Freie Universität Berlin

INSTITUT FÜR MATHEMATIK II

Computerorientierte Mathematik II

Ralf Kornhuber und Christof Schütte

5. Auflage: Sommersemester 2011
(nach Druck korrigierte Fassung)

Für aufmerksames Durchsehen sei gedankt: Dennis Jentsch & Enrico Cheynubrata

Inhaltsverzeichnis

1 Polynominterpolation

Wir betrachten die folgende Interpolationsaufgabe: Gegeben seien eine Funktion

$$f \in C[a,b] := \{v : [a,b] \to \mathbb{R} \ , \ \text{stetig}\}$$

und ein Gitter von $n+1$ paarweise verschiedenen Stützstellen

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

Gesucht ist ein Polynom p_n höchstens n-ten Grades, d.h.

$$p_n \in P_n = \{v \in C[a,b] \mid v(x) = \sum_{k=0}^{n} a_k x^k \ , \ a_k \in \mathbb{R}\} \subset C[a,b]$$

mit der Eigenschaft

$$p_n(x_k) = f(x_k) \quad \forall k = 0, \ldots, n \ .$$

Unsere Hoffnung ist, auf diese Weise eine gute Approximation der „komplizierten“ Funktion f durch eine „einfache“ Funktion $p_n \in P_n$ zu erhalten. Solche Approximationen sind häufig nützlich. Beispielsweise ist die Integration über p_n einfach durchführbar und könnte eine gute Approximation des Integrals über f liefern.

Bemerkung. Durch die affine Transformation

$$\xi = \frac{2x - (a+b)}{b-a}$$

wird das Intervall von $[a,b]$ in $[-1,1]$ überführt. Es reicht also, die Interpolationsaufgabe nur auf $[-1,1]$ (oder auf jedem anderen festgelegten Intervall) zu betrachten.

1.1 Lagrange-Darstellung und Newtonsche Darstellung

Wir schauen uns nun die Polynome n-ten Grades genauer an.

P_n ist ein linearer Raum über $\mathbb{R}$ bezüglich

Addition: $(p+q)(x) = p(x) + q(x) \quad \forall x \in [a,b]$,
Skalarmultiplikation: $(\lambda p)(x) = \lambda p(x) \quad \forall x \in [a,b]$.

Die sogenannten Monome

$$x^k, \qquad k = 0, \ldots, n \ ,$$

bilden eine Basis von P_n.

Eine andere Basis, die sogenannte Knotenbasis, wird duch die Lagrange-Polynome L_k,

$$L_k(x) = \prod_{\substack{i=0 \\ i \neq k}}^{n} \frac{x - x_i}{x_k - x_i} \ , \qquad k = 0, \ldots, n \ ,$$

aufgespannt. Offenbar gilt

$$L_k(x_i) = \delta_{ik} = \begin{cases} 0 & i \neq k \\ 1 & i = k \end{cases} \quad \text{(Kronecker-}\delta\text{)}$$

und daher

$$p(x) = \sum_{k=0}^{n} p(x_k) L_k(x) \qquad \forall p \in P_n \ .$$

Die Koeffizienten sind also gerade die Werte von p an den Knoten x_k.

Satz 1.1. *(Existenz und Eindeutigkeit)*
Die Interpolationsaufgabe

$$p_n \in P_n : \quad p_n(x_k) = f(x_k) \quad \forall k = 0, \ldots, n \tag{1.1}$$

hat die eindeutig bestimmte Lösung

$$p_n = \sum_{k=0}^{n} f(x_k) L_k \ . \tag{1.2}$$

Beweis. Offenbar ist $p_n \in P_n$. Durch Einsetzen von $x = x_k$, $k = 0, \ldots, n$, prüft man nach, daß die Interpolationsbedingungen erfüllt sind. Also ist p_n eine Lösung.

Sind $p_1, p_2 \in P_n$ zwei Lösungen von (1.1), so gilt

$$q = p_1 - p_2 \in P_n \ , \quad q(x_k) = 0 \quad \forall k = 0, \ldots, n.$$

Also hat $q \in P_n$ $n+1$ Nullstellen. Nach dem Fundamentalsatz der Algebra (Dissertation von Gauß 1799) muß dann $q \equiv 0$ sein. □

Als nächstes wollen wir die Kondition der Polynominterpolation untersuchen. Dazu haben wir festzulegen, wie Störungen in den Eingangsdaten $f \in C[a,b]$ und der Lösung $p_n \in P_n \subset C[a,b]$ gemessen werden sollen.

Da eine stetige Funktion auf einem abgeschlossenen Intervall $[a,b]$ beschränkt ist und ihr Supremum annimmt, ist die sogenannte Maximumsnorm

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

für alle $f \in C[a,b]$ wohldefiniert. Offenbar beschreibt

$$\|f - \tilde{f}\|_\infty = \max_{x \in [a,b]} |f(x) - \tilde{f}(x)|$$

gerade die maximale Abweichung einer gestörten Funktion $\tilde{f} \in C[a,b]$ von $f \in C[a,b]$.

Satz 1.2. *Es sei $\phi_n : C[a,b] \to P_n$ der durch*

$$\phi_n(f) = \sum_{k=0}^{n} f(x_k) L_k \in P_n$$

definierte Interpolationsoperator. Die Abbildung ϕ_n ist linear, d.h. es gilt

$$\phi_n(\alpha f + \beta g) = \alpha\phi_n(f) + \beta\phi_n(g)$$

für alle $f, g \in C[a,b]$ und $\alpha, \beta \in \mathbb{R}$. Weiter gilt

$$\sup_{\substack{f\in C[a,b]\\ f\neq 0}} \frac{\|\phi_n(f)\|_\infty}{\|f\|_\infty} = \Lambda_n \tag{1.3}$$

mit der sogenannten Lebesgue-Konstante

$$\Lambda_n = \left\| \sum_{k=0}^{n} |L_k| \right\|_\infty = \max_{x\in[a,b]} \sum_{k=0}^{n} |L_k(x)| \ .$$

Beweis. Nach Definition ist

$$\begin{aligned}
\phi_n(\alpha f + \beta g) &= \sum_{k=0}^{n} \left(\alpha f(x_k) + \beta g(x_k)\right) L_k \ , \\
&= \alpha \sum_{k=0}^{n} f(x_k) L_k + \beta \sum_{k=0}^{n} g(x_k) L_k \ , \\
&= \alpha\phi_n(f) + \beta\phi_n(g) \qquad \forall f \in C[a,b] \ .
\end{aligned}$$

Wir kommen zum Beweis von (1.3). Es gilt für ein beliebig gewähltes $f \in C[a,b]$

$$\begin{aligned}
\|\phi_n(f)\|_\infty &= \max_{x\in[a,b]} \left| \sum_{k=0}^{n} f(x_k) L_k(x) \right| \\
&\leq \max_{x\in[a,b]} \sum_{k=0}^{n} \max_{\xi\in[a,b]} |f(\xi)| \ |L_k(x)| \\
&= \Lambda_n \ \|f\|_\infty \ .
\end{aligned}$$

Damit haben wir

$$\sup_{\substack{f\in C[a,b]\\ f\neq 0}} \frac{\|\phi_n(f)\|_\infty}{\|f\|_\infty} \leq \Lambda_n$$

gezeigt. Um nachzuweisen, daß sogar Gleichheit gilt, konstruieren wir ein $f^* \in C[a,b]$ mit der Eigenschaft

$$\|\phi_n(f^*)\|_\infty = \Lambda_n \|f^*\|_\infty \ .$$

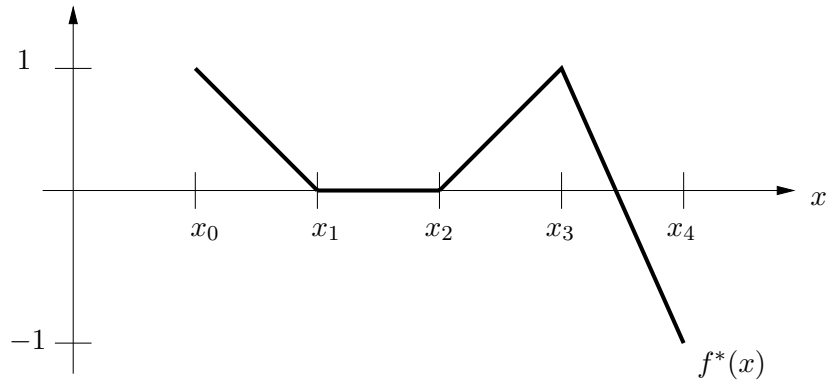
Dazu wählen wir zunächst $x^* \in [a,b]$ so, daß

$$\sum_{k=0}^{n} |L_k(x^*)| = \max_{x\in[a,b]} \sum_{k=0}^{n} |L_k(x)| \ .$$

(Warum ist das möglich?) Wir wählen als nächstes $f^* \in C[a,b]$ mit den Eigenschaften

$$\begin{aligned}
&f^* \text{ linear auf } [x_k, x_{k+1}] \ , \quad k = 0, \ldots, n-1 \ , \\
&f^*(x_k) = \operatorname{sgn}(L_k(x^*)) := \begin{cases} 1 & \text{falls} \quad L_k(x^*) > 0 \\ 0 & \text{falls} \quad L_k(x^*) = 0 \\ -1 & \text{falls} \quad L_k(x^*) < 0 \end{cases} \ .
\end{aligned}$$

Wie das beispielsweise geht, ist in der folgenden Abbildung veranschaulicht.



Nun folgt wegen $\|f^*\|_\infty = 1$

$$\begin{aligned}\|\phi_n(f^*)\|_\infty &= \max_{x\in[a,b]} \left| \sum_{k=0}^{n} \operatorname{sgn}(L_k(x^*))L_k(x) \right| \\ &\geq \left| \sum_{k=0}^{n} \operatorname{sgn}(L_k(x^*))L_k(x^*) \right| = \sum_{k=0}^{n} |L_k(x^*)| = \Lambda_n \|f^*\|_\infty \ .\end{aligned}$$

□

Folgerung: Die absolute Kondition κ_{abs} der Polynominterpolation ist

$$\kappa_{\text{abs}} = \Lambda_n \ .$$

Ist nämlich $p_n = \phi_n(f)$ das Interpolationspolynom zu $f \in C[a,b]$ und ist $\tilde{f} \in C[a,b]$ eine gestörte Funktion mit dem entsprechenden Interpolationspolynom $\tilde{p}_n = \phi_n(\tilde{f})$, so gilt nach Satz 1.2

$$\begin{aligned}\|p_n - \tilde{p}_n\|_\infty &= \|\phi_n(f) - \phi_n(\tilde{f})\|_\infty \\ &= \|\phi_n(f - \tilde{f})\|_\infty \leq \Lambda_n \|f - \tilde{f}\|_\infty \ .\end{aligned}$$

Aus

$$\min_{k=0,\ldots,n-1} |x_{k+1} - x_k| \approx 0$$

folgt $\Lambda_n = \kappa_{\text{abs}} \gg 1$. Im Falle dicht benachbarter Stützstellen ist die Polynominterpolation also schlecht konditioniert (im Falle $x_k = x_{k+1}$ sogar unlösbar!). Soviel zur absoluten Kondition. Wie sieht die relative Kondition κ_{rel} aus (Übung)?

Wir betrachten nun den Aufwand der Polynominterpolation. Dabei verwenden wir

$$\text{Aufwandsmaß} = \text{Anzahl der Punktoperationen.}$$

Verwendet man die sogenannte Lagrange-Darstellung (1.2) des Interpolationspolynoms p_n, so erhält man durch Abzählen folgende Aufwandsabschätzung.

- Die Auswertung des Lagrangeschen Interpolationspolynoms an einer Stelle x erfordert

  $$(n+1)(n+n-1) = \mathcal{O}(n^2)$$

  Punktoperationen.

Wir werden eine Darstellung des Interpolationspolynoms herleiten, welche es erlaubt, die Auswertung mit optimalem Aufwand von n Punktoperationen zu bewerkstelligen. Dazu betrachten wir die sogenannte Newtonsche Darstellung

$$p_n(x) = a_0 + \sum_{i=1}^{n} a_i \prod_{k=0}^{i-1} (x - x_k)$$

des Interpolationspolynoms.

Sind die Koeffizienten $a_k, k = 0, \ldots, n$, bekannt, so läßt sich die Newton-Darstellung nach geschicktem Ausklammern

$$p_n(x) = a_0 + (x - x_0)\,(a_1 + (x - x_1)(a_2 + (\cdots + (x - x_{n-1})a_n)\ldots))$$

rekursiv auswerten.

Algorithmus 1.3. *(Horner-Schema)*

Initialisierung:	$S_n = a_n$.
Rekursion:	*Für* $k = n-1, \ldots, 0$ *berechne* $S_k = S_{k+1}(x - x_k) + a_k$.
Ergebnis:	$p_n(x) = S_0$

Ein entscheidender Vorteil der Newtonschen Darstellung ist also:

- Die Auswertung des Newtonschen Interpolationspolynoms an einer Stelle x mit dem Horner-Schema erfordert n Punktoperationen.

Ein möglicher Nachteil der Newtonschen Darstellung ist, daß die Koeffizienten $a_0, \ldots, a_n$, anders als bei der Lagrange-Darstellung, extra berechnet werden müssen. Wir kümmern uns daher nun um eine möglichst schnelle Berechnungsvorschrift.

Es gilt zunächst

$$f(x_0) = p_n(x_0) = a_0.$$

Sind $a_0, \ldots, a_l$ für ein festes $l < n$ bekannt, so können wir

$$p_l(x) = a_0 + \sum_{i=1}^{l} a_i \prod_{k=0}^{i-1} (x - x_k)$$

auswerten. Die Interpolationsbedingung in x_{l+1} liefert

$$f(x_{l+1}) = p_n(x_{l+1}) = p_l(x_{l+1}) + a_{l+1}(x_{l+1} - x_0) \cdot \ldots \cdot (x_{l+1} - x_l)$$

und damit

$$a_{l+1} = \frac{f(x_{l+1}) - p_l(x_{l+1})}{(x_{l+1} - x_0) \cdot \ldots \cdot (x_{l+1} - x_l)} \ . \tag{1.4}$$

Der Aufwand dieser Berechnungsvorschrift ist

$$\sum_{i=1}^{n}(i+1) = n + \tfrac{1}{2}(n+1)n \ .$$

Jeder Koeffizient a_i hängt also nur von den Werten von f in den Stützstellen $x_0, \ldots, x_i$ ab. Die folgende Bezeichnungsweise hat sich eingebürgert:

Definition 1.4. Die Koeffizienten

$$a_i = f[x_0, \ldots, x_i], \qquad i = 0, \ldots, n,$$

der Newtonschen Darstellung des Interpolationspolynoms heißen Newtonsche dividierte Differenzen i-ter Ordnung.

Hinzufügen einer weiteren Stützstelle x_{n+1} erfordert nur die Berechnung von a_{n+1}. Verwendet man in Berechnungsvorschrift (1.4) das Horner-Schema, so kommt man mit $2n$ Punktoperationen aus. Aber es geht noch schneller.

Satz 1.5. *Die Newtonschen dividierten Differenzen lassen sich aus*

$$f[x_i] = f(x_i) \ , \quad i = 0, \ldots, n,$$

und

$$f[x_i, \ldots, x_k] = \frac{f[x_{i+1}, \ldots, x_k] - f[x_i, \ldots, x_{k-1}]}{x_k - x_i} \ , \qquad 0 \le i < k \le n \ , \tag{1.5}$$

rekursiv berechnen.

Beweis. Wir haben nur (1.5) zu zeigen. Der folgende Beweis geht auf Neville zurück. Es seien i, k mit $0 \le i < k \le n$ beliebig, aber fest gewählt und p_{ik} Lösung der Interpolationsaufgabe

$$p_{ik} \in P_{k-i} \ : \qquad p_{ik}(x_\nu) = f(x_\nu) \qquad \forall \nu = i, i+1, \ldots, k \ . \tag{1.6}$$

Durch Ausmultiplizieren der Newtonschen Darstellung von p_{ik} sieht man, daß $f[x_i, \ldots, x_k]$ der führende Koeffizient von p_{ik} ist, daß also

$$p_{ik}(x) = f[x_i, \ldots, x_k]x^{k-i} + q_0(x)$$

mit einem geeigneten $q_0 \in P_{k-i-1}$ gilt.

Durch Einsetzen von $x = x_i, \ldots, x_k$ bestätigt man, daß

$$p_{ik}(x) = \frac{1}{x_k - x_i}\left((x - x_i)p_{i+1,k}(x) - (x - x_k)p_{i,k-1}(x)\right) \tag{1.7}$$

eine weitere Darstellung der (eindeutig bestimmten!) Lösung von (1.6) ist. Die führenden Koeffizienten von $p_{i+1,k}$ bzw. $p_{i,k-1}$ sind $f[x_{i+1}, \ldots, x_k]$ bzw. $f[x_i, \ldots, x_{k-1}]$. Einsetzen in die Darstellung (1.7) ergibt

$$p_{ik}(x) = \frac{f[x_{i+1}, \ldots, x_k] - f[x_i, \ldots, x_{k-1}]}{x_k - x_i}x^{k-i} + q_0(x) \ .$$

Koeffizientenvergleich liefert die Behauptung. □

Bemerkung. Die dividierten Differenzen sind von der Reihenfolge der Stützstellen unabhängig (Übung).

Es besteht offenbar ein enger Zusammenhang zwischen dividierten Differenzen und Differenzenquotienten. Näheres dazu erfährt man später in der Vorlesung 'Einführung in die Numerische Mathematik' oder, wer das nicht erwarten kann, in Satz 7.12 des Lehrbuchs von Deuflhard und Hohmann [1].

Auf der Basis von Satz 1.5 erfolgt die Berechnung dividierter Differenzen entsprechend dem nachstehenden Dreiecksschema.

Algorithmus 1.6. *(Neville)*

$$
\begin{array}{ccccc}
f[x_0] & & & & \\
 & \searrow & & & \\
f[x_1] & \longrightarrow & f[x_0,x_1] & & \\
 & \searrow & & \searrow & \\
f[x_2] & \longrightarrow & f[x_1,x_2] & \longrightarrow & f[x_0,x_1,x_2]
\end{array}
$$

Jeder waagrechte Pfeil im Dreiecksschema erfordert eine Division. Wird ein weiterer Punkt x_{n+1} hinzugenommen, so muß nur die untere Zeile des Dreiecksschemas neu berechnet werden, um die Newtonsche Darstellung von p_{n+1} zu erhalten.

- Das Aufstellen des Newtonschen Interpolationspolynoms mit dem Nevilleschen Dreiecksschema erfordert
$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$
Punktoperationen.
- Hinzufügen einer weiteren Stützstelle $x_{n+1} \in (a,b)$ erfordert $n+1$ Punktoperationen.

Man vergleiche Lagrange- und Newton-Darstellung hinsichtlich des Rechenaufwands (Übung).

Oft interessiert nur der Wert $p_n(x^*)$ des Interpolationspolynoms an einer einzigen Stelle x^*. Anstatt das Newtonsche Interpolationspolynom aufzustellen und dann das Horner-Schema zu verwenden, kann man $p_n(x^*)$ direkt bestimmen. Nach (1.7) gilt nämlich

$$p_{ik}(x^*) = \frac{1}{x_k - x_i}\left((x^* - x_i)p_{i+1,k}(x^*) - (x^* - x_k)p_{i,k-1}(x^*)\right), \qquad 0 \le i < k \le n. \tag{1.8}$$

Im Falle $i = k$ ist bekanntlich

$$p_{ii}(x^*) = f(x_i), \qquad i = 0, \ldots, n. \tag{1.9}$$

Die rekursive Berechnung des gesuchten Wertes $p_n(x^*) = p_{0n}(x^*)$ erfolgt dann unter Verwendung von (1.8) und (1.9) nach dem folgenden Dreiecksschema:

Algorithmus 1.7. *(Aitken-Neville)*

$$
\begin{array}{ccccc}
p_{00}(x^*) & & & & \\
 & \searrow & & & \\
p_{11}(x^*) & \longrightarrow & p_{01}(x^*) & & \\
 & \searrow & & \searrow & \\
p_{22}(x^*) & \longrightarrow & p_{12}(x^*) & \longrightarrow & p_{02}(x^*)
\end{array}
$$

- Die Berechnung von $p_n(x^*)$ mit dem Dreiecksschema von Aitken-Neville erfordert $3\frac{n(n+1)}{2}$ Punktoperationen.
- Hinzufügen einer weiteren Stützstelle $x_{n+1} \in (a,b)$ erfordert $n+1$ Punktoperationen.

Der zweite Punkt ist vor allem für Extrapolationsverfahren wichtig (Übung).

Bemerkung. Gilt für gewisse $k = 0, \ldots, n-1$

$$f(x_k) \approx f(x_{k+1}) \text{ oder } x_k \approx x_{k+1} \ ,$$

so besteht in den Algorithmen 1.3, 1.6 und 1.7 die Gefahr von Auslöschung! Einige Bemerkungen zur Stabilisierung finden sich in dem Lehrbuch von Stoer [3] im Kapitel über Polynominterpolation.

1.2 Das Restglied bei der Polynominterpolation

Wir wollen nun für ein gegebenes $x^* \in [a,b]$ den Interpolationsfehler

$$f(x^*) - p(x^*)$$

untersuchen. Wir vereinbaren zunächst folgende Notation

$$C^m[a,b] = \{v \in C[a,b] \mid v^{(m)} \in C[a,b]\} \ , \quad m \in \mathbb{N} \ .$$

Wir erinnern an einen wichtigen Satz aus der Analysis.

Satz 1.8. *(Satz von Rolle)*
Sei $f \in C^1[a,b]$ und $f(a) = f(b) = 0$. Dann existiert ein $\xi \in (a,b)$ mit der Eigenschaft

$$f'(\xi) = 0 \ .$$

Beweis. Ist $f(x) = f(a)\ \forall x \in (a,b)$, so gilt trivialerweise $f'(x) = 0\ \forall x \in (a,b)$. Andernfalls gibt es eine Stelle $\xi \in (a,b)$, an der f ein relatives Minimum oder Maximum annimmt. Es folgt $f'(\xi) = 0$. □

Wir können nun das Hauptresultat dieses Abschnittes beweisen.

Satz 1.9. *Sei $p_n \in P_n$ Lösung der Interpolationsaufgabe (1.1) und $f \in C^{n+1}[a,b]$. Dann gibt es zu jedem $x^* \in [a,b]$ ein $\xi = \xi(x^*) \in (a,b)$ mit der Eigenschaft*

$$f(x^*) - p_n(x^*) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{k=0}^{n} (x^* - x_k) \ .$$

Beweis. Ist $x^* = x_k$ für ein $k = 0, \ldots, n$, so haben wir nichts zu beweisen. Es sei also $x^* \neq x_k$ $\forall k = 0, \ldots, n$. Wir betrachten die Funktion g,

$$g(x) = f(x) - p_n(x) - \gamma \prod_{k=0}^{n} (x - x_k) \ ,$$

und bestimmen $\gamma \in \mathbb{R}$ so, daß $g(x^*) = 0$, also

$$\gamma = \frac{f(x^*) - p_n(x^*)}{(x^* - x_0) \cdot \ldots \cdot (x^* - x_n)} \ .$$

Offenbar hat g mindestens $n + 2$ verschiedene Nullstellen in $[a, b]$, nämlich $x_0, \ldots, x_n$ und x^*. Nach dem Satz von Rolle finden wir zwischen jeder dieser Nullstellen von g eine Nullstelle von g'. Induktive Fortsetzung dieser Argumentation ergibt, daß $g^{(n+1)}$ mindestens eine Nullstelle $\xi \in (a, b)$ haben muß. Es gilt also

$$0 = g^{(n+1)}(\xi) = f^{(n+1)}(\xi) - 0 - \gamma(n + 1)!$$

und damit die Behauptung. □

Aus Satz 1.9 erhält man sofort die Fehlerabschätzung

$$\|f - p_n\|_\infty \leq \frac{1}{(n+1)!} \|f^{(n+1)}\|_\infty \|\omega\|_\infty \ ,$$

wobei

$$\omega(x) = \prod_{k=0}^{n} (x - x_k)$$

gesetzt ist. Offenbar hängt der zweite Faktor

$$\|\omega\|_\infty = \max_{x \in [a,b]} |(x - x_0) \cdot \cdots \cdot (x - x_n)|$$

nur von der Wahl der Stützstellen ab.

