NUMERISCHE MATHEMATIK I

Vorlesungsmitschrift[1] zur Vorlesung von Prof. März im Sommersemester 2002, HU-Berlin

Die Numerische Mathematik beschäftigt sich mit der Entwicklung und Analyse (theoretisch und experimentell) von Methoden, Algorithmen und Software zur approximativen Lösung[2] von Problemklassen.

Ziele dieser Lehrveranstaltung sind:

(1) Vermittlung von Grundwissen zu elementaren Problemklassen:
nichtlineare Gleichungen, Minimierungsprobleme, Interpolation, numerische Berechnung von Integralen, Berechnung von Eigenwerten, numerische Lösung von Differentialgleichungen
(2) Erfahrungen zum Umgang mit numerischer Software
(3) algorithmisches konstruktives Denken

Literatur

[SB89] J. Stoer / R. Bulisch: Numerische Mathematik 1, Springer '89, $\geq$5. Auflage
[SB90] J. Stoer / R. Bulisch: Numerische Mathematik 2, Springer '90, $\geq$3. Auflage
[DH93] P. Deuflhard, A. Hohman: Numerische Mathematik I, de Gruyter 1993
[S79] H. Schwetlick, Numerische Lösung nichtlinearer Gleichungen, Dt. Verlag derWissenschaften, Berlin 1979
[S93] H. R. Schwarz, Numerische Mathematik, B.G. Teubner Stuttgart 1993, 3. Auflage

[1] Mitschrift von Stefan Vigerske und Daniel Seifert. Gefundene Fehler bitte an vigerske@mathematik.hu-berlin.de.
[2] wobei das Berechnen dieser Lösung effektiv und zuverlässig sein soll

Inhaltsverzeichnis

Literatur 1
1. Nichtlineare Gleichungen im $\mathbb{R}^m$ 3
1.1. Einfachste Fälle und Lösbarkeit 3
1.2. Newton-Verfahren und Sekantenverfahren 5
1.3. Quasi-Newton-Verfahren 11
1.4. Einbettung (Homotopie) 12
2. Ausgleichsprobleme (Kleinste-Quadrate-Probleme) und Gauß-Newton-Verfahren 15
2.1. Lineare Ausgleichsprobleme / lineare überbestimmte Gleichungen 16
2.2. Nichtlineare Ausgleichsprobleme 17
3. Minimierungsaufgaben (Optimierungsaufgaben) 19
3.1. Freie Minimierungssprobleme 19
3.2. Elementare Ansätze für Minimierung mit Nebenbedingungen 22
4. Interpolation 23
4.1. Interpolationspolynome von Lagrange und Newton 23
4.2. Interpolierende kubische Splinefunktionen 25
5. Numerische Berechnung von Integralen 30
5.1. Interpolationsformeln 30
5.2. Zusammengesetzte Formeln 33
5.3. Extrapolation und Romberg-Integration 34
6. Numerische Lösung gewöhnlicher Anfangswertaufgaben (AWA) 35
6.1. Ansatz linearer Mehrschrittverfahren und Konsistenz 37
6.2. Stabilität und Konvergenz 39
6.3. Reflexion des qualitativen asymptotischen Lösungsverhalten auf $[t_0, \infty)$ 45
6.4. Zur praktischen Realisierung 48
7. Eigenwertaufgaben 50
7.1. Einfache und inverse Vektoriteration zur Bestimmung spezieller Eigenwerte 50
7.2. QR-Verfahren zur Bestimmung aller Eigenwerte einer reellen Matrix 51
7.3. Abschätzung von Eigenwerten 54
8. Iterations-Verfahren zur Lösung großer linearer Gleichungssysteme 55
8.1. Gesamtschritt-, Einzelschritt-, Relaxationsverfahren 56
8.2. Mehrgitter-Verfahren 58
Index 59

1. Nichtlineare Gleichungen im $\mathbb{R}^m$

Ziel ist die Lösung von Gleichungen

$$\boxed{f(x) = 0}$$

mit $f : D \subseteq \mathbb{R}^m \to \mathbb{R}^m$, wobei D eine offene Menge ist und $f \in C^1$. Beispiele:

- Extremalpunkte zu $\varphi : \mathbb{R}^m \to \mathbb{R}$, $\varphi \in C^2$, $\varphi'(x) = 0$ (Analysis)
- Schnittpunkte von Kurven, ...
- Stationäre Lösung von Differentialgleichungen der Form $x'(t) = f(x(t))$, wobei $\bar{x}(t) = c$ ist stationäre Lösung genau dann, wenn $f(c) = 0$.
- Gleichungen in unendlich-dimensionalen Räumen: $F(x) = 0$ / Diskretisierung

1.1. Einfachste Fälle und Lösbarkeit.

- Lineare Gleichungen:

  $$f(x) = Ax - b,$$

  $x \in \mathbb{R}^m$, $A \in L(\mathbb{R}^m)$, $b \in \mathbb{R}^m$, A regulär $\rightsquigarrow \exists! x^* \in \mathbb{R}^m$ mit $f(x^*) = 0$

  $$\rightsquigarrow x_* = A^{-1}b$$

  In der Praxis wird man diese explizite Lösungsangabe aber nicht benutzen, sondern stattdessen die Gleichung $Ax - b = 0$ betrachten.
- $m = 1$, f Polynom der Form $f(x) = a_0 x^p + \ldots + a_p$, $a_i \in \mathbb{R}$. Dies ist für $p \leq 4$ mit Lösungsformel lösbar, allerdings sind auch diese vom numerischen Gesichtspunkt eher unbrauchbar.
- Kurvenschnittstellen:

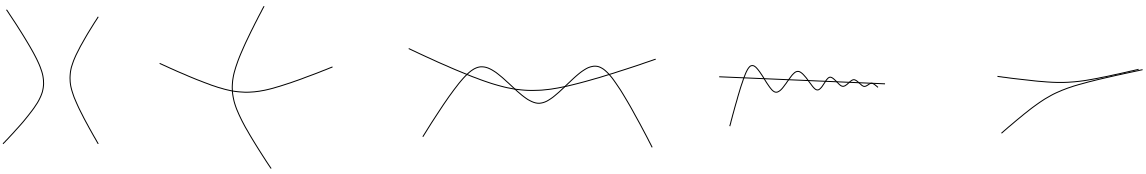


Abbildung 1.1. Schnittstellen von 2 Kurven

Bemerkung. Wir schreiben im folgenden $|.|$ für die Vektornorm auf $\mathbb{R}^m, \mathbb{C}^m$. Desweiteren sei $\|.\|$ die Matrix-Norm auf $L(\mathbb{R}^m)$, induziert durch $|.|$, $\|I\| = 1$:

$$A \in L(\mathbb{R}^m, \mathbb{R}^n): \quad \|A\| := \max_{x \in \mathbb{R}^m, |x|=1} |Ax| = \max_{x \in \mathbb{R}^m, x \neq 0} \frac{|Ax|}{|x|}$$

Die Regularität einer Matrix wird uns im folgenden noch öfter als Kriterium für bestimmte Eigenschaften dienen. Da in der Numerischen Mathematik oftmals mit Werten (inkl. Matrizen) gerechnet wird, welche aufgrund von Fehlern leicht vom korrekten Wert abweichen, ist es interessant zu wissen, wie stark eine Matrix sich ändern darf, ohne ihre Regularität zu verlieren:

Störungslemma

Lemma 1.1. *(Störungslemma, Banach-Lemma)*

Seien $A, C \in L(\mathbb{R}^m)$, A sei regulär, $\|A - C\| \leq \alpha$ und $\|A^{-1}\| \alpha < 1$. Dann ist C regulär und es gilt:

$$\boxed{\|C^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \alpha \|A^{-1}\|}}$$

Beweis.

$$\begin{aligned} C &= A\left(I - \underbrace{A^{-1}(A-C)}_{=:B}\right) = A(I-B) \\ \|B\| &\leq \left\|A^{-1}\right\| \cdot \|A-C\| \leq \left\|A^{-1}\right\| \cdot \alpha < 1 \end{aligned}$$

$\rightsquigarrow I-B$ ist regulär[3] $\rightsquigarrow C$ ist regulär, $C^{-1} = (I-B)^{-1} A^{-1}$

$$\begin{aligned} 1 &= \|I\| = \left\|(I-B)\cdot(I-B)^{-1}\right\| \\ &= \left\|(I-B)^{-1} - B(I-B)^{-1}\right\| \\ &\geq \left\|(I-B)^{-1}\right\| - \left\|B(I-B)^{-1}\right\| \\ &\geq \left\|(I-B)^{-1}\right\| - \|B\| \cdot \left\|(I-B)^{-1}\right\| \\ &= \left\|(I-B)^{-1}\right\| \cdot (1-\|B\|) \\ \rightsquigarrow \left\|(I-B)^{-1}\right\| &\leq \frac{1}{1-\|B\|} \\ \rightsquigarrow \left\|C^{-1}\right\| = \left\|(I-B)^{-1}A^{-1}\right\| &\leq \frac{1}{1-\|B\|} \cdot \left\|A^{-1}\right\| \leq \frac{\left\|A^{-1}\right\|}{1-\alpha\left\|A^{-1}\right\|} \end{aligned}$$

□

Bemerkung. Sei $A(s) = (a_{i,j}(s))_{i,j} \in L(\mathbb{R}^n, \mathbb{R}^m)$, $s \in [a,b]$. Dann ist $\int_a^b A$ wie folgt definiert:

$$\int_a^b A(s)\,ds := \left(\int_a^b a_{i,j}(s)\,ds\right)_{i,j}$$

Weiterhin gilt

$$\left\|\int_a^b A(s)\,ds\right\| \leq \left|\int_a^b \|A(s)\|\,ds\right|$$

Integralmittelwertsatz

Lemma 1.2. *(Integralmittelwertsatz) Sei $f \in C^1(D, \mathbb{R}^m)$, $D \subseteq \mathbb{R}^m$ offen, $x, y \in D, [x,y] \subset D$. Dann gilt:*

$$\boxed{f(x) - f(y) = \int_0^1 f'(sx + (1-s)y)\,ds \cdot (x-y)}$$

Beweis. Wir fixieren $i \in \{1, \ldots, m\}$. Sei

$$g_i(s) := f_i(sx + (1-s)y), \quad s \in [0,1].$$

g_i ist stetig differenzierbar auf dem Intervall $[0,1]$. Hauptsatz der Infinitisimalrechnung:

$$\begin{aligned} g_i(1) - g_i(0) &= \int_0^1 \frac{d}{ds} g_i(s)\,ds \\ \rightsquigarrow f_i(x) - f_i(y) &= \int_0^1 \sum_{j=1}^m \frac{\partial f_i}{\partial x_j}(sx + (1-s)y) \cdot (x_j - y_j)\,ds \\ &= \sum_{j=1}^m \int_0^1 \frac{\partial f_i}{\partial x_j}(sx + (1-s)y)\,ds \cdot (x_j - y_j), \quad i = 1, \ldots, m \end{aligned}$$

□

[3] Wäre $(I-B)z = 0, z \neq 0 \rightsquigarrow Bz = z, z \neq 0 \rightsquigarrow \frac{|Bz|}{|z|} = 1 \rightsquigarrow \|B\| = \max_{z \neq 0} \frac{|Bz|}{|z|} \geq 1$ ↯.

Oder: B ist kontraktiv $\rightsquigarrow$ Es ex. genau ein Fixpunkt. Wegen $B \cdot 0 = 0$ muß dann für alle $z \neq 0$ $Bz \neq z$ gelten.

1.2. **Newton-Verfahren und Sekantenverfahren.**

1.2.1. *Newton-Verfahren.*

Sei zunächst $m = 1$: Gesucht ist die Lösung von $f(x) = 0$. Wir wählen x_0 beliebig und berechnen die Tangente g_0 an $f(x_0)$ (Taylorentwicklung 1. Ordnung):

$$g_0(x) := f(x_0) + f'(x_0)(x - x_0)$$

Sei x_1 die Nullstelle von g_0, d.h. $g_0(x_1) = 0$. Dann gilt

$$x_1 = x_0 - f'(x_0)^{-1} \cdot f(x_0)$$

Wir wiederholen diese Schritte jetzt.

Sei nun $m \geq 1$ beliebig: Wir approximieren $f(x)$ durch

$$g_0(x) := f(x_0) + f'(x_0)(x - x_0), \quad x_0 \in D, x \in \mathbb{R}^m$$

(„Linearisierung") und lösen zuerst das lineare Problem $g_0(x) = 0$, d.h.

$$\begin{aligned} f'(x_0)(x - x_0) &= -f(x_0). \\ \rightsquigarrow x_1 &= x_0 - f'(x_0)^{-1} \cdot f(x_0) \end{aligned}$$

Dabei ist sicherzustellen, daß $f'(x_0)$ regulär ist und $x_1 \in D$ gilt.

Bei dem Newton-Verfahren handelt es sich um ein iteratives Verfahren der linearen Approximierung (iterative Linearisierung) von $f(x)$ durch

$$g_i(x) := f(x_i) + f'(x_i)(x - x_i).$$

Dabei bezeichnen wir x_0 als Startwert.

Newton-Verfahren mit Startwert x_0 : Newton-Verfahren

$$\boxed{\begin{aligned} x_j &:= x_{j-1} + z_j \\ \text{mit } z_j \text{ aus } f'(x_{j-1})\, z_j &= -f(x_{j-1}),\ j \geq 1 \end{aligned}}$$

Definition. Das Newton-Verfahren mit dem Startwert x_0 heißt durchführbar, falls es eine Folge $\{x_j\}_{j\geq 0}$ gibt, so daß $x_j \in D$ und $f'(x_j)$ regulär sind, $j \geq 0$. durchführbar

Satz 1.3. *(lokaler Konvergenzsatz für das Newton-Verfahren)* lokaler Konvergenzsatz

Sei $f \in C^1(D, \mathbb{R}^m)$, $D \subseteq \mathbb{R}^m$ offen, $x_* \in D$, $f(x_*) = 0$ und $f'(x_*)$ regulär. Dann gilt

(1) *Es existiert eine Kugel $\bar{B}(x_*, \rho)$, $\rho > 0$, derart, daß das Newton-Verfahren mit einem Startwert $x_0 \in \bar{B}(x_*, \rho)$ durchführbar ist und eine gegen x_* konvergente Folge $\{x_j\}_{j\geq 0}$ liefert, mit $x_j \in \bar{B}(x_*, \rho)$.*

(2) *Sei $x_0 \in \bar{B}(x_*, \rho)$. Dann gilt*

$$|x_j - x_*| \leq q_j \cdot |x_{j-1} - x_*|, \quad j \geq 1$$

mit einer Nullfolge $\{q_j\}_{j\geq 1}$.

