbb{1}\right\|_\infty \leq \left\|A_h^{-1}A_hW\right\|_\infty \leq \max_{x\in\overline{B_R(0)}} |w(x)| = \frac{R^2}{4},$$

sodass die Behauptung bewiesen ist. □

Folgerung 2.12
Falls eine Lösung des Poisson-Problems

$$-\Delta u = f, \quad u|_{\partial\Omega} = g$$

existiert, so konvergieren die durch die Differenzenquotienten aus Lemma 2.4 und Lemma 2.7 erhaltenen Approximationen u_h gegen die Lösung.

2.3 Finite Differenzen für parabolische Differentialgleichungen

Wir betrachten nun eine parabolische Gleichung der Form

$$u_t(x,t) = u_{xx}(x,t) - b(x,t)u_x(x,t) - c(x,t)u(x,t) + f(x,t)$$

für alle $x \in (0,1)$, $t \in (0,T)$ mit homogenen Dirichletrandbedingungen

$$u(0,t) = 0 = u(1,t) \quad \text{für alle } t \in (0,T)$$

und Anfangsbedingung $u(x,0) = u_0(x)$ für $x \in (0,1)$.

Für die in diesem Abschnitt entwickelte Theorie benötigen wir, dass eine hinreichend oft differenzierbare Lösung $u(x,t)$ existiert und dass b und c beschränkte Funktionen sind mit $c(x) \geq 0$.

Wie in Abschnitt 1.8 diskretisieren wir die räumlichen Ableitungen durch zentrale Differenzenquotienten auf einem äquidistanten Gitter $x_i = ih$, $i = 0,\ldots,n+1$ und erhalten so ein *semi-diskretisiertes* System

$$\partial_t U_h(t) = -L_hU_h(t) + F(t), \quad u_h(0) = u(x,0), \tag{2.5}$$

für Approximationen $U_h(t) : [0,T] \to \mathbb{R}^n$ an die zeitliche Entwicklung der Auswertungen

$$U(t) = (u(x_1,t),\ldots,u(x_n,t))^T$$

der wahren Lösung $u(x,t)$ in den räumlichen Gitterpunkten. Dieses Vorgehen nennt man auch *Linienmethode*, vgl. die in der Vorlesung gemalte Skizze.

Das semi-diskretisierte System (2.5) ist eine gewöhnliche Differentialgleichung und kann mit einem der im ersten Kapitel dieser Vorlesung besprochenen Verfahren gelöst werden. Da die Matrix L_h für kleine h stark negative Eigenwerte besitzt, sollte zur Lösung möglichst ein L-stabiles Verfahren verwendet werden, vgl. Übungsaufgabe 7.4.

Wir analysieren die Konvergenz für die beiden einfachsten uns bekannten Verfahren, dem expliziten und impliziten Euler-Verfahren. Für eine Zeitdiskretisierung $t_j = j\tau$, $j = 0, \ldots, m$ mit Zeitschrittweite $\tau := T/m$ ergeben sich die folgenden Iterationsvorschriften für die Approximationen $U_{h,j} \approx U_h(t_j) \in \mathbb{R}^n$:

(a) Explizites Euler-Verfahren:

$$\begin{aligned} U_{h,j+1} &= U_{h,j} + \tau \left(-L_h U_{h,j} + F(t_j)\right) \\ &= (I - \tau L_h) U_{h,j} + \tau F(t_j), \end{aligned}$$

(b) Implizites Euler-Verfahren:

$$\begin{aligned} U_{h,j+1} &= U_{h,j} + \tau \left(-L_h U_{h,j+1} + F(t_{j+1})\right), \quad \text{also} \\ (I + \tau L_h) U_{h,j+1} &= U_{h,j} + \tau F(t_{j+1}) \end{aligned}$$

wobei jeweils als Startwert $U_{h,0} = (u_0(x_1), \ldots, u_0(x_n))^T$ verwendet wird. Da L_h diagonaldominant ist, ist $(I + \tau L_h)$ offensichtlich strikt diagonaldominant und damit (nach Lemma 1.44) unabhängig von der Zeitschrittweite τ invertierbar.

Zur Konvergenzanalyse untersuchen wir diese Verfahren wieder zunächst auf Konsistenz, also wie gut die wahre Lösung die diskretisierte Gleichung erfüllt:

Lemma 2.13
Die Auswertungen der wahren Lösung $U(t) = (u(x_1, t), \ldots, u(x_n, t))^T$ erfüllen bzgl. der $\|\cdot\|_\infty$-Norm:

(a) $U(t_{j+1}) = (I - \tau L_h) U(t_j) + \tau F(t_j) + O(\tau^2) + O(\tau h^2)$.

(b) $(I + \tau L_h) U(t_{j+1}) = U(t_j) + \tau F(t_{j+1}) + O(\tau^2) + O(\tau h^2)$.

