Vorlesungsmitschrieb

Numerik von PDE II

Prof. Dr. Christian Lubich

im Sommersemester 2010
an der Eberhard-Karls-Universtität Tübingen

gesetzt von Markus Klein mit LaTeX

Letzte Änderung: 4. August 2010

Vorwort

Dieses Scriptum entstand als Live-Mitschrieb im Sommersemester 2010 bei Prof. Dr. Christian Lubich an der Eberhard-Karls-Universität Tübingen.
In der Vorlesung wurden noch zusätzlich numerische Verfahren in der Quantenmechanik behandelt, für die ich auf *From quantum to classical molecular dynamics: reduced models and numerical analysis* von Prof. Lubich verweise.

Dieses Scriptum erhebt keinen Anspruch auf Vollständigkeit oder Richtigkeit. Es ist *nicht* durch Prof. Lubich autorisiert. Weiter hält sich dieser Mitschrieb nicht an die Nummerierung in der Vorlesung.

Ich danke Dr. Ludwig Gauckler für seine Unterlagen und Notizen zur Vorlesung, die sehr hilfreich waren. Weiter danke ich Dhia Mansour, Martin Trick und Dr. Daniel Weiss für zahlreich Korrekturvorschläge und Hinweise.

Bei Fragen, Wünschen oder Verbesserungsvorschlägen freue ich mich über jede E-Mail an
klein@na.uni-tuebingen.de.
Vielen Dank!

Inhaltsverzeichnis

1 Steife Differentialgleichungen

1.1 Einführung

Wir möchten ein Standardverfahren (Runge–Kutta-Verfahren, Mehrschrittverfahren, etc.) mit Schrittweitensteuerung bei Problemen aus der Chemie, elektrischen Netzwerken, Wärmeleitung, etc. anwenden. Bei diesen Probleme laufen sowohl schnelle als auch langsame Phänomone ab. Die Standardverfahren mit adaptiver Schrittweitensteuerung machen auch in „langweiligen“ Bereichen sehr kleine Schrittweiten.

1.1 Wiederholung
Wir erinnern uns an explizite Verfahren aus der Numerik I, mit denen wir *Anfangswertprobleme* (AWP) von Differentialgleichungen,

$$\begin{cases} y'(t) = f(t, y(t)), & t \in [t_0, T], \\ y(t_0) = y_0 \end{cases}$$

behandelten, wobei $y : [t_0, T] \to \mathbb{R}^d$ die gesuchte Lösung ist, $y_0 \in \mathbb{R}^d$ der Startwert und $f : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ beliebig oft differenzierbar.

Beim *expliziten Euler-Verfahren* lösten wir näherungsweise $y(t_n) \approx y_n$ und $y_{n+1} = y_n + h_n f(t_n, y_n)$. Dieses Verfahren konvergierten mit Ordnung 1, i.e. $\|y_n - y(t_n)\| \leq Mh$, wobei $h = \max h_n$ und $M = \frac{e^{L^T} - 1}{L}$ und L die Lipschitz-Konstante von f war.

Für höhere Ordnungen betrachteten wir die Klasse von Runge–Kutta-Verfahren oder Mehrschrittverfahren.

Beim Lösen mit solch expliziten Verfahren wählt das Programm die Schrittweiten h_n sehr klein, obwohl das Lösungsverhalten sehr „glatt“ ist. Bei größeren Schrittweiten wurde numerisch die Lösungs sehr instabil (i.e. sie machte Sprünge), obwohl sie analytisch stabil ist.

Der Grund hierfür liegt im unterschiedlichen Stabilitätsverhalten der Differentialgleichung und der zugehörigen numerischen Methode. Hierzu betrachten wir ein Beispiel.

1.2 Beispiel
Wir betrachten die Differentialgleichung $y'(t) = f(t, y(t))$. Sei z die Lösung zum Anfangswert $z(t_0) = z_0$. Für eine beliebige andere Lösung der Differentialgleichung y gibt es mit dem

Mittelwertsatz ein ξ mit

$$y'(t) = f(t, y(t)) = \underbrace{f(t, z(t))}_{=z'(t)} + \partial_y f(t, \xi) \cdot (y(t) - z(t)). \tag{1.1}$$

Wir machen nun für die weitere Diskussion einige vereinfachten Annahmen, die zumindest in kleinen Umgebungen der Lösung z stimmen: Wir nehmen an, daß $\partial_y f(t, y) = A$ ist, wobei A eine konstante $d \times d$-Matrix ist und daß $z(t) = z_0$. Dann erhalten wir (wegen $z'(t) = 0$) aus (1.1) die Differentialgleichung

$$\begin{cases} y'(t) = A(y(t) - z_0), \\ y(t_0) = y_0, \end{cases}$$

die mit dem expliziten Euler-Verfahren mittels $y_{n+1} = y_n + hA(y_n - z_0)$ für ein gegebenes y_0 gelöst wird.

Die Fortpflanzung des Fehlers $e(t) := y(t) - z(t)$ ist durch $e'(t) = Ae(t)$ und $e(t_0) = e_0 := y_0 - z_0$ gegeben und im expliziten Euler-Verfahren durch $e_{n+1} = e_n + hAe_n$.

Falls A diagonalisierbar ist, gibt es eine Basiswechselmatrix V, so daß

$$V^{-1}AV =: \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$$

ist. Wir setzen nun $u := V^{-1}e$ und erhalten somit die Differentialgleichung

$$\begin{cases} u'(t) = \Lambda u(t), \\ u(t_0) = u_0 := V^{-1}e_0, \end{cases} \tag{1.2}$$

welche sich mit dem expliziten Euler-Verfahren durch $u_{n+1} = u_n + h\Lambda u_n$ lösen läßt.

In vielen Anwendungen gilt dabei für alle i, daß $\operatorname{Re}(\lambda_i) \leq 0$ und es gibt meist ein i mit $\operatorname{Re}(\lambda_i) \ll -1$. Wir betrachten nun die i-te Komponente von u in (1.2) und erhalten (für y die i-te Komponente von u)

$$\begin{cases} y'(t) = \lambda y, \\ y(t_0) = y_0, \end{cases} \tag{1.3}$$

welche numerisch durch $y_{n+1} = y_n + h\lambda y_n$, also $y_n = (1 + h\lambda)^n y_0$ gelöst wird.

Die exakte Lösung von (1.3) ist $y(t) = e^{\lambda(t-t_0)}y_0$, welche wegen $\operatorname{Re}(\lambda) \leq 0$ stets beschränkt ist (bzw. für $t \to \infty$ sogar gegen Null strebt, falls $\operatorname{Re}(\lambda) < 0$ ist). Die numerische Lösung hat dieses Verhalten allerdings nur für $|1 + h\lambda| \leq 1$ bzw. $|1 + h\lambda| < 1$.

Falls also $\operatorname{Re}(\lambda) \leq 0$ und $|\lambda| \gg 1$ ist, haben wir eine Schrittweitenbeschränkung bei der Wahl großer Schrittweiten, i.e. die numerische Lösung divergiert bei großen Schrittweiten.

1.3 Definition (A-Stabilität)
Ein numerisches Verfahren heißt *A-stabil*, falls die numerische Lösung $(y_n)_{n\geq 0}$ beschränkt bleibt, sofern das Verfahren mit beliebiger Schrittweite $h > 0$ auf die Differentialgleichung $y' = \lambda y$, $y(0) = y_0$ und $\operatorname{Re}(\lambda) \leq 0$ angewandt wird. Bei Mehrschrittverfahren muß dies für jede beliebige Wahl von Startwerten gelten.

1.4 Bemerkung
Mit Beispiel 1.2 ist das explizite Euler-Verfahren nicht A-stabil.

1.5 Beispiel (implizites Euler-Verfahren)
Wir betrachten das implizite Euler-Verfaren,

$$y_{n+1} = y_n + hf(t_{n+1}, y_{n+1}),$$

welches wir auf die Differentialgleichung $y' = \lambda y$ anwenden. Hierfür ergibt sich

$$y_{n+1} = y_n + h\lambda y_{n+1}.$$

Durch elementare Umformungen erhalten wir damit

$$y_n = \left(\frac{1}{1-h\lambda}\right)^n y_0,$$

i.e. für die A-Stabilität muß nun $\left|\frac{1}{1-h\lambda}\right| \leq 1$ gelten, was aber für jedes $h > 0$ und jedes λ mit $\mathrm{Re}(\lambda) \leq 0$ gilt.

Damit ist das implizite Euler-Verfahren A-stabil.

1.6 Bemerkung
Für praktische alle Verfahren (Runge–Kutta-Verfahren, Mehrschrittverfahren, etc.) sind die folgenden beiden Diagramme kommutativ, wobei $V^{-1}AV = \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ bezeichnet.

$$\begin{array}{ccc} y' = Ay,\ y_0 & \xrightarrow{y = Vz} & z' = \Lambda z,\ z_0 = V^{-1}y_0 \\ \text{num. Verfahren} \downarrow & & \downarrow \text{num. Verfahren} \\ (y_n)_{n\geq 0} & \xrightarrow{y = Vz} & (z_n)_{n\geq 0} \end{array}$$

$$\begin{array}{ccc} y' = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)y & \xrightarrow{\text{Projektion}} & y_i' = \lambda_i y_i \\ \text{num. Verfahren} \downarrow & & \downarrow \text{num. Verfahren} \\ (y_n)_{n\geq 0} & \xrightarrow{\text{Projektion}} & (y_{n,i})_{n\geq 0} \end{array}$$

Damit haben wir, falls ein Verfahren A-stabil ist, auch lineare Systeme im Griff.

1.2 Stabilitätsbereiche

1.7 Definition (Stabilitätsbereich)
Als *Stabilitätsbereich* eines numerischen Verfahrens bezeichnet man die Menge

$$S := \left\{ z = h\lambda \in \mathbb{C} : \begin{array}{l} \text{Verfahren liefert beschränkte numerische Lösung } (y_n)_{n\geq 0}, \\ \text{wenn es mit Schrittweite } h \text{ auf } y' = \lambda y \text{ mit beliebigen} \\ \text{Startwert } y_0 \text{ angewandt wird.} \end{array} \right\}.$$

1.8 Bemerkung
Die Definition ist nur sinnvoll, wenn die Lösung nur vom Produkt $h\lambda$ und nicht den einzelnen Faktoren h und λ abhängt.

1.9 Bemerkung
Ein Verfahren ist genau dann A-stabil, wenn $\mathbb{C}^- := \{z \in \mathbb{C} : \operatorname{Re} z \leq 0\} \subseteq S$ ist.

1.10 Bemerkung
Ein Mehrschrittverfahren ist genau dann 0-stabil, falls $0 \in S$ ist.

Wir möchten nun ein Verfahren mit unbeschränkten Stabilitätsbereich konstruieren. Hierfür betrachten wir zunächst die beiden Klassen an Verfaren, die wir kennen: Runge–Kutta-Verfahen und Mehrschrittverfaren.

1.11 Beispiel (Stabilitätsbereich beim expliziten Runge–Kutta-Verfahren)
Bei einem expliziten Runge–Kutta-Verfahren haben wir den Ansatz

$$\begin{aligned} y_{n+1} &= y_n + h\sum_{i=1}^{s} b_i f(t_n + c_i h_i Y_{n,i}), \text{ wobei} \\ Y_{n,i} &= y_n + h\sum_{j=1}^{i-1} a_{ij} f(t_n + c_j h_i Y_{n,j}), \qquad i = 1, \ldots, s. \end{aligned}$$

Dadurch, daß wir bei den $Y_{n,i}$ nur bis $i-1$ summieren, ist das Verfahren explizit. Für die Differentialgleichung $y' = \lambda y$ erhalten wir damit

$$\begin{aligned} Y_{n,i} &= y_n + h\lambda\sum_{j=1}^{i-1} a_{ij} Y_{n,j}, \qquad i = 1, \ldots, s, \\ y_{n+1} &= y_n + h\lambda\sum_{i=1}^{s} b_i Y_{n,i}, \end{aligned}$$

i.e. in diese Fall gilt $y_{n+1} = P(h\lambda)y_n$, wobei P ein Polynom vom Grad s ist, i.e. es gilt $y_n = P(h\lambda)^n y_0$. Dies ist aber genau dann beschränkt, wenn $|P(h\lambda)| \leq 1$ ist, also ist der Stabilitätsbereich $S = \{z \in \mathbb{C} : |P(z)| \leq 1\}$.

1.12 Satz
Für explizite Runge–Kutta-Verfahren ist der Stabilitätsbereich stets beschränkt.

1.13 Beispiel (Stabilitätsbereich beim Mehrschrittverfahren)
Bei Mehrschrittverfahren betrachten wir den Ansatz

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f_{n+j},$$

wobei $\alpha_k \neq 0$ und $f_{n+j} = f(t_{n+j}, y_{n+j})$ ist. Ein Mehrschrittverfahren ist genau dann explizit, wenn $\beta_k = 0$ ist. Typische Beispiele hierfür sind etwa das Adams-Verfahren (welches auf Integration beruht) und das BDF-Verfahren(welches auf der Differentiation beruht).

Bei der Anwendung auf $y' = \lambda y$ erhalten wir für $z = h\lambda$ damit

$$\sum_{j=0}^{k} (\alpha_j - z\beta_j) y_{n+j} = 0.$$

Ohne Beschränkung der Allgemeinheit nehmen wir an, daß $\mathrm{ggT}(\alpha, \beta) = 1$ ist (Übung!). Wir betrachten nun das charakteristische Polynom,

$$\alpha(\zeta) - z\beta(\zeta) := \sum_{j=0}^{k} (\alpha_j - z\beta_j)\zeta^j.$$

Dann ist $(y_n)_{n\geq 0}$ für beliebige Startwerte $y_0, \ldots, y_{k-1}$ genau dann beschränkt, wenn

1. Alle Nullstelle von $\alpha(\zeta) - z\beta(\zeta)$ sind betragsmäßig durch 1 beschränkt und
2. Jede Nullstelle, die normiert ist, ist eine einfache Nullstelle.

Der erste Punkte ist dazu äquivalent, daß $\alpha(\zeta) - z\beta(\zeta) \neq 0$ ist für $|\zeta| > 1$ und dies ist dazu äquivalent, daß

1. $z \neq \frac{\alpha(\zeta)}{\beta(\zeta)}$ für alle $|\zeta| > 1$ und $\beta(\zeta) \neq 0$ und
2. $\alpha(\zeta) \neq 0$ für alle Nullstellen ζ von β mit $|\zeta| > 1$.

Da aber $\mathrm{ggT}(\alpha, \beta) = 1$ ist, ist der zweite Punkt irrelevant und damit gilt

$$S \subseteq \mathbb{C} \setminus \left\{ \frac{\alpha(\zeta)}{\beta(\zeta)} : \ |\zeta| > 1 \right\}.$$

Alle wichtigen Verfahren erfüllen die Eigenschaft, daß keine Nullstelle des charakterischen Polynoms normiert ist. Damit gilt für alle wichtigen Verfahren sogar

$$S = \mathbb{C} \setminus \left\{ \frac{\alpha(\zeta)}{\beta(\zeta)} : \ |\zeta| > 1 \right\}.$$

In jedem Fall aber gilt

$$S \supseteq \mathbb{C} \setminus \left\{ \frac{\alpha(\zeta)}{\beta(\zeta)} : \ |\zeta| \geq 1 \right\}.$$

1.14 Satz (Stabilitätsbereich expliziter Mehrschrittverfahren)
Der Stabilitätsbereich expliziter Mehrschrittverfahren ist beschränkt.

Beweis
Wir benutzen die Vorarbeit aus Beispiel 1.13 und betrachten

$$w(\zeta) := \frac{\beta_k + b_{k-1}\zeta + \ldots + \beta_0\zeta^k}{\alpha_k + \alpha_{k-1}\zeta + \ldots + \alpha_0\zeta^k} = \frac{\zeta^k \beta\left(\frac{1}{\zeta}\right)}{\zeta^k \alpha\left(\frac{1}{\zeta}\right)}.$$

Mit Beispiel 1.13 ergibt sich damit

$$S \subseteq \mathbb{C} \setminus \left\{ \frac{\alpha(\zeta)}{\beta(\zeta)} : \ |\zeta| > 1 \right\} = \mathbb{C} \setminus \left\{ \frac{1}{w(\zeta)} : \ |\zeta| < 1 \right\}. \tag{1.4}$$

Da wir ein explizites Verfahren betrachten, ist $\beta_k = 0$, also $w(0) = 0$ und da holomorphe Funktionen offen sind, ist also $\{w(\zeta) : \ |\zeta| < 1\}$ eine Umgebung der 0, i.e.

$$\left\{ \frac{1}{w(\zeta)} : \ |\zeta| < 1 \right\}$$

ist auf der Riemannschen Zahlenkugel eine Umgebung von ∞, also ist

$$\mathbb{C} \setminus \left\{ \frac{1}{w(\zeta)} : \ |\zeta| < 1 \right\}$$

beschränkt. Da S nach (1.4) dort liegt, ist auch S beschränkt. □

1.15 Wiederholung (implizite Adams-Verfahren)
Nach dem Hauptsatz gilt stets

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} \underbrace{f(t, y(t))}\ dt,$$

wobei wir den markierten Term $f(t, y(t))$ mittels eines Interpolationspolynoms durch die Knoten $f_{n+1}, f_n, \ldots, f_{n-k+1}$ legen und dieses dann integrieren.

Dabei ist k die Anzahl der Stützstellen und aus Numerik I wissen wir, daß $k+1$ die Ordnung des Adams-Verfahren ist.

1.16 Beispiel (Stabilitätsbereiche von implizites Adams-Verfahren)
Wir betrachten nun einige Beispiele.

1. Für $k = 0$ hatten wir das implizite Euler-Verfahren, welches $y' = \lambda y$ durch
$$y_{n+1} = y_n + h\lambda y_{n+1} \iff y_{n+1} = \frac{1}{1-z} y_n,$$
wobei $z = h\lambda$ ist, gelöst hat. Hierbei ist der Stabilitätsbereich
$$S = \left\{ z \in \mathbb{C} : \left| \frac{1}{1-z} \right| \leq 1 \right\} = \{ z \in \mathbb{C} : |z-1| \geq 1 \},$$
i.e. das Verfahren ist A-stabil.

2. Für $k = 1$ betrachten wir die Trapezregel zum Lösen der Differentialgleichung $y' = \lambda y$ und erhalten für $z = h\lambda$
$$y_{n+1} = y_n + \frac{h\lambda}{2}(y_n + y_{n+1}) \iff y_{n+1} = \frac{1 + \frac{z}{2}}{1 - \frac{z}{2}} y_n,$$
womit
$$S = \left\{ z \in \mathbb{C} : \left| 1 + \frac{z}{2} \right| \leq \left| 1 - \frac{z}{2} \right| \right\} = \mathbb{C}^- := \{ z \in \mathbb{C} : \operatorname{Re}(z) \leq 0, \},$$
womit auch dieses Verfahren A-stabil ist.

3. Für $k = 2$ betrachten wir
$$y_{n+2} = y_{n+1} + h\lambda \left(\frac{5}{12} y_{n+2} + \frac{8}{12} y_{n+1} - \frac{1}{12} y_n \right),$$
also
$$\alpha(\zeta) = \zeta^2 - \zeta, \beta(\zeta) = \frac{5}{12}\zeta^2 + \frac{8}{12}\zeta - \frac{1}{12}.$$
Dabei hat β die Nullstellen
$$\zeta_{1,2} = -\frac{4}{5} \pm \sqrt{\frac{16}{25} + \frac{1}{5}} = -\frac{4}{5} \pm \frac{\sqrt{21}}{5},$$
wobei $|\zeta_2| > 1$ ist. Damit ist – mit dem gleichen Argument wie in Satz 1.14 – der Stabilitätsbereich dieses Mehrschrittverfahrens beschränkt.

4. Auch für $k \geq 2$ ist der Stabilitätsbereich beschränkt, was ungeeignet für steife Differentialgleichungen ist.

1.3 BDF-Verfahren

Wir betrachten nun eine weitere Klasse von Mehrschrittverfahren: Die so genannten BDF-Verfahren (backward differentation formulas).

1.17 Wiederholung
Beim BDF-Verfahren bestimmen wir ein Interpolationspolynom $p(t)$ durch die Punkte

$$(t_{n+1}, y_{n+1}), (t_n, y_n), \dots, (t_{n-k+1}, y_{n-k+1})$$

und verlangen dann, daß $p'(t_{n+1}) = f(t_{n+1}, p(t_{n+1}))$ ist, i.e. die BDF-Verfahren beruhen auf numerischen Differentatiation. Dabei ist k wieder die Anzahl der Stützstellen des Verfahrens und wir wissen aus Numerik I, daß dann k auch die Ordnung des Verfahrens ist.

1.18 Beispiel (Stabilitätsbereiche von impliziten BDF-Verfahren)
Wir betrachten nun einige Beispiele von impliziten BDF-Verfahren.

1. Für $k = 1$ erhalten wir das implizite Euler-Verfahren, welches A-stabil ist.
2. Für $k = 2$ erhalten wir das Verfahren
   $$\frac{3}{2}y_{n+2} - 2y_{n+1} + \frac{1}{2}y_n = h\lambda y_{n+2}.$$
   Dieses Verfahren ist A-stabil, denn wir erhalten für $\zeta = re^{\imath\theta}$ und $r \geq 1$
   $$\begin{aligned}\operatorname{Re}\frac{\alpha(\zeta)}{\beta(\zeta)} &= \operatorname{Re}\left(\frac{3}{2} - 2\frac{1}{\zeta} + \frac{1}{2\zeta^2}\right) = \operatorname{Re}\left(\frac{3}{2} - \frac{2}{r}e^{-\imath\theta} + \frac{1}{2r^2}e^{-2\imath\theta}\right)\\ &= \frac{3}{2} - \frac{2}{r}\cos\theta + \frac{1}{2r^2}\cos(2\theta),\end{aligned}$$
   was sich wegen $\cos(2\theta) = 2\cos^2\theta - 1$ nun zu
   $$= 1 - 2\frac{\cos\theta}{r} + \left(\frac{\cos\theta}{r}\right)^2 + \frac{1}{2} - \frac{1}{2r^2} = \underbrace{\left(1 - \frac{\cos\theta}{r}\right)^2}_{\geq 0} + \underbrace{\frac{1}{2}\left(1 - \frac{1}{r^2}\right)}_{\geq 0} \geq 0$$
   ergibt, womit nach der Vorarbeit in Beispiel 1.13 dann $S \supseteq \mathbb{C}^-$ ist.
3. Für größere k ist das BDF-Verfahren haben wir in Numerik I gesehen, daß diese (für $k > 6$) nicht einmal 0-stabil sind. Dennoch haben diese Verfahren einen unbeschränkten Stabilitätsbereich, in dem zumindest ein Kegel der Art
   $$\{z \in \mathbb{C} : \ |\arg(-z)| \leq \alpha\} \subseteq S$$
   mit Öffnungswinkel α liegt, was wir mit $A(\alpha)$-stabil bezeichnen. Falls Eigenwerte nahe der imaginären Achse auftreten, haben diese Verfahren jedoch schlechte Ergebnisse.

Für bestimmte k ergeben sich die folgenden Öffnungswinkel.

k	1	2	3	4	5	6
α	90°	90°	88°	73°	51°	18°

BDF-Verfahren gehören zu den beliebtesten Verfahren für steife Differentialgleichungen.

1.4 Ordnungsschranke für A-stabile Mehrschrittverfahren

Alle bisher genannten A-stabilen Mehrschrittverfahen haben Ordnung ≤ 2. Wir sehen mit dem folgenden Satz, daß dies bereits eine optimale Schranke ist.

1.19 Satz (Dahlquist, 1963)
Ein A-stabiles Mehrschrittverfahren hat Ordnung $p \leq 2$.

Beweis
Wenn das Mehrschrittverfahren A-stabil ist, so ist mit der Vorarbeit aus Beispiel 1.13

$$\operatorname{Re}\left(\frac{\alpha(\zeta)}{\beta(\zeta)}\right) \geq 0 \quad \forall |\zeta| \geq 1 \tag{1.5}$$

In Numerik I haben wir mit der Taylor-Entwicklung der Exponentialfunktion bewiesen, daß ein Mehrschrittverfahren genau dann Ordnung p hat, wenn $\alpha(e^h) - h\beta(e^h) = \mathcal{O}(h^{p+1})$ ist, i.e. für $z \to 0$ gibt es ein $z \neq 0$ mit

$$\frac{1}{z}\frac{\alpha(e^z)}{\beta(e^z)} - 1 = cz^p + \mathcal{O}(z^{p+1}). \tag{1.6}$$

Wir betrachten zunächst den Imaginärteil von $\frac{1}{z}\frac{\alpha(e^z)}{\beta(e^z)}$ auf dem Rand des Rechtecks $[0, x_k] \times [0, 2\pi\imath]$.

1. Für $z \in \mathbb{R}^+$ oder $z \in 2\pi\imath + \mathbb{R}^+$ ist $e^z \in \mathbb{R}$ und damit ist $\frac{\alpha(e^z)}{\beta(e^z)} \in \mathbb{R}$, womit wegen (1.5) $\frac{\alpha(e^z)}{\beta(e^z)} \geq 0$ ist, i.e. es gilt

   $$\operatorname{Im}\left(\frac{1}{z}\frac{\alpha(e^z)}{\beta(e^z)}\right) = \underbrace{\operatorname{Im}\frac{1}{z}}_{\leq 0}\underbrace{\frac{\alpha(e^z)}{\beta(e^z)}}_{\geq 0} \leq 0.$$

2. Für $z = x + \imath y$ mit $0 \leq y \leq 2\pi$ und $x \to \infty$ ergibt sich, da nach Satz 1.14 $\beta_k \neq 0$ ist,

   $$\frac{\alpha(e^z)}{\beta(e^z)} = \frac{\alpha_k e^{kz}(1 + \mathcal{O}(e^{-z}))}{\beta_k e^{kz}(1 + \mathcal{O}(e^{-z}))} \to \frac{\alpha_k}{\beta_k}.$$

   Sei $\varepsilon > 0$ beliebig vorgegeben. Dann ist damit für großes x

   $$\operatorname{Im}\left(\frac{1}{z}\frac{\alpha(e^z)}{\beta(e^z)}\right) \leq \varepsilon.$$

3. Für $z = \imath y$ und $0 < y \leq 2\pi$ ergibt sich mit (1.5)

   $$\operatorname{Im}\left(\frac{1}{\imath y}\frac{\alpha(e^z)}{\beta(e^z)}\right) = -\operatorname{Re}\left(\frac{1}{y}\frac{\alpha(e^{\imath y})}{\beta(e^{\imath y})}\right) \leq 0.$$

Mit dem Maximumprinzip für harmonische Funktionen erhalten wir mit allen drei Teilergebnissen auch

$$\operatorname{Im}\left(\frac{1}{z}\frac{\alpha(e^z)}{\beta(e^z)}\right) \leq \varepsilon$$

im Inneren des Rechtecks $[0, x_k] \times [0, 2\pi\imath]$; und da $\varepsilon > 0$ beliebig war, gilt damit

$$\operatorname{Im}\left(\frac{1}{z}\frac{\alpha(e^z)}{\beta(e^z)}\right) \leq 0, \text{ also auch } \operatorname{Im}\left(\frac{1}{z}\frac{\alpha(e^z)}{\beta(e^z)} - 1\right) \leq 0.$$

Mit (1.6) ist damit auch $c \operatorname{Re} z^p + \mathcal{O}(z^{p+1}) \leq 0$ für $z \to 0$ und $0 \leq \arg z \leq \frac{\pi}{2}$, also muß $c < 0$ sein, denn $c \operatorname{Im}(z^p) \leq 0$ ist für $p \geq 3$ nicht möglich. □

1.5 Kollokationsverfahren

Wir möchten im Folgenden klären, ob es A-stabile (implizite) Runge–Kutta-Verfahen der Ordnung $p \geq 3$ gibt.