(3) *Gilt zusätzlich, daß die Jacobi-Matrix Lipschitz-stetig ist, d.h.*

$$\|f'(x) - f'(\bar{x})\| \leq L|x - \bar{x}|, \quad x, \bar{x} \in \bar{B}(x_*, \rho), L \in \mathbb{R}^+$$

so gilt die quadratische Konvergenz:

$$|x_j - x_*| \leq q_j |x_{j-1} - x_*|^2, \quad j \geq 1, q \in \mathbb{R}^+$$

Beweis.

(1) Da f' in x_* stetig ist, existiert zu jedem $\varepsilon > 0$ ein $\rho(\varepsilon) > 0$, so daß

$$\|f'(x) - f'(x_*)\| \leq \varepsilon \quad \text{für } x \in \bar{B}(x_*, \rho(\varepsilon)) \subset D.$$

Sei $\varepsilon > 0$ so, daß $\varepsilon \cdot \left\| f'(x_*)^{-1} \right\| < 1$. Nach Störungslemma (L. 1.1) ist $f'(x)$ regulär und

$$\left\| f'(x)^{-1} \right\| \leq \frac{\left\| f'(x_*)^{-1} \right\|}{1 - \varepsilon \cdot \left\| f'(x_*)^{-1} \right\|}, \quad x \in \bar{B}(x_*, \rho(\varepsilon))$$

Wir wählen einen Startwert $x_0 \in \bar{B}(x_*, \rho(\varepsilon))$. Dann gilt:

$$\begin{aligned}
x_1 - x_* &= x_0 - x_* - f'(x_0)^{-1} \cdot f(x_0). \\
\rightsquigarrow\ x_1 - x_* &\overset{\text{L. 1.2}}{=} x_0 - x_* - f'(x_0)^{-1} \cdot \int_0^1 f'(sx_0 + (1-s)x_*)\, ds \cdot (x_0 - x_*) \\
&= f'(x_0)^{-1} \cdot \left\{ f'(x_0) - \int_0^1 f'(sx_0 + (1-s)x_*)\, ds \right\} \cdot (x_0 - x_*) \\
\rightsquigarrow\ |x_1 - x_*| &\leq \left\| f'(x_0)^{-1} \right\| \cdot \int_0^1 \underbrace{\|f'(x_0) - f'(sx_0 + (1-s)x_*)\|}_{\leq \|f'(x_0) - f'(x_*)\| + \|f'(x_*) - f'(sx_0 + (1-s)x_*)\| < 2\varepsilon} ds \cdot |x_0 - x_*| \\
&\leq \underbrace{\frac{\left\| f'(x_*)^{-1} \right\|}{1 - \varepsilon \left\| f'(x_*)^{-1} \right\|} \cdot 2\varepsilon}_{=:\eta(\varepsilon)} \cdot |x_0 - x_*|
\end{aligned}$$

η ist stetig und $\eta(0) = 0$. Gesucht ist ein hinreichend kleines ε mit $\eta(\varepsilon) < 1$.

Sei $\varepsilon \cdot \left\| f'(x_*)^{-1} \right\| < \frac{1}{3}$. Dann ist $\eta(\varepsilon) < 1$. Wir fixieren ein solches ε.

Sei $\rho := \rho(\varepsilon)$, $\eta := \eta(\varepsilon)$, $x_0 \in \bar{B}(x_*, \rho)$. Damit ist

$$\begin{aligned}
|x_1 - x_*| &\leq \eta |x_0 - x_*| \qquad \rightsquigarrow \quad x_1 \in \bar{B}(x_*, \rho) \\
\rightsquigarrow\ |x_i - x_*| &\leq \eta |x_{i-1} - x_*| \leq \eta^i |x_0 - x_*| \xrightarrow{i \to \infty} 0.
\end{aligned}$$

(2) Es gilt

$$\begin{aligned}
x_j - x_* &= x_{j-1} - x_* - f'(x_{j-1})^{-1} (f(x_{j-1}) - f(x_*)), \text{ da } f(x_*) = 0 \\
&= f'(x_{j-1})^{-1} \left(f'(x_{j-1}) - \int_0^1 f'(sx_{j-1} + (1-s)x_*)\, ds \right) (x_{j-1} - x_*) \\
\rightsquigarrow\ |x_j - x_*| &\leq \underbrace{\frac{\left\| f'(x_*)^{-1} \right\|}{1 - \varepsilon \left\| f'(x_*)^{-1} \right\|} \cdot \int_0^1 \left\| \underbrace{f'(x_{j-1})}_{\to f'(x_*)} - \underbrace{f'(sx_{j-1} + (1-s)x_*)}_{\to f'(x_*)} \right\| ds}_{=:q_j \xrightarrow{j \to \infty} 0} \cdot |x_{j-1} - x_*|
\end{aligned}$$

(3) Wie in (2) erhalten wir

$$\begin{aligned}
|x_j - x_*| &\leq \frac{\left\| f'(x_*)^{-1} \right\|}{1 - \varepsilon \left\| f'(x_*)^{-1} \right\|} \cdot \int_0^1 \underbrace{\|f'(x_{j-1}) - f'(sx_{j-1} + (1-s)x_*)\|}_{\leq (1-s) \cdot L |x_{j-1} - x_*|} ds \cdot |x_{j-1} - x_*| \\
&\leq \frac{\left\| f'(x_*)^{-1} \right\|}{1 - \varepsilon \left\| f'(x_*)^{-1} \right\|} \cdot \frac{1}{2} L \cdot |x_{j-1} - x_*|^2, \qquad q := \frac{\left\| f'(x_*)^{-1} \right\|}{1 - \varepsilon \left\| f'(x_*)^{-1} \right\|} \cdot \frac{1}{2} L
\end{aligned}$$

□

Definition. (Begriffe zur Konvergenzgeschwindigkeit)

- Falls eine Folge $\{x_j\}_{j\geq 0}$, die gegen ein x_* konvergiert, der Ungleichung

  lineare Konvergenz

  $$\boxed{|x_j - x_*| \leq \alpha\,|x_{j-1} - x_*|\,, \quad j \geq 1,\, \alpha \in (0,1)}$$

  genügt, so spricht man von linearer Konvergenz.

- Gilt eine Ungleichung

  überlineare Konvergenz

  $$\boxed{|x_j - x_*| \leq q_j\,|x_{j-1} - x_*|\,, \quad j \geq 1,\, q_j \xrightarrow{j\to\infty} 0}$$

  so spricht man von überlinearer Konvergenz (superlineare Konvergenz).

- Gilt eine Ungleichung

  quadratische Konvergenz

  $$\boxed{|x_j - x_*| \leq q\,|x_{j-1} - x_*|^2\,, \quad j \geq 1,\, q > 0}$$

  so spricht man von quadratischer Konvergenz.

Definition. Der Einzugsbereich der Nullstelle x_* der Funktion f ist die Menge der Startwerte $x_0 \in D$, für die das Newton-Verfahren durchführbar ist und eine Folge $\{x_j\}_{j\geq 0} \subseteq D$ mit $x_j \xrightarrow{j\to\infty} x_*$ liefert. Einzugsbereich

Die Teilmenge des Einzugsbereiches, für welche die Ungleichung $|x_j - x_*| \leq q|x_{j-1} - x_*|^2$ mit $q > 0$ und $j \geq 1$ relevant ist, bezeichnen wir als den Bereich der quadratischen Konvergenz.

