Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig

Der Stabilitätsbegriff in der Numerik

by

Wolfgang Hackbusch

Lecture note no.: 20 2003

Der Stabilitätsbegriff in der Numerik*

Wolfgang Hackbusch
Max-Planck-Institut *Mathematik in den Naturwissenschaften*
Inselstr. 22-26, D-04103 Leipzig, Germany
email: wh@mis.mpg.de

Zusammenfassung

In Kapitel 1 behandeln wir die Kondition einer Aufgabe und die Stabilität eines Algorithmus. Hier ist die Verstärkung von Eingabe- bzw. Gleitkommafehlern das Maß der Kondition bzw. Stabilität. Die Begriffe bleiben aber noch vage, da zwischen Verstärkungsfaktoren der Größenordnung 1 und großen Verstärkungsfaktoren nicht exakt unterschieden werden kann.

In Kapitel 2 beschäftigen wir uns mit Quadraturverfahren, genauer gesagt mit einer Familie von Quadraturen Q_n, wobei mit wachsendem n die Qualität zunehmen soll. Letztere wird mittels der *Konsistenz* beschrieben. Die *Stabilität* wird wieder anhand der Eingabefehlerverstärkung definiert. Anders als in §1 kann diese eindeutig definiert werden, da die vagen Begriffe "klein" und "groß" dadurch ersetzt werden, dass eine Größe $\sup C_n$ endlich oder unendlich ist. Obwohl sich die Stabilitätsdefinition an numerischen Phänomenen orientiert, eignet sie sich auch für analytische Zwecke. Stabilität ist fast äquivalent zur Konvergenz des Quadraturergebnisses $Q_n(f)$ gegen das exakte Integral $\int f dx$. Entsprechend wird analytisches Werkzeug aus der Funktionalanalysis benötigt: der Approximationssatz von Weierstraß und der Satz von der gleichmäßigen Beschränktheit.

Die in Kapitel 3 behandelte Interpolation folgt dem gleichen Schema wie schon §2. In beiden Kapiteln kann man sich die folgende Frage stellen: Auch wenn die Stabilität durch eine Aussage der Form $\sup C_n < \infty$ beschrieben wird und notwendig für die Konvergenz ist, so sagt das wenig darüber, ob man sinnvollerweise eine Quadratur bzw. Interpolation für ein *festes* n auch ohne Stabilitätsvoraussetzung verwenden kann.

Dies ist anders in Kapitel 4, in dem es um Ein- und Mehrschrittverfahren zur Lösung gewöhnlicher Anfangswertprobleme geht. Bei der Berechnung der Näherungen $y(x_0 + jh)$ ergibt sich fast immer die Notwendigkeit, größere j zu verwenden (entweder weil im Grenzprozess die Schrittweite h gegen null geht und deshalb $j = (x - x_0)/h \to \infty$ oder weil h konstant gehalten wird, aber y auf vielen Gitterpunkten $x_0 + jh$ approximiert werden soll).

Während bei gewöhnlichen Differentialgleichungen Stabilitätsprobleme erst bei echten Mehrschrittverfahren auftreten und Einschrittverfahren stets stabil sind, ändert sich dies bei den partiellen Differentialgleichungen, die in §5 behandelt werden. Es werden Differenzenverfahren für hyperbolische und parabolische Differentialgleichungen behandelt. Stabilität drückt sich hier mittels der gleichmäßigen Beschränktheit von Potenzen des Differenzenoperators aus.

Auch im Falle elliptischer Differentialgleichungen stellt sich die Frage der Stabilität. Wie in §6 ausgeführt, besteht die Stabilität in einer schrittweiten-unabhängigen Beschränkung der *Inversen* des Differenzenoperators bzw. der Finite-Element-Matrix.

*Als zweistündige Vorlesung mit Übungen gehalten an der Christian-Albrechts-Universität zu Kiel im Sommersemester 2003

Inhaltsverzeichnis

1 Stabilität bzw. Kondition bei endlichen Algorithmen

1.1 Einige Begriffe

Ein Algorithmus dient zur Lösung einer Aufgabe. In der mathematischen Formulierung ist eine "*Aufgabe*" (auch "*Problem*" genannt) eine Abbildung $\Phi : X \to Y$, die numerisch zu realisieren ist.[1]

Damit ein Algorithmus durchführbar ist, muss er aus programmtechnisch realisierbaren Bausteinen zusammengesetzt sein. Letztere heißen "*elementare Operationen*" und sind die arithmetischen Grundoperationen $+, -, *, /$ im Bereich der reellen oder ganzen Zahlen. Dazu kommen die in Programmiersprachen direkt aufrufbare Funktionen wie z.B. sin, cos, exp, $\sqrt{\cdot}, \ldots$

Ein "*Algorithmus*" ist die Komposition von elementaren Operationen. Dabei zeichnet sich ein "*endlicher Algorithmus*" dadurch aus, dass er nur eine endlicher Zahl von elementaren Operationen benötigt. Ein Algorithmus ist eine Realisierung der Abbildung

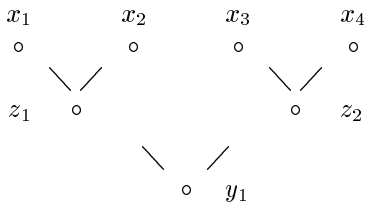
$$\Phi : (x_1, \ldots, x_n) \in X = \text{Urbild}(\Phi) \mapsto (y_1, \ldots, y_m) \in Y = \text{Bild}(\Phi). \tag{1.1.1}$$

Zu einem Φ gibt es (wenn überhaupt) unendlich viele Algorithmen, die Φ realisieren.

Ein endlicher Algorithmus kann durch einen Graphen dargestellt werden. Dessen Knoten bestehen u.a. aus

- $\{x_1, \ldots, x_n\}$: Eingabewerte (im Definitionsbereich elementarer Operationen; $n \in \mathbb{N}_0$),
- $\{y_1, \ldots, y_m\}$: Ausgabewerte (im Bildbereich elementarer Operationen; $m \in \mathbb{N}$),
- $\{z_1, \ldots\}$: Zwischenresultate.

Der gerichtete Graph (V, E) besteht aus der Knotenmenge V ("vortices") und der Kantenmenge E ("edges"). Dabei zeigen die Kanten von Argumentwerten zum Bildwert jeder elementaren Operationen (siehe Beispiel).



Graph zur Aufgabe $y_1 = x_1x_2 + x_3x_4$ mit Zwischenwerten $z_1 := x_1x_2$, $z_2 := x_3x_4$

1.2 Genauigkeit der elementaren Operationen

Grundlegend für die Computerarithmetik ist die Gleitkomma-Arithmetik, die von einen festen Mantissenlänge charakterisiert wird. Bei t Stellen zur Basis b ($t \geq 1, b \geq 2$) ist die Menge $\mathcal{M}$ der *Maschinenzahlen* gegeben durch die Null (die eine Sonderstellung einnimmt) und alle Zahlen der Form

$$x = \pm 0.d_1 \ldots d_t * b^E = \pm b^E \sum_{k=1}^{t} d_k b^{-k}, \text{ wobei } 1 \leq d_1 < b \text{ und } 0 \leq d_k < b \text{ für } k > 1,\ E \in \mathbb{Z}$$

(die Beschränktheit des Exponenten E, die zu Überlauf bzw. Unterlauf führen kann, wird hier ignoriert).

Übungsaufgabe 1.2.1 *a) Was ist die beste Schranke ε in*

$$\sup_{\xi \in \mathbb{R}} \min_{x \in \mathcal{M}} |x - \xi| \leq \varepsilon |x| \ ?$$

b) Wie lautet ε im Spezialfall $b = 2$ der Dualbasis?

[1] Vgl. Kapitel 1 in Stoer: *Einführung in die Numerische Mathematik I.* Springer-Verlag, Berlin, 8. Auflage, 1999

Das minimierende x aus $\min_{x\in\mathcal{M}}|x-\xi|$ ist die beste "Rundung" von $\xi\in\mathbb{R}$. Wenn wir diese Abbildung mit rd bezeichnen, gilt die folgende Abschätzung mit eps $=\varepsilon$:

$$|x - rd(x)| \leq \text{eps}|x|$$

Im folgenden nehmen wir an, dass eps so groß ist, dass alle elementaren Operationen "op" der folgenden Ungleichung genügen:

$$|gl\,(a \text{ op } b) - (a \text{ op } b)| \leq \text{eps}|a \text{ op } b| \qquad \text{für alle } a, b \in \mathcal{M}. \tag{1.2.1}$$

Dabei bezeichnet $gl(\ldots)$ die Durchführung der Operationen im Argument im Sinne der Computer-Gleitkomma-Arithmetik. Entsprechendes gelte für die einstelligen Operationen. Ferner sei eps $\ll 1$ unterstellt.

Man beachte, dass (1.2.1) den *relativen* Fehler kontrolliert.

Bemerkung 1.2.2 *Die Ungleichung (1.2.1) könnte als Konsistenz bezeichnet werden.*

1.3 Beispiel

Die *n-te Bessel-Funktion*[2] (auch *Zylinderfunktion* genannt) hat die Darstellungen

$$J_n(x) = \sum_{k=0}^{\infty} \frac{(-1)^k\,(x/2)^{n+2k}}{k!\,(n+k)!} = \frac{(-1)^n}{\pi}\int_0^{\pi} e^{ix\cos(\varphi)}\cos(n\varphi)d\varphi \qquad \text{für } n\in\mathbb{N}_0. \tag{1.3.1}$$

Die Aufgabe sei, $J_5(0.6)$ zu bestimmen.

Ausgehend von der Tatsache, dass J_0 und J_1 tabelliert[3] sind:

$$J_0(0.6) = 0.9120,\ J_1(0.6) = 0.2867$$

und die Rekursion

$$J_{n+1}(x) = -J_{n-1}(x) + \frac{2n}{x}J_n(x) \tag{1.3.2}$$

gilt, bietet sich der folgende Algorithmus an:

$$\begin{aligned}
J_2(0.6) &= -J_0(0.6) + \frac{2}{0.6}J_1(0.6) = -0.9120 + \frac{2}{0.6}0.2867 &&= 4.36667_{10}\text{-}2,\\
J_3(0.6) &= -J_1(0.6) + \frac{4}{0.6}J_2(0.6) = -0.2867 + \frac{4}{0.6}4.36667_{10}\text{-}2 &&= 4.41111_{10}\text{-}3,\\
J_4(0.6) &= -J_2(0.6) + \frac{6}{0.6}J_3(0.6) = -4.36667_{10}\text{-}2 + \frac{6}{0.6}4.41111_{10}\text{-}3 &&= 4.44444_{10}\text{-}4,\\
J_5(0.6) &= -J_3(0.6) + \frac{8}{0.6}J_4(0.6) = -4.41111_{10}\text{-}3 + \frac{8}{0.6}4.44444_{10}\text{-}4 &&= 1.51481_{10}\text{-}3.
\end{aligned}$$

Das Resultat erhält man mit nur 8 Elementaroperationen auf Grund exakter Gleichungen. Trotzdem ist das berechnete Resultat $J_5(0.6)$ auch in der Größenordnung völlig falsch. Das korrekte Ergebnis lautet $J_5(0.6) = 1.99482_{10}\text{-}5$. Insgesamt gilt *exakt*

$$\begin{aligned}
J_0(0.6) &= 0.912005, &\qquad J_1(0.6) &= 0.286701,\\
J_2(0.6) &= 4.36651_{10}\text{-}2, &\qquad J_3(0.6) &= 4.39966_{10}\text{-}3,\\
J_4(0.6) &= 3.31470_{10}\text{-}3, &\qquad J_5(0.6) &= 1.99482_{10}\text{-}5.
\end{aligned}$$

Den Fehlschlag der Berechnung werden wir anschließend erklären. Zuvor als positives Beispiel eine erfolgreiche Approximation.

[2] Friedrich Wilhelm Bessel, geb. 22. Juli 1784 in Minden, gest. 17. März 1846 in Königsberg

[3] Seite 103 im *Teubner-Taschenbuch der Mathematik*, Teubner, Stuttgart, 1996

Definiert man für negative $-n$ die Bessel-Funktionen $J_{-n}(x) := (-1)^n J_n(x)$, so gilt die Summenformel

$$\sum_{n=-\infty}^{\infty} J_n(x) = 1 \qquad \text{für alle } x \in \mathbb{R}.$$

Elimination der $J_{-n}(x)$ liefert

$$J_0(x) + 2J_2(x) + 2J_4(x) + \ldots = 1. \tag{1.3.3}$$

Übungsaufgabe 1.3.1 *Gegeben sei die Reihendarstellung aus (1.3.1).*
a) Man zeige die Konvergenz für alle $x \in \mathbb{C}$ (d.h. J_n ist eine ganze Funktion).
b) Man beweise (1.3.3). Hinweis hierzu: Per Koeffizientenvergleich zeige man, dass (1.3.3) äquivalent ist zu

$$\sum_{m=0}^{\ell} \frac{(-1)^m}{(\ell-m)!\,(\ell+m)!} * \left\{ \begin{array}{l} 2, \text{ falls } m > 0 \\ 1, \text{ falls } m = 0 \end{array} \right\} = 0 \qquad \text{für } \ell > 0.$$

Zum Nachweis dieser Gleichheiten verwende man die Identität $e^x e^{-x} = 1$.

Als erste Approximation ersetzen wir (1.3.3) durch

$$\tilde{J}_0(x) + 2\tilde{J}_2(x) + 2\tilde{J}_4(x) + \ldots + 2\tilde{J}_m(x) = 1 \tag{1.3.4}$$

für ein geeignetes gerades m. Wir wollen die Rekursion (1.3.2) in umgekehrter Richtung anwenden. Dazu brauchen wir Anfangswerte für $\tilde{J}_m, \tilde{J}_{m-1}$. Hierzu setzen wir (zweite Approximation)

$$\tilde{J}_m = \tilde{J}_{m-1} = c$$

mit einer noch zu bestimmenden Konstanten c. Startend mit $\tilde{J}_m = \tilde{J}_{m-1} = c$ liefert die Rekursion $\tilde{J}_{n-1} = -\tilde{J}_{n+1} + \frac{2n}{0.6}\tilde{J}_n$ alle $\tilde{J}_k$, $0 \le k \le m$. Das Resultat für $c = 1$ liefere $\hat{J}_k$, $0 \le k \le m$. Sei

$$C := \hat{J}_0(x) + 2\hat{J}_2(x) + 2\hat{J}_4(x) + \ldots + 2\hat{J}_m(x).$$

Dann liefert die Wahl $c := 1/C$ Werte $\tilde{J}_k$, die (1.3.4) erfüllen.

Beispiel: Zu $m = 10$ ergibt sich $\tilde{J}_{10} = \tilde{J}_9 = 5.55\ldots_{10}\text{-}11$ und $\tilde{J}_5 = 1.994820_{10}\text{-}5$ (alle angegebenen Stellen exakt).

1.4 Fehlerverstärkung

1.4.1 Auslöschung

Die Ursache der katastrophalen Resultate für $J_5(0.6)$ nach dem ersten Algorithmus kann man anhand einer einzigen Operation diskutieren:

$$y = x_1 - x_2. \tag{1.4.1}$$

Einerseits ist die Operation *harmlos*, denn die "Konsistenz" (1.2.1) der Subtraktion garantiert, dass für $\tilde{y} = gl(x_1 - x_2)$ die Abschätzung

$$|\tilde{y} - y| \le \text{eps}|y|$$

gilt. Andererseits gilt dies nur für $x_1, x_2 \in \mathcal{M}$. Die realistische Aufgabe lautet

$$\eta = \xi_1 - \xi_2, \qquad x_1 = rd(\xi_1), x_2 = rd(\xi_2), \qquad y = x_1 - x_2.$$

Der interessierende Fehler ist $|\eta - \tilde{y}|$ für $\tilde{y} = gl(x_1 - x_2)$. Der absolute Fehler beträgt

$$\begin{aligned} |\eta - \tilde{y}| &\le |x_1 - x_2 - gl(x_1 - x_2)| + |\Delta x_1| + |\Delta x_2| \\ &\le \text{eps}\,(|x_1 - x_2| + |x_1| + |x_2|) \qquad \text{mit } \Delta x_i := \xi_i - x_i. \end{aligned}$$

Solange $|\eta| \sim |x_i|$ $(i = 1, 2)$, gibt es kein Problem, da auch der relative Fehler von der Größenordnung eps ist. Dramatisch ist jedoch der Fall der Auslöschung, wenn $|\eta| \ll |x_1| + |x_2|$. In diesem Fall ist der relative Fehler $\lesssim \text{eps}\frac{|x_1|+|x_2|}{|\eta|}$. Im schlimmsten Fall ist $|\eta|$ von der Größenordnung eps $(|x_1| + |x_2|)$, sodass der relative Fehler $\mathcal{O}(1)$ beträgt.

1.4.2 Lineare (differentielle) Fehleranalyse

Zu untersuchen ist, wie sich Fehler Δx_i der Eingabe $\tilde{x}_i = x_i + \Delta x_i$ auf die Ausgabe $\tilde{y} = \Phi(\tilde{x})$ auswirkt.

Die Abbildung Φ aus (1.1.1) sei stetig differenzierbar. Die Taylor-Entwicklung angewandt auf $\tilde{y}_j = \Phi_j(x_1, \ldots, x_i + \Delta x_i, \ldots, x_n)$ liefert $\tilde{y}_j = \Phi_j(x_1, \ldots, x_n) + \frac{\partial \Phi_j}{\partial x_i} \Delta x_i + o(\Delta x_i)$. Damit ist

$$\frac{\partial \Phi_j(x_1, \ldots, x_n)}{\partial x_i}$$

der Verstärkungsfaktor[4] des Eingabefehlers Δx_i, wobei die *absoluten* Fehler zugrundegelegt sind. Die Verstärkung des *relativen* Fehlers lautet

$$\frac{\partial \Phi_j(x_1, \ldots, x_n)}{\partial x_i} \times \frac{|x_i|}{|y_j|}.$$

Welcher Fehler entscheidend ist (absolut oder relativ, weitere Alternative: relativ bezüglich einer Norm), hängt von der jeweiligen Aufgabe ab.

1.4.3 Kondition und Stabilität

Ein Problem heißt *gut konditioniert*, falls die Verstärkung der Eingabefehler von der Größenordnung 1 ist. Wird der Fehler dagegen stark verstärkt, heißt die Aufgabe *schlecht konditioniert*. Die maximale Fehlerverstärkung wird die Kondition(szahl) genannt.

Gegeben sei ein Algorithmus, der die Aufgabe Φ realisiert. Neben den Knoten $\{x_1, \ldots, x_n\} \subset V$ des zugehörigen Graphen, gibt es Zwischenresultate ξ_α. Die Restabbildung R beschreibt, wie sich das Endresultat y aus den Zwischenresultaten berechnet. Entsprechend dem Verstärkungsfaktor $\partial \Phi_j / \partial x_i$ zum Eingabefehler Δx_i von x_i, ist $\partial R_j / \partial \xi_\alpha$ der Verstärkungsfaktor des Fehlers $\Delta \xi_\alpha$ von ξ_α. Da ξ_α das Resultat einer Elementaroperation ist, ist der relative Fehler von ξ_α durch (1.2.1) beschrieben.

Sei κ die Konditionszahl des Problems Φ. Wenn alle Verstärkungsfaktoren $\partial R_j / \partial \xi_\alpha$ höchstens von der Größenordnung κ sind, heißt der Algorithmus *stabil*. Andernfalls heißt er *instabil*.

Kondition und Stabilität basieren beide auf der Verstärkung von Eingabefehlern. Die Unterscheidung zwischen Kondition und Stabilität ergibt sich aus der Unterscheidung der Knoten im Graphen des Algorithmus nach Eingabe- und Zwischenwert-Knoten. Der Algorithmus ist nur für die Zwischenwert-Knoten verantwortlich, während die Eingabeknoten $\{x_1, \ldots, x_n\}$ ein Teil der Aufgabendefinition sind. Außerdem ist noch der folgende Unterschied zu beachten: die Fehler $\Delta \xi_\alpha$ genügen stets (1.2.1). Dagegen kann der Fehler Δx_i viele Ursachen haben. Im Falle des Beispiels aus §1.3 ist Δx_i dadurch bedingt, dass die Tabellenwerte nur auf 4 Dezimalstellen gegeben sind. Weiterhin können Messfehler Ursache größerer Fehler sein. Bestensfalls ist Δx_i durch Rundung einer exakten Nicht-Maschinenzahl entstanden (dann gilt ebenfalls (1.2.1)).

Die Definitionen der Kondition und Stabilität sind absichtlich vage[5] gehalten: Ob relative oder absolute Fehler zu Grunde gelegt werden sollen, ob einzelne Komponenten oder irgendwelche Normen verglichen werden sollen, welche Fehlerverstärkung als groß eingestuft wird, ist nicht eindeutig festgelegt.

1.4.4 Diskussion des Beispiels aus §1.3

Die Aufgabe hat die Form $\Phi : (J_0, J_1) \mapsto J_5$. Der Algorithmus hat u.a. die Zwischenwerte J_2, J_3, J_4. Lässt man (wie in der Definition gefordert) nur elementare Operationen zu, erfordert die Berechnung von J_2 aus J_0, J_1 die weiteren Zwischenwerte

$$F_1 := 2/0.6, \qquad P_1 := F_1 * J_1,$$

die dann zu $J_2 := P_1 - J_0$ führen (vgl. $J_{n+1}(x) = -J_{n-1}(x) + \frac{2n}{x} J_n(x)$).

Da $\Delta J_0 = 0.000005$ und $\Delta J_1 = 0.000001$ zu $\Delta J_5 = 0.0015$ führen, müssen die Verstärkungsfaktoren $\partial J_5 / \partial J_i$ $(i = 1, 2)$ groß sein, d.h. die Aufgabe[6], J_5 aus J_0, J_1 zu berechnen, ist schlecht konditioniert.

[4] Man spricht auch dann von einem "Verstärkungsfaktor", wenn dieser < 1 ist und deshalb eher "Abschwächungsfaktor" heissen sollte.

[5] Es gibt Versuche einer systematischen Definition; vgl. L.S. de Jong: Towards a formal definition of numerical stability. Numer. Math. **28** (1977) 211-219.

[6] Hier wird $\Phi : (J_0, J_1) \mapsto J_5$ als die "Aufgabe" bezeichnet. Diese Sicht ist insofern richtig, als Eingabewerte für J_0, J_1 gegeben werden und die Gleitkommafehler der Zwischenrechnungen des Algorithmus harmlos sind. Andererseits ist die wirkliche Aufgabe, $J_5(0.6)$ zu berechnen, wozu J_0, J_1 nicht notwendigerweise benötigt werden.

Übungsaufgabe 1.4.1 *Man berechne die Verstärkungsfaktoren* $\partial J_5/\partial J_i$ $(i = 1, 2)$.

Die zweite Berechnungsweise für $J_5(0.6)$ aus §1.3 lässt sich besser als die Aufgabe $\Phi_m : x \mapsto J_5$ beschreiben, da $J_m, J_{m-1}, \ldots, J_0$ als gleichberechtigte Zwischenresultate auftreten. Zu beachten ist, dass Φ_m nicht zu wirklich exaktem $J_5(x)$ führt, sondern aufgrund der Approximationsfehler (Summenabbruch und $J_m = J_{m-1}$) nur eine Näherung $J_5(x) + \varepsilon_m(x)$ liefert. Die exakten Werte J_m, J_{m-1} sind zwar nicht gleich, aber für $m \gg 1$ beide sehr klein. Zudem ist der Verstärkungsfaktor von ΔJ_m (d.h. die Ableitung $\partial R_m/\partial J_m$ der Restabbildung $R_m : (J_m, J_{m-1}) \mapsto J_5$) klein.

Übungsaufgabe 1.4.2 *Man berechne die Verstärkungsfaktoren* $\partial R_m/\partial J_m$ *für* $R_m : (J_m, J_{m-1}) \mapsto J_5$ *und die Werte* $m = 10, 20$.

1.4.5 Weiteres Beispiel für einen instabilen Algorithmus

Ist das Problem Φ schlecht konditioniert, darf auch der Algorithmus eine größere Fehlerverstärkung aufweisen, um immer noch stabil sein zu können. Ist dagegen Φ gut konditioniert, darf der Algorithmus keine Gleitkommafehler extrem vergrößern.

Das folgende Beispiel zeigt, dass Konzepte, die man etwa aus der Linearen Algebra kennt, zu katastrophal instabilen Algorithmen führen können. Die gestellte Aufgabe sei die Eigenwertberechnung symmetrischer Matrizen (auch hier ist eine Approximation unvermeidlich, da im Allgemeinen die Eigenwerte nicht mit endlich vielen Elementaroperationen exakt berechenbar sind). Die Eigenwertberechnung symmetrischer Matrizen ist gut konditioniert, wie der folgende Satz[7] zeigt:

Satz 1.4.3 *Seien* $A, \Delta A \in \mathbb{R}^{n\times n}$ *symmetrische Matrizen und* $\tilde{A} := A + \Delta A$. $\tilde{\lambda}$ *sei ein Eigenwert von* $\tilde{A}$. *Dann gibt es einen Eigenwert* λ *von* A, *sodass* $|\tilde{\lambda} - \lambda| \le \|\Delta A\|_2$ ($\|\cdot\|_2$ *ist die Spektralnorm*).

Da die Eigenwerte in der Linearen Algebra über das charakteristische Polynom eingeführt werden, könnte man auf die folgende Idee kommen:

a) Man berechne zunächst das charakteristische Polynom $P(x) = \sum_{k=0}^{n} a_k x^k$.

b) Anschließend berechne man die Eigenwerte mittels Nullstellensuche für $P(x) = 0$.

Wir unterstellen in optimistischer (und unrealistischen) Weise, dass die Koeffizienten a_k sich fast exakt aus A berechnen lassen (mindestens entsteht ein relativer Fehler eps, da man nicht erwarten kann, dass $a_k \in \mathcal{M}$, und daher eine Rundung notwendig ist). Es bleibt der zweite Teil zu untersuchen: Wie reagieren die Nullstellen von P auf Fehler in a_k? Hierzu ist ein Beispiel[8] von Wilkinson[9] aufschlussreich.

Die Eigenwerte (Nullstellen) $1, 2, \ldots, 20$ seien vorgegeben. Sie sind die Nullstellen von

$$P(x) = \prod_{k=1}^{20} (i - x) = a_0 + \ldots + a_{19}x^{19} + a_{20}x^{20}$$

($a_0 = 20! = 2\,432\,902\,008\,176\,640\,000, \ldots, a_{19} = 190$, $a_{20} = 1$). Der große Wert von a_0 zeigt die Gefahr des Überlaufs bei Rechnungen mit Polynomen. Die Nullstellenbestimmung von P sieht eher einfach aus, denn P hat nur einfache Nullstellen und diese sind zudem deutlich getrennt.

Wir stören nur den Koeffizienten a_{19} in $\tilde{a}_{19} = a_{19} - 2^{-23}$ ($2^{-23} = 1.192 \times 10^{-7}$ ist die Maschengenauigkeit der "einfach genauen" Rechnung). Die Nullstellen des gestörten Polynoms $\tilde{P}$ sind

$$1, \quad 2, \quad 3, \quad 4, \quad 4.\underbrace{999999}_{7}28, \quad 6.\underbrace{000006}_{5}9, \quad 6.\underbrace{999}_{3}69, \quad 8.\underbrace{00}_{2}72, \quad 8.917,$$

$$10.09 \pm 0.64i, \quad 11.79 \pm 1.65i, \quad 13.99 \pm 2.5i, \quad 16.73 \pm 2.8i, \quad 19.5 \pm 1.9i, \quad 20.84.$$

Die (absoluten wie relativen) Fehler sind nicht nur von der Größenordnung $O(1)$, sondern zudem ist die Struktur der reellen Nullstellen zerstört.

Die konjugiert komplexen Paare von Nullstellen treten bei hinreichend kleinen Störungen nicht mehr auf. Dazu betrachten wir jetzt die Störung $\tilde{a}_{19} = a_{19} - 2^{-55}$ ($2^{-55} = 2.776 \times 10^{-17}$). Die Nullstellen sind dann

$$1, \ldots, 10, \quad 11 - 10^{-10}, \quad 12 + 6_{10} - 9, \quad 13 - 17_{10} - 9, \quad 14 + 37_{10} - 9, \quad 15 - 59_{10} - 9, \quad 16 + 47_{10} - 9, \ldots$$

Die Störung der Nullstelle 15 zeigt z.B. die Fehlerverstärkung $59 \times 10^{-9}/2^{-55} = 2.1 \times 10^{9}$.

[7] Eine allgemeinere Formulierung mit Beweis findet man als Satz (6.9.6) in: Stoer - Bulirsch: Einführung in die Numerische Mathematik II. Springer-Verlag, Berlin.

[8] Seiten 54ff in: J.H. Wilkinson: *Rundungsfehler*. Springer-Verlag 1969

[9] James Hardy Wilkinson, geb. 27. Sept. 1919 in Strood, gest. 5. Okt. 1986 in London

2 Quadratur

2.1 Hintergrund

Als Quadratur wird die numerische Approximation eines Integrals bezeichnet.[10] Die Integrationsaufgabe wird hier in der einfachsten Form gestellt.

Aufgabe 2.1.1 *Es sei vorausgesetzt, dass $f \in C([0,1])$ oder dass f höhere Differentiationseigenschaften besitzt. Die Aufgabe besteht in der näherungsweisen Berechnung von $\int_0^1 f(x)dx$.*

Andere Intervalle und eventuelle Gewichtsfunktionen sind möglich, aber für unsere Zwecke uninteressant. Zu $n \in \mathbb{N}_0$ definieren wir Quadraturformeln

$$Q_n(f) = \sum_{i=0}^{n} a_{i,n} f(x_{i,n}). \tag{2.1.1}$$

Dabei sind $a_{i,n}$ die Quadraturgewichte und $x_{i,n}$ die (disjunkten) Stützstellen. Eine Folge $\{Q_n : n \in \mathbb{N}_0\}$ bildet eine *Familie von Quadraturformeln.*

Man erhält (2.1.1) im Allgemeinen als "interpolatorische Quadratur" über eine Interpolation $f(x) \approx f_n(x) := \sum_{i=0}^{n} f(x_{i,n})\Phi_{i,n}(x)$ mit Lagrange-Funktionen (d.h. $\Phi_{i,n}(x_{j,n}) = \delta_{ij}$; vgl. (3.1.1)). Indem man über f_n anstelle von f integriert, erhält man

$$\int_0^1 f(x)dx \approx \int_0^1 f_n(x)dx = \sum_{i=0}^{n} f(x_{i,n}) \underbrace{\int_0^1 \Phi_{i,n}(x)dx}_{=:a_{i,n}}.$$

Standardwahl bei der Interpolation ist die Polynominterpolation. In diesem Falle sind $\Phi_{i,n} = L_{i,n}$ die Lagrange[11]-Polynome.

Übungsaufgabe 2.1.2 *Die Definition $a_{i,n} = \int_0^1 L_{i,n}(x)dx$ ist für die konkrete Berechnung nicht so hilfreich. Man zeige, dass man stattdessen die $a_{i,n}$ über das folgende Gleichungssystem berechnen kann:*

$$\sum_{i=0}^{n} a_{i,n} x_{i,n}^k = \frac{1}{k+1} \qquad \text{für } k = 0, 1, \ldots, n.$$

Bisher waren die Stützstellen $x_{i,n}$ noch frei. Ihre Wahl bestimmt die (Familie der) Quadraturformeln:

- $x_{i,n} = \frac{i}{n}$ führt auf die *Newton*[12]*-Cotes*[13]*-Quadratur*[14],
- $x_{i,n}$: Nullstellen des Legendre-Polynoms[15] P_{n+1} ergibt die *Gauß*[16]*-Quadratur.*

2.2 Konsistenz

Im Zusammenhang mit interpolatorische Quadratur definiert über Polynominterpolation verwendet man die folgende Konsistenzdefinition.

[10] Vgl. Kapitel 3 in Stoer: *Einführung in die Numerische Mathematik I.* Springer-Verlag, Berlin, 8. Auflage, 1999

[11] Joseph-Louis Lagrange, geb. 25. Jan. 1736 in Turin, gest. 10. April 1813 in Paris

[12] Sir Isaac Newton, geb. 4. Jan. 1643 in Woolsthorpe, gest. 31. März 1727 in London

[13] Roger Cotes, geb. 10. Juli 1682 in Burbage, gest. 5. Juni 1716 in Cambridge

[14] Die Newton-Cotes-Quadraturfamilie ist nur für $n \in \mathbb{N}$ definiert, nicht für $n = 0$.

[15] Genauer muss man vom transformierten Legendre-Polynom sprechen. Das eigentliche Legendre-Polynom ist über $[-1,1]$ definiert und muss für unsere Zwecke auf das Integrationsintervall $[0,1]$ transformiert werden. Sind ξ_i die originalen Nullstellen in $[-1,1]$, so ist $x_{i,n} := (1+\xi_i)/2$.
Adrien-Marie Legendre, geb. 18. Sept. 1752 in Paris, gest. 10. Jan. 1833 in Paris

[16] Johann Carl Friedrich Gauß, geb. 30. April 1777 in Braunschweig, gest. 23. Feb. 1855 in Göttingen

Definition 2.2.1 *Eine Familie* $\{Q_n : n \in \mathbb{N}_0\}$ *heißt* konsistent*, falls es eine Funktion* $g : \mathbb{N}_0 \to \mathbb{N}$ *mit* $g(n) \to \infty$ *für* $n \to \infty$ *gibt, sodass*

$$Q_n(P) = \int_0^1 P(x)dx \qquad \text{für alle Polynome } P \text{ mit } \operatorname{grad}(P) \le g(n). \tag{2.2.1}$$

Eine unmittelbare Folge ist das

Korollar 2.2.2 $\{Q_n : n \in \mathbb{N}_0\}$ *sei konsistent. Dann gilt für jedes Polynom* P*, dass* $\lim Q_n(P) = \int_0^1 P(x)dx$.

Beweis. Sei $\gamma := grad(P)$. Wegen $g(n) \to \infty$ für $n \to \infty$ gibt es ein n_0 mit $g(n) \ge \gamma$ für alle $n \ge n_0$. Daher zeigt $Q_n(P) = \int_0^1 P(x)dx$ für $n \ge n_0$ die Behauptung. ∎

Wie aus der Numerik-Grundvorlesung bekannt ist, gilt für die oben erwähnten Quadratur-Familien:

$$g(n) = \left\{ \begin{array}{ll} n & n \text{ ungerade} \\ n+1 & n \text{ gerade} \end{array} \right\} \text{ für die Newton-Cotes-Quadratur,}$$

$$g(n) = 2n+1 \qquad \text{für die Gauß-Quadratur.}$$

Später werden wir noch eine alternative (allgemeinere) Konsistenzbedingung formulieren (vgl. §2.9).

2.3 Konvergenz

Zur Konvergenz lassen sich unterschiedliche Fragen stellen. Gegeben eine Funktion f, wird man sich zunächst für den Fehler

$$\varepsilon_n(f) := \left| Q_n(f) - \int_0^1 f(x)dx \right|$$

interessieren. Neben der Konvergenzfrage (gilt $\lim \varepsilon_n(f) = 0$?) ist die Abschätzung von $\varepsilon_n(f)$ für ein *festes* n oft die Frage von stärkerem praktischen Interesse. Eine (eventuell zu grobe) Abschätzung erhält über den Interpolationsfehler $f - f_n$. Genauere Antworten liefert die Verwendung des Peano-Kernes[17]. In jedem Falle erhält man Schranken der Form

$$\varepsilon_n(f) \le c_n \|f^{(k_n)}\|_\infty, \tag{2.3.1}$$

die Ableitungen von f der Ordnung $k_n \le g(n)+1$ verwenden.

Offenbar hängt die rechte Seite in (2.3.1) von c_n und f (bzw. seinen Ableitungen) ab. Die Quadraturfamilie $\{Q_n\}$ ist nur für die Konstanten c_n zuständig, sodass die Konvergenz $c_n\|f^{(k_n)}\|_\infty \to 0$ nicht als eine Eigenschaft von $\{Q_n : n \in \mathbb{N}_0\}$ angesprochen werden kann. Für ein $f \in C^k([0,1])$ ist (2.3.1) zudem nur für solche n mit $k_n \le k$ anwendbar. Im letzten Fall ist überhaupt keine unendliche Folge von Schranken $\{c_n\|f^{(k_n)}\|_\infty\}$ formulierbar.

Man kann sich fragen, ob man Fehlerabschätzungen (2.3.1) mit $k_n = 0$ finden kann: $\varepsilon_n(f) \le c_n \|f\|_\infty$. Die Antwort lautet, dass in diesem Fall aber nicht $c_n \to 0$ gelten kann. Hierzu modifiziere man die konstante Funktion $f = 1$ in den η-Umgebungen der Stützstellen $x_{i,n}$ $(0 \le i \le n)$ so, dass $0 \le \tilde{f} \le f$ und $\tilde{f}(x_{i,n}) = 0$. Wegen $\tilde{f}(x_{i,n}) = 0$ folgt $Q_n(\tilde{f}) = 0$, während sich $\int_0^1 \tilde{f}(x)dx$ beliebig wenig von $\int_0^1 f(x)dx = 1$ unterscheidet, wenn η hinreichend klein ist. Da $\|f\|_\infty = \|\tilde{f}\|_\infty = 1$, gewinnen wir keine bessere Fehlerabschätzung als $\varepsilon_n(\tilde{f}) = \int_0^1 \tilde{f}(x)dx \le 1 = 1\,\|f\|_\infty$, d.h. $c_n = 1$. Damit erhalten wir die

Bemerkung 2.3.1 *Abschätzungen der Form* $\varepsilon_n(f) \le c_n \|f\|_\infty$ *können nur mit Konstanten* $c_n \ge 1$ *gelten.*

Damit können die rechten Seiten $c_n \|f\|_\infty$ keine Nullfolge bilden. Dies schließt aber nicht aus, dass trotzdem $\varepsilon_n(f) \to 0$ gelten mag. Letzteres ist der Inhalt der folgenden

Definition 2.3.2 *Eine Familie* $\{Q_n : n \in \mathbb{N}_0\}$ *von Quadraturformeln heißt* konvergent*, falls*

$$\text{für alle } f \in C([0,1]) \text{ gilt: } \quad Q_n(f) \to \int_0^1 f(x)dx. \tag{2.3.2}$$

Man beachte, dass (2.3.2) zu $\varepsilon_n(f) \to 0$ äquivalent ist.

[17] Vgl. Kapitel 3.2 in Stoer: *Einführung in die Numerische Mathematik I.* Springer-Verlag, Berlin, 8. Auflage, 1999

2.4 Stabilität

2.4.1 Verstärkung der Eingabefehler

Zunächst wird die Stabilität aufgrund numerischer Begründungen eingeführt. Sei $f \in C([0,1])$ die zu integrierende Funktion. Die Eingabedaten für $Q_n(f)$ sind $\{f(x_{i,n}) : 0 \le i \le n\}$. Sei $\tilde{f}$ die Gleitkomma-Realisierung[18] von f, sodass

$$\tilde{f}_{i,n} := \tilde{f}(x_{i,n}) = f(x_{i,n}) + \delta f_{i,n} \qquad (0 \le i \le n).$$

Es gilt

$$|\delta f_{i,n}| \le \|f - \tilde{f}\|_\infty.$$

Zunächst untersuchen wir die Kondition und finden die Konditionszahl eins:

Bemerkung 2.4.1 *Die Aufgabe $f \mapsto I(f) := \int_0^1 f(x)dx$ ist gut konditioniert. Die Fehlerabschätzung ist*

$$\left|I(f) - I(\tilde{f})\right| \le \|f - \tilde{f}\|_\infty \qquad \textit{für alle } f, \tilde{f} \in C([0,1]). \tag{2.4.1}$$

Beweis. $I(f) - I(\tilde{f}) = \int_0^1 f(x)dx - \int_0^1 \tilde{f}(x)dx = \int_0^1 \left[f(x) - \tilde{f}(x)\right] dx$ und
$\left|I(f) - I(\tilde{f})\right| \le \int_0^1 \left|f(x) - \tilde{f}(x)\right| dx \le \int_0^1 \|f - \tilde{f}\|_\infty dx = \|f - \tilde{f}\|_\infty$. ∎

Eine entsprechende Fehlerabschätzung möchte man für die Quadraturen Q_n erhalten. Für ein festes n rechnet man nach, dass

$$Q_n(\tilde{f}) = \sum_{i=0}^n a_{i,n}\tilde{f}_{i,n} = \sum_{i=0}^n a_{i,n}f(x_{i,n}) + \sum_{i=0}^n a_{i,n}\,\delta f_{i,n} = Q_n(f) + \sum_{i=0}^n a_{i,n}\,\delta f_{i,n}, \text{ also}$$

$$|Q_n(\tilde{f}) - Q_n(f)| = \left|\sum_{i=0}^n a_{i,n}\,\delta f_{i,n}\right| \le \sum_{i=0}^n |a_{i,n}|\,|\delta f_{i,n}| \le \left(\sum_{i=0}^n |a_{i,n}|\right) \|f - \tilde{f}\|_\infty.$$

Dies beweist die Abschätzung

$$|Q_n(\tilde{f}) - Q_n(f)| \le C_n\|f - \tilde{f}\|_\infty \text{ mit} \tag{2.4.2}$$

$$C_n := \sum_{i=0}^n |a_{i,n}|\,. \tag{2.4.3}$$

Da Q_n ein *lineares* Funktional ist, gilt $Q_n(f) - Q_n(\tilde{f}) = Q_n(f - \tilde{f})$. In der Abschätzung $|Q_n(f - \tilde{f})| \le C_n\|f - \tilde{f}\|_\infty$ für alle $f, \tilde{f} \in C([0,1])$ kann man aber $f - \tilde{f}$ durch ein $g \in C([0,1])$ ersetzen und erhält die zu (2.4.2) äquivalente Bedingung

$$|Q_n(g)| \le C_n\|g\|_\infty \qquad \text{für alle } g \in C([0,1]). \tag{2.4.4}$$

Übungsaufgabe 2.4.2 *Man zeige, dass C_n aus (2.4.3) die kleinstmögliche Konstante in (2.4.4) ist.*

2.4.2 Stabilitätsdefinition

Die Konstante C_n ist der Fehlerverstärkungsfaktor der Quadratur Q_n. Um eine wachsende Fehlerverstärkung zu vermeiden, ist es naheliegend zu fordern, dass C_n gleichmäßig beschränkt ist. Dies kann in der folgenden Form geschrieben werden:

$$C_{\text{stab}} := \sup_{n \in \mathbb{N}_0} C_n < \infty. \tag{2.4.5}$$

Damit ist bereits die Stabilität definiert:

Definition 2.4.3 *Eine Quadraturformelfamilie $\{Q_n : n \in \mathbb{N}_0\}$ heißt* stabil, *falls* $\sup_{n\in\mathbb{N}_0} \sum_{i=0}^n |a_{i,n}| < \infty$.

[18] Genaugenommen, ist $\tilde{f}$ nur auf den Maschinenzahlen $\mathcal{M}$ definiert. Für theoretische Zwecke lässt sie sich dazwischen stetig ergänzen.

Hierbei gilt offenbar $\sup_{n\in\mathbb{N}_0}\sum_{i=0}^{n}|a_{i,n}| = C_{\text{stab}}$ mit C_{stab} aus (2.4.5). Eine äquivalente Formulierung enthält

Bemerkung 2.4.4 *a) Eine Familie $\{Q_n : n\in\mathbb{N}_0\}$ von Quadraturformeln ist genau dann stabil, wenn*

$$|Q_n(g)| \le C\|g\|_\infty \qquad \text{für alle } g\in C([0,1]) \text{ und alle } n\in\mathbb{N}_0. \tag{2.4.6}$$

b) Dabei ist C_{stab} aus (2.4.5) die kleinstmögliche Konstante C in (2.4.6).

Beweis. Gemäß Übungsaufgabe 2.4.2 ist für alle $n\in\mathbb{N}_0$ $C_n \le C$ für C aus (2.4.6), da sonst C eine bessere Konstante wäre. Damit gilt auch $C_{\text{stab}} := \sup_{n\in\mathbb{N}_0} C_n \le C$.

Zu a) Wenn daher (2.4.6) (mit endlichem C) gilt, ist (2.4.5) erfüllt, d.h. $\{Q_n\}$ ist stabil.

Zu b) $C_{\text{stab}} \le C$ ist bereits gezeigt. Aus (2.4.4) folgt $|Q_n(g)| \le C_n\|g\|_\infty \le \sup_{n\in\mathbb{N}_0} C_n\|g\|_\infty = C_{\text{stab}}\|g\|_\infty$, d.h. (2.4.6) ist auch mit $C := C_{\text{stab}}$ erfüllt. ∎

Aus Teil b) der Bemerkung und nach Ersetzung von g durch $f-g$ erhält man in Analogie zu (2.4.2):

Korollar 2.4.5 *Es gilt $|Q_n(f)-Q_n(g)| \le C_{\text{stab}}\|f-g\|_\infty$ für alle $f,g\in C([0,1])$ und alle $n\in\mathbb{N}_0$.*

Im Weiteren diskutieren wir die Folgen der Stabilität bzw. Instabilität im numerischen Kontext. Der Gesamtfehler $Q_n(\tilde f) - \int_0^1 f(x)dx$ wird mit der Dreiecksungleichung abgeschätzt:

$$\left|Q_n(\tilde f) - \int_0^1 f(x)dx\right| \le \left|Q_n(\tilde f) - Q_n(f)\right| + \left|Q_n(f) - \int_0^1 f(x)dx\right| \le C_n\|\tilde f - f\|_\infty + \varepsilon_n(f).$$

Im Falle der Stabilität ist der Fehler $\le C_{\text{stab}}\|\tilde f - f\|_\infty + \varepsilon_n(f)$ (vgl. Korollar 2.4.5). Unter der Annahme $\varepsilon_n(f)\to 0$ nähert sich der Gesamtfehler dem "Rauschpegel" $C_{\text{stab}}\|\tilde f - f\|_\infty$, der durch die Eingabefehler $\|\tilde f - f\|_\infty$ verursacht wird und unvermeidlich ist.