Daß man bei der Polynominterpolation unangenehme Überraschungen erleben kann, zeigen folgende Beispiele.

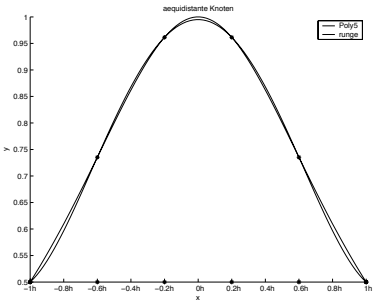
Beispiel. Wir betrachten die beliebig oft differenzierbare Funktion

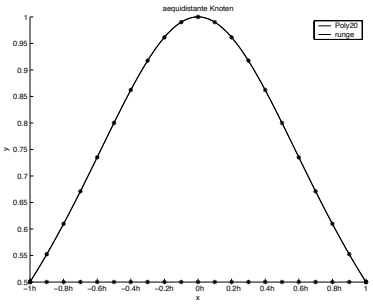
$$f(x) = \frac{1}{1 + x^2}$$

auf dem Intervall $[-1, 1]$. Zur Interpolation verwenden wir äquidistante Stützstellen

$$x_k = -1 + k \cdot h, \quad h = \frac{2}{n} \ .$$

Die folgenden Abbildungen zeigen die Funktion f und das Interpolationspolynom p_n, jeweils für $n = 5$ und $n = 20$. Für $n = 20$ kann man mit dem Auge keinen Unterschied mehr erkennen. Auch weitere Versuche mit wachsendem n deuten auf Konvergenz hin.

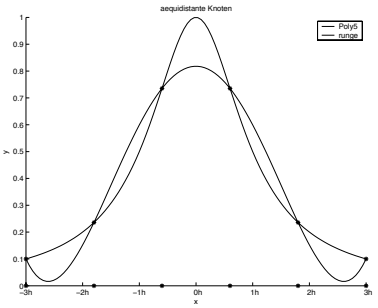


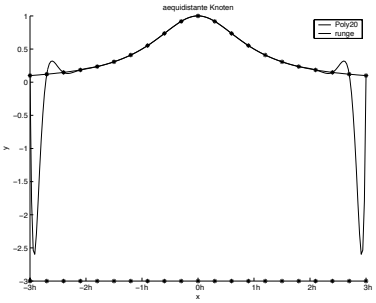


Wiederholen wir unser Experiment auf dem Intervall $[-3, 3]$, wieder mit äquidistanten Stützstellen

$$x_k = -3 + k \cdot h, \quad h = \frac{6}{n} \ ,$$

wieder für $n = 5$ und $n = 20$, so sieht das Ergebnis nicht mehr so gut aus, wie auf den nächsten Abbildungen zu sehen ist.





Weitere Versuche zeigen, daß der Fehler mit wachsendem Polynomgrad n immer größer statt kleiner wird!

Zur Aufklärung verweisen wir auf die Vorlesung 'Einführung in die Numerische Mathematik'. Ungeduldige können sich schon vorher, beispielsweise bei Hämmerlin und Hoffmann [2] in Kapitel 5, § 4, informieren.

Weiterführende Fragen.

- Kann man Stützstellen $x_0, \ldots, x_n$ angeben, so daß

  $$\|\omega\|_\infty \leq \|p\|_\infty \qquad \forall p \in P_{n+1}$$

  (Stichwort: Tschebyscheff-Polynome)
- Unter welchen Bedingungen an f und/oder an die Stützstellen kann man Konvergenz

  $$\lim_{n\to\infty} \|f - p_n\|_\infty = 0$$

  zeigen (Stichwort: Approximationstheorie)?

Literatur

[1] P. Deuflhard und A. Hohmann. *Numerische Mathematik I.* de Gruyter, 1993. Hier wird ein anspruchsvollerer Zugang zu dividierten Differenzen über die Hermite–Genocchi–Formel verfolgt. Auf diese Weise bekommt man eine Newtonsche Darstellung der Lösung von Hermite-Interpolationspolynomen (nicht nur Funktionswert, sondern auch Ableitungen werden vorgeschrieben) gleich mitgeliefert.

[2] G. Hämmerlin und K.-H. Hoffmann. *Numerische Mathematik.* Springer, 3. Auflage, 1992. Ein eher analytisch orientiertes Lehrbuch. Unter anderem enthält es je ein Kapitel über Approximation, Interpolation und Integration.

[3] J. Stoer. *Numerische Mathematik 1.* Springer, 1999. 8. Auflage. Ein Standardwerk.

2 Numerische Quadratur

2.1 Vorbemerkungen zum Integrationsoperator

Gegeben sei eine stetige Funktion, also

$$f \in C[a,b] \ .$$

Wir wollen das Integral

$$I(f) = \int_a^b f(x)\,dx \tag{2.10}$$

berechnen. Nach dem Hauptsatz der Differential- und Integralrechnung ist

$$I(f) = F(b) - F(a) \ ,$$

wobei F eine Stammfunktion von f, d.h. eine Funktion mit der Eigenschaft

$$F'(x) = f(x) \qquad \forall x \in [a,b]$$

darstellt. In den Grundvorlesungen zur Analysis lernt man Regeln kennen, nach denen Stammfunktionen zu Polynomen, rationalen Funktionen, Exponentialfunktion, usw. ermittelt werden können. Diese Regeln kann man auch einem Computer beibringen. Das Ergebnis sind *Symbolikprogramme*, wie zum Beispiel MAPLE. Zu gegebenem f wird eine Stammfunktion in geschlossener Form, d.h. als rationale Funktion von (komplexer) Exponentialfunktion und Logarithmus, ermittelt, falls dies möglich ist. Allein die Frage nach der Existenz einer Stammfunktion führt tief hinein in die Algebra (differentielle Galoistheorie).

Beispiel. Zu berechnen ist das Integral

$$\int_0^1 x^2\,dx \ .$$

Nach Starten von MAPLE liefert die Eingabe

```
> int(x^2,x=0..1);
```

das richtige Resultat

```
                              1/3
```

Versucht man in gleicher Weise

$$\int_0^1 e^{-x^2}\ dx$$

mit MAPLE zu berechnen, so erhält man auf die Eingabe

```
> int(exp(-x^2),x=0..1);
```

hin die Antwort

```
                          1/2 erf(1) Pi^1/2
```

Nun ist der Wert der sogenannten *error function* `erf` an der Stelle 1 gerade definiert durch

$$\texttt{erf}(1) = \frac{2}{\sqrt{\pi}} \int_0^1 e^{-x^2}\, dx \ .$$

Mit dieser Antwort sind wir also keinen Schritt weitergekommen.

Hintergrund des zweiten Beispiels ist, daß keine geschlossene Darstellung der Stammfunktion von e^{-x^2} existiert. In solchen Fällen bleibt nur die näherungsweise Berechnung mittels Näherungsverfahren, sogenannter Quadraturformeln. Von einem numerischen Verfahren können wir *nicht die exakte Lösung*, sondern nur eine *beliebig genaue Näherung* erwarten. Wir stellen uns daher folgende Aufgabe:

Zu einem gegebenen $\varepsilon > 0$ finde man eine Näherung $\tilde{I}(f)$ mit der Eigenschaft

$$|I(f) - \tilde{I}(f)| < \varepsilon \ . \tag{2.11}$$

Der Fehler $|I(f) - \tilde{I}(f)|$ heißt Diskretisierungsfehler. Den Aufwand zur Lösung von (2.11) beschreiben wir durch die Anzahl N der benötigten Auswertungen von f. Eine Quadraturformel $\tilde{I}_N(f)$ mit N f-Auswertungen heißt konvergent, falls

$$\lim_{N\to\infty} \tilde{I}_N(f) = I(f)$$

gilt und von r-ter Ordnung, falls es eine positive Konstante c gibt, die nicht von N abhängt, so daß

$$|I(f) - \tilde{I}_N(f)| \le cN^{-r}$$

richtig ist. Die Effizienz von $\tilde{I}_N(f)$ wird gemäß

$$\text{Effizienz} \ = \ \text{Genauigkeit pro Aufwand} \ = |I(f) - \tilde{I}_N(f)|^{-1} N^{-1}$$

definiert. Vor diesem Hintergrund sind wir nur an konvergenten Quadraturformeln interessiert und wünschen uns eine möglichst hohe Ordnung r bei möglichst kleinem c.

Ein wichtiges allgemeines Prinzip bei der Entwicklung numerischer Verfahren ist die Erhaltung wesentlicher Struktureigenschaften des kontinuierlichen Modells (hier des bestimmten Integrals) durch die Diskretisierung (hier der Quadraturformel). Wir stellen also einige wesentliche Eigenschaften des bestimmten Integrals zusammen.

Struktureigenschaften von I:

- Der Integrationsoperator
$$I : f \in C[a,b] \to I(f) \in \mathbb{R}$$
ist linear, d.h. es gilt
$$I(\alpha f + \beta g) = \alpha I(f) + \beta I(g)$$
für alle $f, g \in C[a,b]$, $\alpha, \beta \in \mathbb{R}$.

- Die absolute Kondition von I bezüglich der Norm $\|\cdot\|_\infty$ ist $\kappa_{\text{abs}} = b-a$, denn es gilt für jede gestörte Funktion $\tilde{f} \in C[a,b]$

$$\begin{aligned} |I(f) - I(\tilde{f})| &= \left| \int_a^b f(x) - \tilde{f}(x)\ dx \right| \leq \int_a^b |f(x) - \tilde{f}(x)|\ dx \\ &\leq (b-a) \max_{x\in[a,b]} |f(x) - \tilde{f}(x)| = (b-a)\|f - \tilde{f}\|_\infty\ . \end{aligned}$$

  Die Abschätzung ist scharf, denn im Falle von $\tilde{f}(x) = 1 + f(x)$ gilt Gleichheit. Wie sieht die relative Kondition aus (Übung)?

- Der Integrationsoperator I ist positiv, d.h.

$$f(x) \geq 0\ \forall x \in [a,b] \Rightarrow I(f) \geq 0\ .$$

2.2 Globale Newton-Côtes-Formeln

Bei der Konstruktion von Quadraturformeln gehen wir von folgender Grundidee aus: Erst wird die "komplizierte" Funktion f durch eine "einfache" Funktion approximiert. Dann liefert die Integration dieser einfachen Funktion eine Näherung für $I(f)$.

Als Approximation von $f \in C[a,b]$ wollen wir das Interpolationspolynom p_n zu den Stützstellen

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

verwenden. Wir erhalten dann als Approximation von $I(f)$

$$I_n(f) = \int_a^b p_n(x)\ dx\ .$$

Einsetzen der Lagrange-Darstellung

$$p_n = \sum_{k=0}^n f(x_k) L_k$$

liefert

$$I_n(f) = (b-a) \sum_{k=0}^n f(x_k)\lambda_k \tag{2.12}$$

mit den Gewichten λ_k,

$$\lambda_k = \frac{1}{b-a} \int_a^b L_k(x)\ dx\ , \qquad k = 0, \ldots, n\ . \tag{2.13}$$

Nach Konstruktion ist die Quadraturformel (2.12) exakt für alle $f \in P_n$. Diese Aussage läßt sich umkehren.

Satz 2.1. *Die Quadraturformel* (2.12) *ist genau dann exakt für alle $f \in P_n$, wenn* (2.13) *gilt.*

Beweis. Sei (2.12) exakt für alle $f \in P_n$. Wir setzen $f = L_k$ und erhalten aus $L_k(x_j) = \delta_{kj}$ (Kronecker-δ)

$$\int_a^b L_k(x)\ dx = (b-a) \sum_{j=0}^n L_k(x_j)\lambda_j = (b-a)\lambda_k$$

und damit (2.13). □

Wir vergleichen nun strukturelle Eigenschaften von I und I_n.

Struktureigenschaften von I_n:

- $I_n : f \in C[a,b] \to I_n(f) \in \mathbb{R}$ ist linear.
- Unter der Voraussetzung $\lambda_k \geq 0 \ \forall k = 0, \ldots, n$ ist $\kappa_{\text{abs}} = b - a$ die absolute Kondition von I_n. Für jedes $\tilde{f} \in C[a,b]$ gilt nämlich

$$\begin{aligned} |I_n(f) - I_n(\tilde{f})| &\leq (b-a) \sum_{k=0}^{n} |f(x_k) - \tilde{f}(x_k)| \ |\lambda_k| \\ &\leq (b-a)\|f - \tilde{f}\|_\infty \sum_{k=0}^{n} |\lambda_k| = (b-a)\|f - \tilde{f}\|_\infty \ , \end{aligned}$$

  wegen $\sum_{k=0}^{n} \lambda_k = 1$. Gleichheit erhält man bei Wahl von $\tilde{f}(x) = 1 + f(x)$.
- I_n ist genau dann positiv, wenn $\lambda_k \geq 0 \ \forall k = 0, \ldots, n$ gilt .

Vor diesem Hintergrund sind wir nur an Verfahren mit nichtnegativen Gewichten λ_k interessiert.

Definition 2.2. Quadraturformeln der Gestalt (2.12) mit Gewichten aus (2.13) und äquidistanten Stützstellen

$$x_k = a + kh \ , \qquad k = 0, \ldots, n \ , \qquad h = \frac{b-a}{n}$$

heißen (globale) Newton-Côtes-Formeln.

Die Gewichte λ_k von Newton-Côtes-Formeln lassen sich wie folgt berechnen

$$\lambda_k = \frac{1}{b-a} \int_a^b L_k(x)\, dx = \frac{1}{b-a} \int_a^b \prod_{\substack{j=0 \\ j \neq k}}^{n} \frac{x - x_j}{x_k - x_j}\, dx = \frac{1}{n} \int_0^n \prod_{\substack{j=0 \\ j \neq k}}^{n} \frac{s - j}{k - j}\, ds \ .$$

Offenbar hängen die Gewichte nicht von a und b ab. Das war der Sinn der Normierung in (2.13). Man erhält folgende Zahlenwerte:

n	λ_k							Name
1	$\frac{1}{2}$	$\frac{1}{2}$						Trapezregel
2	$\frac{1}{6}$	$\frac{4}{6}$	$\frac{1}{6}$					Simpson-Regel
3	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$				Newtonsche $\frac{3}{8}$-Regel
4	$\frac{7}{90}$	$\frac{32}{90}$	$\frac{12}{90}$	$\frac{32}{90}$	$\frac{7}{90}$			Milne-Regel
5	$\frac{19}{288}$	$\frac{75}{288}$	$\frac{50}{288}$	$\frac{50}{288}$	$\frac{75}{288}$	$\frac{19}{288}$		———
6	$\frac{41}{840}$	$\frac{216}{840}$	$\frac{27}{840}$	$\frac{272}{840}$	$\frac{27}{840}$	$\frac{216}{840}$	$\frac{41}{840}$	Weddle-Regel

Ab $n = 8$ treten negative Gewichte auf, und die entsprechenden Newton-Côtes-Formeln werden unbrauchbar!

Wir wollen nun den Diskretisierungsfehler $|I(f) - I_n(f)|$ untersuchen und beginnen mit dem Fall $n = 1$.

Satz 2.3. *Die Trapezregel I_1 ist für alle $f \in P_1$ exakt, und es gilt für alle $f \in C^2[a,b]$:*

$$|I(f) - I_1(f)| \leq \frac{1}{12}\|f''\|_\infty (b-a)^3$$

Beweis. Ist $f \in P_1$, so folgt $p_1(f) = f$ und somit $I_1(f) = I(f)$.

Es sei nun $f \in C^2[a,b]$. Nach Satz 1.9 gibt es dann zu jedem $x \in [a,b]$ ein $\xi(x) \in (a,b)$ mit der Eigenschaft

$$f(x) = p_1(x) + \frac{f''(\xi(x))}{2}(x-a)(x-b) \ .$$

Es folgt

$$\begin{aligned}
|I(f) - I_1(f)| &= \left| \int_a^b p_1(x) + \frac{f''(\xi(x))}{2}(x-a)(x-b)\,dx - \int_a^b p_1(x)\,dx \right| \\
&\leq \frac{1}{2}\int_a^b |f''(\xi(x))|\ |(x-a)(x-b)|\,dx \\
&\leq \frac{1}{2}\|f''\|_\infty \underbrace{\int_a^b (x-a)(b-x)\,dx}_{\frac{1}{6}(b-a)^3} \\
&= \frac{1}{12}\|f''\|_\infty (b-a)^3 \ .
\end{aligned}$$

□

Im Falle $n = 2$ erwartet man, daß Funktionen $f \in P_2$ exakt integriert werden. Wir erleben aber eine freudige Überraschung.

Satz 2.4. *Die Simpson-Regel I_2 ist für alle $f \in P_3$ exakt und es gilt für alle $f \in C^4[a,b]$*

$$|I(f) - I_2(f)| \leq \frac{1}{4!}\|f^{(4)}\|_\infty (b-a)^5 \ . \tag{2.14}$$

Beweis. Sei $f \in P_3$. Dann haben wir nach Satz 1.8 die Restglieddarstellung

$$f(x) = p_2(x) + \gamma(x-a)\left(x - \tfrac{a+b}{2}\right)(x-b)$$

mit einer *Konstanten* $\gamma = \frac{f'''(\xi)}{3!}$. Somit ist

$$I(f) - I_2(f) = \gamma \int_a^b (x-a)\left(x - \tfrac{a+b}{2}\right)(x-b)\,dx = \gamma \cdot 0 \ .$$

Sei nun $f \in C^4[a,b]$. Zum Beweis der Fehlerabschätzung (2.14) wählen wir einfach eine weitere Stützstelle, z.B. $x^* = a + \frac{b-a}{4}$ und erhalten aus der Restglieddarstellung für das zugehörige Interpolationspolynom $p_3 \in P_3$, also aus

$$f(x) = p_3(x) + \frac{f^{(4)}(\xi(x))}{4!}(x-a)(x-x^*)\left(x-\tfrac{a+b}{2}\right)(x-b)\ ,$$

sofort

$$\begin{aligned} |I(f) - I_2(f)| &= |I(f) - I_2(p_3)| = |I(f) - I(p_3)| \qquad (2.15)\\ &\le \frac{1}{4!}\int_a^b |f^{(4)}(\xi(x))|\ |(x-a)(x-x^*)\left(x-\tfrac{a+b}{2}\right)(x-b)|\,dx \\ &\le \frac{1}{4!}\|f^{(4)}\|_\infty (b-a)^5\ . \qquad (2.16)\end{aligned}$$

□

Bemerkung. Der Faktor $\frac{1}{4!}$ in der Abschätzung (2.14) läßt sich verbessern. Wählt man im Beweis das sogenannte Hermite-Polynom $p_3 \in P_3$, das durch die Bedingungen

$$p_3(x_k) = f(x_k)\ , \quad k = 0,1,2 \qquad \text{und} \qquad p_3'(x_1) = f'(x_1)$$

charakterisiert ist, so gilt die Restglieddarstellung

$$f(x) = p_3(x) + \frac{f^{(4)}(\xi(x))}{4!}(x-a)\left(x-\tfrac{a+b}{2}\right)^2(x-b)\ .$$

Mit Hilfe dieser Formel erhält man eine Abschätzung der Form (2.14) mit dem Faktor $\frac{1}{90}$.

Noch genauer kann man zeigen, daß ein $\xi \in (a,b)$ existiert, so daß gilt

$$I_n(f) - I(f) = \frac{f^{(4)}(\xi)}{90}(b-a)^5.$$

Ähnliche Resultate liegen für die anderen Newton-Côtes-Formeln vor (siehe etwa Tabelle 1). Beweise dieser weitergehenden Aussagen finden sich z.B. in Kapitel 9 des Lehrbuches von Deuflhard und Hohmann [1].

Insgesamt ist das Ergebnis dieses Abschnittes unbefriedigend. Es ist uns nicht gelungen, eine konvergente Folge von Quadraturformeln zu finden, mit der wir unsere Aufgabe (2.11) für jedes vorgegebene ε lösen können. Aber wir sind nahe dran.

2.3 Summierte Newton-Côtes-Formeln

Die rettende Idee besteht darin, die Newton-Côtes-Formeln nicht auf dem gesamten Intervall $[a,b]$, sondern auf geeigneten Teilintervallen anzuwenden.

Dazu wählen wir Gitterpunkte z_k, $k = 0,\dots,n$,

$$a = z_0 < z_1 < \cdots < z_n = b$$

Name	Fehlerdarstellung
Trapezregel	$\frac{f''(\xi)}{12}(b-a)^3$
Simpson-Regel	$\frac{f^{(4)}(\xi)}{90}\left(\frac{b-a}{2}\right)^5$
Newtonsche $\frac{3}{8}$-Regel	$\frac{3f^{(4)}(\xi)}{80}\left(\frac{b-a}{4}\right)^5$
Milne-Regel	$\frac{8f^{(6)}(\xi)}{945}\left(\frac{b-a}{5}\right)^7$

Tabelle 1: Fehlerdarstellung globaler Newton-Côtes-Formeln.

und zerlegen das Intervall $[a,b]$ in entsprechende Teilintervalle

$$[a,b] = \bigcup_{k=0}^{n-1} V_k \qquad V_k = [z_k, z_{k+1}] \ .$$

Wir erhalten eine entsprechende Zerlegung des Integrals

$$I(f) = \int_a^b f(x)\,dx = \sum_{k=0}^{n-1} \int_{V_k} f(x)\,dx \ .$$

Summierte Quadraturformeln basieren auf Anwendung gegebener Quadraturformeln (z.B. Newton-Côtes-Formeln) auf die Teilintegrale,

$$I_{V_k}(f) \approx \int_{V_k} f(x)\ dx \ ,$$

und Aufsummieren

$$S_n(f) = \sum_{k=0}^{n-1} I_{V_k}(f) \approx I(f) \ .$$

Wir betrachten als erstes die *summierte Trapezregel* für *äquidistante Stützstellen*

$$\begin{array}{lll} x_k = a + kh \ , & k = 0, \ldots, n \ , & h = \frac{b-a}{n} \ , \\ z_k = x_k \ , & k = 0, \ldots, n \ , & \\ V_k = [x_k, x_{k+1}] \ , & k = 0, \ldots, n-1 \ . & \end{array}$$

Anwendung der Trapezregel auf V_k liefert

$$I_{V_k}(f) = \frac{h}{2}(f(x_k) + f(x_{k+1})) \ ,$$

und Aufsummieren ergibt die summierte Trapezregel $S_n^{(1)}$,

$$S_n^{(1)}(f) = \frac{h}{2}(f(a) + f(b)) + h\sum_{k=1}^{n-1} f(x_k) \ . \tag{2.17}$$

Satz 2.5. *Es sei $f \in C^2[a,b]$. Dann gilt für die summierte Trapezregel $S_n^{(1)}$ die Fehlerabschätzung*

$$|I(f) - S_n^{(1)}(f)| \leq \frac{h^2}{12}(b-a)\|f''\|_\infty \ .$$

Beweis. Nach Satz 2.3 ist

$$\left| \int_{V_k} f(x)\ dx - I_{V_k}(f) \right| \leq \frac{h^3}{12} \max_{x \in V_k} |f''(x)|,$$

und es folgt

$$\begin{aligned} |I(f) - S_n^{(1)}(f)| &\leq \sum_{k=0}^{n-1} \left| \int_{V_k} f(x)\ dx - I_{V_k}(f) \right| \\ &\leq \frac{h^2}{12} \frac{b-a}{n} \sum_{k=0}^{n-1} \max_{x \in [a,b]} |f''(x)| \leq \frac{h^2}{12}(b-a)\|f''\|_\infty \ . \end{aligned}$$

□

Die Anzahl der f-Auswertungen von $S_n^{(1)}$ ist $N^{(1)} = n+1 \leq 2n$. Einsetzen der Abschätzung $h = \frac{b-a}{N^{(1)}-1} \leq 2\frac{b-a}{N^{(1)}}$ in die Fehlerabschätzung von Satz 2.5 liefert

$$|I(f) - S_n^{(1)}(f)| \leq \frac{(b-a)^3}{3}\|f''\|_\infty \big(N^{(1)}\big)^{-2} \ .$$

Die summierte Trapezregel mit äquidistanten Stützstellen ist also für alle $f \in C^2[a,b]$ von zweiter Ordnung. Der Aufwand $N^{(1)}(\varepsilon)$ zur Reduktion des Diskretisierungsfehlers unter eine gegebene Schranke $\varepsilon > 0$ ist dann

$$N^{(1)}(\varepsilon) = \left[\left(\frac{(b-a)^3}{3} \|f''\|_\infty \right)^{\frac{1}{2}} \varepsilon^{-\frac{1}{2}} \right] \ . \tag{2.18}$$

Dabei ist mit der sogenannten Gauß-Klammer $[x]$, $x \in \mathbb{R}$, die kleinste natürliche Zahl mit der Eigenschaft $[x] \geq x$ gemeint.