1.20 Algorithmus (Kollokationsverfahren)
Wir betrachten die Differentialgleichung

$$\begin{cases} y'(t) = f(t, y(t)), \\ y(t_0) = y_0. \end{cases}$$

Seien $c_1, \ldots, c_s \in [0, 1]$ alle verschieden gegeben und h die Schrittweite.

Wir suchen nun ein Polynom $u(t)$ vom Grad $\leq s$, für das die folgenden *Kollokationsbedingungen* gelten.

$$\begin{cases} u'(t) = f(t, u(t)), \quad t = t_0 + c_i h, \\ u(t_0) = y_0. \end{cases}$$

Wir setzen dann

$$y_1 := u(t_0 + h) \approx y(t_0 + h)$$

und nehmen y_1 als Startwert für den nächsten Schritt.

1.21 Satz
Das Kollokationsverfahren ist äquivalent zum impliziten Runge–Kutta-Verfahren mit den Koeffizienten

$$a_{ij} = \int_0^{c_i} \ell_j(x)\,dx, \qquad b_j = \int_0^1 \ell_j(x)\,dx,$$

wobei ℓ_j das Lagrange-Polynom $\deg \ell_j \le s-1$ zum Knoten c_j mit der Eigenschaft $\ell_j(c_i) = \delta_{ij}$.

Beweis
Wir setzen in der Notation eines allgemeinen Runge–Kutta-Verfahrens $Y_i' := u'(t_0 + c_i h)$ und $Y_i := u(t_0 + c_i h)$. Dann erhalten wir aus der Kollokatonsbedingung, daß $Y_i' = f(t_0 + c_i h, Y_i)$ ist. Da u' ein Polynom vom Grad $\le s-1$ ist, können wir mit der Lagrange-Inerpolation

$$u'(t_0 + xh) = \sum_{j=1}^{s} Y_j' \ell_j(x)$$

schreiben und erhalten damit

$$Y_i = u(t_+ c_i h) = u(t_0) + h \int_0^{c_i} u'(t_0 + xh)\,dx = y_0 + h \sum_{j=1}^{s} Y_j' \underbrace{\int_0^{c_i} \ell_j(x)\,dx}_{=a_{ij}}.$$

Mit den gleichen Argumenten erhalten wir

$$y_1 = u(t_0 + h) = \ldots = u(t_0) + h \sum_{j=1}^{s} Y_j' \underbrace{\int_0^1 \ell_j(x)\,dx}_{=b_j}.$$

□

1.22 Bemerkung
Damit erhalten wir, daß für genügend kleine h das Kollokektionsverfahren eindeutig existiet.

1.23 Bemerkung
Die Koeffizienten a_{ij} und b_j erfüllen die Bedingungen

$$\sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{c_j^k}{k}, \qquad k = 1, \ldots, s,$$

$$\sum_{j=1}^{s} b_j c_j^{k-1} = \frac{1}{k}, \qquad k = 1, \ldots, s.$$

Damit sind a_{ij} und b_j aus einem linearen Gleichungssystem mit einer Vandermonde-Matrix zu berechnen.

Aus Numerik von Differentialgleichungen I erinnern wir uns an den folgenden Satz.

1.24 Satz
Das Kollokationsverfahren hat dieselbe Ordnung wie die zugehörige Quadraturformel.

1.25 Bemerkung
Satz 1.24 besagt also, daß das impliziten Runge–Kutta-Verfahren aus Satz 1.21 hat genau dann Ordnung p (i.e. $y_1 - y(t_0 + h) = \mathcal{O}(h^{p+1})$), falls die Quadraturformel

$$\sum_{i=1}^{s} b_j g(c_i) \approx \int_0^1 g(x)\,dx$$

Ordnung p hat (also exakt für Polynome vom Grad $\leq p-1$ ist).

1.26 Herleitung (Charakterisierung der A-Stabilität)
Wir betrachten die Differentialgleichung $y' = \lambda y$ mit $\operatorname{Re}\lambda \leq 0$ und möchten im Folgenden ihren Stabilitätsbereich untersuchen. Hierzu wenden wir das implizite Runge–Kutta-Verfahren auf die Testgleichung an und erhalten für $z = h\lambda$

$$\begin{aligned} y_1 &= y_0 + z\sum_{j=1}^{s} b_j Y_j, \\ Y_i &= y_0 + h\lambda\sum_{j=1}^{s} a_{ij} Y_j. \end{aligned}$$

Für $\mathfrak{A} := (a_{ij})$ erhalten wir nun die Gleichung

$$\underbrace{\begin{pmatrix} Y_1 \\ \dots \\ Y_s \end{pmatrix}}_{=:Y} = y_0 \underbrace{\begin{pmatrix} 1 \\ \dots \\ 1 \end{pmatrix}}_{=\mathbb{1}} + z\mathfrak{A}\begin{pmatrix} Y_1 \\ \dots \\ Y_s \end{pmatrix},$$

also $(I - z\mathfrak{A})Y = y_0\mathbb{1}$. Falls $I - z\mathfrak{A}$ invertierbar ist, dann gilt für $b^T := (b_1, \dots, b_s)$

$$y_1 = y_0 + zb^T Y = y_0 + zb^T(I - z\mathfrak{A})^{-1} y_0 \mathbb{1} = \underbrace{(1 + zb^T(I - z\mathfrak{A})^{-1}\mathbb{1})}_{=:R(z)} y_0,$$

wobei $R(z) = \frac{P(z)}{Q(z)}$ eine rationale Funktion, die so genannte *Stabilitätsfunktion* in z ist mit $\deg P, \deg Q \leq s$. Wir erhalten dann induktiv $y_n = (R(z))^n y_0$, also ist der Stabilitätsbereich

$$S = \{z \in \mathbb{C} : \ |R(z)| \leq 1\},$$

i.e. das implizite Runge–Kutta-Verfahren ist genau dann A-stabil, wenn $|R(z)| \leq 1$ für alle z mit $\mathrm{Re}(z) \leq 0$ ist.

1.27 Beispiel (Kollokationsverfahren mit Gaußscher Quadraturformel)
Bei der Gaußschen Quadraturformel haben wir die Ordnung $p = 2s$ und betrachten nun für verschiedene s die Konstruktion.

$s = 1$: In diesem Fall gilt $c_1 = \frac{1}{2}$, $b_1 = 1$ und $a_{11} = \frac{1}{2}$. Damit erhalten wir die implizite Mittelpunktsregel,

$$y_1 = y_0 + hf\left(t_0 + \frac{h}{2}, \frac{y_0 + y_1}{2}\right).$$

Wir betrachten nun die Gleichung $y' = \lambda y$ und für $z = h\lambda$ erhalten wir

$$y_1 = y_0 + z\frac{y_0 + y_1}{2} \iff y_1 = \underbrace{\frac{1 + \frac{z}{2}}{1 - \frac{z}{2}}}_{=:R(z)} y_0.$$

Per Inspektion sehen wir – ähnlich wie in Beispiel 1.16, daß $|R(z)| \leq 1$ für $\mathrm{Re}(z) \leq 0$ ist, i.e. dieses Verfahren ist A-stabil.

$s = 2$: In diesem Fall gilt $c_{1,2} = \frac{1}{2} \mp \frac{\sqrt{3}}{6}$ und $b_1 = b_2 = \frac{1}{2}$. Aus den Kollokationsbedingungen erhalten wir (Übung!), daß wir das folgende Runge–Kutta-Tableau bearbeiten müssen.

$$\begin{array}{c|cc} \frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

Eine einfache Rechnung ergibt

$$y_1 = \underbrace{\frac{1 + \frac{z}{2} + \frac{z^2}{12}}{1 - \frac{z}{2} + \frac{z^2}{12}}}_{=:R(z)} y_0.$$

Wir prüfen nun nach, ob $|R(z)| \leq 1$ für $\mathrm{Re}(z) \leq 0$ gilt. Es gilt

$$|R(\imath y)|^2 = \frac{\left(1 - \frac{y^2}{12}\right) + \left(\frac{y}{2}\right)^2}{\left(1 - \frac{y^2}{12}\right) + \left(-\frac{y}{2}\right)^2} = 1 \quad \forall y \in \mathbb{R}. \tag{1.7}$$

Die Pole, die durch $1 - \frac{z}{2} + \frac{z^2}{12} = 0$ gegeben sind, befinden sich bei $z_{1,2} = 3 \pm \imath\sqrt{3}$, liegen also beide in der rechten Halbebene. Mit dem Maximumsprinzip für holomorphe Funktionen und (1.7) gilt damit $|R(z)| \leq 1$ für $\mathrm{Re}(z) \leq 0$. Also ist auch dieses Verfahren A-stabil.

Wir werden später zeigen, daß sämtliche Gauß-Kollokationsverfahren Ordnungen A-stabil sind. Der Nachteil ist allerdings, daß $|R(\imath y)| = 1$ ist, also auch $\lim_{z\to-\infty} |R(z)| = 1$, womit auch Dämpfungseigenschaften verloren gehen, da die Exponentialfunktion als echte Lösung im unendlichen verschwindet.

Wir möchten nun ein A-stabiles Verfahren erhalten, welches $|R(-\infty)| < 1$, am besten jedoch $R(-\infty) = 0$. Hierzu betrachten wir als nächstes Beispiel Kollokationsverfahren, die auf Radau-Quadraturformeln beruhen.

1.28 Beispiel (Kollokationsverfahren mit Radau-Quadraturformeln)
Bei einer Radau-Quadraturformel ist $c_s = 1$ und $c_1, \dots, c_{s-1}$ so gegeben, daß die Ordnung der Quadraturformel $p = 2s - 1$ wird. Wir betrachten nun die Quadraturformel für einige s.

$s = 1$: In diesem Falle ist $c_1 = 1$, $b_1 = 1$ und $a_{11} = 1$. Dann ergibt sich das implizite-Euler-Verfahren,

$$y_1 = y_0 + hf(t_0 + h, y_1),$$

als Verfahren, welches A-stabil ist.

$s = 2$: Hier ergibt sich $c_1 = \frac{1}{3}$ und $c_2 = 1$ und das folgende Runge–Kutta-Tableau.

$$\begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array}$$

Dieses Verfahren hat Ordnung 3 und es gilt

$$R(z) = \frac{1 + \frac{z}{3}}{1 - \frac{2z}{3} + \frac{z^2}{6}},$$

also $R(-\infty) = 0$ und auf der imaginären Achse gilt

$$|R(\imath y)|^2 = \frac{1 + \frac{y^2}{9}}{\left(1 - \frac{y^2}{6}\right)^2 + \frac{4}{9}y^2} = \frac{1 + \frac{y^2}{9}}{1 + \frac{y^2}{9} + \frac{y^4}{36}} \leq 1 \quad \forall y \in \mathbb{R}. \tag{1.8}$$

Die Pole der Funktion liegen bei $z_{1,2} = 2 \pm \imath\sqrt{2}$, also in der rechten Halbebene und aus dem Maximumsprinzip erhalten wir mit (1.8), daß $|R(z)| \leq 1$ für $\mathrm{Re}(z) \leq 0$, also ist das Verfahren A-stabil.

1.29 Bemerkung (Dämpfungseigenschaft der Radau-Kollokationsverfahren)
Wegen $c_s = 1$ gilt (aufgrund der Integration von Lagrange-Polynomen), daß $a_{sj} = b_j$ ist. Falls nun die Matrix $\mathfrak{A} := (a_{ij})$ invertierbar ist, gilt

$$\begin{aligned} \lim_{z\to\infty} R(z) &= \lim_{z\to\infty} (1 + b^T z(I - z\mathfrak{A})^{-1}\mathbb{1}) \\ &= \lim_{z\to\infty} \left(1 + b^T \left(\frac{1}{z}I - \mathfrak{A}\right)^{-1} \mathbb{1}\right) \\ &= 1 - b^T\mathfrak{A}^{-1}\mathbb{1} = 1 - e_s^T\mathbb{1} = 0, \end{aligned}$$

wobei wir $a_{sj} = b_j$ benutzt haben. Damit gilt stets die Dämpfungseigenschaft, die wir gefordert haben.

Die Radau-Kollokationsverfahren werden die in der Praxis häufig benutzt, da sie A-stabil sind und die gewünschte Dämpfungseigenschaft haben.

Wir betrachten nun, wie wir die auftretenden (nichtlinearen) Gleichungssysteme beim impliziten Runge–Kutta-Verfahren lösen können.

1.30 Bemerkung (Lösen der auftretenden Gleichungssysteme)
Wir haben die Differentialgleichung $y'(t) = f(t, y(t))$ mit dem Anfangswert $y(t_0) = y_0 \in \mathbb{R}^d$ gegeben und betrachten nun das implizite Runge–Kutta-Verfahren,

$$Y_i = y_0 + h\sum_{j=1}^{s} a_{ij} f(t_0 + c_j h, Y_j), \qquad i = 1, \dots, s,$$
$$y_1 = y_0 + h\sum_{i=1}^{s} b_i f(t_0 + c_i h, Y_i),$$

welches sich als nichtlinearen Gleichungssysteme der Dimension $s \cdot d$ herausstellt. Zum Lösen betrachten wir das vereinfachte Newton-Verfahren und bezeichnen mit $J \approx \partial_y f(t_0, y_0)$ die Jacobi-Matrix im Anfangswert und berechnen für $k = 0, 1, \dots$ die Iterierten

$$Y_i^{(k+1)} = Y_i^{(k)} + \Delta Y_i^{(k)},$$

wobei

$$\Delta Y_i^{(k)} - h\sum_{j=1}^{s} a_{ij} J \Delta Y_j^{(k)} = \underbrace{-Y_i^{(k)} + y_0 + h\sum_{j=1}^{s} a_{ij} f(t_0 + c_j h, Y_j^{(k)})}_{=: r_i^{(k)}}.$$

In Matrixform schreiben wir zunächst

$$\Delta Y := \begin{pmatrix} \Delta Y_1 \\ \dots \\ \Delta Y_s \end{pmatrix}, \qquad r := \begin{pmatrix} r_1 \\ \dots \\ r_s \end{pmatrix}, \qquad \mathfrak{A} \otimes J := \begin{pmatrix} a_{11}J & \dots & a_{1s}J \\ \vdots & & \vdots \\ a_{s1}J & \dots & a_{ss}J \end{pmatrix},$$

wobei $\otimes$ das wie oben definierte $s \cdot d$-dimensionale Kronecker-Produkt ist. Aus diesem Definitionen ergibt sich das lineare Gleichungssystem

$$(I_{sd} - h\mathfrak{A} \otimes J)\Delta Y^{(k)} = r^{(k)}$$

bzw.

$$\left(\frac{1}{h} I_s \otimes I_d - \mathfrak{A} \otimes J\right) = \frac{1}{h} r^{(k)}. \tag{1.9}$$

Eine direkte Lösung hiervon benötigt $\frac{1}{3}(sd)^3$ Operationen, wobei d^3 nicht vermeidbar ist, uns jedoch s^3 stört.

Falls $\mathfrak{A}$ diagonalisierbar ist (was nicht von der Differentialgleichung, sondern ausschließlich vom gewählten Verfahren abhängt!), haben wir deutlich weniger Rechenaufwand, was die folgende Rechnung zeigt. Beispielsweise tritt dies beim Radau-Verfahren mit $s = 3$ auf.

Wir nehmen nun an, daß $T^{-1}\mathfrak{A}T = \Lambda := \operatorname{diag}(\lambda_1, \ldots, \lambda_s)$ ist. Dann ergibt sich die linke Seite von (1.9) zu

$$\frac{1}{h} I_s \otimes I_d - \mathfrak{A} \otimes J = (T \otimes I_d) \left(\frac{1}{h} I_s \otimes I_d - \Lambda \otimes J \right) (T^{-1} \otimes I_d).$$

Wir berechnen nun $\Delta Y^{(k)}$ wie folgt. Wir berechnen

$$q^{(k)} := (T^{-1} \otimes I_d) r^{(k)} = \left(\sum_{j=1}^{s} a_{ij} r_j^{(k)} \right)_{i=1}^{s},$$

was ds^2 Operationen kostet und berechnen anschließend $Z_i^{(k)}$ durch

$$\left(\frac{1}{h} I_d - \lambda_i J \right) \Delta Z_i^{(k)} = \frac{1}{h} q_i^{(k)}$$

bzw.

$$\left(\frac{1}{h\lambda_i} I_d - J \right) \Delta Z_i^{(k)} = \frac{1}{h\lambda_i} q_i^{(k)}, \tag{1.10}$$

was durch 5 LR-Zerlegungen in $s \cdot \frac{1}{3} d^3$ Operationen berechnet werden kann.

Dieser Schritt läßt sich noch besser machen, wenn wir J mit orthogonalen Matrizen Q in Hessenberg-Form H via $J = QHQ^T$ schreiben. Hierfür benötigen wir d^3 Operationen und dann ergibt sich die linke Seite von (1.10) zu

$$\frac{1}{h\lambda_i} I_d - J = Q^T \underbrace{\left(\frac{1}{h\lambda_i} I_d - H \right)} Q,$$

wobei wir dann zum Lösen nur vom hervorgehobenen Ausdruck eine LR-Zerlegung benötigen, die aufgrund der speziellen Struktur der Hessenberg-Matrizen nur d^2 Operationen benötigt.

In beiden Fällen muß abschließend noch

$$\Delta Y^{(k)} = (T \otimes I_{ds}) \Delta Z^{(k)}$$

berechnet werden, was mit sd^2 Operationen geschieht. Wir können also, wenn die Matrix $\mathfrak{A}$ des Verfahrens gut ist, Rechenaufwand sparen. Der Gesamtaufwand besteht dann aus $d^3 + \mathcal{O}(sd^2)$ für einen Zeitschritt.

1.6 Kontraktive Runge–Kutta-Verfahren

Um die A-Stabilität *aller* Gauß- und Radau-Kollokationsverfahren zeigen zu können, führen wir einen stärkeren Begriff, die Kontraktivität, als die A-Stabilität ein, der sich aber leichter allgemein beweisen läßt und werden später zeigen, daß sowohl die Gauß- als auch die Radau-Kollokationsverfahren kontraktiv, also auch A-stabil sind.

1.31 Definition (kontraktive Differentialgleichung)
Eine Differentialgleichung $y'(t) = f(t, y(t))$ heißt *kontraktiv*, wenn für zwei Lösngen y und $\tilde{y}$ (zu unterschiedlichen Anfangswerten y_0 und $\tilde{y}_0$) gilt

$$\|y(t_1) - \tilde{y}(t_1)\| \leq \|y(t_0) - \tilde{y}(t_0)\| \quad \forall t_1 \geq t_0.$$

Wir möchten nun auch die Kontraktivität für das numerische Verfahren bei beliebigen Schrittweiten $h > 0$ haben. Wir leiten im Folgenden hierfür Bedingungen her.

1.32 Herleitung (Charakterisierung der Kontraktivität)
Wir haben die Differentialgleichung $y'(t) = f(t, y(t))$ für ein $f : \mathbb{R} \times \mathbb{C}^d \to \mathbb{C}^d$ gegeben. Wir bezeichnen mit $\langle ., . \rangle$ ein Skalarprodukt auf $\mathbb{C}^d$ und mit $\|.\|$ die dazugehörige Norm. Für zwei Lösungen y und $\tilde{y}$ der Differentialgleichung gilt dann mit der Produktregel und der Tatsache, daß diese die Differentialgleichung lösen, daß

$$\begin{aligned}
\frac{d}{dt}\|y(t) - \tilde{y}(t)\|^2 &= \frac{d}{dt}\langle y(t) - \tilde{y}(t), y(t) - \tilde{y}(t)\rangle \\
&= \langle y'(t) - \tilde{y}'(t), y(t) - \tilde{y}(t)\rangle + \langle y(t) - \tilde{y}(t), y'(t) - \tilde{y}'(t)\rangle \\
&= 2\,\mathrm{Re}\langle y'(t) - \tilde{y}'(t), y(t) - \tilde{y}(t)\rangle \\
&= 2\,\mathrm{Re}\langle f(t, y(t)) - f(t, \tilde{y}(t)), y(t) - \tilde{y}(t)\rangle.
\end{aligned}$$

Damit die Differentialgleichung kontraktiv ist, muß dieser Ausdruck nichtpositiv sein. Wir fordern im Folgenden also für eine kontraktive Differentialgleichung, daß

$$\mathrm{Re}\langle f(t, y(t)) - f(t, \tilde{y}(t)), y(t) - \tilde{y}(t)\rangle \leq 0 \quad \forall t \in \mathbb{R},\ \forall y, \tilde{y} \in \mathbb{C}^d. \tag{1.11}$$

1.33 Definition (Kontraktives Runge–Kutta-Verfahren)
Ein Runge–Kutta-Verfahren heißt *kontraktiv*, wenn für jede Differentialgleichung, die (1.11) erfüllt, gilt, für die numerischen Lösungen $y_1, \tilde{y}_1$ zu den Startwerten $y_0, \tilde{y}_0$ für beliebige Schrittweite $h > 0$ gilt, daß $\|y_1 - \tilde{y}_1\| \leq \|y_0 - \tilde{y}_0\|$.

1.34 Proposition
Ist ein Verfahren kontraktiv, dann ist es auch A-stabil.

Beweis
Die Differentialgleichung $y'(t) = \lambda y =: f(t, y)$ mit $\mathrm{Re}\,\lambda \leq 0$ erfüllt die Bedingung (1.11). □

1.35 Beispiel
Das implizite Euler-Verfahren ist kontraktiv, denn mit Cauchy–Schwarz und (1.11) gilt

$$\|y_1 - z_1\|^2 = \mathrm{Re}\langle y_1 - z_1, y_0 - z_0 + hf(t_1, y_1) - hf(t_1, z_1)\rangle \leq \|y_1 - z_1\|\|y_0 - z_0\|.$$

1.36 Satz (Bedingung an kontraktives Runge–Kutta-Verfaren)
Ein implizites Runge–Kutta-Verfahren mit den Knoten c_i, den Gewichten b_j und der Matrix a_{ij} sei *algebraisch stabil*, i.e.

1. $b_j > 0$ für alle $j = 1, \dots, s$ und
2. Die Matrix $M := (m_{ij})$ mit $m_{ij} := b_i a_{ij} + b_j a_{ji} - b_i b_j$ ist positiv semidefinit.

Dann ist das Runge–Kutta-Verfahren kontraktiv.

Beweis
Wir schreiben zunächst das Runge–Kutta-Verfahren auf. Es gilt

$$Y_i = y_0 + h \sum_{j=1}^{s} a_{ij} Y_j',$$
$$y_1 = y_0 + h \sum_{i=1}^{s} b_i Y_i',$$

wobei $Y_j' = f(t_0 + c_j h, Y_j)$ ist. Für $\tilde{y}$ benutzen wir eine entsprechende Notation. Wir schreiben

$$y_1 - \tilde{y}_1 = (y_0 - \tilde{y}_0) + h \sum_{i=1}^{s} b_i (Y_i' - \tilde{Y}_i'). \tag{1.12}$$

Für $u = v + w$ gilt stets $\|u\|^2 = \|v\|^2 + \|w\|^2 + 2\,\mathrm{Re}\langle v, w\rangle$, also erhalten wir aus (1.12) und $y_0 - \tilde{y}_0 = Y_i - \tilde{Y}_i - h \sum_{j=1}^{s} a_{ij}(Y_j' - \tilde{Y}_j')$

$$\begin{aligned}
\|y_1 - \tilde{y}_1\|^2 &= \|y_0 - \tilde{y}_0\| + 2h \sum_{i=1}^{s} b_i \,\mathrm{Re}\langle Y_i' - \tilde{Y}_i', y_0 - \tilde{y}_0\rangle \\
&\quad + h^2 \sum_{i=1}^{s}\sum_{j=1}^{s} b_i b_j \langle Y_i' - \tilde{Y}_i', Y_j' - \tilde{Y}_j'\rangle \\
&= \|y_0 - \tilde{y}_0\|^2 + 2h \sum_{i=1}^{s} \underbrace{b_i \,\mathrm{Re}\langle f(t_0 + c_i h, Y_i) - f(t_0 + c_i h, \tilde{Y}_i), Y_i - \tilde{Y}_i\rangle}_{=:T_1} \\
&\quad \underbrace{-h^2 \sum_{i=1}^{s}\sum_{j=1}^{s} \underbrace{(b_i a_{ij} - b_j a_{ji} - b_i b_j)}_{=m_{ij}} \langle Y_i' - \tilde{Y}_i, Y_j' - \tilde{Y}_j'\rangle}_{=:T_2}.
\end{aligned}$$

Wegen $b_i > 0$ und (1.11) ist $T_1 \le 0$. Für den Term T_2 betrachten wir $v_i := Y_i' - \tilde{Y}_i$ und rechnen

$$\begin{aligned}
\sum_{i,j=1}^{s} m_{ij} \langle v_i, v_j\rangle &= \sum_{i,j=1}^{s} m_{ij} (\overline{v_{i1}} v_{j1} + \ldots + \overline{v_{is}} v_{js}) \\
&= \underbrace{\sum_{i,j=1}^{s} \overline{v_{i1}} m_{ij} v_{j1}}_{\ge 0} + \ldots + \underbrace{\sum_{i,j=1}^{s} \overline{v_{is}} m_{ij} v_{js}}_{\ge 0} \ge 0,
\end{aligned}$$

wobei wir benutzt haben, daß M positiv semidefinit ist, also ist auch $T_2 \leq 0$ und insgesamt gilt damit $\|y_1 - \tilde{y}_1\|^2 \leq \|y_0 - \tilde{y}_0\|$. □

1.37 Bemerkung
Falls das Runge–Kutta-Vefahren kontraktiv ist und die c_i alle verschieden sind, so ist es algebraisch stabil. (Übung!).