Im Falle der Instabilität gilt dagegen $C_n\to\infty$. Während der Summand $\varepsilon_n(f)$ gegen null fällt, steigt $C_n\|\tilde f - f\|_\infty$ gegen unendlich. Damit ist die Vergrößerung von n keine Garantie für ein besseres Ergebnis. Wenn man keine weitere Information über das Verhalten von $\varepsilon_n(f)$ besitzt, ist es schwierig, ein n zu finden, sodass der Gesamtfehler möglichst klein ist.

2.4.3 Stabilität konkreter Quadraturformeln

Unter der minimalen Bedingung, dass Q_n für Polynome nullter Ordnung exakt ist (d.h. $g(n)\ge 0$ in (2.2.1)), folgt $1 = \int_0^1 dx = Q_n(1) = \sum_{i=0}^n a_{i,n}$:

$$\sum_{i=0}^{n} a_{i,n} = 1. \tag{2.4.7}$$

Folgerung 2.4.6 *Es gelte (2.4.7). a) Falls $a_{i,n}\ge 0$ für alle i,n, ist die Familie $\{Q_n\}$ stabil mit $C_{\text{stab}} = C_n = 1$ für alle n.*

b) Stets gilt $C_{\text{stab}}\ge 1$.

Beweis. a) Wegen $\sum_{i=0}^n |a_{i,n}| = \sum_{i=0}^n a_{i,n} = 1$. b) Wegen $\sum_{i=0}^n |a_{i,n}| \ge |\sum_{i=0}^n a_{i,n}| = |1| = 1$. ∎

Die vorhergehenden Überlegungen beruhen auf (2.4.7). Eine Abschwächung enthält die

Übungsaufgabe 2.4.7 *Folgerung 2.4.6 bleibt richtig, wenn (2.4.7) durch die Konvergenz $Q_n(1)\to 1$ ersetzt wird.*

Da in der Numerik-Vorlesung[19] $a_{i,n}\ge 0$ für die Gauß-Quadratur bewiesen wird, folgt: **Die Familie der Gauß-Quadraturen ist stabil** (und $C_{\text{stab}} = 1$).

Die Newton-Cotes-Formeln erfüllen $a_{i,n}\ge 0$ nur für $n\in\{1,2,3,4,5,6,7,9\}$. Allerdings folgt aus der Existenz von negativen $a_{i,n}\ge 0$ noch keineswegs die Instabilität. Die folgende Tabelle zeigt die Werte von C_n aus (2.4.3):

[19] Vgl. Kapitel 3.6 in Stoer: *Einführung in die Numerische Mathematik I.* Springer-Verlag, Berlin, 8. Auflage, 1999

n	$1,\ldots,7$	8	9	10	11	12	14	16	18	20	22	24
C_n	1	1.45	1	3.065	1.589	7.532	20.34	58.46	175.5	544.2	1606	9923

Offenbar steigt C_n exponentiell gegen unendlich, d.h. die Newton-Cotes-Formeln sind nach allem Anschein instabil. Ein exakter Beweis der Instabilität kann die folgende asymptotische Aussage[20] verwenden:

$$a_{i,n} = \frac{(-1)^{i-1}\, n!}{i!\,(n-i)!n^2 \log^2 n} \left(\frac{1}{i} + \frac{(-1)^n}{n-i}\right) \left(1 + \mathcal{O}\left(\frac{1}{\log n}\right)\right) \text{ für } 1 \le i \le n-1.$$

Offenbar gilt für ein gerades n die Ungleichung

$$C_n = \sum_{i=0}^{n} |a_{i,n}| \ge \left|a_{\frac{n}{2},n}\right|.$$

Übungsaufgabe 2.4.8 *a) Man schlage die Stirlingsche*[21] *Formel zur asymptotischen Darstellung von $n!$ nach.*
b) Unter Verwendung von a) untersuche man das Verhalten von $\left|a_{\frac{n}{2},n}\right|$ und schließe daraus, dass ***die Familie der Newton-Cotes-Formeln instabil ist.***

Ein weiteres Beispiel folgt im nächsten Unterkapitel.

2.4.4 Romberg-Quadratur

Die Existenz negativer Gewichte $a_{i,n}$ ist noch kein Grund zur Instabilität, solange $\sum_{i=0}^{n} |a_{i,n}|$ gleichmäßig beschränkt bleibt. Dass es Verfahren gibt, die tatsächlich negative Gewichte aufweisen und stabil sind, wird anhand der Romberg-Quadratur[22] gezeigt.

Für $h = 1/N$ mit $N \in \mathbb{N}$ definiert

$$T(f,h) := h\left[\frac{1}{2}f(0) + f(h) + f(2h) + \ldots + f(1-h) + \frac{1}{2}f(1)\right]$$

die *summierte Trapezformel*. Unter der Voraussetzung $f \in C^m([0,1])$, m gerade, kann man die *asymptotische Entwicklung*

$$T(f,h) = \int_0^1 f(x)dx + h^2 e_2(f) + \ldots + h^{m-2} e_{m-2}(f) + \mathcal{O}(h^m \|f^{(m)}\|_\infty) \tag{2.4.8}$$

beweisen[23]. Daher ist die Richardson[24]-Extrapolation anwendbar: Man berechnet $T(f,h_i)$ für verschiedene h_i, $i = 0,\ldots,n$, und extrapoliert $\left\{\left(h_i^2, T(f,h_i)\right) : i = 0,\ldots,n\right\}$ auf $h = 0$. Das Resultat lässt sich mit Hilfe der Lagrange-Polynome explizit darstellen (vgl. Übungsaufgabe 3.1.2):

$$Q_n(f) := \sum_{i=0}^{n} T(f,h_i) \underbrace{\prod_{\substack{\nu=0 \\ \nu\neq i}}^{n} \frac{h_\nu^2}{h_\nu^2 - h_i^2}}_{=:c_{i,n}} = \sum_{i=0}^{n} c_{i,n} T(f,h_i). \tag{2.4.9}$$

Wir fixieren eine unendliche Schrittweitenfolge

$$h_0 > h_1 > h_2 > \ldots, \qquad h_i = 1/N_i, \quad N_i \in \mathbb{N},$$

mit der Eigenschaft

$$h_{i+1} \le \alpha h_i \quad \text{mit } 0 < \alpha < 1 \text{ für alle } i \ge 0. \tag{2.4.10}$$

Die eigentliche Romberg-Quadratur fordert $\alpha = 1/2$. Die Bedingung (2.4.10) erzwingt $h_i \to 0$ für $i \to \infty$. Man folgert aus (2.4.10), dass

$$h_i/h_j \le \alpha^{i-j} \text{ für } j \le i \qquad \text{und} \quad h_i/h_j \ge \alpha^{i-j} \text{ für } j \ge i. \tag{2.4.11}$$

[20] Zu finden auf Seite 64 in: Davis - Rabinowitz: *Methods of numerical integration.* Academic Press, New York, 1975
[21] James Stirling, geb. Mai 1692 in Garden (Schottland), gest. 5. Dez. 1770 in Edinburgh
[22] Werner Romberg, geb. 16. Mai 1909 in Berlin, gest. 20. Febr. 2003 in Heidelberg
[23] R. Bulirsch: Bemerkungen zur Romberg-Integration. Numer. Math. **6** (1964) 6-16
[24] Lewis Fry Richardson, geb. 11. Okt. 1881 in Newcastle upon Tyne, gest. 30. Sept. 1953 in Kilmun (Schottland)

Lemma 2.4.9 *Es gibt ein $B < \infty$, sodass $\sum_{i=0}^{n} |c_{i,n}| \le B$ für alle $n \in \mathbb{N}_0$.*

Übungsaufgabe 2.4.10 *Man zeige a) $1 + x \le e^x$ für alle reellen x, b) $\frac{1}{1-x} \le 1 + \vartheta x$ mit $\vartheta = \frac{1}{1-x_0}$ für alle $0 \le x \le x_0 < 1$.*

Beweis von Lemma 2.4.9. a) Übungsaufgabe 2.4.10 zeigt mit $\vartheta = \frac{1}{1-\alpha^2}$, dass

$$\prod_{j=1}^{m} \frac{1}{1-\alpha^{2j}} \le \prod_{j=1}^{m} \left(1 + \vartheta\alpha^{2j}\right) \le \prod_{j=1}^{m} \exp\left(\vartheta\alpha^{2j}\right) \le \prod_{j=1}^{\infty} \exp\left(\vartheta\alpha^{2j}\right) =: A,$$

wobei $A = \exp\left(\sum_{j=1}^{\infty} \vartheta\alpha^{2j}\right) = \exp\frac{\alpha^2\vartheta}{1-\alpha^2} = \exp\left(\alpha^2\vartheta^2\right)$.

b) Damit gilt auch $\prod_{j=1}^{m} \frac{\alpha^{2j}}{1-\alpha^{2j}} \le A \prod_{j=1}^{m} \alpha^{2j} = A\alpha^{m(m+1)} \le A\alpha^m$ für alle $m \ge 0$.

c) Das Produkt $\prod_{\nu\neq i}$ in (2.4.9) wird in die Teilprodukte $\prod_{\nu=0}^{i-1}$ und $\prod_{\nu=i+1}^{n}$ aufgespalten. Für das erste gilt

$$\left|\prod_{\nu=0}^{i-1} \frac{h_\nu^2}{h_\nu^2 - h_i^2}\right| = \prod_{\nu=0}^{i-1} \frac{1}{1 - h_i^2/h_\nu^2} \underset{(2.4.11)}{\le} \prod_{\nu=0}^{i-1} \frac{1}{1-\alpha^{2(i-\nu)}} = \prod_{j=1}^{i} \frac{1}{1-\alpha^{2j}} \underset{\text{Teil a)}}{\le} A,$$

das zweite erfüllt

$$\left|\prod_{\nu=i+1}^{n} \frac{h_\nu^2}{h_\nu^2 - h_i^2}\right| = \prod_{\nu=i+1}^{n} \frac{1}{h_i^2/h_\nu^2 - 1} \underset{(2.4.11)}{\le} \prod_{\nu=i+1}^{n} \frac{1}{\alpha^{2(i-\nu)} - 1} = \prod_{j=1}^{n-i} \frac{1}{\alpha^{-2j} - 1} = \prod_{j=1}^{n-i} \frac{\alpha^{2j}}{1-\alpha^{2j}} \underset{\text{Teil b)}}{\le} A\alpha^{n-i}.$$

d) Die Abschätzung

$$\sum_{i=0}^{n} |c_{i,n}| = \sum_{i=0}^{n} \left|\prod_{\nu=0}^{i-1} \frac{h_\nu^2}{h_\nu^2 - h_i^2}\right| \times \left|\prod_{\nu=i+1}^{n} \frac{h_\nu^2}{h_\nu^2 - h_i^2}\right| \underset{\text{Teil c)}}{\le} A^2 \sum_{i=0}^{n} \alpha^{n-i} < A^2 \sum_{j=0}^{\infty} \alpha^j = \frac{A^2}{1-\alpha} =: B$$

beweist die Behauptung. ∎

Die von Q_n verwendeten Stützstellen $\{x_{j,n}\}$ sind $\bigcup_{i=0}^{n}\{0, h_i, 2h_i, \ldots, 1\}$. Wählt man die $h_i = 1/N_i$ so, dass N_i kein Teiler von N_{i+1}, ist $a_{j,n} = \frac{1}{2}h_{n-1}c_{n-1,n}$ das Gewicht zur Stützstelle $x_{j,n} = h_{n-1}$. Wegen $\operatorname{sign}(c_{i,n}) = (-1)^{n-i}$ folgt, dass Q_n negative Gewichte enthält.

Lemma 2.4.11 *Die Familie der Quadraturen $\{Q_n\}$ aus (2.4.9) ist stabil.*

Beweis. Die summierte Trapezformel $T(f, h_i) = \sum_{k=0}^{N_i} \tau_{k,i} f(kh_i)$ hat die Gewichte $\tau_{k,i} = h_i$ für $0 < k < N_i$ und $\tau_{k,i} = h_i/2$ für $k = 0, N_i$. Insbesondere gilt $\sum_{k=0}^{N_i} |\tau_{k,i}| = \sum_{k=0}^{N_i} \tau_{k,i} = 1$. Die Quadraturformel Q_n ist definiert als $Q_n(f) = \sum_i c_{i,n} \sum_{k=0}^{N_i} \tau_{k,i} f(kh_i) = \sum_j a_{j,n} f(x_{j,n})$ mit $a_{j,n} := \sum_{(i,k):kh_i = x_{j,n}} c_{i,n}\tau_{k,i}$.

$$\sum_j |a_{j,n}| \le \sum_i |c_{i,n}| \sum_{k=0}^{N_i} |\tau_{k,i}| = \sum_i |c_{i,n}| \underset{\text{Lemma 2.4.9}}{\le} B$$

beweist die Stabilitätsbedingung. ∎

Lemma 2.4.12 *Die Familie der Quadraturen $\{Q_n\}$ aus (2.4.9) ist konsistent.*

Beweis. Sei P ein Polynom vom Grad $\le g(n) := 2n+1$. In (2.4.8) mit $m := 2n+2$ verschwindet der Restterm, da $P^{(m)} = 0$. Damit ist $T(f, h)$ ein Polynom vom Grad $\le n$ in der Variablen h^2. Die Extrapolation tilgt die Terme $h_i^j e_j(P)$, $j = 2, 4, \ldots, 2n$, sodass $Q_n(P) = \int_0^1 P(x)dx$. Da $g(n) = 2n + 1 \to \infty$ für $n \to \infty$, ist die Konsistenz gemäß Definition 2.2.1 gezeigt. ∎

Stabilität und Konsistenz beweisen die Konvergenz der Romberg-Quadratur.

Übungsaufgabe 2.4.13 *Die Bedingung $h_{i+1} \le \alpha h_i$ in (2.4.10) kann abgeschwächt werden. Man zeige: Wenn ein $\ell \in \mathbb{N}$ und ein $\alpha \in (0, 1)$ existieren, sodass $h_{i+\ell} \le \alpha h_i$ für alle $i \ge 0$, so gilt Lemma 2.4.9 ebenfalls.*

2.5 Approximationssatz von Weierstraß

Für den nächsten Beweisschritt benötigen wir den folgenden Satz:

Satz 2.5.1 (Approximationssatz von Weierstraß) [25] *Für alle $\varepsilon > 0$ und alle $f \in C([0,1])$ existiert ein Polynom $P = P_{\varepsilon,f}$ mit*

$$\|f - P\|_\infty \leq \varepsilon.$$

Eine äquivalente Formulierung lautet: Die Menge $\mathcal{P} := \{P : \text{Polynom beschränkt auf } [0,1]\}$ ist eine *dichte Teilmenge* von $C([0,1])$.

Im folgenden beweisen wir eine allgemeinere Form des Weierstraßschen Approximationssatzes. Das nächste Lemma benutzt das punktweise Maximum bzw. Minimum $\mathrm{Max}(f,g)(x) := \max(f(x), g(x))$, $\mathrm{Min}(f,g)(x) := \min(f(x), g(x))$ zweier Funktionen. Die Bedingung i) beschreibt, dass $\mathcal{F}$ unter Maximum- und Minimumbildung abgeschlossen ist. Bedingung ii) ist die Approximierbarkeit in zwei Punkten.

Lemma 2.5.2 *Sei $Q \subset \mathbb{R}^d$ kompakt. Sei $\mathcal{F} \subset C(Q)$ eine Familie von stetigen Funktionen mit den folgenden beiden Eigenschaften:*

i) *Für alle $f_1, f_2 \in \mathcal{F}$ gilt $\mathrm{Max}(f_1, f_2), \mathrm{Min}(f_1, f_2) \in \mathcal{F}$.*

ii) *Zu allen $x', x'' \in Q$, allen $\varepsilon > 0$ und allen $g \in C(Q)$ gilt es ein $\varphi \in \mathcal{F}$ mit $|\varphi(x') - g(x')| < \varepsilon$ und $|\varphi(x'') - g(x'')| < \varepsilon$*

Dann gibt es zu jedem $\varepsilon > 0$ und jedem $g \in C(Q)$ eine Funktion $f \in \mathcal{F}$ mit $\|f - g\|_\infty < \varepsilon$ (d.h. $\mathcal{F}$ liegt dicht in $C(Q)$).

Beweis. a) Seien $\varepsilon > 0$ und $g \in C(Q)$ gegeben. Ferner sei $x' = x_0 \in Q$ fest vorgegeben, während der zweite Punkt $x'' = y \in Q$ im Folgenden variabel gehalten wird. Nach Voraussetzung ii) existiert eine Funktion $h = h(\cdot\,; x_0, y, \varepsilon)$ mit

$$|h(x_0) - g(x_0)| < \varepsilon, \qquad |h(y) - g(y)| < \varepsilon.$$

Aus der letzten Ungleichung folgt insbesondere $g(y) - \varepsilon < h(y)$. Wegen der Stetigkeit von h, g gilt diese Ungleichung in einer gesamten Umgebung $U(y)$ von y :

$$g(x) - \varepsilon < h(x) \qquad \text{für alle } x \in U(y).$$

Für jedes $y \in Q$ existiert eine Funktion $h = h(\cdot\,; x_0, y, \varepsilon)$, die diese Ungleichung in einer bestimmten Umgebung erfüllt. Da $\bigcup_{y \in Q} U(y)$ die kompakte Menge Q überdeckt, kann man eine endliche Teilmenge $\{U(y_i) : i = 1, \ldots, n\}$ auswählen, die Q ebenfalls überdeckt:

$$\bigcup_{i=1,\ldots,n} U(y_i) \supset Q.$$

Zu jedem y_i gehört eine Funktion $h(\cdot\,; x_0, y_i, \varepsilon) \in \mathcal{F}$ mit

$$g(x) - \varepsilon < h(x; x_0, y_i, \varepsilon) \qquad \text{für alle } x \in U(y_i).$$

Nach Voraussetzung i) gehört $h(\cdot\,; x, \varepsilon) := \mathrm{Max}_{i=1,\ldots,n} h(\cdot\,; x_0, y_i, \varepsilon)$ wieder zu $\mathcal{F}$ und erfüllt

$$g(x) - \varepsilon < h(x; x_0, \varepsilon) \qquad \text{für alle } x \in Q.$$

b) Nun wird auch der Parameter x_0 zu einer Variablen in Q gemacht. Da alle $h(\cdot\,; x_0, y_i, \varepsilon)$ die Funktion g auch in x_0 approximieren, gilt $g(x_0) + \varepsilon > h(x_0; x_0, \varepsilon)$. Wieder gibt es eine Umgebung $V(x_0)$, sodass

$$g(x) + \varepsilon > h(x; x_0, \varepsilon) \qquad \text{für alle } x \in V(x_0).$$

Wie im Teil a) findet man eine endliche Überdeckung $\{V(x_i) : i = 1, \ldots, m\}$ von Q. Die Funktion

$$f := \mathrm{Min}_{i=1,\ldots,m} h(\cdot\,; x_i, \varepsilon)$$

gehört wieder zu $\mathcal{F}$ und erfüllt $g + \varepsilon > f$. Da jedes $h(\cdot\,; x_i, \varepsilon)$ der Ungleichung $g - \varepsilon < h(\cdot\,; x_i, \varepsilon)$ aus Teil a) genügt, folgt auch $g - \varepsilon < f$. Zusammen erhält man $\|f - g\|_\infty < \varepsilon$, d.h. $f \in \mathcal{F}$ ist die gesuchte Approximierende. ∎

[25] Karl Theodor Wilhelm Weierstraß, geb. 31. Okt. 1815 in Ostenfelde, gest. 19. Febr. 1897 in Berlin

Bemerkung 2.5.3 *Anstelle des Abschlusses bezüglich* Max *und* Min *kann man äquivalent fordern, dass $\mathcal{F}$ bezüglich der Absolutbildung abgeschlossen ist:*

$$f \in \mathcal{F} \quad \Rightarrow \quad |f| \in \mathcal{F},$$

wobei $|f|$ punktweise definiert ist: $|f|(x) := |f(x)|$ für alle $x \in Q$.

Beweis. Man beachte $\mathrm{Max}(f,g) = \frac{1}{2}(f+g)+\frac{1}{2}|f-g|$ und $\mathrm{Min}(f,g) = \frac{1}{2}(f+g)-\frac{1}{2}|f-g|$ bzw. in umgekehrter Richtung $|f| = \mathrm{Max}(f,-f)$. ∎

Die Addition und Multiplikation von Funktionen sei punktweise definiert: $(f+g)(x) = f(x)+g(x)$, $(f \cdot g)(x) = f(x)g(x)$, $x \in Q$. Entsprechend ist die Multiplikation mit Zahlen aus dem Körper $\mathbb{K}$ punktweise definiert: $(\lambda f)(x) = \lambda f(x)$.

Definition 2.5.4 (Algebra von Funktionen) *Eine Menge $\mathcal{A}$ von Funktionen heißt eine Algebra, falls sie (ohne die Multiplikation) ein Vektorraum über $\mathbb{K}$ ist und zusätzlich die Multiplikation erlaubt.*

Beispiel 2.5.5 *Algebren werden gebildet von allen a) stetigen Funktionen auf $Q \subset \mathbb{R}^d$ (keine Kompaktheit vorausgesetzt), b) beschränkten Funktionen auf $Q \subset \mathbb{R}^d$, c) Polynomen, d) trigonometrischen Funktionen.*

Im Falle von d) hat man zu zeigen, dass z.B. das Produkt $\sin(nx)\cos(mx)$ $(n,m \in \mathbb{N}_0)$ wieder eine trigonometrische Funktion ist. Dies folgt aber aus $2\sin(nx)\cos(mx) = \sin((n+m)x) + \sin((n-m)x)$.

Ist $\mathcal{A} \subset C(Q)$ eine Algebra, so wird der Abschluss $\bar{\mathcal{A}}$ (bezüglich der Maximumnorm $\|\cdot\|_\infty$) als *abgeschlossene Hülle der Algebra $\mathcal{A}$* bezeichnet.

Übungsaufgabe 2.5.6 *Mit $\mathcal{A}$ ist auch $\bar{\mathcal{A}}$ eine Algebra stetiger Funktionen, d.h. $f,g \in \bar{\mathcal{A}}$ impliziert $f+g \in \bar{\mathcal{A}}$ und $f \cdot g \in \bar{\mathcal{A}}$.*

Lemma 2.5.7 (Weierstraß) *Sei $\mathcal{A} \subset C(Q)$ eine Algebra. Dann gilt für alle $f \in \mathcal{A}$, dass $|f| \in \bar{\mathcal{A}}$.*

Beweis. a) Eine einfache Skalierungsüberlegung ergibt, dass es reicht, die Behauptung für $f \in \mathcal{A}$ mit $\|f\|_\infty \leq 1$ zu zeigen.

b) Sei $\varepsilon > 0$ gegeben. Die Funktion $T(\zeta) := \sqrt{\zeta + \varepsilon^2}$ ist in der komplexen Halbebene $\Re e\zeta > -\varepsilon^2$ holomorph. Die Taylor-Reihe $\sum \alpha_\nu \left(x - \frac{1}{2}\right)^\nu$ für $T(x)$ hat den Konvergenzradius $\frac{1}{2} + \varepsilon^2$ und konvergiert im Intervall $[0,1]$ gleichmäßig. Es gibt daher ein endliches Taylor-Polynom P_n vom Grad n, sodass

$$\left|\sqrt{x+\varepsilon^2} - P_n(x)\right| \leq \varepsilon \qquad \text{für alle } 0 \leq x \leq 1.$$

c) Ersetzen von x durch x^2 liefert

$$\left|\sqrt{x^2+\varepsilon^2} - P_n(x^2)\right| \leq \varepsilon \qquad \text{für alle } -1 \leq x \leq 1.$$

Der Spezialfall $x = 0$ zeigt $|\varepsilon - P_n(0)| \leq \varepsilon$ und damit $|P_n(0)| \leq 2\varepsilon$. Das Polynom $Q_{2n}(x) := P_n(x^2) - P_n(0)$ vom Grad $2n$ hat einen verschwindenden absoluten Term und erfüllt

$$\left|\sqrt{x^2+\varepsilon^2} - Q_{2n}(x)\right| \leq \left|\sqrt{x^2+\varepsilon^2} - P_n(x^2)\right| + |P_n(0)| \leq 3\varepsilon \qquad \text{für alle } -1 \leq x \leq 1.$$

Mit

$$\left|\sqrt{x^2+\varepsilon^2} - |x|\right| = \frac{\varepsilon^2}{\left|\sqrt{x^2+\varepsilon^2} + |x|\right|} \leq \frac{\varepsilon^2}{\varepsilon} = \varepsilon \qquad \text{für alle } -1 \leq x \leq 1$$

ergibt sich schließlich die Ungleichung

$$\left||x| - Q_{2n}(x)\right| \leq 4\varepsilon \qquad \text{für alle } -1 \leq x \leq 1.$$

d) Für f mit $\|f\|_\infty \leq 1$ erfüllt $f(\xi)$ $(\xi \in Q)$ die Ungleichung $-1 \leq f(\xi) \leq 1$, sodass in der vorhergehenden Ungleichung $x = f(\xi)$ gesetzt werden kann:

$$\left||f(\xi)| - Q_{2n}(f(\xi))\right| \leq 4\varepsilon \qquad \text{für alle } \xi \in Q.$$

Wegen[26] $Q_{2n}(f(\xi)) = \sum_{\nu=1}^{n} q_\nu \left(f(\xi)\right)^{2\nu} = \left(\sum_{\nu=1}^{n} q_\nu f^{2\nu}\right)(\xi)$, gehört $Q_{2n}(f)$ wieder zu $\mathcal{A}$ und erfüllt die Abschätzung $||f| - Q_{2n}(f)| \leq 4\varepsilon$. Da $\varepsilon > 0$ beliebig, liegt $|f|$ im Abschluss von $\mathcal{A}$. ∎

Wir beweisen nun den Satz von Weierstraß in der verallgemeinerten Form von Stone[27]:

Satz 2.5.8 (Stone - Weierstraß) *Vorausgesetzt seien*

i) $Q \subset \mathbb{R}^d$ *sei kompakt,*

ii) $\mathcal{A}$ *sei eine Algebra von auf* Q *stetigen Funktionen (d.h.* $\mathcal{A} \subset C(Q)$*),*

iii) $\mathcal{A}$ *trenne die Punkte von* Q*, d.h. zu jedem Punktepaar* $x', x'' \in Q$ *mit* $x' \neq x''$ *existiert ein* $f \in \mathcal{A}$ *mit* $f(x') \neq f(x'')$*.*

Dann gilt für die abgeschlossene Hülle $\bar{\mathcal{A}}$ *entweder* $\bar{\mathcal{A}} = C(Q)$ *oder es gibt ein* $x_0 \in Q$*, sodass*

$$\bar{\mathcal{A}} = \{f \in C(Q) : f(x_0) = 0\}. \tag{2.5.1}$$

Beweis. a) Nach Lemma 2.5.7 und Bemerkung 2.5.3 erfüllt $\mathcal{F} = \bar{\mathcal{A}}$ die Voraussetzung i) aus Lemma 2.5.2. Sobald auch ii) aus Lemma 2.5.2 gezeigt ist, folgt $\bar{\mathcal{F}} = C(Q)$. Da $\mathcal{F} = \bar{\mathcal{A}}$ schon abgeschlossen ist, wäre der erste Fall $\bar{\mathcal{A}} = C(Q)$ bewiesen.

b) Zum Beweis von ii) aus Lemma 2.5.2 betrachten wir die folgende Alternative: *Entweder* gibt es zu jedem $x \in Q$ ein $f \in \mathcal{A}$ mit $f(x) \neq 0$ *oder* es existiert ein $x_0 \in Q$ mit $f(x_0) = 0$ für alle $f \in \mathcal{A}$. Die erste Alternative wird in Teil c) untersucht, die zweite in Teil d).

c) Die erste Alternative sei vorausgesetzt. Wir werden zunächst beweisen:
Zwischenbehauptung: Zu Punkten $x', x'' \in Q$ mit $x' \neq x''$ aus Voraussetzung ii) von Lemma 2.5.2 existiert ein $f \in \mathcal{A}$ mit $0 \neq f(x') \neq f(x'') \neq 0$.
Beweis der Zwischenbehauptung. Die Trennungseigenschaft iii) aus der Voraussetzung zeigt $f(x') \neq f(x'')$ für ein geeignetes f. Sei $f(x') = 0$ angenommen, was $f(x'') \neq 0$ impliziert (Der Fall $f(x'') = 0$ ist völlig analog). Die erste Alternative aus Teil b) garantiert die Existenz eines $f_0 \in \mathcal{A}$ mit $f_0(x') \neq 0$. Die Funktion

$$f_\lambda := f - \lambda f_0$$

hat die Eigenschaften

$$\begin{aligned} 0 \neq f_\lambda(x'') &\quad \text{für genügend kleine } \lambda, && (\text{wegen } f(x'') \neq 0) \\ f_\lambda(x'') \neq f_\lambda(x') &\quad \text{für genügend kleine } \lambda, && (\text{wegen } f(x') \neq f(x'')) \\ f_\lambda(x') = \lambda f_0(x') \neq 0 &\quad \text{für alle } \lambda \neq 0. \end{aligned}$$

Damit gilt für hinreichend kleines, aber positives λ

$$0 \neq f_\lambda(x') \neq f_\lambda(x'') \neq 0, \qquad f_\lambda \in \mathcal{A}.$$

Indem wir f aus der Zwischenbehauptung durch f_λ ersetzen, ist diese beweisen. ∎

Sei $g \in C(Q)$ die Funktion aus Voraussetzung ii) von Lemma 2.5.2. Für das dort geforderte φ machen wir den Ansatz $\varphi = \alpha f + \beta f^2$ mit f aus der Zwischenbehauptung. Mit f gehören auch f^2 und φ zu $\mathcal{A}$. Aus den Bedingungen $\varphi(x') = g(x')$ und $\varphi(x'') = g(x'')$ ergibt sich ein 2×2-Gleichungssystem für α, β. Da die Determinante $f(x')f^2(x'') - f(x'')f^2(x') = f(x')f(x'')\left[f(x'') - f(x')\right]$ von null verschieden ist, ist die Lösbarkeit gesichert. Damit ist die Voraussetzung ii) von Lemma 2.5.2 sogar für $\varepsilon = 0$ erfüllt und Teil a) zeigt $\bar{\mathcal{A}} = C(Q)$.

d) Die zweite Alternative sei angenommen: Es existiert ein $x_0 \in Q$ mit $f(x_0) = 0$ für alle $f \in \mathcal{A}$. Sei $\mathbb{1} \in C(Q)$ die Funktion mit dem konstanten Wert 1. Sei $\mathcal{A}^*$ die von $\mathcal{A}$ und $\{\mathbb{1}\}$ erzeugte Algebra: Sie lautet explizit $\mathcal{A}^* = \{f = g + \lambda\mathbb{1}\colon g \in \mathcal{A}, \lambda \in \mathbb{K}\}$. Offenbar existiert zu jedem $x \in Q$ ein $f \in \mathcal{A}^*$ mit $f(x) \neq 0$ (nämlich $f = \mathbb{1}$). Damit trifft auf $\mathcal{A}^*$ die erste Alternative zu. Der bisherige Beweis zeigt $\overline{\mathcal{A}^*} = C(Q)$.

[26] Hier wird wesentlich ausgenutzt, dass $q_0 = 0$, da $f^0 = 1$ nicht notwendigerweise der Algebra $\mathcal{A}$ angehören muss.
[27] Marshall Harvey Stone, geb. 8. April 1903 in New York, gest. 9. Jan. 1989 in Madras

Sei $g \in C(Q)$ eine beliebige Funktion mit $g(x_0) = 0$, d.h. aus der rechten Seite von (2.5.1). Wegen $g \in C(Q) = \overline{\mathcal{A}^*}$ existiert zu jedem $\varepsilon > 0$ ein $f^* \in \mathcal{A}^*$ mit $\|g - f^*\|_\infty < \varepsilon$. Nach Definition von $\mathcal{A}^*$ lässt sich f^* als $f^* = f + \lambda \mathbb{1}$ mit $f \in \mathcal{A}$ schreiben. Damit gilt

$$\|g - f - \lambda \mathbb{1}\|_\infty < \varepsilon.$$

Insbesondere gilt bei x_0, dass $|g(x_0) - f(x_0) - \lambda| = |\lambda| < \varepsilon$. Zusammen erhält man $\|g - f\|_\infty < 2\varepsilon$, wobei $f \in \mathcal{A}$. Dies beweist (2.5.1). ∎

Für den Beweis des Satzes 2.5.1 setze man $Q = [0,1]$ (kompakte Teilmenge des $\mathbb{R}^1$) und $\mathcal{A}$ als Algebra aller Polynome. Für diese Algebra ist die Voraussetzung iii) aus Satz 2.5.8 mit $f(x) = x$ erfüllt. Also muss eine der beiden Alternativen $\bar{\mathcal{A}} = C(Q)$ oder (2.5.1) gelten. Da die konstante Funktion $\mathbb{1}$ zu $\mathcal{A}$ gehört, ist (2.5.1) ausgeschlossen und $\bar{\mathcal{A}} = C(Q)$ gezeigt.

2.6 Konvergenzbeweis

An die Definition 2.3.2 sei erinnert: Q_n ist konvergent, falls für jedes $f \in C([0,1])$ gilt, dass $Q_n(f) \to \int_0^1 f(x)dx$. Bisher war noch offen, ob diese Eigenschaft erfüllbar ist. Das positive Resultat enthält der folgende Satz, der dem Muster

$$\text{Konsistenz} + \text{Stabilität} \implies \text{Konvergenz} \tag{2.6.1}$$

folgt.

Satz 2.6.1 (Konvergenzsatz) *Ist die Familie $\{Q_n : n \in \mathbb{N}_0\}$ von Quadraturformeln konsistent und stabil, so ist sie auch konvergent.*

Beweis. Sei $\varepsilon > 0$ gegeben. Zu zeigen ist, dass für alle $f \in C([0,1])$ ein n_0 existiert, sodass $\left|Q_n(f) - \int_0^1 f(x)dx\right| \leq \varepsilon$ für $n \geq n_0$. Mit einem beliebigen Polynom P gilt dank der Dreiecksungleichung

$$\begin{aligned}\left|Q_n(f) - \int_0^1 f(x)dx\right| &= \left|Q_n(f) - Q_n(P) + Q_n(P) - \int_0^1 P(x)dx + \int_0^1 P(x)dx - \int_0^1 f(x)dx\right| \\ &\leq |Q_n(f) - Q_n(P)| + \left|Q_n(P) - \int_0^1 P(x)dx\right| + \left|\int_0^1 P(x)dx - \int_0^1 f(x)dx\right|.\end{aligned}$$

Wir wählen P gemäß Satz 2.5.1 so, dass $\|f - P\|_\infty \leq \varepsilon / (1 + C_{\text{stab}})$, wobei C_{stab} die Stabilitätskonstante sei. Nun kann der erste Summand $|Q_n(f) - Q_n(P)|$ laut Korollar 2.4.5 durch $C_{\text{stab}} \|f - P\|_\infty \leq \varepsilon C_{\text{stab}} / (1 + C_{\text{stab}})$ abgeschätzt werden.

Mit P ist auch $\text{grad}(P)$ fixiert. Wegen $g(n) \to \infty$ gibt es ein n_0, sodass $g(n) \geq \text{grad}(P)$ für alle $n \geq n_0$. Die Konsistenz garantiert damit Exaktheit der Quadratur: $\left|Q_n(P) - \int_0^1 P(x)dx\right| = 0$.

Bemerkung 2.4.1 liefert $\left|\int_0^1 P(x)dx - \int_0^1 f(x)dx\right| \leq \|f - P\|_\infty \leq \varepsilon / (1 + C_{\text{stab}})$ für den letzten Summanden.

Damit ist die Summe der drei Terme beschränkt durch $\frac{\varepsilon C_{\text{stab}}}{1+C_{\text{stab}}} + \frac{\varepsilon}{1+C_{\text{stab}}} = \varepsilon$. ∎

Gemäß Satz 2.6.1 ist die Stabilität hinreichend für Konvergenz. Als Nächstes soll gezeigt werden, dass die Stabilitätsbedingung (2.4.5) auch *notwendig* für die Konvergenz ist. Hierfür benötigen wir einen weiteren Satz der Funktionalanalysis.

2.7 Satz von der gleichmäßigen Beschränktheit

Zunächst sei an einige Notationen erinnert.

X heißt *normierter Raum* (in voller Schreibweise wäre X durch $(X, \|\cdot\|)$ zu ersetzen), falls auf dem Vektorraum X eine Norm $\|\cdot\|$ definiert ist. Im Zweifelsfall wird diese Norm als $\|\cdot\|_X$ notiert.

X heißt *Banach-Raum*, falls X normiert und vollständig ist ("vollständig" bedeutet, dass alle Cauchy-Folgen in X konvergieren).

Mit $L(X,Y)$ wird die Menge der linearen und stetigen Abbildung von X nach Y bezeichnet.

Bemerkung 2.7.1 *a) Sind X, Y normiert, so ist auch $L(X,Y)$ normiert. Die zugehörige "Operatornorm" eines $T \in L(X,Y)$ lautet*

$$\|T\| := \sup_{x \in X \setminus \{0\}} \frac{\|Tx\|_Y}{\|x\|_X}. \tag{2.7.1}$$

b) Die definitionsgemäß stetigen $T : X \to Y$ aus $L(X,Y)$ führen immer auf ein endliches Supremum (2.7.1). Ist umgekehrt für ein lineares $T : X \to Y$ der Ausdruck in (2.7.1) endlich (d.h. T beschränkt), so ist T auch stetig.

Der Satz von der gleichmäßigen Beschränktheit geht auf Banach[28] und Steinhaus[29] zurück und lautet:

Satz 2.7.2 *Vorausgesetzt seien (a) X ist Banach-Raum, (b) Y ist normierter Raum, (c) $\mathcal{T} \subset L(X,Y)$ ist eine (im Allgemeinen unendliche) Teilmenge von Abbildungen, (d) für alle $x \in X$ gelte $\sup_{T \in \mathcal{T}} \|Tx\|_Y < \infty$. Dann ist $\mathcal{T}$ gleichmäßig beschränkt, d.h. $\sup_{T \in \mathcal{T}} \|T\| < \infty$.*

Zunächst einige Anmerkungen. Sei K die Einheitskugel $K := \{x \in X : \|x\| = 1\}$. Die Definition (2.7.1) führt sofort auf

$$\|T\| = \sup_{x \in K} \|Tx\|_Y .$$

Die Aussage des Satzes lautet $\sup_{T \in \mathcal{T}} \sup_{x \in K} \|Tx\| < \infty$. Da sich die Suprema vertauschen lassen, kann man auch $\sup_{x \in K} \sup_{T \in \mathcal{T}} \|Tx\| < \infty$ schreiben. Vorausgesetzt ist nur $C(x) := \sup_{T \in \mathcal{T}} \|Tx\| < \infty$, d.h. die Funktion $C(x)$ ist punktweise beschränkt. Dass $C(x)$ auf K gleichmäßig beschränkt ist, ist eine erstaunliche Eigenschaft. Nur in dem Falle, dass X ein endlich-dimensionaler Vektorraum ist, ließe sich ein einfacher Beweis finden, indem man $\sup_{T \in \mathcal{T}} \|Tb_i\|_Y < \infty$ für alle Basisvektoren b_i $(i = 1, \ldots, n)$ ausnutzt.

In den späteren Anwendungsfällen wird eine speziellere Variante des Satzes verwandt.

Korollar 2.7.3 *X, Y seien wie in Satz 2.7.2. Ferner gelte für $T, T_n \in L(X,Y)$ $(n \in \mathbb{N})$ entweder a) für alle $x \in X$ ist $\{T_n x\}$ eine Cauchy-Folge oder b), es gebe ein $T \in L(X,Y)$ mit $T_n x \to Tx$ für alle $x \in X$. Dann gilt $\sup_{n \in \mathbb{N}} \|T_n\| < \infty$.*

Beweis. In diesem Fall ist $\mathcal{T} = \{T_n \in L(X,Y) : n \in \mathbb{N}\}$ abzählbar unendlich. a) Da jede Cauchy-Folge beschränkt ist, folgt die Beschränktheit $\sup_{n \in \mathbb{N}} \|T_n x\|_Y < \infty$, sodass Satz 2.7.2 anwendbar ist. Die Voraussetzung b) impliziert a). ∎

Der Beweis des Satzes 2.7.2 benötigt zwei weitere Sätze zur Vorbereitung.

Satz 2.7.4 (Bairescher Kategoriensatz)[30] *$X \neq \emptyset$ sei vollständiger metrischer Raum. Ferner habe X die Darstellung*

$$X = \bigcup_{k \in \mathbb{N}} A_k \qquad \text{mit abgeschlossenen Mengen } A_k.$$

Dann existiert mindestens ein $k_0 \in \mathbb{N}$, sodass $\mathring{A}_{k_0} \neq \emptyset$ ($\mathring{A}_{k_0}$ bezeichnet das Innere von A_{k_0}).

[28] Stefan Banach, geb. 30. März 1892 in Krakau, gest. 31. Aug. 1945 in Lemberg (Lwow)
[29] Hugo Dyonizy Steinhaus, geb. 14. Jan. 1887 in Jaslo (Galizien), gest. 25. Feb. 1972 in Breslau
[30] René-Louis Baire, geb. 21. Jan. 1874 in Paris, gest. gest. 5. Juli 1932 in Chambéry

Beweis. a) Für den indirekten Beweis nehmen wir $\mathring{A}_k = \emptyset$ für alle k an. Wir wählen eine nichtleere, offene Menge $U \subset X$ und ein $k \in \mathbb{N}$. Da A_k abgeschlossen, ist $U \backslash A_k$ wieder offen und nichtleer (sonst würde A_k die offene Menge U enthalten, d.h. $\mathring{A}_k \supset U \neq \emptyset$). Da $U \backslash A_k$ offen, enthält es eine abgeschlossene Kugel $\overline{K_\varepsilon(x)}$ mit Radius $\varepsilon > 0$ und Mittelpunkt x. O.B.d.A. kann $\varepsilon \leq 1/k$ gewählt werden.

b) Ausgehend von $\varepsilon_0 := 1$, $x_0 := 0$ wählen wir gemäß a) per Induktion

$$\overline{K_{\varepsilon_k}(x_k)} \subset K_{\varepsilon_{k-1}}(x_{k-1}) \backslash A_k \text{ und } \varepsilon_k \leq 1/k.$$

Da $x_\ell \in K_{\varepsilon_k}(x_k)$ für $\ell \geq k$ und $\varepsilon_k \to 0$ $(k \to \infty)$, ist $\{x_k\}$ eine Cauchy-Folge, die wegen der Vollständigkeit gegen ein $x := \lim x_k \in X$ konvergieren muss und für alle k zu $\overline{K_{\varepsilon_k}(x_k)}$ gehört. Da $\overline{K_{\varepsilon_k}(x_k)} \cap A_k = \emptyset$ nach Konstruktion, folgt $x \notin \bigcup_{k \in \mathbb{N}} A_k = X$, was den Widerspruch ergibt. ∎

Satz 2.7.5 *Sei X sei vollständiger metrischer Raum und Y ein normierter Raum. Für eine Teilmenge $\mathcal{F} \subset C^0(X,Y)$ der stetigen Abbildungen gelte $\sup_{f \in \mathcal{F}} \|f(x)\|_Y < \infty$ für alle $x \in X$. Dann existieren $x_0 \in X$ und $\varepsilon_0 > 0$ so, dass*

$$\sup_{x \in \overline{K_{\varepsilon_0}(x_0)}} \sup_{f \in \mathcal{F}} \|f(x)\|_Y < \infty. \tag{2.7.2}$$

Beweis. Man setze $A_k := \bigcap_{f \in \mathcal{F}} \{x \in X : \|f(x)\|_Y \leq k\}$ für $k \in \mathbb{N}$ und überprüfe, dass A_k abgeschlossen ist. Gemäß Voraussetzung muss jedes $x \in X$ zu einem der A_k gehören, d.h. $X = \bigcup_{k \in \mathbb{N}} A_k$. Damit sind die Voraussetzungen von Satz 2.7.4 erfüllt. Demnach gilt $\mathring{A}_{k_0} \neq \emptyset$ für mindestens ein $k_0 \in \mathbb{N}$. Nach Definition der A_k gilt zudem $\sup_{x \in A_{k_0}} \sup_{f \in \mathcal{F}} \|f(x)\|_Y \leq k_0$. Man wähle eine Kugel mit $\overline{K_{\varepsilon_0}(x_0)} \subset A_{k_0}$. Dies liefert die gesuchte Ungleichung (2.7.2) mit der Schranke $\leq k_0$. ∎

Um den Satz 2.7.2 zu beweisen, beachten wir, dass ein Banach-Raum auch ein vollständiger metrischer Raum ist und $L(X,Y) \subset C^0(X,Y)$, sodass $\mathcal{T} = \mathcal{F}$ gesetzt werden kann. Die Voraussetzung $\sup_{f \in \mathcal{F}} \|f(x)\|_Y < \infty$ ist zu $\sup_{T \in \mathcal{T}} \|Tx\|_Y < \infty$ äquivalent. Das Resultat (2.7.2) wird zu $\sup_{x \in \overline{K_{\varepsilon_0}(x_0)}} \sup_{T \in \mathcal{T}} \|Tx\|_Y < \infty$. Für ein beliebiges $\xi \in X \backslash \{0\}$ gehört $x_\xi := x_0 + \frac{\varepsilon_0}{\|\xi\|_X} \xi$ zu $\overline{K_{\varepsilon_0}(x_0)}$, sodass

$$\frac{\|T\xi\|_Y}{\|\xi\|_X} = \frac{1}{\varepsilon_0} \|T(x_\xi - x_0)\|_Y \leq \frac{1}{\varepsilon_0} \left(\|Tx_\xi\|_Y + \|Tx_0\|_Y \right)$$

für alle $T \in \mathcal{T}$ und alle $\xi \in X \backslash \{0\}$ gleichmäßig beschränkt ist. Damit ist $\sup_{T \in \mathcal{T}} \sup_{\xi \in X \backslash \{0\}} \frac{\|T\xi\|_Y}{\|\xi\|_X} = \sup_{T \in \mathcal{T}} \|T\| < \infty$.