Beweis: Für die Auswertungen der wahren Lösung $U(t)$ gilt

$$U(t_{j+1}) = U(t_j) + \tau \partial_t U(t_j) + O(\tau^2)$$

$\partial_t U(t_j)$ enthält die Einträge $\partial_t u(x_i, t_j)$. Da

$$\partial_t u(x_i, t_j) = u''(x_i, t_j) - b(x_i, t_j) u'(x_i, t_j) - c(x_i, t_j) u(x_i, t_j) + f(x, t),$$

folgt mit der Konsistenzabschätzung in Lemma 1.42 für die zentralen finite Differenzen im Ort, dass

$$\partial_t U(t_j) = -L_h U(t_j) + F(t_j) + O(h^2).$$

Damit ist

$$\begin{aligned} U(t_{j+1}) &= U(t_j) + \tau \partial_t U(t_j) + O(\tau^2) \\ &= U(t_j) + \tau(-L_h U(t_j) + F(t_j) + O(h^2)) + O(\tau^2), \end{aligned}$$

womit (a) bewiesen ist.

Mit $U(t_j) = U(t_{j+1}) - \tau \partial_t U(t_{j+1}) + O(\tau^2)$ folgt genauso

$$\begin{aligned} U(t_{j+1}) &= U(t_j) + \tau \partial_t U(t_{j+1}) + O(\tau^2) \\ &= U(t_j) + \tau(-L_h U(t_{j+1}) + F(t_{j+1}) + O(h^2)) + O(\tau^2), \end{aligned}$$

so dass auch (b) bewiesen ist. □

Um mit diesen Konsistenzabschätzungen auch Konvergenz zu zeigen, benötigen wir noch die folgenden Stabilitätsabschätzungen:

Lemma 2.14
(a) Es ist $\|I - \tau L_h\|_\infty \leq 1$ *für alle hinreichend kleinen* $h, \tau > 0$ *mit* $\frac{\tau}{h^2} < \frac{1}{2}$.

(b) Es ist $\|(I + \tau L_h)^{-1}\|_\infty \leq 1$ *für alle* $h, \tau > 0$.

Beweis: (a) Mit den Bezeichnungen aus Abschnitt 1.8.2 ist

$$\|I - \tau L_h\|_\infty = \max_{i=1,\ldots,n} \left(\frac{\tau}{h^2}|r_i| + |1 - \frac{\tau}{h^2} d_i| + \frac{\tau}{h^2}|s_i| \right),$$

wobei

$$d_i := 2 + h^2 c(x_i), \quad r_i := -1 - hb(x_i)/2 \quad \text{und} \quad s_i = -1 + hb(x_i)/2.$$

Für hinreichend kleine $h > 0$ ist $r_i < 0$ und $s_i < 0$. Außerdem gilt für alle hinreichend kleinen $h, \tau > 0$ mit $\tau/h^2 < \frac{1}{2}$

$$1 - \frac{\tau}{h^2} d_i = 1 - 2\frac{\tau}{h^2} - \tau c(x_i) > 0$$

und es folgt, dass

$$\begin{aligned} \|I - \tau L_h\|_\infty &= \max_{i=1,\ldots,n} \left(\frac{\tau}{h^2}(-r_i - s_i) + 1 - \frac{\tau}{h^2} d_i \right) \\ &= \max_{i=1,\ldots,n} (1 - \tau c(x_i)) \leq 1. \end{aligned}$$

(b) Da L_h diagonaldominant ist, ist jede Komponente von $L_h\mathbb{1}$ nicht-negativ und es folgt, dass $(I+\tau L_h)\mathbb{1} \geq \mathbb{1}$. Da $I+\tau L_h$ strikt diagonaldominant mit positiven Diagonalelementen und nicht-positiven Nebendiagonalelementen ist, folgt mit der Monotonieeigenschaft aus Lemma 1.44

$$\mathbb{1} = (I+\tau L_h)^{-1}(I+\tau L_h)\mathbb{1} \geq (I+\tau L_h)^{-1}\mathbb{1} \geq 0$$

und damit

$$\left\|(I+\tau L_h)^{-1}\mathbb{1}\right\|_\infty \leq \|\mathbb{1}\|_\infty = 1$$

für alle $h, \tau > 0$. □

Satz 2.15
Die Approximationen

$$U_{h,j} \approx (u(x_1,t_j),\ldots,u(x_n,t_j))^T = U(t_j)$$

des expliziten und impliziten Euler-Verfahrens konvergieren in folgendem Sinne gegen die Auswertungen der wahren Lösung der in diesem Abschnitt betrachteten parabolischen Differentialgleichung.