Bemerkung. (zu Satz 1.3):

(1) Lokale Konvergenz ist naturgemäß bei nichtlinearen Problemen die einzige mögliche Aussage. Globale Konvergenzsätze haben stets eine Einschränkung des Definitionsbereiches beziehungsweise der betrachteten Funktionen zur Folge.

(2) Kann man auf die Regularität von $f'(x_*)$ verzichten? Die Antwort darauf ist „ja", allerdings sollte man in der Regel nicht auf die Regularität verzichten, wie wir am folgenden Beispiel sehen werden.

Beispiel. Sei $m = 1$, $f(x) := (x-1)^k$, $k > 1 \rightsquigarrow x_* = 1$

$f'(x) = k \cdot (x-1)^{k-1}$, $x_0 \neq 1$

$$\begin{aligned} x_j &= x_{j-1} - \frac{f(x_{j-1})}{f'(x_{j-1})} = x_{j-1} - \frac{x_{j-1} - 1}{k} \\ x_j - x_* = x_j - 1 &= \underbrace{\left(1 - \frac{1}{k}\right)}_{=:\alpha} (x_{j-1} - 1) \\ \rightsquigarrow x_j &\xrightarrow{j\to\infty} 1 \end{aligned}$$

Das Verfahren konvergiert also trotz fehlender Regularität, allerdings ist nur lineare Konvergenz möglich.

1.2.2. *Sekantenverfahren.*

Zunächst sei $m = 1$: Gesucht ist die Lösung von $f(x) = 0$. Wir wählen x_0 und $x_1 = x_0 + h$ und berechnen die Sekante durch $f(x_0)$ und $f(x_0 + h)$:

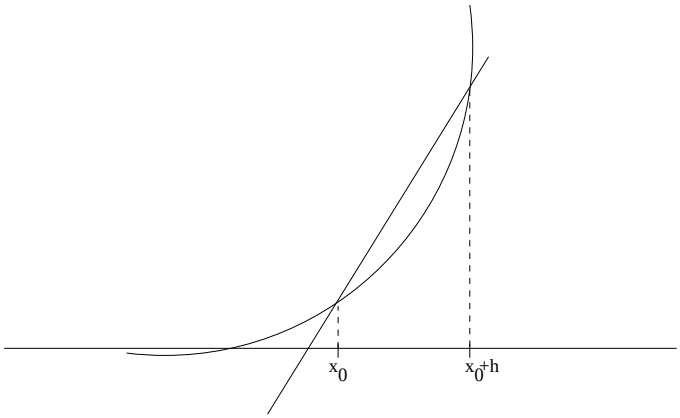


ABBILDUNG 1.2. $g_0(x) = f(x_0) + \frac{1}{h}(f(x_0 + h) - f(x_0))(x - x_0)$

Für $m \geq 1$ definieren wir die Differenzenquotientenmatrix $F(x,h) \in L(\mathbb{R}^m)$ mit $x \in D$ und hinreichend kleiner Schrittweitenmatrix $h \in L(\mathbb{R}^m)$ als

$$\boxed{(F(x,h))_{i,j} := \begin{cases} \frac{1}{h_{i,j}}(f_i(x + h_{i,j}e_j) - f_i(x)) & \text{falls } h_{i,j} \neq 0 \\ \frac{\partial f_i}{\partial x_j}(x) & \text{falls } h_{i,j} = 0 \end{cases} \qquad i,j = 1,\ldots,m}$$

$F(x,h)$ ist stetig: $F(x_*,0) = f'(x_*)$ und es gilt nach dem Hauptsatz der Infinitisimalrechnung:

$$\begin{aligned} (F(x,h))_{i,j} &= \int_0^1 f_i'(x + sh_{ij}e_j)\,ds \\ F(x,0) &= f'(x) \end{aligned}$$

Sei $x_* \in D$, $f(x_*) = 0$, $f'(x_*)$ regulär. Dann ist $F(x_*,0) = f'(x_*)$ regulär.

Iterative Linearisierung:

$$g_j(x) := f(x_j) + F(x_j,h_j)(x - x_j)$$

$f(x) = 0$ wird ersetzt durch $g_j(x) = 0$ (lineare Gleichung) mit Lösung x_{j+1}:

$$F(x_j,h_j)(x_{j+1} - x_j) = -f(x_j)$$

Sekanten-verfahren

Algorithmus 1.4. *Sekantenverfahren mit Startwert $x_0 \in D$, $j \geq 0$:*

(1) *bestimme* $F(x_j,h_j)$

(2) *bestimme* z_{j+1} *aus*

$$\boxed{F(x_j,h_j)\,z_{j+1} = -f(x_j)}$$

(3) *setze*

$$\boxed{x_{j+1} := x_j + z_{j+1}}$$

durch-führbar

Definition. Das Sekantenverfahren heißt durchführbar, wenn $x_j \in D$ und $F(x_j,h_j)$ regulär sind.

Satz 1.5. *Sei $f \in C^1(D, \mathbb{R}^m)$, $D \subseteq \mathbb{R}^m$ offen, $x_* \in D$, $f(x_*) = 0$ und $f'(x_*)$ regulär. Dann gilt:* lokaler Konvergenzsatz

(1) *Es gibt $\rho > 0$, $\gamma > 0$ derart, daß das Sekantenverfahren mit Startwerten $x_0 \in \bar{B}(x_*, \rho)$ und Schrittweiten aus dem Intervall $[-\gamma, \gamma]$ durchführbar ist und gegen x_* konvergente Folgen liefert.*

(2) *Falls zusätzlich $h_j \xrightarrow{j\to\infty} 0$ gilt, so konvergiert die Folge $\{x_j\}_{j\geq 0}$ überlinear gegen x_*.*

Beweis. Wir wählen $\varepsilon > 0$ mit $\varepsilon \cdot \left\| f'(x_*)^{-1} \right\| < \frac{1}{3}$. Wegen der Stetigkeit von F existieren $\rho(\varepsilon) > 0$ und $\gamma(\varepsilon) > 0$ mit

$$\left\| F(x,h) - \underbrace{F(x_*, 0)}_{f'(x_*)} \right\| \leq \varepsilon, \quad x \in \bar{B}(x_*, \rho(\varepsilon)), |h_{i,j}| \leq \gamma(\varepsilon)$$

Nach dem Störungslemma (L. 1.1) ist $F(x,h)$ regulär für $x \in \bar{B}(x_*, \rho(\varepsilon))$, $|h_{i,j}| \leq \gamma(\varepsilon)$:

$$\left\| F(x,h)^{-1} \right\| \leq \frac{\left\| f'(x_*)^{-1} \right\|}{1 - \varepsilon \left\| f'(x_*)^{-1} \right\|}, \quad x \in \bar{B}(x_*, \rho(\varepsilon)), |h_{i,j}| \leq \gamma(\varepsilon)$$

Sei $\rho := \rho(\varepsilon)$, $\gamma := \gamma(\varepsilon)$, $x_0 \in \bar{B}(x_*, \rho)$ und $|h_{0_{i,k}}| \leq \gamma$. $F(x_0, h_0)$ ist regulär. Dann ist

$$\begin{aligned}
x_1 - x_* &= x_0 - x_* - F(x_0,h_0)^{-1} \{f(x_0) - f(x_*)\} \\
&\overset{\text{L. 1.2}}{=} F(x_0,h_0)^{-1} \left\{ F(x_0,h_0) - f'(x_*) + f'(x_*) - \int_0^1 f'(sx_0 + (1-s)x_*)\, ds \right\} (x_0 - x_*) \\
|x_1 - x_*| &\leq \frac{\left\| f'(x_*)^{-1} \right\|}{1 - \varepsilon \left\| f'(x_*)^{-1} \right\|} \{\varepsilon + \varepsilon\} |x_0 - x_*| \\
&= \eta(\varepsilon) |x_0 - x_*|, \quad \eta(\varepsilon) < 1 \\
|x_j - x_*| &\leq \eta(\varepsilon)^j |x_0 - x_*| \xrightarrow{j\to\infty} 0
\end{aligned}$$