2.8 Notwendigkeit der Stabilitätsbedingung, Äquivalenzsatz

$X = C([0,1])$ zusammen mit der Maximumnorm $\|\cdot\|_\infty$ ist ein Banach-Raum, $Y := \mathbb{R}$ ist normiert (Norm ist der Absolutbetrag). Die Abbildungen $f \in C([0,1]) \mapsto I(f) := \int_0^1 f(x)dx \in \mathbb{R}$ und $f \in C([0,1]) \mapsto Q_n(f) \in \mathbb{R}$ sind lineare und stetige Abbildungen, gehören also zu $L(X,Y)$. Nach Bemerkung 2.7.1b ist die Stetigkeit zur Beschränktheit äquivalent, die wir im folgenden Lemma quantifizieren.

Lemma 2.8.1 *Die Operatornormen von I und $Q_n \in L(X,Y)$ sind*

$$\|I\| = 1, \qquad \|Q_n\| = C_n := \sum_{i=0}^{n} |a_{i,n}|.$$

Beweis. Die Abschätzungen $\|I\| \leq 1$ und $\|Q_n\| \leq C_n$ sind mit den Abschätzungen (2.4.1) aus Bemerkung 2.4.1 und (2.4.4) äquivalent. Gemäß Übungsaufgabe 2.4.2 ist C_n die kleinstmögliche Konstante, was $\|Q_n\| = C_n$ impliziert. Das Beispiel $f = 1$ zeigt, dass 1 die beste Schranke für $\|If\|_\infty / \|f\|_\infty$ ist, d.h. $\|I\| = 1$. ∎

Nun setzen wir in Korollar 2.7.3 $T := I$ und $T_n := Q_n$. Die Konvergenz der Quadraturfamilie $\{Q_n\}$ schreibt sich als $Q_n(f) \to I(f)$. Aus Korollar 2.7.3 folgern wir $\sup_{n \in \mathbb{N}} \|T_n\| < \infty$. Nach obigem Lemma bedeutet dies $\sup_{n \in \mathbb{N}} C_n < \infty$ und ist mit der Stabilitätsbedingung aus Definition 2.4.3 identisch. Damit ist der folgende Satz bewiesen:

Satz 2.8.2 (Stabilitätssatz) *Die Familie $\{Q_n\}$ von Quadraturverfahren sei konvergent. Dann ist $\{Q_n\}$ auch stabil.*

Es wurde bereits "Konsistenz + Stabilität $\Longrightarrow$ Konvergenz" (vgl. (2.6.1)) bewiesen. In Satz 2.8.2 haben wir "Stabilität $\Longleftarrow$ Konvergenz" bewiesen. Zusammen ergibt sich der Äquivalenzsatz.

Satz 2.8.3 (Äquivalenzsatz) *Die Konsistenz der Familie $\{Q_n\}$ von Quadraturverfahren sei vorausgesetzt. Dann sind Stabilität und Konvergenz äquivalent.*

2.9 Modifizierte Definitionen für Konsistenz und Konvergenz

Die Begriffe 'Konsistenz' und 'Konvergenz' können noch etwas besser getrennt werden, ohne dass die bisherigen Aussagen falsch werden.

Der bisherige Begriff der Konvergenz enthielt nicht nur, dass die Folge $Q_n(f)$ konvergent ist, sondern auch dass sie das gesuchte Integral $\int_0^1 f(x)dx$ als Grenzwert besitzt. Dieser zweite Teil wird in der folgenden, modifizierten Form ausgespart:

$$\{Q_n\} \text{ ist konvergent, wenn für alle } f \in C([0,1]) \text{ die Folge } \{Q_n(f)\} \text{ eine Cauchy-Folge ist.} \tag{2.9.1}$$

Im bisherigen Begriff der Konsistenz spielten die Polynome eine Schlüsselrolle. Die Polynome kommen ins Spiel, da wir von der interpolatorischen Quadratur ausgehen, die auf der Polynominterpolation beruht. Für eine interpolatorischen Quadratur, die z.B. die trigonometrische Interpolation verwendet, würde die Konsistenz im Sinne von Definition 2.2.1 nicht gelten. Wie anschließend an Satz 2.5.1 vermerkt, zählt bei den Polynomen die Eigenschaft, in $C([0,1])$ dicht zu liegen. Man kann die Polynome durch jede andere dichte Teilmenge ersetzen. Damit erhält man eine Verallgemeinerung des Konsistenzbegriffes:

$$\{Q_n\} \text{ ist konsistent, wenn es eine dichte Teilmenge } X_0 \subset C([0,1]) \text{ gibt, sodass} \tag{2.9.2}$$
$$Q_n(f) \to \int_0^1 f(x)dx \text{ für alle } f \in X_0.$$

Man beachte, dass wir gleichzeitig die Exaktheit $Q_n(f) = \int_0^1 f(x)dx$ für $n \geq n_0$ durch die allgemeinere Konvergenz ersetzt haben.

Die Stabilitätseigenschaft bleibt unverändert.

Die bisherigen Sätze lassen sich dann wie folgt umformulieren.

Satz 2.9.1 *a) $\{Q_n\}$ sei konsistent im Sinne von (2.9.2) und stabil. Dann ist es konvergent im Sinne von (2.9.1) und darüber hinaus ist $\lim_{n\to\infty} Q_n(f) = \int_0^1 f(x)dx$ der gewünschte Integralwert.*

b) $\{Q_n\}$ sei konvergent im Sinne von (2.9.1). Dann ist $\{Q_n\}$ auch stabil.

c) Unter der Voraussetzung der Konsistenz im Sinne von (2.9.2), sind Stabilität und die Konvergenz (2.9.1) äquivalent.

Beweis. a) Es braucht nur $\lim_{n\to\infty} Q_n(f) = \int_0^1 f(x)dx$ gezeigt zu werden, da dies (2.9.1) impliziert. Seien $f \in C([0,1])$ und $\varepsilon > 0$ vorgegeben. Da X_0 aus (2.9.2) dicht ist, gibt es ein $g \in X_0$ mit $\|f-g\|_\infty \leq \frac{\varepsilon}{2(1+C_{\text{stab}})}$, wobei C_{stab} die Stabilitätskonstante ist. Gemäß (2.9.2) gibt es ein n_0, sodass $\left|Q_n(g) - \int_0^1 g(x)dx\right| \leq \frac{\varepsilon}{2}$ für alle $n \geq n_0$. Die Dreiecksungleichung liefert die gewünschte Abschätzung

$$\begin{aligned}\left|Q_n(f) - \int_0^1 f(x)dx\right| &\leq |Q_n(f) - Q_n(g)| + \left|Q_n(g) - \int_0^1 g(x)dx\right| + \left|\int_0^1 g(x)dx - \int_0^1 f(x)dx\right| \\ &\leq C_{\text{stab}} \|f-g\|_\infty + \frac{\varepsilon}{2} + \|f-g\|_\infty \underset{\|f-g\|_\infty \leq \varepsilon/[2(1+C_{\text{stab}})]}{\leq} \varepsilon.\end{aligned}$$

b) Konvergenz im Sinne von (2.9.1) garantiert, dass $\{Q_n(f)\}$ für alle $f \in C([0,1])$ eine Cauchy-Folge ist. Da Cauchy-Folgen beschränkt sind, ist Korollar 2.7.3 anwendbar und liefert $\sup_{n\in\mathbb{N}_0} \|Q_n\| = \sup_{n\in\mathbb{N}_0} C_n < \infty$, d.h. die Stabilität.

c) Der Teil c) folgt aus den Teilen a) und b). ∎

Abschließend sei eine mögliche Anwendung der verallgemeinerten Begriffe gegeben. Um die Schwierigkeiten, die sich aus der Instabilität der Newton-Cotes-Formeln ergeben, zu vermeiden, werden gerne summierte Formeln angewandt. Das bekannteste Beispiel ist die *summierte Trapezformel*, bei der auf jedem

Teilintervall $[i/n, (i+1)/n]$ die Trapezformel verwendet wird. Die summierten Trapezformeln definieren wieder eine Familie $\{Q_n\}$. Sie ist nicht konsistent im Sinne der Definition 2.2.1, da außer den konstanten und linearen Funktionen keine weiteren Polynome exakt integriert werden. Stattdessen gehen wir zurück zur Formulierung (2.3.1) des Quadraturfehlers. Die Standardabschätzung besagt, dass

$$\left| Q_n(f) - \int_0^1 f(x)dx \right| \leq \frac{1}{12n^2} \|f''\|_\infty \to 0 \tag{2.9.3}$$

für alle $f \in C^2([0,1])$. Die Teilmenge $C^2([0,1])$ ist dicht in $C([0,1])$ (einfachster Beweis: $C^2([0,1]) \supset \{\text{Polynome}\}$ und letztere sind nach Satz 2.5.1 schon dicht). Damit erfüllen die summierten Trapezformeln $\{Q_n\}$ die Konsistenzbedingung (2.9.2) mit $X_0 = C^2([0,1])$. Die Stabilität von $\{Q_n\}$ ergibt sich sofort aus Folgerung 2.4.6a, da alle Gewichte positiv sind und $Q_n(1) = 1$. Aus Satz 2.9.1a folgern wir die Eigenschaft $Q_n(f) \to \int_0^1 f(x)dx$ für alle stetigen f.

2.10 Weitere Anmerkungen

Während Abschätzungen wie (2.9.3) die Konvergenz in ihrer quantitativen Form beschreiben, fehlt diese ganz für die Aussage $Q_n(f) \to \int_0^1 f(x)dx$. Nicht quantitative Konvergenzbeschreibungen sind in numerischen Anwendung wenig aussagekräftig. Wenn man nicht weiß, ob ein Fehler $\varepsilon = 0.01$ schon bei $n = 5$, bei $n = 10^6$ oder vielleicht erst bei $n = 10^{10^{10}}$ erreicht wird, ist wenig geholfen.

Deshalb sei abschließend die Frage gestellt, ob etwas über die Konvergenzgeschwindigkeit von $Q_n(f) \to \int_0^1 f(x)dx$ bei allgemeinem $f \in C([0,1])$ ausgesagt werden kann. Die Konvergenzgeschwindigkeit könnte durch eine Nullfolge ε_n beschrieben werden: $\left| Q_n(f) - \int_0^1 f(x)dx \right| \leq \varepsilon_n \|f\|_\infty$. Die Antwort hierauf ist bereits in Bemerkung 2.3.1 gegeben worden: Wenn $\left| Q_n(f) - \int_0^1 f(x)dx \right| \leq \varepsilon_n \|f\|_\infty$ für alle $f \in C([0,1])$ gelten soll, muss notwendigerweise $\varepsilon_n \geq 1$ gelten, was jede Nullfolge ε_n ausschließt. Folglich kann die Konvergenz $Q_n(f) \to \int_0^1 f(x)dx$ *beliebig langsam* erfolgen.

Eine andere Fragestellung, die positiv beantwortet werden kann, ist die folgende. Aufgrund weiterer Störungen mag es sein, dass wir nicht $Q_n(f)$, sondern $Q_n(f_n)$ mit $f_n \to f$ berechnen. Lässt sich wieder $\lim_{n\to\infty} Q_n(f_n) = \int_0^1 f(x)dx$ zeigen?

Satz 2.10.1 *Die Familie $\{Q_n : n \in \mathbb{N}_0\}$ von Quadraturformeln sei konsistent und stabil, ferner gelte $f_n \to f$ für eine Folge von $f_n \in C([0,1])$. Dann konvergiert auch $Q_n(f_n)$ gegen $\int_0^1 f(x)dx$.*

Beweis. Die bisherigen Überlegungen garantieren $Q_n(f) \to \int_0^1 f(x)dx$. Dank der Stabilität ist die Störung $|Q_n(f_n) - Q_n(f)|$ durch $C_{\text{stab}} \|f_n - f\|_\infty \to 0$ beschränkt und eine Nullfolge. ∎

Ein mögliche Anwendung dieses Resultates, bei der Q_n auf unterschiedliche f_n angewandt wird, könnte wie folgt aussehen. Sei $t_n \to \infty$ eine Folge von natürlichen Zahlen und f_n die Computer-Realisierung von f in einer t_n-stelligen Arithmetik, sodass $\|f_n - f\|_\infty \leq C2^{-t_n}$. Mit der Erhöhung der Zahl der Quadraturstützstellen wird also gleichzeitig die Maschinengenauigkeit erhöht.

3 Interpolation

3.1 Interpolationsaufgabe

Die übliche lineare[31] Interpolationsaufgabe ist durch einen Vektorraum $V_n \subset C([0,1])$ und eine disjunkte Stützstellenmenge $\{x_{i,n} \in [0,1] : 0 \le i \le n\}$ charakterisiert. Zu einem Tupel $\{y_i : 0 \le i \le n\}$ von "Funktionswerten" ist eine Interpolierende $g \in V_n$ mit der Eigenschaft $g(x_{i,n}) = y_i$ $(0 \le i \le n)$ gesucht.

Übungsaufgabe 3.1.1 *a) Die Interpolationsaufgabe ist genau dann für alle Tupel* $\{y_i : 0 \le i \le n\}$ *lösbar, wenn* $X_n := \{(g(x_{i,n}))_{i=0}^n \in \mathbb{R}^{n+1} : g \in V_n\}$ *die Dimension* $n+1$ *hat.*
b) Die Interpolationsaufgabe ist eindeutig lösbar, wenn außerdem $\dim V_n = n+1$.

Im Falle $\dim V_n = n+1$ kann die Interpolationsaufgabe auf ein Gleichungssystem der Dimension $n+1$ zurückgeführt werden, sodass es bekanntlich zwei Alternativen gibt: Die Interpolationsaufgabe ist allgemein lösbar (d.h. für beliebige y_i gibt es genau eine Interpolierende) oder nicht (dann braucht keine Interpolierende zu existieren, wenn sie existiert, ist sie nicht eindeutig).[32]

Die *Polynominterpolation* ergibt sich für $V_n = \{\text{Polynome vom Grad} \le n\}$ und ist stets lösbar.

Im Weiteren wird die eindeutige Lösbarkeit angenommen.

Indem man die speziellen Werte $y_i = \delta_{ij}$ (j fest, δ_{ij} Kronecker-Symbol) interpoliert, erhält man die Interpolierenden $\Phi_{j,n} \in V_n$, die hier *Lagrange-Funktionen* genannt seien.

Übungsaufgabe 3.1.2 *a) Die Interpolierende zu beliebigen* y_i $(0 \le i \le n)$ *ist*

$$\Phi = \sum_{i=0}^{n} y_i \Phi_{i,n} \in V_n.$$

b) Im Falle der Polynominterpolation ist

$$L_{i,n}(x) := \Phi_{i,n}(x) := \prod_{j \in \{0,\dots n\} \setminus \{i\}} \frac{x - x_j}{x_i - x_j} \tag{3.1.1}$$

das Lagrange-Polynom.

Zu einer stetigen Funktion f definieren wir

$$I_n(f) := \sum_{i=0}^{n} f(x_{i,n}) \Phi_{i,n} \tag{3.1.2}$$

als Interpolierende zu $y_i = f(x_{i,n})$. Damit ist $I_n : C([0,1]) \to C([0,1])$ eine Abbildung der stetigen Funktionen in sich.

Übungsaufgabe 3.1.3 *a)* $I_n : C([0,1]) \to C([0,1])$ *ist stetig und linear, also* $I_n \in L(X,X)$ *für* $X = C([0,1])$.
b) I_n *ist Projektion, d.h.* $I_n I_n = I_n$.

Nach den Vorbereitungen des vorigen Kapitels können wir die Begriffe Konvergenz, Konsistenz und Stabilität schnell einführen. Hierbei ist wesentlich, dass wir nicht nur eine Interpolation I_n haben, sondern eine Familie $\{I_n : n \in \mathbb{N}_0\}$ von Interpolationen.

3.2 Konvergenz

Definition 3.2.1 *Eine Familie* $\{I_n : n \in \mathbb{N}_0\}$ *von Interpolationen heißt* konvergent, *falls*

$$I_n(f) \text{ für alle } f \in C([0,1]) \text{ eine Cauchy-Folge bildet.}$$

Die Absicht ist natürlich, dass $I_n(f) \to f$, aber hier reicht die Konvergenz ohne Fixierung des Grenzwertes.

[31] Das Wort "linear" bezieht sich auf den zugrundeliegenden linearen Raum, nicht auf lineare Funktionen.
[32] Vgl. Kapitel 2 in Stoer: *Einführung in die Numerische Mathematik I.* Springer-Verlag, Berlin, 8. Auflage, 1999

3.3 Konsistenz

Wir folgen dem Vorbild von (2.9.2):

Definition 3.3.1 *Eine Familie* $\{I_n : n \in \mathbb{N}_0\}$ *von Interpolationen heißt* konsistent, *wenn es eine dichte Teilmenge* $X_0 \subset C([0,1])$ *gibt, sodass*

$$I_n(g) \to g \qquad \text{für alle } g \in X_0.$$

Übungsaufgabe 3.3.2 *Sei* $\{I_n\}$ *die Interpolation mit Polynomen vom Grad* $\leq n$. *Man zeige, dass die Definition 3.3.1 für* $X_0 := \{Polynome\}$ *zutrifft.*

3.4 Stabilität

Zunächst charakterisieren wir die Operatornorm $\|I_n\|$.

Lemma 3.4.1 *Es gilt* $\|I_n\| = \|\sum_{i=0}^n |\Phi_{i,n}(\cdot)|\|_\infty$ *mit* $\Phi_{i,n}$ *aus (3.1.2).*

Beweis. a) Wir setzen $C_n := \|\sum_{i=0}^n |\Phi_{i,n}(\cdot)|\|_\infty$. Für beliebiges $f \in C([0,1])$ gilt

$$|I_n(f)(x)| = \left|\sum_{i=0}^n f(x_{i,n})\Phi_{i,n}(x)\right| \leq \sum_{i=0}^n \underbrace{|f(x_{i,n})|}_{\leq \|f\|_\infty} |\Phi_{i,n}(x)| \leq \|f\|_\infty \sum_{i=0}^n |\Phi_{i,n}(x)| \leq \|f\|_\infty C_n.$$

Da dies für alle $x \in [0,1]$ gilt, folgt $\|I_n(f)\| \leq C_n \|f\|_\infty$ und $\|I_n\| \leq C_n$.

b) Die Funktion $\sum_{i=0}^n |\Phi_{i,n}(\cdot)|$ sei bei x_0 maximal: $\sum_{i=0}^n |\Phi_{i,n}(x_0)| = C_n$. Man wähle $f \in C([0,1])$ mit $\|f\|_\infty = 1$ und $f(x_{i,n}) = \text{sign}(\Phi_{i,n}(x_0))$. Dann ist

$$|I_n(f)(x_0)| = \left|\sum_{i=0}^n f(x_{i,n})\Phi_{i,n}(x_0)\right| = \sum_{i=0}^n |\Phi_{i,n}(x_0)| = C_n = C_n \|f\|_\infty,$$

d.h. $\|I_n(f)\|_\infty = C_n \|f\|_\infty$ für dieses f. Für $\|I_n\| = \sup_{f \in C([0,1])\setminus\{0\}} \|I_n(f)\|_\infty / \|f\|_\infty$ folgt $\|I_n\| \geq C_n$.

c) Die Teile a,b ergeben zusammen die Behauptung $\|I_n\| = C_n$. ∎

Die Stabilität besteht wieder in der Beschränktheit der Folge $\{C_n\}$:

Definition 3.4.2 *Eine Familie* $\{I_n : n \in \mathbb{N}_0\}$ *von Interpolationen heißt* stabil, *wenn*

$$C_{\text{stab}} := \sup_{n \in \mathbb{N}_0} C_n < \infty \qquad \text{für} \quad C_n = \|I_n\| = \left\|\sum_{i=0}^n |\Phi_{i,n}(\cdot)|\right\|_\infty$$

(C_n aus Lemma 3.4.1) gilt.

3.5 Sätze

Dem Schema (2.6.1) folgend gilt der

Satz 3.5.1 (Konvergenzsatz) *Die Familie* $\{I_n : n \in \mathbb{N}_0\}$ *von Interpolationen sei konsistent und stabil. Dann ist sie auch konvergent und darüber hinaus gilt* $I_n(f) \to f$.

Beweis. a) Seien $f \in C([0,1])$ und $\varepsilon > 0$ vorgegeben. Es gibt ein $g \in X_0$ mit $\|f - g\|_\infty \leq \frac{\varepsilon}{2(1+C_{\text{stab}})}$, wobei C_{stab} die Stabilitätskonstante ist. Gemäß Definition 3.3.1 gibt es ein n_0, sodass $\|I_n(g) - g\|_\infty \leq \frac{\varepsilon}{2}$ für alle $n \geq n_0$. Die Dreiecksungleichung liefert die gewünschte Abschätzung

$$\begin{aligned}\|I_n(f) - f\|_\infty &\leq \|I_n(f) - I_n(g)\|_\infty + \|I_n(g) - g\|_\infty + \|g - f\|_\infty \\ &\leq C_{\text{stab}} \|f - g\|_\infty + \frac{\varepsilon}{2} + \|f - g\|_\infty \underset{\|f-g\|_\infty \leq \varepsilon/[2(1+C_{\text{stab}})]}{\leq} \varepsilon.\end{aligned}$$

∎

Die Stabilitätsbedingung ist wieder notwendig:

Lemma 3.5.2 *Eine konvergente Familie $\{I_n : n \in \mathbb{N}_0\}$ von Interpolationen ist stabil.*

Beweis. Da $I_n(f)$ eine Cauchy-Folge bilden, sind sie beschränkt. Man wende Korollar 2.7.3 mit $X = Y = C([0,1])$ und $T_n := I_n \in L(X;Y)$ an. ∎

Zusammen ergibt sich der Äquivalenzsatz:

Satz 3.5.3 (Äquivalenzsatz) *Die Familie $\{I_n : n \in \mathbb{N}_0\}$ von Interpolationen sei konsistent. Dann sind Konvergenz und Stabilität äquivalent.*

3.6 Instabilität der Polynominterpolation

Wir wählen die äquidistanten Stützstellen $x_{i,n} = i/n$ und beschränken uns auf gerade n. Das Lagrange-Polynom $L_{\frac{n}{2},n}$ ist im Teilintervall $(0, 1/n)$ besonders groß. In seinem Mittelpunkt gilt

$$\left|L_{\frac{n}{2},n}(\frac{1}{2n})\right| = \left|\prod_{\substack{j=0\\ j\neq\frac{n}{2}}}^{n} \frac{\frac{1}{2n}-\frac{j}{n}}{\frac{1}{2}-\frac{j}{n}}\right| = \left|\prod_{\substack{j=0\\ j\neq\frac{n}{2}}}^{n} \frac{\frac{1}{2}-j}{\frac{n}{2}-j}\right| = \frac{\frac{1}{2}\times\frac{1}{2}\times\frac{3}{2}\times\ldots\times\left(\frac{n}{2}-\frac{3}{2}\right)\times\left(\frac{n}{2}+\frac{1}{2}\right)\times\ldots\times\left(n-\frac{1}{2}\right)}{\left[\left(\frac{n}{2}\right)!\right]^2}.$$

Übungsaufgabe 3.6.1 *Man zeige, dass der obige Ausdruck exponentiell divergiert.*

Wegen $C_n = \|\sum_{i=0}^{n} |L_{i,n}(\cdot)|\|_\infty \geq \|L_{\frac{n}{2},n}\|_\infty \geq \left|L_{\frac{n}{2},n}(\frac{1}{2n})\right|$ kann die Interpolation nicht stabil sein.

3.7 Stabilität der stückweisen Polynominterpolation

Das Intervall $[0,1]$ sei wieder in n Teilintervalle $[i/n, (i+1)/n]$ zerlegt und in jedem Teilintervall werde beispielsweise die kubische Polynominterpolation in den Stützstellen $\left(i+\frac{j}{3}\right)/n$ $(j = 0,1,2,3)$ angewandt. Diese definiere die Interpolation I_n. Aufgrund der üblichen Fehlerabschätzung gilt

$$\|I_n(g) - g\|_\infty \leq Cn^{-4}\|g^{(4)}\|_\infty \to 0.$$

Wir wählen daher $X_0 = C^4([0,1])$ als dichten Teilraum von $C([0,1])$ und erhalten so die Konsistenz.

Übungsaufgabe 3.7.1 *Man bestimme die Stabilitätskonstante der stückweise kubischen Interpolation.*

Zusammen mit der Stabilität ergibt Satz 3.5.1 die Konvergenz $I_n(f) \to f$ für alle $C([0,1])$.

3.8 Von punktweiser Konvergenz zur Operatornormkonvergenz

Wie in §2.10 schon für die Quadratur ausgeführt, hat man nur die *punktweise Konvergenz* $I_n(f) \to f$ $(f \in X)$, nicht aber die Operatornormkonvergenz $\|I_n - id\| \to 0$. Allerdings gibt es Situationen, wo sich die punktweise Konvergenz in Operatornormkonvergenz umwandeln läßt.

Ein Operator $K : X \to Y$ heißt *kompakt*, wenn das Bild $B := \{Kf : \|f\|_X \leq 1\}$ präkompakt[33] ist. Der folgende Satz wird für eine beliebige, punktweise konvergente Folge $A_n : Y \to Z$ formuliert. Man kann sowohl an den Interpolationsoperator $A_n = I_n$ $(Y = Z = C([0,1]))$ als auch an die Quadratur $A_n = Q_n$ $(Y = C([0,1]),\ Z = \mathbb{R})$ denken.

Satz 3.8.1 *X, Y, Z seien Banach-Räume. Für $A, A_n \in L(Y,Z)$ gelte die punktweise Konvergenz $A_n y \to Ay$ für alle $y \in Y$. Ferner sei $K : X \to Y$ kompakt. Dann konvergieren die Produkte $P_n := A_n K$ in der Operatornorm gegen $P := AK$, d.h. $\|P_n - P\| \to 0$.*

Der Beweis wird mit dem folgenden Lemma vorbereitet.

Lemma 3.8.2 *$M \subset X$ sei eine präkompakte Teilmenge des Banach-Raumes X. Die Operatoren $A_n \in L(Y,Z)$ seien punktweise konvergent gegen $A \in L(Y,Z)$ (d.h. $A_n y \to Ay$ für alle $y \in Y$). Dann konvergieren die Folgen $\{A_n x\}$ für alle $x \in M$ gleichmäßig, d.h.*

$$\sup_{x\in M} \|A_n x - Ax\|_Z \to 0 \qquad \text{für } n \to \infty. \tag{3.8.1}$$

[33] B präkompakt $\iff$ $\bar{B}$ kompakt $\iff$ alle Folgen $\{f_k\} \subset B$ besitzen eine konvergente Teilfolge: $\lim_{n\to\infty} f_{k_n} \in \bar{B}$.

Beweis. a) $C := \sup\{\|A_n\| : n \in \mathbb{N}\}$ ist endlich ("Stabilität", vgl. Korollar 2.7.3). Ferner ist $\|A\| \leq C$ eine einfache Folgerung.

b) Für den indirekten Beweis wird die Negation von (3.8.1) angenommen: Es gibt ein $\varepsilon > 0$ und eine Teilfolge $\mathbb{N}' \subset \mathbb{N}$, sodass $\sup_{y \in M} \|A_n y - Ay\|_Z \geq \varepsilon$ für alle $n \in \mathbb{N}'$. Daher existieren $y_n \in M$ mit

$$\|A_n y_n - A y_n\|_Z \geq \varepsilon/2 \qquad \text{für alle } n \in \mathbb{N}'.$$

Da M präkompakt ist, gibt es eine weitere Teilfolge $\mathbb{N}'' \subset \mathbb{N}'$, sodass $\lim_{n \in \mathbb{N}''} y_n =: \eta$ existiert. Man wähle $n \in \mathbb{N}''$ mit $\|y_n - \eta\|_Y \leq \varepsilon/(8C)$ und $\|A_n \eta - A\eta\|_Z < \varepsilon/4$. Für dieses n ist

$$\|A_n y_n - A y_n\|_Z \leq \|(A_n - A)(y_n - \eta)\|_Z + \|(A_n - A)\eta\| < \underbrace{(\|A_n\| + \|A\|)}_{\leq 2C} \|y_n - \eta\|_Y + \varepsilon/4 \leq \varepsilon/2$$

im Widerspruch zur vorherigen Abschätzung. ∎

Beweis des Satzes 3.8.1. $M := \{Kx : \|x\|_X \leq 1\} \subset Y$ ist wegen der Kompaktheit von K präkompakt, sodass das vorhergehende Lemma anwendbar ist:

$$\begin{aligned}\|P_n - P\| &= \sup\{\|P_n x - Px\|_Z : x \in X, \|x\| \leq 1\} \\ &= \sup\{\|A_n(Kx) - A(Kx)\|_Z : x \in X, \|x\| \leq 1\} = \sup\{\|A_n y - Ay\|_Z : y \in M\} \overset{(3.8.1)}{\to} 0.\end{aligned}$$

∎

4 Gewöhnliche Differentialgleichungen

4.1 Einführung

4.1.1 Anfangswertaufgabe

Sei $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ eine stetige Funktion[34]. Gesucht sind im Folgenden stetig differenzierbare Funktionen $y(x)$, die die *gewöhnliche Differentialgleichung*

$$y'(x) = f(x, y(x)) \tag{4.1.1a}$$

erfüllen. Gegeben einen "Anfangswert" y_0 an einer Stelle x_0 besteht die "Anfangswertaufgabe" darin, eine Lösung y von (4.1.1a) zu finden, die zusätzlich

$$y(x_0) = y_0 \tag{4.1.1b}$$

erfüllt. Üblicherweise sucht man die Lösung y nur rechts von x_0: entweder in einem endlichen[35] Intervall $I := [x_0, x_E]$ oder im unbeschränkten Bereich $I := [x_0, \infty)$. Entsprechend braucht f nur auf $I \times \mathbb{R}$ definiert zu sein.

Ist f wie oben beschrieben nur stetig, existiert zwar nach dem Satz von Peano[36] eine Lösung von (4.1.1a,b) (in einer Umgebung $[x_0, x_0 + \varepsilon)$), allerdings braucht diese nicht eindeutig zu sein. Eindeutigkeit erzielt man durch die Voraussetzung, dass f im zweiten Argument Lipschitz-stetig sei. Wir formulieren hier der Einfachheit halber die globale *Lipschitz-Stetigkeit*:[37]

$$|f(x, y_1) - f(x, y_2)| \le L\,|y_1 - y_2| \qquad \text{für alle } x \in I,\ y_1 - y_2 \in \mathbb{R}. \tag{4.1.2}$$

Die Lipschitz-Konstante L wird in den weiteren Untersuchungen mehrfach auftreten. Die eindeutige Lösbarkeit wird ein Korollar eines späteren Satzes sein (vgl. Korollar 4.3.3 zu Satz 4.3.1).

4.1.2 Einschrittverfahren

Wir wählen eine feste Schrittweite[38] $h > 0$ und dazu die Stützstellen

$$x_i := x_0 + ih \qquad (i \in \mathbb{N}_0, x_i \in I).$$

Im Folgenden sollen Näherungen η_i von $y(x_i)$ berechnet werden. Die Schreibweise η_i unterstellt eine feste und bekannte Schrittweite h. Wenn es nötig ist, schreiben wir statt η_i auch $\eta(x_0 + ih; h)$. Man beachte, dass $\eta(x; h)$ nur auf dem Gitter $\{x_0 + ih : i \in \mathbb{N}_0\}$ definiert ist. Die Hoffnung ist, dass $\eta(x; h) \approx y(x)$ bis auf einen Fehler gilt, der mit h gegen null geht.

Aufgrund des bekannten Anfangswertes y_0 wird stets

$$\eta_0 := y_0 \tag{4.1.3}$$

gewählt.

Der Prototyp des Einschrittverfahrens ist das *Euler-Verfahren*[39], das mit (4.1.3) startet und rekursiv mittels

$$\eta_{i+1} := \eta_i + hf(x_i, \eta_i) \tag{4.1.4}$$

definiert ist.

[34] Bei einem *System* von Differentialgleichungen liegt der Definitionsbereich von f in $\mathbb{R} \times \mathbb{R}^n$ und die Lösung $y \in C^1(\mathbb{R}, \mathbb{R}^n)$ ist vektorwertig. Für unsere Betrachtungen reicht es aber, den skalaren Fall $n = 1$ zu betrachten.

[35] Es kann durchaus sein, dass die Lösung nur auf einem kleineren Intervall $[x_0, x_S) \subset [x_0, x_E]$ existiert.

[36] Giuseppe Peano, geb. 27. Aug. 1858 in Cuneo (Italien), gest. 20. April 1932 in Turin

[37] Rudolf Otto Sigismund Lipschitz, geb. 14. Mai 1832 in Königsberg, gest. 7. Okt. 1903 in Bonn

[38] In wirklichen Implementierungen ist es ganz wesentlich, variierende Schrittweiten $h_i = x_{i+1} - x_i$ zuzulassen, die in geeigneter Weise adaptiv gewählt werden. Für die anstehende Diskussion kann aber eine konstante Schrittweite angenommen werden.

[39] Leonhard Euler, geb. 15. April 1707 in Basel, gest. 18. Sept. 1783 in St. Petersburg

Übungsaufgabe 4.1.1 *Man betrachte die Differentialgleichung $y' = ay$ (d.h. $f(x,y) = ay$) mit dem Anfangswert $y_0 = 1$ bei $x_0 = 0$, die bekanntlich $y(x) = e^{ax}$ als Lösung besitzt. Welche Lösung ergibt sich für (4.1.4)? Gilt $y(x) - \eta(x;h) \to 0$ für $h := x/n > 0$ beim Grenzübergang $n \to \infty$ und $h \to 0$ mit fixiertem $nh = x$?*

Im allgemeinen Fall ersetzt man die rechte Seite in (4.1.4) durch einen allgemeineren Ausdruck $h\phi(x_i, \eta_i, h; f)$. Dabei bedeutet das Argument f, dass die Funktion f für beliebige Auswertungen innerhalb seines Definitionsbereiches vor Verfügung steht.

Definition 4.1.2 *Ein allgemeines explizites Einschrittverfahren hat die Form*

$$\eta_{i+1} := \eta_i + h\phi(x_i, \eta_i, h; f). \tag{4.1.5}$$

Die praktische Auswertung von ϕ wird häufig in mehreren Teilschritten vollzogen. Das *Heun-Verfahren* verwendet z.B. den Zwischenschritt $\eta_{i+1/2}$:

$$\eta_{i+1/2} := \eta_i + \frac{h}{2} f(x_i, \eta_i),\ \eta_{i+1} := \eta_i + hf(x_i + \frac{h}{2}, \eta_{i+1/2}).$$

Offenbar ist in diesem Fall $\phi(x_i, \eta_i, h; f) := f(x_i + \frac{h}{2}, \eta_i + \frac{h}{2} f(x_i, \eta_i))$. Das klassische *Runge-Kutta-Verfahren*[40] verwendet vier Zwischenschritte.

4.1.3 Mehrschrittverfahren

Der Name "Einschrittverfahren" bezieht sich darauf, dass man ausgehend von (x_i, η_i) in einem Schritt zu (x_{i+1}, η_{i+1}) gelangt. Die Vergangenheit $\{\eta_j : j < i\}$ spielt keine explizite Rolle. Da andererseits die Werte $\eta_{i-r}, \eta_{i-r+1}, \ldots, \eta_{i-1}$ neben η_i zur Verfügung stehen (r sei eine fixierte natürliche Zahl), kann man sich fragen, ob sie für die Berechnung hilfreich sind. Tatsächlich erreicht man mit ihrer Hilfe, dass mehr Parameter zur Verfügung stehen, was man ausnutzen kann, um die Ordnung des Verfahrens möglichst hochzutreiben.

Das lineare r-Schrittverfahren hat die Form

$$\eta_{j+r} := -\sum_{\nu=0}^{r-1} \alpha_\nu \eta_{j+\nu} + h\phi(x_j, \eta_{j+r-1}, \ldots, \eta_j, h; f) \tag{4.1.6}$$

(genau genommen ist dies die explizite Form) mit zusätzlichen Konstanten $\alpha_\nu \in \mathbb{R}$ für $j = 0, \ldots, r-1$. Wie wir sehen werden, stellt

$$\sum_{\nu=0}^{r-1} \alpha_\nu = -1 \tag{4.1.7}$$

eine erste Konsistenzbedingung dar.

Bemerkung 4.1.3 *Wegen (4.1.7) reduziert sich das Mehrschrittverfahren (4.1.6) für $r = 1$ auf das Einschrittverfahren (4.1.5).*

Bemerkung 4.1.4 *Im Falle $r \geq 2$ kann das Mehrschrittverfahren (4.1.6) nur zur Berechnung der η_i für $i \geq r$ verwendet werden. Die Berechnung der $\eta_1, \eta_2, \ldots, \eta_{r-1}$ muss in anderer Weise definiert werden (z.B. durch ein Einschrittverfahren).*

Ein Beispiel für ein Zweischrittverfahren ist die *Mittelpunktsformel*

$$\eta_{j+2} = \eta_j + 2hf(x_{j+1}, \eta_{j+1}), \tag{4.1.8}$$

d.h. $r = 2$, $\alpha_0 = -1$, $\alpha_1 = 0$, $\phi(x_j, \eta_{j+1}, \eta_j, h; f) = 2f(x_j + h, \eta_{j+1})$.

[40] Die Runge-Kutta-Methode wurde 1901 veröffentlicht. Die Autoren sind:
Carle David Tolmé Runge, geb. 30. Aug. 1856 in Bremen, gest. 3. Jan. 1927 in Göttingen;
Martin Wilhelm Kutta, geb. 3. Nov. 1867 in Pitschen (Oberschlesien), gest. 25. Dez. 1944 in Fürstenfeldbruck.

Ein eher zweifelhafter Vorschlag ist die Extrapolation

$$\eta_{j+2} = 2\eta_{j+1} - \eta_j, \tag{4.1.9}$$

d.h. $r = 2$, $\alpha_0 = -2$, $\alpha_1 = 1$, $\phi(x_j, \eta_{j+1}, \eta_j, h; f) = 0$. Dass dieses Zweischrittverfahren nicht zum Ziel führen kann, sieht man ohne eine quantitative Fehleranalyse: Da (4.1.9) nicht von f abhängt, liefert es für alle Differentialgleichungen die lineare Funktion $\eta_j = \eta_0 + j\,(\eta_1 - \eta_0)$, während es offenbar Differentialgleichungen mit andersartigen Lösungen gibt.

4.2 Fixpunktsatz und rekursive Ungleichungen

Als Vorbereitung benötigen wir den

Satz 4.2.1 (Banachscher Fixpunktsatz) *Seien X ein Banach-Raum und $\Psi : X \to X$ eine stetige und kontrahierende Abbildung, d.h. es gibt ein $L_\Psi \in [0,1)$, sodass*

$$\|\Psi(x') - \Psi(x'')\|_X \le L_\Psi \|x' - x''\|_X$$

für alle $x', x'' \in X$. Dann hat die Fixpunktgleichung $x = \Psi(x)$ *genau eine Lösung, und für alle Startwerte $x_0 \in X$ konvergiert die* Fixpunktiteration $x_{n+1} = \Psi(x_n)$ *gegen diese Lösung.*

Beweis. a) Seien x', x'' zwei Lösungen der Fixpunktgleichung, d.h. $x' = \Psi(x')$ und $x'' = \Psi(x'')$. Differenzenbildung und Anwendung der Kontraktionseigenschaft mit $L = L_\Psi$ liefert

$$\|x' - x''\|_X = \|\Psi(x') - \Psi(x'')\|_X \le L \|x' - x''\|_X ,$$

was wegen $L < 1$ nur $\|x' - x''\|_X = 0$ zulässt, d.h. es gilt die Eindeutigkeit $x' = x''$.

b) Die Iterierten x_n der Fixpunktiteration erfüllen die Ungleichung

$$\|x_{n+1} - x_n\|_X = \|\Psi(x_n) - \Psi(x_{n-1})\|_X \le L \|x_n - x_{n-1}\|_X$$

und damit $\|x_{n+1} - x_n\|_X \le L^n \|x_1 - x_0\|_X$. Die mehrfache Dreiecksungleichung liefert für beliebige $n > m$ die Abschätzung

$$\|x_n - x_m\|_X \le \sum_{j=m+1}^{n} \|x_j - x_{j-1}\|_X \le \sum_{j=m+1}^{n} L^{j-1} \|x_1 - x_0\|_X \le \sum_{j=m}^{\infty} L^j \|x_1 - x_0\|_X = L^m \frac{\|x_1 - x_0\|_X}{1 - L_\Psi},$$

d.h. $\{x_n\}$ ist eine Cauchy-Folge. Da X als Banach-Raum vollständig ist, gilt es ein $x^* = \lim x_n$. Limesbildung in $x_{n+1} = \Psi(x_n)$ ergibt wegen der Stetigkeit von Ψ, dass $x^* = \Psi(x^*)$. Damit ist x^* aber die eindeutige Fixpunktlösung. ∎

Im Folgenden werden rekursive Ungleichungen der folgenden Form auftreten:

$$a_{\nu+1} \le (1 + hL)\, a_\nu + h^k B \qquad \text{für alle } \nu \ge 0, \text{ wobei } L, B, h, k, a_0 \ge 0. \tag{4.2.1}$$

Dabei wird L eine der Lipschitz-Konstanten sein, $h > 0$ die Schrittweite und k die lokale Konsistenzordnung.

Lemma 4.2.2 *Jede Lösung der Ungleichungen (4.2.1) genügt der Abschätzung*

$$a_\nu \le e^{\nu h L} a_0 + h^{k-1} B \times \left\{ \begin{array}{ll} \nu h & \text{für } L = 0 \\ \dfrac{e^{\nu h L} - 1}{L} & \text{für } L > 0 \end{array} \right\} \qquad (\nu \in \mathbb{N}_0).$$

Beweis. a) Wir stellen die folgende Induktionsbehauptung auf:

$$a_\nu \le A_\nu := \sum_{\mu=0}^{\nu-1} (1 + hL)^\mu\, h^k B + (1 + hL)^\nu\, a_0. \tag{4.2.2}$$

Der Induktionsanfang ist mit $A_0 = a_0$ gezeigt. Induktion $\nu \to \nu+1$: Einsetzen von $a_\nu \le A_\nu$ in (4.2.1) liefert

$$\begin{aligned} a_{\nu+1} \le (1+hL)\, A_\nu + h^k B &= \sum_{\mu=1}^{\nu} (1+hL)^\mu\, h^k B + (1+hL)^{\nu+1}\, a_0 + h^k B \\ &= \sum_{\mu=0}^{\nu} (1+hL)^\mu\, h^k B + (1+hL)^{\nu+1}\, a_0 = A_{\nu+1}. \end{aligned}$$

Mit Übungsaufgabe 2.4.10a) folgt $(1+hL)^\nu \le \left(e^{hL}\right)^\nu = e^{hL\nu}$. Für $L > 0$ liefert die geometrische Summe den Wert

$$h^k B \sum_{\mu=0}^{\nu-1} (1+hL)^\mu = h^k B \frac{(1+hL)^\nu - 1}{(1+hL) - 1} = h^{k-1} \frac{B}{L} \left[(1+hL)^\nu - 1\right] \le h^{k-1} \frac{B}{L} \left[e^{hL\nu} - 1\right].$$

Damit lässt sich A_ν aus (4.2.2) durch $A_\nu \le h^{k-1} \frac{B}{L} \left[e^{hL\nu} - 1\right] + e^{hL\nu} a_0$ abschätzen. Der Sonderfall $L = 0$ kann etwa extra oder aus $L > 0$ mittels Grenzwertbildung behandelt werden. ∎

4.3 Wohlkonditioniertheit der Anfangswertaufgabe

Bevor wir mit der numerischen Lösung beginnen, ist zu fragen, ob die Anfangswertaufgabe (d.h. die Abbildung $(y_0, f) \mapsto y$) überhaupt wohlkonditioniert ist. Nach §1.4.3 ist die Verstärkung einer Störung der Eingabedaten zu untersuchen. Im vorliegenden Falle kann sowohl der Anfangswert y_0 gestört sein als auch die Funktion f. Der erste Fall wird im folgenden Satz analysiert.