(a) Bei Anwendungen des expliziten Euler-Verfahrens

$$\|U_{h,j} - U(t_j)\|_\infty = O(h^2) + O(\tau) \quad \textit{für } h,\tau \to 0 \quad \textit{und} \quad \tau/h^2 < \frac{1}{2}.$$

(b) Bei Anwendungen des impliziten Euler-Verfahrens

$$\|U_{h,j} - U(t_j)\|_\infty = O(h^2) + O(\tau) \quad \textit{für } h,\tau \to 0.$$

Beweis: Für das explizite Euler-Verfahren folgt aus Lemma 2.13 und Lemma 2.14

$$\begin{aligned}
&\|U_{h,j} - U(t_j)\|_\infty \\
&= \|(I-\tau L_h)U_{h,j-1} + \tau F(t_j - 1) \\
&\qquad -(I-\tau L_h)U(t_{j-1}) - \tau F(t_{j-1}) + O(\tau^2) + O(\tau h^2)\|_\infty \\
&\leq \|I-\tau L_h\|_\infty \|U_{h,j-1} - U(t_{j-1})\|_\infty + O(\tau^2) + O(\tau h^2) \\
&\leq \|U_{h,j-1} - U(t_{j-1})\|_\infty + O(\tau^2) + O(\tau h^2) \\
&\leq \ldots \leq \|(U_{h,0} - U(t_0))\|_\infty + j(O(\tau^2) + O(\tau h^2)) \\
&\leq O(\tau) + O(h^2),
\end{aligned}$$

wobei wir im letzten Schritt $U_{h,0} = U(t_0)$ und $j \leq m = T/\tau$ verwendet haben.

Genauso erhalten wir für das implizite Euler-Verfahren aus Lemma 2.13 und Lemma 2.14

$$\begin{aligned}
&\|U_{h,j} - U(t_j)\|_\infty \\
&= \big\|(I + \tau L_h)^{-1}(U_{h,j-1} + \tau F(t_j)) \\
&\qquad -(I + \tau L_h)^{-1}(U(t_{j-1}) + \tau F(t_j) + O(\tau^2) + O(\tau h^2)\big\|_\infty \\
&\leq \big\|(I + \tau L_h)^{-1}\big\|_\infty \big\|U_{h,j-1} - U(t_{j-1}) + O(\tau^2) + O(\tau h^2)\big\|_\infty \\
&\leq \|U_{h,j-1} - U(t_{j-1})\|_\infty + O(\tau^2) + O(\tau h^2) \\
&\leq \ldots \leq \|(U_{h,0} - U(t_0))\|_\infty + j(O(\tau^2) + O(\tau h^2)) \\
&\leq O(\tau) + O(h^2).
\end{aligned}$$

wobei wir wieder im letzten Schritt $U_{h,0} = U(t_0)$ und $j \leq m = T/\tau$ verwendet haben. □

Bemerkung 2.16

(a) Im Gegensatz zu den gewöhnlichen DGL hat die bessere Stabilität des impliziten Euler-Verfahrens hier auch Auswirkungen auf die Konvergenz. Das explizite Euler-Verfahren konvergiert nur unter der Zusatzbedingung, dass Zeit- und Ortsschrittweite entsprechend der Vorschrift $\tau/h^2 < \frac{1}{2}$ gewählt sind. Das implizite Euler-Verfahren konvergiert ohne Zusatzbedingung für $\tau, h \to 0$.

(b) Da der Gesamtfehler jedoch in der Größenordnung $O(\tau) + O(h^2)$ liegt, ist es auch beim impliziten Euler-Verfahren sinnvoll $\tau \approx h^2$ zu wählen.

(c) In der Praxis ist auch die Anwendung des Crank-Nicolson-Verfahren zur Zeitdiskretisierung populär. Man kann zeigen, dass dieses ohne Zusatzbedingung konvergiert mit einem Gesamtfehler in der Größenordnung $O(\tau^2) + O(h^2)$. Das Verfahren ist jedoch lediglich A- und nicht L-stabil, sodass für kleine h Eigenvektoren von L_h zu sehr negativen Eigenwerten nicht schnell genug weggedämpft werden und Oszillationen auftreten können, vgl. wieder Übungsaufgabe 7.4.

Literaturverzeichnis

[FulfordBroadbridge] G. R. Fulford, P. Broadbridge: *Industrial Mathematics: Case Studies in the Diffusion of Heat and Matter*, Australian Mathematical Society Lecture Series 16, Cambridge University Press, Cambridge, 2002.

[Hanke] M. Hanke-Bourgeois: *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*, Teubner Verlag, Wiesbaden, 2009.

[HairerNorsettWanner] E. Hairer, S. P. Nørsett, G. Wanner: *Solving ordinary differential equations I. Nonstiff problems*, Springer, 1987.

[NumerikWS1718] B. Harrach: Vorlesungsskript *Numerische Mathematik*, Goethe-Universität Frankfurt am Main, WS17/18.
`http://numerical.solutions`

[Heuser] H. Heuser: *Gewöhnliche Differentialgleichungen*, Vieweg+Teubner Verlag, Wiesbaden, 2009.