1.38 Beispiel
Bei Gauß-Kollokationsverfahren gilt stets $M = 0$ und beim Radau-Kollokationsverfahren ist M stets positiv definit.

1.7 Kontraktivität von Gauß- und Radau-Kollokationsverfahren

Wir zeigen in diesem Abschnitt, daß die Gauß- und Radau-Kollokationsverfahren aus Abschnitt 1.5 kontraktiv sind.

1.39 Hilfssatz (Gewichte der Gauß- und Radau-Quadraturformeln)
Die Gewichte b_i der Gauß- und der Radau-Quadraturformeln sind strikt positiv.

Beweis
In beiden Fällen ist die Ordnung der Quadraturformel $p \geq 2s - 1$, i.e. die Quadraturformel ist exakt für Polynome vom Grad $\leq 2s - 2$. Sei nun ℓ_i das Lagrange-Polynom durch die Knoten c_i vom Grad $s - 1$ mit der Eigenschaft $\ell_i(c_j) = \delta_{ij}$. Dann gilt

$$0 < \int_0^1 (\ell_i(x))^2 \, dx = \sum_{j=1}^{s} b_j \ell_i(c_j)^2 = b_i.$$

□

1.40 Satz (Kontraktivität der Gauß-Kollokationsverfahren)
Die Gauß-Kollokationsverfahren sind kontraktiv.

Beweis
Seien u bzw. $\tilde{u}$ Kollokationspolynome (vom Grad $\leq s$) zu den Startwerten y_0 und $\tilde{y}_0$. Dann gilt $y_1 - \tilde{y}_1 = u(t_0 + h) - \tilde{u}(t_0 + h)$.

Wir betrachten das Polynom $m(t) := \|u(t) - \tilde{u}(t)\|^2$ vom Grad $\leq 2s$ und sehen, daß mit der Definition des Kollokationsverfahrens und (1.11) für die Kollokationspunkte $\tau_i := t_0 + c_i h$

$$\begin{aligned} m'(\tau_i) &= 2\,\mathrm{Re}\langle u'(\tau_i) - \tilde{u}'(\tau_i), u(\tau_i) - \tilde{u}(\tau_i)\rangle \\ &= 2\,\mathrm{Re}\langle f(\tau_i, u(\tau_i)) - f(\tau_i, \tilde{u}(\tau_i)), u(\tau_i) - \tilde{u}(\tau_i)\rangle \leq 0 \end{aligned}$$

gilt. Mit dem Hauptsatz und diesem Resultat erhalten wir schließlich wegen $\deg m'(\tau) \leq 2s-1$

$$\|y_1 - \tilde{y}_1\|^2 = \|u(t_0+h) - \tilde{u}(t_0+h)\|^2 = m(t_0+h) = m(t_0) + \int_{t_0}^{t_0+h} m'(t)\,dt$$
$$= m(t_0) + h\sum_{i=1}^{s} \underbrace{b_i}_{>0}\underbrace{m'(\tau_i)}_{\leq 0} \leq m(t_0) = \|u(t_0) - \tilde{u}(t_0)\|^2 = \|y_0 - \tilde{y}_0\|^2,$$

wobei wir benutzt haben, daß wegen Hilfssatz 1.39 alle b_i positiv sind. □

1.41 Satz (Kontraktivität der Radau-Kollokationsverfahren)
Die Radau-Kollokationsverfahren sind kontraktiv.

Beweis
Wir betrachten wie im Beweis von Satz 1.40 das Polynom $m(t) := \|u(t) - \tilde{u}(t)\|^2$ mit $\deg m \leq 2s$ und erhalten wie dort $m'(\tau_i) \leq 0$. Wir schreiben nun $m(t) = c(t-t_0)^{2s} + r(t)$ mit $\deg r \leq 2s-1$ und $c \geq 0$ (da m positiv ist). Damit ergibt sich

$$m'(t) = 2sc(t-t_0)^{2s-1} + r'(t) = 2sc(t-\tau_1)^2 \cdot \ldots \cdot (t-\tau_{s-1})^2(t-\tau_s) + q(t),$$

wobei $\tau_i = t_0 + c_i h$ und $\tau_s = t_0 + h$ (wegen $c_s = 1$) ist und $\deg q \leq 2s-2$. Wir rechnen nun mit dem Hauptsatz und der Tatsache, daß die Radau-Quadraturformel Ordnung $2s-1$ hat

$$\|y_1 - \tilde{y}_1\|^2 = \|u(t_0+h) - \tilde{u}(t_0+h)\|^2 = m(t_0+h) = m(t_0) + \int_{t_0}^{t_0+h} m'(t)\,dt$$
$$= m(t_0) + \underbrace{\int_{t_0}^{t_0+h} \underbrace{2sc(t-\tau_1)^2 \cdot \ldots \cdot (t-\tau_{s-1})^2}_{\geq 0}\underbrace{(t-t_0-h)}_{\leq 0}\,dt}_{\leq 0} + \underbrace{\int_{t_0}^{t_0+h} q(t)\,dt}_{=h\sum_{i=1}^{s} b_i q(\tau_i)}$$
$$\leq m(t_0) + \underbrace{h\sum_{i=1}^{s} b_i \underbrace{q(\tau_i)}_{=m'(\tau_i)\leq 0}}_{\leq 0} \leq m(t_0) = \|y_0 - \tilde{y}_0\|^2,$$

wobei wir benutzt haben, daß wegen Hilfssatz 1.39 alle b_i positiv sind. □

1.42 Korollar (A-Stabilität)
Die Gauß- und die Radau-Kollokationsvefahren sind A-stabil.

2 Parabolische Differentialgleichungen

2.1 Beispiel (Wärmeleitungsgleichung)
Das typische Beispiel für eine parabolische Differentialgleichung ist die Wärmeleitungsgleichung: Wir suchen $u = u(x,t)$, wobei $x \in \Omega \subseteq \mathbb{R}^d$ der Ort und t die Zeit ist und möchten nun die folgende Differentialgleichung lösen.

$$\begin{cases} \partial_t u(x,t) = \Delta u(x,t) + f(x,t), & x \in \Omega,\ 0 < t \leq T, \\ u(x,t) = g(x,t) & x \in \partial\Omega,\ 0 < t \leq T, \\ u(x,0) = u_0(x), & x \in \Omega, \end{cases}$$

wobei f, g und u_0 gegebene Funktionen. Die zweite Zeile beschreibt die Randbedingungen und die dritte Zeile die Anfangsbedingungen.

2.1 Einführendes Beispiel, Linienmethode

2.2 Herleitung (der Semidiskretisierung mit der Linienmethode)
Wir betrachten im Folgenden $\Omega := (0,1)^d$ und $f := 0$ sowie $g := 0$. Bei der *Linienmethode* führen wir eine Semidiskretisierung im Raum durch (i.e. die Zeit bleibt kontinuierlich). Hierfür betrachten wir in $d = 2$ finite Differenzen und den diskreten Laplace-Operator Δ_h auf dem Gitter für $h = \frac{1}{N+1}$

$$\Omega_h := \{(kh, lh) :\ 1 \leq k, l \leq N :\ k, l \in \mathbb{N}\}.$$

Für den diskreten Laplace-Operator, der für $x \in \Omega$ durch

$$\begin{aligned} \Delta_h u_h(x,t) := &\frac{u_h(x + (h,0), t) - 2u_h(x,t) + u_h(x - (h,0), t)}{h^2} \\ &+ \frac{u_h(x + (0,h), t) - 2u_h(x,t) + u_h(x - (0,h), t)}{h^2} \end{aligned}$$

gegeben ist, was dem 5-Punkte-Stern entspricht, suchen wir nun $u_h : \overline{\Omega}_h \times [0,T] \to \mathbb{R}$ mit $u_h(x,t) \approx u(x,t)$ für $x \in \Omega_h$ suchen wir $u_h : \overline{\Omega}_h \times [0,T] \to \mathbb{R}$ als Lösung von

$$\begin{cases} \partial_t u_h = \Delta_h u_h, & x \in \Omega_h,\ 0 < t \leq T, \\ u_h(x,0) = 0, & x \in \Gamma_h,\ 0 < t \leq T, \\ u_h(x,0) = u_0(x), & x \in \Omega_h, \end{cases}$$

wobei $\overline{\Omega}_h := \Omega_h \cup \Gamma_h$ und $\Gamma_h = \partial\Omega_h$ ist.

Dies ergibt ein großes Systeme von gewöhnlichen Differentialgleichungen, wobei der betragsgrößte Eigenwert von Δ_h etwa $-\frac{c}{h^2} \to -\infty$ ist (Übung!). Dabei gilt $c = 4$ für $d = 1$.

Für $d = 2$ erhalten wir für $(N+1)h = 1$ und $U = (U_k)_{k=1}^{N^2}$ das System

$$U_{i+N_j}(t) = u_h(ih, jh, t).$$

Anschließend müssen wir das gewöhnliche Differentialgleichungssystem $\partial_t U = -AU$ zum Anfangswert $U(0) = U_0$ in der Zeit lösen und hierfür die Zeit diskretisieren.

2.3 Beispiel (Zeitdiskretisierung mit dem expliziten Euler-Verfahren)
Hierfür betrachten wir zunächst das explizite Euler-Verfahren, um $U^n \approx U(t_0)$ zu lösen. Wir iterieren für $n = 0, 1, \ldots$ und $n\tau \le T$

$$U^{n+1} = U^n - \tau A U^n. \tag{2.1}$$

Wir zeigen nun, daß dies nicht stabil ist. Hierfür diagonalisieren wir A via $A = Q\Lambda Q^{-1}$ mit $\Lambda = \operatorname{diag}(\lambda_j)$, wobei die λ_j alle reell sind mit $0 < \lambda_j \le \frac{8}{h^2}$. Durch Multiplikation des Euler-Verfahrens, (2.1), mit Q^{-1} erhalten wir für $Q^{-1}U^n =: Y^n$

$$Y^{n+1} = Y^n - \tau \Lambda Y^n.$$

Dabei gilt für die k-te Komponente hiervon

$$y_k^{n+1} = y_k^n - \tau\lambda_k y_k^n = (1 - \tau\lambda_k) y_k^n.$$

Es ist $|1 - \tau\lambda_k| \le 1$ nur, falls $\tau\lambda_k \le 2$ für alle k ist, i.e. wir haben die Schrittweitenbeschränkung $\tau \le \frac{h^2}{4}$, sonst wird das Verfahren instabil.

Ähnliches gilt auch für explizite Runge–Kutta-Verfahren oder Mehrschrittverfahren.

Wir betrachten nun $A(\alpha)$- bzw. A-stabile Verfahren zur Zeitdiskretisierung.

2.4 Beispiel (Zeitdiskretisierung mit dem impliziten Euler-Verfahren)
Wir betrachten nun das implizite Euler-Verfahren, $U^{n+1} = U^n - \tau A U^{n+1}$, und müssen damit in jedem Zeitschritt das lineare Gleichungssystem $(I + \tau A)U^{n+1} = U^n$ lösen, i.e. für eine Komponente hiervon gilt $(1 + \tau\lambda_k) y_k^{n+1} = y_k^n$ bzw. da alle $\lambda_k > 0$ sind

$$y_k^{n+1} = \underbrace{\frac{1}{1 + \tau\lambda_k}}_{<1 \ \forall t>0} y_k^n.$$

Damit ist dieses Verfahren stabil für beliebige Zeitschritte $\tau > 0$ und unabhängig von h (i.e. wir haben keine Zeitschrittweitenbeschränkung durch die Raumdiskretisierung).

2.5 Beispiel (Crank–Nicolson-Verfahren)
Wir betrachten das Crank–Nicholson-Verfahren,

$$\frac{U^{n+1}-U^n}{\tau} = -A\frac{U^{n+1}+U^n}{2},$$

welches der impliziten Mittelpunktsregel entspricht und das äquivalent durch

$$\left(I+\frac{\tau}{2}A\right)U^{n+1} = \left(I-\frac{\tau}{2}A\right)U^n$$

charakterisiert wird. Für eine Komponente hiervon gilt

$$\left(1+\frac{\tau}{2}\lambda_k\right)y_k^{n+1} = \left(1-\frac{\tau}{2}\lambda_k\right)y_k^n,$$

also $y_k^{n+1} = R(-\tau\lambda_k)y_k^n$ für

$$R(z) := \frac{1+\frac{z}{2}}{1-\frac{z}{2}}.$$

Wir beachten, daß $R(-\tau\lambda) \to -1$ für $\lambda \to \infty$ gilt. Insgesamt ist dieses Verfahren für beliebige $\tau > 0$ und unabhängig von h stabil.

2.2 Schwache Formulierung parabolischer DGL

2.6 Wiederholung (schwache Formulierung)
Wir betrachten als typisches Beispiel die Wärmeleitungsgleichung ohne Randwert, i.e.

$$\begin{cases} \partial_t u = \Delta u + f, & \text{in } \Omega\times(0,T), \\ u = 0, & \text{auf } \partial\Omega\times(0,T), \\ u = u_0, & \text{in } \Omega\times\{0\}. \end{cases}$$

Wir multiplizieren nun die Gleichung mit $v \in H_0^1(\Omega)$, integrieren über Ω und benutzen die partielle Integration, womit wir die *schwache Formulierung*

$$\int_\Omega \partial_t uv + \underbrace{\int_\Omega \nabla u\cdot\nabla v}_{=:a(u,v)} = \int_\Omega fv$$

erhalten. Wir kennen bereits die elliptische Bilinearform a auf $H_0^1(\Omega)$. Dabei ist

$$H_0^1(\Omega) := \overline{\mathcal{C}_0^\infty(\Omega)}^{\|\cdot\|_{H^1(\Omega)}}$$

der Abschluß von $\mathcal{C}_0^\infty(\Omega)$ bezüglich der H^1-Norm; er kann mit dem Spursatz durch

$$H_0^1(\Omega) = \left\{ v : \Omega \to \mathbb{R} : \int_\Omega v^2 < \infty,\ \int_\Omega |\nabla v|^2 < \infty,\ v = 0 \text{ auf } \partial\Omega \right\}$$

charakterisiert werden. Weiter haben wir auf dem Raum das Skalarprodukt

$$(u, v) := \int_\Omega uv + \int_\Omega \nabla u \cdot \nabla v,$$

und die induzierte Norm ist (mit der Poincare-Ungleichung) äquivalent zur Norm

$$|u|_1 := \int_\Omega |\nabla u|^2.$$

Wir zeigen nun einige grundlegene Eigenschaften eines solchen Problems für abstraktere Situationen. Die folgenden Annahmen seien für den Rest des Kapitels gültig.

2.7 Definition
Hierzu sei V ein Hilbertraum mit der Norm $\|.\|$, $a : V \times V \to \mathbb{R}$ eine Bilinearform, die symmetrisch und V-elliptisch ist, i.e. es gibt $0 < \alpha, M < \infty$ mit

$$\begin{aligned} a(v, w) &= a(w, v) && \forall v, w \in V, \\ |a(v, w)| &\le M \cdot \|v\|\|w\| && \forall v, w \in V, \\ a(v, v) &\ge \alpha \|v\|^2 && \forall v \in V. \end{aligned}$$

Damit ist $a(v, .) : V \to \mathbb{R}$ linear und stetig, i.e. $a(v, .) \in V'$, der mit der Norm

$$\|\varphi\|_* := \|\varphi\|_{V'} := \sup_{\|v\| \le 1} |\varphi(v)|$$

für $\varphi \in V'$ ausgestattet ist. Wir schreiben oft auch die duale Paarung von $\langle \varphi, v \rangle := \varphi(v) \in \mathbb{R}$.

Schließlich sei noch $v \mapsto a(v, .) : V \to V'$ linear und stetig. Wir definieren nun $A : V \to V'$ via $Av := a(v, .)$ und erhalten automatisch damit $\langle Av, w \rangle = a(v, w)$.

2.8 Hilfssatz (Eigenschaften von A)
Die Abbildung $A : V \to V'$ aus Definition 2.7 ist linear, stetig und bijektiv.

Beweis
Klarerweise ist A linear. Wir rechnen zunächst

$$\|Av\|_{V'} = \sup_{\|w\|=1} |\langle Av, w \rangle| = \sup_{\|w\|=1} |a(v, w)| \le M\|v\| \quad \forall v \in V,$$

i.e. A ist stetig. Sei nun $\varphi \in V'$. Wir zeigen, daß es genau ein $u \in V$ gibt mit $Au = \varphi$. Dies ist dazu äquivalent, daß

$$\langle Au, v \rangle = \langle \varphi, v \rangle \quad \forall v \in V \iff a(u, v) = \varphi(v) \quad \forall v \in V,$$

wobei letztere Aussage gerade mit dem Satz von Lax-Milgram wahr ist. □

Wir betrachten nun einen weiteren Hilbertraum H mit der Norm $|.|$ und identifizieren ihn gemäß Frechet-Riesz mit seinem eigenen Dualraum H' via $v = (v, .)_H$. Wir fordern weiter, daß $V \subseteq H$ dicht ist und durch die Interpretation via $\langle w, v\rangle = (w, v)$ für $w \in H$ und $v \in V \subseteq H$ gilt $H \subseteq V'$. Weiter sei $u_0 \in H$ und $f : [0, T] \to H$ stetig.

Wir haben also zusammenfassend folgende Situation: Gegeben sind die Hilberträume $(H, |.|)$, $(V, \|.\|)$ und $(V', \|.\|_{V'})$ mit den dichten stetigen Einbettungen (Übung für die zweite!)

$$V \subseteq H \cong H' \subseteq V'.$$

Ein solches Tripel (V, H, V') heißt auch Gelfand-Tripel. Wir werden die Theorie später auf $V = H_0^1(\Omega)$, $H = L^2(\Omega)$ und $f(t) = f(., t)$ anwenden.

2.9 Problem (Schwache Formulierung)
Für $u_0 \in H$ und $f : [0, T] \to H$ suchen wir $u : (0, T) \to V$ stetig differenzierbar mit

$$\begin{cases} (\partial_t u(t), v) + a(u, v) = (f(t), v) & \forall v \in V,\ 0 < t < T, \\ |u(t) - u_0| \to 0 & \text{für } t \searrow 0. \end{cases}$$

Wegen $\langle \partial_t u, v\rangle + \langle Au, v\rangle = (\partial_t u, v) + a(u, v)$ ist dies äquivalent, $u : (0, T) \to V$ zu finden mit

$$\begin{cases} \partial_t u(t) + Au = f(t) & \text{in } V', \\ u(0+) = u_0 & \text{in } H. \end{cases}$$

Wir betrachten nun aufgrund der Linearität zwei getrennte Differentialgleichungen. Zum einen die homogene Differentialgleichung $\partial_t v + Av = 0$ mit $v(0+) = u_0$ und die inhomogene Differentialgleichung mit 0 als Anfangswert, $\partial_t w + Aw = f(t)$ mit $w(0+) = 0$.

Falls v und w Lösungen dieser Gleichungen sind, dann ist $u := v + w$ eine Lösung von 2.9.

2.10 Wiederholung
Falls wir $A \in \mathbb{R}^{N\times N}$ als Matrix haben, so hat die Differentialgleichung

$$\begin{cases} \partial_t u + Au = f(t), \\ u(., 0) = u_0 \end{cases}$$

die Lösung

$$u(t) = e^{-tA}u_0 + \int_0^t e^{-(t-s)A} f(s)\, ds.$$

Dieses Konzept möchten wir im Folgenden auf stetige Operatoren in Banachräumen verallgemeinern und eine Lösungsformel angeben. Für einen Operator $A : V \to V'$ ist

$$e^{-tA} = \sum_{n=0}^{\infty} \frac{(-tA)^n}{n!}$$

nicht notwendigerweise definiert. Für endlich-dimensionale Fälle gilt allerdings, falls alle Eigenwerte von A im Inneren des Kreisrandes Γ liegen, daß (Übung!)

$$e^{-tA} = \frac{1}{2\pi\imath} \int_\Gamma e^{\lambda t}(\lambda I + A)^{-1}\, d\lambda,$$

wobei wir $(\lambda I + A)^{-1}$ die Resolvente zu λ nennen.

Der folgende Satz stellt das Hauptresultat dieses Abschnitts dar und ergibt sich aus einer Reihe von Lemmata.

2.11 Satz (Lösung des homogenen Problems)
Das homogene Problem

$$\begin{cases} \partial_t u(t) + Au = 0 & \text{in } V', \\ u(0+) = u_0 & \text{in } H \end{cases}$$

hat für jedes $u_0 \in H$ eine eindeutig bestimmte Lösung $u : (0,T) \to V$, welche dort analytisch ist. Dabei gilt $Au(t) \in H$ und es gelten für alle $t \in (0,T)$ die folgenden a-priori Abschätzungen für alle $k \in \mathbb{N}$

$$|u(t)|^2 + 2\alpha \int_0^t \|u(s)\|^2\, ds \le |u_0|^2, \qquad \|u(t)\| \le \frac{C}{\sqrt{t}}|u_0|,$$

$$|Au(t)| \le \frac{C}{t}|u_0|, \qquad \left|\partial_t^k u(t)\right| \le \frac{C_k}{t^k}|u_0| \qquad \forall k \in \mathbb{N}.$$

2.12 Bemerkung
Wir erhalten, daß die Lösung für beliebiges $t > 0$ bereits analytisch ist, i.e. die Lösung wird (selbst bei irregulärem Anfangswert u_0) instantan glatt und die Lösung sowie alle Ableitungen hiervon fallen mit den Abschätzungen im Unendlichen stark ab.

2.13 Definition (Komplexifizierung von A)
Wir betrachten nun die *Komplexifizierung* der linearen Abbildung A. Wir setzen hierfür $V_\mathbb{C} := V \oplus \imath V$, wobei wir die Multiplikation via

$$\underbrace{(\alpha_1 + \imath\alpha_2)}_{\in\mathbb{C}}\underbrace{(v_1 \oplus \imath v_2)}_{\in V_\mathbb{C}} := (\alpha_1 v_1 - \alpha_2 v_2) \oplus \imath(\alpha_2 v_1 + \alpha_1 v_2)$$

definieren. Wir definieren dann $a_\mathbb{C} : V_\mathbb{C} \to \mathbb{C}$ durch

$$a_\mathbb{C}(v_1 + \imath v_2, w_1 + \imath w_2) := a(v_1, w_1) - a(v_2, w_2) + \imath a(v_2, w_1) + \imath a(v_1, w_2).$$

Durch diese Definition ist $a_{\mathbb{C}}$ sesquilinear, i.e.

$$a_{\mathbb{C}}(\alpha v, w) = \bar{\alpha} a_{\mathbb{C}}(v, w), \qquad a_{\mathbb{C}}(v, \alpha w) = \alpha a_{\mathbb{C}}(v, w).$$

Da a symmetrisch ist, wird $a_{\mathbb{C}}$ hermitesch, i.e. $a_{\mathbb{C}}(v, w) = \overline{a_{\mathbb{C}}(w, v)}$ und aus der V-Elliptizität von a erhalten wir $a_{\mathbb{C}}(v, v) \geq \alpha \|v\|^2 \quad \forall v \in V_{\mathbb{C}}$ für alle $v \in V_{\mathbb{C}}$.

Wir schreiben im Folgenden der Übersicht halber wieder V anstelle von $V_{\mathbb{C}}$ und a anstelle von $a_{\mathbb{C}}$, meinen aber die Komplexifizierung des Raums bzw. der Abbildung.