Satz 4.3.1 *Seien $y_1, y_2 \in C^1(I)$ zwei Lösungen der Differentialgleichung (4.1.1a) zu den Anfangswerten*

$$y_1(x_0) = y_{0,1} \quad \textit{bzw.} \quad y_2(x_0) = y_{0,2}.$$

$f \in C(I \times \mathbb{R})$ erfülle (4.1.2). Dann gilt in I die Abschätzung

$$|y_1(x) - y_2(x)| \le |y_{0,1} - y_{0,2}|\, e^{L(x-x_0)} \qquad \textit{mit } L \textit{ aus (4.1.2).} \tag{4.3.1}$$

Beweis. Für $i = 1, 2$ gilt $y_i(x) = y_{0,i} + \int_{x_0}^{x} f(t, y_i(t))dt$, sodass

$$\begin{aligned} |y_1(x) - y_2(x)| &= \left| y_{0,1} - y_{0,2} + \int_{x_0}^{x} \left[f(t, y_1(t)) - f(t, y_2(t))\right] dt \right| \\ &\le |y_{0,1} - y_{0,2}| + \int_{x_0}^{x} |f(t, y_1(t)) - f(t, y_2(t))|\, dt \\ &\overset{(4.1.2)}{\le} |y_{0,1} - y_{0,2}| + \int_{x_0}^{x} L\, |y_1(t) - y_2(t)|\, dt. \end{aligned}$$

Übungsaufgabe 4.3.2 *Man zeige, dass jede Lösung φ der Ungleichung*

$$\varphi(x) \le \varphi_0 + L \int_{x_0}^{x} \varphi(t)dt$$

durch Φ nach oben beschränkt ist (d.h. $\varphi(x) \le \Phi(x)$), wobei Φ Lösung der Integralgleichung

$$\Phi(x) = \varphi_0 + L \int_{x_0}^{x} \Phi(t)dt.$$

Hinweis: *a) Mit $\Psi(\Phi)(x) := \varphi_0 + L \int_{x_0}^{x} \Phi(t)dt$ ist $\Phi = \Psi(\Phi)$ eine Fixpunktgleichung für $\Phi \in C(I)$, $I = [x_0, x_E]$. Man zeige zunächst, dass Ψ kontrahierend ist bezüglich der Norm*

$$\|\psi\| := \max\{|\psi(t)| \exp\left(-2L(t-x_0)\right) : t \in I\},$$

wobei die Kontraktionszahl $L_\Psi = \frac{1}{2}$ beträgt (unabhängig von der Länge des Intervalles I, auch für $I = [x_0, \infty)$ gültig).
b) Man wende die Fixpunktiteration $\Phi_{n+1} := \Psi(\Phi_n)$ mit $\Phi_0 := \varphi$ an und zeige $\Phi_n \ge \varphi$ für alle $n \ge 0$.

Die Funktion $\Phi(x) := |y_{0,1} - y_{0,2}|\, e^{L(x-x_0)}$ erfüllt in I die Gleichung $\Phi(x) = |y_{0,1} - y_{0,2}| + \int_{x_0}^{x} L\Phi(t)dt$. Also beweist das Resultat der Übungsaufgabe $|y_1(x) - y_2(x)| \le \Phi(x)$, d.h. (4.3.1). ∎

Korollar 4.3.3 *Die Voraussetzung (4.1.2) sichert die Eindeutigkeit der Anfangswertaufgabe (4.1.1a,b).*

Beweis. Sind $y_1, y_2 \in C^1(I)$ zwei Lösungen der Anfangswertaufgabe (4.1.1a,b), so zeigt Satz 4.3.1, dass $|y_1(x) - y_2(x)| \le |y_{0,1} - y_{0,2}| \exp(L(x-x_0)) \underset{y_{0,1}=y_{0,2}}{=} 0$, also $y_1(x) = y_2(x)$ in I. ∎

Schreibt man die Lösung der Anfangswertaufgabe (4.1.1a,b) als $y(x; y_0)$ mit y_0 als zweitem Argument, so besagt (4.3.1), dass $y(\cdot;\cdot)$ ebenso wie $f(\cdot,\cdot)$ Lipschitz-stetig im zweiten Argument ist. Dies kann verallgemeinert werden:

Übungsaufgabe 4.3.4 *Ist $f(\cdot,\cdot)$ k-fach stetig differenzierbar bezüglich des zweiten Argumentes y, so auch die Lösung $y(\cdot;\cdot)$ bezüglich y_0.*

Die Störung in der rechten Seite f der Differentialgleichung (4.1.1a) ist Gegenstand von

Satz 4.3.5 *Seien y und $\tilde{y}$ Lösungen von $y' = f(x,y)$ bzw. $\tilde{y}' = \tilde{f}(x,\tilde{y})$ zum gleichen Anfangswert $y(x_0) = \tilde{y}(x_0) = y_0$. Dabei braucht nur f die Lipschitz-Bedingung (4.1.2) zu erfüllen, während*

$$\left| f(x,y) - \tilde{f}(x,y) \right| \le \varepsilon \qquad \text{für alle } x \in I, y \in \mathbb{R}.$$

Dann gilt $|y(x) - \tilde{y}(x)| \le \frac{\varepsilon}{L}\left(e^{L(x-x_0)} - 1\right)$, falls in (4.1.2) $L > 0$ gilt; sonst – für den trivialen Fall $L = 0$ – $|y(x) - \tilde{y}(x)| \le \varepsilon(x - x_0)$.

Beweis. Man setze $\delta(x) := |y(x) - \tilde{y}(x)|$ und beachte, dass

$$\begin{aligned}
\delta(x) &= \left| \int_{x_0}^{x} \left[f(\xi, y(\xi)) - \tilde{f}(\xi, \tilde{y}(\xi)) \right] d\xi \right| \\
&\le \int_{x_0}^{x} \left| \left[f(\xi, y(\xi)) - \tilde{f}(\xi, \tilde{y}(\xi)) \right] \right| d\xi = \int_{x_0}^{x} \left| f(\xi, y(\xi)) - f(\xi, \tilde{y}(\xi)) + f(\xi, \tilde{y}(\xi)) - \tilde{f}(\xi, y(\xi)) \right| d\xi \\
&\le \int_{x_0}^{x} |f(\xi, y(\xi)) - f(\xi, \tilde{y}(\xi))|\, d\xi + \int_{x_0}^{x} \left| f(\xi, \tilde{y}(\xi)) - \tilde{f}(\xi, y(\xi)) \right| d\xi \\
&\le \int_{x_0}^{x} L\delta(\xi) d\xi + (x - x_0)\varepsilon.
\end{aligned}$$

Wieder erhält man eine Majorante als Lösung von $d(x) = \int_{x_0}^{x} Ld(\xi)d\xi + (x - x_0)\varepsilon$, die $\frac{\varepsilon}{L}\left(e^{L(x-x_0)} - 1\right)$ im Falle $L > 0$ lautet. ∎

4.4 Analyse von Einschrittverfahren

Vor der eigentlichen Analyse begründen wir noch, warum wir uns auf explizite Einschrittverfahren beschränken können (§4.4.1) und diskutieren die Lipschitz-Stetigkeit von ϕ (§4.4.2). Wir werden danach stets von einem expliziten Einschrittverfahren (4.1.5) ausgehen, wobei ϕ zumindest für hinreichend kleines h ($h \le h_0$) die Lipschitz-Bedingung (4.4.3) erfüllt.

4.4.1 Implizite Verfahren

Erweitert man $\phi(x_i, \eta_i, h; f)$ in (4.1.5) um das Argument η_{i+1}, gelangt man zu den *impliziten Verfahren*.

Definition 4.4.1 *Das allgemeine implizite Einschrittverfahren lautet*

$$\eta_{i+1} := \eta_i + h\phi(x_i, \eta_i, \eta_{i+1}, h; f). \tag{4.4.1}$$

Ein Beispiel ist das implizite Euler-Verfahren mit

$$\phi(x_i, \eta_i, \eta_{i+1}, h; f) = f(x_i + h, \eta_{i+1}). \tag{4.4.2}$$

Im Weiteren wird angenommen, dass ϕ für $x_i \in I$, $\eta_i \in \mathbb{R}$ (bzw. zusätzlich $\eta_{i+1} \in \mathbb{R}$) und hinreichend kleine h ($0 < h \leq h_0$) definiert ist.

Übungsaufgabe 4.4.2 *ϕ aus (4.4.1) sei bezüglich η_{i+1} Lipschitz-stetig:*

$$|\phi(x_i, \eta_i, \eta_{i+1}, h; f) - \phi(x_i, \eta_i, \hat{\eta}_{i+1}, h; f)| \leq L\,|\eta_{i+1} - \hat{\eta}_{i+1}|\,.$$

Man zeige, dass die Fixpunktgleichung (4.4.1) für $h < 1/L$ eindeutig lösbar ist.

Die eindeutige Lösbarkeit von (4.4.1) vorausgesetzt, kann man implizit die Funktion $\eta_{i+1} = \Psi(x_i, \eta_i, h; f)$ definieren. Indem man in $\phi(x_i, \eta_i, \eta_{i+1}, h; f)$ das Argument durch $\eta_{i+1} = \Psi(x_i, \eta_i, h; f)$ ersetzt, erhält man formal ein explizites Einschrittverfahren (4.1.5) mit $\hat{\phi}(x_i, \eta_i, h; f) := \phi(x_i, \eta_i, \Psi(x_i, \eta_i, h; f), h; f)$. Deshalb darf man sich für die theoretische Betrachtungen auf den expliziten Fall (4.1.5) beschränken. Die einzige Einschränkung ist, dass man die Schrittweiten auf $0 < h \leq h_0$, d.h. hinreichend kleine h beschränken muss. Dies ist aber zumindest bei der theoretischen Untersuchung harmlos, da dort der Grenzprozess $h \to 0$ zu studieren ist.

4.4.2 Lipschitz-Stetigkeit von ϕ

In Analogie zur Lipschitz-Bedingung (4.1.2) werden wir

$$|\phi(x_i, \eta', h; f) - \phi(x_i, \eta'', h; f)| \leq L_\phi\,|\eta' - \eta''| \qquad \text{für alle } x_i \in I,\ \eta', \eta'' \in \mathbb{R},\ h \leq h_0 \tag{4.4.3}$$

als eine essentielle Bedingung an ϕ benötigen.

Übungsaufgabe 4.4.3 *a) Für das Euler- und Heun-Verfahren zeige man, dass (4.4.3) aus (4.1.2) folgt.*
b) Das implizite Euler-Verfahren (4.4.2) führt gemäß der Diskussion am Ende von §4.4.1 zu einem expliziten Verfahren mit $\hat{\phi}(x_i, \eta_i, h; f)$. Man zeige für $\hat{\phi}$ die analoge Aussage für hinreichend kleines h.

4.4.3 Konsistenz

Zunächst wird die Konsistenzbedingung motiviert. Setzt man anstelle der diskreten Lösung η_i von $\eta_{i+1} := \eta_i + h\phi(x_i, \eta_i, h; f)$ die exakte Lösung $y(x_i)$ der Differentialgleichung ein, ergibt sich der sogenannte *lokale Diskretisierungsfehler* τ:

$$y(x_{i+1}) = y(x_i) + h\,[\phi(x_i, y(x_i), h; f) + \tau(x_i, y(x_i); h)]\,.$$

Offenbar wird man vermuten dürfen, dass das Einschrittverfahren (4.1.5) um so besser ist, je kleiner τ ist, denn für $\tau = 0$ würde das Idealergebnis $\eta_i = y(x_i)$ folgen.

Definition 4.4.4 *Für $\xi \in I$ und $\eta \in \mathbb{R}$ bezeichne $Y(\cdot\,; \xi, \eta)$ die Lösung von (4.1.1a) zur Anfangswertbedingung $Y(\xi; \xi, \eta) = \eta$. a) Dann heißt*

$$\tau(x, \eta; h) := \frac{Y(x + h; x, \eta) - \eta}{h} - \phi(x, \eta, h; f) \tag{4.4.4}$$

der "lokale Diskretisierungsfehler" bei (x, η).

b) Das durch ϕ charakterisierte Einschrittverfahren heißt konsistent, falls (für alle $f \in C(I \times \mathbb{R})$)

$$\tau(x, y(x); h) \to 0 \qquad \text{für } h \to 0 \tag{4.4.5}$$

gleichmäßig auf $x \in I$ gilt. Hierbei ist y die Lösung von (4.1.1a,b).

c) Weiterhin heißt ϕ konsistent von der (Konsistenz-)Ordnung p, falls $\tau(x, y(x); h) = \mathcal{O}(h^p)$ gleichmäßig für $h \to 0$ auf $x \in I$ für alle hinreichend glatten f gilt.

Wenn man f als hinreichend glatt voraussetzt[41], kann man den Ausdruck $\frac{y(x+h;x,\eta)-\eta}{h}$ mittels Taylor entwickeln und so äquivalente Konsistenzbedingungen erhalten. Mit

$$y(x+h;x,\eta) = y(x;x,\eta) + hy'(x;x,\eta) + o(h) = \eta + hf(x,\eta) + o(h)$$

wird aus (4.4.5) die Bedingung $\phi(x,\eta,h;f) \to f(x,\eta)$. Man prüft sofort nach, dass diese Bedingung für die Verfahren von Euler und Heun erfüllt sind.

Die simple Extrapolation $\eta_{i+1} := \eta_i$ (d.h. $\phi = 0$) führt dagegen im Allgemeinen zu $\tau(x,\eta;h) = \mathcal{O}(1)$ und ist nicht konsistent.

4.4.4 Konvergenz

Wir erinnern an die Schreibweise $\eta_i = \eta(x_i,h)$ und die Zielvorstellung, dass $\eta(x,h) \approx y(x)$. Beim Grenzübergang $h \to 0$ beschränken wir uns stillschweigend auf (eine Teilfolge von) $h_n := (x-x_0)/n$, da dann $x = nh_n$ stets zum Gitter gehört, auf dem $\eta(\cdot,h_n)$ definiert ist.

Definition 4.4.5 (Konvergenz) *Das Einschrittverfahren heißt konvergent, falls für alle Lipschitz-stetigen f und alle $x \in I$ gilt, dass $\lim_{h\to 0}\eta(x,h) = y(x)$ (y Lösung von (4.1.1a,b)). Ein Einschrittverfahren hat die Konvergenzordnung p, falls $\eta(x,h) = y(x) + \mathcal{O}(h^p)$ für hinreichend glatte f.*

4.4.5 Stabilität

Die Konsistenz kontrolliert den Fehler im i-ten Schritt von x_i nach x_{i+1} unter der Voraussetzung, dass η_i der exakter Startwert ist. Beim Start ist tatsächlich $\eta_0 = y_0$ exakt, sodass der Fehler $\varepsilon_1 := \eta_1 - y_1$ gemäß Bedingung (4.4.5) von der Größe $o(h)$ bzw. $\mathcal{O}(h^{p+1})$ ist.

Das in den Schritten für $i \geq 1$ entstehende Problem ist, dass sich z.B. der Konsistenzfehler bei x_1 auf die Größen $\eta_2, \eta_3, \ldots$ fortpflanzt. Da man $x = x_n$ erreichen will, hat man $n = \mathcal{O}(1/h)$ Schritt durchzuführen. Würde sich der Fehler pro Schritt um einem Faktor $c > 1$ (c unabhängig von h) verstärken, hätte η_n einen Fehler $\mathcal{O}(c^n) = \mathcal{O}(c^{1/h})$, der offenbar exponentiell mit $h \to \infty$ explodieren würde. Zusätzlich ist zu beachten, dass nicht nur die Verstärkung des Konsistenzfehlers ε_1 auftritt, sondern bei jedem x_i neue Konsistenzfehler entstehen.

Wie werden aber zeigen, dass dank der Lipschitz-Bedingung (4.4.3) die Fehler im gewünschten Rahmen bleiben.

Lemma 4.4.6 (Stabilität von Einschrittverfahren) *Es gelte die Lipschitz-Bedingung (4.4.3) mit L_ϕ. Die lokalen Diskretisierungsfehler seien durch $|\tau(x_i,y(x_i);h)| \leq T_h$ beschränkt (vgl. (4.4.4)). Dann ist der globale Diskretisierungsfehler beschränkt durch*

$$|\eta(x,h) - y(x)| \leq T_h \frac{e^{(x-x_0)L_\phi} - 1}{L_\phi}. \tag{4.4.6}$$

Beweis. Sei $\delta_i := |\eta_i - y(x_i)|$ der *globale Fehler*. Der lokale Diskretisierungsfehler sei mit $\tau_i = \tau(x_i,y(x_i);h)$ bezeichnet. Startend mit $\delta_0 = 0$ erhalten wir die Rekursionsformel

$$\begin{aligned}
\delta_{i+1} &= |\eta_{i+1} - y(x_{i+1})| = |\eta_i + h\phi(x_i,\eta_i,h;f) - y(x_{i+1})| \\
&= \left|\eta_i - y(x_i) - h\left[\frac{y(x_{i+1})-y(x_i)}{h} - \phi(x_i,\eta_i,h;f)\right]\right| \\
&= \left|\eta_i - y(x_i) - h\left[\frac{y(x_{i+1})-y(x_i)}{h} - \phi(x_i,y(x_i),h;f)\right] + h\left[\phi(x_i,\eta_i,h;f) - \phi(x_i,y(x_i),h;f)\right]\right| \\
&\leq |\eta_i - y(x_i)| + h\left|\frac{y(x_{i+1})-y(x_i)}{h} - \phi(x_i,y(x_i),h;f)\right| + h\left|\phi(x_i,\eta_i,h;f) - \phi(x_i,y(x_i),h;f)\right| \\
&\leq \delta_i + h\left|\tau_i\right| + hL_\phi\delta_i = (1+hL_\phi)\,\delta_i + hT_h\,,
\end{aligned}$$

[41] Ohne die Voraussetzung der p-fachen stetigen Differenzierbarkeit von f läßt sich $\tau(x,\eta;h) = \mathcal{O}(h^p)$ nicht verifizieren.

die mit (4.2.1) übereinstimmen, wenn wir dort $a_\nu = \delta_\nu, h = h, L = L_\phi, k = 1, B = T_h$ wählen. Lemma 4.2.2 beweist $\delta_\nu \le T_h \dfrac{e^{\nu h L_\phi} - 1}{L_\phi}$ wegen $a_0 = \delta_0 = 0$. ∎

Noch wurde die Konsistenz nicht vorausgesetzt. Mit ihrer Hilfe folgt die Konvergenz. Zudem stimmen Konsistenzordnung und Konvergenzordnung überein.

Satz 4.4.7 *Das Einschrittverfahren (4.1.5) erfülle die Lipschitz-Bedingung (4.4.3) und sei konsistent. Dann ist es auch konvergent:* $\lim_{h\to 0} \eta(x,h) = y(x)$. *Liegt darüberhinaus die Konsistenzordnung p vor, so folgt auch Konvergenz von der Ordnung p.*

Beweis. Im Falle der einfachen Konsistenz gilt $T_h \to 0$, sodass die Konvergenz aus (4.4.6) folgt. Bei Konsistenz von der Ordnung p ist $T_h \le Ch^p$ und damit auch $|\eta(x,h) - y(x)| \le \mathcal{O}(h^p)$. ∎

Es sei angemerkt, dass im Allgemeinen die Lipschitz-Bedingung (4.4.3) nur lokal gilt. In diesem Fall geht man wie folgt vor. $G := \{(x,y) : x \in [x_0, x_E], |y - y(x)| \le 1\}$ ist kompakt. Es reicht, (4.4.3) auf G zu fordern[42]. Für hinreichend kleines h ist $T_h \frac{e^{(x-x_0)L_\phi}-1}{L_\phi}$ in (4.4.6) durch 1 beschränkt und damit $(x, \eta(x,h)) \in G$. Ein Blick auf den Beweis von Lemma 4.4.6 zeigt, dass alle zwischenzeitlichen Argumente in G liegen und somit (4.4.3) anwendbar ist.

4.5 Analyse von Mehrschrittverfahren

4.5.1 Lokaler Diskretisierungsfehler, Konsistenz

Indem wir formal $\alpha_r := 1$ zusätzlich einführen, schreibt sich das r-Schrittverfahren als

$$\sum_{\nu=0}^{r} \alpha_\nu \eta_{j+\nu} = h\phi(x_j, \eta_{j+r-1}, \dots, \eta_j, h; f)\,. \tag{4.5.1}$$

Der *lokale Diskretisierungsfehler* ist

$$\tau(x,y;h) := \frac{1}{h}\Big[\sum_{\nu=0}^{r} \alpha_\nu Y(x_{j+\nu}; x_j, y) - h\phi\Big(x_j, Y(x_{j+\nu-1}; x_j, y), \dots, \underbrace{Y(x_j; x_j, y)}_{=y}, h; f\Big)\Big], \tag{4.5.2}$$

wobei $Y(x;\xi,\eta)$ wie in Definition 4.4.4 erklärt ist.

Definition 4.5.1 *Ein Mehrschrittverfahren heißt konsistent, wenn für alle* $f \in C(I \times \mathbb{R})$ *mit (4.1.2) gilt, dass* $\sup_{x\in I} \tau(x, y(x); h) \to 0$ *für* $h \to 0$. *Hierbei ist* $y(x)$ *die Lösung von (4.1.1a,b). Weiterhin heißt (4.5.1) konsistent von der (Konsistenz-)Ordnung p, falls* $|\tau(x, y(x); h)| = \mathcal{O}(h^p)$ *für hinreichend glatte* f.

Wählt man insbesondere $f = 0$ und den Anfangswert $y_0 = 1$, so lautet die Lösung $y(x) = 1$ und $\tau(x, y(x); h) \to 0$ vereinfacht sich in diesem Spezialfall zu $\left(\sum_{\nu=0}^{r} \alpha_\nu - h\phi\right)/h \to 0$, was $\sum_{\nu=0}^{r} \alpha_\nu = 0$ – also die Bedingung (4.1.7) – impliziert.

4.5.2 Konvergenz

Anders als beim Einschrittverfahren können wir nicht von exakten Startwerten $\eta_1, \dots, \eta_{r-1}$ ausgehen. Wir werden deshalb alle als gestört annehmen:

$$\eta_j = y(x_j) + \varepsilon_j, \qquad \vec{\varepsilon} = (\varepsilon_j)_{j=0,\dots,r-1}\,.$$

Die Lösung zu diesen Startwerten schreiben wir als $\eta(x; \vec{\varepsilon}, h)$.

Definition 4.5.2 *Ein Mehrschrittverfahren heißt* konvergent, *falls für alle* $f \in C(I \times \mathbb{R})$ *mit (4.1.2) und alle Startwerte* y_0 *gilt, dass* $\sup_{x\in I} |\eta(x; \vec{\varepsilon}, h) - y(x)| \to 0$ *für* $h \to 0$ *und* $\|\vec{\varepsilon}\|_\infty \to 0$.

[42] Lokal Lipschitz-stetige Funktionen sind auf einem Kompaktum gleichmäßig Lipschitz-stetig.

Eine weitergehende Forderung ist, dass wir auch die Gleichungen (4.5.1) um ε_j $(j \geq r)$ stören:

$$\sum_{\nu=0}^{r} \alpha_\nu \eta_{j+\nu} = h\phi(x_j, \eta_{j+r-1}, \ldots, \eta_j, h; f) + h\varepsilon_{j+r} \qquad \text{für } j \geq 0. \tag{4.5.3}$$

In diesem Fall ist $\vec{\varepsilon} = (\varepsilon_j)_{j\geq 0}$ ein Tupel mit soviel Elementen, wie Gitterpunkte existieren (man beachte, dass die Größen ε_j für $j < r$ und $j \geq r$ eine ganz unterschiedliche Bedeutung haben!). Wieder kann $\eta(x; \vec{\varepsilon}, h) \to y(x)$ für $h \to 0$ und $\|\vec{\varepsilon}\|_\infty \to 0$ gefordert werden.

4.5.3 Stabilität

Die Koeffizienten α_ν aus (4.5.1) definieren das charakteristische Polynom

$$\psi(\zeta) := \sum_{\nu=0}^{r} \alpha_\nu \zeta^\nu. \tag{4.5.4}$$

Definition 4.5.3 *Das Mehrschrittverfahren (4.5.1) heißt* stabil, *falls für alle Nullstellen ζ des charakteristischen Polynoms ψ gilt: Entweder ist $|\zeta| < 1$ oder ζ ist eine einfache Nullstelle mit $|\zeta| = 1$.*

Wir prüfen einige Spezialfälle:

- Für die Mittelpunktsregel (4.1.8) ergibt sich $\psi(\zeta) = \zeta^2 - 1$. Beide Nullstellen $\zeta = \pm 1$ sind einfach mit $|\zeta| = 1$. Daher ist die Mittelpunktsregel stabil.
- Dem Zweischrittverfahren (4.1.9) entspricht $\psi(\zeta) = \zeta^2 - 2\zeta + 1 = (\zeta - 1)^2$. Damit ist $\zeta = 1$ eine doppelte Nullstelle und führt zur Instabilität.
- Im Falle eines Einschrittverfahrens (d.h. $r = 1$) ist wegen $\alpha_r = 1$ und (4.1.7) nur $\psi(\zeta) = \zeta - 1$ möglich. Die einzige Nullstelle $\zeta = 1$ erfüllt die zweite Bedingung in Definition 4.5.3. Dies beweist die folgende Bemerkung, die dem Resultat aus Lemma 4.4.6 entspricht.

Bemerkung 4.5.4 *Einschrittverfahren sind immer stabil im Sinne der Definition 4.5.3.*

Der Zusammenhang zwischen der Stabilitätsbedingung aus Definition 4.5.3 und dem Mehrschrittverfahren (4.5.1) ist nicht unmittelbar einsichtig. Das Verbindungsglied bilden die Differenzengleichungen, die in §4.5.5 untersucht werden. Als Vorbereitung werden anschließend die potenzbeschränkten Matrizen diskutiert.

4.5.4 Potenzbeschränkte Matrizen

Definition 4.5.5 *Sei $\|\cdot\|$ eine Matrixnorm. Eine $d \times d$-Matrix A ist eine potenzbeschränkte Matrix, wenn*

$$\sup\{\|A^n\| : n \in \mathbb{N}\} < \infty. \tag{4.5.5}$$

Man beachte, dass wegen der Normäquivalenz die Wahl der Matrixnorm $\|\cdot\|$ für die Definition irrelevant ist. Eine Charakterisierung der potenzbeschränkten Matrizen gibt der

Satz 4.5.6 *A ist genau dann eine potenzbeschränkte Matrix, wenn für seine Eigenwerte gilt: entweder a) $|\lambda| < 1$ oder b) $|\lambda| = 1$ und λ hat übereinstimmende algebraische und geometrische Vielfachheit.*[43]

Eine weitere äquivalente Formulierung gibt das

Lemma 4.5.7 *A ist genau dann eine potenzbeschränkte Matrix, wenn es eine zugeordnete Matrixnorm*[44] *gibt, sodass $\|A\| \leq 1$.*

[43] λ hat die algebraische Vielfachheit $k \in \mathbb{N}_0$, falls das charakteristische Polynom $\det(xI - A)$ den Faktor $(x - \lambda)^k$, aber nicht $(x - \lambda)^{k+1}$ enthält. λ hat die geometrische Vielfachheit $k \in \mathbb{N}_0$, falls $\dim\{e \in C^d : Ae = \lambda e\} = k$. Es gilt stets geometrische Vielfachheit $\leq$ algebraische Vielfachheit.

[44] $\|\cdot\|$ ist zugeordnete Matrixnorm, falls es eine Vektornorm $|||\cdot|||$ gibt mit $\|A\| = \sup\{|||Ax||| \,/\, |||x|||: x \neq 0\}$.

Den Beweis von Satz 4.5.6 und Lemma 4.5.7 geben wir gemeinsam, indem wir den folgenden Ringschluss benutzen: (4.5.5) $\Rightarrow$ Charakterisierung aus Satz 4.5.6 $\Rightarrow$ Charakterisierung aus Lemma 4.5.7 $\Rightarrow$ (4.5.5).

Beweis. Sei $C_{\text{stab}} := \sup\{\|A^n\| : n \in \mathbb{N}\}$, falls (4.5.5) zutrifft. $\sigma(M) := \{\lambda \in \mathbb{C} : \lambda \text{ Eigenwert von } M\}$ bezeichne das Spektrum einer quadratischen Matrix M.

a) "(4.5.5) $\Rightarrow$ Charakterisierung aus Satz 4.5.6": Sei $\|\cdot\|$ als zugeordnete Matrixnorm angenommen. Für diese gilt stets: $|\lambda| \le \|B\|$ für alle $\lambda \in \sigma(B)$. Mit $B = A^n$ erhalten wir $|\lambda^n| = |\lambda|^n \le \|A^n\| \le C_{\text{stab}}$ für alle $n \in \mathbb{N}$, also $|\lambda| \le 1$. Sei nun $|\lambda| = 1$ angenommen. Wenn λ eine höhere algebraische als geometrische Vielfachheit besätze, gäbe es einen Eigenvektor $e \neq 0$ und einen Hauptvektor h, sodass $Ae = \lambda e$ und $Ah = e + \lambda h$. Hieraus folgt $A^n h = \lambda^{n-1}(ne + \lambda h)$ und $\|A^n h\| = \|ne + \lambda h\| \ge n\|e\| - \|\lambda h\| \to \infty$ für $n \to \infty$ im Widerspruch zu $\|A^n h\| \le C_{\text{stab}} \|h\|$. Also muss die Charakterisierung aus Satz 4.5.6 zutreffen.

b) "Charakterisierung aus Satz 4.5.6 $\Rightarrow$ Charakterisierung aus Lemma 4.5.7": Die Eigenwerte λ_i von A seien in der Reihenfolge $|\lambda_1| \le |\lambda_2| \le \ldots < |\lambda_{d-m+1}| = \ldots = |\lambda_d| = 1$ sortiert, wobei $m \ge 0$ die Zahl der Nullstellen mit Betrag eins ist. Sei T die Transformation auf Jordan-Normalform:

$$J = T^{-1}AT = \begin{bmatrix} J_1 & 0 \\ 0 & D \end{bmatrix} \text{ mit } J_1 = \begin{bmatrix} \lambda_1 & * & & \\ & \lambda_2 & * & \\ & & \ddots & * \\ & & & \lambda_{d-m} \end{bmatrix},\ D = \operatorname{diag}\{\lambda_{d-m+1}, \ldots, \lambda_d\},$$

wobei die Einträge $*$ entweder null oder eins sind. Da für $\lambda_{d-m+1}, \ldots, \lambda_d$ die algebraischen und geometrischen Vielfachheiten übereinstimmen, ist D eine diagonale $m \times m$-Matrix, während alle Eigenwerte λ_i ($i = 1, \ldots, r-m$) einen Betrag < 1 haben. Sei $\Delta_\varepsilon := \operatorname{diag}\{1, \varepsilon, \varepsilon^2, \ldots, \varepsilon^{r-1}\}$ mit $\varepsilon \in (0, 1 - |\lambda_{r-m}|]$. Man prüft nach, dass $\Delta_\varepsilon^{-1} J \Delta_\varepsilon$ die Zeilensummennorm $\|\Delta_\varepsilon^{-1} J \Delta_\varepsilon\|_\infty \le 1$ hat. Damit liefert die Transformation mit $S := T\Delta_\varepsilon$ die Norm $\|S^{-1}AS\|_\infty \le 1$. Man prüft nach, dass $\|A\| := \|S^{-1}AS\|_\infty$ die zugeordnete Matrixnorm zur Vektornorm $\|x\| := \|Sx\|_\infty$ ist.

c) "Charakterisierung aus Lemma 4.5.7 $\Rightarrow$ (4.5.5)": Für zugeordnete Matrixnormen gilt die Submultiplikativität $\|A^n\| \le \|A\|^n$, sodass man aus $\|A\| \le 1$ auf $C_{\text{stab}} = 1 < \infty$ schließt. ∎

4.5.5 Differenzengleichungen

4.5.5.1 Lösungsraum $\mathcal{F}_0$

Mit $\mathbf{x} = (x_j)_{j \in \mathbb{N}_0}$ werden Folgen komplexer Zahlen bezeichnet. $\mathcal{F} = \mathbb{C}^{\mathbb{N}_0}$ sei die Menge aller Folgen. Wir suchen die Folgen $\mathbf{x} \in \mathcal{F}$, die die folgende Differenzengleichung erfüllen:

$$\sum_{\nu=0}^{r} \alpha_\nu x_{j+\nu} = 0 \qquad \text{für alle } j \ge 0, \text{ wobei } \alpha_r = 1. \tag{4.5.6}$$

Lemma 4.5.8 *a) $\mathcal{F}$ bildet einen linearen Vektorraum (mit komponentenweiser Addition und Multiplikation).*
b) Die Teilmenge $\mathcal{F}_0 := \{\mathbf{x} \in \mathcal{F}$ erfüllt (4.5.6)$\}$ ist ein linearer Teilraum von $\mathcal{F}$ mit der Dimension r.

Beweis. a) Dass $\mathcal{F}$ und $\mathcal{F}_0$ lineare Räume bilden ist trivial. Es bleibt $\dim \mathcal{F}_0 = r$ zu beweisen.

b) Wir definieren $\mathbf{x}^{(i)} \in \mathcal{F}$ für $i = 0, 1, \ldots, r-1$ durch die Anfangswerte $x_j^{(i)} = \delta_{ij}$ für $j \in \{0, \ldots, r-1\}$. Für $j \ge r$ kann (4.5.6) benutzt werden, um

$$x_j^{(i)} := \sum_{\nu=0}^{r-1} \alpha_\nu x_{j-r+\nu}^{(i)} \qquad \text{für alle } j \ge r \tag{4.5.7}$$

zu definieren. Damit erfüllt $\mathbf{x}^{(i)}$ die Differenzengleichung (4.5.6), d.h. $\mathbf{x}^{(i)} \in \mathcal{F}_0$ für $i = 0, 1, \ldots, r-1$.

c) Zum Nachweis der linearen Unabhängigkeit nehmen wir $\sum_i \beta_i \mathbf{x}^{(i)} = 0$ an. Auswertung der Komponenten $j \in \{0, \ldots, r-1\}$ liefert $0 = \sum_i \beta_i x_j^{(i)} = \sum_i \beta_i \delta_{ij} = \beta_j$, was die lineare Unabhängigkeit beweist.

d) Zu jedem $\mathbf{x} \in \mathcal{F}_0$ lässt sich $\mathbf{y} := \mathbf{x} - \sum_{i=0}^{r-1} x_i \mathbf{x}^{(i)} \in \mathcal{F}_0$ definieren, für das definitionsgemäß $y_0 = y_1 = \ldots = y_{r-1} = 0$ gilt. Analog zu (4.5.7) erzeugen diese Anfangswerte $y_j = 0$ für alle $j \ge r$. $\mathbf{y} = 0$ beweist, dass $\{\mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(r-1)}\}$ bereits $\mathcal{F}_0$ aufspannt. Also ist $\dim \mathcal{F}_0 = r$. ∎

4.5.5.2 Darstellung der Lösungen

Bemerkung 4.5.9 *a) Das Polynom $\psi(\zeta) = \sum_{\nu=0}^{r} \alpha_\nu \zeta^\nu$ habe die Nullstelle $\zeta_0 \in \mathbb{C}$. Dann ist $\mathbf{x} = (\zeta_0^j)_{j\in\mathbb{N}_0}$ eine Lösung von (4.5.6).*

b) ψ habe r verschiedene Nullstellen $\zeta_i \in \mathbb{C}$, $i = 1, \ldots, r$. Dann spannen die Lösungen $\mathbf{x}^{(i)} = (\zeta_i^j)_{j\in\mathbb{N}_0}$ den Raum $\mathcal{F}_0$ auf.

Beweis. a) Einsetzen von $x_j = \zeta_0^j$ in (4.5.6) liefert

$$\sum_{\nu=0}^{r} \alpha_\nu x_{j+\nu} = \sum_{\nu=0}^{r} \alpha_\nu \zeta_0^{j+\nu} = \zeta_0^j \sum_{\nu=0}^{r} \alpha_\nu \zeta_0^\nu = \zeta_0^j \psi(\zeta_0) = 0,$$

d.h. $\mathbf{x} \in \mathcal{F}_0$.

b) Nach Teil a) gilt $\mathbf{x}^{(i)} \in \mathcal{F}_0$. Man zeigt leicht, dass die $\mathbf{x}^{(i)}$ linear unabhängig sind. Wegen $\dim \mathcal{F}_0 = r$ bilden die $\mathbf{x}^{(i)}$, $i = 1, \ldots, r$, eine Basis. ∎

Im Falle der Bemerkung 4.5.9 liegen einfache Nullstellen vor. Es bleibt der Fall mehrfacher Nullstellen zu diskutieren.

Es sei daran erinnert, dass ψ genau dann eine (mindestens) k-fache Nullstelle ζ_0 besitzt, wenn $\psi(\zeta_0) = \psi'(\zeta_0) = \ldots = \psi^{(k-1)}(\zeta_0) = 0$. Mit der Leibniz-Regel erhält man $\left(\frac{d}{d\zeta}\right)^\ell \left(\zeta^j \psi(\zeta)\right) = 0$ für alle $0 \le \ell \le k-1$. Die explizite Darstellung von $(\frac{d}{d\zeta})^\ell \left(\zeta^j \psi(\zeta)\right)$ lautet

$$\left(\zeta^j \psi(\zeta)\right)^{(\ell)}\Big|_{\zeta=\zeta_0} = \sum_{\nu=0}^{r} \alpha_\nu \zeta_0^{j+\nu-\ell} \, (j+\nu)\,(j+\nu-1) \times \ldots \times (j+\nu-\ell+1)\,. \tag{4.5.8}$$

Man definiere $\mathbf{x}$ mittels $x_j = \zeta_0^j j\,(j-1) \times \ldots \times (j-\ell+1)$ mit $0 \le \ell \le k-1$. Einsetzen in die Differenzengleichung (4.5.6) liefert $\sum_{\nu=0}^{r} \alpha_\nu x_{j+\nu} = \sum_{\nu=0}^{r} \alpha_\nu \zeta_0^{j+\nu} \, (j+\nu)\,(j+\nu-1) \times \ldots \times (j+\nu-\ell+1)$. Dieser Ausdruck ist das Produkt von (4.5.8) mit ζ_0^ℓ, also $\sum_{\nu=0}^{r} \alpha_\nu x_{j+\nu} = 0$, d.h. $\mathbf{x} \in \mathcal{F}_0$.

Bemerkung 4.5.10 *Sei $\zeta_0 \neq 0$ eine Nullstelle von ψ mit der Vielfachheit k. a) Dann sind $\mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(k-1)}$ mit $x_j^{(\ell)} = \zeta_0^j \prod_{\nu=0}^{\ell-1}(j-\nu)$ $(\ell = 0, \ldots, k-1)$ k linear unabhängige Lösungen von (4.5.6).*

b) Ebenso bilden $\hat{x}_j^{(\ell)} = \zeta_0^j j^\ell$ k linear unabhängige Lösungen von (4.5.6).

Beweis. a) $\mathbf{x}^{(\ell)} \in \mathcal{F}_0$ ist bereits gezeigt worden. Die lineare Unabhängigkeit für $\ell = 0, \ldots, k-1$ folgt z.B. aus dem unterschiedlichen Wachstum von $x_j^{(\ell)}/\zeta_0^j$ für $j \to \infty$.

b) $\left\{\prod_{\nu=0}^{\ell-1}(x-\nu) : \ell = 0, \ldots, k-1\right\}$ ist ebenso wie $\left\{x^\ell : \ell = 0, \ldots, k-1\right\}$ eine Basis der Polynome vom Grad $\le k-1$. Daher spannen $\left\{\mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(k-1)}\right\}$ und $\left\{\hat{\mathbf{x}}^{(0)}, \ldots, \hat{\mathbf{x}}^{(k-1)}\right\}$ die gleichen Räume auf. ∎

Der Fall $\zeta_0 = 0$ ist in Bemerkung 4.5.10 ausgenommen, da die bisherige Definition zu $x_j^{(\ell)} = 0$ für $j \ge \min\{1-\ell, 0\}$ und damit nicht zu linear unabhängigen Lösungen führt.

Bemerkung 4.5.11 *Sei $\zeta_0 = 0$ eine k-fache Nullstelle von ψ. Dann sind $\mathbf{x}^{(i)}$ mit $x_j^{(i)} = (\delta_{ij})_{j\in\mathbb{N}_0}$ $(i = 0, \ldots, k-1)$ k linear unabhängige Lösungen von (4.5.6).*

Beweis. Aus $\psi(\zeta_0) = \psi'(\zeta_0) = \ldots = \psi^{(k-1)}(\zeta_0) = 0$ für $\zeta_0 = 0$ erhält man $\alpha_0 = \ldots = \alpha_{k-1} = 0$. Damit ist $\sum_{\nu=0}^{r} \alpha_\nu x_{j+\nu}^{(i)} = \sum_{\nu=k}^{r} \alpha_\nu x_{j+\nu}^{(i)}$. Definitionsgemäß ist aber $x_{j+\nu}^{(i)} = 0$ für $\nu \ge k$, d.h. (4.5.6) ist erfüllt. Offenbar sind die $\mathbf{x}^{(i)}$ linear unabhängig. ∎

Satz 4.5.12 *Das Polynom $\psi(\zeta) = \sum_{\nu=0}^{r} \alpha_\nu \zeta^\nu$ habe die verschiedenen Nullstellen ζ_i $(i = 1, \ldots, m)$ mit den jeweiligen Vielfachheiten k_i. Zu ζ_i gehören jeweils k_i Lösungen von (4.5.6), die im Falle von $\zeta_i \neq 0$ in Bemerkung 4.5.10 und für $\zeta_i = 0$ in Bemerkung 4.5.11 definiert sind. Insgesamt ergeben sich r linear unabhängige Lösungen, die $\mathcal{F}_0$ aufspannen.*

Beweis. Die lineare Unabhängigkeit der konstruierten Lösungen wird als Übungsaufgabe überlassen. Die Gesamtzahl der gefundenen Lösungen ist $\sum k_i = r$. Da nach Lemma 4.5.8 $\dim \mathcal{F}_0 = r$ gilt, ist eine Basis von $\mathcal{F}_0$ gefunden. ∎

4.5.5.3 Stabilität

Definition 4.5.13 *Die Differenzengleichung (4.5.6) heißt* stabil, *wenn jede Lösung von (4.5.6) in der Supremumsnorm beschränkt ist:*

$$\|\mathbf{x}\|_\infty := \sup_{j\in\mathbb{N}_0} |x_j| < \infty \qquad \textit{für alle } \mathbf{x} \in \mathcal{F}_0.$$

Bemerkung 4.5.14 *Äquivalent zur Definition 4.5.13 ist: Es gibt eine Konstante C, sodass*

$$\|\mathbf{x}\|_\infty \le C \max_{j=0,\dots,r-1} |x_j| \qquad \textit{für alle } \mathbf{x} \in \mathcal{F}_0. \tag{4.5.9}$$

Beweis. a) Gilt (4.5.9), so folgt $\|\mathbf{x}\|_\infty < \infty$ aus der Tatsache, dass $\max_{j=0,\dots,r-1} |x_j|$ immer endlich ist.

b) Wir wählen die Basis $\mathbf{x}^{(i)} \in \mathcal{F}_0$ wie in Beweisteil b) des Lemmas 4.5.8. Für diese ist $\mathbf{x} = \sum_{i=0}^{r-1} x_i \mathbf{x}^{(i)}$. Gilt Stabilität im Sinne von Definition 4.5.13, so sind $C_i := \|\mathbf{x}^{(i)}\|_\infty < \infty$ und damit auch $C := \sum_{i=0}^{r-1} C_i < \infty$. Mit diesem C folgt $\|\mathbf{x}\|_\infty = \|\sum_{i=0}^{r-1} x_i \mathbf{x}^{(i)}\|_\infty \le \sum_{i=0}^{r-1} |x_i| \, \|\mathbf{x}^{(i)}\|_\infty = \sum_{i=0}^{r-1} C_i \, |x_i| \le C \max_{j=0,\dots,r-1} |x_j|$, d.h. (4.5.9). ∎

Offenbar sind alle Lösungen von (4.5.6) genau dann beschränkt, wenn alle in Satz 4.5.12 angegebenen (Basis-) Lösungen beschränkt sind. Die folgende vollständige Fallunterscheidung bezieht sich auf die Nullstellen ζ_i von ψ und ihre Vielfachheit k_i.