Gilt nun zusätzlich $h_j \xrightarrow{j\to\infty} 0$, so ist

$$\begin{aligned}
x_j - x_* &= \underbrace{F(x_j,h_j)^{-1}}_{\xrightarrow{j\to\infty} f'(x_*)^{-1}} \left\{ \underbrace{F(x_j,h_j)}_{\xrightarrow{j\to\infty} F(x_*,0) = f'(x_*)} - \int_0^1 \underbrace{f'(sx_j + (1-s)x_*)}_{\xrightarrow{j\to\infty} f'(x_*)}\, ds \right\} (x_{j-1} - x_*) \\
\rightsquigarrow |x_j - x_*| &\leq \underbrace{\left\| F(x_j,h_j)^{-1} \right\|}_{\xrightarrow{j\to\infty} \left\| f'(x_*)^{-1} \right\|} \cdot \underbrace{\left\| F(x_j,h_j) - \int_0^1 f'(sx_j + (1-s)x_*)\, ds \right\|}_{\xrightarrow{j\to\infty} 0} \cdot \|x_{j-1} - x_*\|
\end{aligned}$$

□

Bemerkung. keine quadratische Konvergenz

1.2.3. *Praktische Schrittweitenwahl h_j.*

Bei der Wahl der Schrittweite h_j in Abhängigkeit von f und x_* gilt zu beachten, daß nur eine schlechte Approximation erzielt wird, wenn h_j zu groß gewählt ist. Ist h_j zu niedrig, kommt es zu Auslöschungen und Fehlern.

Standardvariante: Bezeiche ε_{abs} die absolute und ε_{rel} die relative Fehlervorgabe. Sei

$$\tau_j^k := \begin{cases} x_{j-1,k} - x_{j,k} & \text{falls } |x_{j-1,k} - x_{j,k}| \geq \varepsilon_{\text{rel}} \cdot |x_{j,k}| + \varepsilon_{\text{abs}} \\ |x_{j-1} - x_j| & \text{sonst} \end{cases} \qquad k = 1, \ldots, m$$

$$h_j := \begin{pmatrix} \tau_j^1 & \cdots & \tau_j^m \\ \vdots & & \vdots \\ \tau_j^1 & \cdots & \tau_j^m \end{pmatrix}$$

$$\begin{aligned} g_j(x) &= f(x_j) + F(x_j, h_j)(x - x_j) \\ \rightsquigarrow\ g_j(x_j) &= f(x_j) \\ g_j\left(x_j + \tau_j^k e_k\right) &= f(x_j) + F(x_j, h_j)\,\tau_j^k e_k \\ &= f(x_j) + \tau_j^k \frac{1}{\tau_j^k}\left(f\left(x_j + \tau_j^k e_k\right) - f(x_j)\right) \\ &= f\left(x_j + \tau_j^k e_k\right), \qquad k = 1, \ldots, m \end{aligned}$$

1.2.4. *Modifikationen des Newton-Verfahrens und Sekantenverfahrens.*

(1) Wird die Faktorisierung (der F-Matrix) über mehrere Iterationen hinweg beibehalten, so läßt sich der Rechenaufwand verringern.

$$\begin{aligned} x_{j+1} &= x_j + z_{j+1} \\ z_{j+1} \text{ aus } F\left(x_{k_j}, h_{k_j}\right) z_{j+1} &= -f(x_j), \quad 0 \leq k_j \leq j \end{aligned}$$

Extremfall: $k_j = 0$ „Modifizierte Newton-Verfahren"

(2) Dämpfung

$$\begin{aligned} x_{j+1} &= x_j + \alpha_{j+1} z_{j+1} \\ F(x_j, h_j)\, z_{j+1} &= -f(x_j) \end{aligned}$$

Der Dämpfungsparameter $\alpha_{j+1} \in (0, 1]$ wird so bestimmt (z.B. durch Halbieren von α_j), daß

$$|f(x_{j+1})| < |f(x_j)|$$

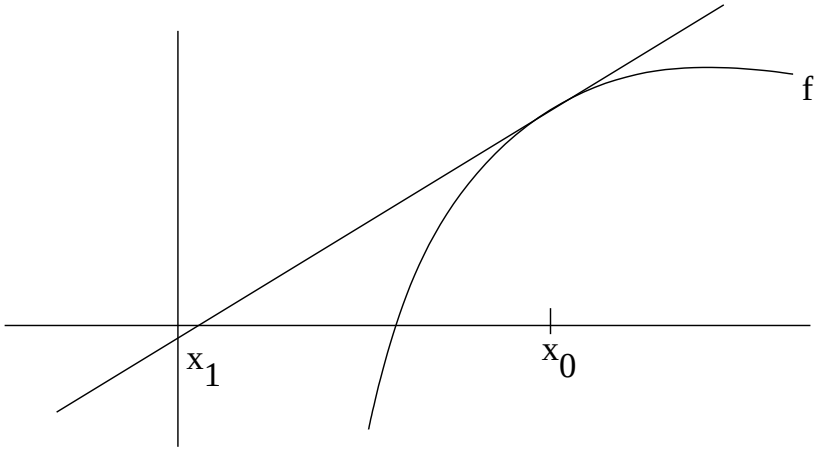


ABBILDUNG 1.3. ohne Dämpfung

Der Vorteil der modifizierten Verfahren ist, daß sie nicht so stark vom Einzugsbereich abhängig sind wie das klassische Newton-Verfahren.

1.3. **Quasi-Newton-Verfahren.**

Wir haben eine Funktion $f(x)$ linear approximiert durch

$$g_j(x) := f(x_j) + A_j(x - x_j)$$

zur Berechnung von x_j. Im Newton-Verfahren war dabei A_j die Jacobi-Matrix ($A_j := f'(x_j)$), im Sekantenverfahren war $A_j := F(x_j, h_j)$.

Bei Funktionen mit hohen Kosten möchte man aber die Zahl der Funktionsaufrufe minimieren.

$$g_j(x) := f(x_j) + A_j(x - x_j)$$

benutzt zur Berechnung von x_{j+1} wieder ein A_j und benötigt damit Funktionsaufrufe, welche wir uns sparen wollen. Nach Broyden (1965) können wir A_j aus A_{j-1} erhalten, ohne dabei auf die Funktion f zurückgreifen zu müssen (sogenannte Aufdatierung).

Festlegung: Quasi-Newton-Bedingung: Quasi-Newton-Bedingung

$$\boxed{g_j(x_{j-1}) = f(x_{j-1})}$$

Alle Verfahren, in denen die Quasi-Newton-Bedingung gilt, bezeichnen wir als Quasi-Newton-Verfahren. Durch diese Bedingung können wir A_j aus A_{j-1} erhalten, ohne dabei auf die Funktion f zurückgreifen zu müssen.