2.14 Hilfssatz (nicht symmetrische, komplexe Version von Lax-Milgram)
Sei V ein komplexer Hilbertraum mit der Norm $\|.\|$ und $b : V \times V \to \mathbb{C}$ eine Sesquilinearform, die stetig und V-elliptisch ist, i.e. es gibt $0 < \alpha, M < \infty$ mit

$$\begin{aligned} |b(v, w)| &\leq M\|v\|\|w\| && \forall v, w \in V, \\ \operatorname{Re} b(v, v) &\geq \alpha \|v\|^2 && \forall v \in V. \end{aligned}$$

Sei weiter $l : V \to \mathbb{C}$ eine stetige Linearform. Dann gibt es genau ein $u \in V$ mit

$$b(u, v) = l(v) \quad \forall v \in V.$$

Beweis
Wir bezeichnen mit $((.,.))$ das Skalarprodukt auf V zur Norm $\|.\|$. Nach der symmetrischen Version von Lax-Milgram gibt es genau ein $\varphi \in V$ mit

$$((\varphi, w)) = l(w) \quad \forall w \in V$$

und zu jedem $v \in V$ gibt es genau ein $Bv \in V$ mit $((Bv, w)) = b(v, w)$. Damit ist $B : V \to V$ linear und stetig, da wir mit den Voraussetzungen an b

$$\|Bv\| = \sup_{\|w\|=1} |((Bv, w))| = \sup_{\|w\|=1} |b(v, w)| \leq M\|v\| \tag{2.2}$$

erhalten. Damit gilt

$$\begin{aligned} b(u, v) = l(v) \quad \forall v \in V &\iff ((Bu - \varphi, v)) = 0 \quad \forall v \in V \\ &\iff Bu = \varphi \\ &\iff \text{für beliebige } \varepsilon > 0 \text{ gilt } \varepsilon Bu + u = u + \varepsilon\varphi. \end{aligned}$$

Wir betrachten nun die Fixpunktiteration

$$u^{(k+1)} := u^{(k)} - \varepsilon B u^{(k)} + \varepsilon\varphi = (I - \varepsilon B)u^{(k)} + \varepsilon\varphi,$$

welche einen eindeutig bestimmten Fixpunkt hat, falls $\|I - \varepsilon B\| < 1$ ist. Mit der Elliptizität von b und (2.2) erhalten wir

$$\|(I - \varepsilon B)v\|^2 = \|v\|^2 - 2\varepsilon \underbrace{\operatorname{Re}((Bv, v))}_{=\operatorname{Re} b(v,v) \geq \alpha\|v\|^2} + \varepsilon^2 \underbrace{\|Bv\|^2}_{\leq M^2\|v\|^2} \leq \underbrace{(1 - 2\varepsilon\alpha + \varepsilon^2 M^2)}\, \|v\|^2,$$

wobei der hervorgehobene Ausdruck für kleine ε echt kleine als 1 ist. Damit hat die Fixpunktiteration einen eindeutigen Fixpunkt u mit dem Banachschen Fixpunktsatz und $u^{(k)} \to u \in V$ und mit der Stetigkeit von von B gilt damit $Bu = \varphi$. □

2.15 Hilfssatz
Die Gleichung $\lambda \cdot (u,v) + a(u,v) = \langle f, v\rangle \quad \forall v \in V$ hat für $\lambda \in \mathbb{C}$ mit $|\arg\lambda| \le \pi - \theta$ für $0 < \theta < \frac{\pi}{2}$ und $f \in V'$ eine eindeutige Lösung $u \in V$ mit den a-priori Abschätzungen

$$\|u\| \le \frac{1}{\alpha} \underbrace{\frac{1}{\sin\theta}}_{=:c_\theta} \|f\|_{V'},$$
$$|u| \le \frac{c_\theta}{|\lambda|}|f|, \text{ falls } f \in H.$$

Beweis
Wir möchten die Aussage auf Hilfssatz 2.14 zurückführen und konstruieren noch die Abbildung b hierfür. Es gilt für $\lambda(v,v) = \lambda|v|^2$ und $|\arg\lambda| \le \pi - \theta$ (o.B.d.A. $\operatorname{Im}\lambda \ge 0$), daß

$$\operatorname{Re} e^{-\imath\left(\frac{\pi}{2}-\theta\right)}(\lambda\cdot(v,v) + a(v,v)) \ge \alpha\sin\theta\|v\|^2.$$

Wir wählen nun

$$b(u,v) := e^{-\imath\left(\frac{\pi}{2}-\theta\right)}(\lambda\cdot(u,v) + a(u,v))$$

und erhalten damit

$$|b(u,v)| \le |\lambda| \underbrace{|u||v|}_{\le C\|u\|\|v\|} + M\|u\|\|v\| \le M(\lambda)\|u\|\|v\| \quad \forall u,v \in V,$$
$$\operatorname{Re} b(u,v) \ge \underbrace{\alpha\sin\theta}_{>0}\|v\|^2 \quad \forall v \in V.$$

Mit Hilfssatz 2.14 gibt es damit genau ein $u \in V$ mit

$$b(u,v) = \left\langle e^{-\imath\left(\frac{\pi}{2}-\theta\right)} f, v\right\rangle \quad \forall v \in V, \tag{2.3}$$

also auch eine Lösung unserer Gleichung (per Multiplikation mit dem Vorfaktor). Wir müssen nun noch die Abschätzung für u zeigen. Für $v = u$ in (2.3) erhalten wir

$$\alpha\sin\theta\|u\|^2 \le \operatorname{Re} b(u,u) \le |b(u,u)| = |\langle f, u\rangle| \le \|f\|_{V'}\|u\|,$$

also gilt

$$\|u\| \le \frac{1}{\alpha\sin\theta}\|f\|_{V'}.$$

Für $f \in H$ gilt mit der Cauchy–Schwarz-Ungleichung und per Konstruktion

$$\underbrace{|\lambda|u|^2 + a(u,u)|}_{=|(f,u)|\le|f||u|} \ge |\lambda||u|^2\sin\theta,$$

also gilt auch

$$|u| \leq \frac{1}{|\lambda| \sin\theta} |f|.$$

□

2.16 Bemerkung
Nach Hilfssatz 2.15 hat also $\lambda u + Au = f$ in V' eine eindeutige Lösung u. Wir definieren nun den Ausdruck $(\lambda + A)^{-1} f := u$. Damit ist $(\lambda + A)^{-1} : V' \to V$ der Lösungsoperator der Gleichung.

2.17 Hilfssatz
Für $u_0 \in H$ und $\Gamma : \{\lambda \in \mathbb{C} : \ |\arg \lambda| = \pi - \theta\}$ ist der Ausdruck

$$u(t) := \frac{1}{2\pi\imath} \int_\Gamma e^{\lambda t} (\lambda + A)^{-1} u_0 \, d\lambda$$

als V-wertiges uneigentliches Riemann-Integral wohldefiniert und reellanalytisch auf $(0, \infty)$ und es gilt $u' + Au = 0$ in V.

Beweis
Nach Hilfssatz 2.15 gilt

$$\|(\lambda + A)^{-1} u_0\| \leq \frac{1}{\alpha \sin\theta} \|u_0\|_{V'} \leq \frac{1}{\alpha \sin\theta} |u_0|,$$

also ist der Integrand gleichmäßig in Γ abgeschätzt und $e^{\lambda t}$ fällt für $t > 0$ entlang Γ ab. Damit existiert das Integral. Mit der Vertauschung von Differentation und Integration gilt

$$u'(t) = \frac{1}{2\pi\imath} \int_\Gamma e^{\lambda t} \underbrace{\lambda(\lambda + A)^{-1}}_{I - A(\lambda + A)^{-1}} u_0 \, d\lambda = \frac{1}{2\pi\imath} \underbrace{\int_\Gamma e^{\lambda t} u_0}_{=0} - A \underbrace{\frac{1}{2\pi\imath} \int_\Gamma e^{\lambda t} (\lambda + A)^{-1} u_0 \, d\lambda}_{= u(t)},$$

wobei das erste Integral aufgrund des Cauchyschen Integralsatzes verschwindet. Insgesamt erhalten wir damit, daß $u'(t) + Au(t) = 0$ für alle $t > 0$ gilt. □

2.18 Hilfssatz
Sei u wie in Hilfssatz 2.17 definiert.

1. Falls $u_0 \in D(A) := \{v \in V : \ Av \in H\}$ ist, dann gilt
$$|u(t) - u_0| \leq Ct |Au_0| \quad \forall t > 0.$$
2. Für beliebige $u_0 \in H$ gilt
$$\lim_{t \searrow 0} |u(t) - u_0| = 0.$$

Beweis

1. Mit der Definition von u und für

$$\Gamma_t := \left\{ \lambda \in \mathbb{C} : \left| \arg\left(\lambda + \frac{1}{t}\right) \right| = \pi - \theta \right\}$$

eine Verschiebung von Γ um $\frac{1}{t}$ nach rechts rechnen wir

$$\begin{aligned} |u(t) - u_0| &= \left| \frac{1}{2\pi i} \int_{\Gamma_t} e^{\lambda t} (\lambda + A)^{-1} u_0 \, d\lambda - \underbrace{\frac{1}{2\pi i} \int_{\Gamma_t} e^{\lambda t} \frac{1}{\lambda} u_0 \, d\lambda}_{=u_0 \text{ mit dem Residuensatz}} \right| \\ &\leq \frac{1}{2\pi} \int_{\Gamma_t} |e^{\lambda t}| \underbrace{|(\lambda + A)^{-1} u_0 - \lambda^{-1} u_0|}_{=\lambda^{-1}(\lambda+A)^{-1} A u_0} \, d|\lambda|, \end{aligned}$$

wobei $d|\lambda|$ bedeutet, daß wir eine feste Parametrisierung genommen haben und hiervon den Ausdruck betragsmäßig abschätzen. Mit Hilfssatz 2.15 für $f := Au_0$ und damit $|(\lambda + A)^{-1} A_0| \leq \frac{1}{|\lambda|} c_\theta |Au_0|$ erhalten wir dann

$$\leq \frac{1}{2\pi} \int_{\Gamma_t} |e^{\lambda t}| \frac{c_\theta}{|\lambda|^2} |Au_0| \, d|\lambda| = t \underbrace{\frac{1}{2\pi} \int_{\Gamma_1} |e^{\mu}| \frac{1}{|\mu|^2} \, d|\mu|}_{=:C} |Au_0|,$$

wobei wir im letzten Schritt $\mu = \lambda t$ eingesetzt haben.

2. Wir zeigen, daß $D(A)$ dicht in H ist.

Wir wissen bereits, daß $H \subseteq V'$ dicht ist und mit Hilfssatz 2.15 für $\lambda = 0$ ist $A^{-1} : V' \to V$ stetig und bijektiv, i.e. $D(A) = A^{-1}(H)$ ist dicht in $A^{-1}(V') = V$, also ist $D(A)$ dicht in V und V ist nach Voraussetzung dicht in H, womit $D(A)$ dicht in H ist.

Sei nun $(u_{0,n})$ eine Folge in $D(A)$ mit $|u_{0,n} - u_0| \leq \frac{1}{n}$ und seien $u_n(t)$ wie $u(t)$ konstruiert, alledings durch $u_{0,n}$ anstelle von u_0. Dann gilt mit dem ersten Teil

$$|u(t) - u_0| \leq |u(t) - u_n(t)| + \underbrace{|u_n(t) - u_{0,n}|}_{\leq Ct|Au_{0,n}|} + \underbrace{|u_{0,n} - u_0|}_{\leq \frac{1}{n}} \tag{2.4}$$

Wir rechnen weiter

$$\begin{aligned} |u(t) - u_n(t)| &= \left| \frac{1}{2\pi i} \int_{\Gamma_t} e^{\lambda t} (\lambda + A)^{-1} (u_0 - u_{0,n}) \, d\lambda \right| \\ &\leq \frac{1}{2\pi} \int_{\Gamma_t} |e^{\lambda t}| \underbrace{\left| (\lambda + A)^{-1} (u_0 - u_{0,n}) \right|}_{\leq \frac{C_\theta}{|\lambda|} |u_0 - u_{0,n}| \leq \frac{C_\theta}{|\lambda|} \frac{1}{n}} \, d|\lambda|, \end{aligned}$$

wobei wir Hilfssatz 2.15 benutzt haben. Für $\mu = \lambda t$ gilt nun schließlich

$$= \underbrace{\frac{1}{2\pi} \int_{\Gamma_1} |e^{\mu}| \frac{C_\theta}{|\mu|} \, d|\mu|}_{=:C^*} \frac{1}{n} = C^* \frac{1}{n},$$

also erhalten wir aus (2.4) für $t \leq \frac{1}{|Au_{0,n}|}$

$$|u(t) - u_0| \leq \frac{C^* + C + 1}{n} \to 0 \quad \text{für } t \searrow 0.$$

□

2.19 Hilfssatz (A-priori Abschätzung)
Für jede Lösung $u(t)$ von

$$\begin{cases} \partial_t u + Au = 0 & \text{in } V', \\ u(0+) = u_0 & \text{in } H \end{cases}$$

gilt für alle $t \in (0, T)$ die a-priori-Abschätzung

$$|u(t)|^2 + 2\alpha \int_0^t \|u(s)\|^2 \, ds \leq |u_0|^2.$$

Beweis
Da $\partial_t u(t) + Au(t) = 0$ in V' gilt, gilt auch

$$\underbrace{\langle \partial_t u(t), v \rangle}_{=(\partial_t u(t), v)} + \underbrace{\langle Au(t), v \rangle}_{=a(u(t),v)} = 0 \quad \forall v \in V.$$

Insbesondere gilt dies für $v = u(t)$, womit wir

$$\underbrace{(\partial_t u(t), u(t))}_{=\frac{1}{2}\partial_t |u(t)|^2} + \underbrace{a(u(t), u(t))}_{\geq \alpha \|u(t)\|^2} = 0$$

erhalten. Integration von 0 bis t liefert schließlich

$$\frac{1}{2}(|u(t)|^2 - |u_0|^2) + \alpha \int_0^t \|u(s)\|^2 \, ds \leq 0.$$

□

2.20 Bemerkung
Wir haben damit Satz 2.11 gezeigt: Das homogene Problem hat eine Lösung (cf. Hilfssatz 2.17). Diese Lösung ist über eine Resolventenformel gemäß Hilfssatz 2.17 gegeben. Die Lösung nimmt den Anfangswert wie gefordert an (cf. Hilfssatz 2.18) und für die homogene Lösung gilt für a-priori-Abschätzung (cf. Hilfssatz 2.19). Die restlichen Abschätzungen sind eine Übungsaufgabe.

2.21 Definition
Wir sehen, daß die Lösung u vom Anfangswert u_0 linear abhängt, also bezeichnen wir im Folgenden $u(t) =: U(t)u_0$, wobei damit $U(t) : H \to V$ linear ist für $t > 0$.

2.22 Satz (Lösung des inhomogenen Problems)
Sei $f : [0,T] \to H$ stetig differenzierbar. Dann hat das inhomogene Problem,

$$\begin{cases} \partial_t u(t) + Au(t) = f(t) & \text{in } V', \\ u(0+) = u_0 & \text{in } H \end{cases}$$

eine eindeutige Lösung

$$u(t) = U(t)u_0 + \int_0^t U(t-s)f(s)\,ds$$

mit der a-priori Energie-Abschätzung

$$|u(t)|^2 + \alpha \int_0^t \|u(s)\|^2\,ds \le |u_0|^2 + \frac{1}{\alpha}\int_0^t \|f(s)\|_{V'}^2\,ds.$$

Beweis

1. Wir betrachten den Ansatz

$$u(t) = U(t)u_0 + \int_0^t U(s)f(t-s)\,ds. \tag{2.5}$$

Für die Ableitung hiervon ergibt sich aufgrund der Definition von $U(t)$

$$\partial_t u(t) = -AU(t)u_0 + U(t)f(0) + \int_0^t U(s)\partial_t f(t-s)\,ds$$

und bei Anwendung des Operators A ergibt sich mit partieller Integration (wieder mit der Definition von $U(t)$)

$$\begin{aligned} Au(t) &= AU(t)u_0 + \int_0^t AU(s)f(t-s)\,ds \\ &= AU(t)u_0 - \underbrace{U(s)f(t-s)\Big|_{s=0}^t}_{=U(t)f(0)-U(0)f(t)} - \int_0^t U(s)\partial_t f(t-s)\,ds \\ &= AU(t)u_0 - U(t)f(0) + f(t) - \int_0^t U(s)\partial_t f(t-s)\,ds, \end{aligned}$$

wobei wir beachten, daß $U(0) = \mathrm{id}$ ist. Addition dieser beiden Zeilen ergibt schließlich

$$\partial_t u(t) + Au(t) = f(t) \quad \text{in } V'.$$

Für $t \searrow 0$ in (2.5) ergibt sich außerdem $u(0+) = u_0$ in H.

Damit existiert eine Lösung und sie berechnet sich über (2.5). Die Eindeutigkeit aus der Linearität der Gleichung folgt direkt aus der Linearität und Satz 2.11.

2. Wenn wir die Operatoren auf die Lösung u anwenden, erhalten wir aus der Gleichung

$$\underbrace{(\partial_t u(t), u(t))}_{\frac{1}{2}\partial_t |u(t)|^2} + \underbrace{a(u(t), u(t))}_{\geq \alpha \|u\|^2} = \langle f, u \rangle.$$

Durch Integration von 0 bis t ergibt erhalten wir damit schließlich

$$\begin{aligned}\frac{1}{2}|\partial_t u(t)|^2 - \frac{1}{2}|u_0|^2 + \alpha \int_0^t \|u(s)\|^2 \, ds &\leq \int_0^t \langle f(s), u(s) \rangle \, ds \leq \int_0^t \|f(s)\|_{V'} \|u(s)\| \, ds \\ &\leq \int_0^t \left(\frac{1}{2\alpha} \|f(s)\|_{V'}^2 + \frac{\alpha}{2} \|u(s)\|^2 \right) ds,\end{aligned}$$

wobei wir im zuletzt die Cauchy-Ungleichung $2ab \leq a^2 + b^2$ verwendet haben. □

2.23 Bemerkung

1. Die Lösungsformel für die inhomogene Gleichung wird auch – in Analogie zum endlich-dimensionalen Fall – „Variation der Konstanten"- Formel genannt.
2. Die im zweiten Teil von Satz 2.22 benutzte Technik der Energieabschätzung wird im Folgenden oft verwendet.

2.3 Finite-Elemente Semidiskretisierung im Raum

2.24 Herleitung (der Semidiskretisierung)
Wir betrachten nun die schwache Formulierung des parabolischen Problems,

$$\begin{cases} (\partial_t u(t), v) + a(u(t), v) = (f(t), v) & \forall v \in V, \ \forall 0 < t \leq T, \\ (u(0+), w) = (u_0, w) & \forall w \in H, \end{cases} \tag{2.6}$$

wobei a eine V-elliptische Bilinearform ist und ersetzen V durch einen endlich-dimensionalen Unterraum V_h. Typische Beispiele sind etwa $V = H_0^1(\Omega)$, $H = L^2(\Omega)$ und V_h ist ein Finite-Elemente-Raum, e.g. V_h beinhaltet lokal affine, global stetige Funktionen mit Randwert 0. Wir suchen nun $u_h : [0, T] \to V_h$ mit

$$\begin{cases} (\partial_t u_h(t), v_h) + a(u_h(t), v_h) = (f(t), v_h) & \forall v_h \in V_h, \\ (u_h(0), w_h) = (u_0, w_h) & \forall w_h \in V_h. \end{cases} \tag{2.7}$$

Wir beachten, daß wir damit auch $H_h := V_h$ als Diskretisierung des Raumes H gewählt haben, was aufgrund der Dichtheit von $V \subseteq H$ kein Problem darstellt.

Wir betrachten nun die Basisdarstellung. Sei $(\varphi_1, \dots, \varphi_N)$ die Basis von V_h. Dann definieren wir die Massematrix M bzw. die Stefigkeitsmatrix K durch

$$M = (m_{ij})_{i,j=1}^N := ((\varphi_i, \varphi_j))_{i,j=1}^N, \qquad K = (k_{ij})_{i,j=1}^N := (a(\varphi_i, \varphi_j))_{i,j=1}^N.$$

Beide Matrizen sind symmetrisch und positiv definit. Wir schreiben nun

$$u_h(t) = \sum_{i=1}^N \mu_i(t)\varphi_i$$

und erhalten für den Vektor $\mu(t) := (\mu_1(t), \dots, \mu_N(t))^T$ ein System von gewöhnlichen Differentialgleichungen im $\mathbb{R}^n$ zum Anfangswert $\mu(0) = \mu_0$,

$$M\partial_t\mu(t) + K\mu(t) = l(t),$$

wobei $\mu_{0,i} = (u_0, \varphi_i)$ und $l_i(t) = (f(t), \varphi_i)$ ist. Dieses System ist typischerweise steif.

2.25 Bemerkung
Mit der Cholesky-Zerlegung $M = CC^T$ können wir dieses System für $y := C^T\mu$ sowie $A := C^{-1}KC^{-T}$, $g := C^{-1}l$ und $y_0 := C^T\mu_0$ äquivalent in $\partial_t y(t) + Ay(t) = g(t)$ zum Anfangswert $y(0) = y_0$ schreiben. Wir beachten jedoch, daß diese Zerlegung die schwache Besetzung von M zerstört und wir deswegen in der Praxis eine solche Umformung nicht durchführen.

2.26 Hilfssatz (Stabilität der Semidiskretisierung)
Sei $e_h : [0, T] \to V_h$ eine Lösung von

$$\begin{cases} (\partial_t e_h(t), v_h) + a(e_h(t), v_h) = (d(t), v_h) & \forall v_h \in V_h, \\ (e_h(0), v_h) = (e_0, v_h) & \forall v_h \in V_h. \end{cases}$$

Dann gilt

$$|e_h(t)|^2 + \alpha \int_0^t \|e_h(s)\|^2 \, ds \le |e_0|^2 + \frac{1}{\alpha} \int_0^t \|d(s)\|_{V'}^2 \, ds.$$

Beweis
Wir setzen $v_h = e_h$ in die Gleichung und erhalten

$$\underbrace{(\partial_t e_h(t), e_h(t))}_{=\frac{1}{2}\partial_t |e_h(t)|^2} + \underbrace{a(e_h, e_h)}_{\ge \alpha \|e_h\|^2} = (d, e_h(t)) = \langle d, e_h(t) \rangle \le \|d\|_{V'} \|e_h(t)\| \le \frac{1}{2\alpha} \|d\|_{V'}^2 + \frac{\alpha}{2} \|e_h(t)\|^2,$$

also

$$\partial_t |e_h(t)|^2 + \alpha \|e_h\|^2 \leq \frac{1}{\alpha} \|d\|^2_{V'}.$$

Integration von 0 bis t liefert schließlich die Behauptung, wenn wir beachten, daß aus

$$|e_h(0)|^2 = (e_h(0), e_h(0)) = (e_h(0), e_0) \leq |e_h(0)||e_0|$$

stets $|e_h(0)| \leq |e_0|$ folgt. □

Unser nächstes Ziel ist die Konvergenz der Semidiskretisierung für $h \to 0$. Hierfür definieren wir nun zwei Projektionen.

2.27 Wiederholung (orthogonale Projektionen)
Wir erinnern uns an die beiden orthogonalen Projektionen.

1. H-orthogonale Projektion: Für $u \in H$ definieren wir $P_h u \in V_h$ durch
$$(P_h u, v_h) = (u, v_h) \quad \forall v_h \in V_h.$$
2. V-orthogonale Projektion: Für $u \in V$ definieren wir $R_h u \in V_h$ durch
$$a(R_h u, v_h) = a(u, v_h) \quad \forall v_h \in V_h.$$

Nach dem Satz von Lax-Milgram existieren solche Projektionen.

Für $l(v_h) := a(u, v_h)$ ist $R_h u$ die Finite-Elemente Lösung von $a(u_h, v_h) = l(v_h) \quad \forall v_h \in V_h$ zum Problem $a(u, v) = l(v) \quad \forall v \in V$.

Wir wissen für stationäre lineare Finite-Elemente für $V = H^1_0(\Omega)$ und $H = L^2(\Omega)$, daß für die V-orthogonale Projektion die folgenden Abschätzungen gelten, falls $u \in H^2(\Omega)$ ist.

$$\begin{cases} \|R_h u - u\|_{H^1(\Omega)} \leq Ch \|\nabla^2 u\|_{L^2(\Omega)}, \\ \|R_h u - u\|_{L^2(\Omega)} \leq Ch^2 \|\nabla u\|_{L^2(\Omega)}, \\ \|P_h u - u\|_{L^2(\Omega)} = \min\limits_{v_h \in V_h} |v_h - u| \leq Ch^2 \|\nabla^2 u\|_{L^2(\Omega)}. \end{cases} \tag{2.8}$$

2.28 Satz (Konvergenz)
Sei u die Lösung von des kontinuierlichen Problems (2.6) genügend regulär, u_h die Lösung des diskreten Problems (2.7), und für den Finite-Elemente-Raum gelte (2.8). Dann gibt es Konstanten $0 < c_1, c_2 < \infty$ (die von u abhängen) mit

$$\left(\int_0^T \|u_h(t) - u(t)\|^2 \, dt \right)^{\frac{1}{2}} \leq c_1 h,$$

$$\max_{0 \leq t \leq T} |u_h(t) - u(t)| \leq c_2 h^2.$$

Beweis
Wegen der Gültigkeit der kontinuierlichen Gleichung (2.6) gilt

$$(R_h\partial_t u, v_h) + a(R_h u, v_h) = (f, v_h) + (R_h\partial_t u - \partial_t u, v_h) \quad \forall v_h \in V_h.$$

Wir ziehen nun von dieser Gleichung die diskrete Gleichung (2.7) ab und erhalten

$$(R_h\partial_t u - \partial_t u, v_h) + a(R_h u - u_h, v_h) = (R_h\partial_t u - \partial_t u, v_h) \quad \forall v_h \in V_h. \tag{2.9}$$

Für $e_h := u_h - R_h u \in V_h$ gilt also und $R_h\partial_t u = \partial_t(R_h u)$

$$(\partial_t e_h, v_h) + a(e_h, v_h) = (\partial_t u - R_h\partial_t u, v_h) \quad \forall v_h \in V_h.$$

Es gilt nun

$$e_h(0) = \underbrace{P_h u_0}_{=u_h(0)} - \underbrace{R_h u_0}_{=u(0)} = \underbrace{(P_h u_0 - u_0)}_{=\mathcal{O}(h^2)} - \underbrace{(R_h u_0 - u_0)}_{=\mathcal{O}(h^2)} = \mathcal{O}(h^2),$$

wobei die erste Abschätzung eine Übung ist und die zweite Abschätzung aufgrund von (2.8) gilt. Damit ist $|e_h(0)| \leq Ch^2$.

Aus der stetigen Einbettung $H \subseteq V'$ und (2.8) erhalten wir

$$\|R_h\partial_t u - \partial_t u\|_{V'} \leq C|R_h\partial_t u - \partial_t u| \leq Ch^2.$$

Damit und mit dem Hilfssatz 2.26 gilt schließlich

$$|e_h(t)|^2 + \alpha \int_0^t \|e_h(s)\|^2 \, ds \leq (Ch^2)^2,$$

also gilt wieder mit (2.8)

$$u_h - u = \underbrace{e_h}_{=u_h - R_h u = \mathcal{O}(h^2)} + \underbrace{R_h u - u}_{=\mathcal{O}(h^2)},$$

womit die Behauptung folgt. □

2.4 Vollständige Diskretisierung mit dem Euler-Verfahren

2.29 Herleitung (der Volldiskretisierung)
Wir diskretisieren in diesem Abschnitt die raumdiskretisierte Gleichung (2.7) auch in der Zeit durch das implizite Euler-Verfahren und erhalten damit das lineare Gleichungssystem

$$\begin{cases} \left(\frac{u_n - u_{n-1}}{\tau}, v_h\right) + a(u_n, v_h) = (f(t_n), v_h) \quad \forall v_h \in V_h,\ n = 1, 2, \ldots, \\ (u_0, w_0) = (u(0), w_h) \quad \forall w_h \in V_h. \end{cases} \tag{2.10}$$

Dabei ist $t_n = n\tau$ für die Zeitschrittweite $\tau > 0$ und wir erhalten $u_n \in V_h$ mit $u_n \approx u(t_n)$. Mit der zweiten Zeile projezieren wir den Startwert $u(0)$ in V_h und erhalten die Anfangsiterierte u_0.

2.30 Bemerkung (lineare Algebra der Volldiskretisierung)
Wenn wir

$$u_n = \sum_{i=1}^{N} \mu_{n,i} \varphi_i$$

mit $\mu_n = (\mu_{n,i})_{i=1}^N \in \mathbb{R}^N$ schreiben, ist (2.10) äquivalent zu

$$M \frac{\mu_n - \mu_{n-1}}{\tau} + K\mu_n = b(t_n),$$

wobei $M = ((\varphi_i, \varphi_j))_{i,j=1}^N$ die Massematrix, $K = (a(\varphi_i, \varphi_j))_{i,j=1}^N$ die Steifigkeitsmatrix und $b(t) = ((f(t), \varphi_i))_{i=1}^N$ ist. Dabei ist $\frac{1}{\tau}M + K$ symmetrisch und positiv definit.