1. Fall $|\zeta_i| < 1$: Alle Folgen $(\zeta_i^j j^\ell)_{j\in\mathbb{N}_0}$ mit $0 \le \ell < k_i$ sind Nullfolgen und daher beschränkt.
2. Fall $|\zeta_i| > 1$: Für alle Folgen gilt $\lim |\zeta_i^j j^\ell| = \infty$, d.h. sie sind unbeschränkt.
3. Fall $|\zeta_i| = 1$ und $k_i = 1$ (einfache Nullstelle): $(\zeta_i^j)_{j\in\mathbb{N}_0}$ ist betragsmäßig durch 1 beschränkt.
4. Fall $|\zeta_i| = 1$ und $k_i > 1$ (mehrfache Nullstelle): Für $1 \le \ell \le k_i - 1$ gilt $\lim |\zeta_i^j j^\ell| = \infty$, d.h. die Folgen sind unbeschränkt.

Damit charakterisieren der 1. und 3. Fall die stabile Situation, während der 2. und 4. Fall zur Instabilität führt. Dies beweist den

Satz 4.5.15 *Genau dann, wenn ψ die Stabilitätsbedingung aus Definition 4.5.3 erfüllt, ist die Differenzengleichung (4.5.6) stabil.*

4.5.5.4 Begleitmatrix

Definition 4.5.16 *Die Begleitmatrix zum Polynom $\psi(\zeta) = \sum_{\nu=0}^{r} \alpha_\nu \zeta^\nu$ hat das Format $r \times r$ und lautet*

$$A = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -\alpha_0 & \dots & -\alpha_{r-2} & -\alpha_{r-1} \end{bmatrix}. \tag{4.5.10}$$

Bemerkung 4.5.17 *a)* $\det(\zeta I - A) = \psi(\zeta)$. *b) Jede Lösung $(x_j)_{j\in\mathbb{N}_0}$ der Differenzengleichung (4.5.6) erfüllt (4.5.11) und umgekehrt:*

$$\begin{bmatrix} x_{j+1} \\ x_{j+2} \\ \vdots \\ x_{j+r} \end{bmatrix} = A \begin{bmatrix} x_j \\ x_{j+1} \\ \vdots \\ x_{j+r-1} \end{bmatrix}. \tag{4.5.11}$$

Mit (4.5.11) können wir das r-Schrittverfahren formal umformulieren als ein Einschrittverfahren für die r-Tupel $X_j := (x_j, \dots, x_{j+r-1})^\top$. Das Stabilitätsverhalten muss daher mit Eigenschaften von A ausgedrückt werden können. Ein offensichtlicher Zusammenhang wird in der nächsten Bemerkung beschrieben.

Bemerkung 4.5.18 *Es bezeichne $\|\cdot\|$ eine Vektornorm und die zugeordnete Matrixnorm. Das Differenzengleichung (4.5.6) ist genau dann stabil, wenn A eine potenzbeschränkte Matrix ist.*

Beweis. Die Tupel X_j erfüllen $X_j = AX_{j-1}$ (vgl. (4.5.11)). Insbesondere ist $X_n = A^n X_0$. Gleichmäßige Beschränktheit von $|x_j|$ und $\|X_j\|$ sind äquivalent. Wenn (4.5.6) stabil ist, ist $\|A^n X_0\|$ für alle X_0 mit $\|X_0\| \le 1$ und alle $n \in \mathbb{N}$ gleichmäßig beschränkt, d.h. A ist eine potenzbeschränkte Matrix. Umgekehrt liefert $C_{\text{stab}} := \sup\{\|A^n\| : n \in \mathbb{N}\} < \infty$ die Abschätzung $\|X_n\| \le C_{\text{stab}} \|X_0\|$ und damit die Stabilität. ∎

Mit Lemma 4.5.7 erhalten wir die Aussage:

Lemma 4.5.19 *Das Differenzengleichung (4.5.6) ist genau dann stabil, wenn es eine zugeordnete Matrixnorm gibt, sodass $\|A\| \le 1$ für die Begleitmatrix A aus (4.5.10) gilt.*

4.5.5.5 Abschätzungen inhomogener Lösungen

Satz 4.5.20 *Die Differenzengleichung sei stabil. Für die Anfangswerte gelte $|x_j| \le \alpha$ $(0 \le j \le r-1)$. Die Folge x_j erfülle die inhomogene Differenzengleichung $\sum_{\nu=0}^{r} \alpha_\nu x_{j+\nu} = \beta_{j+r}$, wobei*

$$|\beta_{j+r}| \le \beta + \gamma \max\{|x_\mu| : 0 \le \mu \le j+r-1\}$$

für ein $\gamma \ge 0$ gelte. Dann existieren k, k', sodass

$$|x_j| \le kk'\alpha e^{jk\gamma} + \begin{cases} jk\beta & \text{für } \gamma = 0, \\ \frac{\beta}{\gamma}\left(e^{j\gamma k} - 1\right) & \text{für } \gamma > 0. \end{cases}$$

Beweis. a) A sei die Begleitmatrix. Mit $\|\cdot\|$ wird die Vektornorm in $\mathbb{R}^r$ wie auch die zugeordnete Matrixnorm $\|\cdot\|$ aus Lemma 4.5.19 bezeichnet. Wegen der Normäquivalenz gilt für geeignete k, k'

$$\|X\|_\infty \le k\,\|X\|, \quad \|X\| \le k'\,\|X\|_\infty \qquad \text{für alle } X \in \mathbb{R}^r.$$

b) Setze $X_j := (x_j, \ldots, x_{j+r-1})^\top$ und $\mathbf{e} = (0, \ldots, 0, 1)^\top \in \mathbb{R}^r$. Gemäß Lemma 4.5.19 gilt $\|A\| \le 1$. Da eine Skalierung der Vektornorm die zugeordnete Matrixnorm nicht ändert, kann o.B.d.A. $\|\mathbf{e}\| = 1$ angenommen werden. Die Differenzengleichung $\sum_{\nu=0}^{r} \alpha_\nu x_{j+\nu} = \beta_{j+r}$ ist äquivalent zu

$$X_{j+1} = AX_j + \beta_{j+r}\mathbf{e}.$$

c) Setze $\xi_j := \max\{|x_\mu| : 0 \le \mu \le j\}$. Nach Definition von X_μ gilt auch $\xi_j = \max_{0 \le \mu \le j-r+1} \|X_\mu\|_\infty$ für $j \ge r-1$. Es folgt

$$\|X_{j+1}\| = \|AX_j + \beta_{j+r}\mathbf{e}\| \le \underbrace{\|A\|}_{\le 1}\|X_j\| + |\beta_{j+r}|\underbrace{\|\mathbf{e}\|}_{=1} \le \|X_j\| + |\beta_{j+r}| \le \|X_j\| + \beta + \gamma\xi_{j+r-1}.$$

Man definiere η_j durch

$$\eta_0 := \|X_0\|, \qquad \eta_{j+1} := \eta_j + \beta + \gamma\xi_{j+r-1}. \tag{4.5.12}$$

Offenbar gilt $\|X_j\| \le \eta_j$ und $\eta_{j+1} \ge \eta_j$ $(j \ge 0)$. Die Abschätzung

$$\xi_{j+r-1} = \max_{0 \le \mu \le j+r-1} |x_\mu| = \max_{0 \le \mu \le j} \|X_\mu\|_\infty \le \max_{0 \le \mu \le j} k\,\|X_\mu\| \le k \max_{0 \le \mu \le j} \eta_\mu \underset{\eta_{j+1} \ge \eta_j}{=} k\eta_j,$$

zusammen mit der Definition (4.5.12) von η_{j+1} ergibt

$$\eta_{j+1} \le (1 + \gamma k)\,\eta_j + \beta.$$

Hierauf lässt sich Lemma 4.2.2 anwenden. Die Entsprechung der Größen in (4.2.1) lautet: $\nu \equiv j$, $a_\nu \equiv \eta_j$, $h \equiv 1$, $L \equiv k\gamma$, $B \equiv \beta$. Das Resultat des Lemmas 4.2.2 ist

$$\eta_j \le \eta_0 e^{jk\gamma} + \left\{ \begin{array}{ll} j\beta & \text{falls } \gamma = 0 \\ \frac{\beta}{k\gamma}\left(e^{jk\gamma} - 1\right) & \text{falls } \gamma > 0 \end{array} \right\} \qquad (j \in \mathbb{N}_0).$$

Weiter gelten $\eta_0 = \|X_0\| \le k'\,\|X_0\|_\infty \le k'\alpha$ und $|x_j| \le \|X_j\|_\infty \le k\,\|X_j\| \le k\eta_j$. Zusammen folgt die Behauptung des Satzes. ∎

4.5.6 Sätze

Wir werden zeigen, dass Konvergenz und Stabilität von Mehrschrittverfahren fast äquivalent sind. Für exakte Aussagen braucht man noch Annahmen über den Zusammenhang von $\phi(x_j, \eta_{j+r-1}, \ldots, \eta_j, h; f)$ und f. Eine sehr schwache Voraussetzung lautet:

$$f = 0 \quad \Longrightarrow \quad \phi(x_j, \eta_{j+r-1}, \ldots, \eta_j, h; f) = 0. \tag{4.5.13}$$

Diese Voraussetzung ist insbesondere für die wichtige Klasse der *linearen r-Schrittverfahren* erfüllt:

$$\phi(x_j, \eta_{j+r-1}, \ldots, \eta_j, h; f) = \sum_{\mu=0}^{r-1} b_\mu f_{j+\mu} \qquad \text{mit } f_k = f(x_k, \eta_k). \tag{4.5.14}$$

Satz 4.5.21 (Stabilitätssatz) *Voraussetzung (4.5.13) sei erfüllt. Dann impliziert die Konvergenz aus Definition 4.5.2 die Stabilität.*

Beweis. a) Wir wählen $f = 0$ und den Startwert $y_0 = 0$. Damit ist $y = 0$ die exakte Lösung des Anfangsrandwertproblems, während die diskrete Lösung den Gleichungen $\sum_{\nu=0}^{r} \alpha_\nu \eta_{j+\nu} = 0$ genügt, wobei die Anfangswerte $\eta_0 = \varepsilon_0, \ldots, \eta_{r-1} = \varepsilon_{r-1}$ lauten (vgl. Definition 4.5.2). Da $\sum_{\nu=0}^{r} \alpha_\nu \eta_{j+\nu} = 0$ die Differenzengleichung (4.5.6) ist, gilt $(\eta_j)_{j\in\mathbb{N}_0} \in \mathcal{F}_0$. Allerdings ist zu beachten, dass das Mehrschrittverfahren nur den endlichen Abschnitt $(\eta_j)_{0\le j\le J(h)}$ mit $J(h) = \lfloor (x_E - x_0)/h \rfloor$ berücksichtigt, da nur für solche j $x_j \in I$.

b) Zum indirekten Beweis sei Instabilität angenommen. Aufgrund dieser Annahme existiert eine unbeschränkte Lösung $\mathbf{x} \in \mathcal{F}_0$. Die Divergenz

$$C(h) := \max\{|x_j| : 0 \le j \le J(h)\} \to \infty \quad \text{für } h \to 0$$

folgt aus $J(h) \to \infty$ für $h \to 0$ und $\|\mathbf{x}\|_\infty = \infty$. Man wähle die Anfangsstörungen $\vec{\varepsilon} = (\varepsilon_j)_{j=0,\ldots,r-1} := (x_j/C(h))_{j=0,\ldots,r-1}$. Offenbar gilt $\|\vec{\varepsilon}\|_\infty \to 0$ für $h \to 0$. Das Mehrschrittverfahren produziert für diese Anfangsstörungen die Lösung $(\eta_j)_{0\le j\le J(h)}$ mit $\eta_j = \frac{1}{C(h)} x_j$. Da

$$\sup_{x\in I} |\eta(x; \vec{\varepsilon}, h) - y(x)| = \frac{1}{C(h)} \max\{|x_j| : 0 \le j \le J(h)\} = 1,$$

geht dieser Fehler für den Grenzübergang $h \to 0$, $\|\vec{\varepsilon}\|_\infty \to 0$ nicht gegen 0 im Widerspruch zur vorausgesetzten Konvergenz. ∎

Für die umgekehrte Richtung brauchen wir eine Lipschitz-Bedingung, die (4.4.3) im Falle eines Einschrittverfahrens entspricht:

$$\begin{aligned} &\text{Für jedes } f \in C(I \times \mathbb{R}) \text{ mit (4.1.2) existiere } L_\phi, \text{ sodass} \\ &|\phi(x_j, u_{r-1}, \ldots, u_0, h; f) - \phi(x_j, v_{r-1}, \ldots, v_0, h; f)| \le L_\phi \max_{i=0,\ldots,r-1} |u_i - v_i| . \end{aligned} \tag{4.5.15}$$

Bemerkung 4.5.22 *Bedingung (4.5.15) ist für lineare r-Schrittverfahren (4.5.14) erfüllt.*

Satz 4.5.23 (Konvergenzsatz) *Es gelte Bedingung (4.5.15). Außerdem sei das Mehrschrittverfahren konsistent und stabil. Dann ist es konvergent [sogar im stärkeren Sinne wie im Anschluss an Definition 4.5.2 diskutiert].*

Beweis. Die Anfangsfehler seien mittels $\eta_j = y(x_j) + \varepsilon_j$ für $j = 0, \ldots, r-1$ definiert. Die Mehrschrittformel mit zusätzlichen Fehlern $h\varepsilon_{j+r}$ lautet (4.5.3). Die Fehlernorm ist $\|\vec{\varepsilon}\|_\infty := \max\{|\varepsilon_j| : 0 \le j \le J(h)\}$ mit $J(h)$ wie im vorherigen Beweis (die übliche Konvergenz führt auf $\varepsilon_j = 0$ für $r \le j \le J(h)$, nur im Falle des Klammerzusatzes [...] kann $\varepsilon_j \ne 0$ für $r \le j \le J(h)$ auftreten). Zu zeigen ist $\eta(x; \vec{\varepsilon}, h) \to y(x)$ für $h \to 0$, $\|\vec{\varepsilon}\|_\infty \to 0$.

Der Fehler wird mit $e_j := \eta_j - y(x_j)$ notiert. Für $0 \le j \le r-1$ ist $e_j = \varepsilon_j$. Für $j \ge r$ ergibt die Differenz zwischen (4.5.3) und der mit (4.5.2) äquivalenten Gleichung

$$\sum_{\nu=0}^{r} \alpha_\nu y(x_{j+\nu}) - h\phi(x_j, y(x_{j+r-1}), \ldots, y(x_j), h; f) = h\tau(x_j, y(x_j); h) =: h\tau_{j+r}$$

mit dem lokalen Diskretisierungsfehler τ_{j+r}, dass

$$\sum_{\nu=0}^{r} \alpha_\nu e_{j+\nu} = \beta_{j+r} := h\left[\phi(x_j, \eta_{j+r-1}, \ldots, \eta_j, h; f) - \phi(x_j, y(x_{j+r-1}), \ldots, y(x_j), h; f)\right] + h\left(\varepsilon_{j+r} - \tau_{j+r}\right).$$

Aus (4.5.15) folgert man

$$|\beta_{j+r}| \le hL_\phi \max_{j\le\mu\le,j+r-1} |e_\mu| + h\left(\|\vec{\varepsilon}\|_\infty + \|\overrightarrow{\tau}\|_\infty\right), \quad \text{wobei } \overrightarrow{\tau} = (\tau_j)_{r\le j\le J(h)}\,,\ \|\overrightarrow{\tau}\|_\infty = \max_{r\le j\le J(h)} |\tau_j|.$$

Die Konsistenzbedingung $\sup_{x\in I} \tau(x, y(x); h) \to 0$ impliziert $\|\overrightarrow{\tau}\|_\infty \to 0$ für $h \to 0$.

Die Voraussetzungen von Satz 4.5.20 gelten mit

$$x_j = e_j, \quad \alpha = \|\vec{\varepsilon}\|_\infty, \quad \beta = h\left(\|\vec{\varepsilon}\|_\infty + \|\overrightarrow{\tau}\|_\infty\right), \quad \gamma = hL_\phi,$$

und der Satz liefert

$$|e_j| \le kk' \|\vec{\varepsilon}\|_\infty e^{jhL_\phi k} + \frac{\|\vec{\varepsilon}\|_\infty + \|\overrightarrow{\tau}\|_\infty}{L_\phi}\left(e^{jhL_\phi k} - 1\right) \tag{4.5.16}$$

für den Fall $hL_\phi > 0$ (der Fall $hL_\phi = 0$ ist analog). Das Produkt jh im Exponenten ist als $x_j - x_0$ zu interpretieren und damit durch $x_E - x_0$ beschränkt (bzw. es ist konstant gleich x für den Grenzwertprozess $j = n \to \infty$, $h := (x - x_0)/n$). Für $h \to 0$ gilt $\|\overrightarrow{\tau}\|_\infty \to 0$ (Konsistenz) und $\|\vec{\varepsilon}\|_\infty \to 0$ (Teil der Konvergenzdefinition). Nach (4.5.16) konvergiert auch e_j gleichmäßig gegen null, d.h. $\sup\limits_{x\in I} |\eta(x; \vec{\varepsilon}, h) - y(x)| \to 0$. ■

Korollar 4.5.24 *Zusätzlich zu den Voraussetzungen von Satz 4.5.23 sei Konsistenz von der Ordnung p angenommen. Dann liegt auch Konvergenz der gleichen Ordnung p vor, falls die Anfangsfehler hinreichend klein sind:*

$$|\eta(x; \vec{\varepsilon}, h) - y(x)| \le C\left(h^p + \max_{j=0}^{r-1} |\varepsilon_j|\right).$$

Beweis. Es ist $\|\vec{\varepsilon}\|_\infty = \max_{j=0}^{r-1} |\varepsilon_j|$ und $\|\overrightarrow{\tau}\|_\infty \le \mathcal{O}(h^p)$. (4.5.16) beweist die gewünschte Fehlerschranke. ■

4.6 Konstruktion optimaler Mehrschrittverfahren

4.6.1 Beispiele

Die Adams-Bashforth-Verfahren[45] sind explizite lineare r-Schrittverfahren der Form

$$\eta_{j+r} = \eta_{j+r-1} + h\sum_{\mu=0}^{r-1} b_\mu f_{j+\mu}. \tag{4.6.1}$$

Das zugehörige charakteristische Polynom ist $\psi(\zeta) = \zeta^r - \zeta^{r-1} = \zeta^{r-1}(\zeta - 1)$. Die Nullstellen lauten $\zeta_1 = \ldots = \zeta_{r-1} = 0$, $\zeta_r = 1$, sodass das Adams-Bashforth-Verfahren stabil ist. Die Koeffizienten b_μ $(\mu = 0, \ldots, r-1)$ stehen zur Verfügung, um den lokalen Konsistenzfehler möglichst klein zu machen. Bei optimaler Wahl entsteht ein Mehrschrittverfahren der Ordnung[46] r.

Übungsaufgabe 4.6.1 *Man zeige: a) Für $r = 1$ ist das Euler-Verfahren das optimale Adams-Bashforth-Verfahren. b) Wie lautet die optimalen Koeffizienten b_0, b_1 aus (4.6.1) für $r = 2$?*

Das allgemeine explizite lineare Zweischrittverfahren lautet

$$\eta_{j+2} = -\alpha_1\eta_{j+1} - \alpha_0\eta_j + h\left[b_1 f_{j+1} + b_0 f_j\right].$$

[45] John Couch Adams, geb. 5. Juni 1819 in Laneast, gest. 21. Jan. 1892 in Cambridge

[46] Man beachte den folgenden Vorteil des Mehrschrittverfahrens über Einschrittverfahren: Trotz der erhöhten Konsistenzordnung r braucht in jedem Stützpunkt x_j nur ein Funktionswert $f_j = f(x_j, \eta_j)$ ausgewertet zu werden (der in $r-1$ weiteren Berechnungsschritten wiederverwendet werden kann).

Da (4.1.7) (also $\alpha_0 + \alpha_1 = 1$) als eine Nebenbedingung von der Zahl der Freiheitsgrad abzuziehen ist, lässt sich bei optimaler Wahl der Koeffizienten noch die Konsistenzordnung $p = 3$ erreichen. Das entstehende Verfahren lautet

$$\eta_{j+2} = -4_1\eta_{j+1} + 5\eta_j + h\left[4f_{j+1} + 2f_j\right]. \tag{4.6.2}$$

Das dazugehörige Polynom ist

$$\psi(\zeta) = \zeta^2 + 4\zeta - 5 = (\zeta - 1)(\zeta + 5).$$

Die Nullstelle -5 führt zur Instabilität. Dass die Instabilität auch in der Praxis deutlich zu sehen ist, zeigt die Anwendung von (4.6.2) auf das Anfangswertproblem $y' = -y$, $y(0) = 1 =: \eta_0$ ($\Rightarrow y(x) = e^{-x}$). Wir wählen $\eta_1 := e^{-h}$ in Übereinstimmung mit der exakten Lösung. Als Schrittweite in $I = [0, 1]$ wird $h = 0.01$ gewählt:

j	x_j	$\eta_j - y(x_j)$
2	0.02	-0.16_{10}-8
3	0.03	$+0.50_{10}$-8
4	0.04	-0.30_{10}-7
5	0.05	$+0.14_{10}$-6
$\vdots$	$\vdots$	$\vdots$
99	0.99	$+0.13_{10}$+60
100	1.00	-0.65_{10}+60

Die Wurzel $\zeta = -5$ ist für die wechselnden Vorzeichen des Fehlers und für das explosionsartige Anwachsen verantwortlich ($5^{98} = 3.155_{10}68$).

4.6.2 Optimale Ordnung stabiler Mehrschrittverfahren

Das vorherige Beispiel zeigt, dass man nicht alle Koeffizienten α_ν, b_μ aus (4.1.6) und (4.5.14) verwenden darf, um die Konsistenzordnung zu maximieren. Daneben hat man als Nebenbedingung die Stabilität zu erfüllen. Die Charakterisierung der optimalen stabilen Mehrschrittverfahren stammt von Dahlquist[47] und sei hier ohne Beweis wiedergegeben:

Satz 4.6.2 (Dahlquist) *a) $r \geq 1$ sei ungerade. Die höchste Konsistenzordnung eines stabilen r-Schrittverfahrens ist $p = r + 1$.*

b) Sei $r \geq 2$ gerade. Die höchste Konsistenzordnung eines stabilen r-Schrittverfahrens ist $p = r + 2$. In diesem Falle haben alle Wurzeln des charakteristischen Polynoms ψ den Betrag 1.

4.7 Andere Stabilitätsbegriffe

Im Bereich der gewöhnlichen Differentialgleichungen gibt es eine Vielzahl weiterer Stabilitätsbegriffe, unter anderem solche, die sich auf das Phänomen der steifen Differentialgleichungen beziehen oder die sich mit dem Verhalten der Diskretisierungsfehler für große x beschäftigen.[48] Schließlich gibt es Stabilitätsbegriffe (z.B. Ljapunow-Stabilität), die sich nicht auf die Diskretisierung, sondern auf die Differentialgleichung selbst beziehen.[49]

[47] Germund Dahlquist, geb. 16. Jan. 1925, Emeritus an der Universität Stockholm

[48] Vgl. die Monographien *Stetter:* Analysis of discretization methods for ordinary differential equations. Springer-Verlag, Berlin, 1973, und *Hairer - Wanner:* Solving ordinary differential equations II. Springer-Verlag, Berlin, 1991

[49] Vgl. H. Heuser: *Gewöhnliche Differentialgleichungen.* Teubner, Stuttgart, 1989

5 Partielle Differenzengleichungen

5.1 Notation, Aufgabenstellung, Funktionenräume

Geht man von den skalaren gewöhnlichen Differentialgleichung zu den linearen Systemen gewöhnlicher Differentialgleichungen über, erhält man $y' = Ay + f$, wobei A im einfachsten Fall eine konstante $N \times N$-Matrix ist und y und f Werte in $\mathbb{R}^N$ (oder $\mathbb{C}^N$) haben. Die vorherigen Resultate zu Ein- und Mehrschrittverfahren lassen sich leicht auf diesen Fall verallgemeinern. Anders wird es, wenn man die (beschränkte) Matrix A durch einen unbeschränkten Differentialoperator ersetzt, was im folgenden geschieht.

Die unabhängige Variable, auf die sich das Differentiationssymbol $'$ bezieht, sei mit t (statt x) bezeichnet (die Vorstellung ist, dass t die Zeit beschreibt). Der Differentialoperator A wird dagegen auf Ortsvariablen $x_1, \ldots, x_d$ angewandt. Obwohl $d = 3$ der realistische Fall ist, reicht es für die Zwecke dieser Vorlesung, $d = 1$ zu analysieren.

Notation 5.1.1 *Die gesuchte Lösung wird mit u (statt früher y) bezeichnet. Die unabhängigen Variablen sind t und x. Die klassische Schreibweise für u ist deshalb $u(t, x)$. Sei B ein Raum von Funktionen in der Variablen x. Dann bezeichnet $u(t)$ die (partiell in t ausgewertete) Funktion $u(t, \cdot) \in B$. Damit sind $u(t, x)$ und $u(t)(x)$ gleichbedeutende Notationen. Sei $I = [0, T]$ das Zeitintervall, in dem t variieren soll. Ist $D_A \subset B$ der Definitionsbereich[50] des Differentialoperators A, so lautet die partielle Differentialgleichung: Gesucht ist eine stetige Funktion $u : I \to D_A \subset B$, sodass*

$$\frac{\partial}{\partial t} u(t) = Au(t) \qquad \text{für alle } t \in I. \tag{5.1.1a}$$

Die Anfangswertbedingung lautet

$$u(0) = u_0 \qquad \text{für ein } u_0 \in D_A \subset B. \tag{5.1.1b}$$

Für den Differentialoperator A werden zwei unterschiedliche Fälle diskutiert:

$$A := a\frac{\partial}{dx} \quad (a \neq 0) \qquad \text{oder} \qquad A := a\frac{\partial^2}{dx^2} \quad (a > 0). \tag{5.1.2}$$

Im folgenden wird der Definitionsbereich von $u(\cdot, \cdot)$ der Streifen

$$\Sigma = I \times \mathbb{R} \qquad \text{mit } I = [0, T] \tag{5.1.3}$$

sein. Dabei variiert die Zeit t in $I = [0, T]$, während die Ortsvariable x in $\mathbb{R}$ variiert. I entspricht dem Intervall $I = [x_0, x_E]$ aus §4.1.1. Der Ortsbereich $\mathbb{R}$ wird als unbeschränkt angenommen, da so Randbedingungen vermieden werden[51].

Als Kandidaten für den Funktionenraum B bieten sich zwei Banach-Räume an:

- $B = C(\mathbb{R})$, Raum der komplexwertigen, stetigen Funktionen mit der Supremumsnorm $\|v\|_B = \|v\|_\infty = \sup\{|v(x)| : x \in \mathbb{R}\}$.
- $B = L^2(\mathbb{R})$, Raum der komplexwertigen, messbaren und quadratintegrablen Funktionen. Letzteres bedeutet, dass die Norm $\|v\|_B = \|v\|_2 = \sqrt{\int_{\mathbb{R}} |v(x)|^2 \, dx}$ endlich ist. Dieser Banach-Raum ist zugleich Hilbert-Raum mit dem Skalarprodukt $\langle u, v \rangle_2 := \int_{\mathbb{R}} u(x)\overline{v(x)}dx$.

Wir werden für die beiden Fälle aus (5.1.2) zeigen, dass die Aufgabe lösbar ist

[50] Der *Definitionsbereich* des Differentialoperators A ist $D_A = \{v \in B : Av \text{ ist definiert und gehört zu } B\}$. Oft reicht es, eine kleinere, dichte Menge $B_0 \subset D_A$ zu wählen und die Resultate durch stetige Fortsetzung auf D_A auszuweiten.

[51] Eine ähnliche Situation ergibt sich, wenn die Lösungen 2π-periodisch in x angenommen werden. Dies entspricht der Lösung im beschränkten Gebiet $\Sigma = I \times [0, 2\pi]$ mit der *periodischen Randbedingung* $u(t, 0) = u(t, 2\pi)$. Im 2π-periodischen Fall lauten die Räume

$$C_{\text{per}}(\mathbb{R}) := \{v \in C(\mathbb{R}) : v(x) = v(x + 2\pi) \text{ für alle } x \in \mathbb{R}\}, \quad L^2_{\text{per}}(\mathbb{R}) := \{v \in L^2(\mathbb{R}) : v(x) = v(x + 2\pi) \text{ für fast alle } x \in \mathbb{R}\}.$$

5.2 Der hyperbolische Fall $A = a\frac{\partial}{dx}$

Wir wählen zuerst $B = C(\mathbb{R})$. Der Definitionsbereich von A ist der in B dichte Unterraum $B_0 = C^1(\mathbb{R})$. Die partielle Differentialgleichung $\frac{\partial}{\partial t}u = a\frac{\partial}{dx}u$ (oder einfacher als $u_t = au_x$ bezeichnet) gehört zum Typ[52] der hyperbolischen Differentialgleichungen. Die Lösung der Aufgabe (5.1.1a,b) ist direkt angebbar:

Lemma 5.2.1 *Für jedes $u_0 \in B_0 = C^1(\mathbb{R})$ ist $u(t,x) := u_0(x+at)$ eine eindeutige Lösung des Anfangswertproblems (5.1.1a,b).*

Beweis. a) Da $\frac{\partial}{\partial t}u = \frac{\partial}{\partial t}u_0(x+at) = au_0'(x+at)$ und $a\frac{\partial}{dx}u_0(x+at) = au_0'(x+at)$, ist die Differentialgleichung (5.1.1a) erfüllt. Zudem ist der Anfangswert $u(0,x) := u_0(x)$.

b) Zum Eindeutigkeitsbeweis führt man die Transformation auf die (charakteristische) Richtung ein: $\xi = x+at$, $\tau = t$ und $U(\tau,\xi) := u(t(\tau,\xi), x(\tau,\xi))$ mit der Umkehrtransformation $t(\tau,\xi) = \tau$, $x(\tau,\xi) = \xi - a\tau$. Die Kettenregel liefert $\frac{\partial}{\partial\tau}U = \frac{\partial}{\partial t}u\frac{\partial}{\partial\tau}t + \frac{\partial}{\partial x}u\frac{\partial}{\partial\tau}x = \frac{\partial}{\partial t}u - a\frac{\partial}{\partial x}u$. Da sind die Lösungen von (5.1.1a) in den neuen Variablen durch die gewöhnliche Differentialgleichung $\frac{\partial}{\partial\tau}U(\tau,\xi) = 0$ (für jedes $\xi \in \mathbb{R}$) beschrieben. Dies hat offenbar genau die konstante Lösung $U(\tau,\xi) = U(0,\xi) = u(t(0,\xi), x(0,\xi)) = u(0,\xi) = u_0(\xi)$. ∎

Im Falle des Raumes $B = L^2(\mathbb{R})$ wählen wir die in B dichte Teilmenge $C_0^\infty(\mathbb{R})$ als Definitionsbereich. Dabei ist $C_0^\infty(\mathbb{R})$ die Menge aller unendlich oft differenzierbaren Funktionen mit kompaktem Träger (der Träger einer Funktion ist $Tr(\varphi) := \overline{\{x \in \mathbb{R} : \varphi(x) \neq 0\}}$).

Bemerkung 5.2.2 *Sei $t \geq 0$ beliebig. Gehört der Anfangswert u_0 zu $C^1(\mathbb{R})$ oder $C_0^\infty(\mathbb{R})$, so gehört auch die Lösung $u(t)$ zu $C^1(\mathbb{R})$ bzw. $C_0^\infty(\mathbb{R})$. Außerdem gilt $\|u(t)\|_B = \|u_0\|_B$ sowohl für $\|\cdot\|_B = \|\cdot\|_\infty$ als auch $\|\cdot\|_B = \|\cdot\|_2$.*

Beweis. Da $u(t)$ eine verschobene Version von u_0 ist und eine Verschiebung die Zugehörigkeit zu B und die Norm $\|\cdot\|_B$ nicht ändert, folgen die Behauptungen. ∎

5.3 Der parabolische Fall $A = \frac{\partial^2}{dx^2}$

Die parabolische Differentialgleichung $\frac{\partial}{\partial t}u = a\frac{\partial^2 u}{dx^2}$ heißt auch *Wärmeleitungsgleichung*, da sie die Entwicklung der Temperatur u zur Zeit t am Ort x im Falle eines unendlichen Drahtes (eindimensionaler Fall!) beschreibt. Der Faktor $a > 0$ ist der Wärmeleitungskoeffizient. Durch eine Transformation $t \mapsto at$ oder $x \mapsto \sqrt{a}x$ kann man stets $a = 1$ erreichen, daher beschränken[53] wir uns auf

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{dx^2} \qquad \text{für } t > 0. \tag{5.3.1}$$

Lemma 5.3.1 *Die Lösung von (5.3.1) zu einem stetigen Anfangswert (5.1.1b) lautet*

$$u(t,x) = \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} u_0(\xi) \exp\left(\frac{-(x-\xi)^2}{4t}\right) d\xi \qquad \textit{für } t > 0 \textit{ und } x \in \mathbb{R}. \tag{5.3.2}$$

Beweis. a) Man prüft nach, dass $\frac{1}{\sqrt{4\pi t}} \exp\left(\frac{-(x-\xi)^2}{4t}\right)$ für jedes $\xi \in \mathbb{R}$ und $t > 0$ eine Lösung von (5.3.1) ist. Da der Integrand in (5.3.2) exponentiell fällt, kann man auch hier Integration und Differentiation vertauschen und erhält so, dass u aus (5.3.2) die Wärmeleitungsgleichung (5.3.1) erfüllt.

b) Es bleibt $\lim_{t\searrow 0} u(t,x) = u_0(x)$ zu zeigen. Ein Blick in einschlägige Formelsammlungen[54] zeigt, dass

$$\frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} \exp\left(\frac{-\zeta^2}{4t}\right) d\zeta = 1 \qquad \text{für } t > 0. \tag{5.3.3}$$

[52] Näheres zur Typeneinteilung der partiellen Differentialgleichungen enthält Kapitel 1 in W. Hackbusch: *Theorie und Numerik elliptischer Differentialgleichungen*. Teubner, Stuttgart 1996

[53] Die allgemeine Form einer parabolischen Differentialgleichung ist $\frac{\partial u}{\partial t} = Au$ mit einem *elliptischen Differentialoperator* A, dessen Spektrum $\sigma(A)$ einen nach oben beschränkten Realteil besitzen muss: $\sup\{\Re e\, z : z \in \sigma(A)\} < \infty$. Zur Definition der Elliptizität siehe Definition 1.2.1 im Buch aus Fußnote 52. Die Bedingung an $\sigma(A)$ legt das Vorzeichen von A fest.

[54] Zum Beispiel Seite 185 im *Teubner-Taschenbuch der Mathematik*, Teubner, Stuttgart, 1996

Substitution $\zeta = \xi - x$ liefert die Umformung

$$\begin{aligned} u(t,x) &= \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} u_0(\xi) \exp\left(\frac{-(x-\xi)^2}{4t}\right) d\xi \\ &= u_0(x) + \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} [u_0(\xi) - u_0(x)] \exp\left(\frac{-(x-\xi)^2}{4t}\right) d\xi. \end{aligned}$$

Um $\lim_{t \searrow 0} u(t,x) = u_0(x)$ zu zeigen, muß bewiesen werden, dass der letzte Summand für $t \searrow 0$ gegen null strebt.

Seien x und $\varepsilon > 0$ fest gewählt. Wegen der Stetigkeit von u_0 gibt es ein $\delta > 0$, so dass $|u_0(\xi) - u_0(x)| \leq \varepsilon/2$ für alle $|\xi - x| \leq \delta$. Wir zerlegen das Integral in die Summe von

$$\begin{aligned} I_1(t,x) &:= \frac{1}{\sqrt{4\pi t}} \int_{x-\delta}^{x+\delta} [u_0(\xi) - u_0(x)] \exp\left(\frac{-(x-\xi)^2}{4t}\right) d\xi, \\ I_2(t,x) &:= \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{x-\delta} [u_0(\xi) - u_0(x)] \exp\left(\frac{-(x-\xi)^2}{4t}\right) d\xi, \\ I_3(t,x) &:= \frac{1}{\sqrt{4\pi t}} \int_{x+\delta}^{\infty} [u_0(\xi) - u_0(x)] \exp\left(\frac{-(x-\xi)^2}{4t}\right) d\xi. \end{aligned}$$

Das erste Integral ist beschränkt durch

$$\begin{aligned} |I_1(t,x)| &\leq \frac{1}{\sqrt{4\pi t}} \int_{x-\delta}^{x+\delta} |u_0(\xi) - u_0(x)| \exp\left(\frac{-(x-\xi)^2}{4t}\right) d\xi \\ &\leq \frac{\varepsilon/2}{\sqrt{4\pi t}} \int_{x-\delta}^{x+\delta} \exp\left(\frac{-(x-\xi)^2}{4t}\right) d\xi \leq \frac{\varepsilon/2}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} \exp\left(\frac{-(x-\xi)^2}{4t}\right) d\xi \underset{(5.3.3)}{=} \frac{\varepsilon}{2}. \end{aligned}$$

Sei $C := \sup_{x \in \mathbb{R}} |u_0(x)| < \infty$. Dann ist I_2 beschränkt durch

$$\begin{aligned} |I_2(t,x)| &\leq \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{x-\delta} |u_0(\xi) - u_0(x)| \exp\left(\frac{-(x-\xi)^2}{4t}\right) d\xi \\ &\leq \frac{2C}{\sqrt{4\pi t}} \int_{-\infty}^{x-\delta} \exp\left(\frac{-(x-\xi)^2}{4t}\right) d\xi \underset{\tau=(x-\xi)/\sqrt{4t}}{\overset{\text{Substitution}}{=}} \frac{2C}{\sqrt{\pi}} \int_{\delta/\sqrt{4t}}^{\infty} \exp\left(-\tau^2\right) d\tau. \end{aligned}$$

Da das uneigentliche Integral $\int_{-\infty}^{\infty} \exp\left(-\tau^2\right) d\tau$ existiert, gilt $\int_{R}^{\infty} \exp\left(-\tau^2\right) d\tau \to 0$ für $R \to \infty$. Damit gilt für ein hinreichend kleines $t > 0$, dass $|I_2(t,x)| \leq \frac{\varepsilon}{4}$. Für I_3 erhalten wir die gleiche Schranke $|I_3(t,x)| \leq \frac{\varepsilon}{4}$. Zusammen gilt also $\left| \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} [u_0(\xi) - u_0(x)] \exp \frac{-(x-\xi)^2}{4t} d\xi \right| \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \varepsilon$ für hinreichend kleines $t > 0$. Da x und ε beliebig gewählt waren, gilt für alle x, dass

$$\lim_{t \searrow 0} \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} [u_0(\xi) - u_0(x)] \exp\left(\frac{-(x-\xi)^2}{4t}\right) d\xi = 0.$$

■

Wie im Lemma formuliert, führt bereits ein nur stetiger Anfangswert u_0 zu einer unendlich oft differenzierbaren Lösung $u(t)$ bei $t > 0$. Allerdings existiert die Lösung nur für $t > 0$, nicht für $t < 0$. Im Gegensatz dazu gilt im hyperbolischen Fall die Lösungsdarstellung aus Lemma 5.2.1 genauso gut für $t < 0$.

Im hyperbolischen Fall waren die Normen $\|u(t)\|_B$ unabhängig von t. Im parabolischen Fall gelten dagegen nur Monotonieaussagen.

Lemma 5.3.2 *$u(t) \in B = C(\mathbb{R})$ sei Lösung von (5.3.1). Dann gilt $\|u(t)\|_\infty \leq \|u(0)\|_\infty$ für alle $t \geq 0$.*

Beweis. Sei $C := \|u(0)\|_\infty$. Da für $t = 0$ nichts zu beweisen ist, sei $t > 0$. Aufgrund von (5.3.2) erhalten wir

$$|u(t,x)| \le \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} |u_0(\xi)| \exp\left(\frac{-(x-\xi)^2}{4t}\right) d\xi \le \frac{C}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} \exp\left(\frac{-(x-\xi)^2}{4t}\right) d\xi \underset{(5.3.3)}{=} C,$$

also $\|u(t)\|_\infty \le C$. ∎

Lemma 5.3.3 *$u(t) \in B = L^2(\mathbb{R})$ sei Lösung von (5.3.1). Dann gilt $\|u(t)\|_2 \le \|u(0)\|_2$ für alle $t \ge 0$.*

Beweis. Man schließt entweder aus (5.3.2) oder aus allgemeinen Überlegungen (vgl. Bemerkung 5.4.2a), dass $\frac{\partial^2 u}{dx^2} \in L^2(\mathbb{R})$, sodass die folgenden Integrale existieren. Sei $t'' \ge t' \ge 0$. Wegen

$$\int_{\mathbb{R}} u(t'',x)^2 dx - \int_{\mathbb{R}} u(t',x)^2 dx = \int_{\mathbb{R}} \int_{t'}^{t''} \frac{\partial}{\partial t} \left(u(t',x)^2\right) dt dx = 2 \int_{\mathbb{R}} \int_{t'}^{t''} u(t,x) \frac{\partial u(t,x)}{\partial t} dt dx$$
$$= 2 \int_{\mathbb{R}} \int_{t'}^{t''} u(t,x) \frac{\partial^2 u(t,x)}{\partial x^2} dt dx = 2 \int_{t'}^{t''} \int_{\mathbb{R}} u(t,x) \frac{\partial^2 u(t,x)}{\partial x^2} dx dt = -2 \int_{t'}^{t''} \int_{\mathbb{R}} \left(\frac{\partial u(t,x)}{\partial x}\right)^2 dx dt \le 0$$

ist $\|u(t)\|_2^2$ schwach monoton fallend. ∎

5.4 Halbgruppe der Lösungsoperatoren

Sei $B_0 \subset D_A \subset B$ ein dichter Teilraum von B (z.B. $C^1(\mathbb{R})$, $C_0^\infty(\mathbb{R})$ oder $C^\infty(\mathbb{R})$). Für jeden Anfangswert $u_0 \in B_0$ existiere eine Lösung $u(t) \in B_0$ der Anfangswertaufgabe (5.1.1a,b). Die Abbildung $u_0 \mapsto u(t)$ für ein festes $t \ge 0$ ist offenbar linear. Damit ist der Operator $T(t)$ mittels

$$T(t)u_0 = u(t)$$

auf B_0 definiert und wird *Lösungsoperator* genannt.

Im vorherigen Abschnitt haben wir die Ungleichung $\|u(t)\|_B \le \|u(0)\|_\infty = \|u_0\|_\infty$ bewiesen. Selbst wenn $\|u(t)\|_B \le K(t) \|u_0\|_\infty$ mit einer nur von t abhängigen Konstante gelten würde, ist $T(t)$ auf B_0 beschränkt, d.h. $T(t) \in L(B_0, B)$.

Bemerkung 5.4.1 *Sei $T(t) \in L(B_0, B)$ und B_0 dicht in B. Dann kann $T(t)$ eindeutig auf B stetig fortgesetzt werden, wobei das fortgesetzte $T(t) \in L(B,B)$ und das ursprüngliche $T(t) \in L(B_0, B)$ die gleiche Operatornorm haben (d.h. $\sup\{\|T(t)v\|_B / \|v\|_B : 0 \ne v \in B_0\} = \sup\{\|T(t)v\|_B / \|v\|_B : 0 \ne v \in B\}$).*

Beweis. Sei $u_0 \in B$. Es gibt eine Folge $v_{0,n} \in B_0$ mit $\lim_{n\to\infty} \|u_0 - v_{0,n}\|_B = 0$. Man prüft nach, dass die $v_n := T(t)v_{0,n}$ eine Cauchy-Folge bilden und daher ein eindeutiges $u := \lim_{n\to\infty} v_n$ definieren. Die Definition $T(t)u_0 =: u$ definiert in der gewünschten Weise die stetige Fortsetzung zu $T(t) \in L(B,B)$. ∎

Im Falle von $B = C^1(\mathbb{R})$ wähle man $B_0 := C^\infty(\mathbb{R})$ als dichte Teilmenge von D_A. Es lässt sich allgemein zeigen, dass die Vertauschbarkeit $AT(t) = T(t)A$ gilt. Sei $u_0 \in D_A$, so folgt $Au_0 \in B$ und $T(t)Au_0 \in B$ gemäß Bemerkung 5.4.1. Wegen $AT(t) = T(t)A$ gilt aber auch $T(t)Au_0 = AT(t)u_0 \in B$, d.h. $T(t)u_0 \in D_A$. Dies beweist den ersten Teil der folgenden Bemerkung. Der Beweis des zweiten Teils folgt später.

Bemerkung 5.4.2 *a) Sei $t \ge 0$. Der Lösungsoperator $T(t)$ bildet den Definitionsbereich D_A in sich ab.*
b) $u(t) = T(t)u_0$ ist für alle $u_0 \in D_A$ Lipschitz-stetig.