Zum Abschluß betrachten wir noch die *summierte Simpson-Regel* für ein äquidistantes Gitter

$$\begin{array}{lcll} z_k & = & a + kh \ , & k = 0, \ldots, n \ , \qquad h = \frac{b-a}{n} \ , \\ x_{2k+j} & = & z_k + j\frac{h}{2} \ , & j = 0, \ldots, 2 \ , \\ V_k & = & [z_k, z_{k+1}] = [x_{2k}, x_{2k+2}] \ , & k = 0, \ldots, n-1 \ . \end{array}$$

Anwendung der Simpson-Regel auf V_k liefert

$$I_{V_k} = \frac{h}{6}\left(f(x_{2k}) + 4f(x_{2k+1}) + f(x_{2k+2}) \right) \ ,$$

und Aufsummieren ergibt die summierte Simpson-Regel $S_n^{(2)}$,

$$\begin{aligned} S_n^{(2)} &= \frac{h}{6} \sum_{k=0}^{n-1} (f(x_{2k}) + 4f(x_{2k+1}) + f(x_{2k+2})) \\ &= \frac{h}{6}\left(f(x_0) + 4f(x_1) + 2f(x_2) + \cdots + 4f(x_{2n-1}) + f(x_{2n}) \right) \ . \end{aligned} \tag{2.19}$$

Satz 2.6. *Es sei $f \in C^4[a,b]$. Dann gilt für die summierte Simpson-Regel $S_n^{(2)}$ die Fehlerabschätzung*

$$\left|I(f) - S_n^{(2)}(f)\right| \leq \frac{1}{90} h^4 (b-a) \|f^{(4)}\|_\infty \ .$$

Beweis. Nach der Bemerkung im Anschluß an Satz 2.4 gilt

$$\left|\int_{V_k} f(x)\,dx - I_{V_k}(f)\right| \leq \frac{h^5}{90} \max_{x \in V_k} |f^{(4)}(x)| \ ,$$

und es folgt wie im Beweis zu Satz 2.5

$$|I(f) - S_n(f)| \leq \frac{h^4}{90} \frac{b-a}{n} \sum_{k=0}^{n-1} \max_{x \in V_k} |f^{(4)}(x)| \leq \frac{1}{90} h^4 (b-a) \|f^{(4)}\|_\infty \ .$$

□

Umschreiben der Fehlerabschätzung aus Satz 2.6 in die Anzahl $N^{(2)} = 2n+1 \leq 3n$ der f-Auswertungen von $S_n^{(2)}$ liefert wegen $h = (b-a)/n \leq 3(b-a)/N^{(2)}$ die Abschätzung

$$|I(f) - S_n^{(2)}(f)| \leq \frac{9}{10} (b-a)^5 \|f^{(4)}\|_\infty \big(N^{(2)}\big)^{-4} \ .$$

Die summierte Simpson-Regel ist also für jedes $f \in C^4[a,b]$ von *vierter Ordnung*. Der Aufwand $N^{(2)}(\varepsilon)$ zur Reduktion des Diskretisierungsfehlers unter eine gegebene Schranke $\varepsilon > 0$ ist dann

$$N^{(2)}(\varepsilon) = \left[\left(\frac{9(b-a)^5}{10} \|f^{(4)}\|_\infty \right)^{\frac{1}{4}} \varepsilon^{-\frac{1}{4}} \right] \ . \tag{2.20}$$

Für eine gegebene Funktion $f \in C^4[a,b]$ wollen wir die Effizienz von summierter Trapezregel und summierter Simpson-Regel vergleichen. Bei vorgegebener Genauigkeit ε^{-1} verhält sich die Effizienz $E^{(1)}$ der summierten Trapezregel zur Effizienz $E^{(2)}$ der summierten Simpson-Regel wie $N^{(2)}(\varepsilon)$ zu $N^{(1)}(\varepsilon)$. Nach Einsetzen von (2.18) und (2.20) bringen wir mit der Beobachtung $[x] = x - \xi$, $0 \leq \xi < 1$, die Gauß–Klammern wieder zum Verschwinden und erhalten nach etwas Bruchrechnung

$$\frac{E^{(1)}}{E^{(2)}} = \frac{N^{(2)}(\varepsilon)}{N^{(1)}(\varepsilon)} = \left(\frac{81}{10} \frac{1}{(b-a)} \frac{\|f^{(4)}\|_\infty}{\|f''\|_\infty^2} \right)^{\frac{1}{4}} \varepsilon^{\frac{1}{4}} + \mathcal{O}(\varepsilon^{\frac{1}{2}}) \ .$$

Für hohe Genauigkeiten, also kleine Diskretisierungsfehler ε, ist somit die summierte Simpson-Regel vorzuziehen. (Warum?) Andererseits kann für festes ε auch die summierte Trapezregel besser sein, falls $\|f^{(4)}\|_\infty \gg \|f''\|_\infty^2$ gilt.

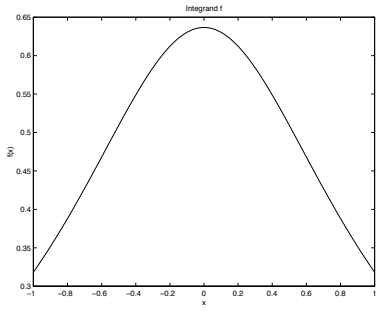
Die Herleitung von Fehlerabschätzungen für summierte Versionen von Newton-Côtes-Formeln höherer Ordnung sei dem Leser überlassen. Wie passt sich die Riemannsche Summe in den Rahmen der summierten Newton-Côtes-Formeln ein?

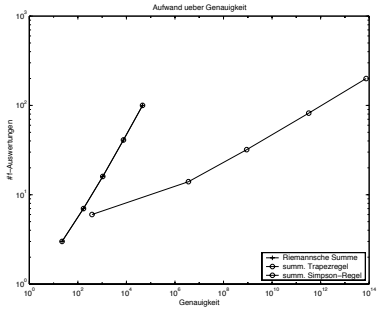
Beispiel. Es gilt für jedes $\alpha \neq 0$

$$\frac{1}{2\arctan(\alpha)} \int_{-1}^{1} \frac{\alpha}{1+\alpha^2 x^2}\,dx = 1 \ .$$

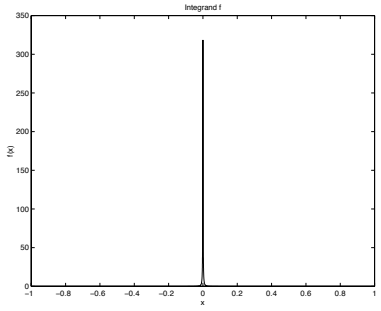
Wir wollen an diesem Beispiel Riemannsche Summe, summierte Trapezregel und summierte Simpson-Regel vergleichen.

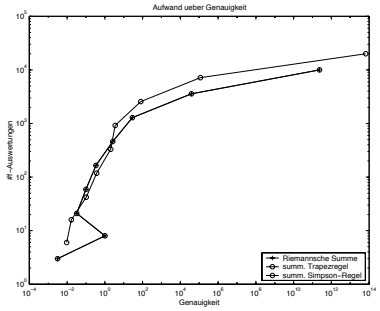
Als erstes betrachten wir den Fall $\alpha = 1$. Während der Integrand links zu sehen ist, zeigt die rechte Abbildung den benötigten Aufwand über der erreichten Genauigkeit. Der lineare Verlauf der Kurven spiegelt die theoretischen Resultate der Sätze 2.5 und 2.6 wieder. Daß Riemannsche Summe und summierte Trapezregel gleichauf liegen, liegt an der Symmetrie des Integranden. Mit 100 f-Auswertungen erreicht die summierte Simpson-Regel einen Fehler von 10^{-12}, wohingegen Riemannsche Summe und summierte Trapezregel nur bei 10^{-4} landen. Wie erwartet schneidet also die summierte Simpson-Regel am besten ab. Ein Vergleich mit der summierten Milne-Regel wäre interessant.





Wie auf der nächsten Abbildung links zu sehen ist, hat die Wahl von $\alpha = 1000$ nun zur Folge, daß unser Integrand eine starke Variation in der Nähe von 0 aufweist, im übrigen aber fast konstant ist. Die Maximumsnorm der k-ten Ableitung wächst mit α^k. Damit wird die höhere Ordnung der summierten Simpson-Regel durch den vergrößerten Vorfaktor wieder zunichte gemacht. Auf dem rechten Bild sieht man, daß tatsächlich alle drei Quadraturformeln gleich schlecht arbeiten. Um einen Fehler von 10^{-12} zu erreichen sind 10 000 f-Auswertungen nötig.





Weiterführende Fragen.

- Kann man durch geschickte Wahl der $n + 1$ Stützstellen erreichen, daß Quadraturformeln der Gestalt (2.12) sogar für $f \in P_{2n+1}$ exakt sind (Stichwort: Gauß-Christoffel-Quadratur)?
- Kann man die Zerlegung V_k, $k = 0, \ldots, n-1$ automatisch an die Funktion anpassen und so einen kleineren Diskretisierungsfehler erreichen, als im äquidistanten Fall (Stichwort: adaptive Romberg-Quadratur)?

- Kann man bei freier Wahl der Zerlegung die Bedingung $f \in C^4[a,b]$ in Satz 2.6 abschwächen? Wie sehen schwächere Bedingungen aus (Stichwort: Approximationstheorie)?

Wieder verweisen wir auf weiterführende Vorlesungen und die Literatur, z.B. die Lehrbücher von Deuflhard und Hohmann [1] oder Hämmerlin und Hoffmann [2].

Literatur

[1] P. Deuflhard und A. Hohmann. *Numerische Mathematik I.* de Gruyter, 1993. Wer außer den Newton-Côtes- Formeln noch Antworten auf die obigen weiterführenden Fragen sucht, kann in Kapitel 9 nachschauen (oder die Vorlesung 'Einführung in die Numerische Mathematik' besuchen).

[2] G. Hämmerlin und K.-H. Hoffmann. *Numerische Mathematik.* Springer, 3. Auflage, 1992. Natürlich kommen die Newton-Côtes-Formeln auch hier zur Sprache. Interessant sind insbesondere die zahlreichen historischen Bemerkungen.

3 Lineare gewöhnliche Differentialgleichungen

3.1 Motivation

Um den naturwissenschaftlichen Hintergrund von Differentialgleichungen zumindest anzudeuten, geben wir hier zwei einfache Beispiele. Darüberhinaus kann man sich beispielsweise im Anfangskapitel von Deuflhard und Bornemann [1] über Anwendungen in der Newton'schen Himmelsmechanik, der klassischen Moleküldynamik und der Schaltkreissimulation informieren. Ein Beispiel aus der chemischen Reaktionskinetik, welches auch diesem Buch entnommen wurde, findet sich im Anhang zu diesem Kapitel.

3.1.1 Bewegung eines Teilchens

Gegeben sei ein Teilchen mit gegebener *konstanter Geschwindigkeit* $v_0 \in \mathbb{R}$ in Richtung der x-Achse. Gesucht ist der

$$x(t): \text{ Aufenthaltsort des Teilchens zum Zeitpunkt } t \geq 0 \, .$$

Bekanntlich ist

$$v_0 = \lim_{\Delta t \to 0} \frac{x(t+\Delta t) - x(t)}{\Delta t} = x'(t) \, .$$

Damit ist also $x(t)$ eine Lösung der Differentialgleichung

$$x'(t) = v_0 \, .$$

Offenbar ist

$$x(t) = v_0 t + \alpha$$

für jedes $\alpha \in \mathbb{R}$ eine Lösung. Um die Eindeutigkeit der Lösung zu sichern, müssen wir den Aufenthaltsort des Teilchens zu einem festen Zeitpunkt kennen, etwa für $t = 0$. Das resultierende Anfangswertproblem (AWP)

$$x'(t) = v_0 \quad \forall t > 0, \qquad x(0) = x_0 \tag{3.21}$$

hat dann die eindeutig bestimmte Lösung

$$x(t) = v_0 t + x_0 \, .$$

Dabei folgt die Eindeutigkeit z.B. aus dem Hauptsatz der Differential- und Integralrechnung (Übung).

Hängt die Geschwindigkeit

$$v = v(x,t)$$

vom aktuellen Aufenthaltsort x und vom Zeitpunkt t ab, so erhält man eine Differentialgleichung der Form

$$x'(t) = v(x(t), t) \, .$$

Kann sich das Teilchen im gesamten Raum $\mathbb{R}^3$ bewegen, so hat man den zugehörigen Koordinatenvektor $\vec{x}(t) \in \mathbb{R}^3$ zu bestimmen. In diesem Fall ist auch die Geschwindigkeit $\vec{v}(\vec{x}, t) \in \mathbb{R}^3$ ein Vektor, und es gilt wieder

$$\vec{x}'(t) = \vec{v}(\vec{x}(t), t) \, .$$

Das ist ein System nichtlinearer Differentialgleichungen. Aus der physikalischen Anschauung heraus erwarten wir, daß auch in diesem Fall eine Anfangsbedingung

$$\vec{x}(0) = \vec{x}_0$$

mit gegebenem $\vec{x}_0 \in \mathbb{R}^3$ notwendig ist, um Eindeutigkeit zu gewährleisten.

3.1.2 Radioaktiver Zerfall

Gegeben sei die Anzahl $x_0 \in \mathbb{R}$ der Atome eines radioaktiven Materials zum Zeitpunkt $t = 0$. Gesucht ist

$$x(t): \text{ Anzahl der Atome zum Zeitpunkt } t \geq 0 \, .$$

Es sei

$$p\Delta t$$

die Wahrscheinlichkeit, daß ein Atom während eines „kleinen“ Zeitraumes Δt zerfällt. Die Zerfallskonstante p ist materialabhängig. Von $x(t)$ Teilchen bleiben dann zum Zeitpunkt $t + \Delta t$ noch $x(t) - p\Delta t \, x(t)$ Teilchen übrig. Also gilt

$$x(t + \Delta t) = x(t) - p\Delta t \, x(t)$$

oder gleichbedeutend ($\Delta t > 0$)

$$\frac{x(t + \Delta t) - x(t)}{\Delta t} = -px(t) \, .$$

Grenzübergang $\Delta t \to 0$ liefert die Differentialgleichung

$$x'(t) = -px(t) \quad t > 0 \, . \tag{3.22}$$

Man beachte, daß dieser Grenzübergang und damit auch die resultierende Differentialgleichung (3.22) nur für *reellwertige Funktionen* $x(t)$ sinnvoll ist! Erst nach Rundung von $x(t)$ auf ganze Zahlen werden Lösungen von (3.22) im Sinne unserer ursprünglichen Aufgabenstellung interpretierbar. Man nimmt dabei (oft stillschweigend) an, daß der damit verbundene *Modellierungsfehler* genügend klein ist. Diese Annahme heißt Kontinuumshypothese.

Offenbar ist die Kontinuumshypothese vertretbar, wenn

$$x(t) >> 1$$

vorliegt. Die Differentialgleichung (3.22) bildet dann zusammen mit der Anfangsbedingung

$$x(0) = x_0$$

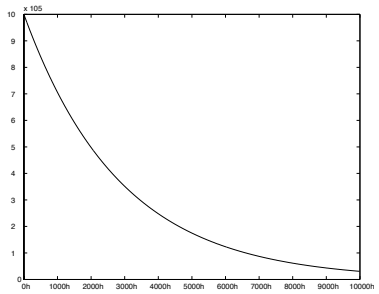
ein Anfangswertproblem (AWP) zur Bestimmung von x.

Durch Einsetzen bestätigt man, daß

$$x(t) = x_0 e^{-pt} \tag{3.23}$$

eine Lösung dieses AWPs ist. Wir werden später sehen, daß es keine weiteren Lösungen gibt.

Beispiel. Im Falle von C^{14} ist $p = p_{C^{14}} \approx 0,000348675/\text{Jahr}$. Durch Einsetzen in (3.23) erfährt man, daß beispielsweise nach 10.000 Jahren die Anzahl der C^{14}-Isotope von 1.000.000 auf 30.696 gefallen ist.



Die gesamte Evolution ist in der obigen Abbildung zu sehen.

Wenn, wie in unserem Beispiel, $x(0) >> 1$ gilt, ist die Kontinuumshypothese offenbar zu Beginn des Zerfallsprozesses vertretbar. Da aber $x(t) \to 0$ für $t \to \infty$ gilt, führt die Evolution zwangsläufig aus dem Gültigkeitsbereich unseres Modells heraus. Von einem gewissen Zeitpunkt ab werden also die Resultate, obwohl mathematisch richtig, *physikalisch fragwürdig*.

3.2 Lineare Differentialgleichungen 1. Ordnung

3.2.1 Existenz, Eindeutigkeit, Kondition

Wir betrachten zunächst die Differentialgleichung

$$x'(t) = \lambda x(t) \qquad t > 0 \tag{3.24}$$

mit gegebenem $\lambda \in \mathbb{R}$. Die Differentialgleichung (3.24) heißt

- *gewöhnlich*, denn es treten nur Ableitungen nach einer Variablen auf,
- *1. Ordnung*, denn es treten nur Ableitungen höchstens 1. Ordnung auf,
- *linear*, denn Linearkombinationen von Lösungen sind wieder Lösung.

Aus der Linearität folgt, daß die *Nullfunktion* $x = 0$, oft auch $x \equiv 0$ geschrieben, d.h. die Funktion mit den Werten $x(t) = 0 \; \forall t > 0$, eine Lösung ist. Differentialgleichungen mit dieser Eigenschaft heißen *homogen*.

Offenbar ist

$$x = \alpha e^{\lambda t} \tag{3.25}$$

für jedes beliebige $\alpha \in \mathbb{R}$ eine Lösung von (3.24). Es stellt sich die Frage, ob es noch weitere Lösungen gibt. Eine Antwort gibt der folgende

Satz 3.1. *Alle Lösungen von* (3.24) *haben die Gestalt* (3.25).

Beweis. Sei $w(t)$ eine Lösung von (3.24), also

$$w' = \lambda w.$$

Dann gilt:

$$\frac{d}{dt}(we^{-\lambda t}) = w'e^{-\lambda t} - \lambda we^{-\lambda t} = \lambda we^{-\lambda t} - \lambda we^{-\lambda t} = 0 \quad \forall t > 0$$

Also muß

$$we^{-\lambda t} = \alpha \in \mathbb{R} \qquad \forall t > 0$$

gelten und es folgt die Behauptung. □

Bemerkung. Die Lösungen der linearen Differentialgleichung 1. Ordnung (3.24) bilden nach Satz 3.1 einen *eindimensionalen linearen Raum*, welcher von der Basisfunktion $\psi_1(t) = e^{\lambda t}$ aufgespannt wird.

Als nächstes betrachten wir ein etwas schwierigeres Problem, nämlich die Differentialgleichung

$$x'(t) = \lambda x(t) + f(t) \qquad t > 0 \tag{3.26}$$

mit einer gegebenen Funktion $f \in C[0, \infty)$. Ist $f \neq 0$, also nicht, wie in (3.24), die Nullfunktion, so ist die Nullfunktion $x = 0$ auch keine Lösung von (3.26). Die Differentialgleichung ist also i.a. *inhomogen*. Die Funktion f heißt in diesem Zusammenhang oft *Inhomogenität* oder *rechte Seite*.

Als erstes wollen wir nun die Lösungen von (3.26) bestimmen. Dazu machen wir den Ansatz

$$x(t) = \alpha(t)e^{\lambda t} \,,$$

der oft als *Variation der Konstanten* bezeichnet wird. Einsetzen in (3.26) liefert

$$\lambda x(t) + f(t) = x'(t) = \alpha'(t)e^{\lambda t} + \lambda\alpha(t)e^{\lambda t} = \alpha'(t)e^{\lambda t} + \lambda x(t),$$

und es folgt

$$\alpha'(t) = f(t)e^{-\lambda t} \qquad t > 0 \,.$$

Der Hauptsatz der Differential- und Integralrechnung ergibt nun

$$\alpha(t) = \alpha(0) + \int_0^t f(\eta)e^{-\lambda\eta}\, d\eta \,.$$

Satz 3.2. *Alle Lösungen von* (3.26) *haben die Gestalt*

$$x(t) = \alpha e^{\lambda t} + \int_0^t f(\eta)e^{\lambda(t-\eta)}\, d\eta \tag{3.27}$$

mit einer beliebigen Konstanten $\alpha \in \mathbb{R}$.

Beweis. Übung. □

Bemerkung. Die Lösungen von (3.26) bilden nach Satz 3.2 einen *eindimensionalen affinen Raum.*

Wir betrachten jetzt das zugehörige Anfangswertproblem

$$\begin{aligned} x'(t) &= \lambda x(t) + f(t) \qquad 0 < t \leq T \\ x(0) &= x_0 \, . \end{aligned} \tag{3.28}$$

Zunächst klären wir die Frage nach **Existenz** und **Eindeutigkeit** einer Lösung.

Satz 3.3. *Das Anfangswertproblem* (3.28) *hat die eindeutig bestimmte Lösung*

$$x(t) = x_0 e^{\lambda t} + \int_0^t f(\eta) e^{\lambda(t-\eta)} \, d\eta \, . \tag{3.29}$$

Beweis. Jede Lösung der Differentialgleichung (3.26) hat die Gestalt (3.27). Einsetzen von $t = 0$ liefert $x_0 = x(0) = \alpha \cdot 1 + 0 = \alpha$. □

Als nächstes stellt sich die Frage, wie Störungen des Anfangswertes x_0 und der rechten Seite f sich auf die Lösung $x(t)$ auswirken. Wir untersuchen also die **Kondition** von (3.28). Dabei beginnen wir mit dem homogenen Fall, setzen also $f = 0$ voraus.

Satz 3.4. *Seien x, $\tilde{x}$ die Lösungen des AWPs* (3.28) *zu den Anfangswerten x_0, $\tilde{x}_0 \in \mathbb{R}$ und übereinstimmenden rechten Seiten $f = \tilde{f}$.*

Dann gilt im Fall $\lambda < 0$

$$\|x - \tilde{x}\|_\infty = |x_0 - \tilde{x}_0| \tag{3.30}$$

und im Fall $\lambda \geq 0$

$$\|x - \tilde{x}\|_\infty = e^{\lambda T} |x_0 - \tilde{x}_0| \, . \tag{3.31}$$

Beweis. Aus Satz 3.3 folgt

$$x(t) = x_0 e^{\lambda t} + \int_0^t f(\eta) e^{\lambda(t-\eta)} \, d\eta, \qquad \tilde{x}(t) = \tilde{x}_0 e^{\lambda t} + \int_0^t f(\eta) e^{\lambda(t-\eta)} \, d\eta \, .$$

Einsetzen ergibt

$$\|x - \tilde{x}\|_\infty = \max_{t \in [0,T]} e^{\lambda t} |x_0 - \tilde{x}_0|$$

und damit die Behauptung. □

Nun betrachten wir das AWP (3.28) in seiner vollen Allgemeinheit. Insbesondere lassen wir jetzt auch Störungen der rechten Seite $f \in C[0,T]$ zu.