Eine alternative Schreibweise für die Quasi-Newton-Bedingung ist:

$$\boxed{A_j \underbrace{(x_j - x_{j-1})}_{=:z_j} = \underbrace{f(x_j) - f(x_{j-1})}_{=:y_j}}$$

Ist $z_j = 0$, so war $f(x_{j-1}) = 0$, also x_{j-1} schon Nullstelle.

$g'_{j-1}(x) = A_{j-1}$, $g'_j(x) = A_j$

Wir versuchen folgende Festlegung: $g'_j(x)\, z = g'_{j-1}(x)\, z$, $\forall z \perp z_j$.

Dies bedeutet gerade, daß für alle $z \perp z_j$ mit $z_j \neq 0$ gilt:

$$(A_j - A_{j-1})\, z = 0$$

Ansatz :

$$A_j - A_{j-1} = u_j z_j^T$$

mit noch zu bestimmenden $u_j \in \mathbb{R}^m$. Daraus folgt:

$$\begin{aligned} (A_j - A_{j-1})\, z &= u_j z_j^T z = u_j \langle z_j, z \rangle = 0 \quad \text{für } z \perp z_j \\ (A_j - A_{j-1})\, z_j &= |z_j|_2^2 \cdot u_j \\ \rightsquigarrow\ u_j &= \frac{1}{|z_j|_2^2} \cdot (y_j - A_{j-1} z_j) \end{aligned}$$

Dies führt dann zur Rang-Eins-Aufdatierungsformel von Broyden ($\operatorname{rg}(u_j z_j^T) = 1$): Aufdatierungsformel Broyden

$$\boxed{A_j = A_{j-1} + \frac{1}{|z_j|_2^2} \cdot (y_j - A_{j-1} z_j) \cdot z_j^T}$$

Satz 1.6. *Sei $f \in C^1(D, \mathbb{R}^m)$, $D \subseteq \mathbb{R}^m$ offen, $x_* \in D$, $f(x_*) = 0$ und $f'(x_*)$ regulär. Dann gilt:* lokaler Konvergenzsatz

(1) *Für $x_0 \in \bar{B}(x_*, \rho)$, $A_0 \in \bar{B}_{L(\mathbb{R}^m)}(f'(x_*), \delta)$ und $\rho, \delta > 0$ hinreichend klein ist das Quasi-Newton-Verfahren mit Rang-Eins-Aufdatierung nach Broyden entweder durchführbar mit $x_j \overset{j\to\infty}{\longrightarrow} x_*$ oder es endet für ein $j_* \in \mathbb{N}$ mit $x_{j_*} = x_*$.*

(2) *Die Konvergenz ist überlinear.*

Beweis. [S79, S. 143] □

Das Quasi-Newton-Verfahren besitzt nur eine „langsame" überlineare Konvergenz, daher wird es in der Praxis zusammen mit anderen Verfahren benutzt. Anzutreffen ist zum Beispiel die Kombination von Quasi-Newton-Schritten mit Sekanten-Schritten:

- Startwert x_0, $A_0 = F(x_0, h_0)$
- x_1 als Sekantenschritt ($x_1 = x_0 + z_1$, z_1 aus $A_0 z_1 = -f(x_0)$), A_1 aus A_0 durch Aufdatierung
- Quasi-Newton-Schritte solange, wie die Korrekturen z_{j+1} groß genug sind
- Wenn z_{j+1} klein ist, dann neuer Sekantenschritt (anstatt Aufdatierung, sogenannte „Reinitialisierung").

Weitere Aufdatierungen sind Rang-2, inverse Matrix, strukturerhaltende und symmetrieerhaltende Aufdatierungen.

1.4. **Einbettung (Homotopie).**

Bisher mußte x_0 immer hinreichend nahe an einer Nullstelle liegen, damit die Verfahren zu einer Lösung kamen (lokale Konvergenz). In diesem Abschnitt behandeln wir die Frage, wie wir ein geeignetes x_0 finden.

$f(x) = 0$, $f : D \subseteq \mathbb{R}^m \to \mathbb{R}^m$,

$H : D \times [0,1] \to \mathbb{R}^m$, $H(x,1) \equiv f(x)$

Die Lösung der Gleichung $H(x,0) = 0$ sei x_0 und „einfach" zu beschaffen.

Betrachten die Gleichungen $H(x,t) = 0$ für alle $t \in [0,1]$ mit den Lösungen $x(t) \in D$

Wenn $x(t)$ stetig ist, dann existiert für jedes $\bar{t}$ ein $U(\bar{t}) \subseteq [0,1]$, so daß x_0 im Einzugsbereich von $H(x,t) = 0$, $t \in U(\bar{t})$, liegt.

Beispiel. künstliche Einbettung

- $H(x,t) = f(x) + (t-1) f(x_0)$, $x_0 \in D$ fest
- $H(x,t) = tf(x) + (1-t)(Ax - b)$

Voraussetzungen:

- $H : D \times [0,1] \to \mathbb{R}^m$ stetig
- $D_x H(x,t)$ stetig in (x,t)
- Es existiert eine Abbildung $x \in C([0,1], \mathbb{R}^m)$ stetig mit $x(t) \in D$, $H(x(t),t) = 0$, $D_x H(x(t),t)$ regulär, $t \in [0,1]$ und es gilt $x_0 = x(0), x_* = x(1)$. (Lokale Eindeutigkeit und Existenz von $x(t)$ folgt aus implizitem Funktionentheorem.)

Wir betrachten eine Zerlegung $0 = t_0 < t_1 < \ldots < t_N = 1$ und Iterationsschritte $k_1, \ldots, k_N$.

Wir lösen die Gleichung $H(x,t_1) = 0$ mit dem Startwert x_0, führen k_1 Iterationsschritte aus und erhalten k_1 Iterationswerte $x_{1,j}$ ($j \in \{1, \ldots, k_1\}$). Anschließend lösen wir $H(x,t_2)$ mit dem Startwert $x_{2,0} = x_{1,k_1}$ usw.

Wir nähern uns damit der eigentlichen Problemstellung schrittweise.

eingebettetes Newton-Verfahren

Für die Iterationen zur Lösung der Gleichungen $H(x,t_i) = 0$ kann man ein beliebiges Verfahren benutzen (Newton, Sekanten, etc.). An dieser Stelle betrachten wir das eingebettete Newton-Verfahren:

Bestimme $x_{1,0}$ mit $H(x_{1,0}, 0) = 0$. Für $i > 1$ sei $x_{i,0} := x_{i-1,k_{i-1}}$.

$$\begin{array}{rcll} x_{i,j+1} & := & x_{i,j} + z_{i,j+1} & \\ \text{mit } z_{i,j+1} \text{ aus } H_x(x_{i,j}, t_i)\, z_{i,j+1} & = & -H_{i,j}(x_{i,j}, t_i) & j = 0, \ldots, k_i - 1,\ i = 1, \ldots, N \end{array}$$

Beispiel.

$$\begin{aligned} H(x,t) &= f(x) + (t-1) f(x_0) \\ D_x H(x,t) &= f'(x) \\ \rightsquigarrow f'(x_{i,j}) z_{i,j+1} &= -(f(x_{i,j}) + (t_i - 1) f(x_0)) \end{aligned}$$

Satz 1.7. *Sei $H : D \times [0,1] \to \mathbb{R}^m$ stetig, $D \subseteq \mathbb{R}^m$ offen, $D_x H$ stetig und $H(x,1) = f(x)$. Es existiere eine Funktion $x \in C([0,1], \mathbb{R}^m)$ mit $x(t) \in D$, $H(x(t), t) = 0$, $D_x H(x(t), t)$ regulär für $t \in [0,1]$.* Konvergenzsatz
Dann gilt: Für $k \in \mathbb{N}$ und hinreichend feine Zerlegungen $0 < t_0 < t_1 < \ldots < t_N = 1$, $k_i = k$, $i = 1, \ldots, N$, ist das eingebettete Newton-Verfahren durchführbar und liefert mit $x_{N,k}$ einen Wert, der zum Einzugsbereich der Nullstelle $x(1)$ der Gleichung $H(x(1), 1) = 0$ des Newton-Verfahrens gehört.