Wir müssen also in jedem Zeitschritt ein lineares Gleichungssystem mit der symmetrischen, positiv definiten Matrix $\frac{1}{\tau}M + K$ lösen.

- In einer Raumdimension ist diese Matrix tridiagonal. Wir können sie einfach mit der Cholesky-Zerlegung lösen.
- In zwei Raumdimensionen ist die Matrix nur schwach besetzt und wir können sie noch immer direkt lösen.
- In drei Raumdimensionen benutzen wir iterativer Verfahren, etwa Mehrgitterverfahren oder das vorkonditionierte cg-Verfahren.

2.31 Hilfssatz (Stabilität der Volldiskretisierung)
Sei $e_0 \in V_h$ gegeben und $e_n \in V_h$ (für $n \geq 1$) die Lösung von

$$\left(\frac{e_n - e_{n-1}}{\tau}, v_h\right) + a(e_n, v_h) = (d_n, v_h) \quad \forall v_h \in V_h. \tag{2.11}$$

Dann gilt die Fehlerabschätzung

$$|e_n|^2 + \alpha\tau \sum_{j=0}^{n} \|e_j\|^2 \leq |e_0|^2 + \frac{1}{\alpha}\tau \sum_{j=1}^{n} \|d_j\|_{V'}^2.$$

Beweis
Wir setzen $v_h = e_h$ in (2.11) und erhalten damit

$$\left(\frac{e_n - e_{n-1}}{\tau}, e_n\right) + \underbrace{a(e_n, e_n)}_{\geq \alpha \|e_n\|^2} = (d_n, e_n) \leq \|d_n\|_{V'} \|e_n\| \leq \frac{\alpha}{2}\|e_n\|^2 + \frac{1}{2\alpha}\|d_n\|_{V'}^2.$$

Dies multiplizieren wir mit τ und summieren über n, womit wir

$$\sum_{j=1}^{n}(e_j - e_{j-1}, e_j) + \frac{\alpha}{2}\tau \sum_{j=1}^{n} \|e_j\|^2 \leq \frac{1}{2\alpha}\tau \sum_{j=1}^{n} \|d_j\|_{V'}^2 \tag{2.12}$$

erhalten. Für die Summe gilt mit Cauchy–Schwarz und der binomischen Formel

$$\begin{aligned}\sum_{j=1}^{n}(e_j - e_{j-1}, e_j) &= \frac{1}{2}|e_n| + \underbrace{\left(\frac{1}{2}|e_n|^2 \underbrace{-(e_n, e_{n-1})}_{\geq -|e_n||e_{n-1}|} + \frac{1}{2}|e_{n-1}|^2\right)}_{\geq \frac{1}{2}(|e_n| - |e_{n-1}|)^2 \geq 0} \\ &\quad + \ldots + \left(\frac{1}{2}|e_1|^2 - (e_1, e_0) + \frac{1}{2}|e_0|^2\right) - \frac{1}{2}|e_0|^2 \\ &\geq \frac{1}{2}|e_n|^2 - \frac{1}{2}|e_0|^2,\end{aligned}$$

womit wir mit (2.12) die Aussage erhalten. □

2.32 Satz (Konvergenz)
Sei die Lösung des kontinuierlichen Problems (2.6) genügend regulär und es gelte (2.8) für die Ritz-Projektion R_h. Dann gibt es eine Konstante $C = C(u)$ und für $n\tau \leq T$ gelten die beiden Abschätzungen

$$\begin{aligned}|u_n - u(t_n)| &\leq C(h^2 + \tau), \\ \left(\tau \sum_{j=1}^{n} \|u_j - u(t_j)\|^2\right)^{\frac{1}{2}} &\leq C(h + \tau).\end{aligned}$$

Beweis
Mit (2.9) und der Gültigkeit der Gleichung gilt für alle $v_h \in V_h$, daß

$$\begin{aligned}&\left(\frac{R_h u(t_n) - R_h u(t_{n-1})}{\tau}, v_h\right) + a(R_h u(t_n), v_h) \\ &= f(t_n, v_h) + \left(\frac{R_h u(t_n) - R_h u(t_{n-1})}{\tau} - R_h \partial_t u(t_n), v_h\right) + (R_h \partial_t u(t_n) - \partial_t u(t_n), v_h).\end{aligned} \tag{2.13}$$

Für $e_n := u_n - R_h u(t_n)$ erhalten wir mit der Stetigkeit der Ritz-Projektion R_h für

$$d_n := (R_h \partial_t u(t_n) - \partial_t u(t_n)) + \frac{1}{\tau}\Big(R_h u(t_n) - R_h u(t_{n-1})\Big) - R_h \partial_t u(t_n)$$

daß $\|d_n\|_{V'} \leq C|d_n| \leq C(h^2 + \tau)$ ist. Die Gleichung (2.13) ist äquivalent zu

$$\left(\frac{e_n - e_{n-1}}{\tau}, v_h\right) + a(e_n, v_h) = (d_n, v_h) \quad \forall v_h \in V_h,$$

also erhalten wir aus Hilfssatz 2.31, daß

$$|e_n|^2 + \alpha\tau \sum_{j=1}^{n} \|e_j\|^2 \leq |e_0|^2 + \frac{1}{\alpha}\tau \sum_{j=1}^{n} \|d_j\|_{V'}^2$$

und die Behauptung folgt aus $e_0 = 0$, der obigen Abschätzung von $\|d_n\|_{V'}$ sowie der Definition von e_n. □

2.33 Bemerkung
Wir haben nun mit der Raumdiskretisierung mit einem linearen Finite-Elemente Ansatz und der Zeitdiskretisierung mit dem impliziten Euler-Verfahren ein volldiskretes Verfahren kennengelernt, welches konvergiert. Wir sehen, daß die Konvergenz der Volldiskretisierung von der Raum- und Zeitdiskretisierung abhängt. Für eine optimale Konvergenz müssen die Parameter h und τ gekoppelt werden.

Um eine höhere Ordnung als die in Satz 2.32 zu erreichen, müssen wir im Finite-Elemente Ansatz Polynome eines höheren Grades betrachten bzw. ein Zeitschrittverfahren höhererer Ordnung, etwa das BDF-Verfahren oder das Radau-Verfahren als Beispiel eines impliziten Runge–Kutta-Verfahrens. Wir betrachten im Rest des Kapitels letztere Variante.

2.5 Zeitdiskretisierung mit BDF-Vefahren

2.34 Wiederholung
Wir haben in Abschnitt 1.3 die BDF-Verfahren kennengelernt, die für k Stützstellen Ordnung k haben, für $k > 6$ instabil sind und für $k \leq 6$ gemäß Beispiel 1.18 $A(\theta)$-stabil sind.

Wir wenden nun das BDF-Verfahren auf die parabolische Differentialgleichung,

$$\begin{cases} \partial_t u(t) + Au(t) = f(t) & \text{in } V', \\ u(0+) = u_0 & \text{in } H, \end{cases}$$

an und zeigen im Folgenden eine Fehlerabschätzung der Semidiskretisierung in der Zeit (Übungsaufgabe für die Volldiskretisierung!), i.e. zu gegebenen Startwerten $u_0, \dots, u_{k-1} \in V$ bestimmen wir iterativ für $n = k, k+1, \dots$ die Größen $u_k, u_{k+1}, \dots$ so, daß

$$\frac{1}{\tau} \sum_{j=0}^{k} \delta_j u_{n-j} + A u_n = f(t_n), \quad n \geq k. \tag{2.14}$$

Dabei sind die δ_j die Gewichte der BDF-Formel.

2.35 Hilfssatz (Stabilität)
Seien $e_0, \dots, e_{k-1} \in V$ und $d_j \in V'$ für $j \geq k$. Für $e_n \in V$ gelte

$$\frac{1}{\tau} \sum_{j=0}^{k} \delta_j e_{n-j} + A e_n = d_n. \tag{2.15}$$

Dann gilt

$$\tau \sum_{j=k}^{n} \|e_j\|^2 \leq C \left(t_n \sum_{i=0}^{k-1} \|e_i\|^2 + \tau \sum_{j=k}^{n} \|d_j\|_{V'}^2 \right). \tag{2.16}$$

Beweis

1. Seien zunächst $e_0 = \dots = e_{k-1} = 0$. Wir betrachten nun die folgenden Erzeugendenfunktionen als formale Reihe

   $$e(\zeta) = \sum_{n=0}^{\infty} e_n \zeta^n, \qquad d(\zeta) = \sum_{n=0}^{\infty} d_n \zeta^n$$

   und entsprechend definieren wir $\delta(\zeta)$. Wir zeigen nun (2.16) für $n \leq N$ für ein festes N und können dann ohne Beschränkung $d_n = 0$ für $n > N$ betrachten (wodurch die Reihe $\delta(\zeta)$ endlich wird). Wir multiplizieren (2.15) mit ζ^n, summieren über n und erhalten daraus (mit dem Cauchy-Produkt)

   $$\frac{1}{\tau} \delta(z) e(\zeta) + A e(\zeta) = d(\zeta). \tag{2.17}$$

   Die $A(\theta)$-Stabilität bedeutet in diesem Fall, daß $|\arg \delta(\zeta)| \leq \pi - \theta$ für $|\zeta| \leq 1$ ist.

   Mit Hilfssatz 2.15 gibt es für $|\zeta| \leq 1$ genau eine Lösung $e(\zeta) \in V$ von (2.17) mit

   $$e(\zeta) = \left(\frac{\delta(\zeta)}{\tau} + A \right)^{-1} d(\zeta),$$
   $$\|e(\zeta)\| \leq \frac{1}{\alpha \sin(\theta)} \|d(\zeta)\|_{V'}.$$

   Die Funktion $\zeta \mapsto e(\zeta)$ ist analytisch in $|\zeta| < 1$ und stetig auf $|\zeta| \leq 1$ (Übung!).

   Die Fourierkoeffizienten von $\hat{e}(\varphi) = e(e^{\imath\varphi})$ sind $(0, \dots, 0, e_k, e_{k+1}, \dots)$, während die Fourier-Koeffizienten von $\hat{d}(\varphi) = d(e^{\imath\varphi})$ die Folge $(0, \dots, 0, d_k, d_{k+1}, \dots)$ ist. Mit zweifacher Anwendung der Parseval-Gleichung gilt damit

   $$\sum_{j=k}^{\infty} \|e_j\|^2 = \frac{1}{2\pi} \int_0^{2\pi} \|\hat{e}(\varphi)\|^2 \, d\varphi \leq \frac{1}{2\pi} \int_0^{2\pi} \|\hat{d}(\varphi)\|_{V'}^2 \, d\varphi \frac{1}{\alpha^2 \sin^2 \theta} = \frac{1}{\alpha^2 \sin^2 \theta} \sum_{j=k}^{N} \|d_j\|_{V'}^2.$$

Da die $e_0, \ldots, e_{k+1}, \ldots, e_n$ nicht von $d_{n+1}, d_{n+2}, \ldots$ abhängen, können wir diese auf der rechten Seite weglassen und erhalten (durch Multiplikation mit τ)

$$\tau \sum_{j=k}^{N_n} \|e_j\|^2 \leq \frac{1}{\alpha^2 \sin^2 \theta} \tau \sum_{j=k}^{n} \|d_k\|_{V'}^2.$$

2. Seien $e_0, \ldots, e_{k-1} \in V$ beliebig und $d_n = 0$ für alle n. Wir erhalten wie oben die Gleichung

$$\frac{1}{\tau}\delta(\zeta)e(\zeta) + Ae(\zeta) = \frac{1}{\tau}e^0(\zeta), \tag{2.18}$$

wobei e_j^0 eine Linearkombination von $e_0, \ldots, e_{k-1}$ ist und

$$e^0(\zeta) = \sum_{j=0}^{2k-1} e_j^0 \zeta^j, \text{ und es gilt } \|e_j^0\| \leq C \sum_{i=0}^{k-1} \|e_i\|.$$

Wie oben gibt es eine eindeutige Lösung von (2.18)

$$e(\zeta) = \left(\frac{\delta(\zeta)}{\tau} + A\right)^{-1} \frac{1}{\tau} e^0(\zeta). \tag{2.19}$$

Für $|\arg(\lambda)| \leq \pi - \theta$ und $v \in V$ erhalten wir

$$\|(\lambda + A)^{-1} v\| \leq \frac{C}{|\lambda|}\|v\|, \tag{2.20}$$

denn wir rechnen für ein solches λ und v

$$\begin{aligned}\|\lambda(\lambda + A)^{-1}v\| &= \|(\lambda + A)(\lambda + A)^{-1}v - A(\lambda + A)^{-1}v\| \\ &\leq \|v\| + \|(\lambda + A)^{-1}Av\| \leq \|v\| + \frac{1}{\alpha \sin\theta}\|Av\|_{V'} \\ &\leq \|v\| + \frac{1}{\sin\theta} C_A \|v\| = \left(1 + \frac{C_A}{\sin\theta}\right)\|v\|,\end{aligned}$$

wobei wir im letzten Schritt die Stetigkeit von $A : V \to V'$ ausgenutzt haben. Damit gilt (2.20) und aus dieser Gleichung erhalten wir, angewandt auf (2.19) gilt

$$\|e(\zeta)\| \leq C \frac{\tau}{|\delta(\zeta)|} \frac{1}{\tau} \|e^0(\zeta)\| = C \Big\| \underbrace{\frac{e^0(\zeta)}{\delta(\zeta)}}_{=:g(\zeta)} \Big\|. \tag{2.21}$$

Da 1 die einzige Nullstelle von $\delta(\zeta)$ vom Betrag ≤ 1 ist, können wir $\delta(\zeta) = (1 - \zeta)\mu(\zeta)$ schreiben mit $\mu(\zeta) \neq 0$ für $|\zeta| \leq 1$. Damit gilt

$$\frac{e^0(\zeta)}{\delta(\zeta)} = \frac{e^0(\zeta)}{1 - \zeta} \frac{1}{\mu(\zeta)}.$$

Wir rechnen nun mit dem Wissen, daß $e_j = 0$ für $j > 2k - 1$ ist,

$$\sum_{n=0}^{\infty} \zeta^n e^0(\zeta) = \frac{1}{1 - \zeta} e^0(\zeta) = e_0^0 + (e_0^0 + e_1^0)\zeta + \ldots + (e_0^0 + \ldots + e_{2k-1}^0)\zeta^n + \ldots,$$

also gilt

$$g(\zeta) = \sum_{n=0}^{\infty} g_n \zeta^n \text{ mit } \|g_n\| \leq C(\|e_0\| + \ldots + \|e_{k-1}\|).$$

Da die $e_k, \ldots, e_n$ wie oben unabhängig von $g_{n+1}, g_{n+2}, \ldots$ sind, setzen wir $g_j = 0$ für $j > n$ und betrachten stattdessen

$$\tilde{g}(\zeta) = \sum_{j=0}^{n} g_j \zeta^j \text{ sowie } \tilde{e}(\zeta) = \left(\frac{\delta(\zeta)}{\tau} + A\right)^{-1} \frac{\delta(\zeta)}{\tau} \tilde{g}(\zeta).$$

Wie im ersten Teil erhalten wir mit diesen neuen Definitionen, (2.21) sowie zweifacher Anwendung der Parseval-Gleichung

$$\sum_{j=0}^{n} \underbrace{\|e_j\|^2}_{=\|\tilde{e}_j\|^2} = \frac{1}{2\pi} \int_0^{2\pi} \|\tilde{e}(e^{\imath\varphi})\|^2 \, d\varphi \leq C \frac{1}{2\pi} \int_0^{2\pi} \|\tilde{g}(e^{\imath\varphi})\|^2 \, d\varphi = C \sum_{j=0}^{n} \|g_j\|^2.$$

Also gilt

$$\tau \sum_{j=k}^{n} \|e_j\|^2 \leq C \underbrace{n\tau}_{=t_n} \underbrace{(\|e_0\| + \ldots + \|e_{k-1}\|)^2}_{\leq \sum_{i=0}^{k-1} \|e_i\|^2}$$

und die Behauptung folgt aus der Linearität der Gleichung (2.15). □

2.36 Satz
Das parabolische kontinuierliche Problem habe genügend oft differenzierbare Lösungen $u : [0, T] \to V$ und für die Startwerte gelte

$$\|u_j - u(t_j)\| \leq C_0 \tau^k \quad \forall j \leq k - 1.$$

Dann gilt für den Fehler der BDF-Semidiskretisierung (für $k \leq 6$)

$$|u_n - u(t_n)| + \left(\tau \sum_{j=0}^{n} \|u_j - u(t_j)\|^2\right)^{\frac{1}{2}} \leq C\tau^k \text{ für } n\tau \leq T.$$

Beweis (nur in der $\|.\|$-Norm auf V)
Der Fehler in der $|.|$-Norm auf H ist eine Übungsaufgabe. Das Umschreiben von (2.14) ergibt

$$\frac{1}{\tau} \sum_{j=0}^{k} \delta_j u(t_{n-j}) + Au(t_n) = f(t_n) + \frac{1}{\tau} \sum_{j=0}^{k} \delta_j u(t_{n-j}) - \partial_t u(t_n) =: f(t_n) + d_n,$$

was für $e_n := u_n - u(t_n)$ sich zu

$$\frac{1}{\tau}\sum_{j=0}^{k}\delta_j e_{n-j} + Ae_n = d_n.$$

ergibt. Wir erfüllen damit die Voraussetzungen an die Rekursion aus Hilfssatz 2.35. Mit unseren Voraussetzungen sehen wir durch direkte Rechnung, daß $\|d_n\|_{V'} \leq C_0\tau^k$ ist. Damit erhalten wir für $t_n \leq T$

$$\left(\tau\sum_{j=k}^{n}\|e_j\|^2\right)^{\frac{1}{2}} \leq C\tau^k.$$

□

2.6 Runge–Kutta-Zeitdiskretisierung einer nichtlinearen parabolsichen DGL

Die bisher entwickelte Theorie hat sich auf lineare parabolische Differentialgleichungen beschränkt. Wir betrachten im Folgenden nichtlineare Gleichungen und ein Verfahren hierfür.

2.37 Beispiel
Wir suchen $u : \Omega \times [0,T] \to \mathbb{R}$ mit

$$\begin{cases} \partial_t u(x,t) = \sum_{i,j=1}^{d} \partial_{x_i}\Big(a_{ij}(u(x,t))\partial_{x_j}u(x,t)\Big) + f(x,t), & x \in \Omega, \\ u = 0 & \text{auf } \partial\Omega \times (0,T), \\ u(.,t=0) = u_0 & \text{in } \Omega. \end{cases} \tag{2.22}$$

2.38 Konstruktion (schwache Formulierung und weitere Voraussetzungen)
Wir formulieren diese nichtlineare parabolische Differentialgleichung schwach auf $H := L^2(\Omega)$ und $V := H_0^1(\Omega)$ mit $A(u) : V \to V'$, welches durch

$$\langle A(u)v, w\rangle := \sum_{i,j=1}^{d}\int_\Omega a_{ij}(u(x,t))\partial_{x_j}v\partial_{x_i}w\,dx$$

definiert ist. Im abstrakten Raum ist (2.22) äquivalent zu

$$\partial_t u(t) + A(u)u = f(t). \tag{2.23}$$

Wir fordern dabei, daß es M und α unabhängig von u gibt mit

$$\begin{aligned} \langle A(u)v, v\rangle &\geq \alpha\|v\|^2 && \forall u, v \in V, \\ |\langle A(u)v, w\rangle| &\leq M\|v\|\|w\| && \forall u, v, w \in V. \end{aligned}$$

Weiter gebe es $S \subseteq V$, so daß für alle $\delta > 0$ ein $L = L(\delta, S)$ gibt mit

$$\|(A(v) - A(w))u\|_{V'} \leq \delta\|v - w\| + L|v - w| \quad \forall u \in S,\ \forall v, w \in V.$$

2.39 Bemerkung
Diese Eigenschaft ist im Beispiel etwa erfüllt, wenn a_{ij} Lipschitz-stetig ist (mit Konstante L_0) und wir

$$S := S(r) := \left\{ u \in H_0^1(\Omega) : \sup_{x\in\Omega} |\nabla u|(x)| \le r \right\}$$

wählen, denn es gilt mit der Cauchy–Schwarz-Ungleichung für $u \in S$

$$\begin{aligned} |\langle (A(v) - A(w))u, \varphi\rangle| &= \left| \int_\Omega \sum_{i,j=1}^d (a_{ij}(v(x)) - a_{ij}(w(x)))\partial_{x_j} u \partial_{x_i}\varphi \, dx \right| \\ &\le \int_\Omega \sum_{i,j=1}^d L_0 |v(x) - w(x)| r |\partial_{x_i}\varphi| \, dx \\ &\le L_0 r \|v - w\|_{L^2(\Omega)} \|\varphi\|_{H_0^1(\Omega)}, \end{aligned}$$

also gilt

$$\|(A(v) - A(w))u\|_{V'} \le Lr\|v - w\|_{L^2(\Omega)}.$$

2.40 Konstruktion (Zeitdiskretisierung mit Radau-Kollokationsverfahren)
Mit einem impliziten Runge–Kutta-Verfahren lösen wir mit der Zeitschrittweite $h > 0$

$$\begin{aligned} u_{n+1} &= u_n + h \sum_{j=1}^s b_j U'_{nj}, \\ U_{ni} &= u_n + h \sum_{j=1}^s a_{ij} U'_{nj}, \\ U'_{ni} + A(U_{ni})U_{ni} &= f(t_n + c_i h). \end{aligned}$$

Wir setzen voraus, daß die Ordnung des Verfahrens p ist und die Stufenordnung q, i.e.

$$\sum_{j=1}^s a_{ij} c_j^{k-1} = \frac{c_i^k}{k} \quad \forall k = 1, \dots, q \; \forall i = 1, \dots s,$$

was bei Kollokationsverfahren $q = s$ bedeutet. Weiter sei das Verfahren kontraktiv, insbesondere $b_i > 0$ und $(b_i a_{ij} + b_j a_{ji} - b_i b_j)$ ist positiv semidefinit und es gelte $|R(\infty)| < 1$.

Für das Radau-Kollokationsverfahren sind alle Voraussetzungen erfüllt. Es gilt $p = 2s - 1$, $q = s$ und $R(\infty) = 0$. Für $s > 1$ gilt dabei $p \ge q + 1$.

2.41 Satz (Konvergenz)
Die exakte Lösung $u : [0,T] \to V$ von (2.23) existiere, sei hinreichend oft differenzierbar und erfülle $u(t) \in S$ für alle $t \in [0,T]$. Dann gilt unter den obigen Voraussetzungen für $h \le h_0$ (wobei $h_0 = h_0(\alpha, M, L)$ eine feste Konstante ist) und $Nh \le T$, daß

$$\left(h \sum_{n=0}^{N} \|u_n - u(t_n)\|^2 \right)^{\frac{1}{2}} + \max_{0 \le n \le N} |u_n - u(t_n)| \le C h^{q+1}$$

ist, wobei C von α, M, L, T sowie den Ableitungen von u abhängt.