Satz 5.4.3 *D_A sei dicht in B. Die Menge $\{T(t) : t \ge 0\}$ bildet eine Halbgruppe mit neutralem Element, d.h. $T(0) = I$ ist die Identität, während $T(t)T(s) = T(t+s)$ für alle $t, s \ge 0$ gilt.*

Beweis. a) Definitionsgemäß gilt $T(0)u_0 = u(0)$. Wegen (5.1.1b) ist zudem $u(0) = u_0$. Dies beweist $T(0) = I$.

b) Sei $u_0 \in D_A$ und $u(\tau) = T(\tau)u_0$ für alle $\tau \ge 0$. Man setze $U(t) := u(\tau + s) \in D_A$ (vgl. Bemerkung 5.4.2). U ist wieder Lösung von (5.1.1a) und besitzt den Anfangswert $u_s := u(s) = T(s)u_0 \in D_A$. Also ist $U(t) = T(t)u_s$. Zusammen gilt

$$T(t+s)u_0 = u(\tau + s) = U(t) = T(t)u_s = T(t)T(s)u_0$$

für alle $u_0 \in D_A$ und alle $s, t \geq 0$. Da D_A dicht in B, folgt $T(t+s) = T(t)T(s)$. ∎

Man nennt $\{T(t) : t \geq 0\}$ die von A erzeugte Halbgruppe. Das erzeugende A kann für alle $v \in D_A$ mittels $Av = \lim_{t \searrow 0} [(T(t)v - v)/t]$ zurückgewonnen werden. Eine andere Schreibweise für $T(t)$ ist e^{tA}.

Im Weiteren benötigen wir die folgenden Eigenschaften von $T(t)$:

Voraussetzung 5.4.4 *$\{T(t) \in L(B,B) : t \geq 0\}$ sei eine Halbgruppe mit neutralem Element $T(0)$. Ferner sei $T(t)$ auf $I = [0,T]$ gleichmäßig beschränkt:*

$$\|T(t)\|_{B \leftarrow B} \leq K_T \qquad \text{für alle } t \in I = [0,T]. \tag{5.4.1}$$

Für die Modellbeispiele $A = a\frac{\partial}{dx}$ und $A = \frac{\partial^2}{dx^2}$ wurde bereits $\|T(t)\|_{B \leftarrow B} \leq 1$ gezeigt, d.h. $K_T = 1$.

Übungsaufgabe 5.4.5 *a) Für ein $\tau > 0$ sei $K_\tau := \sup_{0 \leq t \leq \tau} \|T(t)\|_{B \leftarrow B} < \infty$. Ferner sei $t \geq 0$. Man zeige die Abschätzung $\|T(t)\|_{B \leftarrow B} \leq K_\tau^{\lceil t/\tau \rceil}$, wobei $\lceil t/\tau \rceil := \inf\{n \in \mathbb{N} : t/\tau \leq n\}$ die Aufrundung bezeichnet.*

Beweis zu Bemerkung 5.4.2b. Für jedes $0 \leq \delta \leq 1$ mit $t, t+\delta \in I = [0,T]$ gilt $u(t+\delta) - u(t) = \int_t^{t+\delta} \frac{\partial u}{\partial t} dt = \int_t^{t+\delta} Au(t) dt$. Sei $U(t)$ die Lösung zum Anfangswert $U_0 := Au_0$. Aufgrund der Vertauschbarkeit gilt $Au(t) = AT(t)u_0 = T(t)Au_0 = U(t)$. Die gleichmäßige Beschränktheit von $U(\cdot)$ auf I zeigt die Lipschitz-Stetigkeit $\|u(t+\delta) - u(t)\|_B \leq K\delta \|Au_0\|_B$. ∎

Bemerkung 5.4.6 *Bisher wurde nur die homogene Gleichung $\frac{\partial}{\partial t}u(t) = Au(t)$ erwähnt. Im Falle der inhomogenen Gleichung $\frac{\partial}{\partial t}u(t) = Au(t) + f(t)$ erhält man die Lösung als $u(t) = T(t)u_0 + \int_0^t T(t-s)f(s)ds$.*

5.5 Diskretisierung der partiellen Differentialgleichung

5.5.1 Notationen

Die reelle Achse $\mathbb{R}$ der x-Variablen wird durch ein beidseitig unendliches Gitter der Schrittweite $\Delta x > 0$ ersetzt:

$$G_{\Delta x} = \{x = \nu \Delta x : \nu \in \mathbb{Z}\}. \tag{5.5.1}$$

Entsprechend wird das Intervall $I = [0,T]$ durch das endliche Gitter der Schrittweite $\Delta t > 0$ ersetzt:

$$I_{\Delta t} = \{t = \mu \Delta t \leq T : \mu \in \mathbb{N}_0\}.$$

Das Produkt beider Gitter liefert das Rechtecksgitter

$$\Sigma_{\Delta x}^{\Delta t} := I_{\Delta t} \times G_{\Delta x} = \{(t,x) \in \Sigma : \ x/\Delta x \in \mathbb{Z},\ t/\Delta t \in \mathbb{N}_0\}$$

(vgl. (5.1.3)). Die Schrittweiten $\Delta x, \Delta t$ werden im Allgemeinen nicht unabhängig voneinander gewählt, sondern mittels eines Parameters λ verbunden (die Potenz von Δx entspricht der Ordnung des Differentialoperators A):

$$\lambda = \begin{cases} \Delta t/\Delta x & \text{im hyperbolischen Fall } A = a\frac{\partial}{dx}, \\ \Delta t/\Delta x^2 & \text{im parabolischen Fall } A = \frac{\partial^2}{dx^2}. \end{cases} \tag{5.5.2}$$

Für eine auf $\Sigma_{\Delta x}^{\Delta t}$ definierte Gitterfunktion $U : \Sigma_{\Delta x}^{\Delta t} \to \mathbb{C}$ verwenden wir die Notation

$$U_\nu^\mu := U(\mu \Delta t, \nu \Delta x) \qquad \text{für } (\mu \Delta t, \nu \Delta x) \in \Sigma_{\Delta x}^{\Delta t}.$$

Mit $U^\mu := (U_\nu^\mu)_{\nu \in \mathbb{Z}}$ werden alle Gitterwerte bei $t = \mu \Delta t$ zusammengefasst. Sei $\ell = \mathbb{C}^{\mathbb{Z}}$ der lineare Raum der beidseitig unendlichen Folgen mit komponentenweiser Addition und Multiplikation mit Skalen. Den Banach-Räumen $B = C(\mathbb{R})$ (oder $L^\infty(\mathbb{R})$) und $B = L^2(\mathbb{R})$ entsprechen die Folgenräume ℓ^∞ und ℓ^2:

- $\ell^\infty = \mathbb{C}^{\mathbb{Z}}$ mit der Norm $\|U\|_{\ell^\infty} = \sup\{|U_\nu| : \nu \in \mathbb{Z}\}$ ist Banach-Raum,
- $\ell^2 = \mathbb{C}^{\mathbb{Z}}$ mit der Norm $\|U\|_{\ell^2} = \sqrt{\Delta x \sum_{\nu \in \mathbb{Z}} |U_\nu|^2}$ ist Hilbert-Raum.

Als allgemeines Symbol wird ℓ^p verwendet ($p \in \{2, \infty\}$).

5.5.2 Transferoperatoren r, p

Der kontinuierliche Banach-Raum B und der diskrete Raum ℓ^p der Gitterfunktionen werden mittels

$$r = r_{\Delta x} : B \to \ell^p \tag{5.5.3a}$$

verbunden. Der Buchstabe r steht für "Restriktion". Der Index Δx wird meist weggelassen, da sich die Schrittweite aus dem Zusammenhang eindeutig ergibt.

Im Falle von $B = C(\mathbb{R})$ liegt die Interpolation in den Gitterpunkten von $G_{\Delta x}$ nahe:

$$u \in C(\mathbb{R}) \Rightarrow ru \in \ell^\infty \text{ mit } (ru)_j = u(j\Delta x) \quad \text{für } j \in \mathbb{Z}. \tag{5.5.3b}$$

Im Falle von $B = L^2(\mathbb{R})$ ist keine Interpolation möglich, da Funktionen aus $L^2(\mathbb{R})$ keine wohldefinierten Punktauswertungen zulassen. Man kann aber Mittelwerte bilden:

$$u \in L^2(\mathbb{R}) \Rightarrow ru \in \ell^2 \text{ mit } (ru)_j = \frac{1}{\Delta x} \int_{(j-1/2)\Delta x}^{(j+1/2)\Delta x} u(x)dx \quad \text{für } j \in \mathbb{Z}. \tag{5.5.3c}$$

Wir setzen voraus, dass $r = r_{\Delta x} : B \to \ell^p$ beschränkt ist:

$$\|r_{\Delta x}\|_{\ell^p \leftarrow B} \le C_r \qquad \text{für alle } \Delta x > 0. \tag{5.5.4}$$

Lemma 5.5.1 *Die Restriktionen (5.5.3b,c) erfüllen die Bedingung (5.5.4) mit $C_r = 1$ in den Normen $\|\cdot\|_{\ell^\infty \leftarrow C(\mathbb{R})}$ bzw. $\|\cdot\|_{\ell^2 \leftarrow L^2(\mathbb{R})}$.*

Beweis. a) Fall von (5.5.3b): $\|ru\|_{\ell^\infty} = \sup\{|u(j\Delta x)| : j \in \mathbb{Z}\} \le \sup\{|u(x)| : x \in \mathbb{R}\} = \|u\|_{C(\mathbb{R})}$.

b) Fall von (5.5.3c): $\|ru\|_{\ell^2}^2 = \|ru\|_2^2 = \Delta x \sum_j |(ru)_j|^2 = \frac{1}{\Delta x} \sum_j \left| \int_{(j-1/2)\Delta x}^{(j+1/2)\Delta x} u(x)dx \right|^2$ und die Schwarzsche Ungleichung

$$\left| \int_{(j-1/2)\Delta x}^{(j+1/2)\Delta x} u(x)dx \right|^2 \le \left(\int_{(j-1/2)\Delta x}^{(j+1/2)\Delta x} 1dx \right) \cdot \left(\int_{(j-1/2)\Delta x}^{(j+1/2)\Delta x} |u(x)|^2 dx \right) = \Delta x \int_{(j-1/2)\Delta x}^{(j+1/2)\Delta x} |u(x)|^2 dx$$

ergeben $\|ru\|_{\ell^2}^2 \le \sum_j \int_{(j-1/2)\Delta x}^{(j+1/2)\Delta x} |u(x)|^2 dx = \int_{\mathbb{R}} |u(x)|^2 dx = \|u\|_{L^2(\mathbb{R})}$. ∎

Die "Prolongation" p wirkt in der umgekehrten Richtung $p = p_{\Delta x} : \ell^p \to B$. Wir setzen voraus, dass ein p mit folgender Eigenschaft existiert:

$$\|p_{\Delta x}\|_{B \leftarrow \ell^p} \le C_p \text{ für alle } \Delta x > 0 \qquad \text{und} \qquad rp = I. \tag{5.5.5}$$

Die Bedingung $rp = I$ besagt, dass p die Rechtsinverse von r ist.

Übungsaufgabe 5.5.2 *Man prüfe nach, dass in den Fällen (5.5.3b,c) die folgenden p die Bedingung (5.5.5) mit $C_p = 1$ erfüllen:*

p ist stückweise lineare Interpolation: (5.5.6a)
$v \in \ell^\infty \mapsto pv \in C(\mathbb{R})$ mit $(pv)(x) = \Theta v_j + (1-\Theta) v_{j+1}$, wobei $x = (j+\Theta)\Delta x$, $j \in \mathbb{Z}$, $\Theta \in [0,1)$,

bzw.

p ist stückweise konstante Interpolation: (5.5.6b)
$v \in \ell^2 \mapsto pv \in L^2(\mathbb{R})$ mit $(pv)(x) = v_j$, wobei $x \in \left[\left(j - \frac{1}{2}\right)\Delta x, \left(j + \frac{1}{2}\right) \right)$, $j \in \mathbb{Z}$.

5.5.3 Differenzengleichung

Am Anfang sind die Werte U^0 mittels $U^0_\nu = ru_0$ gegeben (r aus (5.5.3a)). Die explizite Differenzengleichung

$$U^{\mu+1}_\nu = \sum_{j\in\mathbb{Z}} a_j U^\mu_{\nu+j} \tag{5.5.7}$$

erlaubt dann, den nächsten Zeitschritt $U^{\mu+1}$ aus U^μ zu berechnen. In der Praxis ist $\sum_{j\in\mathbb{Z}}$ eine endliche Summe, d.h. fast alle a_j verschwinden.

Beispiel 5.5.3 *a) Im hyperbolischen Falle ersetzt man* $\frac{\partial}{\partial t}u$ *durch den Differenzenquotienten* $\frac{u(t+\Delta t,x)-u(t,x)}{\Delta t}$ *und* $A = a\frac{\partial u}{dx}$ *durch* $a\frac{u(t,x+\Delta x)-u(t,x)}{\Delta x}$. *Löst man* $\frac{u(t+\Delta t,x)-u(t,x)}{\Delta t} = a\frac{u(t,x+\Delta x)-u(t,x)}{\Delta x}$ *nach* $u(t+\Delta t,x)$ *auf, erhält man mit (5.5.2)*

$$u(t+\Delta t,x) = u(t,x) + \frac{a\Delta t}{\Delta x}\left(u(t,x+\Delta x) - u(t,x)\right), \quad \textit{d.h.} \quad U^{\mu+1}_\nu = (1-a\lambda)\,U^\mu_\nu + a\lambda U^\mu_{\nu+1}. \tag{5.5.8a}$$

b) Ersetzt man die rechtsseitige Differenz $\frac{u(t,x+\Delta x)-u(t,x)}{\Delta x}$ *durch die symmetrische Differenz* $\frac{u(t,x+\Delta x)-u(t,x-\Delta x)}{2\Delta x}$, *gelangt man zur Differenzengleichung*

$$U^{\mu+1}_\nu = -\frac{a\lambda}{2}U^\mu_{\nu-1} + U^\mu_\nu + \frac{a\lambda}{2}U^\mu_{\nu+1}. \tag{5.5.8b}$$

c) Eine weitere Ersetzung von $\frac{u(t+\Delta t,x)-u(t,x)}{\Delta t}$ *durch* $\frac{u(t+\Delta t,x)-[u(t,x+\Delta x)+u(t,x-\Delta x)]/2}{\Delta t}$ *liefert*

$$U^{\mu+1}_\nu = \frac{1-a\lambda}{2}U^\mu_{\nu-1} + \frac{1+a\lambda}{2}U^\mu_{\nu+1}. \tag{5.5.8c}$$

d) Im parabolischen Falle $A = \frac{\partial^2}{dx^2}$ *liegen die Ersetzungen* $\frac{u(t+\Delta t,x)-u(t,x)}{\Delta t}$ *für* $\frac{\partial}{\partial t}u$ *und der zweite Differenzenquotient* $\frac{u(t,x-\Delta x)-2u(t,x)+u(t,x+\Delta x)}{\Delta x^2}$ *für* $\frac{\partial^2 u}{dx^2}$ *nahe und führen mit der Definition* $\lambda = \Delta t/\Delta x^2$ *aus (5.5.2) auf* $\frac{u(t+\Delta t,x)-u(t,x)}{\Delta t} = \frac{2u(t,x)-u(t,x-\Delta x)-u(t,x+\Delta x)}{\Delta x^2}$, *d.h.*

$$U^{\mu+1}_\nu = \lambda U^\mu_{\nu-1} + (1-2\lambda)\,U^\mu_\nu + \lambda U^\mu_{\nu+1}. \tag{5.5.9}$$

Die Differenzengleichung (5.5.7) beschreibt eine lineare Abbildung $U^\mu \mapsto U^{\mu+1}$ und definiert den linearen Operator

$$C : \ell^p \to \ell^p, \quad (CU)_\nu := \sum_{j\in\mathbb{Z}} a_j U_{\nu+j} \text{ für } U \in \ell^p. \tag{5.5.10}$$

Voraussetzung 5.5.4 *Im folgenden darf der Differenzenoperator C (und damit auch die Koeffizienten a_j) vom Parameter λ und der Schrittweite Δt abhängen: $C = C(\lambda,\Delta t)$ (die Abhängigkeit von Δx ergibt sich automatisch mittels (5.5.2)).*

Ein Beispiel für einen Δt-abhängigen Differenzenoperator ergibt sich in

Beispiel 5.5.5 *Zu* $A = a\frac{\partial}{dx}$ *sei* $C(\lambda)$ *ein passender Differenzenoperator. Zu* $A = a\frac{\partial}{dx} + b$ *ergibt sich dann* $C'(\lambda,\Delta t) := C(\lambda) + \Delta t\cdot b$, *d.h.* a_0 *aus (5.5.7) wird durch* $a'_0 := a_0 + \Delta t\cdot b$ *ersetzt.*

Übungsaufgabe 5.5.6 *Man zeige: a) Falls* $\sum_{j\in\mathbb{Z}}|a_j| < \infty$, *ist* $C(\lambda,\Delta t) \in L\left(\ell^\infty,\ell^\infty\right)$,
b) falls $\sup_{j\in\mathbb{Z}}|a_j| < \infty$, *ist* $C(\lambda,\Delta t) \in L\left(\ell^2,\ell^2\right)$.

Die μ-fache Anwendung von $C(\lambda,\Delta t)$ liefert $U^\mu = C^\mu U^0$ und damit $U(\mu\Delta t,\nu\Delta x) = \left(C^\mu U^0\right)_\nu$.

Bemerkung 5.5.7 *Hier sind die Koeffizienten a_j Zahlen. Falls anstelle der skalaren Gleichung $u_t = Au$ mit $u : I\times\mathbb{R}\to\mathbb{R}$ eine vektorwertige Gleichung mit $u : I\times\mathbb{R}\to\mathbb{R}^N$ vorliegt, ergeben sich für die Koeffizienten $N\times N$-Matrizen a_j. Der vektorwertige Fall sind in §5.13 diskutiert werden.*

5.6 Konsistenz, Konvergenz und Stabilität

Im Folgenden wird die Restriktion r aus §5.5.2 verwendet. Man beachte, dass man zwischen Konsistenz bezüglich ℓ^2 und ℓ^∞ zu unterscheiden hat.

Definition 5.6.1 (Konsistenz) *$B_0 \subset D_A$ sei eine dichte Teilmenge von B. ℓ^p sei passend zu B gewählt. Mit $u(t) = T(t)u_0$, $u_0 \in B_0$, sei die Lösung von (5.1.1a,b) bezeichnet. Der* lokale Diskretisierungsfehler *τ wird mittels*

$$\tau(t) = \frac{1}{\Delta t}\left[ru(t+\Delta t) - C(\lambda,\Delta t)ru(t)\right]$$

definiert. Das Differenzenschema $C(\lambda,\Delta t)$ heißt konsistent *(bezüglich ℓ^p), falls*

$$\sup\left\{\|\tau(t)\|_{\ell^p} : 0 \le t \le T - \Delta t\right\} \to 0 \qquad \text{für } \Delta t \to 0 \text{ und alle } u_0 \in B_0.$$

Die letzte Bedingung ist äquivalent zu $\sup\left\{\|r\left[T(\Delta t) - C(\lambda,\Delta t)\right]T(t)u_0\|_{\ell^p} : 0 \le t \le T - \Delta t\right\} = o(\Delta t)$ für alle $u_0 \in B_0$.

Man beachte, dass die folgende Konvergenzdefinition den gesamten Banach-Raum B und keinen dichten Teilraum B_0 verwendet.

Definition 5.6.2 (Konvergenz) *Für alle $u_0 \in B$ bezeichne $u(t) = T(t)u_0$ die exakte Lösung. Das Differenzenschema $C(\lambda,\Delta t)$ heißt* konvergent *(bezüglich ℓ^p), falls*

$$\|ru(t) - C(\lambda,\Delta t)^\mu ru_0\|_{\ell^p} \to 0 \qquad \text{für } \Delta t \to 0 \text{ und } \mu\Delta t \to t \in I = [0,T].$$

Definition 5.6.3 (Stabilität) *Das Differenzenschema $C(\lambda,\Delta t)$ heißt* stabil *(bezüglich ℓ^p), falls*

$$\sup\{\|C(\lambda,\Delta t)^\mu\|_{\ell^p \leftarrow \ell^p} :\ \Delta t \ge 0,\ \mu \in \mathbb{N}_0,\ 0 \le \mu\Delta t \le T\} < \infty. \tag{5.6.1}$$

Wenn (5.6.1) nur für gewisse Werte von λ zutrifft, heißt das Verfahren bedingt stabil. *Gilt (5.6.1) dagegen für alle $\lambda > 0$, heißt das Verfahren* unbedingt stabil.

Im Falle von (5.6.1) ist die *Stabilitätskonstante* definiert durch

$$K = K(\lambda) := \sup\{\|C(\lambda,\Delta t)^\mu\|_{\ell^p \leftarrow \ell^p} :\ \Delta t \ge 0,\ \mu \in \mathbb{N}_0,\ 0 \le \mu\Delta t \le T\}.$$

5.7 Sätze

Zunächst wird gezeigt, dass Konsistenz und Stabilität zusammen mit einigen technischen Voraussetzungen die Konvergenz impliziert.

Satz 5.7.1 (Konvergenzsatz) *Vorausgesetzt werden: (a) Bedingung (5.5.4) bezüglich ℓ^p an r, (b) Voraussetzung 5.4.4 an $T(t)$ (äquivalent zu (5.4.1)), (c) Stabilität des Differenzenschema $C(\lambda,\Delta t)$ bezüglich ℓ^p, (d) Konsistenz bezüglich ℓ^p. Dann ist das Differenzenverfahren konvergent bezüglich ℓ^p.*

Beweis. a) Zu einem Anfangswert $u_0 \in B_0 \subset D_A$ (B_0 dichte Teilmenge von B) sei $u(t) = T(t)u_0$ definiert. Wir spalten wie folgt auf:

$$\begin{aligned} ru(t) - C(\lambda,\Delta t)^\mu ru_0 &= r\left[u(t) - u(\mu\Delta t)\right] + \left[rT(\mu\Delta t) - C(\lambda,\Delta t)^\mu r\right]u_0 \\ &= r\left[u(t) - u(\mu\Delta t)\right] + \left[rT(\Delta t)^\mu - C(\lambda,\Delta t)^\mu r\right]u_0. \end{aligned}$$

Es gilt die teleskopartige Darstellung $rA^\mu - B^\mu r = \sum_{\nu=0}^{\mu-1} B^\nu\left[rA - Br\right]A^{\mu-\nu-1}$. Mit $A := T(\Delta t)$ und $B := C(\lambda,\Delta t)$ folgt

$$\begin{aligned} &\|ru(t) - C(\lambda,\Delta t)^\mu ru_0\|_{\ell^p} \\ &\le \|r\|_{\ell^p \leftarrow B}\,\|u(t) - u(\mu\Delta t)\|_B + \sum_{\nu=0}^{\mu-1} \|C(\lambda,\Delta t)^\nu\|_{\ell^p \leftarrow \ell^p}\,\|\left[rT(\Delta t) - C(\lambda,\Delta t)r\right]u((\mu-\nu-1)\,\Delta t)\|_{\ell^p} \\ &\le K_r\,\|u(t) - u(\mu\Delta t)\|_B + \sum_{\nu=0}^{\mu-1} K(\lambda)\Delta t\,\|\tau((\mu-\nu-1)\,\Delta t)\|_{\ell^p} \end{aligned}$$

mit der Stabilitätskonstanten $K(\lambda)$. Da $u_0 \in B_0 \subset D_A$, ist $u(t) = T(t)u_0$ stetig (vgl. Bemerkung 5.4.2b), sodass $\|u(t) - u(\mu\Delta t)\|_B \to 0$ für $\mu\Delta t \to t$. Damit ist der erste Summand eine Nullfolge.

Der lokale Diskretisierungsfehler τ strebt aufgrund der Konsistenzvoraussetzung gleichmäßig gegen null. Mit $\|\tau((\mu-\nu-1)\,\Delta t)\|_{\ell^p} \le \varepsilon$ gilt aber auch $\sum_{\nu=0}^{\mu-1} K(\lambda)\Delta t\,\|\tau((\mu-\nu-1)\,\Delta t)\|_{\ell^p} \le \mu K(\lambda)\Delta t\varepsilon \le K(\lambda)T\varepsilon$ wegen $\mu\Delta t \le T$, sodass die gesamte Summe gegen null strebt. Dies beweist die Konvergenz im Falle eines Anfangswertes $u_0 \in B_0$.

b) Für einen allgemeinen Anfangswert $u_0 \in B$ und ein beliebiges $\varepsilon > 0$ findet man ein $u_0^* \in B_0$ mit $\|u_0 - u_0^*\|_B \le \varepsilon/\left(3K_r \max\left\{\|T(t)\|_{B\leftarrow B}, K(\lambda)\right\}\right)$. Die zugehörige Lösung $u^*(t) = T(t)u_0^*$ erfüllt

$$\|r\,[u(t) - u^*(t)]\|_{\ell^p} \le K_r\,\|T(t)\,[u_0 - u_0^*]\|_B \le K_r\,\|T(t)\|_{B\leftarrow B}\,\|u_0 - u_0^*\|_B \le \varepsilon/3$$

und

$$\|C(\lambda,\Delta t)^\mu r u_0 - C(\lambda,\Delta t)^\mu r u_0^*\|_{\ell^p} = \|C(\lambda,\Delta t)^\mu r\,[u_0 - u_0^*]\|_{\ell^p} \le K(\lambda)K_r\,\|u_0 - u_0^*\|_B \le \varepsilon/3.$$

Zusammen mit $\|ru^*(t) - C(\lambda,\Delta t)^\mu r u_0^*\|_{\ell^p} \le \varepsilon/3$ nach Teil a) für hinreichend kleine Δt und $t - \mu\Delta t$, folgt $\|ru(t) - C(\lambda,\Delta t)^\mu r u_0\|_{\ell^p} \le \varepsilon$, sodass auch für allgemeine Anfangswerte $u_0 \in B$ Konvergenz gezeigt ist. ∎

Als nächstes zeigen wir, dass die Stabilität auch notwendig für die Konvergenz ist.

Satz 5.7.2 (Stabilitätssatz) *B und ℓ^p seien passend gewählt. Es gelte (5.5.4), (5.5.5) und (5.4.1). Dann impliziert die Konvergenz (bzgl. ℓ^p) des Differenzenschema die Stabilität (bzgl. ℓ^p).*

Beweis. Der Beweis wird indirekt geführt. Wenn das Differenzenschema nicht stabil ist, gibt es Folgen $\Delta t_\nu > 0, \mu_\nu \in \mathbb{N}_0$ mit $0 \le \mu_\nu\Delta t_\nu \le T$, sodass $\|C(\lambda,\Delta t_\nu)^{\mu_\nu}\|_{\ell^p\leftarrow\ell^p} \to \infty$ für $\nu \to \infty$. Da das Intervall $I = [0,T]$ kompakt ist, können wir zu einer Teilfolge übergehen, sodass $\mu_\nu\Delta t_\nu \to t \in I$. Aus der Konvergenz schließt man

$$\|ru(t) - C(\lambda,\Delta t_\nu)^{\mu_\nu} r u_0\|_{\ell^p} \to 0 \qquad \text{für alle } u_0 \in B,$$

also insbesondere

$$\begin{aligned}\|C(\lambda,\Delta t_\nu)^{\mu_\nu} r u_0\|_{\ell^p} &\le 1 + \|ru(t)\|_{\ell^p} \underset{u(t)=T(t)\,u_0}{\le} 1 + \|r\|_{\ell^p\leftarrow B}\,\|T(t)\|_{B\leftarrow B}\,\|u_0\|_B \\ &\underset{(5.5.4),\,(5.4.1)}{\le} K_r K_T\,\|u_0\|_B =: K_1(u_0) \qquad \text{für } \nu \ge \nu_0\end{aligned}$$

mit hinreichend großem $\nu_0 = \nu_0(u_0)$. Man schließt hieraus, dass $C_\nu := C(\lambda,\Delta t_\nu)^{\mu_\nu} r$ eine Folge von Operatoren ist, die punktweise beschränkt ist, und kann das Korollar 2.7.3 zum Satz über die gleichmäßige Beschränktheit verwenden. Hiernach ist C_ν gleichmäßig beschränkt: Es gibt ein K mit

$$\|C(\lambda,\Delta t_\nu)^{\mu_\nu} r\|_{\ell^p\leftarrow B} \le K \qquad \text{für alle } \nu \in \mathbb{N}.$$

Da nach (5.5.5) p eine beschränkte Rechtsinverse von r darstellt, folgt

$$\|C(\lambda,\Delta t_\nu)^{\mu_\nu}\|_{\ell^p\leftarrow\ell^p} = \|C(\lambda,\Delta t_\nu)^{\mu_\nu} rp\|_{\ell^p\leftarrow\ell^p} \le \|C(\lambda,\Delta t_\nu)^{\mu_\nu} r\|_{\ell^p\leftarrow B}\,\|p\|_{B\leftarrow\ell^p} \underset{(5.5.5)}{\le} KK_p$$

im Widerspruch zur Annahme $\|C(\lambda,\Delta t_\nu)^{\mu_\nu}\|_{\ell^p\leftarrow\ell^p} \to \infty$. ∎

Die Folgerung aus den beiden vorangegangenen Sätzen ist der Äquivalenzsatz:

Satz 5.7.3 (Äquivalenzsatz) *Vorausgesetzt seien (5.4.1), (5.5.4), (5.5.5) und die Konsistenz bezüglich ℓ^p. Dann sind Konvergenz und Stabilität (jeweils bezüglich ℓ^p) äquivalent.*

5.8 Hinreichende und notwendige Bedingungen für Stabilität

Die folgenden Resultate gehören zu der klassischen Stabilitätstheorie von Lax-Richtmyer[55].

[55] P.D. Lax und R.D. Richtmyer: *Survey of the stability of linear difference equations.* Comm. Pure and Appl. Math., **9** (1956) 267-293

Kriterium 5.8.1 *Falls* $\|C(\lambda,\Delta t)\|_{\ell^p\leftarrow\ell^p} \le 1+K_\lambda\Delta t$ *für alle* $\Delta t>0$, *so ist das Differenzenverfahren stabil mit der Stabilitätskonstanten* $K(\lambda) := e^{TK_\lambda}$.

Beweis. $\|C(\lambda,\Delta t)^\mu\|_{\ell^p\leftarrow\ell^p} \le \|C(\lambda,\Delta t)\|^\mu_{\ell^p\leftarrow\ell^p} \le (1+K_\lambda\Delta t)^\mu \underset{\text{Üb. 2.4.10a}}{\le} \left(e^{K_\lambda\Delta t}\right)^\mu = e^{K_\lambda\mu\Delta t} \underset{\mu\Delta t\le T}{\le} e^{K_\lambda T}$. ∎

Zur Abschätzung von $\|C(\lambda,\Delta t)\|_{\ell^p\leftarrow\ell^p}$ dient die

Bemerkung 5.8.2 *Das Differenzenverfahren (5.5.10) erfüllt* $\|C(\lambda,\Delta t)\|_{\ell^p\leftarrow\ell^p} \le \sum|a_j|$.

Beweis. Es ist $C(\lambda,\Delta t)=\sum a_jE_j$, wobei der Verschiebungsoperator E_j definiert ist durch $(E_jU)_\nu := U_{j+\nu}$. Da Verschiebungen die ℓ^p-Norm von U invariant lassen, gilt $\|E_j\|_{\ell^p\leftarrow\ell^p}=1$. Daher $\|C(\lambda,\Delta t)\|_{\ell^p\leftarrow\ell^p} = \|\sum a_jE_j\|_{\ell^p\leftarrow\ell^p} \le \sum|a_j|\,\|E_j\|_{\ell^p\leftarrow\ell^p} = \sum|a_j|$. ∎

Die Kombination der vorhergehenden Resultate ergibt das

Korollar 5.8.3 *Wenn* $\sum|a_j| \le 1+K_\lambda\Delta t$ *für alle* $\Delta t>0$, *so ist das Differenzenverfahren stabil (bezüglich der* ℓ^2- *und* ℓ^∞-*Norm) mit der Stabilitätskonstanten* $K(\lambda) := e^{TK_\lambda}$.

Definition 5.8.4 *Das Differenzenverfahren (5.5.10) heißt positiv, falls für alle Koeffizienten* $a_j \ge 0$ *gilt.*

Positive Differenzenverfahren haben die Eigenschaft, nichtnegative Gitterfunktionen $U\in\ell^p$ (d.h. $U_j\ge 0$ für alle j) in nichtnegative Gitterfunktionen $C(\lambda,\Delta t)U$ abzubilden. Das Korollar 5.8.3 liefert das

Kriterium 5.8.5 *Positive Differenzenverfahren (5.5.10) mit* $\sum a_j = 1+\mathcal{O}(\Delta t)$ *sind stabil.*

Wir können auch ein Kriterium für Instabilität formulieren:

Kriterium 5.8.6 *Sei* $\sum a_j \ge 1+\Delta t\,c(\Delta t)$ *mit* $\lim_{\Delta t\to 0} c(\Delta t)=\infty$ (*z.B.* $\sum a_j \ge konst > 1$). *Dann ist das Differenzenverfahren (5.5.10) instabil (bezüglich der* ℓ^2- *und* ℓ^∞-*Norm).*

Beweis. a) Im Falle von ℓ^∞ wähle man $U^0\in\ell^\infty$ als die Konstante 1, d.h. $U^0_j=1$ für alle $j\in\mathbb{Z}$. Für $U^1 = C(\lambda,\Delta t)U^0$ findet man ζU^0 mit $\zeta := \sum a_j$ und entsprechend $U^\mu = C(\lambda,\Delta t)^\mu U^0 = \zeta^\mu U^0$. Also ist $\|C(\lambda,\Delta t)^\mu\|_{\ell^p\leftarrow\ell^p} \ge \zeta^\mu \ge [1+\Delta t c(\Delta t)]^\mu$. Mit Übungsaufgabe 5.8.7a folgt die Behauptung.

b) Im Falle von ℓ^2 ist das obige U^0 nicht zulässig, da es keine endliche ℓ^2-Norm besitzt. Stattdessen werden wir den Beweis im Anschluss an Satz 5.10.1 nachliefern. ∎

Übungsaufgabe 5.8.7 *Es gelte* $\lim_{\Delta t\to 0} c(\Delta t)=\infty$. *a) Man beweise*

$$\sup\{[1+\Delta t\,c(\Delta t)]^\mu :\ \mu\in\mathbb{N}_0,\ \Delta t>0,\ \mu\Delta t\le T\}=\infty.$$

b) Ferner zeige man: Für jede Konstante $K>0$ *und alle* $\mu\in\mathbb{N}$ *mit* $\mu\le T/\Delta t$ *erfüllt* $\sqrt[\mu]{K}$ *die Ungleichung* $\sqrt[\mu]{K}\le 1+C_\rho\Delta t$. Hinweis: *Für* $c(\Delta t) := \left(\sqrt[\mu]{K(\lambda)}-1\right)/\Delta t$ *beweist man, dass* $C_\rho := \sup_{\Delta t\ge 0} c(\Delta t) < \infty$.

Wir versuchen, die Kriterien auf die Beispiele aus §5.5.3 anzuwenden.

Beispiel 5.8.8 *a) In (5.5.8a) lauten die Nicht-Null-Koeffizienten* $a_0 = 1-a\lambda$ *und* $a_1 = a\lambda$. *Für* λ *mit* $0\le a\lambda\le 1$ *ist (5.5.8a) ein positives Verfahren mit* $\sum a_j = 1$, *also stabil gemäß Korollar 5.8.3.*

b) In (5.5.8b) lauten die Nicht-Null-Koeffizienten $a_{-1} = -\frac{a\lambda}{2}$, $a_0=1$, $a_1=\frac{a\lambda}{2}$. *Der triviale Fall* $a=0$ *sei ausgeschlossen. Korollar 5.8.3 lässt sich nicht anwenden, da* $\sum|a_j| = 1+a\lambda$. *Es ist kein positives Verfahren.*

c) In (5.5.8c) lauten die Nicht-Null-Koeffizienten $a_{-1} = \frac{1-a\lambda}{2}$ *und* $a_1 = \frac{1+a\lambda}{2}$. *Unter der Bedingung* $|a\lambda|\le 1$ *ist das Verfahren positiv, und Kriterium 5.8.5 sichert die Stabilität* ($\sum a_j = 1$).

d) In (5.5.9) lauten die Nicht-Null-Koeffizienten $a_{-1}=\lambda$, $a_0 = 1-2\lambda$, $a_1=\lambda$. *Für* $\lambda\in(0,1/2]$ *ist das Verfahren positiv, und Kriterium 5.8.5 sichert die Stabilität.*

Eine interessante Frage ist, inwieweit ein stabiles Verfahren nach einer Störung stabil bleibt.

Lemma 5.8.9 (Störungslemma) *Sei $C(\lambda,\Delta t)$ stabil in ℓ^p mit der Stabilitätskonstanten $K(\lambda)$. Die Störung $D(\lambda,\Delta t)$ sei beschränkt durch $\|D(\lambda,\Delta t)\|_{\ell^p\leftarrow\ell^p}\le C_D\Delta t$. Dann ist $C'(\lambda,\Delta t):=C(\lambda,\Delta t)+D(\lambda,\Delta t)$ wieder stabil in ℓ^p mit der Stabilitätskonstanten*[56] *$K'(\lambda)\le K(\lambda)e^{K(\lambda)C_DT}$, wobei $I=[0,T]$ das gegebene Zeitintervall sei. Dies Resultat gilt auch in dem Falle, dass $C(\lambda,\Delta t)$ und $D(\lambda,\Delta t)$ nichtvertauschbare*[57] *Operatoren sind.*

Beweis. Wir haben $C'(\lambda,\Delta t)^\mu$ abzuschätzen. Die einfache binomische Formel gilt nur für vertauschbare Summanden. Im allgemeinen Fall ist

$$[C+D]^\mu=\sum_{m=0}^{\mu}\sum_{\alpha_1+\ldots+\alpha_{\mu+1}=\mu-m} C^{\alpha_1}DC^{\alpha_2}D\cdots C^{\alpha_m}DC^{\alpha_{m+1}},$$

wobei die zweite Summe über alle $\alpha_j\in\mathbb{N}_0$ mit $\alpha_1+\ldots+\alpha_{\mu+1}=\mu-m$ geführt wird. Jeder Summand $C^{\alpha_1}DC^{\alpha_2}D\cdots C^{\alpha_m}DC^{\alpha_{m+1}}$ enthält m Faktoren D und $\mu-m$ Faktoren C. Die Anzahl dieser Summanden zu $m\in[0,\mu]$ ist $\binom{\mu}{m}$ (vgl. binomische Formel). Zusammen mit der Abschätzung

$$\|C^{\alpha_1}DC^{\alpha_2}D\cdots C^{\alpha_\mu}DC^{\alpha_{\mu+1}}\|\le\|C^{\alpha_1}\|\,\|D\|\,\|C^{\alpha_2}\|\,\|D\|\,\cdots\,\|C^{\alpha_m}\|\,\|D\|\,\|C^{\alpha_{m+1}}\|\le K(\lambda)^{m+1}\,(C_D\Delta t)^m$$

erreichen wir die Ungleichung

$$\begin{aligned}\|C'(\lambda,\Delta t)^\mu\|_{\ell^p\leftarrow\ell^p} &= \|[C(\lambda,\Delta t)+D(\lambda,\Delta t)]^\mu\|_{\ell^p\leftarrow\ell^p} \le \sum_{m=0}^{\mu}\tbinom{\mu}{m}K(\lambda)^{m+1}\,(C_D\Delta t)^m\\ &= K(\lambda)\sum_{m=0}^{\mu}\tbinom{\mu}{m}\,(K(\lambda)C_D\Delta t)^m = K(\lambda)\,(1+K(\lambda)C_D\Delta t)^\mu\,.\end{aligned}$$

Mit $(1+K(\lambda)C_D\Delta t)^\mu\le\exp(K(\lambda)C_D\mu\Delta t)\underset{\mu\Delta t\le T}{\le}e^{K(\lambda)C_DT}$ (vgl. Übung 2.4.10) folgt die Behauptung. ∎

Bemerkung 5.8.10 *a) Eine einfache Anwendung von Lemma 5.8.9 lautet wie folgt: Der Differentialoperator A in $\frac{\partial}{\partial t}u=Au$ (vgl. (5.1.1a)) sei $A=A_1+A_0$, wobei A_1 Ableitungen mindestens erster Ordnung enthält, sodass $A_1 1=0$, während $A_0u=a_0u$ der Term nullter Ordnung ist. Die Diskretisierung spalte sich entsprechend in $C(\lambda,\Delta t)=C_1(\lambda,\Delta t)+C_0(\lambda,\Delta t)$ auf. Da für C_0 die Abschätzung $\|C_0(\lambda,\Delta t)\|_{\ell^p\leftarrow\ell^p}=\mathcal{O}(\Delta t)$ gilt (wenn die Diskretisierung konsistent sein soll), ergibt sich die Stabilität von $C(\lambda,\Delta t)$ aus der von $C_1(\lambda,\Delta t)$. Dies bedeutet, dass o.B.d.A. der Differentialoperator A ohne Terme nullter Ordnung untersucht zu werden braucht.*

b) Sei $A=A_1$. Die Eigenschaft $A1=0$ (siehe Teil a) zeigt, dass $u=1$ eine Lösung ist. Hieraus leitet man die spezielle Konsistenzbedingung

$$\sum_{j\in\mathbb{Z}}a_j=1\qquad(=I\text{ im matrixwertigen Fall})\tag{5.8.1}$$

für die Koeffizienten von $C(\lambda,\Delta t)$ (vgl. (5.5.10)) ab.

Ein notwendiges Kriterium kann mittels des *Spektralradius*[58]

$$\rho(A):=\sup\{|\lambda|:\ \lambda\text{ ist singulärer Wert von }A\}$$

formuliert werden. Man beachte, dass die Stabilität von der Wahl des Banach-Raumes ℓ^p abhängig sein kann, aber $\rho(C(\lambda,\Delta t))$ nicht von ℓ^p abhängt.

Kriterium 5.8.11 *Eine notwendige Bedingung für Stabilität ist*

$$\rho(C(\lambda,\Delta t))\le 1+\mathcal{O}(\Delta t).$$

[56] Im Falle der Vertauschbarkeit $CD=DC$ verbessert sich die Stabilitätskonstante zu $K'(\lambda)\le K(\lambda)e^{C_DT}$.

[57] Nichtvertauschbarkeit tritt im Allgemeinen auf, wenn die Koeffizienten a_j Matrizen sind (vgl. §5.13).

[58] λ ist regulärer Wert von A, falls $\lambda I-A$ bijektiv ist und die Inverse $(\lambda I-A)^{-1}\in L(B,B)$ existiert. Andernfalls ist λ ein singulärer Wert von A. Im Falle endlich-dimensionaler Vektorräume (z.B. Matrizen) fallen die Begriffe "singulärer Wert" und "Eigenwert" zusammen.

Beweis. Für $\mu \in \mathbb{N}$ gilt $\rho(A^\mu) = \rho(A)^\mu$. Andererseits gilt $\rho(A) \leq \|A\|$ für jede zugeordnete Norm. Damit liefert die Stabilität

$$\rho(C(\lambda,\Delta t))^\mu = \rho(C(\lambda,\Delta t)^\mu) \leq \|C(\lambda,\Delta t)^\mu\|_{\ell^p \leftarrow \ell^p} \leq \|C(\lambda,\Delta t)^\mu\|_{\ell^p \leftarrow \ell^p} \leq K(\lambda)$$

für alle $\mu, \Delta t$ mit $\mu \Delta t \leq T$. Mit Übungsaufgabe 5.8.7b beweist man $\rho(C(\lambda,\Delta t)) \leq 1 + C_\rho \Delta t$. ∎

Bemerkung 5.8.12 *Der Operator $C(\lambda,\Delta t) \in L(\ell^2,\ell^2)$ sei normal, d.h. $C(\lambda,\Delta t)$ vertausche mit dem adjungierte Operator $C(\lambda,\Delta t)^*$. Dann gilt $\rho(C(\lambda,\Delta t)) = \|C(\lambda,\Delta t)\|_{\ell^2 \leftarrow \ell^2}$ und Stabilität liegt genau dann vor, wenn $\rho(C(\lambda,\Delta t)) \leq 1 + \mathcal{O}(\Delta t)$.*

Beweis. a) Der normale Operator sei mit A bezeichnet. Zuerst zeigen wir, dass $\|A^2\|_{\ell^2 \leftarrow \ell^2} = \|A^*A\|_{\ell^2 \leftarrow \ell^2}$. Mit $\langle AAu, AAu\rangle_{\ell^2} = \langle A^*AAu, Au\rangle_{\ell^2} = \langle AA^*Au, Au\rangle_{\ell^2} = \langle A^*Au, A^*Au\rangle_{\ell^2}$ folgt

$$\|A^2\|^2_{\ell^2 \leftarrow \ell^2} = \sup_{\|u\|_{\ell^2}=1} \langle AAu, AAu\rangle_{\ell^2} = \sup_{\|u\|_{\ell^2}=1} \langle A^*Au, A^*Au\rangle_{\ell^2} = \|A^*A\|^2_{\ell^2 \leftarrow \ell^2}.$$

Da auch

$$\|A^*A\|_{\ell^2 \leftarrow \ell^2} = \sup_{\|u\|_{\ell^2}=\|v\|_{\ell^2}=1} \langle u, A^*Av\rangle_{\ell^2} = \sup_{\|u\|_{\ell^2}=\|v\|_{\ell^2}=1} \langle Au, Av\rangle_{\ell^2} \underset{u=v}{\geq} \sup_{\|u\|_{\ell^2}=1} \langle Au, Au\rangle_{\ell^2} = \|A\|^2_{\ell^2 \leftarrow \ell^2},$$

ist $\|A\|^2_{\ell^2 \leftarrow \ell^2} \geq \|A^2\|_{\ell^2 \leftarrow \ell^2}$ gezeigt. Wegen $\|A^2\|_{\ell^2 \leftarrow \ell^2} \leq \|A\|^2_{\ell^2 \leftarrow \ell^2}$ ist zusammen die Gleichheit $\|A^2\|_{\ell^2 \leftarrow \ell^2} = \|A^*A\|_{\ell^2 \leftarrow \ell^2}$ bewiesen. Analog folgt $\|A\|^n_{\ell^2 \leftarrow \ell^2} = \|A^n\|_{\ell^2 \leftarrow \ell^2}$ für alle $n = 2^k$ $(k \in \mathbb{N})$ und dann für alle $n \in \mathbb{N}$.

b) Es gilt die Charakterisierung $\rho(A) = \lim_{n\to\infty} \sqrt[n]{\|A^n\|_{\ell^2 \leftarrow \ell^2}}$. Nach Teil a) also $\rho(A) = \|A\|_{\ell^2 \leftarrow \ell^2}$.

c) Nach Kriterium 5.8.11 ist $\rho(C(\lambda,\Delta t)) \leq 1 + \mathcal{O}(\Delta t)$ notwendig, während $\|C(\lambda,\Delta t)\|_{\ell^2 \leftarrow \ell^2} \leq 1 + \mathcal{O}(\Delta t)$ nach Kriterium 5.8.1 hinreichend ist. Da $\rho(C(\lambda,\Delta t)) = \|C(\lambda,\Delta t)\|_{\ell^2 \leftarrow \ell^2}$, sind beide Ungleichungen aber identisch. ∎

Da Bemerkung 5.8.12 ein relativ einfaches Stabilitätskriterium an die Hand gibt, kann man sich fragen, ob Ähnliches auch für allgemeinere Operatoren gilt. Dazu führen wir "fast normale Operatoren" ein:

$C(\lambda,\Delta t)$ ist *fast normal*, falls $\|C(\lambda,\Delta t)C(\lambda,\Delta t)^* - C(\lambda,\Delta t)^*C(\lambda,\Delta t)\|_{\ell^2 \leftarrow \ell^2} \leq M\,(\Delta t)^2\,\|C(\lambda,\Delta t)\|^2_{\ell^2 \leftarrow \ell^2}$

für eine Konstante $M < \infty$.