Beweis. Sei $T = \{x(t) | t \in [0,1]\}$ kompakt, $M \subseteq D \subseteq \mathbb{R}^m$ kompakt mit $T \subset M \setminus \partial M$.

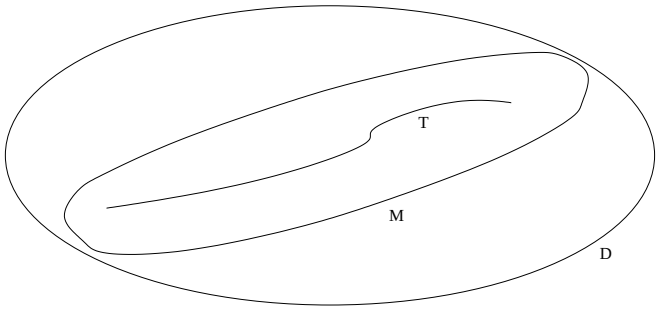


ABBILDUNG 1.4. $T \subset M \setminus \partial M \subseteq D \subseteq \mathbb{R}^m$

$D_x H(x,t)$ ist auf $M \times [0,1]$ stetig und sogar gleichmäßig stetig (da M kompakt), d.h.

$$\forall \varepsilon > 0 \, \exists \delta(\varepsilon) > 0 \, \forall (x,t), (\bar{x}, \bar{t}), |x - \bar{x}| + |t - \bar{t}| \leq \delta(\varepsilon) : \quad \|D_x H(x,t) - D_x H(\bar{x}, \bar{t})\| \leq \varepsilon.$$

Sei

$$G(t) := D_x H(x(t), t)^{-1}, \quad t \in [0,1]$$

G ist eine stetige Matrixfunktion. Da eine stetige Funktion auf einem kompakten Intervall beschränkt ist (Weierstraß), existiert ein $K > 0$ mit

$$\|G(t)\| \leq K, \quad t \in [0,1].$$

Wir wählen ein hinreichend kleines $\varepsilon > 0$ mit $\varepsilon \cdot K < 1$. Aufgrund der gleichmäßigen Stetigkeit gilt dann für $x \in \bar{B}(x(t), \delta(\varepsilon)) \subseteq M$ und $t \in [0,1]$:

$$\|D_x H(x,t) - D_x H(x(t), t)\| \leq \varepsilon$$

Aus dem Störungslemma (L. 1.1) erhalten wir die Aussage, daß $D_x H(x,t)$ regulär ist und das für $x \in \bar{B}(x(t), \delta(\varepsilon))$, $t \in [0,1]$, gilt:

$$\left\| D_x H(x,t)^{-1} \right\| \leq \frac{K}{1 - \varepsilon \cdot K}$$

Wir betrachten das Newton-Verfahren für $H(x,t)=0$ mit festem t: Sei $x_{t,0}\in\bar{B}(x(t),\delta(\varepsilon))$.

$$\begin{aligned}
x_{t,1}-x(t) &= x_{t,0}-x(t)-D_xH(x_{t,0},t)^{-1}H(x_{t,0},t)\\
&= x_{t,0}-x(t)-D_xH(x_{t,0},t)^{-1}\left(H(x_{t,0},t)-\underbrace{H(x(t),t)}_{=0}\right)\\
&\overset{(L.\,1.2)}{=} D_xH(x_{t,0},t)^{-1}\left(D_xH(x_{t,0},t)-\int_0^1 D_xH(sx_{t,0}+(1-s)x(t),t)\,ds\right)\cdot(x_{t,0}-x(t))\\
|x_{t,1}-x(t)| &\le \frac{K}{1-\varepsilon\cdot K}\cdot\underbrace{\left\|D_xH(x_{t,0},t)-\int_0^1 D_xH(sx_{t,0}+(1-s)x(t),t)\,ds\right\|}_{\le\varepsilon}\cdot|x_{t,0}-x(t)|\\
&\le \kappa(\varepsilon)\cdot|x_{t,0}-x(t)| \text{ mit } \kappa(\varepsilon):=\frac{\varepsilon\cdot K}{1-\varepsilon\cdot K}
\end{aligned}$$

Für $\varepsilon\cdot K<\frac{1}{2}$ ist $\kappa(\varepsilon)<1$. Wir wählen dann ε so, daß $\varepsilon\cdot K<\frac{1}{2}$. Das Newton-Verfahren ist durchführbar und es gilt

$$|x_{t,j+1}-x(t)|\le\kappa(\varepsilon)^{j+1}\cdot|x_{t,0}-x(t)|\le\kappa(\varepsilon)^{j+1}\cdot\delta(\varepsilon)\qquad j=1,\dots,k-1 \tag{1.1}$$

Wir setzen $\delta_1:=\delta(\varepsilon)$. Da $x\in C([0,1],\mathbb{R}^m)$ stetig ist (auf der kompakten Menge $[0,1]$), ist x auch gleichmäßig stetig.

Zu $\delta>0$ existiert eine Zerlegung $0=t_0^{(\delta)}<t_1^{(\delta)}<\ldots<t_{N^{(\delta)}}^{(\delta)}=1$ mit der Eigenschaft, daß

$$\left|x\left(t_i^{(\delta)}\right)-x\left(t_{i+1}^{(\delta)}\right)\right|\le\delta,\quad i=0,\dots,N^{(\delta)}-1$$

Wir wählen $\delta_2\le\left(1-\kappa(\varepsilon)^k\right)\cdot\delta_1$ (da aus (1.1) folgt, daß $|x_{t,k}-x(t)|\le\kappa(\varepsilon)^k\cdot\delta(\varepsilon)$) und fixieren eine zugehörige Zerlegung $0=t_0<t_1<\ldots<t_N=1$. Dann ist

$$\begin{aligned}
|x_{i,k}-x(t_{i+1})| &\le |x_{i,k}-x(t_i)|+|x(t_i)-x(t_{i+1})|\\
&\le \kappa(\varepsilon)^k\cdot\delta_1+\delta_2\le\delta_1
\end{aligned}$$

Zu Beginn ist: $x_{1,0}=x_0$, also

$$|x_0-x(t_1)|\le\delta_2\le\delta_1.$$

□

1.4.1. *Kurvenverfolgung.*

Gesamte Lösungsmenge $S:=\{(x,t)|\,H(x,t)=0\}$

Literatur: [DH93, §4.4 (S. 115-131) Parameterabhängige nichtlineare Gleichungen]

2. Ausgleichsprobleme (Kleinste-Quadrate-Probleme) und Gauss-Newton-Verfahren

Sei $f \in C^1(D, \mathbb{R}^n)$, $D \subseteq \mathbb{R}^m$ offen und $n \gg m$.