Satz 3.5. *Seien x, $\tilde{x}$ die Lösungen des AWPs* (3.28) *zu den Anfangswerten x_0, $\tilde{x}_0 \in \mathbb{R}$ sowie den rechten Seiten f, $\tilde{f} \in C[0,T]$.*

Dann gilt im Falle $\lambda < 0$

$$\|x - \tilde{x}\|_\infty \leq (1 + T) \max \left\{ |x_0 - \tilde{x}_0|, \|f - \tilde{f}\|_\infty \right\} \tag{3.32}$$

und im Falle $\lambda \geq 0$

$$\|x - \tilde{x}\|_\infty \leq (1+T)e^{\lambda T} \max\left\{|x_0 - \tilde{x}_0|, \|f - \tilde{f}\|_\infty\right\} . \tag{3.33}$$

Beweis. Einsetzen der Lösungsdarstellung (3.29) liefert mit Dreiecksungleichung und Homogenität des Betrags die Abschätzung

$$\|x - \tilde{x}\|_\infty \leq \max_{t\in[0,T]} \left(e^{\lambda t}|x_0 - \tilde{x}_0| + \int_0^t |f(\eta) - \tilde{f}(\eta)| e^{\lambda(t-\eta)}\, d\eta \right)$$

und aus

$$\int_0^t |f(\eta) - \tilde{f}(\eta)| e^{\lambda(t-\eta)}\, d\eta \leq T \max_{t\in[0,T]} e^{\lambda t}\, \|f - \tilde{f}\|_\infty$$

folgt die Behauptung. □

Aus Satz 3.4 und Satz 3.5 gewinnt man direkt die folgenden Abschätzungen der absoluten Kondition $\kappa_{abs}(\text{AWP})$ des AWPs (3.28). Im Falle $\lambda \leq 0$ gilt nämlich

$$1 \leq \kappa_{abs}(\text{AWP}) \leq 1 + T \tag{3.34}$$

und im Falle $\lambda > 0$ erhält man

$$e^{\lambda T} \leq \kappa_{abs}(\text{AWP}) \leq (1+T)e^{\lambda T} . \tag{3.35}$$

Während für $\lambda > 0$ eine *exponentielle Fehlerverstärkung* eintritt, ist im Vergleich dazu das AWP im Falle $\lambda \leq 0$ *gut konditioniert*.

Bemerkung. Nach Existenz und Eindeutigkeit aus Satz 3.3 sichert Satz 3.5 die stetige Abhängigkeit der Lösung x (bezüglich der Norm $\|\cdot\|_\infty$) von den Anfangswerten (bezüglich der Norm $|\cdot|$) und der rechten Seite (bezüglich der Norm $\|\cdot\|_\infty$). Damit ist das AWP (3.28) *korrekt gestellt* (im Sinne von Hadamard).

3.2.2 Euler-Verfahren

Wir wollen nun numerische Verfahren zur näherungsweisen Lösung des AWPs (3.28) mit der exakten Lösung (3.29) konstruieren und analysieren. Dazu wählen wir zunächst ein Gitter

$$\Delta = \{0 = t_0 < t_1 < \cdots < t_n = T\} .$$

Der Einfachheit halber sei die Schrittweite

$$\tau = t_{k+1} - t_k, \qquad k = 0, \ldots, n-1 ,$$

konstant. In diesem Fall spricht man von einem *äquidistanten Gitter*. Unter der Annahme, daß die Differentialgleichung sich auf $t = 0$ fortsetzen lässt, ist die Tangente $g(\eta)$ an $x(t)$ in $t_0 = 0$ gegeben durch

$$g(\eta) = x_0 + \eta x'(0) = x_0 + \eta\big(\lambda x_0 + f(t_0)\big) .$$

Einsetzen von $\eta = \tau$ liefert die Näherungslösung

$$x_1 = x_0 + \tau\big(\lambda x_0 + f(t_0)\big) \approx x(t_1) \ .$$

Indem wir auf die gleiche Weise fortfahren, erhalten wir das Euler'sche Polygonzugverfahren

$$x_{k+1} = x_k + \tau\big(\lambda x_k + f(t_k)\big), \quad k = 0, \ldots, n-1 \ , \tag{3.36}$$

zur Berechnung der *Gitterfunktion*

$$x_\Delta(t_k) = x_k, \quad k = 0, \ldots, n \ ,$$

aus der Anfangsbedingung $x_0 \in \mathbb{R}$. Da (3.36) eine *explizite* Vorschrift zur Berechnung von x_{k+1} liefert, nennt man dieses Verfahren auch explizites Euler-Verfahren.

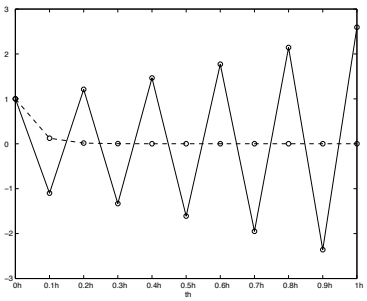
Das explizite Euler-Verfahren lässt sich auch als Diskretisierung der Ableitung $x'(t)$ durch den vorwärtsgenommenen Differenzenquotienten interpretieren. Auf diese Weise erhält man nämlich

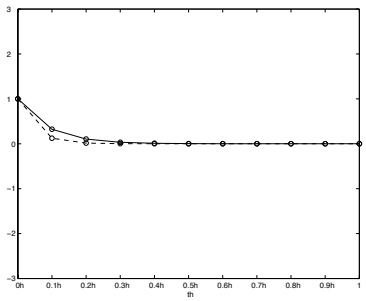
$$\frac{x_{k+1} - x_k}{\tau} = \lambda x_k + f(t_k) \ , \quad k = 0, \ldots, n-1 \ ,$$

und Auflösen nach x_{k+1} liefert (3.36). Verwendet man stattdessen den rückwärtsgenommenen Differenzenquotienten, so ergibt sich das implizite Euler-Verfahren

$$x_{k+1} = x_k + \tau(\lambda x_{k+1} + f(t_{k+1})), \quad k = 0, \ldots, n-1 \ . \tag{3.37}$$

Beispiel. Wir betrachten das AWP (3.28) im homogenen Fall $f = 0$ mit $\lambda = -21$, $x_0 = 1$ und $T = 1$. Zur Diskretisierung wählen wir ein äquidistantes Gitter Δ zur Schrittweite $\tau = 1/10$. Die mittels explizitem und implizitem Euler-Verfahren berechneten Näherungslösungen sind zusammen mit der exakten Lösung in der folgenden linken (explizit) und rechten (implizit) Abbildung zu sehen.





Im Gegensatz zum impliziten Euler-Verfahren liefert das explizite Verfahren völlig unbrauchbare Ergebnisse.

Um das unterschiedliche Verhalten der beiden Verfahren zu verstehen, untersuchen wir die Auswirkung von Störungen des Anfangswerts und der rechten Seite auf die berechnete Gitterfunktion. Es geht also um die **diskrete Kondition**. Wie im kontinuierlichen Fall beginnen wir mit $f = 0$.

Satz 3.6. *Seien x_Δ, $\tilde{x}_\Delta$ die mit dem expliziten Euler-Verfahren* (3.36) *berechneten Näherungslösungen zu den Anfangswerten x_0, $\tilde{x}_0 \in \mathbb{R}$ und übereinstimmenden diskreten rechten Seite $f_\Delta = \tilde{f}_\Delta$.*

Dann gilt im Falle $\lambda < 0$ die Abschätzung

$$\|x_\Delta - \tilde{x}_\Delta\|_\infty \leq |x_0 - \tilde{x}_0| \tag{3.38}$$

genau dann, wenn die Schrittweite τ die Stabilitätsbedingung

$$0 < \tau \leq \frac{2}{|\lambda|} \tag{3.39}$$

erfüllt.

Im Falle $\lambda \geq 0$ gilt für jede beliebige Schrittweite $\tau > 0$ die Abschätzung

$$\|x_\Delta - \tilde{x}_\Delta\|_\infty \leq e^{\lambda T}|x_0 - \tilde{x}_0| \; . \tag{3.40}$$

Beweis. Die Differenz $d_k = x_k - \tilde{x}_k$ genügt der Gleichung

$$d_{k+1} = (1+\lambda)d_k \; , \qquad k = 0, \ldots, n-1 \; .$$

Mit vollständiger Induktion bestätigt man leicht die Lösungsdarstellung

$$x_k - \tilde{x}_k = d_k = d_0(1+\tau\lambda)^k = (x_0 - \tilde{x}_0)(1+\tau\lambda)^k, \quad k = 0, \ldots, n \; . \tag{3.41}$$

Maximumbildung ergibt

$$\|x_\Delta - \tilde{x}_\Delta\|_\infty = \max_{k=0,\ldots,n} |1+\tau\lambda|^k |x_0 - \tilde{x}_0| \; .$$

Im Falle $\lambda < 0$ folgt nun die Behauptung aus

$$|1+\tau\lambda| \leq 1 \quad \Longleftrightarrow \quad -1 \leq 1+\tau\lambda \quad \Longleftrightarrow \quad \tau \leq \frac{-2}{\lambda} \; .$$

Im Falle $\lambda \geq 0$ erhält man aus der Potenzreihenentwicklung der Exponentialfunktion die Abschätzung

$$|1+\tau\lambda|^k = (1+\tau\lambda)^k \leq \left(1 + \tau\lambda + \frac{(\tau\lambda)^2}{2!} + \frac{(\tau\lambda)^3}{3!} + \ldots\right)^k = e^{\lambda t_k} \leq e^{\lambda T}$$

und damit die Behauptung. □

In Verbindung mit Satz 3.4 folgt aus Satz 3.6, daß im homogenen Fall $f = 0$ die diskrete Kondition κ(exEuler) des expliziten Euler-Verfahrens durch die Kondition κ(AWP) des AWPs abgeschätzt werden kann. Es gilt also

$$\kappa(\text{exEuler}) \leq \kappa(\text{AWP}) \; .$$

Im Vergleich zum kontinuierlichen Problem findet also *keine zusätzliche Fehlerverstärkung* statt. Das explizite Euler-Verfahren ist damit stabil. Im Falle $\lambda < 0$ ist die Stabilität allerdings an die Bedingung (3.39) geknüpft. Ist diese verletzt, also $|1+\tau\lambda| > 1$, so folgt

$$\|x_\Delta - \tilde{x}_\Delta\|_\infty = \sigma|x_0 - \tilde{x}_0| \tag{3.42}$$

mit dem Verstärkungsfaktor $\sigma = |1 + \tau\lambda|^n \gg 1$, also *Instabilität*.

Genau das ist in unserem Beispiel passiert: Angesichts von $\lambda = -21$ führt die Wahl von $\tau = 1/10$ auf $|\tau\lambda| = 21/10 > 2$ oder gleichbedeutend $|1 + \tau\lambda| = 1.1 > 1$. Damit haben wir nicht nur die zusätzliche Verstärkung von Fehlern in den Anfangswerten (3.42). Offenbar gilt

$$|x_k| = |1 + \tau\lambda|^k \to \infty \quad \text{für} \quad k \to \infty$$

im Gegensatz zu

$$x(t_k) \to 0 \quad \text{für} \quad t_k \to \infty \, .$$

Mit wachsendem k haben exakte Lösung und Näherung also nichts mehr miteinander zu tun. Problematisch sind dabei sicher *nicht irgendwelche Rundungsfehler*, z.B. bei der Eingabe des Anfangswertes. Stabilität scheint notwendig für die *Konvergenz des Verfahrens*!

Bevor wir diesen Zusammenhang näher untersuchen, betrachten wir noch den inhomogenen Fall.

Satz 3.7. *Seien x_Δ, $\tilde{x}_\Delta$ die mit dem expliziten Euler-Verfahren* (3.36) *berechneten Näherungslösungen zu den Anfangswerten x_0, $\tilde{x}_0 \in \mathbb{R}$ sowie den diskreten rechten Seiten f_Δ, $\tilde{f}_\Delta$ mit $f_\Delta(t_k) = f_k$, $\tilde{f}_\Delta(t_k) = \tilde{f}_k$, $k = 0, \ldots, n-1$.*

Dann gilt im Falle $\lambda < 0$

$$\|x_\Delta - \tilde{x}_\Delta\|_\infty = (1 + T) \max \left\{ |x_0 - \tilde{x}_0|, \|f_\Delta - \tilde{f}_\Delta\|_\infty \right\} \, , \tag{3.43}$$

falls die Schrittweite τ die Stabilitätsbedingung (3.39) *erfüllt.*

Im Falle $\lambda \geq 0$ gilt für jede beliebige Schrittweite $\tau > 0$

$$\|x_\Delta - \tilde{x}_\Delta\|_\infty \leq (1 + T) e^{\lambda T} \max \left\{ |x_0 - \tilde{x}_0|, \|f_\Delta - \tilde{f}_\Delta\|_\infty \right\} \, . \tag{3.44}$$

Beweis. In Analogie zu Satz 3.3 wollen wir zunächst eine geschlossene Darstellung der Gitterfunktion x_Δ in Abhängigkeit von Anfangsbedingung und rechter Seite f_Δ herleiten. Ausgehend von der diskreten Lösung (3.41) für $f_\Delta = 0$ machen wir für $f_\Delta \neq 0$ den Ansatz (Variation der Konstanten!)

$$x_k = \alpha_k (1 + \tau\lambda)^k, \quad k = 0, \ldots, n,$$

und haben nun die Koeffizienten α_k zu bestimmen. Einsetzen in (3.36) liefert

$$\alpha_{k+1}(1 + \tau\lambda)^{k+1} = x_{k+1} = x_k + \tau(\lambda x_k + f_k) = \alpha_k (1 + \tau\lambda)^{k+1} + \tau f_k \, .$$

Ist $1 + \tau\lambda \neq 0$, so ergibt Division durch $(1 + \tau\lambda)^{k+1}$ die folgende Rekursion für die α_k

$$\alpha_{k+1} = \alpha_k + \tau f_k (1 + \tau\lambda)^{-(k+1)}$$

mit der Lösung

$$\alpha_k = \alpha_0 + \tau \sum_{j=0}^{k-1} f_j (1 + \tau\lambda)^{-(j+1)}$$

und aus der Anfangsbedingung folgt

$$x_0 = \alpha_0 (1 + \tau\lambda)^0 = \alpha_0 \, .$$

Einsetzen in unseren Ansatz liefert die gesuchte Darstellung

$$x_k = x_0(1+\tau\lambda)^k + \tau \sum_{j=0}^{k-1} f_j (1+\tau\lambda)^{k-(j+1)} \; . \tag{3.45}$$

Mit der Vereinbarung $(1+\tau\lambda)^0 = 1 \; \forall 1+\tau\lambda \in \mathbb{R}$, ist diese Darstellung auch im Falle $1+\tau\lambda$ richtig.

Nun können wir die behaupteten Abschätzungen beweisen. Einsetzen der Lösungsdarstellungen (3.45) liefert

$$\max_{k=0,\ldots,n} |x_k - \tilde{x}_k| \leq \max_{k=0,\ldots,n} |1+\tau\lambda|^k |x_0 - \tilde{x}_0| + T \max_{k=0,\ldots,n-1} |1+\tau\lambda|^k \max_{j=0,\ldots,n-1} |f_k - \tilde{f}_k|$$

und es folgt

$$\|x_\Delta - \tilde{x}_\Delta\|_\infty \leq (1+T) \max_{k=0,\ldots,n} |1+\tau\lambda|^k \max \left\{ |x_0 - \tilde{x}_0|, \|f_\Delta - \tilde{f}_\Delta\|_\infty \right\} \; .$$

Die Abschätzung von $|1+\tau\lambda|^k$ haben wir bereits im Beweis zu Satz 3.6 durchgeführt. □

In Verbindung mit Satz 3.6 gewinnt man aus Satz 3.7 direkt die folgenden Abschätzungen der diskreten absoluten Kondition $\kappa_{abs}(\text{exEuler})$ des Euler-Verfahrens (3.36). Im Falle $\lambda < 0$ gilt nämlich

$$1 \leq \kappa_{abs}(\text{exEuler}) \leq 1 + T \tag{3.46}$$

und im Falle $\lambda \geq 0$ erhält man

$$e^{\lambda T} \leq \kappa_{abs}(\text{exEuler}) \leq (1+T) e^{\lambda T} \; . \tag{3.47}$$

Diese Abschätzungen der diskreten Kondition sind identisch mit den Abschätzungen (3.46) und (3.47) im kontinuierlichen Fall. Daraus folgt

$$\kappa_{abs}(\text{exEuler}) = \sigma \kappa_{abs}(\text{AWP})$$

mit einem zusätzlichen Fehlerverstärkungsfaktor

$$\sigma \leq 1 + T \; .$$

Im Vergleich zum exponentiellen Term $e^{\lambda T}$ betrachtet man $1 + T \approx 1$ als klein. Das explizite Euler-Verfahren ist also auch im inhomogenen Fall stabil.

Im Falle $\lambda < 0$ ist dazu leider die Stabilitätsbedingung (3.39) erforderlich. In Anbetracht der Tatsache, daß gerade für $|\lambda| \gg 1$ die exakte Lösung $x(t) = x_0 e^{\lambda t}$ bald kaum noch von Null zu unterscheiden und damit praktisch konstant ist, kann die Stabilitätsbedingung unsinnig kleine Schrittweiten τ notwendig machen.

Vor diesem Hintergrund untersuchen wir nun das implizite Euler-Verfahren (3.37) und beschränken uns dabei auf den kritischen Fall $\lambda \leq 0$. Wie zuvor beginnen wir mit $f = 0$.

Satz 3.8. *Seien x_Δ, $\tilde{x}_\Delta$ die mit dem impliziten Euler-Verfahren* (3.37) *berechneten Näherungslösungen zu den Anfangswerten x_0, $\tilde{x}_0 \in \mathbb{R}$ und der diskreten rechten Seite $f_\Delta = 0$.*

Ist $\lambda \leq 0$, so gilt für jede beliebige Schrittweite $\tau > 0$

$$\|x_\Delta - \tilde{x}_\Delta\|_\infty \leq |x_0 - \tilde{x}_0| \; . \tag{3.48}$$

Beweis. Mit vollständiger Induktion bestätigt man leicht die Lösungsdarstellung

$$x_k = x_0(1-\tau\lambda)^{-k}, \quad k = 0, \ldots, n\ .$$

Einsetzen ergibt

$$\|x_\Delta - \tilde{x}_\Delta\|_\infty = \max_{k=0,\ldots,n} |1-\tau\lambda|^{-k} |x_0 - \tilde{x}_0|\ .$$

Wegen $\lambda \leq 0$ gilt nun für jede beliebige Schrittweite $\tau > 0$

$$|1-\tau\lambda|^{-1} = (1-\tau\lambda)^{-1} \leq 1,$$

und es folgt die Behauptung. □

Im inhomogenen Fall sieht es genauso gut aus.

Satz 3.9. *Seien x_Δ, $\tilde{x}_\Delta$ die mit dem impliziten Euler-Verfahren* (3.37) *berechneten Näherungslösungen zu den Anfangswerten x_0, $\tilde{x}_0 \in \mathbb{R}$ sowie den diskreten rechten Seiten f_Δ, $\tilde{f}_\Delta$ mit $f_\Delta(t_k) = f_k$, $\tilde{f}_\Delta(t_k) = \tilde{f}_k$, $k = 1, \ldots, n$. Ist dann $\lambda \leq 0$, so gilt*

$$\|x_\Delta - \tilde{x}_\Delta\|_\infty \leq (1+T) \max\left\{ |x_0 - \tilde{x}_0|, \|f_\Delta - \tilde{f}_\Delta\|_\infty \right\}\ . \tag{3.49}$$

Beweis. Übung. □

In Verbindung mit Satz 3.8 können wir aus Satz 3.9 direkt die unbedingte Stabilität des impliziten Euler-Verfahrens im Falle $\lambda \leq 0$ ablesen, denn es gilt ja

$$1 \leq \kappa_{abs}(\text{imEuler}) \leq 1 + T \tag{3.50}$$

für beliebige Schrittweiten $\tau > 0$. Diese Aussage steht im Einklang mit unseren Beobachtungen im einführenden Beispiel.

Bemerkung. Im Falle $\lambda > 0$ gibt es beim impliziten Euler-Verfahren Probleme: Um Oszillationen zu vermeiden und auch um die Durchführbarkeit des Verfahrens zu sichern, muß die *Stabilitätsbedingung*

$$\tau < \frac{1}{\lambda}$$

erfüllt sein. Das war beim expliziten Euler-Verfahren nicht nötig.

Das Fazit unserer Stabilitätsbetrachtungen:

- Im Falle $\lambda > 0$ ist das explizite Euler-Verfahren unbedingt stabil und das implizite Euler-Verfahren nicht.
- Im Falle $\lambda < 0$ ist das implizite Euler-Verfahren unbedingt stabil und das explizite Euler-Verfahren nicht.

Die Stabilitätseigenschaften der beiden Verfahren sind also komplementär.

3.2.3 Konvergenz der Euler-Verfahren

Wir erwarten, daß immer kleinere Schrittweiten τ zu immer genaueren Approximationen der exakten Lösung führen.

Definition 3.10. Eine Folge von Gitterfunktionen x_Δ heißt konvergent gegen $x \in C[0,T]$, falls

$$\lim_{n\to\infty} \max_{k=0,\dots,n} |x(t_k) - x_k| = 0$$

und konvergent mit der Ordnung p, falls es ein $\tau_0 > 0$ und eine von τ unabhängige Konstante $C \in \mathbb{R}$ gibt, so daß gilt

$$\max_{k=0,\dots,n} |x(t_k) - x_k| \le C\,\tau^p \qquad \forall \tau < \tau_0\ . \tag{3.51}$$

Im Falle des expliziten Euler-Verfahrens (3.36) approximieren wir in jedem Schritt die exakte Lösung durch die Tangente in (t_k, x_k): Das bewirkt selbst im Falle $x_k = x(t_k)$ einen *zusätzlichen Fehler* im aktuellen k-ten Schritt. Aufgrund vorausgegangener Fehler ist aber i.a. $x_k \neq x(t_k)$: Dieser Eingangsfehler kann eventuell durch die weitere Rechnung verstärkt werden.

Der lokale Diskretisierungsfehler $|x(t_k) - x_k|$ an der Stelle t_k setzt sich also aus zwei Beiträgen zusammen:

- dem Fehler, der im aktuellen Zeitschritt zusätzlich gemacht wird,
- dem Fehler, der aus allen Fehlern in vorausgehenden Zeitschritten resultiert.

Wir wenden uns zunächst dem zusätzlichen Fehler im aktuellen Zeitschritt zu.

Definition 3.11. Es sei x die exakte Lösung des AWPs (3.28). Dann heisst der Ausdruck

$$\varepsilon(t_k, \tau) = x(t_k + \tau) - x(t_k) - \tau\,(\lambda x(t_k) + f(t_k)) \tag{3.52}$$

Konsistenzfehler des expliziten Euler-Verfahrens (3.36). Das explizite Euler-Verfahren (3.36) heisst konsistent mit der Konsistenzordnung p, falls es ein $\tau_0 > 0$ und eine von τ unabhängige Konstante $C \in \mathbb{R}$ gibt, so dass gilt:

$$\max_{k=0,\dots,n-1} |\varepsilon(t_k, \tau)| \le C\tau^{p+1} \qquad \forall \tau < \tau_0\ . \tag{3.53}$$

Satz 3.12. *Sei $x \in C^2[0,T]$. Dann ist das explizite Euler-Verfahren* (3.36) *konsistent mit der Ordnung $p = 1$.*

Beweis. Die *Taylor-Entwicklung* von x lautet

$$x(t_k + \tau) = x(t_k) + x'(t_k)\tau + \frac{1}{2}x''(\xi)\tau^2 \quad \text{mit } \xi \in (t_k, t_k + \tau).$$

Daher gilt

$$\begin{aligned}\varepsilon(t_k, \tau) &= (x(t_k + \tau) - x(t_k)) - \tau(\lambda x(t_k) + f(t_k)) \\ &= \tau\left(x'(t_k) - (\lambda x(t_k) + f(t_k))\right) + \tfrac{1}{2}x''(\xi)\tau^2 \\ &= \tfrac{1}{2}x''(\xi)\tau^2\ .\end{aligned}$$

Daraus folgt

$$\max_{k=0,\dots,n-1} |\varepsilon(t_k,\tau)| \le \tfrac{1}{2} \max_{t\in[0,T]} |x''(t)|\tau^2 \ .$$

Das ist gerade die gewünschte Abschätzung (3.53) mit $C = \frac{1}{2}\|x''\|_\infty$. □

Nun können wir die Konvergenz des expliziten Euler-Verfahrens (3.36) (genauer die Konvergenz der mit Hilfe des expliziten Euler-Verfahrens berechneten Gitterfunktion x_Δ) formulieren.