Kriterium 5.8.13 *Wenn $C(\lambda,\Delta t)$ fast normal ist, sind Stabilität bezüglich ℓ^2 und die Abschätzung $\|C(\lambda,\Delta t)\|_{\ell^2 \leftarrow \ell^2} \leq 1 + \mathcal{O}(\Delta t)$ äquivalent.*

Beweis. a) Da $\|C(\lambda,\Delta t)\|_{\ell^2 \leftarrow \ell^2} \leq 1 + \mathcal{O}(\Delta t)$ nach Kriterium 5.8.1 hinreichend ist, bleibt nur die Notwendigkeit zu beweisen. Im folgenden schreiben wir kurz A anstelle von $C(\lambda,\Delta t)$.

b) Zunächst beweisen wir per Induktion, dass $(A^*)^\mu A^\mu$ in höchstens $\mu^2/2$ Schritten in $(A^*A)^\mu$ umgeordnet werden kann. Im Falle von $\mu = 1$ ist keine Umordnung nötig, und $0 \leq \binom{1}{2}/2$ beweist den Induktionsanfang. *Induktionsschritt*: Da

$$(A^*A)^{\mu+1} \underset{\mu^2/2 \text{ Schritte}}{\hookrightarrow} (A^*)^\mu\, A^\mu A^* A \underset{\mu \text{ Schritte}}{\hookrightarrow} (A^*)^{\mu+1}\, A^{\mu+1}.$$

und $\frac{\mu^2}{2} + \mu \leq \frac{(\mu+1)^2}{2}$, ist die Behauptung gezeigt.

c) Pro Vertauschung ändert sich die Matrixnorm höchstens um $M\,(\Delta t)^2\,\|A\|^{2\mu}_{\ell^2 \leftarrow \ell^2}$. Dazu seien in folgendem Produkt die Faktoren A_j, $1 \leq j \leq 2\mu$, entweder A oder A^*:

$$\begin{aligned}
&\|A_1 \cdots A_\nu A A^* A_{\nu+3} \cdots A_{2\mu} - A_1 \cdots A_\nu A^* A A_{\nu+3} \cdots A_{2\mu}\|_{\ell^2 \leftarrow \ell^2} \\
&\leq \|A_1 \cdots A_\nu\|_{\ell^2 \leftarrow \ell^2}\, \|AA^* - A^*A\|_{\ell^2 \leftarrow \ell^2}\, \|A_{\nu+3} \cdots A_{2\mu}\|_{\ell^2 \leftarrow \ell^2} \\
&\leq \|A\|^\nu_{\ell^2 \leftarrow \ell^2} \left[M\,(\Delta t)^2\,\|A\|^2_{\ell^2 \leftarrow \ell^2}\right] \|A\|^{2\mu-\nu-2}_{\ell^2 \leftarrow \ell^2} = M\,(\Delta t)^2\,\|A\|^{2\mu}_{\ell^2 \leftarrow \ell^2} \qquad \text{für alle } 0 \leq \nu \leq 2\mu - 2,
\end{aligned}$$

wobei die Fast-Normalität und $\|A^*\|_{\ell^2 \leftarrow \ell^2} = \|A\|_{\ell^2 \leftarrow \ell^2}$ ausgenutzt wurden.

d) Bildet man in

$$\langle A^\mu u, A^\mu u\rangle_{\ell^2} = \langle u, (A^*)^\mu\, A^\mu u\rangle_{\ell^2} \underset{\text{nach b) und c)}}{\geq} \langle u, (A^*A)^\mu\, u\rangle_{\ell^2} - \frac{M}{2}\,(\mu\Delta t)^2\,\|A\|^{2\mu}_{\ell^2\leftarrow\ell^2}\,\|u\|^2_{\ell^2}$$

das Supremum über alle $u \in \ell^2$ mit $\|u\|_{\ell^2} = 1$, wird die linke Seite zu $\|A^\mu\|^2_{\ell^2\leftarrow\ell^2}$ und die rechte Seite zu $\left(1 - \frac{M}{2}\,(\mu\Delta t)^2\right)\|A\|^{2\mu}_{\ell^2\leftarrow\ell^2}$ wegen $\sup \langle u, (A^*A)^\mu\, u\rangle_{\ell^2} = \|(A^*A)^\mu\|_{\ell^2\leftarrow\ell^2} \underset{A^*A \text{ normal}}{=} \|A^*A\|^\mu_{\ell^2\leftarrow\ell^2} = \|A\|^{2\mu}_{\ell^2\leftarrow\ell^2}$. Dies zeigt zusammen mit der Stabilität, dass

$$\left(1 - \frac{M}{2}\,(\mu\Delta t)^2\right)\|A\|^{2\mu}_{\ell^2\leftarrow\ell^2} \leq \|A^\mu\|^2_{\ell^2\leftarrow\ell^2} \leq K(\lambda)^2. \tag{5.8.2}$$

Wir schränken $\mu, \Delta t$ durch $\mu\Delta t \leq \min\{1/\sqrt{M}, T\}$ ein, sodass $1 - \frac{M}{2}(\mu\Delta t)^2 \geq \frac{1}{2}$. Ungleichung (5.8.2) impliziert $\|A\|_{\ell^2\leftarrow\ell^2} \leq \sqrt[\mu]{2K(\lambda)^2}$. Übungsaufgabe 5.8.7b liefert $\|A\|_{\ell^2\leftarrow\ell^2} = \|C(\lambda,\Delta t)\|_{\ell^2\leftarrow\ell^2} \leq 1 + \mathcal{O}(\Delta t)$. ∎

Für weitergehende Stabilitätskriterien im Falle von ℓ^2 wird das Hilfsmittel der Fourier-Reihen eingesetzt.

5.9 Fourier-Analyse

Eine 2π-periodische Funktionen $f \in L^2_{2\pi}(\mathbb{R})$ ist die 2π-periodische Fortsetzung von $f \in L^2(0, 2\pi)$. Die zugehörige Fourier-Reihe lautet

$$\frac{1}{\sqrt{2\pi}}\sum_{\alpha\in\mathbb{Z}} \varphi_\alpha e^{i\alpha\xi} \qquad \text{mit } \varphi_\alpha := \frac{1}{\sqrt{2\pi}}\int_0^{2\pi} f(\xi)e^{-i\alpha\xi}d\xi \quad (\alpha \in \mathbb{Z}),$$

wobei i die imaginäre Einheit ist. Für glatte, 2π-periodische Funktionen f zeigt man, dass die Fourier-Reihe gleichmäßig (in der Maximumnorm) gegen f konvergiert. Geht man zu L^2-Funktionen über, gilt die Konvergenz nur im Sinne der L^2-Norm. In diesem Sinne gilt $f(\xi) = \frac{1}{\sqrt{2\pi}}\sum_{\alpha\in\mathbb{Z}} \varphi_\alpha e^{i\alpha\xi}$. Maßgebend für diese Eigenschaft ist die Isometrie (Norm-Gleichheit):

$$\|f\|_{L^2(0,2\pi)} = \|\varphi\|_{\ell^2}\,, \qquad \text{wobei } \varphi = (\varphi_\alpha)_{\alpha\in\mathbb{Z}}\,. \tag{5.9.1}$$

Der Übergang von $f \in L^2(0, 2\pi)$ zu seinen Fourier-Koeffizienten $\varphi \in \ell^2$ ist die *Fourier-Analyse*, die im Folgenden mit $\mathcal{F}$ bezeichnet wird:

$$\mathcal{F}f = \varphi.$$

Dagegen heißt die Abbildung $\varphi \mapsto \frac{1}{\sqrt{2\pi}}\sum_{\alpha\in\mathbb{Z}} \varphi_\alpha e^{i\alpha\xi} = f$ die *Fourier-Synthese* und ist die Inverse $\mathcal{F}^{-1}$.

Die Lösungen U^μ des Differenzenschemas sind für $p = 2$ ℓ^2-Folgen. Wir können ihnen daher die 2π-periodischen Funktionen $\hat{U}^\mu$ zuordnen:

$$\hat{U}^\mu := \mathcal{F}^{-1}U^\mu, \qquad \hat{U}^\mu(\xi) = \frac{1}{\sqrt{2\pi}}\sum_{\alpha\in\mathbb{Z}} U^\mu_\alpha\, e^{i\alpha\xi}.$$

Das Differenzenschema $U^{\mu+1} = C(\lambda,\Delta t)U^\mu$ ist äquivalent zu

$$\hat{U}^{\mu+1} = \mathcal{F}^{-1}U^{\mu+1} = \mathcal{F}^{-1}C(\lambda,\Delta t)U^\mu = \mathcal{F}^{-1}C(\lambda,\Delta t)\mathcal{F}\mathcal{F}^{-1}U^\mu = \hat{C}(\lambda,\Delta t)\hat{U}^\mu$$
$$\text{mit} \qquad \hat{C}(\lambda,\Delta t) := \mathcal{F}^{-1}C(\lambda,\Delta t)\mathcal{F}.$$

Während $C(\lambda,\Delta t) \in L(\ell^2,\ell^2)$, ist $\hat{C}(\lambda,\Delta t)$ eine Abbildung aus $L(L^2(0,2\pi), L^2(0,2\pi))$.

Übungsaufgabe 5.9.1 *Aus der Isometrie (5.9.1) schließe man auf*

$$\|\mathcal{F}\|_{\ell^2\leftarrow L^2(0,2\pi)} = \|\mathcal{F}^{-1}\|_{L^2(0,2\pi)\leftarrow\ell^2} = 1 \tag{5.9.2}$$

(d.h. $\mathcal{F}$ und $\mathcal{F}^{-1}$ sind unitär) und zeige $\|A\|_{\ell^2\leftarrow\ell^2} = \|\mathcal{F}^{-1}A\|_{L^2(0,2\pi)\leftarrow\ell^2}$ für alle $A \in L(\ell^2,\ell^2)$ und $\|B\|_{L^2(0,2\pi)\leftarrow\ell^2} = \|B\mathcal{F}\|_{L^2(0,2\pi)\leftarrow L^2(0,2\pi)}$ für alle $B \in L(\ell^2, L^2(0,2\pi))$.

Die für die Stabilität wichtige Norm $\|C(\lambda,\Delta t)^\mu\|_{\ell^2\leftarrow\ell^2}$ überträgt sich auf $\hat{C}(\lambda,\Delta t)$, indem man die Gleichheiten der vorangegangenen Übungsaufgabe nutzt:

$$\begin{aligned}\|C(\lambda,\Delta t)^\mu\|_{\ell^2\leftarrow\ell^2} &= \left\|\mathcal{F}^{-1}C(\lambda,\Delta t)^\mu\right\|_{L^2(0,2\pi)\leftarrow\ell^2} = \|\mathcal{F}^{-1}C(\lambda,\Delta t)^\mu\mathcal{F}\|_{L^2(0,2\pi)\leftarrow L^2(0,2\pi)} \\ &= \|\left[\mathcal{F}C(\lambda,\Delta t)\mathcal{F}^{-1}\right]^\mu\|_{L^2(0,2\pi)\leftarrow L^2(0,2\pi)} = \|\hat{C}(\lambda,\Delta t)^\mu\|_{L^2(0,2\pi)\leftarrow L^2(0,2\pi)}.\end{aligned} \tag{5.9.3}$$

Damit erhält man die

Bemerkung 5.9.2 *Eine äquivalente Definition der ℓ^2-Stabilität mit der Stabilitätskonstanten $K(\lambda)$ ist $\|\hat{C}(\lambda,\Delta t)^\mu\|_{L^2(0,2\pi)\leftarrow L^2(0,2\pi)} \le K(\lambda)$ für alle $\mu\in\mathbb{N}_0$ und $\Delta t>0$ mit $\mu\Delta t\le T$.*

Zur konkreten Bestimmung von $\hat{C}(\lambda,\Delta t)$ betrachten wir einen Summanden C_j aus $C(\lambda,\Delta t)=\sum_j C_j$, wobei C_j durch

$$(C_jU)_\nu := a_jU_{\nu+j} \quad \text{für } U\in\ell^2$$

definiert ist (vgl. (5.5.10)). Wie oben dargestellt, gilt $\hat{C}_j\hat{U}=\mathcal{F}^{-1}C_jU$ mit $\hat{U}=\mathcal{F}^{-1}U$, d.h. $\hat{U}\in L^2(0,2\pi)$ ist die Funktion $\frac{1}{\sqrt{2\pi}}\sum_{\alpha\in\mathbb{Z}}U_\alpha e^{i\alpha\xi}$ zu den Koeffizienten $U=(U_\alpha)_{\alpha\in\mathbb{Z}}\in\ell^2$. Definitionsgemäß ist C_jU die verschobene und mit a_j multiplizierte Folge $a_j\,(U_{\alpha+j})_{\alpha\in\mathbb{Z}}$. Die Fourier-Synthese liefert

$$\begin{aligned}\mathcal{F}^{-1}C_jU &= \frac{1}{\sqrt{2\pi}}\sum_{\alpha\in\mathbb{Z}}(a_jU_{\alpha+j})\,e^{i\alpha\xi} = a_j\frac{1}{\sqrt{2\pi}}\sum_{\alpha\in\mathbb{Z}}U_{\alpha+j}e^{i\alpha\xi} \underset{\beta=\alpha+j}{=} a_j\frac{1}{\sqrt{2\pi}}\sum_{\beta\in\mathbb{Z}}U_\beta e^{i(\beta-j)\xi} \\ &= a_je^{-ij\xi}\frac{1}{\sqrt{2\pi}}\sum_{\beta\in\mathbb{Z}}U_\beta e^{i\beta\xi} = a_je^{-ij\xi}\hat{U}.\end{aligned}$$

Der Vergleich mit $\mathcal{F}^{-1}C_jU=\hat{C}_j\hat{U}$ zeigt $\hat{C}_j=a_je^{-ij\xi}$, d.h. die lineare Abbildung $\hat{C}_j: L^2(0,2\pi)\to L^2(0,2\pi)$ ist die Multiplikation mit der Funktion $a_je^{-ij\xi}$. Da $\hat{C}(\lambda,\Delta t)=\mathcal{F}^{-1}C(\lambda,\Delta t)\mathcal{F}=\mathcal{F}^{-1}\sum_j C_j\mathcal{F}=\sum_j\mathcal{F}^{-1}C_j\mathcal{F}=\sum_j\hat{C}_j$, erhalten wir die

Bemerkung 5.9.3 *Der Fourier-transformierte Differenzenoperator lautet $\hat{C}(\lambda,\Delta t)=\sum_{j\in\mathbb{Z}}a_je^{-ij\xi}$ und bedeutet die Multiplikation mit dem trigonometrischen Polynom*

$$G(\xi) := \sum_{j\in\mathbb{Z}}a_je^{-ij\xi}, \tag{5.9.4}$$

das auch charakteristische Funktion (gelegentlich auch "Symbol") zu $C(\lambda,\Delta t)$ heißt.

Übungsaufgabe 5.9.4 *Sei $\phi\in C(\mathbb{R})$ eine beschränkte stetige Funktion. Der Multiplikationsoperator $\Phi: B\to B$ sei definiert mittels $(\Phi(f))(\xi) := \phi(\xi)f(\xi)$ für alle $\xi\in\mathbb{R}$. Man zeige in den beiden Fällen $B=C(\mathbb{R})$ oder $B=L^2(\mathbb{R})$, dass $\Phi\in L(B,B)$ mit der Operatornorm $\|\Phi\|_{B\leftarrow B}=\|\phi\|_\infty$.*

Die Gleichheit (5.9.3) wird damit zu

$$\|C(\lambda,\Delta t)^\mu\|_{\ell^2\leftarrow\ell^2} = \sup\{|G(\xi)^\mu| : \xi\in\mathbb{R}\}. \tag{5.9.5}$$

Dabei kann $\xi\in\mathbb{R}$ auch durch $\xi\in[0,1)$ ersetzt werden, da G 2π-periodisch ist.[59]

5.10 Weitere Kriterien

Die Anwendung der letzten Übungsaufgabe auf $\hat{C}(\lambda,\Delta t)^\mu=\left(\sum_{j\in\mathbb{Z}}a_je^{-ij\xi}\right)^\mu$ liefert

$$\|\hat{C}(\lambda,\Delta t)^\mu\|_{L^2(0,2\pi)\leftarrow L^2(0,2\pi)} = \left\|\left(\sum\nolimits_{j\in\mathbb{Z}}a_je^{-ij\xi}\right)^\mu\right\|_\infty = \left\|\sum\nolimits_{j\in\mathbb{Z}}a_je^{-ij\xi}\right\|_\infty^\mu, \tag{5.10.1}$$

was zusammen mit Bemerkung 5.9.2 zum folgenden Satz führt. Man beachte, dass (5.10.1) nur für skalare Koeffizienten a_j gültig ist.

[59] Wenn die Koeffizienten a_j Matrizen sind (vgl. Bemerkung 5.5.7), ist $|G(\xi)^\mu|$ durch die Spektralnorm $\|G(\xi)^\mu\|_2$ zu ersetzen.

Satz 5.10.1 *Das Differenzenverfahren (5.5.10) ist genau dann in ℓ^2 stabil, wenn die charakteristische Funktion $G(\xi) = \sum_{j\in\mathbb{Z}} a_j e^{-ij\xi}$ der Abschätzung (5.10.2) mit geeignetem K_λ genügt:*

$$|G(\xi)| \le 1 + K_\lambda \Delta t \qquad \text{für alle } \xi \in \mathbb{R}. \tag{5.10.2}$$

Beweis. a) Sei $G(\xi) = \sum_{j\in\mathbb{Z}} a_j e^{-ij\xi}$. Falls (5.10.2) für eine Konstante K_λ zutrifft, ist $\|C(\lambda,\Delta t)\|_{\ell^2\leftarrow\ell^2} = \|\hat{C}(\lambda,\Delta t)\|_{L^2(0,2\pi)\leftarrow L^2(0,2\pi)} = \|G\|_\infty \le 1 + K_\lambda\Delta t$, womit das Korollar 5.8.3 die Stabilität beweist.

b) Man setze $c(\Delta t) := (\|G\|_\infty - 1)/\Delta t$. Falls keine Konstante K_λ mit (5.10.2) existiert, folgt $c(\Delta t) \to \infty$ für $\Delta t \to 0$. Mit Übungsaufgabe 5.8.7a und (5.10.1) folgt die Instabilität. ∎

Beweis von Kriterium 5.8.6 im Falle von ℓ^2. Da $\sum a_j \ge 1 + \Delta t\, c(\Delta t)$ mit $\lim_{\Delta t\to\infty} c(\Delta t) = \infty$ vorausgesetzt ist und $\|G\|_\infty \ge G|_{\xi=0} = \sum a_j$ gilt, kann (5.10.2) nicht gelten, was Instabilität in ℓ^2 beweist. ∎

Die bisherige Analyse bezog sich auf den ℓ^2. Die charakteristische Funktion hat aber auch Konsequenzen bezüglich ℓ^∞.

Lemma 5.10.2 $\|C(\lambda,\Delta t)^\mu\|_{\ell^\infty\leftarrow\ell^\infty} \ge \|G^\mu\|_\infty = \|C(\lambda,\Delta t)^\mu\|_{\ell^2\leftarrow\ell^2}$ *für alle* $\mu \in \mathbb{N}_0$.

Beweis. Man wähle den speziellen Anfangswert U^0 mit $U^0_\nu = \exp(i\nu\xi)$, wobei $\xi \in \mathbb{R}$ durch $|G(\xi)| = \|G\|_\infty$ charakterisiert sei. Man beachte $U^0 \in \ell^\infty$ und $\|U^0\|_\infty = 1$. Anwendung von $C(\lambda,\Delta t)$ liefert

$$U^1 = C(\lambda,\Delta t)U^0 \quad \text{mit } U^1_\nu = \sum_{j\in\mathbb{Z}} a_j U^0_{\nu+j} = \sum_{j\in\mathbb{Z}} a_j e^{i(\nu+j)\xi} = e^{i\xi\nu}\sum_{j\in\mathbb{Z}} a_j e^{ij\xi} = G(\xi)U^0_\nu,$$

sodass $C(\lambda,\Delta t)^\mu U^0 = G(\xi)^\mu U^0$ und $\|C(\lambda,\Delta t)^\mu U^0\|_\infty = \|G(\xi)^\mu U^0\|_\infty = |G(\xi)|^\mu \|U^0\|_\infty = \|G^\mu\|_\infty \|U^0\|_\infty$. Dies zeigt die Behauptung. ∎

Korollar 5.10.3 *Wenn das Differenzenverfahren (5.5.10) bezüglich ℓ^2 instabil ist, so auch bezüglich ℓ^∞.*

Beweis. Nach Lemma 5.10.2 impliziert $\sup \|C(\lambda,\Delta t)^\mu\|_{\ell^2\leftarrow\ell^2} = \infty$ auch $\sup \|C(\lambda,\Delta t)^\mu\|_{\ell^\infty\leftarrow\ell^\infty} = \infty$. ∎

Die Beispiele aus §5.5.3 werden erneut auf Stabilität bzw. Instabilität untersucht.

Beispiel 5.10.4 *a)* ***Verfahren (5.5.8a)*** *mit $a_0 = 1 - a\lambda$, $a_1 = a\lambda$. Für λ mit $0 \le a\lambda \le 1$ wurde bereits die Stabilität in Beispiel 5.8.8a bestätigt. Die zugehörige charakteristische Funktion*

$$G(\xi) := 1 - a\lambda + a\lambda e^{-i\xi} = 1 - a\lambda(1-\cos\xi) - ia\lambda\sin\xi$$

hat den Betrag $|G(\pi)| = |1 - 2a\lambda|$ bei $\xi = \pi$. Da $|G(\pi)| > 1$ für alle $a\lambda \notin [0,1]$, beweisen Satz 5.10.1 und Korollar 5.10.3 in diesem Bereich die Instabilität bezüglich ℓ^2 und ℓ^∞. Das Verfahren (5.5.8a) ist daher nur bedingt *stabil mit der Einschränkung $a\lambda \in [0,1]$.*

b) ***Verfahren (5.5.8b)*** *mit $a_{-1} = -\frac{a\lambda}{2}$, $a_0 = 1$, $a_1 = \frac{a\lambda}{2}$. Die zugehörige charakteristische Funktion $G(\xi) := -\frac{a\lambda}{2}e^{i\xi} + 1 + \frac{a\lambda}{2}e^{-i\xi} = 1 - ia\lambda\sin\xi$ hat die Maximumnorm $\|G\|_\infty = \sqrt{1+|a\lambda|^2}$ und ist daher bis auf den trivialen Fall $a = 0$ immer instabil (bezüglich ℓ^2 und ℓ^∞).*

c) ***Verfahren (5.5.8c)*** *mit $a_{-1} = \frac{1-a\lambda}{2}$, $a_1 = \frac{1+a\lambda}{2}$. Für $|a\lambda| \le 1$ bestätigte Beispiel 5.8.8c bereits die Stabilität. Wegen $G(\xi) := \frac{1-a\lambda}{2}e^{i\xi} + \frac{1+a\lambda}{2}e^{-i\xi}$ und $|G(\xi)|^2 = \cos^2\xi + |a\lambda|^2\sin^2\xi$ folgt $\|G\|_\infty = \max\{1, |a\lambda|^2\}$. Somit ist das Verfahren bedingt stabil für $|a\lambda| \le 1$, aber instabil für $|a\lambda| > 1$ (bezüglich ℓ^2 und ℓ^∞).*

d) ***Verfahren (5.5.9)*** *mit $a_{-1} = \lambda$, $a_0 = 1 - 2\lambda$, $a_1 = \lambda$. Für $\lambda \in (0, 1/2]$ ist die Stabilität in Beispiel 5.8.8d gezeigt. Wegen $G(\xi) := \lambda e^{i\xi} + 1 - 2\lambda + \lambda e^{-i\xi} = 1 - 2\lambda(1-\cos\xi)$ und $\|G\|_\infty = |G(\pi)| = |1 - 4\lambda|$ ist das Verfahren bedingt stabil für $\lambda \in (0, 1/2]$, aber instabil für $\lambda > 1/2$ (bezüglich ℓ^2 und ℓ^∞).*

5.11 CFL-Bedingung

Eine Bedingung sehr spezieller Art ist die CFL-Bedingung, wobei das Kürzel die Namen Courant-Friedrichs-Lewy bezeichnet[60]. Es ist im eigentlichen Sinne ein Kriterium für Nicht-Konvergenz.

[60] Courant-Friedrichs-Lewy: *Über die partiellen Differenzengleichungen der mathematischen Physik.* Math. Ann. 100 (1928) 32-74

Kriterium 5.11.1 (Courant-Friedrichs-Lewy) *Zu der hyperbolischen Differentialgleichung* $\frac{\partial}{\partial t}u = a\frac{\partial}{dx}u$ *gehöre das explizite Differenzenschema (5.5.7), wobei sich die Summe* $\sum_{j\in\mathbb{Z}} a_j U^\mu_{\nu+j}$ *auf die endliche Summe* $\sum_{J_1\le j\le J_2} a_j U^\mu_{\nu+j}$ *reduziere. Falls* $a\lambda \notin [J_1, J_2]$ *ist das Verfahren nicht konvergent. Damit ist* $a\lambda \in [J_1, J_2]$ *eine notwendige Konvergenzbedingung.*

Wegen des Äquivalenzsatzes kann für $a\lambda \notin [J_1, J_2]$ auch keine Stabilität vorliegen, weshalb im Allgemeinen von dem CFL-*Stabilitäts*kriterium gesprochen wird.
Beweis des Kriteriums 5.11.1. Zu einem beliebigen Raum-Zeit-Punkt (t, x) mit $t > 0$ lautet die Lösung gemäß Lemma 5.2.1 $u(t, x) = u_0(x + at)$. Für das Differenzenverfahren stellt man fest, dass in einem Gitterpunkt $(t, x) = (\mu\Delta t, \nu\Delta x)$ die Lösung U^μ_ν ausschließlich von den Anfangswerten U^0_k mit $k \in [\nu + \mu J_1, \nu + \mu J_2]$ abhängt. Multiplikation mit Δx zeigt

$$x_k = k\Delta x \in [x + \mu\Delta x J_1, x + \mu\Delta x J_2] = [x + \mu\Delta x J_1, x + \mu\Delta x J_2] = [x + tJ_1/\lambda, x + tJ_2/\lambda].$$

Falls $a\lambda \notin [J_1, J_2]$ folgt $x + at \notin [x + tJ_1/\lambda, x + tJ_2/\lambda]$. Damit benutzt die Berechnung von U^μ_ν nicht die Daten, von denen die Lösung $u(t, x)$ einzig abhängt. Folglich kann U^μ_ν nicht gegen $u(t, x)$ konvergieren. ∎

Das Verfahren (5.5.8a) sei als Beispiel genommen. Da $a_0 = 1 - a\lambda$, $a_1 = a\lambda$, lauten die Schranken $J_1 = 0$, $J_2 = 1$. Die notwendige CFL-Bedingung $a\lambda \in [0, 1]$ stimmt gemäß Beispiel 5.10.4a mit der exakten Stabilitätseigenschaft überein.

5.12 Implizite Verfahren

Bisher waren alle Verfahren bestenfalls bedingt stabil. Um *unbedingt stabile* Verfahren zu erreichen, muss man implizite Differenzenverfahren zulassen.[61]

Im parabolischen Fall $\frac{\partial u}{\partial t} = \frac{\partial^2 u}{dx^2}$ kann die zweite x-Differenz auf der Zeitstufe $t + \Delta t$ gebildet werden (d.h. $\frac{\partial^2 u}{dx^2} \approx \frac{u(t+\Delta t,x-\Delta x)-2u(t+\Delta t,x)+u(t+\Delta t,x+\Delta x)}{\Delta x^2}$), während die Zeitableitung $\frac{\partial}{\partial t}u$ wie bisher durch $\frac{u(t+\Delta t,x)-u(t,x)}{\Delta t}$ diskretisiert. Dies führt auf

$$-\lambda U^{\mu+1}_{\nu-1} + (1 + 2\lambda)\, U^{\mu+1}_\nu - \lambda U^{\mu+1}_{\nu+1} = U^\mu_\nu. \tag{5.12.1}$$

Anstelle der expliziten Form $U^{\mu+1} = C(\lambda, \Delta t)U^\mu$ erhält man jetzt ein implizites Verfahren der Form

$$C_1(\lambda, \Delta t)U^{\mu+1} = C_2(\lambda, \Delta t)U^\mu, \tag{5.12.2}$$

wobei im vorliegenden Falle $(C_1(\lambda, \Delta t)U)_\nu = \sum_{j\in\mathbb{Z}} a_{1,j}U_{\nu+j}$ mit $a_{1,-1} = a_{1,1} = -\lambda$, $a_{1,0} = 1 + 2\lambda$ und $C_2(\lambda, \Delta t)U = \sum_{j\in\mathbb{Z}} a_{2,j}U_{\nu+j}$ mit $a_{2,0} = 1$ (sonstige Koeffizienten $= 0$). Man möchte nach $U^{\mu+1}$ auflösen:

$$U^{\mu+1} = C(\lambda, \Delta t)U^\mu \qquad \text{mit } C(\lambda, \Delta t) := [C_1(\lambda, \Delta t)]^{-1}\, C_2(\lambda, \Delta t). \tag{5.12.3}$$

Hierzu ist die Existenz der Inversen $[C_1(\lambda, \Delta t)]^{-1}$ zu klären.

Lemma 5.12.1 *a) Wenn die Koeffizienten* $a_{1,j}$ *von* $C_1(\lambda, \Delta t)$ *einer Ungleichung*

$$\Big(\sum_{j\in\mathbb{Z}\setminus\{0\}} |a_{1,j}|\Big) / |a_{1,0}| \le 1 - \varepsilon < 1$$

genügen, existiert die Inverse und erfüllt $\| [C_1(\lambda, \Delta t)]^{-1} \|_{\ell^p\leftarrow\ell^p} \le 1/(\varepsilon\,|a_{1,0}|)$.
b) Die Inverse $[C_1(\lambda, \Delta t)]^{-1}$ *existiert in* $L(\ell^2, \ell^2)$ *genau dann, wenn die zu* C_1 *gehörende charakteristische Funktion* $G_1(\xi)$ *eine Ungleichung* $|G_1(\xi)| \ge \varepsilon > 0$ *erfüllt. Die Norm ist*

$$\| [C_1(\lambda, \Delta t)]^{-1} \|_{\ell^2\leftarrow\ell^2} = 1/\inf_{\xi\in\mathbb{R}} |G_1(\xi)|.$$

[61] Die CFL-Bedingung ist im impliziten Fall nicht anwendbar, da ein implizites Verfahren formal als explizites mit unendlicher Summe (d.h. $J_1 = -\infty$, $J_2 = \infty$) geschrieben werden kann. Da dann stets $a\lambda \in [J_1, J_2] = \mathbb{R}$ gilt, ist die CFL-Bedingung immer erfüllt.

Beweis. a) Die Gleichung $C_1 V = U$ ist äquivalent zur Fixpunktgleichung $V = \frac{1}{a_{1,0}}(U - C_1' V) =: \Phi(V)$ mit $C_1' := C_1 - a_{1,0} I$. Die Kontraktionskonstante von Φ ist $\|C_1'\|_{\ell^p \leftarrow \ell^p} / |a_{1,0}|$. Nach Bemerkung 5.8.2 ist $\|C_1'\|_{\ell^p \leftarrow \ell^p} / |a_{1,0}| \le \left(\sum_{j \in \mathbb{Z} \setminus \{0\}} |a_{1,j}|\right) / |a_{1,0}| \le 1 - \varepsilon$, so dass eine eindeutige Inverse existiert. Die Abschätzung $\|V\|_{\ell^p} \le \frac{1}{|a_{1,0}|} \|U\|_{\ell^p} + (1-\varepsilon) \|V\|_{\ell^p}$ zeigt $\varepsilon \|V\|_{\ell^p} \le \frac{1}{|a_{1,0}|} \|U\|_{\ell^p}$, sodass die Schranke für $\| [C_1(\lambda, \Delta t)]^{-1} \|_{\ell^p \leftarrow \ell^p}$ folgt.

b) Sei $\hat{C}_1(\lambda, \Delta t)$ die Fourier-Transformierte $\mathcal{F}^{-1} C_1(\lambda, \Delta t) \mathcal{F}$ zu C_1. $\hat{C}_1(\lambda, \Delta t)$ ist der Operator $\hat{U} \mapsto G_1 \hat{U}$. Offenbar ist der Multiplikationsoperator $M \in L(\ell^2, \ell^2)$ mit $M\hat{U} := (1/G_1)\hat{U}$ die Inverse zu $\hat{C}_1$. Die Norm von M ist $\|M\|_{\ell^2 \leftarrow \ell^2} = \|\hat{C}_1^{-1}\|_{\ell^2 \leftarrow \ell^2} = \|1/G_1\|_\infty \le 1/\varepsilon$. Umgekehrt kann man nachprüfen, dass $\|\hat{C}_1^{-1}\|_{\ell^2 \leftarrow \ell^2}$ nicht endlich sein kann, wenn $\inf |G_1| = 0$. Da die Fourier-Transformation die Norm nicht ändert, folgt $\|C_1^{-1}\|_{\ell^2 \leftarrow \ell^2} = \|\hat{C}_1^{-1}\|_{\ell^2 \leftarrow \ell^2} = \|1/G_1\|_\infty = 1/\inf_{\xi \in \mathbb{R}} |G_1(\xi)|$. ∎

Beispiel 5.12.2 *Das Verfahren (5.12.1) ist für alle $\lambda = \Delta t / \Delta x^2$ stabil bezüglich ℓ^2, d.h. unbedingt stabil.*

Beweis. Zu C_1 gehört $G_1(\xi) = -\lambda e^{i\xi} + 1 + 2\lambda - \lambda e^{-i\xi} = 1 + 2\lambda(1 - \cos x) \ge \inf_{\xi \in \mathbb{R}} |G_1(\xi)| = 1$. Da $C_2 = I$, ist $C(\lambda, \Delta t) := [C_1(\lambda, \Delta t)]^{-1} C_2(\lambda, \Delta t) = [C_1(\lambda, \Delta t)]^{-1}$ und $\|C_1^{-1}\|_{\ell^2 \leftarrow \ell^2} = 1/\inf_{\xi \in \mathbb{R}} |G_1(\xi)| = 1$. Die Stabilität folgt nach Kriterium 5.8.1. ∎

Den allgemeinen Fall eines impliziten Verfahrens (5.12.2) behandelt

Kriterium 5.12.3 *Das Verfahren (5.12.2) ist genau dann ℓ^2-stabil, wenn die charakteristische Funktion $G(\xi) := G_2(\xi)/G_1(\xi)$ die Bedingung (5.10.2) erfüllt.*

Beispiel 5.12.4 *Eine Modifikation von (5.12.1) ist das sogenannte Theta-Verfahren*

$$-\lambda\Theta U_{\nu-1}^{\mu+1} + (1 + 2\lambda\Theta) U_\nu^{\mu+1} - \lambda\Theta U_{\nu+1}^{\mu+1} = \lambda(1-\Theta) U_{\nu-1}^\mu + (1 - 2\lambda(1-\Theta)) U_\nu^\mu + \lambda(1-\Theta) U_{\nu+1}^\mu \quad (5.12.4)$$

für ein $\Theta \in [0,1]$. $\Theta = 0$ liefert (5.5.9) und $\Theta = 1$ liefert (5.12.1) zurück. Für $\Theta = 1/2$ ist (5.12.4) das Crank-Nicolson-Schema. Das Verfahren (5.12.4) ist für $\Theta \in [1/2, 1]$ unbedingt ℓ^2-stabil, während es im Falle von $\Theta \in [0, 1/2)$ bedingt ℓ^2-stabil für $\lambda \le 1/(2(1-2\Theta))$ ist. Im Stabilitätsfall gilt stets $\|C(\lambda, \Delta t)\|_{\ell^2 \leftarrow \ell^2} = 1$.

Beweis. Sei $(DU)_\nu = -U_{\nu-1} + 2U_\nu - U_{\nu+1}$ der Operator der zweiten (negativen) Differenz. Die charakteristische Funktion zu D ist $G_D(\xi) = 2 - 2\cos\xi = 4\sin^2(\xi/2)$. Die Operatoren C_1, C_2 aus (5.12.2) lauten im Falle von (5.12.4)

$$C_1(\lambda, \Delta t) = I + \lambda\Theta D \quad \text{und} \quad C_2(\lambda, \Delta t) = I - \lambda(1-\Theta) D.$$

Die zugehörigen Funktionen sind $G_1(\xi) = 1 + \lambda\Theta G_D(\xi)$ und $G_2(\xi) = 1 - \lambda(1-\Theta) G_D(\xi)$, sodass

$$G(\xi) = \frac{1 - \lambda(1-\Theta) G_D(\xi)}{1 + \lambda\Theta G_D(\xi)}.$$

Die Funktion $\frac{1-\lambda(1-\Theta)X}{1+\lambda\Theta X}$ ist bezüglich X monoton fallend, sodass die Randextrema bei $X = 0 = G_D(0)$ und $X = 4 = G_D(\pi)$ zu untersuchen sind:

$$\|G\|_\infty = \max\{G(0), -G(\pi)\} = \max\left\{1, \frac{4\lambda(1-\Theta) - 1}{1 + 4\lambda\Theta}\right\}.$$

Für $\Theta \in [1/2, 1]$ ist $-G(\pi) = \frac{4\lambda(1-\Theta)-1}{1+4\lambda\Theta} = \frac{4\lambda}{1+4\lambda\Theta} - 1 \underset{\Theta \ge 1/2}{\le} \frac{4\lambda}{1+2\lambda} - 1 = \frac{2\lambda - 1}{2\lambda + 1} \le 1$ und beweist $\|G\|_\infty = 1$.

Im Falle von $\Theta \in [0, 1/2)$ muss die Wahl von λ die Abschätzung $-G(\pi) \le 1$ gewährleisten. Äquivalent sind $4\lambda(1-\Theta) - 1 \le 1 + 4\lambda\Theta \quad \Leftrightarrow \quad 4\lambda(1-2\Theta) \le 2 \quad \Leftrightarrow \quad \lambda \le 1/(2(1-2\Theta))$. ∎

5.13 Vektorwertige Gitterfunktionen

Bisher war ℓ^p die Menge der komplexwertigen Folgen $(U_\nu)_{\nu \in \mathbb{Z}}$, $U_\nu \in \mathbb{C}$. Wenn die Gleichung $\frac{\partial}{\partial t} u(t) = Au(t)$ aus (5.1.1a) vektorwertig ist (Werte in $\mathbb{C}^N$), müssen auch die Gitterfunktionen $(U_\nu)_{\nu \in \mathbb{Z}}$ vektorwertig sein:

$$\ell^p = \{U = (U_\nu)_{\nu \in \mathbb{Z}} : U_\nu \in \mathbb{C}^N\} \text{ mit den Normen } \|U\|_2 = \sqrt{\sum_{\nu \in \mathbb{Z}} \|U_\nu\|_2^2}, \quad \|U\|_\infty = \sup_{\nu \in \mathbb{Z}} \|U_\nu\|_\infty,$$

wobei $\|U_\nu\|_p$ die Euklidische Norm in $\mathbb{C}^N$ $(p=2)$ bzw. die Maximumnorm in $\mathbb{C}^N$ ist (vgl. Bemerkung 5.5.7).

Im Differenzenverfahren (5.5.10) sind die Koeffizienten a_j dementsprechend $N \times N$-Matrizen. Die Aussagen zur Konsistenz, Konvergenz und Stabilität bleiben unverändert (nur die Normen sind anders zu interpretieren). Dagegen sind die Kriterien ab §5.8 auf den Fall $N > 1$ zu verallgemeinern.

Kriterium 5.8.1 ist unverändert gültig.

In Bemerkung 5.8.2 muss es nun $\|C(\lambda,\Delta t)\|_{\ell^p \leftarrow \ell^p} \le \sum_j \|a_j\|_p$ heißen, wobei $\|.\|_2$ die Spektralnorm und $\|.\|_\infty$ die Zeilensummennorm für $N \times N$-Matrizen sind.

In Korollar 5.8.3 ist $\sum |a_j|$ durch $\sum_j \|a_j\|_p$ zu ersetzen.

Wenn wir mittels der Fourier-Transformation $G(\xi) = \sum_{j\in\mathbb{Z}} a_j e^{-ij\xi}$ bilden, muss jetzt beachtet werden, dass $G(\xi)$ $N \times N$-Matrizen sind. Die Gleichheit (5.9.5) wird zu

$$\|C(\lambda,\Delta t)^\mu\|_{\ell^2 \leftarrow \ell^2} = \sup\{\|G(\xi)^\mu\|_2 : \xi \in [0,2\pi)\}. \tag{5.13.1}$$

Anstelle eines relativ abstrakten Operators $C(\lambda,\Delta t)$ hat man nun die Beschränktheit der $N \times N$-Matrizen $G(\xi)^\mu$ zu untersuchen.

Im Kriterium 5.8.11 wurde $\rho(C(\lambda,\Delta t)) \le 1 + \mathcal{O}(\Delta t)$ als eine notwendige Bedingung erkannt. Da die unitäre Fourier-Transformation die Spektren unverändert lässt, haben $C(\lambda,\Delta t)$ und $\{G(\xi) : \xi \in [0,2\pi)\}$ die gleichen Eigenwerte. Damit erhalten wir die von-Neumann-Bedingung.[62] Hierin bezeichnet $\rho(G(\xi))$ den größen Betrag $|\lambda_i|$ der Eigenwerte λ_i von $G(\xi)$. Der Teil b) entspricht der Bemerkung 5.8.12.

Kriterium 5.13.1 (von-Neumann-Bedingung) *a) Eine notwendige Bedingung für Stabilität ist*

$$\sup\{\rho(G(\xi)) : \xi \in [0,2\pi)\} \le 1 + \mathcal{O}(\Delta t).$$

b) Falls alle Matrizen $G(\xi)$ normal sind, ist diese Bedingung auch hinreichend.

Die folgende Aussage benutzt den *numerischen Radius* $r(A) := \sup\limits_{0\neq v\in\mathbb{C}^N} \left|\frac{\langle Av,v\rangle}{\|v\|_2^2}\right|$ einer Matrix, wobei $\langle\cdot,\cdot\rangle$ das Euklidische Skalarprodukt in $\mathbb{C}^N$ bezeichnet.

Lemma 5.13.2 (Lax-Wendroff-Bedingung) *Es existiere eine Konstante K_{LW}, sodass für alle $\xi \in [0,1]$ und alle Vektoren $v \in \mathbb{C}^N$*

$$|\langle G(\xi)v, v\rangle| \le (1 + K_{LW}\Delta t)\,\|v\|_2^2$$

gelte. Dann liegt ℓ^2-Stabilität vor.[63]

Beweis. Der numerische Radius besitzt für allgemeine Matrizen A die Eigenschaften[64]

$$\|A\|_2 \le 2r(A), \qquad r(A^n) \le r(A)^n \quad \text{für } n \in \mathbb{N}_0.$$

Damit gilt $\|G(\xi)^n\|_2 \le 2r(G(\xi)^n) \le 2r(G(\xi))^n \le 2\,(1 + K_{LW}\Delta t)^n \le 2\exp(K_{LW}T)$ für alle n mit $n\Delta t \le T$ und alle $\xi \in [0,2\pi)$. Wegen (5.13.1) ist die Behauptung bewiesen. ∎

In der Definition 5.8.4 der positiven Differenzenverfahren ist $a_j \ge 0$ durch "a_j positiv semidefinit" zu ersetzen.