Beweis

1. Wir bezeichnen die exakten Lösungswerte als

$$\tilde{U}_{ni} = u(t_n + c_i h), \qquad \tilde{U}'_{ni} = u'(t_n + c_i h), \qquad \tilde{u}_n = u(t_n)$$

und setzen diese in das Runge–Kutta-Verfahren ein und erhalten

$$\begin{aligned} \tilde{u}_{n+1} &= \tilde{u}_n + h \sum_{i=1}^{s} b_i \tilde{U}'_{ni} + d_{n+1}, \\ \tilde{U}_{ni} &= \tilde{u}_n + h \sum_{j=1}^{s} a_{ij} \tilde{U}'_{nj} + D_{ni}, \end{aligned}$$

wobei d_{n+1} und D_{ni} die Quadraturformel-Fehler sind mit

$$\|D_{ni}\| = \mathcal{O}(h^{p+1}), \qquad \|d_{n+1}\| = \mathcal{O}(h^{p+1}) = \mathcal{O}(h^{q+2}),$$

wobei wir $p \ge q+1$ benutzt haben. Damit gilt insgesamt

$$B := h \sum_{n=0}^{N} \sum_{j=1}^{s} \|D_{nj}\|^2 + h \sum_{n=0}^{N} \left(\|d_{n+1}\|^2 + \left\| \frac{d_{n+1}}{h} \right\|_{V'}^2 \right) \le C (h^{q+1})^2. \tag{2.24}$$

2. Die Fehler $e_n := u_n - \tilde{u}_n$, $E_{ni} := U_{ni} - \tilde{U}_{ni}$, $E'_{ni} := U'_{ni} - \tilde{U}_{ni}$ erfüllen aufgrund des Runge–Kutta-Verfahrens bzw. der parabolischen Gleichung

$$\begin{cases} E'_{ni} + A(U_{ni}) E_{ni} = -(A(U_{ni}) - A(\tilde{U}_{ni}))\tilde{U}_{ni}, \\ E_{ni} = e_n + h \sum_{j=1}^{s} a_{ij} E'_{nj} - D_{ni}, \\ e_{n+1} = e_n + h \sum_{i=1}^{s} b_i E'_{ni} - d_{n+1}. \end{cases} \tag{2.25}$$

Die letzten Gleichungen von (2.25) in der $|.|$-Norm ergibt

$$|e_{n+1}|^2 = \left| e_n + h \sum_{i=1}^{s} b_i E'_{ni} \right|^2 - \left\langle d_{n+1}, e_n + h \sum_{i=1}^{s} b_i E'_{ni} \right\rangle + |d_{n+1}|^2 \tag{2.26}$$

Wir schätzen nun die Terme von (2.26) getrennt ab. Für den ersten Term ergibt sich

$$\begin{aligned}\left|e_n + h\sum_{i=1}^{s} b_i E'_{ni}\right|^2 &= |e_n|^2 + 2h\sum_{i=1}^{s} b_i\langle E'_{ni}, E_{ni} + D_{ni}\rangle \\ &\quad + \underbrace{h^2\sum_{i=1}^{s}\sum_{j=1}^{s} \underbrace{b_i b_j - b_i a_{ij} - b_j a_{ji}}\langle E'_{ni}, E'_{nj}\rangle}_{\leq 0},\end{aligned} \tag{2.27}$$

wobei wir die Definition der E_{nj}, E'_{nj}, etc. aus (2.25) benutzt haben und der hervorgehobene Ausdruck nach Voraussetzung negativ semidefinit ist, cf. Satz 1.36. Beim mittleren Term in (2.27) lassen wir die Indices n und i weg und es gilt mit (2.25)

$$\langle E', E + D\rangle = -\langle A(U)E, E\rangle - \langle A(U)E, D\rangle - \Big\langle (A(U) - A(\tilde{U}))\tilde{U}, E + D\Big\rangle.$$

Unter den gegebenen Voraussetzungen und $\tilde{U} \in S$ gilt damit für ein $\delta > 0$

$$\leq -\alpha\|E\|^2 + M\|E\|\|D\| + \Big(\delta\|E\| + L|E|\Big)\|E + D\|,$$

womit dann schließlich für hinreichend kleines δ gilt mit der Cauchy-Ungleichung (mit ε) und der Tatsache, daß die $|.|$-Norm schwächer als die $\|.\|$-Norm ist, daß

$$\leq -\frac{\alpha}{2}\|E\|^2 + C|E|^2 + C\|D\|^2.$$

Für den zweiten Term in (2.26) erhalten wir mit der Cauchy-Ungleichung (mit ε)

$$\begin{aligned}\left\langle d_{n+1}, e_n + h\sum_{j=1}^{s} b_j E'_{nj}\right\rangle &\leq \|d_{n+1}\|_{V'} Verte_n\| + \|d_{n+1}\| h\sum_{j=1}^{s} b_j\|E'_{nj}\|_{V'} \\ &\leq \frac{h\delta}{2}\|e_n\|^2 + \frac{1}{2}\frac{h}{\delta}\left\|\frac{d_{n+1}}{h}\right\|_{V'}^2 \\ &\quad + Ch\delta\sum_{j=1}^{s}\|E'_{nj}\|_{V'}^2 + C\frac{h}{\delta}\|d_{n+1}\|^2.\end{aligned}$$

Mit der Definition von E'_{nj} aus (2.25) sowie den Voraussetzungen an A gilt $\|E'_{nj}\|_{V'}^2 \leq C\|E_{nj}\|^2$. Also gilt nun

$$\begin{aligned}&\leq \frac{h\delta}{2}\|e_n\|^2 + \frac{1}{2}\frac{h}{\delta}\left\|\frac{d_{n+1}}{h}\right\|_{V'}^2 \\ &\quad + Ch\delta\sum_{i=1}^{s}\|E_{ni}\|^2 + C\frac{h}{\delta}\|d_{n+1}\|^2.\end{aligned}$$

Für den letzten Term in (2.26) gilt mit der Cauchy-Ungleichung (mit ε)

$$|d_{n+1}|^2 = \langle d_{n+1}, d_{n+1}\rangle \leq \|d_{n+1}\|_{V'}\|d_{n+1}\| \leq \frac{1}{2}h\|d_{n+1}\|^2 + \frac{1}{2}h\left\|\frac{d_{n+1}}{h}\right\|_{V'}^2.$$

Mit der bisher geleisteten Vorarbeit gilt nun (mit geeigneter Wahl von δ)

$$|e_{n+1}|^2 - |e_n|^2 + \frac{\alpha}{4} h \sum_{i=1}^{s} b_i \|E_{ni}\|^2 \\ \leq Ch \left(\delta \|e_n\|^2 + \sum_{i=1}^{s} |E_{ni}|^2 + \sum_{i=1}^{s} \|D_{ni}\|^2 + \|d_{n+1}\|^2 + \left\| \frac{d_{n+1}}{h} \right\|_{V'}^2 \right). \quad (2.28)$$

3. Aus (2.25) erhalten wir

$$e_{n+1} = \underbrace{(1 - b^T \mathfrak{A}^{-1} \mathbb{1})}_{=R(\infty)} e_n + b^T \mathfrak{A}^{-1} (E_n + D_n) d_{n+1},$$

womit wir wegen $|R(\infty)| < 1$ erhalten, daß

$$h \sum_{n=0}^{N} \|e_{n+1}\|^2 \leq Ch \sum_{n=0}^{N} \sum_{i=1}^{s} \|E_{ni} + D_{ni}\|^2 + Ch \sum_{n=0}^{N} \|d_{n+1}\|^2. \quad (2.29)$$

4. Wir summieren (2.28) von 0 bis N auf und erhalten mit (2.29) durch Absorption der Terme $\|E_{ni}\|$ auf die linke Seite

$$|e_{N+1}|^2 + h \sum_{n=0}^{N} \sum_{i=1}^{s} \|E_{ni}\|^2 \leq Ch \sum_{n=0}^{N} \sum_{i=1}^{s} |E_{ni}|^2 + CB \\ \leq Ch \sum_{n=0}^{N} \sum_{i=1}^{s} \left(\frac{\delta}{2} \|E_{ni}\|^2 + \frac{1}{2\delta} \|E_{ni}\|_{V'}^2 \right) + CB. \quad (2.30)$$

Mit (2.25) gilt

$$\|E_{ni}\|_{V'} \leq C|e_n| + Ch \sum_{j=1}^{s} \|E_{ni}\| + \|D_{ni}\|.$$

Damit gilt aus (2.30)

$$h \sum_{n=0}^{N} \sum_{i=1}^{s} |E_{ni}|^2 \leq \left(\frac{\delta}{2} + C \frac{h}{\delta} \right) h \sum_{n=0}^{N} \sum_{j=1}^{s} \|E_{nj}\|^2 + Ch \sum_{n=0}^{N} |e_n|^2 + CB.$$

Für ein geeignetes $\delta > 0$ gilt mit (2.30) für $0 \leq Nh \leq T$

$$|e_{N+1}| \leq Ch \sum_{n=0}^{N} |e_n|^2 + CB$$

erhalten und mit der diskreten Gronwall-Ungleichung (Übung!) gilt $|e_n|^2 \leq CB$. Durch die Abschätzung von B in (2.24) und (2.30) folgt mit (2.29) die Behauptung. □

2.7 Schnelle Runge–Kutta-Approximation für inhomogene lineare parabolische Anfangsrandwertprobleme

2.42 Bemerkung
Aus der parabolischen Differentialgleichung erhalten wir ein großes System von gewöhnlichen Differentialgleichungen $Mu'(t) + Au(t) = g(t)$, wobei die Matrizen M und A symmetrisch und positiv definit und sehr groß sind. Wir setzen im Folgenden $M = \mathrm{id}$ (was etwa bei der Wärmeleitungsgleichung tatsächlich der Fall ist; das Folgende läßt sich leicht auch auf den allgemeinen Fall übertragen).

Mit der Zeitdiskretisierung (etwa dem impliziten Euler-Verfahren) mußten wir $(I+\tau A)u_n = u_{n-1} + \tau g(t_n)$ für $n = 1, \dots, N$ berechnen. Hierfür müssen wir N lineare Gleichungssysteme lösen (und sogar sN Stück für s-stufige impliziten Runge–Kutta-Verfahren).

Wir zeigen, daß bei einer vorgegebenen Genauigkeit von ε sogar $\mathcal{O}\left(\log N \log \frac{1}{\varepsilon}\right)$ Operationen genügen. Für $N = 10^5$ und $\varepsilon = 10^{-5}$ müßten wir in diesem Fall 100 lineare Gleichungssysteme lösen (statt etwa 300 000 für das Radau-Verfahren mit $s = 3$ Stufen). Hierfür stellen wir die wesentlichen Ideen und den Algorithmus vor, u_N zu berechnen – allerdings ohne die vorherigen Zeitschritte $u_1, \dots, u_{N-1}$ zu berechnen. Wir beachten, daß der Algorithmus auf lineare Probleme beschränkt ist.

Wir fordern für den Rest dieses Abschnittes, daß A sektoriell ist, i.e. für $|\arg \lambda| \le \pi - \theta$ mit $\theta < \frac{\pi}{2}$ gilt für ein $\sigma \in \mathbb{C}$

$$\|(\lambda I + A)^{-1}\| \le \frac{M}{|\lambda - \sigma|}.$$

Falls A symmetrisch und positiv definit ist, ist auch A sektoriell. Die Bedingung ist dabei für $M = \frac{1}{\sin\theta}$ und $\sigma = 0$ erfüllt.

2.43 Idee (Zutaten für einen schnellen Algorithmus)
Die wesentlichen Ideen für den schnelle Algorithmus wie oben besprochen sind die folgenden.

1. *Diskrete Variation der Konstanten-Formel:* Im kontinuierlichen Fall haben wir bei der inhomogenen Gleichung eine Variation der Konstanten-Formel,

   $$u(t) = e^{-tA}u_0 + \int_0^t e^{-(t-s)A} g(s)\, ds,$$

   erhalten, zu der wir ein Analogon im Diskreten suchen. Für ein implizites A-stabiles Runge–Kutta-Verfahren (e.g. Radau) betrachten wir wie in Abschnitt 1.5 für $\mathfrak{A} := (a_{ij})$ und $b^T := (b_1, \dots, b_s)$ die Funktionen

   $$R(z) := 1 + zb^T(I - z\mathfrak{A})\mathbb{1}, \qquad Q(z) := (Q_1(z), \dots, Q_s(z)) := b^T(I - z\mathfrak{A})^{-1},$$

   womit sich das Radau-Verfahren als

   $$u_{n+1} = R(-\tau A)u_n + \tau \sum_{i=1}^{s} Q_i(-\tau A) g(t_n + c_i\tau)$$

schreiben läßt. Durch direktes Auflösen der Rekursion ergibt sich

$$u_N = R(-\tau A)^N u_0 + \tau \sum_{j=0}^{N-1} R(-\tau A)^{N-j-1} Q(-\tau A) g_j \text{ mit } g_j := \begin{pmatrix} g(t_j + c_1 h) \\ \dots \\ g(t_j + c_j h) \end{pmatrix}.$$

2. *Kurvenintegrale und ihre Approximation:* Mit der Cauchy-Integralformel gilt

$$R(-\tau A)^n Q(-\tau A) = \frac{1}{2\pi\imath} \int_\Gamma (\lambda I + A)^{-1} R(\tau\lambda)^n Q(\tau\lambda)\, d\lambda, \tag{2.31}$$

wobei Γ eine Hyperbel in den Sektor einschließt. Wir möchten das Integral dabei durch die Trapezregel auf Γ approximieren – wir beachten, daß $R(\tau\lambda)$ im Unendlichen abfällt und damit das Integral abgeschnitten werden kann –, was sich aber global dadurch schwierig gestaltet, da diese Approximation nicht gleichmäßig in n ist.

Wir approximieren daher nur auf Teilintervallen $I_1, \dots, I_L$, etc. mit $I_l := [B^{l-1}\tau, B^l\tau]$ und $B \in \mathbb{N}$. Für $B^L\tau > T$ benötigen wir hierfür also $L = \mathcal{O}(\log_B N)$ Intervalle.

Für $n\tau \in I_l$ parametrisieren wir die Hyperbel Γ_l dann durch

$$\gamma_l(\varphi) = \mu_l(1 - \sin(\beta + \imath\varphi)) + \sigma,$$

wobei $\beta > 0$ und μ_l noch zu wählen sind und sich aus technischen Gründen

$$\mu_l \approx \frac{1}{B^l\tau} \log \frac{1}{\varepsilon}$$

anbietet. Die Approximation von 2.31 mit der Trapezregel auf diesen Hyperbeln ergibt mit K Quadraturpunkten und dem Abschneiden zur Schrittweite δ mit $\frac{1}{\delta} \approx \log \frac{1}{\varepsilon}$.

$$R(-\tau A)^n Q(-\tau A) \approx \sum_{k=-K}^{K} w_k^{(l)} (\lambda_k^{(l)} I + A)^{-1} R(\tau\lambda_k^{(l)})^n Q(\tau\lambda_k^{(l)}),$$

mit

$$\lambda_k^{(l)} := \gamma_l(k\delta), \qquad w_k^{(l)} = -\frac{\delta}{2\pi\imath} \gamma_l'(k\delta).$$

2.44 Satz
Falls $K \approx \log \frac{1}{\varepsilon}$ und $n \geq C \log \frac{1}{\varepsilon}$ ist, ist der Quadraturfehler $\leq \varepsilon$.

Dabei ist K unabhängig von l, n, τ mit $n\tau \leq T$. Weiter hängt K nur in Form von θ, M und σ von A ab.

2.45 Algorithmus (Schneller Algorithmus zur Berechnung)
Sei $u_0 = 0$. Wir zerlegen $u_N = u_N^{(0)} + \dots + u_N^{(L)}$ mit $N \leq B^L$ und

$$u_N^{(0)} = \tau Q(-\tau A) g_{N-1},$$
$$u_N^{(l)} = \tau \sum_{(n-j-1)\tau \in I_l} R(-\tau A)^{N-1-j} Q(-\tau A) g_j,$$

was für $n_l = N - B^l$ und $n_1 = 0$ sich zu

$$\begin{aligned}
&= \tau \sum_{j=n_l}^{n_{l-1}-1} \frac{1}{2\pi\imath} \int_{\Gamma_l} (\lambda I + A)^{-1} R(\tau A)^{N-1-j} Q(\tau\lambda) g_j \, d\lambda \\
&\approx \tau \sum_{j=n_l}^{n_{l-1}-1} \sum_{k=-K}^{K} w_k^{(l)} (\lambda_k^{(l)} I + A)^{-1} R(\tau\lambda_k^{(l)})^{N-1-j} Q(\tau\lambda_k^{(l)}) g_j \\
&= \sum_{k=-K}^{K} \underbrace{w_k^{(l)} R(\tau\lambda_k^{(l)})^{N-n_{l-1}}}_{=:c_k^{(l)}} (\lambda_k^{(l)} I + A)^{-1} \underbrace{\tau \sum_{j=n_l}^{n_{l-1}-1} R(\tau\lambda_k^{(l)})^{n_{l-1}-j-1} Q(\tau\lambda_k^{(l)}) g_j}_{=:y_k^{(l)}}
\end{aligned}$$

ergibt. Wir gehen nun folgendermaßen vor.

1. Wir lösen mit dem Runge–Kutta-Verfahren zur Schrittweite τ das skalaren linearen Anfangswertproblem

   $$\partial_t y_k^{(l)}(t) = \lambda_k^{(l)} y_k^{(l)} + g(t) \text{ für } t \in [n_l\tau, n_{l-1}\tau]$$

   zum Startwert $y(n_l\tau) = 0$ für $k = -K, \ldots, K$, $K \approx \log \frac{1}{\varepsilon}$ und $l = 2, \ldots, L \leq \log_B N + 1$.

2. Wir lösen die $L(K+1)$ linearen Gleichungssysteme

   $$(\lambda_k^{(l)} I + A) x_k^{(l)} = y_k^{(l)},$$

   die sich in $\mathcal{O}\left(\log n \log \frac{1}{\varepsilon}\right)$ berechnen lassen.

3. Wir berechnen für $l = 2, \ldots, L$ die Linearkombination

   $$u_N^{(l)} = \sum_{k=-K}^{K} c_k^{(l)} x_k^{(l)}.$$

4. Wir berechnen $u_N^{(0)} + u_N^{(1)} =: v(N\tau)$ durch das Lösen von sB linearen Gleichungssystem mit dem Runge–Kutta-Verfahren für das Anfangswertproblem

   $$\begin{cases} \partial_t v + Av = g(t), \\ v(N\tau - B\tau) = 0. \end{cases}$$

   Dies benötigt in der Theorie $\mathcal{O}\left(\log \frac{1}{\varepsilon}\right)$ direkte Schritte.

5. Wir berechnen $u_N := u_N^{(0)} + u_N^{(1)} + \ldots + u_N^{(l)}$.

3 Hyperbolische Differentialgleichungen

Wir betrachten im Folgenden hyperbolische Differentialgleichungen, deren Archetyp die Wellengleichung ist. Für die Analyse einer solchen Gleichung müssen wir anders vorgehen als bei elliptischen oder parabolischen Differentialgleichungen. Bei elliptischen und parabolischen Gleichungen hat die Raumdimension keine Rolle gespielt, während dies bei hyperbolischen deutlich komplizierter wird. Wir werden uns in diesem Kapitel auf eine Raumdimension beschränken.

3.1 Die Wellengleichung

3.1 Beispiel (Wellengleichung)
Wir betrachten eine schwingende Seite, die am Ort $x \in [0, \pi]$ zur Zeit t um $u(x,t)$ ausgelenkt ist. In der Physik wird hergeleitet, daß eine solche Situation durch die Wellengleichung,

$$\begin{cases} \partial_{tt} u = c^2 \partial_{xx} u, & \\ u(x,0) = u_0(x) & 0 \le x \le \pi, \\ \partial_t u(x,0) = v_0(x) & 0 \le x \le \pi, \\ u(0,t) = u(\pi,t) = 0 & \forall t > 0, \end{cases}$$

beschrieben wird. Die Variable c ist dabei die Wellengeschwindigkeit, was im Folgenden klar wird. Auch in höheren Dimensionen gibt es Modelle, Wellenausbreitungen zu beschreiben.

3.2 Konstruktion (Lösungskonstruktion mit Fourierreihen)
Falls eine klassische Lösung der Wellengleichung existiert, so setzen wir diese ungerade auf $[-\pi, 0]$ fort und betrachten die Fourierkoeffizienten

$$\hat{u}_k(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} u(x,t) e^{-\imath kx} \, dx$$

und für absolut summierbare Fourierkoeffizienten die Fourierreihe

$$u(x,t) = \sum_{k=-\infty}^{\infty} \hat{u}_k(t) e^{\imath kx}.$$

Mit der Rechenregel $(\widehat{\partial_{xx}u})_k = -k^2\hat{u}_k$ erhalten wir für $k \in \mathbb{Z}$

$$\partial_{tt}\hat{u}_k = c^2(\imath k)^2\hat{u}_k = -c^2k^2\hat{u}_k,$$

was durch $\hat{u}_k = \alpha_k e^{\imath kct} + \beta_k e^{-\imath kct}$ gelöst wird, i.e. für u ergibt sich damit

$$u(x,t) = \sum_{k=-\infty}^{\infty} \hat{u}_k(t)e^{\imath kx} = \sum_{k=-\infty}^{\infty} \alpha_k e^{\imath k(x+ct)} + \sum_{k=-\infty}^{\infty} \beta_k e^{\imath k(x-ct)}.$$

Wir sehen, daß die erste Summe auf $x + ct = \text{const}$ konstant ist und die zweite Summe ist auf $x - ct = \text{const}$ ebenfalls konstant.

Die Koeffizienten α_k und β_k berechnen sich aus den Anfangsbedingungen. Für $t = 0$ gilt

$$\sum_{k=-\infty}^{k} (\alpha_k + \beta_k)e^{\imath kx} = u_0(x) = \sum_{k=-\infty}^{\infty} \mu_k e^{\imath kx},$$
$$\sum_{k=-\infty}^{\infty} (\alpha_k - \beta_k)\imath cke^{\imath kx} = v_0(x) = \sum_{k=-\infty}^{\infty} \nu_k e^{\imath kx},$$

wobei wir u_0 bzw. v_0 ebenfalls in einer Fourierreihe geschrieben haben. Damit müssen wir

$$\alpha_k + \beta_k = \mu_k, \qquad \imath kc(\alpha_k - \beta_k) = \nu_k$$

lösen und für $k = 0$ gilt, da u ungerade ist, $\alpha_k = \beta_k = 0$.

Wir prüfen nun nach, ob die Randbedingung erfüllt ist. Diese ist erfüllt, wenn $\hat{u}_k(t) = -\hat{u}_{-k}(t)$ gilt, was der Fall für $\mu_k = -\mu_{-k}$ und $\nu_k = -\nu_{-k}$ ist.

Falls u_0 und v_0 reell sind, so gilt $\mu_k = \overline{\mu_{-k}}$ und $\nu_k = \overline{\nu_{-k}}$. Damit gilt dann bereits $\alpha_k = \overline{\alpha_{-k}}$ und $\beta_k = \overline{\beta_{-k}}$, also ist dann $u(x,t) \in \mathbb{R}$.

Falls α_k und β_k genügend gut abfallen (aufgrund der obigen Relation also genau so gut wie μ_k und ν_k), e.g. $\mu_k, \nu_k = \mathcal{O}(|k|^{-4})$, dann konvergieren die Fourierreihen für u, $\partial_{xx}u$ und $\partial_{tt}u$ absolut und wir haben eine klassische Lösung $u \in \mathcal{C}^2([0,\pi] \times [0,T])$. Sonst gibt es eine verallgemeinerte Lösung, solange die Fourierreihe für u konvergiert.

Hiermit haben wir die Existenz einer Lösung (unter den obigen Voraussetzungen) bewiesen und aus der Eindeutigkeit der Fourierkoeffizienten ist diese Lösung auch eindeutig.

3.3 Bemerkung (Energieerhaltung)
Wir definieren die Energie E (die aus kinetischer und potentieller Energie besteht) durch

$$E(t) := \frac{1}{2}\int_0^{\pi} (\partial_t u(x,t))^2 + c^2(\partial_x u(x,t))^2\, dx.$$

Hierfür gilt mit Differentiation unter dem Integral und partieller Integration

$$\begin{aligned} E'(t) &= \int_0^\pi \partial_t u(x,t) \cdot \partial_{tt}(x,t) + c^2 \partial_x(x,t) \partial_{xt} u(x,t)\, dx \\ &= \int_0^\pi \partial_t u(x,t) \cdot (\partial_{tt} u(x,t) - c^2 \partial_{xx} u(x,t))\, dx = 0, \end{aligned}$$

wobei wir benutzt haben, daß u die Wellengleichung erfüllt. Also gilt $E(t) = \text{const}$.

3.4 Konstruktion (Approximation mit finiten Differenzen, Leapfrog-Schema) Wir betrachten im Ort eine Gitterweite h und eine Zeitschrittweite τ und approximieren $u_j^n \approx u(jh, n\tau)$. Die Wellengleichung übersetzen wir dabei in das sog. Leapfrog-Schema

$$\frac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{\tau^2} = c^2 \frac{u_{j+1}^n - 2_j^n + u_{j-1}^n}{h^2}$$

mit den Anfangs- und Randbedingungen

$$u_j^0 := u_0(jh), \quad u_j^1 = u_0(jh) + \tau v_0(jh), \quad u_0^n := u_N^n = 0,$$

i.e. wir betrachten dies wie Finite Elemente mit den beiden Komponenten (x, t).

Für $\frac{c\tau}{h} > 1$ sind numerische Experimente des Finite-Differenzen Ansatzes instabil. Wir führen nun eine Stabilitätsuntersuchung durch.

3.5 Bemerkung (Stabilitätsuntersuchung, mit diskreter Fourier-Transformation) Wir setzen die Finite-Differenzen Lösung nun ungerade auf $[-\pi, 0]$ fort, also $u_{-j}^n = -u_j^n$ und betrachten die diskrete Fourier-Transformation $\mathbb{C}^{2N} \to \mathbb{C}^{2N}$, die analog zur kontinuierlichen Fourier-Transformation durch

$$\hat{u}_k^n = \frac{1}{2N} \sum_{j=-N}^{N-1} u_j^n w^{-jk}, \qquad u_j^n = \sum_{k=-N}^{N-1} \hat{u}_k^n w^{jk},$$

definiert ist, wobei w die $2N$-te primitive komplexe Einheitswurzel ist. Dann gilt

$$u_{j+1}^n - 2u_j^n + u_{j-1}^n = \sum_{k=-N}^{N-1} \hat{u}_k^n w^{jk} \underbrace{(w^k - 2 + w^{-k})}_{=2\cos kh - 2 = -4\sin^2 \frac{kh}{2}}$$

und durch Einsetzen in das Leapfrog-Schema gilt

$$\hat{u}_k^{n+1} - 2\hat{u}_k^n + \hat{u}_k^{n-1} = -\left(\frac{c\tau}{h}\right)^2 4\sin^2\left(\frac{kh}{2}\right) \hat{u}_k^n$$

bzw.

$$\hat{u}_k^{n+1} - 2\underbrace{\left(1 - 2r^2 \sin^2\left(\frac{kh}{2}\right)\right)}_{=:a} \hat{u}_k^n + \hat{u}_k^{n-1} = 0,$$

wobei $r = \frac{c\tau}{h}$ ist. Wir erhalten also entkoppelte Differentialgleichungen mit der Rekursion $y_{n+1} - 2ay_n + y_{n-1} = 0$, wobei das charakteristische Polynom hiervon $\zeta^2 - 2a\zeta + 1$ ist und die Nullstellen $\zeta_{1,2} = a \pm \imath\sqrt{1-a^2}$ hat. Die allgemeine Lösung können wir dann also als $y_n = c_1\zeta_1^n + c_2\zeta_2^n$ schreiben und diese Rekursion ist stabil (also beschränkt für $n \to \infty$), solange $|\zeta_1|, |\zeta_2| \leq 1$ ist, also solange $|a| \leq 1$ ist.

3.6 Proposition (Stabiltätsbedingung)
Die Finite-Differenzen Lösung ist genau dann stabil, wenn

$$\left|1 - 2r^2 \sin^2 \frac{kh}{2}\right| \leq 1 \quad \forall k = -N, \ldots, N-1$$

ist. Dies ist insbesondere der Fall, wenn $\left|\frac{c\tau}{h}\right| \leq 1$ gilt.

3.7 Bemerkung
Die Zahl $r := \frac{c\tau}{h}$ wird auch Courant-Zahl genannt.

3.8 Bemerkung (Wellengleichung als System von DGL erster Ordnung)
Sei $c = 1$. Für $\partial_t u = \partial_x v$ und $\partial_t v = \partial_x u$ gilt $\partial_{tt} u = \partial_t \partial_x v = \partial_x \partial_t v = \partial_{xx} u$, also

$$\partial_t \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \partial_x \begin{pmatrix} u \\ v \end{pmatrix} == T \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} T^{-1} \partial_x \begin{pmatrix} u \\ v \end{pmatrix}$$

mit $T = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = T^{-1}$. Für $\begin{pmatrix} p \\ q \end{pmatrix} = T^{-1} \begin{pmatrix} u \\ v \end{pmatrix}$ gilt dann

$$\partial_t p = \partial_x p, \qquad \partial_t q = -\partial_x q.$$

Wir erhalten also zwei entkoppelte partielle Differentialgleichungen erster Ordnung.

3.2 Advektionsgleichungen, Charkateristiken

3.9 Definition (Advektionsgleichung)
Wir betrachten die Advektionsgleichung

$$\begin{cases} \partial_t u + a\partial_x u = 0 & x \in \mathbb{R}, t > 0, \\ u(x,0) = u^0(x) & x \in \mathbb{R}. \end{cases}$$

3.10 Konstruktion (Lösung der Advektionsgleichung, Charakterisitik)

1. Für $a = \text{const} \in \mathbb{R}$ ist die Lösung $u(x,t) = u^0(x - at)$, was dem Transport der Anfangsfunktion mit der Geschwindigkeit a entspricht.