Satz 3.13. *Es sei $x \in C^2[0,T]$. Dann ist das explizite Euler-Verfahren* (3.36) *konvergent mit der Ordnung $p = 1$ und es gilt die a priori Abschätzung*

$$\max_{k=0,\dots,n} |x(t_k) - x_k| \le (1+T)e^{\lambda_+ T}\|x''\|_\infty \ h \ ,$$

wobei die Abkürzung $\lambda_+ = \max\{0,\lambda\}$ verwendet wurde.

Beweis. Mit Blick auf (3.52) genügt die Gitterfunktion x^*_Δ gegeben durch $x^*_k = x(t_k)$ der Rekursionsvorschrift

$$x^*_{k+1} = x^*_k + \tau\left(\lambda x^*_k + f(t_k) + \tau^{-1}\varepsilon(t_k,\tau)\right) \ , \qquad x^*_0 = x_0 \ .$$

Damit ist x^*_Δ Lösung des Euler-Verfahrens (3.36) mit der *gestörten rechten Seite*

$$\tilde f(t_k) = f(t_k) + \tau^{-1}\varepsilon(t_k,\tau), \quad k = 0,\dots,n-1 \ .$$

Aus Satz 3.7 folgt nun unmittelbar

$$\max_{k=0,\dots,n} |x(t_k) - x_k| \le (1+T) \max_{k=0,\dots,n-1} \tau^{-1}|\varepsilon(t_k,\tau)| \qquad \forall\tau,\ 0<\tau\le\tau_0 = \tfrac{2}{|\lambda|} \ ,$$

falls $\lambda < 0$ und

$$\max_{k=0,\dots,n} |x(t_k) - x_k| \le (1+T)e^{\lambda T} \max_{k=0,\dots,n-1} \tau^{-1}|\varepsilon(t_k,\tau)| \qquad \forall\tau<0 \ ,$$

falls $\lambda \ge 0$ vorliegt. In beiden Fällen liefert die Konsistenz aus Satz 3.12 die Behauptung. □

Die *Stabilität* aus Satz 3.7 zusammen mit der *Konsistenz* aus Satz 3.12 ergibt also die *Konvergenz* (3.51).

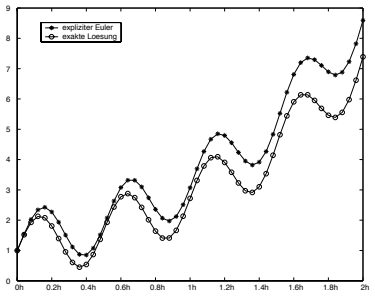
Bemerkung. Auch das implizite Euler-Verfahren (3.37) ist konsistent mit der Ordnung $p = 1$ (Übung). Genau wie oben folgt dann aus der Stabilität (siehe Satz 3.8 und anschließende Bemerkung) die Konvergenz des impliziten Euler-Verfahrens mit der Konvergenzordnung $p = 1$.

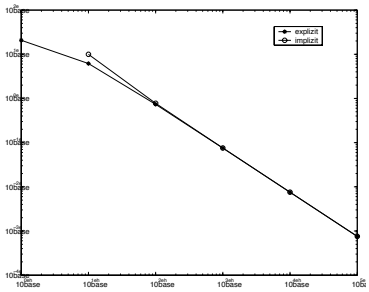
Beispiel. Um unsere theoretischen Konvergenzaussagen numerisch zu illustrieren, betrachten wir das AWP

$$x'(t) = x(t) + 4\pi\cos(4\pi t) - \sin(4\pi t) \ , \quad t\in(0,2] \ , \qquad x_0 = 1 \ ,$$

also gerade unser Problem (3.28) mit $\lambda = 1$ und $f(t) = 4\pi\cos(4\pi t) - \sin(4\pi t)$. Die exakte Lösung ist

$$x(t) = e^t + \sin(4\pi t) \quad t\in[0,2] \ .$$





Das linke Bild zeigt die mit dem expliziten Euler-Verfahren berechnete Näherungslösung zusammen mit der exakten Lösung zur Schrittweite $\tau = T/50$. Man beachte die sukzessive Fehlerverstärkung. Rechts ist der Diskretisierungsfehler $\|x - x_\Delta\|_\infty$ jeweils für explizites und implizites Verfahren in Abhängigkeit von der Anzahl n der Zeitschritte bzw. von der Schrittweite $\tau = T/n$ dargestellt. Für $n = 1$ ist das implizite Verfahren nicht durchführbar. Abgesehen davon sieht man aber auf der doppelt-logarithmischen Skala deutlich, daß eine Verzehnfachung der Zeitschritte n, also eine Reduktion der Schrittweite τ um den Faktor 10^{-1}, genau zu einer Reduktion des Fehlers um denselben Faktor 10^{-1} führt. Das bestätigt insbesondere unsere theoretischen Aussagen in Satz 3.12.

3.3 Systeme linearer Differentialgleichungen mit konstanten Koeffizienten

In diesem Abschnitt wollen wir lineare Systeme von n Differentialgleichungen behandeln. Wir werden sehen, daß durch geschickte Transformation alle wesentlichen Aussagen zu Existenz, Eindeutigkeit, Kondition, Stabilität und Konvergenz vom skalaren Fall auf den Systemfall übertragen werden können. Um dabei keine Langeweile aufkommen zu lassen und auch nicht den Blick aufs Wesentliche durch allzu kompliziert aussehende Formeln zu verstellen, wollen wir uns auf den homogenen Fall $f = 0$ beschränken. Erweiterungen auf den inhomogenen Fall sind eine gute Übung zur Vertiefung der Beweistechniken aus dem vorangegangenen Abschnitt 3.2.

3.3.1 Existenz, Eindeutigkeit, Kondition

Zum Aufwärmen betrachten wir das folgende gekoppelte System von zwei Differentialgleichungen

$$\begin{aligned} x_1'(t) &= a_{11}x_1(t) + a_{12}x_2(t) \\ x_2'(t) &= a_{21}x_1(t) + a_{22}x_2(t) \end{aligned}$$

mit den zwei unbekannten Funktionen x_1 und x_2. Setzt man

$$x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}, \qquad A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \in \mathbb{R}^{2,2},$$

so kann man das System einfacher in der Form

$$x'(t) = Ax(t) \tag{3.54}$$

als Gleichung von Vektoren aus $\mathbb{R}^2$ schreiben. Die Ableitung ist dabei komponentenweise zu verstehen.

Das System (3.54) ist linear, denn Linearkombinationen von Lösungen sind offenbar wieder Lösung. Die Menge V aller Lösungen ist also ein *linearer Raum*. Wie schon im skalaren Fall wollen wir eine Basis von V finden. Eine Basis des Lösungsraums eines linearen Systems von Differentialgleichungen heißt Fundamentalsystem.

Zunächst betrachten wir den Spezialfall einer *Diagonalmatrix* A, also

$$A = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}.$$

Dann reduziert sich (3.54) auf das *entkoppelte System*

$$\begin{aligned} x_1'(t) &= \lambda_1 x_1(t) \\ x_2'(t) &= \lambda_2 x_2(t)\ . \end{aligned} \tag{3.55}$$

Aus Abschnitt 3.2 wissen wir bereits, daß alle Lösungen der beiden Gleichungen (3.55) die Gestalt

$$x_1(t) = \alpha_1 e^{\lambda_1 t}\ , \quad x_2(t) = \alpha_2 e^{\lambda_2 t}$$

mit beliebigen α_1, $\alpha_2 \in \mathbb{R}$ haben. In Vektorschreibweise bedeutet das

$$x(t) = \begin{pmatrix} \alpha_1 e^{\lambda_1 t} \\ \alpha_2 e^{\lambda_2 t} \end{pmatrix} = \alpha_1 e^{\lambda_1 t} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \alpha_2 e^{\lambda_2 t} \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Damit hat in diesem Beispiel der Lösungsraum V die Dimension 2 und

$$\psi_1 = e^{\lambda_1 t} \begin{pmatrix} 1 \\ 0 \end{pmatrix}\ , \quad \psi_2 = e^{\lambda_2 t} \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

ist ein zugehöriges Fundamentalsystem.

Diese Aussage wollen wir verallgemeinern. Motiviert durch unsere bisherigen Erfahrungen machen wir den Lösungsansatz

$$x(t) = e^{\lambda t}\varphi$$

und versuchen aus (3.54) Bedingungen an $\lambda \in \mathbb{R}$ und $\varphi \in \mathbb{R}^2$ zu gewinnen. Einsetzen in (3.54) ergibt

$$\lambda e^{\lambda t}\varphi = x'(t) = Ax(t) = e^{\lambda t} A\varphi$$

und Division durch $e^{\lambda t} \neq 0$ liefert

$$A\varphi = \lambda\varphi\ .$$

Paare λ, φ mit dieser Eigenschaft haben einen Namen.

Definition 3.14. Sei $A \in \mathbb{R}^{m,m}$. Gilt dann

$$A\varphi = \lambda\varphi \tag{3.56}$$

für $\varphi \in \mathbb{R}^m$, $\varphi \neq 0$ und $\lambda \in \mathbb{R}$, so heisst φ Eigenvektor von A zum Eigenwert λ von A.

Anschaulich bedeutet (3.56), daß die Richtung eines Eigenvektors unter Multiplikation mit A invariant bleibt. Die Längenänderung ist gerade der zugehörige Eigenwert.

Beispiel. Wir betrachten das System (3.54) mit der Matrix

$$A = \begin{pmatrix} -1 & 3 \\ 3 & -1 \end{pmatrix}$$

Dann sind

$$\lambda_1 = 2, \quad \varphi_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} , \qquad \lambda_2 = -4, \quad \varphi_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix} ,$$

alle Paare von Eigenwerten und zugehörigen Eigenvektoren von A. Aus unserem Ansatz wissen wir, daß das Differentialgleichungssystem (3.54) dann die Lösungen

$$x = \alpha_1 e^{2t} \varphi_1 + \alpha_2 e^{-4t} \varphi_2$$

mit beliebigen $\alpha_1, \alpha_2 \in \mathbb{R}$ besitzt. Wir werden auf den nächsten Seiten beweisen, daß es keine weiteren Lösungen gibt. Der Lösungsraum V ist also wieder zweidimensional und

$$\psi_1(t) = e^{2t} \begin{pmatrix} 1 \\ 1 \end{pmatrix} , \qquad \psi_2(t) = e^{-4t} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

ist ein Fundamentalsystem.

Wir wollen untersuchen, unter welchen Bedingungen sich dieser Sachverhalt verallgemeinern lässt. Dazu betrachten wir jetzt das System

$$x'(t) = Ax(t) \tag{3.57}$$

mit

$$x(t) = \begin{pmatrix} x_1(t) \\ \vdots \\ x_m(t) \end{pmatrix} , \qquad A = \begin{pmatrix} a_{11} & \dots & a_{1m} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mm} \end{pmatrix} \in \mathbb{R}^{m,m} .$$

Dabei werden wir uns von nun an auf *symmetrische* Koeffizientenmatrizen A beschränken. Der Grund liegt in folgendem zentralen Satz aus der Linearen Algebra.

Satz 3.15. *Sei $A \in \mathbb{R}^{m,m}$ symmetrisch, d.h. $A^T = A$. Dann besitzt A orthonormale Eigenvektoren $\varphi_1, \dots, \varphi_m$ zu paarweise verschiedenen, reellen Eigenwerten $\lambda_1, \dots, \lambda_m$.*

Wir machen also im Folgenden die Voraussetzung

$$A \in \mathbb{R}^{m,m} , \quad \text{symmetrisch} .$$

Der Bequemlichkeit halber seien die Eigenwerte ihrer Größe nach geordnet, d.h.

$$\lambda_m < \lambda_{m-1} < \dots < \lambda_2 < \lambda_1 .$$

Wir definieren noch die Matrizen

$$T = (\varphi_1, \dots, \varphi_m) \in \mathbb{R}^{m,m} , \qquad D = \operatorname{diag}(\lambda_1, \dots, \lambda_m) \in \mathbb{R}^{m,m} .$$

Aus $A\varphi_k = \lambda_k \varphi_k$, $k = 1, \ldots, m$, folgt nämlich unmittelbar

$$\begin{aligned} AT = &\quad (A\varphi_1, ..., A\varphi_m) \\ = &\quad (\lambda_1\varphi_1, ..., \lambda_m\varphi_m) \quad = TD \end{aligned}$$

und somit

$$T^{-1}AT = T^{-1}TD = D\,, \qquad A = TDT^{-1}\,. \tag{3.58}$$

Wegen der Orthonormalität der Eigenvektoren $\varphi_1, \ldots, \varphi_m$ ist die Matrix T *unitär*, d.h. es gilt $T^{-1} = T^T$. Bezeichnet

$$|v|_2 = \left(\sum_{i=1}^{m} v_i^2 \right)^{\frac{1}{2}}\,, \qquad v = (v_1, \ldots, v_m)^T \in \mathbb{R}^m\,,$$

die Euklidische Vektornorm, so ist die zugehörige Euklidische Matrixnorm $|\cdot|_2$ bekanntlich definiert durch

$$|M|_2 = \sup_{\substack{v \in \mathbb{R}^m \\ v \neq 0}} \frac{|Mv|_2}{|v|_2}\,, \qquad M \in \mathbb{R}^{m,m}\,. \tag{3.59}$$

Unitäre Matrizen haben die Euklidische Norm 1. Insbesondere gilt

$$\kappa(T) = |T|_2 |T^{-1}|_2 = 1\,.$$

Nach diesen Vorbereitungen wenden wir uns nun wieder den Differentialgleichungen zu. Als erstes führen wir unser System (3.57) auf ein äquivalentes System mit *diagonaler* Koeffizientenmatrix zurück.

Lemma 3.16. *Die Koeffizientenmatrix $A \in \mathbb{R}^{m,m}$ sei symmetrisch. Dann ist die Funktion x genau dann eine Lösung von* (3.57)*, wenn die Funktion $y = T^{-1}x$ eine Lösung von*

$$y'(t) = Dy(t) \tag{3.60}$$

ist. Man nennt das System (3.60) *eine* Diagonalisierung *von* (3.57)*.*

Beweis. Sei x eine Lösung von (3.57), also $x' = Ax$. Wegen (3.58) folgt dann für $y := T^{-1}x$

$$y' = T^{-1}x' = T^{-1}Ax = T^{-1}ATT^{-1}x = Dy\,.$$

Sie umgekehrt y eine Lösung von (3.60), also $y' = Dy$. Dann folgt in gleicher Weise für $x := Ty$

$$x' = Ty' = TDy = TDT^{-1}Ty = Ax\,.$$

□

Nun können wir die Lösungen von (3.57) angeben.

Satz 3.17. *Die Koeffizientenmatrix $A \in \mathbb{R}^{m,m}$ sei symmetrisch. Dann haben alle Lösungen von* (3.57) *die Gestalt*

$$x(t) = \sum_{k=1}^{m} \alpha_k e^{\lambda_k t} \varphi_k \tag{3.61}$$

mit beliebigen Konstanten $\alpha_k \in \mathbb{R}$, $k = 1, \ldots, m$.

Beweis. Es sei x eine Lösung von (3.57), also $x' = Ax$. Dann ist nach Lemma 3.16 $y := T^{-1}x$ eine Lösung von (3.60), also $y' = Dy$. Damit gilt

$$y_k' = \lambda_k y$$

für $k = 1, \ldots, m$. Aus Satz 3.2 in Abschnitt 3.2 wissen wir, daß dann jeweils

$$y_k(t) = \alpha_k e^{\lambda_k t}$$

mit beliebigem $\alpha_k \in \mathbb{R}$ sein muss. Nach Definition des Matrix/Vektor-Produkts ist daher

$$x = Ty = \sum_{k=1}^{m} \alpha_k e^{\lambda_k t} \varphi_k$$

und das war die Behauptung. □

Bemerkung. Unter der Voraussetzung, daß A symmetrisch ist, ist der Lösungsraum V von (3.57) also m-dimensional und

$$\psi_k = e^{\lambda_k t}\varphi_k \ , \qquad k = 1, ..., m \ ,$$

ein(e) Fundamentalsystem (Basis).

Man kann zeigen, daß dim $V = m$ auch für beliebige $A \in \mathbb{R}^{m,m}$ und sogar im Falle variabler, stetiger Koeffizienten $a_{jk} = a_{jk}(t)$ richtig ist.

Wir betrachten nun das Anfangswertproblem (AWPS)

$$\begin{aligned} x'(t) &= Ax(t) \ , \quad 0 < t \leq T \ , \\ x(0) &= x_0 \ , \end{aligned} \tag{3.62}$$

mit gegebenem $x_0 \in \mathbb{R}^m$. Als erstes klären wir **Existenz** und **Eindeutigkeit**.

Satz 3.18. *Die Koeffizientenmatrix $A \in \mathbb{R}^{m,m}$ sei symmetrisch. Dann hat das Anfangswertproblem* (3.62) *die eindeutig bestimmte Lösung*

$$x(t) = \sum_{k=1}^{m} \alpha_k e^{\lambda_k t} \varphi_k \ . \tag{3.63}$$

Der Koeffizientenvektor $\alpha = (\alpha_1, \ldots, \alpha_m)^T \in \mathbb{R}^m$ ist dabei die eindeutig bestimmte Lösung des linearen Gleichungssystems

$$T\alpha = x_0. \tag{3.64}$$

Beweis. Alle Lösungen von $x' = Ax$ haben die Gestalt (3.63). Aus der Anfangsbedingung folgt

$$x_0 = x(0) = \sum_{k=1}^{m} \alpha_k \varphi_k = T\alpha \ .$$

T ist unitär und damit insbesondere regulär. Damit ist die Lösung α von (3.64) zu beliebigem $x_0 \in \mathbb{R}^m$ eindeutig bestimmt. □

Bemerkung. Definiert man für eine symmetrische Matrix A die *Matrix-Exponentielle* e^A durch

$$e^A := T \operatorname{diag}\left(e^{\lambda_1}, \ldots, e^{\lambda_m}\right) T^{-1} ,$$

so kann man (3.63) auch in der kompakten Form

$$x(t) = e^{tA} x_0$$

schreiben. In dieser Schreibweise hat die Lösung des inhomogenen Systems

$$x'(t) = Ax(t) + f(t) \quad t \in (0, T] , \qquad x(0) = x_0 \tag{3.65}$$

in direkter Analogie zu (3.29) die Gestalt

$$x(t) = e^{tA} x_0 + \int_0^t e^{(t-\eta)A} f(\eta)\, d\eta . \tag{3.66}$$

Die Integration ist dabei komponentenweise zu verstehen.

Wir wollen nun die Auswirkung von Störungen der Anfangsbedingungen x_0 auf die Lösung x untersuchen. Es geht also um die **Kondition**.

Störungen der Anfangswerte $x_0 \in \mathbb{R}^m$ messen wir in der Euklidischen Norm $|\cdot|_2$. Um auch die Auswirkung von Störungen auf die Lösung $x : [0, T] \to \mathbb{R}^m$ quantifizieren zu können, definieren wir außerdem auf $C([0, T], \mathbb{R}^m)$, dem Vektorraum der stetigen Funktionen auf $[0, T]$ mit Werten in $\mathbb{R}^m$, die Norm

$$\|v\|_\infty = \max_{t \in [0,T]} |v(t)|_2 , \qquad v \in C([0, T], \mathbb{R}^m) .$$

Nun sind wir soweit.

Satz 3.19. *Die Koeffizientenmatrix $A \in \mathbb{R}^{m,m}$ sei symmetrisch. Seien x, $\tilde{x}$ die Lösungen des AWPSs* (3.62) *zu den Anfangswerten x_0, $\tilde{x}_0 \in \mathbb{R}^m$. Dann gilt*

$$\|x - \tilde{x}\|_\infty \leq \max_{t \in [0,T]} e^{\lambda_1 t} |x_0 - \tilde{x}_0|_2 . \tag{3.67}$$

Diese Abschätzung ist scharf.

Beweis. Übung. □

Wir haben damit die Kondition

$$\kappa(\text{AWPS}) = \max_{t \in [0,T]} e^{\lambda_1 t}$$

des Anfangswertproblems (3.62) ermittelt. Sind alle Eigenwerte negativ, also $\lambda_1 < 0$, so gilt

$$\kappa(\text{AWPS}) = 1 .$$

Im Falle $\lambda_1 \geq 0$ ist

$$\kappa(\text{AWPS}) = e^{\lambda_1 T} .$$

Man vergleiche das entsprechende Resultat im skalaren Fall aus Satz 3.4. Unter Verwendung der Lösungsdarstellung (3.66) kann man auch Satz 3.5 ohne besondere Schwierigkeiten auf den Systemfall erweitern.

Bemerkung. Nach Satz 3.19 hängt die eindeutig bestimmte Lösung x stetig von den Anfangsbedingungen ab. Das AWP (3.62) ist (bezüglich der verwendeten Normen!) korrekt gestellt.

Wir betrachten zum Schluß dieses Abschnitts ein Beispiel, das im Zusammenhang mit der numerischen Lösung dissipativer partieller Differentialgleichungen steht.

Beispiel. Nach Diskretisierung der sogenannten *Wärmeleitungsgleichung* mit der ebenfalls sogenannten *Linienmethode* zur *Ortsschrittweite* $h = 1/(m+1)$ erhält man ein Anfangswertproblem der Form (3.62) mit der Koeffizientenmatrix

$$A = (m+1)^2 \begin{pmatrix} -2 & 1 & 0 & 0 & \dots & 0 \\ 1 & -2 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 & -2 & 1 \\ 0 & \dots & 0 & 0 & 1 & -2 \end{pmatrix} \in \mathbb{R}^{m,m} .$$

Für $h \to 0$ oder gleichbedeutend $m \to \infty$ konvergieren die so berechneten Näherungen gegen die exakte Lösung.

Physikalisch lassen sich dabei die unbekannten Funktionen $x_k(t)$ als Approximationen der Temperatur eines Stabes S der Länge 1 an den Stellen $S_k = kh$, $h = 1/(m+1)$, $k = 1, \dots, m$, zum Zeitpunkt $t \in [0, T]$ interpretieren. Über die Anfangsbedingung $x(0) = x_0$ kann die Anfangstemperatur $x_k(0) = x_{k,0}$ an den Stellen S_k, $k = 1, \dots, m$, vorgeschrieben werden. In dem vorliegenden mathematischen Modell ist für alle $t \in [0, T]$ die Temperatur $x_0(t) = 0$ am linken Ende des Stabes und die Temperatur $x_{m+1}(t) = 0$ am rechten Ende fest gewählt.

Die Eigenvektoren der Tridiagonalmatrix A sind

$$\varphi_k = \left(\sin\left(k \frac{j}{m+1} \pi \right) \right)_{j=1,\dots,m} , \qquad k = 1, \dots, m ,$$

mit zugehörigen Eigenwerten

$$\lambda_k = -4(m+1)^2 \sin^2\left(\frac{k}{2(m+1)} \pi \right) , \quad k = 1, \dots, m .$$

Wegen

$$\lim_{m\to\infty} 2(m+1) \sin\left(\frac{1}{2(m+1)} \pi \right) = \pi , \qquad \lim_{m\to\infty} \sin\left(\frac{m}{2(m+1)} \pi \right) = 1$$

gilt für große m

$$-4(m+1)^2 \underset{\sim}{<} \lambda_m < \dots < \lambda_2 < \lambda_1 \underset{\sim}{<} -\pi^2 .$$

Nach Satz 3.18 setzt sich die Lösung dieses AWPSs also aus m Komponenten mit *stark unterschiedlichem Abklingverhalten* zusammen. Solche Systeme nennt man oft steif. Die Eigenschaften unseres Anfangswertproblems sind typisch für steife Systeme. Wir werden es daher von nun an kurz als *Modellproblem* bezeichnen.

Da alle Eigenwerte von A negativ sind, gilt offenbar

$$\max_{t\in[0,T]} e^{\lambda_1 t} < 1 .$$

Verwenden wir die Euklidische Norm $|\cdot|$, so ist aufgrund der Symmetrie von A außerdem $\kappa(T) = 1$.