Übungsaufgabe 5.13.3 *Die positiv semidefiniten Koeffizienten a_j seien (i) sämtlich diagonal oder (ii) simultan diagonalisierbar, d.h. es existiert eine Transformation S, sodass die Matrizen $d_j = Sa_jS^{-1}$ für alle j diagonal sind. Man zeige, dass analog zu Kriterium 5.8.5 das Differenzenverfahren (5.5.10) stabil ist, wenn $\sum a_j = I$ (Fall (i)) bzw. $\sum d_j = I$ (Fall (ii)).*

Aber auch ohne simultane Diagonalisierbarkeit lässt sich das Kriterium verallgemeinern, wobei sogar x-abhängige Koeffizienten $a_j = a_j(x)$ zugelassen sind, wie das Kriterium von Friedrichs[65] zeigt.

[62] János von Neumann, geb. 28. Dez. 1903 in Budapest, gest. 8. Febr. 1957 in Washington D.C.

[63] P.D. Lax und B. Wendroff: *Difference schemes for hyperbolic equations with high order of accuracy.* Comm. Pure Appl. Math., 17 (1964), 381-398

[64] Vgl. (2.9.11d,h) in: W. Hackbusch: *Iterative Lösung großer schwachbesetzter Gleichungssysteme.* 2. Auflage, Teubner-Verlag, Stuttgart, 1993.

[65] Kurt Otto Friedrichs, geb. 28. Sept. 1901 in Kiel, gest. 31.Dez. 1982 in New Rochelle (N.Y., USA)

Kriterium 5.13.4 (Friedrichs)[66] *Das Differenzenverfahren (5.5.10) habe positiv semidefinite Koeffizienten a_j mit der Konsistensbedingung $\sum_{j\in\mathbb{Z}} a_j = I$ (vgl. (5.8.1)). Die Koeffizienten a_j seien entweder konstant oder es gelten die drei Bedingungen: (i) der hyperbolische Fall mit $\lambda = \Delta t/\Delta x$ liege vor (vgl. (5.5.2)), (ii) die Koeffizienten $a_j(\cdot)$ seien global Lipschitz-stetig in $\mathbb{R}$ zur Lipschitz-Konstanten L_j, (iii) $B := \sum_{j\in\mathbb{Z}} jL_j < \infty$. Dann gilt $\|C(\lambda,\Delta t)\|_{\ell^2\leftarrow\ell^2} \le 1 + C_L\Delta t$ mit $C_L = B/(2\lambda)$.*

Beweis. Wir beweisen den allgemeinen Fall (konstante Koeffizienten entsprechen $L_j = 0$). In der Darstellung

$$(C(\lambda,\Delta t)V,U)_{\ell^2} = \Delta x \sum_{j\in\mathbb{Z}}\sum_{\nu\in\mathbb{Z}} \langle a_j(\nu\Delta x)V_{\nu+j}, U_\nu\rangle \tag{5.13.2}$$

bezeichnet $\langle\cdot,\cdot\rangle$ das Skalarprodukt zu $\mathbb{C}^N$ (N ist die Dimension von $U_\nu, V_\nu \in \mathbb{C}^N$).

Für jede positiv semidefinite Matrix M gilt $\langle Mx,y\rangle \le \frac{1}{2}\langle Mx,x\rangle + \frac{1}{2}\langle My,y\rangle$ für alle $x,y\in\mathbb{C}^N$. Anwendung auf (5.13.2) liefert

$$\left|\Delta x \sum_{j\in\mathbb{Z}}\sum_{\nu\in\mathbb{Z}} \langle a_j(\nu\Delta x)V_{\nu+j},U_\nu\rangle\right| \le \frac{\Delta x}{2}\sum_{\nu\in\mathbb{Z}}\sum_{j\in\mathbb{Z}}\langle a_j(\nu\Delta x)U_\nu,U_\nu\rangle + \frac{\Delta x}{2}\sum_{\nu\in\mathbb{Z}}\sum_{j\in\mathbb{Z}}\langle a_j(\nu\Delta x)V_{\nu+j},V_{\nu+j}\rangle\,.$$

Der erste Summand ist $\frac{\Delta x}{2}\sum_{\nu\in\mathbb{Z}}\left\langle\sum_{j\in\mathbb{Z}} a_j(\nu\Delta x)U_\nu, U_\nu\right\rangle \underset{(5.8.1)}{=} \frac{\Delta x}{2}\sum_{\nu\in\mathbb{Z}}\|U_\nu\|_2^2 = \frac{1}{2}\|U\|_{\ell^2}^2$. Im zweiten Summanden wird $\mu = \nu + j$ substituiert:

$$\frac{\Delta x}{2}\sum_{\nu\in\mathbb{Z}}\sum_{j\in\mathbb{Z}}\langle a_j(\nu\Delta x)V_{\nu+j},V_{\nu+j}\rangle = \frac{\Delta x}{2}\sum_{\mu\in\mathbb{Z}}\sum_{j\in\mathbb{Z}}\langle a_j((\mu-j)\,\Delta x)V_\mu,V_\mu\rangle\,.$$

Wegen der Lipschitz-Stetigkeit ist $\|a_j((\mu-j)\,\Delta x) - a_j(\mu\Delta x)\|_2 \le L_j j\Delta x$. Daher gilt $\sum_{j\in\mathbb{Z}} jL_j\Delta x = B\Delta x$ und

$$\frac{\Delta x}{2}\sum_{\mu\in\mathbb{Z}}\sum_{j\in\mathbb{Z}}\langle a_j(\mu\Delta x)V_\mu,V_\mu\rangle \le \frac{B(\Delta x)^2}{2}\sum_{\mu\in\mathbb{Z}}\|V_\mu\|_2^2 + \frac{\Delta x}{2}\sum_{\nu\in\mathbb{Z}}\left\langle\Big[\sum_{j\in\mathbb{Z}} a_j(\mu\Delta x)\Big]V_\mu, V_\mu\right\rangle \underset{(5.8.1)}{=} \frac{1+B\Delta x}{2}\|V\|_{\ell^2}^2\,.$$

Zusammen ergibt sich

$$|(CV,U)_{\ell^2}| \le \frac{1}{2}\|U\|_{\ell^2}^2 + \frac{1+B\Delta x}{2}\|V\|_{\ell^2}^2\,.$$

Wegen $\|C(\lambda,\Delta t)\|_{\ell^2\leftarrow\ell^2} = \sup_{\|U\|_{\ell^2}=\|V\|_{\ell^2}=1} |(CV,U)_{\ell^2}|$ folgt $\|C(\lambda,\Delta t)\|_{\ell^2\leftarrow\ell^2} \le 1 + \frac{B\Delta x}{2} \underset{\Delta x=\Delta t/\lambda}{=} 1 + C_L\Delta t$. ■

Zu diesem Kriterium sei noch Folgendes angemerkt.
1) Im parabolischen Fall mit $\lambda = \Delta t/\Delta x^2$ lässt sich zwar $\|C(\lambda,\Delta t)\|_{\ell^2\leftarrow\ell^2} \le 1 + \mathcal{O}(\Delta x) = 1 + \mathcal{O}(\sqrt{\Delta t})$ beweisen, aber dies reicht nicht zur Stabilität.
2) Wenn die Koeffizienten a_j konstant sind, folgt $L_j = 0$ und daher $\|C(\lambda,\Delta t)\|_{\ell^2\leftarrow\ell^2} = 1$.
3) Wenn nur endlich viele Koeffizienten von null verschieden sind, ist wie gefordert $B = \sum_{j\in\mathbb{Z}} jL_j < \infty$.

Ein positives Differenzenverfahren erhält man aus dem *symmetrischen hyperbolischen Gleichungssystem*

$$\frac{\partial}{\partial t}u + A(x)\frac{\partial}{\partial x}u = 0 \qquad (A(x) \text{ symmetrische } N\times N\text{-Matrix, } u\in\mathbb{C}^N), \tag{5.13.3}$$

wenn man $\frac{\partial}{dx}u$ durch $[u(t,x+r\Delta x) - u(t,x-r\Delta x)]\,/\,(2r\Delta x)$ und $\frac{\partial}{dt}u$ durch $[u(t+\Delta t,x) - \bar u(t,x)]\,/\Delta t$ mit $\bar u(t,x) = \frac{1}{2}[u(t,x+r\Delta x) - u(t,x-r\Delta x)]$ ersetzt. Es entsteht das Differenzenverfahren

$$(C(\lambda,\Delta t)U)_\nu = \frac{\lambda}{2}\left\{\left[I - \frac{1}{r}A(\nu\Delta x)\right]U_{\nu-r} + \left[I + \frac{1}{r}A(\nu\Delta x)\right]U_{\nu+r}\right\}.$$

Wählt man $r\in\mathbb{N}$ mit $r \ge \sup_{x\in\mathbb{R}}\|A(x)\|_2$, ist $C(\lambda,\Delta t)$ ein positives Schema, sodass Stabilität folgt, sobald $A(x)$ Lipschitz-stetig ist.

[66] K.O. Friedrichs: *Symmetric hyperbolic linear differential equations.* Comm. Pure and Appl. Math., **7** (1954) 345-392

Kriterium 5.8.11 wie auch Bemerkung 5.8.12 und Kriterium 5.8.13 bleiben auch im vektorwertigen Fall gültig.

Der Fourier-transformierte Differenzenoperator lautet auch im vektorwertigen Fall $\hat{C}(\lambda,\Delta t)(x) = \sum_{j\in\mathbb{Z}} a_j e^{-ijx}$, wobei nun $\hat{C}$ eine 2π-periodische Funktion ist, deren Werte $N\times N$-Matrizen sind. Wie zuvor gilt die Stabilitätsabschätzung $\|C(\lambda,\Delta t)^\mu\|_{\ell^2\leftarrow\ell^2} \le K(\lambda)$ für alle $\mu\Delta t \le T$ genau dann, wenn $\|\hat{C}(\lambda,\Delta t)^\mu\|_{L^2(0,2\pi)\leftarrow L^2(0,2\pi)} \le K(\lambda)$. Allerdings ist die Gleichheit $\left\|\left(\sum_{j\in\mathbb{Z}} a_j e^{-ijx}\right)^\mu\right\|_\infty = \left\|\sum_{j\in\mathbb{Z}} a_j e^{-ijx}\right\|_\infty^\mu$ aus (5.10.1) für den matrizenwertigen Fall im Allgemeinen nicht mehr gültig.

5.14 Verallgemeinerungen

Der Abschnitt 5.13 hat bereits die Verallgemeinerung auf den Fall von Differentialgleichungssystemen mit vektorwertigen Lösungen analysiert. Im Folgenden wird eine Reihe weiterer Verallgemeinerungen diskutiert.

5.14.1 Der Fall mehrerer Ortvariablen

Anstelle einer Ortsvariablen können d Variablen $x = (x_1,\dots,x_d)$ zugelassen werden ($d=3$ ist naheliegend). Die Wärmeleitungsgleichung (5.3.1) wird dann zu $\frac{\partial u}{\partial t} = \Delta u$ für $t>0$ mit dem Laplace-Operator $\Delta = \sum_{k=1}^d \frac{\partial^2}{dx_k^2}$. Die Darstellung (5.3.2) von u lässt sich der d-dimensionalen Situation anpassen.

Die hyperbolische Differentialgleichung $\frac{\partial}{\partial t}u = a\frac{\partial}{dx}u$ ist sofort verallgemeinerungsfähig zu

$$\frac{\partial}{\partial t}u = \sum_{k=1}^d A_k \frac{\partial u}{dx_k}. \tag{5.14.1}$$

Bei der Diskretisierung ist das Gitter $G_{\Delta x} = \{x = \nu\Delta x : \nu\in\mathbb{Z}\}$ aus (5.5.1) durch das d-dimensionale Gitter $G_{\Delta x} = \{x = \nu\Delta x : \nu\in\mathbb{Z}^d\}$ mit Multiindizes $\nu = (\nu_1,\dots,\nu_d)$, $\nu_j\in\mathbb{Z}$, zu ersetzen[67]. Die weiterhin mit ℓ^p bezeichneten Banach-Räume enthalten die Elemente von $\mathbb{C}^{\mathbb{Z}^d}$ mit endlichen Normen $\|U\|_{\ell^\infty} = \sup\{|U_\nu| : \nu\in\mathbb{Z}^d\}$ bzw. $\|U\|_{\ell^2} = \sqrt{\Delta x^d \sum_{\nu\in\mathbb{Z}^d}|U_\nu|^2}$. Die Fourier-Transformierte von $U\in\ell^2$ ist nun $\hat{U}(x) = \frac{1}{(2\pi)^{d/2}}\sum_{\nu\in\mathbb{Z}^d} U_\nu e^{i\nu x}$, wobei $\nu x = \langle\nu,x\rangle = \sum_{k=1}^d \nu_j x_j$ das Euklidische Skalarprodukt bezeichnet.

Bei der Stabilitätsuntersuchung des Systems (5.14.1) mit $N\times N$-Matrizen A_k ergibt sich die folgende Komplikation: Im univariaten Fall $d=1$ sind alle Koeffizientenmatrizen a_j von $C(\lambda,\Delta t)$ nur von einer Matrix A_1 abgeleitet, sodass die a_j üblicherweise paarweise vertauschbar (und damit simultan diagonalisierbar) sind. Für $d>1$ mit nichtvertauschbaren Matrizen A_k in (5.14.1) sind auch die a_j nicht simultan diagonalisierbar.

5.14.2 Der Fall zeitabhängiger Koeffizienten

Die Koeffizienten in der Differentialgleichung (5.1.2) oder (5.14.1) können von t abhängen: $a = a(t)$ bzw. $A_k = A_k(t)$. Das Konzept der Halbgruppe der Lösungsoperatoren (vgl. §5.4) wird in diesem Falle etwas modifiziert. An die Stelle von $T(t)$ tritt der Operator $T(t_1,t_0)$ mit $0\le t_0\le t_1\le T$, der die Überführung eines Anfangswertes zur Zeit t_0 in die Lösung zur Zeit t_1 beschreibt. Die Halbgruppeneigenschaft lautet $T(t_2,t_1)T(t_1,t_0) = T(t_2,t_0)$ und $T(t,t) = I$.

Die Diskretisierung liefert zeitabhängige Differenzenschemata $C(t;\lambda,\Delta t)$ mit $U^\mu = C(\mu\Delta t;\lambda,\Delta t)U^{\mu-1}$,

$$\left(C(t;\lambda,\Delta t)U\right)_\nu = \sum\nolimits_{j\in\mathbb{Z}^d} a_j(t)U_{\nu+j} \qquad \text{für } U\in\ell^p \text{ und } \nu\in\mathbb{Z}^d.$$

Die Stabilitätsdefinition (5.6.1) ist dadurch zu modifizieren, dass $C(\lambda,\Delta t)^\mu$ durch alle Produkte

$$C(t_0+\mu\Delta t;\lambda,\Delta t)C(t_0+(\mu-1)\,\Delta t;\lambda,\Delta t)\cdots C(t_0+\Delta t;\lambda,\Delta t) \quad \text{mit } 0\le t_0\le t_0+\mu\Delta t\le T$$

zu ersetzen ist.

Die Kriterien 5.8.1, 5.8.11 und Lemma 5.8.9 gelten auch im zeitabhängigen Fall.

[67] Im Prinzip sind unterschiedliche Schrittweiten Δx_j sinnvoll. Sie können aber durch die Transformation $x_j \mapsto \frac{\Delta x_1}{\Delta x_j}x_j$ der Ortsvariablen auf eine gemeinsame Ortsschrittweite Δx vereinheitlicht werden.

5.14.3 Der Fall ortsabhängiger Koeffizienten

Die Differentialgleichung kann ortsabhängige Koeffizienten enthalten wie zum Beispiel in der Differentialgleichung (5.13.3). Entsprechend sind dann die Koeffizienten $a_j = a_j(x)$ von $C(\lambda, \Delta t)$ ortsabhängig. Das Kriterium 5.13.4 hat bereits diese Verallgemeinerung im Fall positiver Differenzenverfahren zugelassen.

Verschiedene Kriterien benutzen wieder die Funktion

$$G(x, \xi) := \sum_{j \in \mathbb{Z}} a_j(x) e^{-ij\xi}, \tag{5.14.2}$$

die formal aus (5.9.4) entsteht (es ist aber nicht die Fourier-Transformierte $\mathcal{F}^{-1}C(\lambda, \Delta t)\mathcal{F}$!). Ein Versuch besteht in der Untersuchung der Stabilitätseigenschaften von $G(x, \xi)$ für einen *eingefrorenen* Koeffizienten x_0 (dies entspricht dem Differenzenverfahren $C(x_0; \lambda, \Delta t)$ mit $a_j(x)$ ersetzt durch die konstanten Koeffizienten $a_j(x_0)$). Die Frage, wie die Stabilität des Schemas $C(\lambda, \Delta t)$ mit variablen Koeffizienten einerseits mit der Stabilität von $C(x_0; \lambda, \Delta t)$ für alle $x_0 \in \mathbb{R}$ andererseits zusammenhängt, ist zunächst negativ zu beantworten: Stabilität von $C(x_0; \lambda, \Delta t)$ für alle $x_0 \in \mathbb{R}$ ist im Allgemeinen weder hinreichend noch notwendig für die Stabilität von $C(\lambda, \Delta t)$.

Typische, hinreichende Kriterien[68] verlangen neben der Lipschitz-Stetigkeit der Koeffizienten $a_j(\cdot)$, dass das Schema dissipativ ist. Dabei heißt $C(\lambda, \Delta t)$ *dissipativ von der Ordnung* $2r$ ($r \in \mathbb{N}$), wenn Konstanten $\delta, \tau > 0$ existieren, sodass

$$|\lambda_\nu(x, \Delta t, \xi)| \leq 1 - \delta\,|\xi|^{2r} \quad \text{für alle } x \in \mathbb{R},\ \Delta t \in (0, \tau),\ |\xi| \leq \pi, \tag{5.14.3}$$

wobei λ_ν, $\nu = 1, \ldots, N$, die Eigenwerte von $G(x, \xi)$ sind.

Abschließend sei auf einen Zusammenhang zwischen der Bedingung (5.14.3) und der Stabilitätsdefinition 4.5.3, die für die Wurzeln von ψ fordert, dass λ_ν mit $|\lambda_\nu| = 1$ einfache Nullstellen sind, während ansonsten $|\lambda_\nu| < 1$ gelten muss. Im Falle von (5.14.3) müssen die Potenzen $\|G(x, \xi)^n\|$ gleichmäßig beschränkt sein, während im zweiten Fall nach Bemerkung 4.5.18 die Begleitmatrix $\|A^n\| \leq const$ erfüllen muss. Der Unterschied ist aber, dass im zweiten Falle nur endlich viele Eigenwerte existieren, sodass $\max|\lambda_\nu| < 1$ für alle Eigenwerte mit $|\lambda_\nu| \neq 1$ gilt. Im Falle von $|\lambda_\nu(x, \Delta t, \xi)|$ sind die λ_ν stetige Funktionen von ξ und können betraglich gegen 1 streben. Die Bedingung (5.14.3) beschreibt quantitativ, wie schnell sich die $\lambda_\nu(x, \Delta t, \xi)$ 1 nähern dürfen.

5.15 Dissipativität für parabolische Diskretisierungen

Abschließend sei die Dissipativität (5.14.3) für Diskretisierungen der Wärmeleitungsgleichung (5.3.1) diskutiert. Da der Lösungsoperator hochfrequente Anteile stark dämpft, kommt es zu der glättenden Wirkung von (5.3.2): Anfangswerte u_0, die nur als stetig (oder aus L^∞) angenommen werden, führen zu Lösungen $u(t)$, die für alle $t > 0$ unendlich oft differenzierbar sind. Eine entsprechende Bedingung für die diskreten Schemata ist die Bedingung (5.14.3), die hier

$$|G(\xi)| \leq 1 - \delta\,|\xi|^{2r} \qquad \text{für alle } |\xi| \leq \pi \tag{5.15.1}$$

lautet, da G nicht von x abhängt und 1×1-Matrizen mit ihrem einzigen Eigenwert übereinstimmen.

Übungsaufgabe 5.15.1 *a) Das einfachste Schema (5.5.9) erfüllt (5.15.1) mit den Parametern $r = 1$ und $\delta = \min\{4\lambda, 2\,(1 - 2\lambda)\}/\pi^2$ für $0 < \lambda < 1/2$. Für $\lambda = 1/2$ liegt keine Dissipativität vor.*

b) Das Crank-Nicolson-Schema (das ist (5.12.4) mit $\Theta = 1/2$) ist dissipativ für alle $\lambda > 0$ mit $r = 1$. Für $\lambda \to \infty$ bleibt die Stabilitätseigenschaft gleichmäßig erhalten, aber die Dissipativität geht verloren (d.h. $\delta \to 0$).

[68] Man vergleiche Kapitel 5 in Richtmyer - Morton: *Difference methods for initial-value problems.* 2. Auflage, Interscience Publishers, New York, 1967

6 Stabilität bei elliptischen Diskretisierungen

Im vorangegangenen Kapitel wurden partiellen Differentialgleichungen von hyperbolischen und parabolischen Typ behandelt. Zur Vervollständigung folgt ein kurzes Kapitel zum Begriff der Stabilität bei elliptischen Differentialgleichungen.

6.1 Elliptische Differentialgleichungen

Der Prototyp aller elliptischen Differentialgleichungen ist die *Poisson-Gleichung*[69]

$$\begin{aligned} Lu := u_{xx} + u_{yy} &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{auf } \Gamma, \end{aligned} \tag{6.1.1}$$

wobei Ω ein beschränktes Gebiet des $\mathbb{R}^2$ sei und $\Gamma := \partial\Omega$ sein Rand sei. Die *Randbedingung* $u = 0$ auf Γ könnte auch durch $u = g$ ersetzt werden. Eine allgemeinere lineare Differentialgleichungen in m Variablen $x = (x_1, \ldots, x_d)$ ist

$$Lu = f \quad \text{mit} \quad L = \sum_{i,j=1}^{d} \frac{\partial}{\partial x_i} a_{ij}(x) \frac{\partial}{\partial x_j} + L_0, \tag{6.1.2}$$

wobei L_0 weitere erste und nullte Ableitungen enthalten darf. L heißt gleichmäßig *elliptisch*, wenn $\sum_{i,j=1}^{d} a_{ij}(x)\xi_i\xi_j \geq \delta \|\xi\|_2^2$ für alle $x \in \Omega$ und alle $\xi = (\xi_1, \ldots, \xi_d) \in \mathbb{R}^d$ mit einer positiven Konstanten δ.

6.2 Diskretisierung

Wie in §5.14.1 überzieht man Ω mit einem Gitter. Bei klassischen Differenzenverfahren ist dies ein kartesisches Gitter der Schrittweite h, bei Finite-Element-Diskretisierungen verwendet man allgemeinere Dreiecksgitter. Als Modellbeispiel wählen wir (6.1.1) mit dem Quadrat $\Omega = (0,1)^2$. Das Gitter ist dann

$$\Omega_h := \{(x,y) \in \Omega : x/h, y/h \in \mathbb{N}\} = \{(\nu h, \mu h) : \nu, \mu = 1, \ldots, N\},$$

wobei $N = \frac{1}{h} - 1$ als ganzzahlig angenommen ist. Die Menge der möglichen Schrittweiten ist daher

$$H := \{h = 1/(N+1) : N \in \mathbb{N}\}.$$

Allgemein sei $H \subset (0, \infty)$ mit $0 \in \bar{H}$, sodass Folgen $h_\nu \in H$ mit $h_\nu \to 0$ existieren.

Anstelle von $u(x,y)$ aus (6.1.1) sollen nur Näherungen von u an den Knotenstellen $(\nu h, \mu h) \in \Omega_h$ berechnet werden:

$$u_{\nu,\mu} \approx u(\nu h, \mu h).$$

In jedem Knotenpunkt $(\nu h, \mu h)$ wird die zweite x-Ableitung u_{xx} aus (6.1.1) durch die zweite dividierte Differenz $\frac{1}{h^2}[u_{\nu-1,\mu} - 2u_{\nu,\mu} + u_{\nu+1,\mu}]$ in dieser Richtung ersetzt. Entsprechend wird u_{yy} zu $\frac{1}{h^2}[u_{\nu,\mu-1} - 2u_{\nu,\mu} + u_{\nu,\mu+1}]$. Zusammen ergibt sich der sogenannte *Fünfpunktstern*

$$\frac{1}{h^2}[-4u_{\nu,\mu} + u_{\nu-1,\mu} + u_{\nu+1,\mu} + u_{\nu,\mu-1} + u_{\nu,\mu+1}] = f_{\nu,\mu} \qquad \text{für alle } 1 \leq \nu, \mu \leq N, \tag{6.2.1}$$

wobei $f_{\nu,\mu} := f(\nu h, \mu h)$ die Auswertung[70] der rechten Seite f von (6.1.1) ist. Für $\nu = 1$ enthält die Gleichung (6.2.1) auch den Wert $u_{\nu-1,\mu} = u_{0,\mu}$. Der Punkt $(0, \mu h)$ gehört nicht zu Ω_h, sondern zum Rand $\Gamma_h := \{(x,y) \in \Gamma : x/h, y/h \in \mathbb{Z}\}$. Da hier die Randbedingung $u = 0$ gilt, können alle Werte $u_{\nu',\mu'}$ in (6.2.1) mit $(\nu' h, \mu' h) \in \Gamma_h$ durch null ersetzt werden. Es bleibt ein System von N^2 linearen Gleichungen für die N^2 Unbekannten $\{u_{\nu,\mu} : (\nu h, \mu h) \in \Omega_h\}$:

$$\mathbf{L}_h \mathbf{u}_h = \mathbf{f}_h, \tag{6.2.2}$$

wobei $\mathbf{u}_h = (u_{\nu,\mu})_{(\nu h, \mu h) \in \Omega_h}$ und $\mathbf{f}_h = (f_{\nu,\mu})_{(\nu h, \mu h) \in \Omega_h}$.

Entsprechend kann man allgemeinere Differentialgleichungen wie z.B. (6.1.2) auch in komplizierten Gebieten mit Differenzen- oder Finite-Element-Verfahren diskretisieren.[71]

[69] Siméon Denis Poisson, geb. 21. Juni 1781 in Pithiviers, gest. 25. April 1840 in Sceaux

[70] Im Falle einer Finite-Element-Diskretisierung mit stückweise linearen Funktionen über einem regelmäßigen Dreiecksgitter ergibt sich die gleiche Matrix, nur die rechte Seite $\mathbf{f}_h$ besteht aus Integralmittelwerten $f_{\nu,\mu}$ anstelle der Punktauswertungen.

[71] Details in W. Hackbusch: *Theorie und Numerik elliptischer Differentialgleichungen.* 2. Aufl., Teubner, Stuttgart, 1996

Bemerkung 6.2.1 *Die Matrix $\mathbf{L}_h$ aus (6.2.2) und (6.2.1) hat die folgenden Eigenschaften:*

a) $\mathbf{L}_h$ *ist* schwachbesetzt*, insbesondere gibt es pro Zeile höchstens 5 von null verschiedene Matrixeinträge.*

b) $\mathbf{L}_h$ *ist symmetrisch.*

c) $-\mathbf{L}_h$ *hat die positive Diagonalelemente* $4/h^2$*, alle Außerdiagonalelemente sind* ≤ 0*.*

d) Die Summe der Matrixelemente jeder Zeile von $-\mathbf{L}_h$ *ist* ≥ 0*. Genauer gilt für die Zeilensumme: Falls* $2 \leq \nu, \mu \leq N-1$*, ist die Summe* 0*, in den Eckpunkten* $(\nu, \mu) \in \{(1,1), (1,N), (N,1), (N,N)\}$ *ist die Summe* $2/h^2$*, in den übrigen Punkten mit* ν *oder* μ *in* $\{1, N\}$ *ist die Summe* $1/h^2$*.*

e) Für eine konkrete Darstellung der Matrix $\mathbf{L}_h$ *muss eine Anordnung der Komponenten der Vektoren* $\mathbf{u}_h, \mathbf{f}_h$ *festgelegt werden. Eine mögliche Anordnung ist die lexikographische Sortierung der Indexpaare* (ν, μ) *:* $(1,1), (2,1), \ldots, (N,1),\ (1,2), \ldots, (N,2),\ \ldots,\ (1,N), \ldots, (N,N)$*. In diesem Falle hat* $\mathbf{L}_h$ *die Blockgestalt*

$$\mathbf{L}_h = \frac{1}{h^2} \begin{bmatrix} T & I & & & \\ I & T & I & & \\ & \ddots & \ddots & \ddots & \\ & & I & T & I \\ & & & I & T \end{bmatrix} \qquad \textit{mit} \quad T = \begin{bmatrix} -4 & 1 & & & \\ 1 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 1 \\ & & & 1 & -4 \end{bmatrix},$$

wobei alle Blöcke T, I *das Format* $N \times N$ *haben. Nicht gekennzeichnete Blöcke und Matrixeinträge sind null.*

Im Falle der Matrix $\mathbf{L}_h$ aus (6.2.1) (aber nicht für jede Diskretisierung elliptischer Differentialgleichungen) ist $\mathbf{L}_h$ zudem eine M-Matrix.

Definition 6.2.2 (M-Matrix) *Eine Matrix* $A \in \mathbb{R}^{n \times n}$ *heißt* M*-Matrix, falls a)* $A_{ii} > 0$ $(1 \leq i \leq n)$*, b)* $A_{ij} > 0$ $(1 \leq i \neq j \leq n)$*, c)* A *regulär und* A^{-1} *hat nur nichtnegative Matrixeinträge:* $\left(A^{-1}\right)_{ij} \geq 0$*.*

Bemerkung 6.2.1c zeigt bereits die Eigenschaften a), b) für $A = -\mathbf{L}_h$. Bemerkung 6.2.1d zusammen mit der Irreduzibilität von $\mathbf{L}_h$ beschreibt, dass $A = -\mathbf{L}_h$ *irreduzibel diagonaldominant* ist. Irreduzibel diagonaldominante Matrizen mit den Eigenschaften a), b) besitzen aber bereits die Eigenschaft c)[72], d.h. $-\mathbf{L}_h$ ist eine M-Matrix.

6.3 Konsistenz

Die Konsistenzbedingung muss sicherstellen, dass die Diskretisierungsmatrix $\mathbf{L}_h$ und der Differentialoperator L zusammenpassen. Hierzu sind zunächst die Urbild- und Bildräume von L und $\mathbf{L}_h$ einzuführen: X, Y, X_h, Y_h seien Banach-Räume, sodass

$$L : X \to Y, \qquad \mathbf{L}_h : X_h \to Y_h \qquad (h \in H)$$

stetige Abbildungen sind. Ferner seien

$$R_X^h : X \to X_h, \qquad R_Y^h : Y \to Y_h, \qquad \mathbf{f}_h = R_Y^h f \qquad (h \in H)$$

"Restriktionen" der Funktionenräume X, Y in die Räume X_h, Y_h der "Gitterfunktionen". Die Konsistenzbedingung lautet dann

$$\lim_{H \ni h \to 0} \| \left(\mathbf{L}_h R_X^h - R_Y^h L\right) u \|_{Y_h} = 0 \qquad \text{für alle } u \in X. \tag{6.3.1}$$

Bemerkung 6.3.1 *Die konkrete Wahl der Banach-Räume kann z.B. wie folgt aussehen:*

$$\begin{aligned} X &= \{u \in C^2(\bar{\Omega}) : u(x,y) = 0 \textit{ für } (x,y) \in \Gamma\}, \\ Y &= C(\bar{\Omega}), \\ X_h &= \mathbb{R}^{N^2}, \quad X_h \ni \mathbf{u}_h = (u_{\nu,\mu})_{(\nu h, \mu h) \in \Omega_h}, \\ Y_h &= \mathbb{R}^{N^2}, \quad Y_h \ni \mathbf{f}_h = (f_{\nu,\mu})_{(\nu h, \mu h) \in \Omega_h} \end{aligned}$$

[72] Satz 6.4.10b in W. Hackbusch: *Iterative Lösung großer schwachbesetzter Gleichungssysteme.* Teubner, Stuttgart, 1993

mit den Normen[73]

$$\begin{aligned}
\|u\|_X &= \max\{\|u\|_\infty, \|u_x\|_\infty, \|u_y\|_\infty, \|u_{xx}\|_\infty, \|u_{yy}\|_\infty\},\\
\|u\|_Y &= \|u\|_\infty := \max\{|u(x,y)| : (x,y) \in \bar{\Omega}\},\\
\|\mathbf{u}_h\|_{X_h} &= \max\{\|\mathbf{u}_h\|_\infty, \|\partial_x \mathbf{u}_h\|_\infty, \|\partial_y \mathbf{u}_h\|_\infty, \|\partial_{xx} \mathbf{u}_h\|_\infty, \|\partial_{yy} \mathbf{u}_h\|_\infty\},\\
\|\mathbf{f}_h\|_{X_h} &= \|\mathbf{f}_h\|_\infty := \max\{|u_{\nu,\mu}| : (\nu h, \mu h) \in \Omega_h\},
\end{aligned}$$

wobei $\partial_x, \partial_y, \partial_{xx}, \partial_{yy}$ die ersten und zweiten Differenzenquotienten sind. Die zugehörigen Restriktionen sind die punktweisen Beschränkungen:

$$\left(R_X^h u\right)_{\nu,\mu} := u\left(\nu h, \mu h\right), \qquad \left(R_Y^h f\right)_{\nu,\mu} := f\left(\nu h, \mu h\right).$$

Eine Komponente von $(\mathbf{L}_h R_X^h - R_Y^h L)u$ hat die Darstellung $[\partial_{xx} + \partial_{yy}]\, u(x,y) - [u_{xx}(x,y) + u_{yy}(x,y)]$ mit $(x,y) \in \Omega_h$. Da für $u \in X$ die zweiten Differenzen $\partial_{xx} u$ gleichmäßig gegen die zweite Ableitung u_{xx} konvergiert, folgt $\|(\mathbf{L}_h R_X^h - R_Y^h L)u\|_{X_h} \to 0$ für $h \to 0$. Damit ist die Konsistenzbedingung (6.3.1) verifiziert.

Im Fall der Finite-Element-Diskretisierung haben die Matrix $\mathbf{L}_h$ und die rechte Seite $\mathbf{f}_h$ die Darstellung[71]

$$\mathbf{L}_h = R_h L P_h, \qquad \mathbf{f}_h = R_h f, \tag{6.3.2}$$

wobei $P_h : X_h \to X$ die Knotenwerte $\mathbf{u}_i$ in die Finite-Element-Funktion $\sum \mathbf{u}_i b_i$ (b_i: Basisfunktion des i-ten Knotens[74]) abbildet und R_h die Adjungierte ist: $\int_\Omega v(x,y)\,(P_h \mathbf{u}_h)\,(x,y) dxdy = h^2 \sum_i (R_h v)_i\, \mathbf{u}_i$.

Bemerkung 6.3.2 *Im Finite-Element-Fall ist $R_Y^h := R_h$ die kanonische Wahl, sodass*

$$\mathbf{L}_h R_X^h - R_Y^h L = R_h L \left(P_h R_X^h - I\right).$$

Die Räume X, Y sind die Sobolev-Räume[71] *$X = H_0^1(\Omega)$, $Y = H^{-1}(\Omega)$. Die Normen von X_h, Y_h sind $\|\mathbf{u}_h\|_{X_h} = \|P_h \mathbf{u}_h\|_X$, $\|\mathbf{f}_h\|_{Y_h} = \|P_h \mathbf{f}_h\|_Y$. Mit $R_X^h = R_h$ erhält man $\|\left(P_h R_X^h - I\right) u\|_X \to 0$ für $h \to 0$ und $u \in X$. Da $\|R_h L\|_{Y_h \leftarrow X} \le C$, folgt (6.3.1).*

6.4 Konvergenz und Stabilität

Konvergenz ist durch

$$\|R_X^h u - \mathbf{u}_h\|_{\hat{X}_h} \to 0 \qquad \text{für } h \in H \text{ mit } h \to 0 \tag{6.4.1}$$

definiert, wobei $u \in X$ die Differentialgleichung $Lu = f$ (mit Randbedingung $u = 0$) löst, während $\mathbf{u}_h$ die Lösung von (6.2.2) ist. Die Norm in (6.4.1) gehört zu einem Banach-Raum $\hat{X}_h$. Je schwächer die Norm ist, desto einfacher sollte der Nachweis von (6.4.1) sein. Die stärkste Norm ist die von X_h.

Die *Stabilität* ist durch die gleichmäßige Beschränktheit von $\mathbf{L}_h^{-1}$ charakterisiert:

$$\sup_{h \in H} \|\mathbf{L}_h^{-1}\|_{\hat{X}_h \leftarrow Y_h} \le C_{\text{stab}}. \tag{6.4.2}$$

Satz 6.4.1 *Konsistenz (6.3.1) und Stabilität (6.4.2) implizieren die Konvergenz (6.4.1).*

Beweis. Es gilt die Darstellung

$$R_X^h u - \mathbf{u}_h = R_X^h u - \mathbf{L}_h^{-1} \mathbf{f}_h = R_X^h u - \mathbf{L}_h^{-1} R_Y^h f = \mathbf{L}_h^{-1} \left(\mathbf{L}_h R_X^h - R_Y^h L\right) u.$$

Daher beweist $\|R_X^h u - \mathbf{u}_h\|_{X_h} \le \|\mathbf{L}_h^{-1}\|_{X_h \leftarrow Y_h} \|\left(\mathbf{L}_h R_X^h - R_Y^h L\right) u\|_{Y_h} \le C_{\text{stab}} \|\left(\mathbf{L}_h R_X^h - R_Y^h L\right) u\|_{Y_h} \to 0$ die Behauptung. ∎

Für die klassische Analyse der Differenzenverfahren wählt man die Maximumnorm

$$\|\mathbf{u}_h\|_{\hat{X}_h} = \|\mathbf{u}_h\|_\infty.$$

Damit wird $\|\mathbf{L}_h^{-1}\|_{\hat{X}_h \leftarrow Y_h}$ zur Zeilensummennorm $\|\mathbf{L}_h^{-1}\|_\infty$. Für die Abschätzung von $\|\mathbf{L}_h^{-1}\|_\infty$ machen wir davon Gebrauch, dass im Modellbeispiel $-\mathbf{L}_h$ eine M-Matrix ist. Für M-Matrizen existiert eine konstruktive Möglichkeit, die Zeilensummennorm der Inversen zu bestimmen. Im folgenden Lemma ist $\mathbb{1}$ der Vektor, dessen sämtliche Komponenten gleich 1 sind.

[73] Nur die Norm von Y_h erscheint in (6.3.1) und (6.4.2).

[74] Hier ersetzt i den Index ν, μ oder $(\nu h, \mu h) \in \Omega_h$, da bei Finite-Element-Diskretisierungen im Allgemeinen keine regelmäßigen Gitter verwendet werden.

Lemma 6.4.2 *Sei A eine M-Matrix und w ein Vektor, sodass die Ungleichung $Aw \geq \mathbb{1}$ komponentenweise gilt. Dann ist $\|A^{-1}\|_\infty \leq \|w\|_\infty$.*

Beweis. Zu einem beliebigen Vektor u führen wir die Schreibweise $|u|$ für den Vektor $(|u_i|)_{i=1}^n$ ein. Die folgenden Ungleichungen sind komponentenweise zu verstehen. Es gilt $|u| \leq \|u\|_\infty \mathbb{1} \leq \|u\|_\infty Aw$. Wegen der M-Matrix-Eigenschaft c) gilt $A^{-1} \geq 0$, sodass

$$\left|A^{-1}u\right| \leq A^{-1}\,|u| \leq A^{-1}\,\|u\|_\infty\, Aw = \|u\|_\infty\, w$$

und $\|A^{-1}u\|_\infty / \|u\|_\infty \leq \|w\|_\infty$. Also $\|A^{-1}\|_\infty = \sup_{u\neq 0} \|A^{-1}\|_\infty / \|u\|_\infty \leq \|w\|_\infty$. ∎

Im Falle von $A = -\mathbf{L}_h$ sucht man zunächst eine Funktion $w(x,y)$ mit $Lw(x,y) \geq 1$. Dies ist z.B. $w(x,y) = \frac{1}{2}x\,(1-x)$ mit der Maximumnorm $\|w\|_\infty = 1/8$. Als Vektor wählen wir die punktweise Beschränkung $\mathbf{w}_h = R_X^h w$ auf das Gitter Ω_h. Da zweite Differenzen und zweite Ableitungen im Falle der quadratischen Funktion w identisch sind, folgt $(-\mathbf{L}_h)\,\mathbf{w}_h \geq \mathbb{1}$ und $\|\mathbf{w}_h\|_\infty \leq 1/8$, was die Stabilitätseigenschaft

$$\|\mathbf{L}_h^{-1}\|_\infty \leq 1/8 \qquad \text{für alle } h \in H \tag{6.4.3}$$

mit $C_{\text{stab}} = 1/8$ beweist.

6.5 Höhere Konvergenzordnung

Bisher wurde nur die Konvergenzordnung $o(1)$ gezeigt. Im Allgemeinen möchte man $\|R_X^h u - \mathbf{u}_h\|_{X_h} = \mathcal{O}(h^\kappa)$ für ein $\kappa > 0$ zeigen. Dazu muss die Lösung $u \in X$ zusätzliche Glattheitseigenschaften aufweisen: $u \in Z$ für ein $Z \subset X$.

Im Falle des Differenzenverfahrens war $X = \{u \in C^2(\bar\Omega) : u(x,y) = 0 \text{ für } (x,y) \in \Gamma\}$. Wenn man

$$Z := X \cap C^4(\bar\Omega)$$

wählt, folgt $\|\left(\mathbf{L}_h R_X^h - R_Y^h L\right) u\|_{Y_h} = \mathcal{O}(h^2)$ für alle $u \in Z$. Entsprechend folgt die Konvergenz

$$\|R_X^h u - \mathbf{u}_h\|_{\dot X_h} = \mathcal{O}(h^2).$$

7 Literatur

Die Literaturzitate wurden jeweils in den Fußnoten angegeben und sind hier noch einmal gesammelt.

R. Bulirsch: *Bemerkungen zur Romberg-Integration.* Numer. Math. **6** (1964) 6-16 (Fußnote 23)

R. Courant, K.O. Friedrichs und H. Lewy: *Über die partiellen Differenzengleichungen der mathematischen Physik.* Math. Ann. **100** (1928) 32-74 (Fußnote 60)

Ph. J. Davis und Ph. Rabinowitz: *Methods of numerical integration.* Academic Press, New York, 1975 (Fußnote 20)

K.O. Friedrichs: *Symmetric hyperbolic linear differential equations.* Comm. Pure and Appl. Math., **7** (1954) 345-392 (Fußnote 66)

W. Hackbusch: *Theorie und Numerik elliptischer Differentialgleichungen.* 2. Auflage, Teubner-Verlag, Stuttgart, 1996 (Fußnoten 52, 53, 71)

W. Hackbusch: *Iterative Lösung großer schwachbesetzter Gleichungssysteme.* 2. Auflage, Teubner-Verlag, Stuttgart, 1993 (Fußnoten 64, 72)

E. Hairer und G. Wanner: *Solving ordinary differential equations II.* Springer-Verlag, Berlin, 1991 (Fußnote 48)

H. Heuser: *Gewöhnliche Differentialgleichungen.* Teubner, Stuttgart, 1989 (Fußnote 49)

L.S. de Jong: *Towards a formal definition of numerical stability.* Numer. Math. **28** (1977) 211-219 (Fußnote 5)

P.D. Lax und R.D. Richtmyer: *Survey of the stability of linear difference equations.* Comm. Pure and Appl. Math., **9** (1956) 267-293 (Fußnote 55)

P.D. Lax und B. Wendroff: *Difference schemes for hyperbolic equations with high order of accuracy.* Comm. Pure Appl. Math., **17** (1964) 381-398 (Fußnote 63)

R.D. Richtmyer und K.W. Morton: *Difference methods for initial-value problems.* 2. Auflage, Interscience Publishers, New York, 1967 (Fußnote 68)

H.J. Stetter: *Analysis of discretization methods for ordinary differential equations.* Springer-Verlag, Berlin, 1973 (Fußnote 48)

J. Stoer: *Einführung in die Numerische Mathematik I.* 8. Auflage, Springer-Verlag, Berlin, 1999 (Fußnoten 1, 10, 17, 19, 32)

J. Stoer und R. Bulirsch: *Einführung in die Numerische Mathematik II.* 3. Auflage, Springer-Verlag, Berlin, 1990 (Fußnote 7)

J.H. Wilkinson: *Rundungsfehler.* Springer-Verlag, Berlin, 1969 (Fußnote 8)

E. Zeidler (Hrsg.): *Teubner-Taschenbuch der Mathematik*, Teubner-Verlag, Stuttgart, 1996 (Fußnoten 3, 54)

Index