2. Sei $a = a(x,t)$. Die *Charakteristik* ist die Lösung $x(t)$ der gewöhnlichen Differentialgleichung $\partial_t x(t) = a(x(t),t)$. Die Lösung der Advektionsgleichung ist entlang von Charakteristiken konstant, denn es gilt mit der Kettenregel

   $$\partial_t u(x(t),t) = \partial_x u \partial_t x + \partial_t u = \partial_x u a(x(t),t) + \partial_t u = 0.$$

   Damit gilt $u(x(t),t) = u^0(x(0),0)$.

Wir betrachten nun unser erstes Lösungsverfahren.

3.11 Algorithmus (Charakteristikenmethode)
Wir wählen Punkte $x_0, x_1, \ldots, x_N$ und bestimmen Charakteristiken $x_j(t)$ mit x_j^0 als Anfangswert. Dann gilt

$$u(x_j(t),t) = u^0(x_j(0)) = u^0(x_j^0).$$

3.12 Bemerkung

1. Falls $a = a(x,t)$ ist und $a \in \mathcal{C}^1$ ist, so können sich die Charakteristiken nicht kreuzen, was aus der Eindeutigkeit der Lösungen von gewöhnlichen Differentialgleichungen folgt.

2. Im nichtlinearen Fall (falls $a = a(u)$ ist) gilt für die Charakteristiken $\partial_t x = a(u(x(t),t))$, also ist die Charakteristik eine Gerade $x(t) = a(u)t + c$ und es gilt

   $$u(x,t) = u^0(x - a(u(x(t),t))).$$

   In diesem Fall ist die Charakteristik eine Schar von Geraden, die sich kreuzen können (dies führt zu so genannten Schocks). Die Lösung kann in einem solchen Fall nicht stetig sein, aber es gibt dann schwache Lösungen (die allerdings nicht eindeutig sind). In einem solchen Fall müssen wir uns über so genannte Entropiebedingung eine physikalisch sinnvoll Lösung auswählen, um zur Eindeutigkeit zu gelangen.

3.13 Bemerkung (Systeme)
Wir betrachten nun das System von gewöhnlichen Differentialgleichungen,

$$\begin{cases} \partial_t u + A\partial_x u = 0, \\ \qquad u(x,0) = u_0(x), \end{cases}$$

wobei $A \in \mathbb{R}^{n\times n}$ konstant ist und wir $u = (u_1, \ldots, u_n) \in \mathbb{R}^n$ suchen. Falls A diagonalisierbar ist, i.e. für $\Lambda = \operatorname{diag}(a_1, \ldots, a_n)$ gilt $SA = \Lambda S$, erhalten wir für $r = Su$ die entkoppelten skalaren Advektionsgleichungen

$$\partial_t r + \Lambda \partial_x r = 0.$$

Diese lösen sich dann – wie oben erwähnt – durch $r_j(x,t) = r_j^0(x - a_j t)$ und wir sehen, daß die a_j die Geschwindigkeiten sind, welche reell sein sollten.

3.14 Definition (hyperbolisches System)
Ein System

$$\begin{cases} \partial_t u + A\partial_x u = 0, \\ \qquad u(x,0) = u_0(x) \end{cases}$$

heißt *hyperbolisch*, falls A diagonalisierbar ist mit reellen Eigenwerten.

3.15 Beispiel
Die Wellengleichung hat als Differentialgleichungssystem nach Bemerkung 3.8 die Eigenwerte ± 1 und ist daher hyperbolisch.

Wir betrachten nun nichtlineare Systeme, die beispielsweise bei Erhaltungsgrößen in der Physik, etwa bei der Massen- und Impulserhaltung, auftreten.

3.16 Bemerkung
Das nichtlineare System $\partial_t u + A(u)\partial_x u = 0$ heißt hyperbolisch, falls es ein $S(u)$ und eine Diagonalmatrix $\Lambda(u)$ mit reellen Eigenwerten gibt, so daß $S(u)A(u) = \Lambda(u)S(u)$ gilt. In einem solchen Fall erhalten wir mit

$$S\partial_t u + \Lambda S\partial_x u = 0 \tag{3.1}$$

die so genannte charakteristische Normalform. Falls es eine so genannten Riemann-Invariante $r = r(u)$ mit $\partial_t r = S\partial_t u$ und $\partial_x r = S\partial_x u$ gibt (die zumindest für zweidimensionale Systeme stets existiert), dann erhalten wir wir wie im linearen Fall aus (3.1) die skalare Gleichung

$$\partial_t r + \Lambda\partial_x r = 0.$$

3.3 Differenzenverfahren für die Advektionsgleichung

3.17 Herleitung (Volldiskretisierung)
Wir betrachten die Differentialgleichung

$$\begin{cases} \partial_t u = c\partial_x u, \\ u(x,0) = u^0(x). \end{cases}$$

Dabei ist $c = -a > 0$. Wir betrachten nun das Differenzenschema

$$\frac{u(x,t+\tau) - u(x,t)}{\tau} = c\frac{u(x+h,t) - u(x,t)}{h} + \mathcal{O}(\tau) + \mathcal{O}(h),$$

wobei $\Delta x = h$ die Gitterweite und $\Delta t = \tau$ die Zeitschrittweite ist. Wir schreiben nun $u_j^n \approx u(jh, n\tau)$ und lösen also

$$\frac{u_j^{n+1} - u_j^n}{\tau} = c\frac{u_{j+1}^n - u_j^n}{h}$$

bzw. mit der *Courant-Zahl* $r = \frac{c\tau}{h}$

$$u_j^{n+1} - u_j^n = r(u_{j+1}^n - u_j^n).$$

3.18 Bemerkung (notwendige Bedingung für Konvergenz)
Wir betrachten nun eine notwendige (nicht hinreichende) Bedingung für die Konvergenz des Verfahrens: Der numerische Abhängigkeitsbereich (alle Punkte, die numerisch benötigt werden, um die Funktion in einem bestimmten Punkt (x_j, t_n) zu berechnen) zu einem beliebigen Gitterpunkt (x_j, t_n) muß den Quellpunkt der Charakteristik durch diesen Punkt enthalten.

Für $c < 0$ kann also dieses Verfahren nie konvergent sein, da wir das Schema „in der falschen Richtung“ betrachten.

3.19 Bemerkung (CFL-Bedingung (Courant, Friedrichs, Lewy, 1928))
Die CFL-Bedingung ist in unserem Fall erfüllt, falls $0 \leq r = \frac{c\tau}{h} \leq 1$ ist.

3.20 Bemerkung (CFL-Bedingung beim upwind schema)
Wenn $c < 0$ ist, so schreiben wir

$$\frac{u_j^{n+1} - u_j^n}{\tau} = \begin{cases} c\frac{u_{j+1}^n - u_j^n}{\tau}, & c \geq 0, \\ c\frac{u_j^n - u_{j-1}^n}{\tau}, & c < 0, \end{cases}$$

was ein so genanntes „upwind schema“ ist. Dann gilt

$$r = \frac{c\tau}{h} \leq 1 \iff c \leq \frac{h}{\tau}.$$

3.21 Bemerkung (von Neumann-Stabilität, 1947)
Wir wählen $u^0(x) = e^{\imath\alpha x}$, was bei genügend regulärem Anfangswert aufgrund der Fouriertransformation machbar ist. Dann berechnen wir mit dem Differenzenschema

$$\begin{aligned} u_j^1 &= u_j^0 + r(u_{j+1}^0 - u_j^0) = e^{\imath\alpha x} + r(e^{\imath\alpha h} - 1)e^{\imath\alpha x} = (1 + r(e^{\imath\alpha h} - 1))e^{\imath\alpha x}, \\ u_j^2 &= u_j^1 + r(u_{j+1}^1 - u_j^1) = (1 + r(e^{\imath\alpha h} - 1))^2 e^{\imath\alpha x}, \\ u_j^n &= G(\alpha)^n e^{\imath\alpha x}, \end{aligned}$$

wobei hier $G(\alpha) = 1 + r(e^{\imath\alpha h} - 1)$ ist. Damit die numerische Lösung beschränkt bleibt, fordern wir, daß $|G(\alpha)| \leq 1$ ist für jedes $\alpha \in \mathbb{R}$. Dies ist die von Neumann-Bedingung.

Im Beispiel ist das Verfahren für $r > 1$ instabil, denn es gilt $G(\frac{\pi}{h}) = 1 - 2r < -1$. Für $r < 0$ ist es auch instabil, denn $G(\frac{\pi}{h}) > 1$. Für $0 \leq r \leq 1$ ist es stabil, was der CFL-Bedingung entspricht (Im Allgemeinen sind die beiden Bedingungen aber nicht zueinander äquivalent).

Wir zeigen nun, daß die von Neumann-Bedingung auch hinreichend für die Konvergenz (mit Ordnung 1) ist, sofern die Anfangsdaten glatt genug sind. Die Beweisanalyse wird zeigen, daß die von Neumann-Bedingung (bis auf leichte Modifikation) notwendig für die Konvergenz ist.

3.22 Satz (Konvergenz)
Sei $u^0 \in \mathcal{C}^4(\mathbb{R})$ mit kompakten Träger und $|G(\alpha)| \leq 1$ für alle $\alpha \in \mathbb{R}$. Dann gilt

$$u_j^n - u(x_j, t_n) = \mathcal{O}(h) \text{ gleichmäßig für } n\tau \leq T.$$

Beweis
Wir schreiben

$$u^0 = \int_{\mathbb{R}} \hat{u}^0(\alpha) e^{\imath\alpha x}\, d\alpha \quad \text{mit} \quad \int_{\mathbb{R}} \alpha^2 |\hat{u}^0(\alpha)|\, d\alpha \leq M < \infty,$$

was mit unserer Annahme über u^0 erfüllt ist. Wir können die exakte Lösung durch

$$u(x,t) = u^0(x+ct) = \int_{\mathbb{R}} \hat{u}^0(\alpha) e^{\imath\alpha(x+ct)}\, dx = \int_{\mathbb{R}} \hat{u}^0(\alpha) e^{\imath\alpha x} e^{\imath\alpha ct}\, dx$$

darstellen. Weiter gilt

$$u_j^n = \int_{\mathbb{R}} \hat{u}^0(x) G(\alpha)^n e^{\imath\alpha x_j}\, d\alpha.$$

Wir vergleichen nun die beiden Integranden im Punkt (x_j, t_n) miteinander und schreiben

$$G(\alpha)^n - (e^{\imath\alpha c\tau})^n = (G(\alpha) - e^{\imath\alpha c\tau}) \underbrace{\Big(G(\alpha)^{n-1} + G(\alpha)^{n-2} e^{\imath\alpha c\tau} + \ldots + e^{\imath\alpha c\tau(n-1)}\Big)}_{|.|\leq n}, \quad (3.2)$$

wobei wir $|G(\alpha)| \leq 1$ benutzt haben. Mit der Taylor-Entwicklung gilt

$$\begin{aligned} |G(\alpha) - e^{\imath\alpha c\tau}| &= \left|1 + \frac{c\tau}{h}(e^{\imath\alpha h} - 1) - e^{\imath\alpha c\tau}\right| \\ &= \left|1 + \frac{c\tau}{h}\Big(1 + \imath\alpha h + \mathcal{O}(\alpha^2 h^2) - 1\Big) - \Big(1 + \imath\alpha c\tau + \mathcal{O}(\alpha^2 c^2 \tau^2 h^2)\Big)\right| \\ &= \mathcal{O}(\alpha^2 h\tau). \end{aligned}$$

Damit erhalten wir mit (3.2) die Abschätzung

$$|G(\alpha)^n - (e^{\imath\alpha c\tau})^n| \leq C\alpha^2 h\tau n = \mathcal{O}(\alpha^2 h).$$

Mit der Integraldarstellung der Lösungen gilt schließlich

$$|u_j^n - u(x_j, t_n)| \leq \int_{\mathbb{R}} C\alpha^2 h |\hat{u}^0(\alpha)|\, d\alpha = MCh = \mathcal{O}(h).$$

□

3.23 Beispiel (Lax–Friedrichs-Verfahren, Lax–Wendroff-Verfahren)

1. Für einen symmetrischen Differenzenquotienten im Raum berechnen wir

$$\frac{u_j^{n+1} - u_j^n}{\tau} = c\frac{u_{j+1}^n - u_{j-1}^n}{2h}.$$

Dabei ist $r = \frac{c\tau}{h}$ und die CFL-Bedingung ist $|r| \leq 1$, während dieses Verfahren die von Neumann-Bedingung für beliebiges $r \neq 0$ verletzt. Daher ist das Verfahren instabil.

2. Eine stabiles Verfahren der Ordnung 1 ist das *Lax–Friedrichs-Verfahren*: Wir berechnen

$$\frac{u_j^{n+1} - \frac{1}{2}(u_{j-1}^n + u_{j+1}^n)}{\tau} = c\frac{u_{j+1}^n - u_{j-1}^n}{2h}.$$

Dies erfüllt $|r| \leq 1$ und die von Neumann-Bedingung, was eine Übungsaufgabe ist.

3. Eine stabile Modifikation der Ordnung 2 ist das *Lax–Wendroff-Verfahren*: Wir rekapitulieren zunächst die Taylor-Entwicklung

$$u^{n+1} = u^n + \tau\partial_t u^n + \frac{\tau^2}{2}\partial_{tt}u^n + \dots$$

und berechnen nun

$$\frac{u_j^{n+1} - u_j^n}{\tau} = c\frac{u_{j+1}^n - u_{j-1}^n}{2h} + \frac{1}{2}c^2\tau\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2}.$$

Dieses Verfahren erfüllt $|r| \leq 1$ und ist von Ordnung 2. Dies werden wir später in Form einer allgemeinen Theorie zeigen.

3.24 Bemerkung

1. Falls $c = c(x,t)$ abhängt, wir also einen variablen Koeffizienten haben, dann können wir zeigen, daß die von Neumann-Stabilität punktweise notwendig und (im Wesentlichen) auch hinreichend ist. Auf diesen Punkt gehen wir im Folgenden nicht weiter ein.

2. Falls $\partial_t u = C\partial_x u$ mit einer diagonalisierbaren Matrix C mit reellen Eigenwerten ist, können wir die Verfahren aus dem vorherigen Beispiel direkt anwenden und erhalten eine Lösung.

3. Ebenfalls können die oben besprochenen Verfahren auch nichtlineare Systeme angewendet werden, auf was wir auch nicht weiter eingehen werden.

3.4 Ordnung, Stabilität und Konvergenz von Differenzverfahren

Wir möchten im Folgenden Satz 3.22 verallgemeinern. Hierzu führen wir zunächst einen theoretischen Rahmen ein.

3.25 Bemerkung (Fouriertransformation)
Sei $f : \mathbb{R} \to \mathbb{C}$. Dann ist die Fouriertransformation bekanntlich durch

$$\hat{f}(w) := \frac{1}{2\pi}\int_{\mathbb{R}} e^{-\imath wx} f(x)\,dx$$

definiert. Die Rücktransformation ist dabei gegeben durch

$$f(x) = \int_{\mathbb{R}} e^{\imath w x} \hat{f}(w)\, dw.$$

Diese Formeln sind dabei im Schwartz-Raum,

$$\mathcal{S} := \left\{ f \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{C}) : \ x^j f^{(k)}(x) \text{ ist beschränkt für alle } j, k \in \mathbb{N} \right\},$$

wohldefiniert. Dabei gilt für die Fouriertransformation $\mathcal{F} : \mathcal{S} \to \mathcal{S}$ mit der Identität

$$\int_{\mathbb{R}} |\hat{f}(w)|^2\, dw = \frac{1}{2\pi} \int_{\mathbb{R}} |f(x)|^2\, dx,$$

also ist $\mathcal{F}$ eine beschränkte lineare Abbildung, die sich mit der Dichtheit von $\mathcal{S} \subseteq L^2(\mathbb{R})$ stetig fortsetzen läßt durch $\mathcal{F} : L^2(\mathbb{R}) \to L^2(\mathbb{R})$. Diese Fortsetzung wird auch oft *Plancharel-Transformation* genannt.

Wir betrachten nun eine größere Klasse von Gleichungen.

3.26 Problem
Wir betrachten nun die lineare partielle Differentialgleichung mit konstanten Koeffizienten auf $\mathbb{R} \times [0, T]$, die durch

$$\begin{cases} \partial_t u(x,t) = L(\partial_x) u(x,t), \\ u(x,0) = u_0(x) \end{cases}$$

gegeben ist, wobei

$$L(\partial_x) u = \sum_{k=0}^{K} c_k \partial_x^k u$$

ein gegebenes Polynom mit $c_k \in \mathbb{C}^{n \times n}$ ist.

3.27 Beispiel

1. $\partial_t u = \partial_x u$, die Advektionsgleichung.
2. Das System $\partial_t u = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \partial_x u$, das äquivalent zur Wellengleichung ist.
3. $\partial_t u = \partial_x u + 7u$.
4. $\partial_t u = \partial_{xx} u$, die Wärmeleitungsgleichung.
5. $\partial_t u = -\partial_{xx} u$ macht hier keinen Sinn, was wir weiter unten sehen werden.
6. $\partial_t u = \imath \partial_{xx} u$, die Schrödingergleichung.

3.28 Bemerkung (verallgemeinerte Lösung)
Wir betrachten nun die formale Fourier-Transformation (im Ort), mit der wir die Gleichung

$$\begin{cases} \partial_t \hat{u}(w) = L(\imath w)\hat{u}(w), \\ \hat{u}(w,0) = \hat{u}_0(w) \end{cases}$$

mit $L(z) = \sum_{k=0}^{K} c_k z^k$ erhalten. Die Lösung hiervon ist für $w \in \mathbb{R}$

$$\hat{u}(w,t) = e^{tL(\imath w)}\hat{u}_0(w).$$

Wir möchten den Ausdruck hiervon gleichmäßig (bezüglich w) beschränkt haben und sehen, daß dies der Fall ist, falls $\operatorname{Re} L(\imath w) \leq \mu$ für alle $w \in \mathbb{R}$ ist. Dann gilt $|e^{tL(\imath w)}| \leq e^{t\mu}$, also auch $|\hat{u}(w,t)| \leq e^{t\mu}|\hat{u}_0(w)|$. Für die Fourierrücktransformation erhalten wir dann mit Plancharel

$$\|u(.,t)\|_{L^2} = \sqrt{2\pi}\|\hat{u}(.,t)\|_{L^2} \leq \sqrt{2\pi}e^{t\mu}\|\hat{u}_0\|_{L^2} = e^{t\mu}\|u_0\|_{L^2}.$$

Dies wird etwa erfüllt, wenn die partielle Differentialgleichung für $\hat{u}(.,t) \in \mathcal{S}$ liegt, i.e. falls also $\hat{u}_0 \in \mathcal{S}$, also auch $u_0 \in \mathcal{S}$ ist. In diesem Fall ist die Lösung klassisch.

Für ein beliebiges $u_0 \in L^2$ erhalten wir eine verallgemeinerte Lösung der partiellen Differentialgleichung mit

$$u(.,t) := \mathcal{F}^{-1}(e^{tL(\imath w)}\hat{u}_0(w)).$$

Die Diskussion motiviert die folgende Definition.

3.29 Definition (wohlgestelltes Anfangswertproblem)
Wir sagen, das Anfangswertproblem ist wohlgestellt, falls $\operatorname{Re} L(\imath w) \leq \mu$ für alle $w \in \mathbb{R}$ gilt.

Bei Systemen fordern wir entsprechend, daß $\operatorname{Re}\langle v, L(\imath w)v\rangle \leq \mu\|v\|^2 \quad \forall v \in \mathbb{C}^d$.

3.30 Beispiel

1. Die Differentialgleichung $\partial_t u = \partial_x u$ hat $L(\imath w) = \imath w$ und ist damit wohlgestellt.
2. Das System
$$\partial_t u = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \partial_x u$$
ist ebenfalls wohlgestellt und es gilt $\operatorname{Re}\left\langle v, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \imath w v \right\rangle = 0.$
3. Die Differentialgleichung $\partial_t u = \partial_x u + \gamma u$ erfüllt $L(\imath w) = \imath w + \gamma$, ist also wohlgestellt.
4. Die Wärmeleitungsgleichung $\partial_t u = \partial_{xx} u$ erfüllt $L(\imath w) = (\imath w)^2 = -w^2$, ist also wegen $\operatorname{Re} L(\imath w) \leq 0$ wohlgestellt.
5. Die Gleichung $\partial_t u = -\partial_{xx} u$ ist wegen $L(\imath w) = w^2$ nicht wohlgestellt.

6. Die Schrödingergleichung, $\partial_t u = \imath \partial_{xx} u$, ist wegen $L(\imath w) = \imath(\imath w)^2 = -\imath w^2$ wohlgestellt.

Wir betrachten nun das Differenzverfahren, eine solche Gleichung zu lösen.

3.31 Konstruktion (Differenzverfahren)
Wir betrachten $u_j^n = u(jh, n\tau)$ für $h = \Delta x$, $\tau = \Delta t$, $j \in \mathbb{Z}$ und $n \in \mathbb{N}$. Dabei sind h und τ so gekoppelt, daß $\tau = \varphi(h)$ mit $\varphi(h) \to 0$ für $h \to 0$ geht; etwa durch $r = \frac{c\tau}{h}$, die Courant-Zahl. Für den Vektor $u^n = (u_j^n)_{j\in\mathbb{Z}}$ betrachten wir die Gleichung

$$\sum_{l=-L}^{L} q_l u_{j+l}^{n+1} = \sum_{l=-L}^{L} p_l u_{j+l}^n, \tag{3.3}$$

wobei q_l und p_l von h und τ abhängen. Wir definieren nun die Operatoren

$$Q(z) := \sum_{l=-L}^{L} q_l z^l, \quad P(z) := \sum_{l=-L}^{L} p_l z^l, \quad (Ev)_j := v_{j+1},$$

wobei für den Verschiebungsoperator E gilt, daß $\widehat{Ev}(\alpha) = e^{\imath h\alpha}\hat{v}(\alpha)$. Weiter definieren wir

$$Q(E) := \sum_{l=-L}^{L} q_l E^l, \quad (Q(E)u^{n+1})_j := \sum_{l=-L}^{L} q_l u_{j+l}^{n+1}.$$

Damit läßt sich nun (3.3) kompakt schreiben durch

$$Q(E)u^{n+1} = P(E)u^n. \tag{3.4}$$

Falls nun gilt, daß

$$u_j^n = u^n(jh) = \int_{\mathbb{R}} e^{\imath\alpha x}\hat{u}^n(\alpha)\, d\alpha,$$

dann ist

$$u_j^{n+1} = \int_{\mathbb{R}} e^{\imath\alpha x}\hat{u}^{n+1}(\alpha)\, d\alpha,$$

wobei $\hat{u}^{n+1} = G(\alpha)\hat{u}^n(\alpha)$ durch den Verstärkungsfaktor

$$G(\alpha) := \frac{P(e^{\imath h\alpha})}{Q(e^{\imath h\alpha})}$$

gegeben ist, wobei wir auf (3.4) die Fouriertransformation angewendet haben, um auf die Darstellung für $\hat{u}^{n+1}$ zu kommen.

3.32 Definition (Stabilität und Ordnung von G)
Nach der obigen Herleitung ist die Stabilität und Ordnung des Differenzenverfahrens durch die Funktion G bereits charakterisiert.

1. Wir sagen, G ist *stabil*, wenn $|G(\alpha)| \leq e^{\gamma\tau}$ mit einem von h, τ, α unabhängigen γ gilt und $\tau = \varphi(h)$ gegeben ist.

   In diesem Fall gilt dann auch

   $$|G(\alpha)^n| \leq e^{\gamma\tau n} \leq e^{\gamma T} \quad \text{für } n\tau \leq T.$$

2. Wir sagen, G hat *Ordnung* p, falls

   $$|G(\alpha) - e^{\tau L(\imath\alpha)}| \leq C\tau h^p(1 + |\alpha|^q)$$

   gilt, wobei C und q unabhängig von h, τ, α und p sind und $\tau = \varphi(h)$ gilt.

3.33 Beispiel
Wir betrachten die Gleichung $\partial_t u = c\partial_x u$, womit $L(\imath w) = cw$ ist. Wir überprüfen nun Stabilität des Lax–Wendroff-Schemas aus Beispiel 3.23 sowie die Ordnung des Schemas.

1. Mit der Taylor-Entwicklung und der Gültigkeit der Gleichung erhalten wir das Lax–Wendroff-Schema, wobei

   $$Q(z) = 1, \qquad P(z) = 1 + \frac{r}{2}(z - z^{-1}) + \frac{r^2}{2}(z - 2 + z^{-1})$$

   gegeben ist, also ist

   $$G(z) = 1 + \frac{r}{2}\underbrace{(e^{\imath h\alpha} - e^{-\imath h\alpha})}_{=\imath\sin(h\alpha)} + \frac{r^2}{2}\underbrace{(e^{\imath h\alpha} - 2 + e^{-\imath h\alpha})}_{=2\cos h\alpha - 2}.$$

   Für $|r| \leq 1$ gilt $|G(\alpha)| \leq 1$, also ist das Verfahren stabil im Sinne von Definition 3.32.

2. Mit der Definition von G und der Taylor-Entwicklung von G erhalten wir

   $$|G(\alpha) - e^{\tau c\imath\alpha}| = \mathcal{O}(r(h\alpha)^3) = \mathcal{O}(\tau h^2\alpha^3),$$

   wobei wir $\tau c = rh$ benutzt haben. Damit hat das Verfahren Ordnung 2.

Der folgende Satz liefert gleichmäßige Konvergenz (in der Zeit) mit Ordnung p für glatte Anfangsdaten.

3.34 Satz (Konvergenz für glatte Anfangsdaten)
Sei das Verfahren stabil für $\tau = \varphi(h)$ und von Ordnung p im Sinne von Definition 3.32. Sei weiter $u^0 \in \mathcal{S}$. Dann gibt es eine von h, τ, j, n unabhängige Konstante M mit

$$|u_j^n - u(jh, n\tau)| \leq Mh^p \quad \text{für } n\tau \leq T,$$

Beweis
Der Beweis verläuft gleich wie der Beweis von Satz 3.22 und ist eine Übungsaufgabe. □

Wir möchten nun auch Konvergenz für nichtglatte Anfangsdaten erhalten.