Nach Satz 3.19 ist die Kondition also immer durch Eins beschränkt. Unser Modellproblem ist *gleichmäßig in m* gut konditioniert.

3.3.2 Euler-Verfahren

Wie zuvor in Abschnitt 3.2.2 gehen wir bei der Diskretisierung des AWPS (3.62) aus von einem *äquidistanten Gitter*

$$\Delta = \{0 = t_0 < t_1 < \cdots < t_n = T\}$$

zur konstanten Schrittweite $\tau = t_{k+1} - t_k = T/n$.

Ersetzt man die Ableitung durch den vorwärtsgenommenen Differenzenquotienten, so erhält man analog zu (3.36) das *explizite Euler-Verfahren*

$$x_{k+1} = x_k + \tau A x_k \ , \qquad k = 0, ..., n-1 \ , \tag{3.68}$$

zur Berechnung der Näherungslösungen x_k,

$$x_k \approx x(t_k) \ , \quad k = 0, \ldots, n \ .$$

Als Startwert für (3.68) verwenden wir den exakten Startwert x_0 aus (3.62). Offensichtlich kann man dann x_k, $k = 1, \ldots, n$, in eindeutiger Weise aus (3.68) berechnen. **Existenz** und **Eindeutigkeit** einer diskreten Lösung x_Δ sind also klar.

Wir kommen zur **diskreten Kondition**, bzw. zur **Stabilität**. Zur Quantifizierung der Wirkung von Störungen erklären wir zunächst auf dem linearen Raum der vektorwertigen Gitterfunktionen $v_\Delta : \Delta \to \mathbb{R}^m$ eine Norm durch

$$\|v_\Delta\|_\infty = \max_{k=0,\ldots,n} |v_k|_2 \ .$$

Satz 3.20. *Die Koeffizientenmatrix $A \in \mathbb{R}^{m,m}$ sei symmetrisch. Seien x_Δ, $\tilde{x}_\Delta$ die mit dem expliziten Euler-Verfahren* (3.68) *berechneten Näherungslösungen zu den Anfangswerten x_0, $\tilde{x}_0 \in \mathbb{R}^m$.*

Ist $\lambda_m < 0$, so genüge die Schrittweite τ der Stabilitätsbedingung

$$\tau \le \frac{2}{|\lambda_m|} \ . \tag{3.69}$$

Dann gilt

$$\|x_\Delta - \tilde{x}_\Delta\|_\infty \le \max_{t\in[0,T]} e^{t\lambda_1} |x_0 - \tilde{x}_0|_2 \ . \tag{3.70}$$

Beweis. Aus (3.68) erhält man direkt die Darstellungen

$$x_k = (I + \tau A)^k x_0, \qquad \tilde{x}_k = (I + \tau A)^k \tilde{x}_0$$

mit der Einheitsmatrix $I \in \mathbb{R}^{m,m}$. Daraus folgt sofort

$$\|x_\Delta - \tilde{x}_\Delta\|_\infty = \max_{k=0,\ldots,n} |x_k - \tilde{x}_k|_2 \le \max_{k=0,\ldots,n} |(I + \tau A)^k|_2 |x_0 - \tilde{x}_0|_2 \ .$$

Zur Abschätzung von $|(I+\tau A)^k|_2$ schreiben wir mit Hilfe der Diagonalisierung (3.58)

$$(I+\tau A)^k = (T(I+\tau D)T^{-1})^k = T(I+\tau D)^k T^{-1} .$$

Aus der Submultiplikativität der Matrixnorm und $\kappa(T) = |T|_2|T^{-1}|_2 = 1$ folgt

$$|(I+\tau A)^k|_2 \le |T^{-1}|_2 \, |I+\tau D|_2^k \, |T|_2 = |I+\tau D|_2^k .$$

Durch Einsetzen in die Definition (3.59) bestätigt man

$$|I+\tau D|_2 = \max_{i=1,\dots,m} |1+\tau\lambda_i| .$$

Ist $\lambda_i < 0$, so folgt aus der Stabilitätsbedingung (3.69) sofort

$$1 > 1+\tau\lambda_i \ge 1+\tau\lambda_m \ge -1 ,$$

also $|1+\tau\lambda_i| \le 1$. Im Falle $\lambda_1 < 0$ sind wir damit schon fertig, denn dann gilt ja

$$|I+\tau D|_2 = \max_{i=1,\dots,m} |1+\tau\lambda_i| \le 1 = \max_{t\in[0,T]} e^{\lambda_1 t} .$$

Ist $\lambda_1 \ge 0$, so gilt für jedes $\lambda_i \ge 0$ die Abschätzung

$$1 \le 1+\tau\lambda_i \le 1+\tau\lambda_1$$

und daher

$$|I+\tau D|_2 = \max_{i=1,\dots,m} |1+\tau\lambda_i| = 1+\tau\lambda_1 .$$

Die bekannte Abschätzung

$$(1+\tau\lambda_1)^k \le \left(1+\tau\lambda_m+\tfrac{(\tau\lambda)^2}{2!}+\dots\right)^k = e^{k\tau\lambda_m} \le e^{T\lambda_1}$$

liefert die Behauptung. □

Analog zum skalaren Fall erhalten wir also unter einer Schrittweitenbeschränkung (3.69) die Stabilität

$$\kappa(\text{exEuler}) \le \kappa(\text{AWPS})$$

des expliziten Euler-Verfahrens.

Die Diskretisierung eines inhomogenen Systems der Gestalt (3.65) mit dem expliziten Euler-Verfahren führt auf die Rekursion

$$x_{k+1} = x_k + \tau(Ax_k + f_k) , \quad k = 0,\dots,n-1 .$$

Mit vollständiger Induktion bestätigt man die geschlossene Darstellung

$$x_k = (I+\tau A)^k x_0 + \tau \sum_{j=0}^{k-1} (I+\tau A)^{k-(j+1)} f_j , \quad k = 0,\dots,n .$$

Nun kann man den Beweis von Satz 3.7 übertragen, um entsprechende Stabilitätsabschätzungen auch im Systemfall zu gewinnen.

Den Konsistenzbeweis aus Satz 3.12 kann man fast wörtlich übernehmen.

Aus Konsistenz und Stabilität folgt schließlich die Konvergenz des expliziten Euler-Verfahrens mit Ordnung $p = 1$, also

$$\max_{k=0,\dots,n} |x(t_k) - x_k|_2 \le C\tau \qquad \forall \tau \le \frac{2}{|\lambda_m|} ,$$

genau wie im skalaren Fall (siehe Satz 3.13).

Beispiel. Wie schon im kontinuierlichen Fall ist auch die diskrete Kondition des expliziten Euler-Verfahrens bei Anwendung auf unser Modellproblem besonders gut, denn dann gilt ja *gleichmäßig in m* die Abschätzung

$$\|x_\Delta - \tilde{x}_\Delta\|_\infty \le |x_0 - \tilde{x}_0|_2 .$$

Allerdings ist der Preis dafür unter Umständen sehr hoch. Wegen

$$\tau \le \frac{2}{|\lambda_m|} \approx \frac{1}{2(m+1)^2}$$

sind für große m nur noch äußerst kleine Zeitschrittweiten erlaubt.

Im skalaren Fall war das implizite Euler-Verfahren gerade für $\lambda < 0$ unbedingt stabil. Wir wollen sehen, ob das im Systemfall auch so ist.

Verwendet man den rückwärtsgenommenen Differenzenquotienten zur Diskretisierung der Ableitung nach t, so erhält man das *implizite Euler-Verfahren*

$$x_{k+1} = x_k + \tau A x_{k+1}, \quad k = 0, \dots, n-1 . \tag{3.71}$$

Offenbar haben wir in jedem Zeitschritt das lineare Gleichungssystem

$$(I - \tau A)x_{k+1} = x_k, \qquad k = 0, \dots, n-1 ,$$

zu lösen. Damit stellt sich die Frage, ob es immer eine eindeutig bestimmte Lösung gibt. Es geht also um **Existenz** und **Eindeutigkeit** einer diskreten Lösung von (3.71).

Mit Blick auf unser Modellproblem interessieren wir uns nur für den Fall negativer Eigenwerte.

Satz 3.21. *Die Koeffizientenmatrix $A \in \mathbb{R}^{m,m}$ sei symmetrisch und es gelte $\lambda_1 < 0$.*

Dann ist für jede Schrittweite $\tau > 0$ und jeden Anfangswert $x_0 \in \mathbb{R}^m$ durch das implizite Euler-Verfahren (3.71) *in eindeutiger Weise eine Näherungslösung x_Δ bestimmt.*

Beweis. Wegen (3.58) gilt

$$I - \tau A = T(I - \tau D)T^{-1}$$

und aufgrund der Voraussetzung $\lambda_m < \dots < \lambda_1 < 0$ ist $I - \tau D$ regulär. □

Wir kommen zur **Stabilität**.

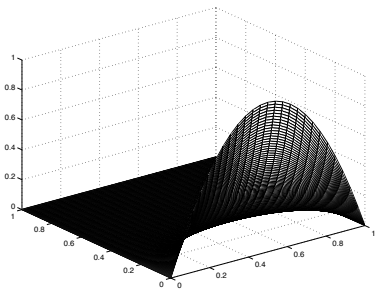
Satz 3.22. *Die Koeffizientenmatrix $A \in \mathbb{R}^{m,m}$ sei symmetrisch und es gelte $\lambda_1 < 0$.*

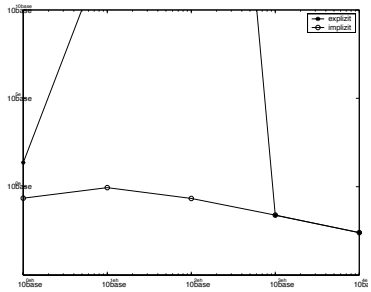
Seien x_Δ, $\tilde{x}_\Delta$ die mit dem impliziten Euler-Verfahren (3.71) berechneten Näherungslösungen zu den Anfangswerten x_0, $\tilde{x}_0 \in \mathbb{R}^m$. Dann gilt für jede Schrittweite $\tau > 0$ die Abschätzung

$$\|x_\Delta - \tilde{x}_\Delta\|_\infty \leq |x_0 - \tilde{x}_0|_2 \; . \tag{3.72}$$

Beweis. Übung. □

Um zu sehen, ob sich die verbesserten Stabilitätseigenschaften des impliziten Euler-Verfahrens gegenüber der expliziten Variante auch numerisch auszahlen, wollen wir beide Verfahren anhand unseres Modellproblems vergleichen.





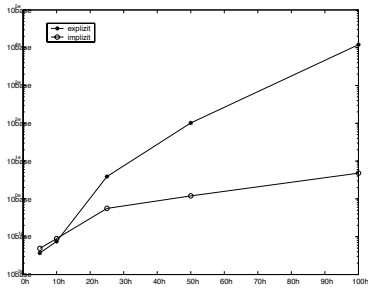
Dazu wählen wir die Anfangsbedingung

$$x_{0,k} = 4S_k(1 - S_k) \; , \quad k = 0, \ldots, m+1 \; ,$$

mit $S_k = kh$ und $h = 1/(m+1)$. Sei zunächst $m = 10$. Im linken Bild ist die schnell abklingende Lösung dargestellt. Auf der rechten Seite sehen wir den Diskretisierungsfehler des expliziten und des impliziten Euler-Verfahrens, jeweils für $n = 1, 10, \ldots, 10^4$. Da für $n < 238$ die Stabilitätsbedingung (3.69) verletzt ist, wird der Diskretisierungsfehler des expliziten Verfahrens zunächst mit kleiner werdender Schrittweite immer größer (bis zu 10^{51} bei $n = 100$)! Erst sobald (3.69) erfüllt ist, verhält sich der Fehler wie $\tau = 1/n$, ganz im Einklang mit unserer Konvergenzanalyse.

Wie mit Blick auf Satz 3.22 zu erwarten war, gibt es beim impliziten Verfahren keine Stabilitätsprobleme. Ab $n = 10$ beobachtet man lineares Abklingen des Diskretisierungsfehlers bei Reduktion von $\tau = 1/n$.

Die besseren Stabilitätseigenschaften des impliziten Euler-Verfahrens müssen mit der Lösung eines linearen Gleichungssystems in jedem Zeitschritt erkauft werden. Wir wollen nun untersuchen, ob sich dieser erhöhte Rechenaufwand lohnt. Dazu geben wir eine feste Genauigkeitsschranke $TOL = 10^{-1}$ vor. Durch Ausprobieren bestimmen wir für jedes $m = 5, 10, 25, 50, 100$ die größte Zeitschrittweite τ, so daß der entsprechende Diskretisierungsfehler für unsere beiden Euler-Verfahren kleiner als TOL ausfällt. Die für dieses τ benötigte Rechenzeit in Abhängigkeit von m ist in der folgenden Abbildung dargestellt.



Während für kleine m die mangelnde Stabilität des expliziten Euler-Verfahrens keine Rolle spielt, sind schließlich für $m = 100$ mehr als 3 Stunden erforderlich, um die Aufgabe zu lösen. Das implizite Verfahren kommt mit 4.8 Sekunden aus und das ist immerhin 5000 mal so schnell.

Fazit: Spezielle Eigenschaften eines Differentialgleichungssystems (hier: steif) erfordern spezielle Diskretisierungsverfahren (hier: implizit)!

Unsere numerischen Resultate legen nahe, daß man auch die Konvergenz des impliziten Euler-Verfahrens mit Ordnung $p = 1$ beweisen kann. Tatsächlich führt der im skalaren Fall durchgeführte und im Systemfall beim expliziten Verfahren angedeutete Schluß

$$\textbf{Konsistenz} + \textbf{Stabilität} = \textbf{Konvergenz}$$

in gleicher Weise zum Ziel wie zuvor.

Weiterführende Fragen.

- Kann man Verfahren höherer Ordnung konstruieren?
  Stichwort: Runge-Kutta-Verfahren (siehe z.B.: Deuflhard und Bornemann [1, Kapitel 4]).

- Wie und in welchem Umfang kann man die vorgestellten Ideen und Ergebnisse auf nichtlineare Anfangswertprobleme der Gestalt

  $$x'(t) = f(x,t)\,, \quad t \in (0,T]\,, \qquad x(0) = x_0$$

  übertragen?
  Stichwort: Flußoperator $x(t) = \phi^{0,t}x_0$ statt Matrix-Exponentieller $x(t) = e^{tA}x_0$ (siehe z.B.: Deuflhard und Bornemann [1, Kapitel 3]).

- Kann man in jedem Schritt die Zeitschrittweite τ_k automatisch so wählen, daß der Diskretisierungsfehler unter einer vorgegebenen Schranke bleibt?
  Stichwort: Adaptive Schrittweitensteuerung (siehe z.B.: Deuflhard und Bornemann [1, Kapitel 5]).

Alle diese Punkte werden u.a. auch in der Vorlesung *Einführung in die Numerische Mathematik (Numerik I)* angesprochen. (siehe Vorlesungsskript: Kornhuber und Schütte [2, Kapitel 5]).

Literatur

[1] P. Deuflhard und F. Bornemann. *Numerische Mathematik II.* de Gruyter, 2. Auflage 2002. Mittlerweile ein Standardlehrbuch über die numerische Lösung gewöhnlicher Differentialgleichungen. Das erste Kapitel vermittelt einen Überblick über aktuelle einige Anwendungsgebiete. In den folgenden Kapiteln 2–5 werden im Prinzip ähnliche Fragestellungen wie in diesem Skript behandelt, allerdings in viel größerer Allgemeinheit und damit auf entsprechend höherem mathematischen Niveau.

[2] R. Kornhuber und Ch. Schütte. *Einführung in die Numerische Mathematik (Numerik I).* FU Berlin, 1. Auflage 2001. Das Vorlesungsskript zur weiterführenden Numerik-Veranstaltung an der FU.

4 Nichtlineare Gleichungssysteme

4.1 Fixpunktiteration

Wir kennen bereits Verfahren, die es erlauben, die Lösung *linearer* Gleichungssysteme in endlich vielen Schritten (bis auf Rundungsfehler) *exakt* auszurechnen (vgl. z.B. Skript CoMa I). Im Falle *nichtlinearer Gleichungssysteme* gelingt dies im allgemeinen nicht mehr. Wir müssen uns mit *iterativen Verfahren* zufriedengeben, mit denen man eine Lösung in endlich vielen Schritten nur *bis auf eine vorgegebene Genauigkeit* berechnen kann. Wir beginnen mit einem einfachen Beispiel.

Beispiel. Gesucht ist eine Nullstelle $x^* \in \mathbb{R}$ von

$$F(x) = x(x-1).$$

Offenbar sind $x_1^* = 0$ und $x_2^* = 1$ Lösungen (Achtung: keine Eindeutigkeit).

Wir wollen nun ein Iterationsverfahren zur näherungsweisen Berechnung einer Nullstelle x^* konstruieren. Dazu betrachten wir anstelle der ursprünglichen nichtlinearen Gleichung $F(x) = 0$ das äquivalente Problem

$$x^* \in \mathbb{R}: \qquad \phi(x^*) = x^* \tag{4.1}$$

wobei

$$\phi(x) = F(x) + x = x^2$$

gesetzt ist. Jede Lösung von (4.1) heißt *Fixpunkt* von ϕ. Wir wollen nun eine Folge $\{x_k\}$ von Näherungslösungen konstruieren, die gegen einen Fixpunkt x^* von ϕ konvergiert. Wir versuchen es einfach mit der *Fixpunktiteration*

$$x_{k+1} = \phi(x_k), \qquad x_0 \in \mathbb{R} \text{ geeignet.}$$

Für $x_0 = 0.4$ erhält man die Folge $x_1 = 0.16$, $x_2 = 0.0256$, $x_3 = 0.0007$, ..., welche offenbar gegen x_1^* konvergiert. Die Wahl $x_0 = 1.1$ liefert hingegen die Folge $x_1 = 1.21$, $x_2 = 1.4641$, $x_3 = 2.1436$, ..., welche offenbar divergiert.

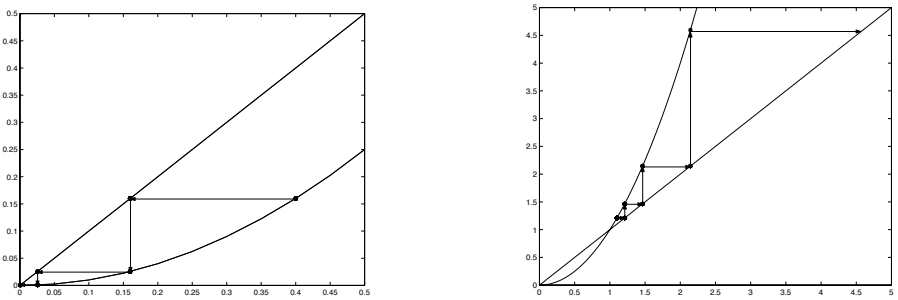


Abbildung 1: Konvergenz und Divergenz der Fixpunktiteration

Die Vorgehensweise ist in Abbildung 1 illustriert.

Je nach Startwert erhält man also Konvergenz oder Divergenz. Wir wollen verstehen, woran das liegt.

Satz 4.1. *Sei $I = [a, b]$ und $\phi : I \to \mathbb{R}$ eine Abbildung mit den beiden folgenden Eigenschaften.*

$$\phi(x) \in I \qquad \forall x \in I. \tag{4.2}$$

$$|\phi(x) - \phi(y)| \leq q|x - y| \qquad \forall x, y \in I, \quad q \in [0, 1). \tag{4.3}$$

Dann besitzt ϕ genau einen Fixpunkt $x^* \in I$ und die Folge

$$x_{k+1} = \phi(x_k)$$

konvergiert für jeden Startwert $x_0 \in I$. Die Fehlerreduktion *erfolgt gemäß*

$$|x^* - x_{k+1}| \leq q|x^* - x_k|.$$

Weiter gelten die a priori *Fehlerabschätzung*

$$|x^* - x_k| \leq \frac{q^k}{1 - q}|x_1 - x_0|$$

und die a posteriori *Fehlerabschätzung*

$$|x^* - x_{k+1}| \leq \frac{q}{1 - q}|x_{k+1} - x_k|.$$

Beweis. a) Durchführbarkeit.
Wegen (4.2) ist $x_k \in I \ \forall k \in \mathbb{N}$, falls $x_0 \in I$. Damit ist die Fixpunktiteration und auch die resultierende Folge $\{x_k\} \subset I$ wohldefiniert.

b) Konvergenz.
Als erstes zeigen wir, daß $\{x_k\}$ eine Cauchy-Folge ist. Wegen (4.3) haben wir

$$|x_{k+1} - x_k| = |\phi(x_k) - \phi(x_{k-1})| \leq q|x_k - x_{k-1}| \leq q^k|x_1 - x_0|$$

und ebenso

$$|x_{k+i+1} - x_{k+i}| \leq q|x_{k+1+(i-1)} - x_{k+(i-1)}| \leq q^i|x_{k+1} - x_k|.$$

Es gilt daher für $k, j \geq 0$

$$|x_{k+j} - x_k| \leq \sum_{i=0}^{j-1} |x_{k+i+1} - x_{k+i}| \leq |x_{k+1} - x_k| \sum_{i=0}^{j-1} q^i \leq \frac{q^k}{1 - q}|x_1 - x_0|.$$

Es gibt also zu jedem $\varepsilon > 0$ ein $k_0 \in \mathbb{N}$, so daß $|x_{k+j} - x_k| \leq \varepsilon \ \forall j \geq 0, k > k_0$. Damit ist $\{x_k\}$ Cauchy-Folge. Da $\mathbb{R}$ vollständig ist, gibt es ein $x^* \in \mathbb{R}$ mit

$$\lim_{k \to \infty} x_k = x^*.$$

Da $I \subset \mathbb{R}$ abgeschlossen ist, gilt $x^* \in I$.

c) x^* ist Lösung.
Aus (4.3) folgt insbesondere die Stetigkeit von ϕ. Daher gilt $\phi(x_k) \to \phi(x^*)$. Andererseits haben wir $\phi(x_k) = x_{k+1} \to x^*$. Aus der Eindeutigkeit des Grenzwerts folgt $\phi(x^*) = x^*$.

d) Eindeutigkeit des Fixpunkts.
Seien x^*, y^* zwei verschiedene Fixpunkte von ϕ. Dann folgt aus (4.3)

$$|x^* - y^*| = |\phi(x^*) - \phi(y^*)| \leq q|x^* - y^*|$$

offenbar $q \geq 1$. Widerspruch.

e) Fehlerreduktion.

$$|x^* - x_{k+1}| = |\phi(x^*) - \phi(x_k)| \leq q|x^* - x_k|.$$

f) a posteriori Fehlerabschätzung.
Die Abschätzung folgt aus

$$|x^* - x_{k+1}| \leq |\phi(x^*) - \phi(x_{k+1})| + |\phi(x_{k+1}) - \phi(x_k)| \leq q|x^* - x_{k+1}| + q|x_{k+1} - x_k|.$$

g) a priori Fehlerabschätzung.

$$|x^* - x_k| \leq \frac{q}{1-q}|x_k - x_{k-1}| \leq \frac{q}{1-q}q^{k-1}|x_1 - x_0|.$$

□

Bemerkung:

Ist $\phi \in C^1(I)$ und $q = \sup_{z \in I} |\phi'(z)|$, so gilt nach dem Mittelwertsatz mit einem $\xi \in I$

$$|\phi(x) - \phi(y)| = |\phi'(\xi)||x - y| \leq q|x - y| \qquad \forall x, y \in I.$$

Beispiel. Wir kommen auf unser Eingangsbeispiel zurück. Dazu betrachten wir $I = [-r, r]$ mit $r < \frac{1}{2}$ und $\phi(x) = x^2$. Dann folgt

$$|x^2 - y^2| = |x + y||x - y| \leq 2r|x - y| \qquad \forall x, y \in I,$$

also gilt (4.3) mit $q = 2r < 1$. Außerdem haben wir

$$|\phi(x)| \leq |x| \leq r \qquad \forall x \in I,$$

also (4.2). Das erklärt die Konvergenz der Fixpunktiteration für $x_0 = 0.4$ gegen $x_1^* = 0$. Übrigens liegt für alle $x_0 < 1$ Konvergenz der Fixpunktiteration gegen $x_1^* = 0$ vor: Die Voraussetzungen von Satz 4.1 sind dafür *hinreichend*, aber *nicht notwendig*.

Manche mögen es gleich gemerkt haben: Satz 4.1 ist ein Spezialfall des *Banachschen Fixpunktsatzes:*

Satz 4.2. *Sei B ein Banachraum mit Norm $\|\cdot\|$, $U \subset B$ abgeschlossen und $\phi : U \to B$ eine Abbildung mit den beiden folgenden Eigenschaften.*

$$\phi(x) \in U \qquad \forall x \in U. \tag{4.4}$$

$$\|\phi(x) - \phi(y)\| \leq q\|x - y\| \qquad \forall x, y \in U, \quad q \in [0, 1). \tag{4.5}$$

Dann besitzt ϕ genau einen Fixpunkt $x^* \in U$ und die Folge

$$x_{k+1} = \phi(x_k)$$

konvergiert für jeden Startwert $x_0 \in U$. Die Fehlerreduktion *erfolgt gemäß*

$$\|x^* - x_{k+1}\| \leq q\|x^* - x_k\|.$$

Weiter gelten die a priori *Fehlerabschätzung*

$$\|x^* - x_k\| \leq \frac{q^k}{1-q}\|x_1 - x_0\|$$

und die a posteriori *Fehlerabschätzung*

$$\|x^* - x_{k+1}\| \leq \frac{q}{1-q}\|x_{k+1} - x_k\|.$$

Der Beweis von Satz 4.2 ist wortwörtlich der gleiche wie der Beweis von Satz 4.1 (Vertrauen ist gut, Kontrolle ist besser).

Die Eigenschaft (4.5) gibt Anlass zu folgenden Definitionen.

Definition 4.3. Sei B ein Banach-Raum. Eine Abbildung $\phi : U \subset B \to B$ heißt Lipschitz-stetig auf U mit Lipschitz-Konstante L, falls gilt

$$\|\phi(x) - \phi(y)\| \leq L\|x - y\| \qquad \forall x, y \in U. \tag{4.6}$$

ϕ heißt kontrahierend auf U, falls $L \in [0, 1)$.

Definition 4.4. Sei B ein Banach-Raum. Eine Folge $\{x_k\} \subset B$ konvergiert linear mit Konvergenzrate q gegen x^*, falls gilt

$$\|x^* - x_{k+1}\| \leq q\|x^* - x_k\| \qquad \forall k \geq 0.$$

Als erste Anwendung von Satz 4.2 betrachten wir die *iterative Lösung linearer Gleichungssysteme.* Überraschenderweise sind iterative Verfahren den sogenannten *direkten* Methoden (z.B. Gaußscher Algorithmus) in wichtigen Fällen tatsächlich überlegen. (Stichwort: diskretisierte partielle Differentialgleichungen).

Vorgelegt sei also das lineare Gleichungssystem

$$Ax = b, \qquad A \in \mathbb{R}^{n,n},\ b \in \mathbb{R}^n. \tag{4.7}$$

Als erstes haben wir unser Gleichungssystem $F(x) = b - Ax = 0$ auf Fixpunktgestalt zu bringen. Die Wahl

$$\phi(x) = b - Ax + x$$

funktioniert meistens nicht (keine Kontraktion). Wir geben daher einen allgemeineren Zugang an. Dazu wählen wir $M \in \mathbb{R}^{n,n}$, regulär, mit der Eigenschaft

$$My = r \qquad \text{ist für alle } r \in \mathbb{R}^n \text{ mit } \mathcal{O}(n) \text{ Punktoperationen lösbar.} \tag{4.8}$$

Durch Addition von Mx erhält man die Fixpunktgestalt

$$Mx = F(x) + Mx = (M - A)x + b$$

und die zugehörige Fixpunktiteration

$$Mx_{k+1} = (M - A)x_k + b, \qquad x_0 \in \mathbb{R}^n. \tag{4.9}$$

In jedem Iterationsschritt hat man also wieder ein lineares Gleichungssystem zu lösen. Nach Voraussetzung ist das aber mit optimalem Aufwand (Ordnung $\mathcal{O}(n)$) möglich.

Mit Hilfe von Satz 4.2 wollen wir nun eine Bedingung an M herleiten, welche die Konvergenz des iterativen Verfahrens (4.9) garantiert. Die Fixpunktiteration (4.9) ist äquivalent zu

$$x_{k+1} = \phi(x_k), \qquad \phi(x) = (I - M^{-1}A)x + M^{-1}b, \tag{4.10}$$

wobei $I \in \mathbb{R}^{n,n}$ die Einheitsmatrix bezeichnet. Mit $\|\cdot\|$ bezeichnen wir sowohl eine *Vektornorm* auf $\mathbb{R}^n$ als auch die *zugehörige Matrixnorm* auf $\mathbb{R}^{n,n}$ (vgl. z.B. Skript CoMa I). Wir fordern nun zusätzlich zu (4.8), daß M auch der Bedingung

$$\|I - M^{-1}A\| = q < 1 \tag{4.11}$$

genügt. Dann sind die Voraussetzungen des Banachschen Fixpunktsatzes 4.2 mit $B = U = \mathbb{R}^n$ und ϕ aus (4.10) erfüllt. Die Lösung x^* des linearen Gleichungssystems (4.7) ist also eindeutig bestimmt (insbesondere ist A regulär!) und die Folge x_k konvergiert für jeden Startwert $x_0 \in \mathbb{R}^n$ mit der Konvergenzrate q gegen x^*.

Jede Vorschrift zur Wahl von M charakterisiert ein iteratives Verfahren für lineare Gleichungssysteme. Die einfachste Wahl $M = I$ führt auf das sogenannte *Richardson-Verfahren*. Die Wahl

$$M = \operatorname{diag}(A)$$

liefert das *Jacobi-Verfahren* (nach einer Arbeit von Carl Gustav Jacobi (1845)). Wie wir wissen (siehe z.B. Skript CoMa I), kann man gestaffelte Gleichungssysteme mit optimalem Aufwand (d.h. $n^2/2$ Punktop.) lösen. Das legt die Wahl

$$M_{ij} = \begin{cases} A_{ij}, & \text{falls } i \geq j \\ 0, & \text{sonst} \end{cases}$$

nahe. Man erhält das *Gauß-Seidel-Verfahren* (nach Carl Friedrich Gauß (1819–1823) und Phillip Ludwig Seidel (1874), übrigens ein Student von Jacobi).

Offenbar ist (4.8) in den obigen drei Fällen erfüllt. Ob auch (4.11) gilt, hängt jeweils von der Matrix A ab! Wir verweisen beispielsweise auf Deuflhard und Hohmann [1, Abschnitt 8.1] oder Stoer und Bulirsch [5, Kapitel 8]. Erheblich trickreichere iterative Verfahren, die sogenannten Mehrgittermethoden, werden später im Zusammenhang mit elliptischen partiellen Differentialgleichungen eine Rolle spielen.

Bislang haben wir nur lineare Konvergenz kennengelernt. Um höhere Konvergenzgeschwindigkeiten zu quantifizieren führen wir jetzt die Begriffe *superlineare Konvergenz* und *Konvergenzordnung* ein.

Definition 4.5. Sei B ein Banach-Raum und $\{x_k\} \subset B$. Die Folge $\{x_k\}$ heißt superlinear konvergent gegen x^*, falls es eine Folge $q_k \geq 0$ mit $\lim_{k\to\infty} q_k = 0$ gibt, so daß gilt

$$\|x_{k+1} - x^*\| \leq q_k \|x_k - x^*\|.$$

Sei $\{x_k\}$ konvergent gegen x^*. Dann heißt $\{x_k\}$ konvergent mit der Ordnung $p \geq 2$, falls es eine von k unabhängige Konstante $C \geq 0$ gibt, so daß gilt

$$\|x_{k+1} - x^*\| \leq C\|x_k - x^*\|^p. \tag{4.12}$$

Im Falle $p = 2$ sprechen wir von quadratischer Konvergenz.

Konvergenz mit Ordnung $p = 1$ ist dasselbe wie lineare Konvergenz (vgl. Definition 4.4). Die durch $x_{k+1} = x_k^2$ erzeugte Folge aus unserem Eingangsbeispiel konvergiert übrigens für $x_0 < 1$ quadratisch gegen $x_1^* = 0$. Das ist ein glücklicher Zufall. Nach Satz 4.1 bzw. Satz 4.2 konvergiert eine Fixpunktiteration nur linear. Ein berühmtes Verfahren, das unter gewissen Voraussetzungen quadratische Konvergenz liefert, diskutieren wir im nächsten Abschnitt.

4.2 Newton-Verfahren

Wir betrachten zunächst die skalare Gleichung

$$x^* \in \mathbb{R}: \qquad F(x^*) = 0 \tag{4.13}$$

mit einer stetig differenzierbaren Funktion $F : \mathbb{R} \to \mathbb{R}$.

Um eine Folge von Näherungslösungen zu berechnen, approximieren wir (4.13) durch eine Folge „einfacherer“ Probleme gleicher Bauart (vgl. iterative Verfahren für lineare Systeme im vorigen Abschnitt). Dazu ersetzen wir F durch eine „einfachere“ Funktion und berechnen deren Nullstelle. Ist $x_0 \in \mathbb{R}$ gegeben, so ist bekanntlich die *Tangente*

$$p(x) = F(x_0) + F'(x_0)(x - x_0)$$

eine gute Approximation von F, zumindest in einer genügend kleinen Umgebung von x_0. Unter der Voraussetzung $F'(x_0) \neq 0$ errechnet man die Nullstelle der Tangente p zu

$$x_1 = x_0 - \frac{F(x_0)}{F'(x_0)}.$$

Sukzessive Anwendung dieser Vorschrift liefert das *Newton-Verfahren*

$$x_{k+1} = x_k - \frac{F(x_k)}{F'(x_k)}, \qquad x_0 \in \mathbb{R} \text{ geeignet.} \tag{4.14}$$

Beispiel. Wir wollen das Newton-Verfahren auf die Funktion

$$F(x) = \arctan(x)$$

anwenden. Die ersten drei Iterationsschritte zum Startwert $x_0 = 1.3$ sind in Abbildung 2 veranschaulicht.

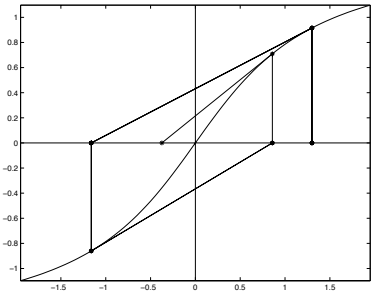


Abbildung 2: Newton-Verfahren

Zahlenwerte für $x_0 = 1.3$ und $x_0 = 1.4$ finden sich in Tabelle 2. Offenbar konvergieren für $x_0 = 1.3$ die

x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1.3	-1.1616	0.8589	-0.3742	0.0342	-2.6240e-05	1.2045e-14	0
1.4	-1.4136	1.4501	-1.5506	1.8471	-2.8936	8.7103	-103.2498

Tabelle 2: Iterierte des Newtonverfahrens

Iterierten mit wachsender Geschwindigkeit gegen die Lösung $x^* = 0$: Ab $k = 3$ verdoppelt sich die Anzahl der gültigen Stellen in jedem Schritt. Das bedeutet *quadratische Konvergenz*! Für $x_0 = 1.4$ sieht die Sache anders aus: Die Iterierten divergieren.

Bemerkung. Das Newton-Verfahren ist eine Fixpunktiteration $x_{k+1} = \phi(x_k)$ für

$$\phi(x) = x - \frac{F(x)}{F'(x)}.$$

Wir können also Satz 4.1 anwenden und erhalten folgendes Konvergenzkriterium. Unter den Voraussetzungen $F \in C^2(\mathbb{R})$ und

$$\sup_{z\in\mathbb{R}} |\phi'(z)| = \sup_{z\in\mathbb{R}} \left| F(z)\frac{F''(z)}{F'(z)^2} \right| = q < 1$$

konvergiert daher das Newton-Verfahren für jeden Startwert $x_0 \in \mathbb{R}$.

Leider erklärt das obige Resultat nicht die lokal wachsende Konvergenzgeschwindigkeit. Bevor wir dazu kommen, wollen wir das Newton-Verfahren auf ein System

$$x^* \in D: \qquad F(x) = 0 \quad , \quad F : D \subset \mathbb{R}^n \to \mathbb{R}^n \tag{4.15}$$

von nichtlinearen Gleichungen erweitern. Die Ableitung (Jacobi-Matrix)

$$F'(x_0) = \begin{pmatrix} \frac{\partial}{\partial x_1} F_1(x_0) & \dots & \frac{\partial}{\partial x_n} F_1(x_0) \\ \vdots & & \vdots \\ \frac{\partial}{\partial x_1} F_n(x_0) & \dots & \frac{\partial}{\partial x_n} F_n(x_0) \end{pmatrix} \in \mathbb{R}^{n,n}$$

von F an der Stelle x_0 ist bekanntlich charakterisiert durch

$$\|F(x) - (F(x_0) + F'(x_0)(x - x_0))\| = o\big(\|x - x_0\|\big) \quad \text{für} \quad x \to x_0.$$

Damit ist die Tangentialebene

$$p(x) = F(x_0) + F'(x_0)(x - x_0)$$

wieder eine gute Approximation von $F(x)$, zumindest in einer genügend kleinen Umgebung von x_0. Unter der Voraussetzung, daß $F'(x_k)$ jeweils regulär ist, hat die Iterationsvorschrift des Newton-Verfahrens für nichtlineare Systeme wieder die Gestalt

$$x_{k+1} = x_k - F'(x_k)^{-1} F(x_k), \qquad x_0 \in \mathbb{R}^n \text{ geeignet.}$$

Um die Berechnung von $F'(x_k)^{-1}$ zu vermeiden, verwendet man die äquivalente Formulierung

$$x_{k+1} = x_k + \Delta x_k, \quad F'(x_k)\Delta x_k = -F(x_k), \qquad x_0 \in \mathbb{R}^n \text{ geeignet.}$$

In jedem Iterationsschritt hat man also anstelle des ursprünglichen, nichtlinearen Problems ein lineares Gleichungssystem mit der Koeffizientenmatrix $F'(x_k)$ zu lösen.

Beispiel. Vorgelegt sei das Gleichungssystem

$$\begin{aligned} \sin(\xi) - \eta &= 0 \\ \xi - \cos(\eta) &= 0 \end{aligned}$$

also

$$\begin{aligned} F_1(\xi, \eta) &= \sin(\xi) - \eta \\ F_2(\xi, \eta) &= \xi - \cos(\eta) \end{aligned}.$$

Man erhält

$$F'(x) = \begin{pmatrix} \cos(\xi) & -1 \\ 1 & \sin(\eta) \end{pmatrix}, \qquad x = \begin{pmatrix} \xi \\ \eta \end{pmatrix}.$$

Achtung: $F'(x)$ kann singulär sein, z.B. für $x = (\pi, \frac{1}{2}\pi)$! Die Iterierten zu $x_0 = (3, 3)^T$ finden sich in Tabelle 3. Ab Schritt $k = 7$ beobachten wir wieder eine Verdopplung der gültigen Stellen in jedem Iterationsschritt, also *quadratische Konvergenz*.

k	ξ_k	η_k
1	-1.16898713819790	4.26838599329955
2	-7.26977629911835	-3.30627645454922
3	-1.87916601032632	2.13896869859183
4	3.42480564811751	-2.56261488791645
5	1.44984477398723	1.61684138641047
6	0.67129464906329	0.89875685831196
7	0.77538096829107	0.70350160297372
8	0.76818082842510	0.69484618466670
9	0.76816915690064	0.69481969089595
10	0.76816915673680	0.69481969073079
11	0.76816915673680	0.69481969073079

Tabelle 3: Iterierte des Newtonverfahrens.

Bemerkung. Offenbar sind (4.15) und

$$x^* \in D: \qquad AF(x^*) = 0 \tag{4.16}$$

für jede reguläre Skalierungsmatrix $A \in \mathbb{R}^{n,n}$ äquivalent. Man spricht von *Affin-Invarianz*. Diese *Struktureigenschaft* wird durch das Newton-Verfahren erhalten! Anwendung auf (4.16) liefert nämlich

$$\big((AF(x_k))'\big)^{-1} AF(x_k) = F'(x_k)^{-1}F(x_k) = \Delta x_k.$$

Konsequenterweise sollten also auch alle Konvergenzaussagen affin-invariant formuliert werden.

Dies berücksichtigen wir gleich bei der Formulierung der Voraussetzungen unseres Konvergenzsatzes.

Satz 4.6. *Sei $D \subset \mathbb{R}^n$ offen und $F : D \to \mathbb{R}^n$.*

Es existiere eine Nullstelle $x^* \in D$ von F.

Für alle $x \in D$ sei F differenzierbar. Die Jacobi-Matrix $F'(x)$ sei invertierbar für alle $x \in D$. Für alle $x \in D$ und $v \in \mathbb{R}^n$ mit $x + sv \in D$ für alle $s \in [0,1]$ gelte die (affin-invariante) Lipschitz-Bedingung

$$\|F'(x)^{-1}\big(F'(x+sv) - F'(x)\big)v\| \le s\omega\|v\|^2 \tag{4.17}$$

mit Lipschitz-Konstante $\omega \ge 0$.

Es sei schließlich $x_0 \in B_\rho(x^*) = \{x \in \mathbb{R}^n \mid \|x - x^*\| < \rho\}$, wobei $\rho > 0$ so gewählt ist, daß die Bedingungen

$$\rho < \tfrac{2}{\omega} \quad \textit{und} \quad B_\rho(x^*) \subseteq D$$

erfüllt sind.

Dann ist x^* die einzige Nullstelle von F in $B_\rho(x^*)$. Die Folge der Newton-Iterierten $\{x_k\}$ liegt in $B_\rho(x^*)$ und konvergiert quadratisch gegen x^*. Insbesondere gilt $\lim_{k\to\infty} x_k = x^*$ und

$$\|x_{k+1} - x^*\| \le \frac{\omega}{2}\|x_k - x^*\|^2, \qquad k = 0, 1, \dots \ .$$

Beweis. Die Fixpunktabbildung

$$\phi(x) = x - F'(x)^{-1}F(x)$$

ist wohldefiniert für alle $x \in D$. Sei $x \in B_\rho(x^*) \subseteq D$ und $s \in [0,1]$. Dann gilt $x + s(x^* - x) \in B_\rho(x^*) \subseteq D\ \forall s \in [0,1]$ und

$$\frac{d}{ds}F(x + s(x^* - x)) = F'(x + s(x^* - x))(x^* - x).$$

Aus dem Hauptsatz der Differential- und Integralrechnung folgt daher

$$0 = F(x^*) = F(x) + \int_0^1 F'(x + s(x^* - x))(x^* - x)\, ds.$$

Daraus ergibt sich

$$\begin{aligned} x^* - \phi(x) &= x^* - x + F'(x)^{-1}F(x) \\ &= F'(x)^{-1}\big(F(x) + F'(x)(x^* - x)\big) \\ &= F'(x)^{-1}\left(-\int_0^1 F'(x + s(x^* - x))(x^* - x)\, ds + F'(x)(x^* - x)\right) \\ &= -\int_0^1 F'(x)^{-1}\big(F'(x + s(x^* - x)) - F'(x)\big)(x^* - x)\, ds. \end{aligned}$$

Aus der Lipschitz-Bedingung und $\rho < \frac{2}{\omega}$ erhalten wir

$$\begin{aligned} \|x^* - \phi(x)\| &\le \int_0^1 \|F'(x)^{-1}\big(F'(x + s(x^* - x)) - F'(x)\big)(x^* - x)\|\, ds \\ &\le \int_0^1 \omega s\|x^* - x\|^2\, ds = \tfrac{1}{2}\omega\|x^* - x\|^2 < q\|x^* - x\| \qquad q := \tfrac{1}{2}\omega\rho < 1. \end{aligned} \tag{4.18}$$

Wegen $x_0 \in B_\rho(x^*)$ folgt daraus induktiv $x_k \subset B_\rho(x^*)$ und

$$\|x^* - x_k\| \le q^k\|x^* - x_0\| \to 0.$$

Auch die quadratische Konvergenz von $\{x_k\}$ folgt direkt aus (4.18).

Im Widerspruch zur Behauptung nehmen wir an, daß $\tilde{x}^* \in B_\rho(x^*)$ eine weitere Nullstelle von F ist. Einsetzen von $x = \tilde{x}^*$ in (4.18) führt dann auf $q > 1$. Widerspruch. □

Bemerkung. Da die quadratische Konvergenz „gute“ Startwerte $x_0 \in B_\rho(x^*)$ voraussetzt, spricht man von *lokal* quadratischer Konvergenz des Newton-Verfahrens. Wir haben schon am Eingangsbeispiel gesehen, daß für „schlechte“ Startwerte überhaupt keine Konvergenz vorzuliegen braucht.

Varianten des obigen Satzes liefern auch die Existenz der oben angenommenen Lösung x^*.

In der Praxis ist es oft nicht möglich, Startwerte zu finden, die auch nur die Konvergenz des Newton-Verfahrens gewährleisten. Um den Konvergenzbereich zu vergrößern, betrachten wir das *gedämpfte Newton-Verfahren*

$$x_{k+1} = x_k + \lambda_k \Delta x_k, \qquad F'(x_k)\Delta x_k = -F(x_k) \tag{4.19}$$

mit einem geeigneten *Dämpfungsparameter* $\lambda_k \in (0, 1]$. Die Wahl von λ_k sollte idealerweise so erfolgen, daß für alle $x_0 \in D$ Konvergenz vorliegt und sich darüberhinaus im Falle $x_k \in B_\rho(x^*)$ automatisch $\lambda_k = 1$ ergibt. So bliebe die lokal quadratische Konvergenz erhalten.

Wir beschreiben eine *affin-invariante Dämpfungsstrategie*, die auf dem sogenannten *natürlichen Monotonietest* beruht.

Algorithmus 4.7. *(Dämpfungsstrategie)*

gegeben: $\Delta x_k = F'(x_k)^{-1}F(x_k)$

1. *setze:* $\lambda_k = 1$

2. *berechne* $\bar{\Delta}x_k = F'(x_k)^{-1}F(x_k + \lambda_k \Delta x_k)$

3. *natürlicher Monotonietest:*
   falls $\|\bar{\Delta}x_k\| \leq (1 - \frac{\lambda_k}{2})\|\Delta x_k\|$ *akzeptiere* λ_k.
   andernfalls setze $\lambda_k = \lambda_k/2$ *und gehe zu Schritt 2.*

Beachte, daß in Schritt 2 jeweils ein lineares Gleichungssystem mit der Koeffizientenmatrix $F'(x_k)$ gelöst werden muß. Hat man Δx_k über eine LR-Zerlegung von $F'(x_k)$ berechnet, so ist dies mit optimalem Aufwand ($\mathcal{O}(n)$) möglich.

Beispiel. Wir betrachten wieder unser skalares Eingangsbeispiel $F(x) = \arctan(x)$. Die Iterierten des gedämpften Newton-Verfahrens für verschiedene Startwerte zeigt Tabelle 4. Offenbar wird der Konvergenz-

k	x_k	λ_k	x_k	λ_k	x_k	λ_k	x_k	λ_k
0	1.4000	0.5	5.0000	0.125	10.000	0.0625	100.00	0.0078
1	-0.0068	1.	0.5364	1.	0.7135	1.	-21.949	0.0312
2	2.1048e-07	1.	-0.0976	1.	-0.2217	1.	1.0620	0.5
3	-6.2469e-21	1.	6.1913e-04	1.	0.0072	1.	0.1944	1.
4	0	1.	-1.5821e-10	1.	-2.4854e-07	1.	-0.0049	1.
5			0	1.	1.0217e-20	1.	7.6666e-08	1.
6					0	1.	-2.9117e-22	1.
7							0	1.

Tabelle 4: Iterierte des Newtonverfahrens.

bereich erheblich erweitert! Man sieht, daß die mit x_0 wachsenden Inversen $|F'(x_0)^{-1}|$ durch immer kleinere Dämpfungsparameter kompensiert werden müssen. Andererseits wird in der Nähe der Lösung nicht mehr gedämpft und die lokal quadratische Konvergenz bleibt erhalten.