3.35 Definition (Interpolation)
Wir definieren die Fortsetzung der Interpolation der numerischen Lösung auf ganz $\mathbb{R}$ für ein $x \in \mathbb{R}$ durch

$$u^n(x) = \int_{\mathbb{R}} e^{\imath\alpha x} \hat{u}^n(\alpha)\, d\alpha,$$

wobei $\hat{u}^{n+1}(\alpha) = G(\alpha)\hat{u}^n(\alpha)$, also $\hat{u}^n(\alpha) = G(\alpha)^n \hat{u}_0(\alpha)$ ist und $\hat{u}^n(\alpha)$ damit auf ganz $\mathbb{R}$ bekannt ist. Wir beachten, daß $u_j^n = u^n(jh)$ für alle $j \in \mathbb{Z}$ gilt.

3.36 Satz (Lax, Richtmyer, 1954)
Sei das Verfahren konsistenz (i.e. es hat Ordnung $p \geq 1$) und stabil im Sinne von Definition 3.32. Sei $u_0 \in L^2(\mathbb{R})$. Dann gilt

$$\|u^n - u(., n\tau)\|_{L^2(\mathbb{R})} \to 0 \quad \text{für } h \to 0,\ \tau = \varphi(h) \to 0,$$

wobei die Konvergenz gleichmäßig (in der Zeit) für $n\tau \leq T$ ist.

Beweis
Wir erhalten mit der Plancharel-Transformation

$$\|u^n - u(., n\tau)\|_{L^2(\mathbb{R})} = \sqrt{2\pi}\|\hat{u}^n - \hat{u}(., n\tau)\|_{L^2(\mathbb{R})}.$$

Wir betrachten nun den Integranden hiervon.

1. Es gilt

$$\hat{u}^n(\alpha) - \hat{u}(\alpha, n\tau) = (G(\alpha)^n - e^{n\tau L(\imath\alpha)})\hat{u}_0(\alpha). \tag{3.5}$$

Dabei ergibt sich mit den Voraussetzungen und dem binomischen Lehrsatz

$$|G(\alpha)^n - e^{n\tau L(\imath\alpha)}| = \underbrace{|G(\alpha) - e^{\tau L(\imath\alpha)}|}_{\leq C\tau h^p(1+|\alpha|^q)} \underbrace{|G(\alpha)^{n-1} + G(\alpha)^{n-2}e^{\tau L(\imath\alpha)} + \ldots + e^{(n-1)\tau L(\imath\alpha)}|}_{\leq n e^{\beta n\tau}},$$

wobei $\beta = \max\{\mu, \gamma\}$ ist. Für $|\alpha|^q h^p \leq \sqrt{h}$ gilt dabei $|G(\alpha)^n - e^{n\tau L(\imath\alpha)}| = \mathcal{O}(\sqrt{h})$. Die Bedingung an α impliziert, daß

$$|\alpha| \leq B_h := h^{\frac{-p+\frac{1}{2}}{q}}.$$

Dabei gilt $B_h \to \infty$ für $h \to 0$. Aus (3.5) und unserer Bemerkung ergibt sich

$$\int_{|\alpha|\leq B_h} |\hat{u}^n(\alpha) - \hat{u}(\alpha, n\tau)|^2\, dx \leq Ch\|\hat{u}^0\|_{L^2(|\alpha|\geq B_h)} \leq Ch\|\hat{u}^0\|_{L^2(\mathbb{R})} = \mathcal{O}(h).$$

2. Es gilt allgemein

$$|G(\alpha)^n - e^{n\tau L(\imath\alpha)}| \leq |G(\alpha)|^n + |e^{\tau L(\imath\alpha)}|^n \leq e^{n\tau\gamma} + e^{n\tau\mu} \leq 2e^{n\tau\beta}$$

mit $\beta = \max\{\gamma, \mu\}$. Damit gilt (für große α) also

$$\int_{|\alpha|\geq B_h} |\hat{u}^n(\alpha) - \hat{u}(\alpha, n\tau)|^2 \, d\alpha \leq 4e^{\beta T} \int_{|\alpha|\geq B_h} |\hat{u}^0(\alpha)|^2 \, d\alpha \to 0,$$

da $B_h \to \infty$ für $h \to 0$ geht.

Für den Grenzübergang $h \to 0$ verschwindet – wie erwähnt – das Integral aus dem zweiten Teil und – mit der gleichmäßigen Abschätzung aus dem ersten Teil – verschwindet das Integral auch insgesamt und es gilt damit die Behauptung. □

3.5 Dissipation, Dispersion und Gruppengeschwindigkeit

3.37 Wiederholung
Wir untersuchen nun die Ausbreitung von Wellen im Differenzenverfahren. Als Modellproblem betrachten wir die Advektionsgleichung $\partial_t u = c\partial_x u$, die für $c \in \mathbb{R}$ die Lösung $u(x,t) = u^0(x + ct)$ hat. Falls $u_0(x) = e^{\imath\alpha x}$ ist, so gilt $|u(x,t)| = |e^{\imath\alpha(x+ct)}| = 1$ für jedes t.

Wir betrachten nun das Differenzenverfahren $u^{n+1} := G(\alpha)u^n$, also auch $u^n = G(\alpha)^n u^0$. Dies ist *stabil*, falls $|G(\alpha)| \leq 1$ ist.

3.38 Definition (Dissipation)
Wir sagen, das Differenzenverfahren ist *dissipativ*, falls $|G(\alpha)| < 1$ gilt.

Bei dissipativen Verfahren werden hohe Frequenzen (i.e. $h\alpha$ weg von 0) gedämpft und dies führt zur Glättung (was nicht unbedingt immer erwünscht ist). Nicht dissipative Verfahren treten hierbei typischerweise bei impliziten Verfahren auf.

3.39 Beispiel

1. Beim upwind-Verfahren für $c > 0$ gilt für $r = \frac{c\tau}{h}$

$$G(\alpha) = 1 + r(e^{\imath h\alpha} - 1).$$

Für $r < 1$ ergibt sich dabei $|G(\alpha)| < 1$, sofern $h\alpha$ kein ganzzahliges Vielfaches von 2π ist. Mit exakter Rechnung erhalten wir ein $\varrho > 0$ mit

$$|G(\alpha)| \leq 1 - \varrho \sin^2 \frac{h\alpha}{2}.$$

Dieses Verhalten nennen wir dissipativ von Ordnung 2.

2. Das Lax–Wendroff-Verfahren ist dissipativ von Ordnung 4 und es gilt hier

$$|G(\alpha)|^2 = 1 - 4r^2(1 - r^2) \sin^4 \frac{h\alpha}{2}.$$

3. Das Crank–Nicholson-Verfahren,

$$\frac{u_j^{n+1}-u_j^n}{\tau}=\frac{c}{2}\left(\frac{u_{j+1}^{n+1}-u_{j-1}^{n+1}}{2h}+\frac{u_{j+1}^n-u_{j-1}^n}{2h}\right),$$

ist nicht dissipativ. Hierfür formen wir die Gleichung in

$$\left(\frac{1}{\tau}-\frac{c}{2}\frac{e^{\imath\alpha h}-e^{-\imath h\alpha}}{2h}\right)\hat{u}^{n+1}(\alpha)=\left(\frac{1}{\tau}+\frac{c}{2}\frac{e^{\imath\alpha h}-e^{-\imath\alpha h}}{2h}\right)\hat{u}^n(\alpha)$$

um und erhalten damit

$$\hat{u}^{n+1}(\alpha)=\frac{1+\frac{r}{2}\imath\sin\alpha h}{1-\frac{r}{2}\imath\sin\alpha h}\hat{u}^n(\alpha)=:G(\alpha)\hat{u}^n(\alpha).$$

Wir sehen direkt, daß $|G(\alpha)|=1$ für jedes $\alpha\in\mathbb{R}$ ist.

3.40 Bemerkung (Phasengeschwindigkeit, Dispersion)
Sei $u_0(x)=e^{\imath\alpha x}$, also $u(x,t)=e^{\imath\alpha(x+ct)}$. Mit dem Differenzenschema gilt $u_j^n=G(\alpha)^n e^{\imath\alpha x}$. Für die *Phasengeschwindigkeit* $\gamma(\alpha)$ schreiben wir

$$G(\alpha)=|G(\alpha)|e^{\imath\alpha\gamma(\alpha)\tau}$$

und damit gilt

$$u_j^n=|G(\alpha)|^n e^{\imath\alpha(x+\gamma(\alpha)t)},$$

wobei der hintere Faktor für $x+\gamma(\alpha)t=\text{const}$ konstant ist (Beim kontinuierlichen Problem mußte hierfür $x+ct$ konstant sein). Dies bedeutet, daß sich beim Differenzverfahren Wellen einer Wellenzahl α unterschiedlich schnell mit den Geschwindigkeiten $\gamma(\alpha)$ ausbreiten. Dieses Phänomen wird auch *Dispersion* genannt. Wir nennen $c-\gamma(\alpha)$ den Phasenfehler.

3.41 Beispiel (Phasengeschwindigkeit bei Lax–Wendroff)
Beim Verfahren von Lax–Wendroff gilt (Übung!)

$$\gamma(\alpha)=c\left(1-\frac{1}{6}(h\alpha)^2(1-r^2)+\mathcal{O}((h\alpha)^4)\right).$$

3.42 Problem
Wir betrachten die Differentialgleichung $\partial_t u=c\partial_x u$ mit dem Anfangswert $u_0(x)=e^{\imath\alpha x}a(x)$, wobei $a\in\mathcal{S}$ ist. Dann ergibt sich für die kontinuierliche Lösung

$$u(x,t)=e^{\imath\alpha(x+ct)}a(x+ct).$$

Ein solcher Anfangswert u_0 heißt auch *Wellenpaket*. Der folgende Satz führt die Gruppengeschwindigkeit ein und bringt sie in Zusammenhang mit der Phasengeschwindigkeit.

3.43 Satz (über die Gruppengeschwindigkeit)
Sei das Differenzenverfahren zum Problem 3.42 stabil und nichtdissipativ (i.e. es gilt $|G(\alpha)| = 1$ für alle $\alpha \in \mathbb{R}$). Dann gilt für $x = jh$ und $t = n\tau$

$$u_j^n = e^{\imath\alpha(x+\gamma(\alpha)t)} a(x + g(\alpha)t) + \mathcal{O}(h)$$

mit der Phasengeschwindigkeit $\gamma(\alpha)$ und der *Gruppengeschwindigkeit* $g(\alpha)$, gegeben durch

$$g(\alpha) := \frac{d}{d\alpha}(\alpha\gamma(\alpha)) = \gamma(\alpha) + \alpha\gamma'(\alpha).$$

Beweis
Mit der Fouriertransformation des Anfangswerts erhalten wir

$$\hat{u}_0(w) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-\imath wx} u_0(x)\, dx = \hat{a}(w - \alpha).$$

Weiter gilt für die Iterierte des Differenzenschemas

$$u_j^{n+1} = \int_{\mathbb{R}} e^{\imath wx} \hat{u}^{n+1}(w)\, dw,$$

wobei $\hat{u}^{n+1}(w) = G(w)\hat{u}^n(w)$ ist. Mit der Nicht-Dissipativität gilt dabei

$$\hat{u}^n(w) = G(w)^n \hat{u}_0(w) = G(w)^n \hat{a}(w - \alpha) = |G(w)|^n e^{\imath w\gamma(w)t} \hat{a}(w - \alpha). \tag{3.6}$$

Weiter gilt

$$e^{\imath w\gamma(w)t} = e^{\imath\alpha\gamma(\alpha)t} e^{\imath(w\gamma(w)t - \alpha\gamma(w)t)}. \tag{3.7}$$

Für $\Phi(w) := w\gamma(w)$ ergibt sich mit Taylor und der Tatsache, daß γ eine glatte Funktion von $h\alpha$ ist und damit $\Phi''(\xi) = \mathcal{O}(h)$ ist,

$$\Phi(w) - \Phi(\alpha) = \underbrace{\Phi'(\alpha)}_{=g(\alpha)}(w - \alpha) + \frac{1}{2}\underbrace{\Phi''(\xi)}_{=\mathcal{O}(h)}(w - \alpha)^2.$$

Damit gilt

$$e^{\imath\Phi(w)t - \imath\Phi(\alpha)t} = e^{\imath g(\alpha)(w-\alpha)t + \mathcal{O}(h(w-\alpha)^2)} = e^{\imath g(\alpha)(w-\alpha)t}(1 + \mathcal{O}(h(w - \alpha)^2)).$$

Mit dieser Information, (3.6) sowie (3.7) rechnen wir schließlich

$$\begin{aligned} u_j^n &= \int_{\mathbb{R}} e^{\imath w x} \hat{u}^n(w)\,dw = \int_{\mathbb{R}} e^{\imath w x} e^{\imath\alpha\gamma(\alpha)t} e^{\imath g(\alpha)(w-\alpha)t}(1+\mathcal{O}(h(w-\alpha)^2))\hat{a}(w-\alpha)\,dw \\ &= e^{\imath\alpha(x+\gamma(\alpha)t)} \int_{\mathbb{R}} e^{\imath(w-\alpha)x} e^{\imath(w-\alpha)g(\alpha)t}\hat{a}(w-\alpha)\,dw + \mathcal{O}(h) \\ &= e^{\imath\alpha(x+\gamma(\alpha)t)} \int_{\mathbb{R}} e^{\imath(x+g(\alpha)t)(w-\alpha)}\hat{a}(w-\alpha)\,dw + \mathcal{O}(h). \end{aligned}$$

Mit der Transformation $\mu = w - \alpha$ ergibt sich nun

$$= e^{\imath\alpha(x+\gamma(\alpha)t)} \underbrace{\int_{\mathbb{R}} e^{\imath(x+g(\alpha)t)\mu}\hat{a}(\mu)\,d\mu}_{=a(x+g(\alpha)t)} + \mathcal{O}(h),$$

wobei wir im letzten Schritt die inverse Fouriertransformation benutzt haben. □

3.44 Bemerkung
Die Gruppengeschwindigkeit repräsentiert daher die Transportgeschwindigkeit der einhüllenden Funktion a des Anfangswert durch die Differentialgleichung, während die Phasengeschwindigkeit die Geschwindigkeit einer bestimmten Frequenz beschreibt.

3.6 Randbedingungen

Randbedingungen sind bei den hyperbolischen Probleme vor allem in der Numerik schwer zu handhaben. Wir betrachten im restlichen Verlauf dieses Kapitels die Advektionsgleichung mit Randbedingungen und werden die Schwierigkeiten dabei untersuchen.

Wir möchten im Folgenden die Advektionsgleichung auf ein bestimmtes Gebiet beschränken und betrachten nun die Gleichung $\partial_t u = c\partial_x u$ für $x \geq 0$ und $t \geq 0$.

- Für $c < 0$ haben wir bei $x = 0$ einen *Einströmrand* und brauchen Randbedingungen, etwa $u(0,t) = g(t)$.
- Für $c > 0$ haben wir bei $x = 0$ einen *Ausströmrand*, können aber keine Randbedingungen vorgeben, da wegen $u(0,t) = u^0(ct)$ bereits diese Werte determiniert sind.

3.45 Beispiel (Randbedingungen beim Lax–Wendroff-Verfahren)
- Für $c < 0$ brauchen wir Randbedingungen, etwa $u_0^{n+1} = g(t_{n+1})$.
- Für $c > 0$ erhalten wir u_0^{n+1} nicht aus dem Verfahren. Hierfür müssen wir den Wert extra definieren. Wir unterscheiden zwei Fälle
  1. $u_0^{n+1} := u_0^n + r(u_1^n - u_0^n)$,
  2. $u_0^{n+1} := u_0^{n-1} + 2r(u_1^n - u_0^n)$.

Wir beachten, daß die Stabilität eines Vefahrens durch Randbedingungen vernichtet werden kann. Während das Lax–Wendroff-Vefahren für Anfangswertprobleme stabil war, bleibt es für die Randwerte bei $c < 0$ stabil. Für $c > 0$ bleibt es für die erste Möglichkeit der Definition stabil, für die zweite wird es allerdings instabil (was wir später noch zeigen werden).

Wir untersuchen dieses Phänomen nun theoretisch. Zunächst wenden wir uns dem Einströmrand ($c < 0$) zu. Wir betrachten im Folgenden ein Verfahren vom Typ

$$\begin{cases} u_j^{n+1} = a_1 u_{j+1}^n + a_0 u_j^n + a_{-1} u_{j-1}^n, \quad j = 1, 2, \dots \\ u_0^{n+1} = g(t_{n+1}) =: g^{n+1}, \\ u_j^0 = 0. \end{cases} \tag{3.8}$$

Da wir hier Randbedingungen untersuchen wollen, nehmen wir keinen Anfangswert an. Für die Funktion G, über die wir Stabilität definiert haben, gilt dabei

$$G(\alpha) := a_1 e^{\imath\alpha h} + a_0 + a_{-1} e^{-\imath\alpha h}.$$

3.46 Beispiel
Das Lax–Wendroff-Verfahren ist vom Typ (3.8). Für $r = \frac{c\tau}{h}$ gilt dabei

$$a_1 = \frac{r^2 + r}{2}, \quad a_0 = 1 - r^2, \quad a_{-1} = \frac{r^2 - r}{2}.$$

Der folgende Satz besagt, daß Probleme mit Einströmrand harmlos in dem Sinne sind, daß sich Stabilität des Anfangswertproblems direkt aus der Stabilität des Anfangsrandwertproblem vererbt.

3.47 Satz (Stabilität für Einströmrand)
Falls das Differenzverfahren stabil für das Anfangswertproblem ist (i.e. $|G(\alpha)| \leq 1$), dann erfüllt es für das Anfangsrandwertproblem (3.8) die Stabilitätsabschätzung

$$\tau \sum_{n=0}^{N} |u_j^n|^2 \leq \tau \sum_{n=0}^{N} |g^n|^2.$$

Beweis
Wir definieren die Erzeugendenfunktionen

$$u_j(z) := \sum_{n=0}^{\infty} u_j^n z^{-n}, \qquad g(z) := \sum_{n=0}^{N} g^n z^{-n}.$$

Wir halten fest, daß die Erzeugendenfunktionen für $|z| > 1$ konvergieren, solange u_j^n und g^n beschränkt bleiben. Für alle anderen z betrachten wir die Potenzreihen zunächst als formale Potenzreihen.

Multiplikation von (3.8) mit z^{-n} und Summation über alle n liefert nun die lineare Rekursion

$$zu_j(z) = a_1 u_{j+1}(z) + a_0 u_j(z) + a_{-1} u_{j-1}(z). \tag{3.9}$$

Wir betrachten das charakteristische Polynom $a_1\zeta^2 + (a_0 - z)\zeta + a_{-1}$ und möchten nun wissen, wo dessen beide Nullstellen $\zeta_1(z), \zeta_2(z)$ liegen.

Da das Verfahren für das Anfangswertproblem stabil ist, gilt $|a_1\zeta + a_0 + a_{-1}\zeta^{-1}| \leq 1$ für $|\zeta| = 1$. Also gilt für $z = a_1\zeta + a_0 + a_{-1}\zeta^{-1}$ und $|\zeta| = 1$, daß $|z| \leq 1$ ist. Damit kann es für $|z| > 1$ keine Nullstellen vom Betrag 1 geben.

Aus dem Satz von Vieta wissen wir, daß $\zeta_1\zeta_2 = \frac{a_{-1}}{a_1} \in \mathbb{R}$ und $\zeta_1 + \zeta_2 = \frac{z-a_0}{a_1}$ gilt, also ist $\zeta_1(z) \to 0$ und $\zeta_2(z) \to \infty$ für $z \to \infty$. Für $|z| > 1$ gibt es also genau eine Nullstelle, die betragsmäßig echt kleiner als 1 ist.

Eine allgemeine Lösung der linearen Regression (3.9) ist $u_j = c_1\zeta_1^j + c_2\zeta_2^j$. Für $j \to \infty$ ist nur eine beschränkte Lösung sinnvoll (was sich direkt aus der Schema ergibt, da wir mit dem von uns geforderten Anfangswert für festes n für große j irgendwann $u_j^n = 0$ erhalten; genauer muß eine solche Lösung sogar gegen Null gehen), i.e. es gilt $c_2 = 0$. Damit gilt

$$u_j(z) = u_0(z)\zeta_1(z)^j = g(z)\zeta_1(z)^j,$$

also gilt wegen $|\zeta_1(z)| \leq 1$ für $|z| \geq 1$ und einem Stetigkeitsargument

$$|u_j(z)| \leq |g(z)| \quad \forall |z| \geq 1.$$

Durch mehrmaliges Anwenden der Parseval-Gleichung erhalten wir damit

$$\sum_{n=0}^{\infty} |u_j^n|^2 = \frac{1}{2\pi} \int_0^{2\pi} |u_j(e^{\imath\theta})|^2 \, d\theta \leq \frac{1}{2\pi} \int_0^{2\pi} |g(e^{\imath\theta})|^2 \, d\theta = \sum_{n=0}^{\infty} |g^n|^2.$$

Dabei beachten wir, daß die Summe nicht von $n = -\infty$, sondern erst ab $n = 0$ beginnt, da u_j^n für negative n analytisch ist und diese daher diese Summanden verschwinden.

Da u_j^n unabhängig von den g^m für $m > n$ ist, können wir diese in der Summe weglassen und erhalten damit die gewünschte Abschätzung, indem wir $g^m = 0$ für $m > N$ setzen und die dann erhaltene Gleichung mit τ multiplizieren. □

Wir betrachten nun den Ausströmrand ($c > 0$). Hier betrachten wir nun ebenfalls ein Verfahren mit Drei-Term-Rekursion

$$u_j^{n+1} = a_1 u_{j+1}^n + a_0 u_j^n + a_{-1} u_{j-1}^n,$$

zu welcher wir die rationale Funktion $a(z, \zeta) := a_1\zeta + a_0 + a_{-1}\zeta^{-1} - z$ assoziieren.

Für $r = \frac{c\tau}{h}$ definieren wir nun zwei Strategien für numerische Randbedingungen.

$$u_0^{n+1} := u_0^n + r(u_1^n - u_0^n), \tag{3.10}$$
$$u_0^{n+1} := u_0^{n-1} + 2r(u_1^n - u_0^n). \tag{3.11}$$

Weiter sei die Anfangsbedingung u_j^0 gegeben, wobei u_j^0 beschränkt ist. Der Einfachheit halber setzen wir $a_0^0 = u_0^{-1} := 0$. Zu den beiden Randwertbedingungen assoziieren wir nun rationale Funktionen. Dabei soll z symbolisch die Verschiebung in der Zeit und ζ die Verschiebung im Ort bezeichnen.

1. Zur Randwertbedingung (3.10) assoziieren wir $b(z, \zeta) := z - 1 - r(\zeta - 1)$.
2. Zur Randwertbedingung (3.11) assoziieren wir $b(z, \zeta) := z - z^{-1} - 2r(\zeta - 1)$.

Für den folgenden Satz sei $b(z, \zeta)$ eine allgemeine rationale Funktion, die wir mit den Randwerten assoziieren.

3.48 Satz (Instabilitätskriterium von Godunov-Ryabenkii)
Falls es $z, \zeta \in \mathbb{C}$ mit $|z| > 1$ und $|\zeta| < 1$ mit $a(z, \zeta) = 0$ und $b(z, \zeta) = 0$ gibt, dann ist das Verfahren instabil.

Beweis
Zu $a(z, \zeta) = 0$ gibt es – wie wir aus dem Beweis von Satz 3.47 wissen – zwei Lösungen $\zeta_1(z)$ und $\zeta_2(z)$ mit $|\zeta_1(z)| < 1$ und $|\zeta_2(z)| > 1$.

Wir wählen $u_j^0 := \zeta_1(z)^j$ für $j \geq 0$. Dann ist $u_j^n = z^n \zeta_1(z)^j$ die Lösung des Differenzenverfahrens (was wir durch Einsetzen überprüfen können). Für $|z| > 1$ wächst u_j^n exponentiell mit der Zeit (also mit n) an und daher ist dies sogar exponentiell instabil. □

3.49 Bemerkung
Die Voraussetzung von Satz 3.48 ist etwa für das Verfahren von Lax–Wendroff mit der Randbedingungsstrategie (3.11) erfüllt (Übung!). Daher ist dieses Verfahren instabil.

3.50 Bemerkung (Stabilitätsuntersuchung für die Strategie (3.10))
Wir untersuchen nun nur noch die Randbedingungsstrategie (3.10). Wir sehen, daß für $r \leq 1$ keine Nullstelle von $b(z, \zeta)$ mit $|z| > 1$ und $|\zeta| < 1$ existiert. Daher können wir mit Satz 3.48 nicht direkt die Instabilität schließen.

Für die Erzeugendenfunktionen gilt aufgrund der Rekursion – wie im Beweis von Satz 3.47

$$\sum_{n=0}^{\infty} u_j^{n+1} z^{-n} = z \sum_{n=0}^{\infty} u_j^{n+1} z^{-(n+1)} = z(u_j(z) - u_j^0),$$

was zur inhomogenen (in j) linearen Rekursion

$$z u_j(z) - u_j^0 = a_1 u_{j+1}(z) + a_0 u_j(z) + a_{-1} u_{j-1}(z)$$

führt. Für $j \to \infty$ haben wir eine beschränkte Lösung und lassen daher bei der allgemeinen Lösung, die ähnlich wie im Beweis von Satz 3.47 hergeleitet wird, die Nullstelle, die Unbeschränktheit implizieren würde, weg. Wir erhalten also

$$u_j(z) = u_0(z)\zeta_1(z)^j + f_j(z), \tag{3.12}$$

wobei f_j eine beschränkte Lösung der inhomogenen Gleichung mit $f_0 \equiv 0$. Mit den Randbedingungen aus (3.10) gilt

$$zu_0(z) = u_0(z) + r(u_1(z) - u_0(z)).$$

Wir nehmen nun der Einfachheit halber an, daß die Anfangswerte verschwinden und erhalten damit durch Einsetzen in (3.12) mit $\zeta = \zeta_1(z)$

$$(z - 1 - r(\zeta - 1))u_0(z) = rf_1(z),$$

also gilt

$$u_0(z) = \frac{rf_1(z)}{z - 1 - r(\zeta - 1)}.$$

Da beim Lax–Wendroff-Verfahren $\zeta_1(z) = 1 + \frac{1}{r}(z-1) + \mathcal{O}((z-1)^2) \to 1$ (für $z \to 1$) gilt, wird diese Summe für $z \to 1$ kritisch. In diesem Fall gilt

$$\frac{1}{z - 1 - r(\zeta - 1)} \approx \frac{1}{(z-1)^2} = \frac{1}{z^2\left(1 - \frac{1}{z}\right)^2} = \frac{1}{z^2}\sum_{n=0}^{\infty} nz^{-n},$$

was lineare Instabilität (immerhin keine exponentielle) bedeuten würde.

Stichwortverzeichnis

S

V

W