Literatur

[1] P. Deuflhard und A. Hohmann. *Numerische Mathematik I.* de Gruyter, 2. Auflage, 1992. Unsere Darstellung folgt im wesentlichen den Abschnitten 4.1 und 4.2. Ein Vergleich lohnt auf alle Fälle. In Abschnitt 8.1 wird auf iterative Verfahren für symmetrische, positiv definite Gleichungssysteme eingegangen.

[2] C.T. Kelley. *Iterative Methods for Linear and Nonlinear Equations.* SIAM, 1995. Eine kompakte Darstellung, die aber trotzdem deutlich über den Stoffumfang allgemeiner Einführungen in die Numerische Mathematik hinausgeht. Zum Beispiel erfährt man, wie eine Sekantenmethode für nichtlineare Systeme aussieht oder worauf man achten muß, wenn man beim Newton-Verfahren die linearen Teilprobleme ihrerseits iterativ löst.

[3] J.M. Ortega und W.C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables.* SIAM, 2000. Ein Reprint des 1970 erstmals erschienenen Standardwerks. Alle grundlegenden Techniken werden ausführlich vorgestellt. Was verständlicherweise fehlt, sind moderne,

vom kontinuierlichen Problem her motivierte Lösungsansätze für diskretisierte Differentialgleichungen. Aber das ist eine andere Geschichte und die soll ohnehin ein andermal erzählt werden.

[4] J. Stoer. *Numerische Mathematik I.* Springer, 8. Auflage, 1999. Ein Standardwerk zur Numerischen Mathematik, das zusammen mit Band II vor ca. 30 Jahren Maßstäbe gesetzt hat. Die Konvergenz des Newton-Verfahrens wird in Abschnitt 5.3 analysiert. Beachte, daß die Voraussetzungen leicht abweichen (Affin-Invarianz?).

[5] J. Stoer und R. Bulirsch. *Numerische Mathematik II.* Springer, 3. Auflage, 1990. Mehr über iterative Verfahren für lineare Gleichungssysteme kann man in Kapitel 8 erfahren.

A Das Riemann-Integral und Anwendungen

A1 Die Fläche unter einer Kurve

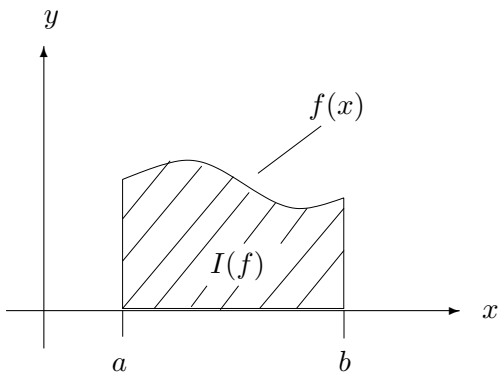
Gegeben sei eine Funktion

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

mit der Eigenschaft

$$0 \le f(x) \le \text{const.} \qquad \forall x \in [a, b] \ .$$

Gesucht ist der Flächeninhalt $I(f)$ unter der Kurve $(x, f(x))$, $x \in [a, b]$.



Im Falle einer konstanten Funktion ist der Fall klar:

$$f(x) = F \in \mathbb{R} \quad \forall x \in [a, b] \qquad \Rightarrow \qquad I(f) = (b - a)F \ .$$

Auch im Falle einer stückweise konstanten Funktion f (Treppenfunktion) kennen wir die Lösung. Ist nämlich beispielsweise

$$f(x) = F_i \in \mathbb{R} \qquad \forall x \in [x_i, x_{i+1}] \ , \qquad i = 0, \ldots, n - 1 \ ,$$

mit

$$a = x_0 < x_1 < \cdots < x_n = b, \tag{1.20}$$

so folgt

$$I(f) = \sum_{i=0}^{n-1} F_i (x_{i+1} - x_i) \ .$$

Den allgemeinen Fall führen wir nun auf Treppenfunktionen zurück.

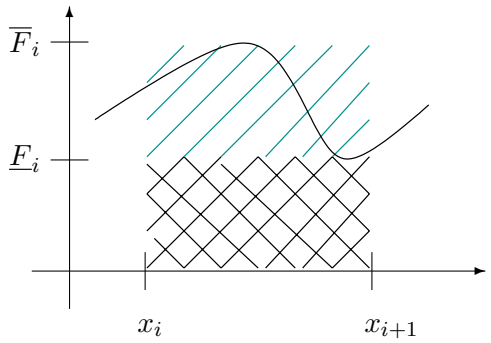
a) Wir wählen Gitterpunkte $x_0, \ldots, x_n$ mit der Eigenschaft (1.20), die wir *Zerlegung*

$$Z = \{x_0, \ldots, x_n\},$$

von $[a, b]$ nennen. Dann approximieren wir f von oben bzw. von unten durch die Treppenfunktionen $\overline{F}_Z$ bzw. $\underline{F}_Z$ mit $\overline{F}_Z(x) = \overline{F}_i$, $\underline{F}_Z(x) = \underline{F}_i$ $\forall x \in [x_i, x_{i+1}]$ und

$$\overline{F}_i = \sup_{\xi \in [x_i, x_{i+1}]} f(\xi) \ , \qquad i = 0, \ldots, n - 1 \ ,$$

$$\underline{F}_i = \inf_{\xi \in [x_i, x_{i+1}]} f(\xi) \ , \qquad i = 0, \ldots, n - 1 \ .$$



Wir definieren nun die *Obersumme*

$$\overline{I}(f,Z) = \sum_{i=0}^{n-1} \overline{F}_i (x_{i+1} - x_i)$$

und die *Untersumme*

$$\underline{I}(f,Z) = \sum_{i=0}^{n-1} \underline{F}_i (x_{i+1} - x_i) \ .$$

b) Offenbar gilt $\underline{I}(f,Z) \leq \overline{I}(f,Z)$. Ist $Z_1 \subset Z_2$ (d.h. Z_2 ist feiner als Z_1), so folgt

$$\underline{I}(f,Z_1) \leq \underline{I}(f,Z_2) \leq \overline{I}(f,Z_2) \leq \overline{I}(f,Z_1) \ .$$

Es existieren daher

$$\underline{I}(f) = \limsup_{\{Z|Z \text{ Zerlegung }\}} \underline{I}(f,Z) \leq \overline{I}(f) = \liminf_{\{Z|Z \text{ Zerlegung}\}} \overline{I}(f,Z) \ .$$

Man bezeichnet $\underline{I}(f)$ bzw. $\overline{I}(f)$ als unteres, bzw. oberes *Riemann-Darboux-Integral*.

Definition A.1. (Riemann-Integral)
Gilt

$$\underline{I}(f) = \overline{I}(f) \ ,$$

so ist f Riemann-integrierbar und

$$I(f) = \int_a^b f(x)\ dx = \underline{I}(f) = \overline{I}(f)$$

heißt Riemann-Integral von f auf $[a,b]$.

Beispiel (Dirichlet-Funktion):

$$f(x) = \begin{cases} 1 & x \text{ rational} \\ 0 & \text{sonst} \end{cases} \ , \qquad x \in [0,1] \ .$$

In jedem Intervall finden sich rationale und irrationale Zahlen, also ist

$$\underline{I}(f,Z) = 0 \ < \ 1 = \overline{I}(f,Z) \ .$$

f ist nicht Riemann-integrierbar!

Satz A.2. *Es sei $f : [a,b] \to \mathbb{R}$ stetig. Dann ist f Riemann-integrierbar und die Riemannsche Summe*

$$\sum_{i=0}^{n^{(\nu)}-1} f(x_i^{(\nu)}) \left(x_{i+1}^{(\nu)} - x_i^{(\nu)}\right) \to \int_a^b f(x)\, dx\ , \qquad \nu \to \infty$$

konvergiert für alle Zerlegungen $Z^{(\nu)} = \{x_0^{(\nu)}, \ldots, x_{n^{(\nu)}}^{(\nu)}\}$ mit

$$\max_{i=0,\ldots,n^{(\nu)}-1} \left(x_{i+1}^{(\nu)} - x_i^{(\nu)}\right) \to 0\ , \qquad \nu \to \infty\ . \tag{1.21}$$

Beweis. Da f stetig auf dem abgeschlossenen Intervall $[a,b]$ ist, ist f sogar gleichmäßig stetig auf $[a,b]$. Zu jedem $\varepsilon > 0$ gibt es daher ein $\delta > 0$, so daß

$$x, y \in [a,b],\ |x-y| < \delta \Rightarrow |f(x) - f(y)| < \frac{\varepsilon}{b-a}\ .$$

Es sei nun $Z^{(\nu)}$ eine Folge von Zerlegungen mit der Eigenschaft (1.21). Dann gibt es ein $\nu_0 \in \mathbb{N}$, so daß

$$\max_{i=0,\ldots,n^{(\nu)}-1} \left(x_{i+1}^{(\nu)} - x_i^{(\nu)}\right) < \delta \qquad \forall \nu \geq \nu_0\ .$$

Daher gilt für $\nu \geq \nu_0$ und $i = 0, \ldots, n^{(\nu)} - 1$

$$\left|\underline{F}_i - \overline{F}_i\right| = \left| \inf_{\xi \in [x_i^{(\nu)}, x_{i+1}^{(\nu)}]} f(\xi) - \sup_{\xi \in [x_i^{(\nu)}, x_{i+1}^{(\nu)}]} f(\xi) \right| < \frac{\varepsilon}{b-a}$$

und somit

$$\left|\underline{I}(f, Z^{(\nu)}) - \overline{I}(f, Z^{(\nu)})\right| \leq \sum_{i=0}^{n^{(\nu)}-1} \left|\underline{F}_i - \overline{F}_i\right| (x_{i+1}^{(\nu)} - x_i^{(\nu)}) \leq \frac{\varepsilon}{b-a} \sum_{i=0}^{n^{(\nu)}-1} (x_{i+1}^{(\nu)} - x_i^{(\nu)}) = \varepsilon\ .$$

Es folgt

$$\limsup_{\nu\to\infty} \underline{I}(f, Z^{(\nu)}) = \liminf_{\nu\to\infty} \overline{I}(f, Z^{(\nu)})$$

und zusammen mit

$$\begin{aligned} \limsup_{\nu\to\infty} \underline{I}(f, Z^{(\nu)}) &\leq \limsup_{\{Z | Z \text{ Zerlegung}\}} \underline{I}(f, Z) \leq \sum_{i=0}^{n^{(\nu)}-1} f(x_i^{(\nu)})(x_{i+1}^{(\nu)} - x_i^{(\nu)}) \\ &\leq \liminf_{\{Z | Z \text{ Zerlegung}\}} \overline{I}(f, Z) \leq \liminf_{\nu\to\infty} I(f, Z^{(\nu)}) \end{aligned}$$

folgt die Behauptung. □

Zerlegungen $Z = \{x_0, \ldots, x_n\}$ mit wachsendem n sind ein Spezialfall von $Z^{(\nu)}$. Im allgemeinen braucht nicht $x_k^{(\nu)} = x_k^{(\nu+1)}$ gelten. Aus (1.21) folgt $n^{(\nu)} \to \infty$ für $\nu \to \infty$.

Satz A.3. *Es sei $f : [a,b] \to \mathbb{R}$ stetig und $F : [a,b] \to \mathbb{R}$ eine Stammfunktion von f, d.h. $F'(x) = f(x)\ \forall x \in [a,b]$. Dann gilt*

$$\int_a^b f(x)\, dx = F(b) - F(a)\ .$$

Beweis. Die Behauptung folgt aus dem Hauptsatz der Differential- und Integralrechnung. □

Beispiele:

$$\int_a^b x^n \, dx = \frac{1}{n+1}\left(b^{n+1} - a^{n+1}\right) \qquad n \in \mathbb{R}, \;\; n \neq -1,$$
$$\int_a^b \sin x \, dx = \cos a - \cos b \; .$$

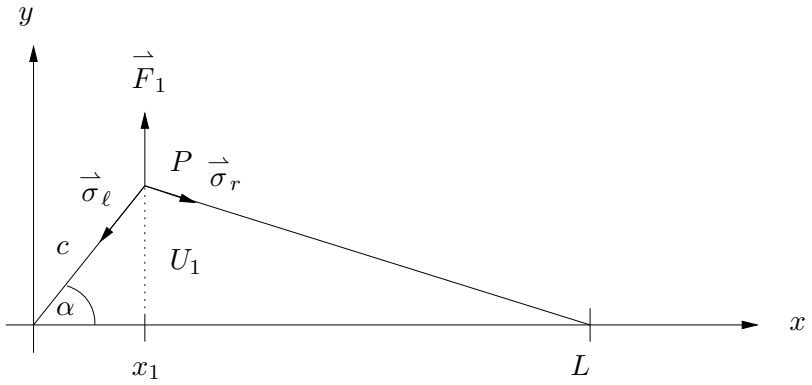
A2 Auslenkung einer Saite

Wir betrachten eine beidseitig eingespannte Saite der Länge $L > 0$. Unser Ziel ist es, ein mathematisches Modell herzuleiten, welches es erlaubt, die Auslenkung $u : [a, b] \to \mathbb{R}$ (Einheit: Meter) durch eine gegebene vertikale Kraftdichte $f : [a, b] \to \mathbb{R}$ (Einheit: $\frac{\text{Newton}}{\text{Meter}}$) zu berechnen.

Wir betrachten als erstes die *Ruhelage*.

In diesem Fall heben sich in jedem Punkt P auf der Saite die angreifenden Käfte $\vec{\sigma}$ und $-\vec{\sigma}$ auf. Wir gehen im folgenden davon aus, daß die *Spannung in Ruhelage* $|\vec{\sigma}|$ bekannt ist.

Als zweites Teilproblem betrachten wir die Auslenkung durch eine vertikale Punktkraft $\vec{F}_1 = \binom{0}{F_1}$ in $x_1 \in (0, L)$



Kräftegleichgewicht in P bedeutet: $\vec{\sigma_l} + \vec{\sigma_r} + \vec{F_1} = \vec{0}$ (Kräftebilanz).

Wir berechnen die y-Komponente von $\vec{\sigma}_l = \binom{\sigma_{l,x}}{\sigma_{l,y}}$. Es gilt

$$\sigma_{l,y} = -|\vec{\sigma_l}| \sin \alpha \; .$$

Für kleine Auslenkungen $U_1 \approx 0$ gilt

$$|\vec{\sigma_l}| \approx |\vec{\sigma}| \qquad \text{(Spannung in Ruhelage)}$$

und

$$\alpha \approx 0 \Rightarrow c \approx x_1 \Rightarrow \sin\alpha = \frac{U_1}{c} \approx \frac{U_1}{x_1} \ ,$$

also insgesamt

$$\sigma_{l,y} \approx -|\vec{\sigma}|\frac{U_1}{x_1} \ .$$

Analog erhält man

$$\sigma_{r,y} \approx -|\vec{\sigma}|\frac{U_1}{L-x_1} \ .$$

Einsetzen in die Kräftebilanz liefert

$$0 = \sigma_{l,y} + \sigma_{r,y} + F_1 \approx -|\vec{\sigma}|U_1\left(\frac{1}{x_1} + \frac{1}{L-x_1}\right) + F_1 \ .$$

Jetzt kommt ein naheliegender Schritt. Wir setzen einfach

$$-|\vec{\sigma}|U_1\left(\frac{1}{x_1} + \frac{1}{L-x_1}\right) + F_1 \stackrel{!}{=} 0 \ . \tag{1.22}$$

Dadurch machen wir einen Fehler, den sogenannten *Modellfehler*. Aus unseren Überlegungen folgt, daß der Modellfehler „klein“ ist, solange U_1 „klein“ ist. Eine quantitative Kontrolle haben wir nicht.

Aus (1.22) können wir U_1 ausrechnen.

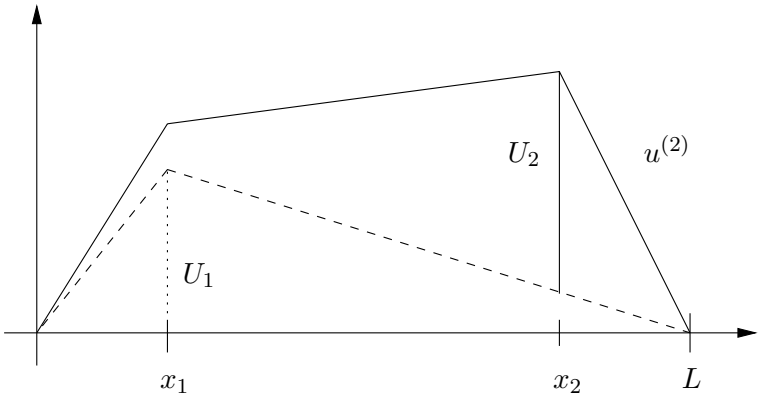
$$U_1 = \frac{(L-x_1)x_1}{|\vec{\sigma}|L}F_1 \ .$$

Durch eine weitere Kräftebilanz bestätigt man, daß die Auslenkung $U_1(x)$ auf $[0, x_1]$ und $[x_1, L]$ linear sein muß (Übung). Das bedeutet

$$U_1(x) = K(x, x_1)F_1 \ ,$$

$$K(x, x_1) = \begin{cases} \frac{(L-x_1)x}{|\vec{\sigma}|L}F_1 & 0 \le x \le x_1 \\ \frac{(L-x)x_1}{|\vec{\sigma}|L}F_1 & x_1 \le x \le L \ . \end{cases}$$

Als drittes Teilproblem betrachten wir nun die Auslenkung durch zwei vertikale Punktkräfte, nämlich $\vec{F}_1 = \binom{0}{F_1}$ in $x_1 \in (0, L)$ und $\vec{F}_2 = \binom{0}{F_2}$ in $x_2 \in (0, L)$.



Im Falle $U_1 \approx 0$ und $U_2 \approx 0$ gilt für das resultierende Inkrement $U_2(x)$

$$U_2(x) \approx K(x, x_2)F_2$$

und daher für die gesamte Auslenkung $u^{(2)}(x)$

$$u^{(2)}(x) \approx U_1(x) + U_2(x) = K(x, x_1)F_1 + K(x, x_2)F_2 \ .$$

Wir machen einen zweiten Modellfehler und setzen

$$u^{(2)}(x) \stackrel{!}{=} U_1(x) + U_2(x) = K(x, x_1)F_1 + K(x, x_2)F_2 \ .$$

Jetzt schließen wir von 2 auf n Punktquellen. Dazu sei

$$0 = x_0 < x_1 < \cdots < x_{n-1} < x_n = L$$

eine Zerlegung von $[0, L]$ und es seien $\vec{F}_0, \ldots, \vec{F}_{n-1}$ vertikale Punktkräfte in $x_0, \ldots, x_{n-1}$. Induktive Anwendung unserer Argumentation aus dem vorigen Schritt liefert für *kleine Auslenkungen* $u^{(n)}(x)$ die Darstellung

$$u^{(n)}(x) = \sum_{i=0}^{n-1} U_i(x) = \sum_{i=0}^{n-1} K(x, x_i)F_i \ .$$

Wir kommen zum letzten Schritt unserer Herleitung. Gegeben sei eine *vertikale Kraftdichte* $f : [a, b] \to \mathbb{R}$. Wir approximieren f durch

$$F_i = f(x_i)(x_{i+1} - x_i) \qquad i = 0, \ldots, n-1 \ .$$

Die Auslenkung durch die entsprechenden vertikalen Punktkräfte ist

$$u^{(n)}(x) = \sum_{i=0}^{n-1} K(x, x_i)f(x_i)(x_{i+1} - x_i) \ . \tag{1.23}$$

Als Modellierer gehen wir davon aus, daß diese Summe für

$$\max_{i=0,\ldots,n-1} (x_{i+1} - x_i) \to 0$$

und jedes feste $x \in [a, b]$ gegen

$$\int_0^L K(x, \xi)f(\xi)d\xi = u(x) \tag{1.24}$$

konvergiert. Das ist unser gesuchtes Modell.

Bemerkungen:

Als Mathematiker wissen wir aus Satz A.2, daß für stetige f die Riemannsche Summe (1.23) gegen das Integral (1.24) konvergiert.

Ist f nicht Riemann-integrierbar, so ist das Integral in (1.24) nicht definiert. In diesem Fall ist unser Modell *mathematisch sinnlos.*

Unser Modell (1.24) liefert für alle stetigen Kraftdichten f eine Auslenkung u, *ist also für alle stetigen f mathematisch sinnvoll.*

Aufgrund der Herleitung wissen wir, daß das Modell *nur für „genügend kleine" f physikalisch sinnvoll* ist. Eine genaue Quantifizierung von „klein" haben wir nicht hergeleitet.

Die Herleitung von (1.24) über (1.23) liefert ein numerisches Verfahren zur Approximation von u gleich mit. Das ist typisch. Dieses Verfahren ist jedoch meist nicht besonders effizient, d.h. Aufwand und Genauigkeit der Approximation stehen in keinem guten Verhältnis. Auch das ist typisch. Effizientere Verfahren werden in Kapitel 2 bereitgestellt.

B Lineare gewöhnliche Differentialgleichungen

A1 Chemische Reaktionssysteme

Wir betrachten eine *monomolekulare Reaktion*

$$A \longrightarrow B$$

zweier Gase in einem Behälter mit konstantem Volumen V. Unsere Reaktion läuft so ab, daß sich beim Zusammenstoß von 2 A-Teilchen eines der beiden in ein B-Teilchen verwandelt. Dabei gehen wir davon aus, daß Temperatur und Druck konstant sind. Außerdem machen wir die sogenannte Durchmischungshypothese, daß die Teilchen zu jedem Zeitpunkt t gleichmäßig im gesamten Behälter verteilt sind. Wir wollen nun ein mathematisches Modell zur Berechnung von

$$\eta_A(t), \eta_B(t): \text{ Anzahl der A-, B-Teilchen zum Zeitpunkt } t > 0$$

herleiten.

Unser Modell basiert auf der (plausiblen) Annahme, daß die Anzahl der Zusammenstöße $Z_A(t)$ zweier A-Teilchen pro „kleinem" Zeitintervall Δt proportional zur Anzahl der Teilchen ist. Es gilt also

$$Z_A(t) = k\eta_A(t)\Delta t \ .$$

Da mit jedem Zusammenstoß ein A-Teilchen verschwindet, ist also

$$\eta_A(t + \Delta t) = \eta_A(t) - Z_A(t) = \eta_A(t) - k\eta_A(t)\Delta t$$

und nach Umordnen

$$\frac{\eta_A(t + \Delta t) - \eta_A(t)}{\Delta t} = -k\eta_A(t) \ .$$

Teilchenerhaltung liefert

$$\eta_B(t+\Delta t) + \eta_A(t+\Delta t) = \eta_B(t) + \eta_A(t)$$

und es folgt

$$\frac{\eta_B(t+\Delta t) - \eta_B(t)}{\Delta t} = -\frac{\eta_A(t+\Delta t) - \eta_A(t)}{\Delta t} = k\eta_A(t) \ .$$

Grenzübergang $\Delta t \to 0$ (Kontinuumshypothese!) liefert das *System linearer Differentialgleichungen*

$$\eta_A' = -k\eta_A, \qquad \eta_B' = k\eta_A \qquad t > 0 \tag{2.25}$$

Anstelle der Teilchenanzahl interessieren meist die *Konzentrationen*, d.h. die Anzahl von Teilchen pro Volumen

$$c_A = \frac{\eta_A}{V}, \qquad c_B = \frac{\eta_B}{V} \ .$$

Da V konstant ist, erhält man sofort aus (2.25) die Differentialgleichungen

$$c_A' = -kc_A, \qquad c_B' = kc_A \qquad t > 0 \ .$$

Dazu treten Anfangsbedingungen

$$c_A(0) = c_{A,0}, \qquad c_B(0) = c_{B,0} \ .$$

Beachte, daß unabhängig von der Wahl der Anfangsbedingungen

$$c_A(t) + c_B(t) = c_A(0) + c_B(0) \quad \forall t > 0$$

gilt. Dies ist die differentielle Form der Teilchenerhaltung, welche häufig (unter Ignorieren des Molekulargewichts) als Massenerhaltung bezeichnet wird.