$f(x) = 0$ ist ein überbestimmtes Gleichungssystem

$$\boxed{\varphi(x) := |f(x)|_2^2, \quad x \in D}$$

Gesucht ist $x_* \in D$ mit

$$\boxed{\varphi(x_*) = \min_{x \in D} \varphi(x)}$$

Kleinste-Quadrate-Problem

x_* heißt kleinste-Quadrate-Lösung

Beispiel. Datenanalyse, Datenkompression (Regressionsanalyse)

(t_j, y_j), $j = 1, \ldots, n$, Meßwerte (n groß: 10^2, 10^3)

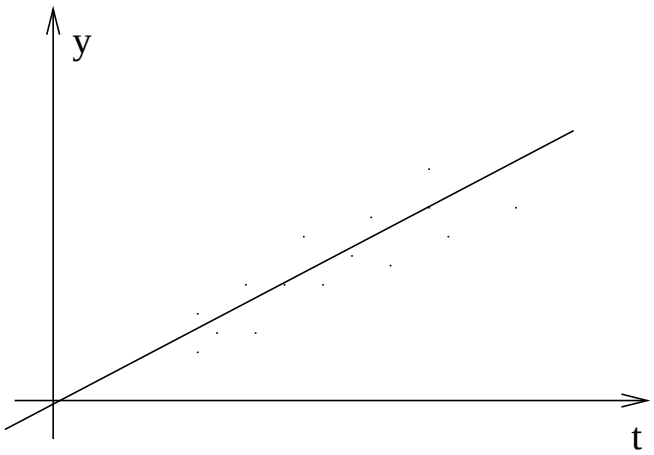


Abbildung 2.1. Meßwerte mit Regressionsgerade

Modellfunktion: $g(t, x_1, \ldots, x_m)$ mit kleiner Anzahl von Parametern ($m = 2, 3$), zum Beispiel:

$$\begin{aligned} g(t, x_1, x_2) &:= x_1 t + x_2 && \text{lineare Regression} \\ g(t, x_1, \ldots, x_m) &:= x_1 t^{m-1} + \ldots + x_m && \text{polynomieller Zusammenhang} \\ g(t, x_1, x_2) &:= e^{t x_1} + x_2 && \text{exponentieller Zusammenhang} \end{aligned}$$

$$\min_{x \in D} \frac{1}{n} \sum_{j=1}^{n} \left(g(t_j, x) - y_j\right)^2$$

$$f(x) := \begin{pmatrix} g(t_1, x) - y_1 \\ \vdots \\ g(t_n, x) - y_n \end{pmatrix}$$

$$\rightsquigarrow \qquad \min_{x \in D} \varphi(x) = \min_{x \in D} |f(x)|_2^2$$

2.1. Lineare Ausgleichsprobleme / lineare überbestimmte Gleichungen.

$f(x) = Ax - b$, $A \in L(\mathbb{R}^m, \mathbb{R}^n)$, $b \in \mathbb{R}^n$

$$\boxed{\varphi(x) := \frac{1}{2}|Ax-b|_2^2, \quad x \in \mathbb{R}^m}$$

φ ist stetig-differenzierbar

Lemma 2.1. *Sei* $\ker A = \{0\}$ $(\mathrm{rg}A = m)$. *Dann besitzt* φ *genau ein Minimalelement* $x_* \in \mathbb{R}^m$ *und es gilt*

$$\boxed{x_* = \left(A^TA\right)^{-1}A^Tb}$$

Beweis. A^TA ist regulär, denn sei $A^TAz = 0$, dann ist $Az \in \ker\left(A^T\right) = (\mathrm{im}A)^T$, also $Az = 0$ und damit $z \in \ker A = \{0\}$.

$$\begin{aligned}
\varphi'(x)z &= \lim_{h\to 0}\frac{1}{h}\{\varphi(x+hz) - \varphi(x)\} \\
&= \frac{1}{2}\lim_{h\to 0}\frac{1}{h}\{\langle A(x+hz)-b, A(x+hz)-b\rangle - \langle Ax-b, Ax-b\rangle\} \\
&= \frac{1}{2}\{\langle Ax-b, Az\rangle + \langle Az, Ax-b\rangle\} \\
&= \langle Ax-b, Az\rangle \\
&= \left\langle A^T(Ax-b), z\right\rangle \\
&= \left[A^T(Ax-b)\right]^T z \\
\rightsquigarrow\ \varphi'(x)^T &= A^T(Ax-b) =: \nabla\varphi(x)
\end{aligned}$$

Also ist

$$\varphi'(x_*) = 0 \quad \Leftrightarrow \quad \underbrace{A^TA}_{\text{reg.}}x_* = A^Tb \quad \Leftrightarrow \quad x_* = \left(A^TA\right)^{-1}A^Tb$$

Die Hesse-Matrix $\nabla^2\varphi(x) = A^TA$ ist positiv definit, weil $\left\langle A^TAx, x\right\rangle = \langle Ax, Ax\rangle > 0\ \forall x \neq 0$. $\square$

Bemerkung. Die Transformation von $Ax = b$ zu $A^TAx = A^Tb$ bezeichnet man auch als Gauß-Transformation und $A^TAx = A^Tb$ heißt „Normalgleichung".

Praktische Bestimmung von x_*:

(1) Hat A kleine Konditionszahl hat so kann man $A^TAx = A^Tb$ durch Cholesky-Zerlegung lösen, denn es ist $\mathrm{cond}_2\left(A^TA\right) = \left(\mathrm{cond}_2(A)\right)^2$.

(2) Ist A schlecht konditioniert, so sollte man das Householder-Verfahren benutzen:

$$\begin{aligned}
QA &= \begin{pmatrix} R \\ 0 \end{pmatrix}, \quad Q \in L(\mathbb{R}^n), Q^T = Q^{-1}, R \in L(\mathbb{R}^m) \text{ reg. und obere } \triangle - \text{Matrix} \\
\varphi(x) &= \frac{1}{2}|Ax-b|_2^2 \\
&= \frac{1}{2}|Q(Ax-b)|_2^2 \\
&= \frac{1}{2}\left|\begin{pmatrix} R \\ 0 \end{pmatrix}x - \begin{pmatrix} \tau_1 \\ \tau_2 \end{pmatrix}\right|_2^2, \quad \begin{pmatrix} \tau_1 \\ \tau_2 \end{pmatrix} := Qb \\
&= \frac{1}{2}|Rx-\tau_1|_2^2 + \frac{1}{2}|\tau_2|_2^2 \\
\rightsquigarrow\ x_* &= R^{-1}\tau_1 \qquad \text{Also Gleichungssystem } Rx = \tau_1 \text{ lösen.}
\end{aligned}$$

Bemerkung. $A^+ := \left(A^TA\right)^{-1}A^T$ heißt Moore-Penrose Inverse

$A^+AA^+ = A^+$, $AA^+A = A$, $A^+A = (A^+A)^T$, $AA^+ = (AA^+)^T \rightsquigarrow \ker A^T \leftrightarrow \mathrm{im}A$

2.2. **Nichtlineare Ausgleichsprobleme.**

$f(x) = 0$, x_0 Startwert

Wir betrachten die zur Linearisierung

$$g_0(x) = f(x_0) + f'(x_0)(x - x_0)$$

gehörige Aufgabe $g_0(x) = 0$. Dazu sei

$$x_1 = x_0 + z_1 \qquad \text{mit } z_1 \text{ aus } f'(x_0) z_1 = -f(x_0).$$

Hat $f'(x_0)$ vollen Rang, so ist die kleinste Quadrate Lösung dieser Gleichung gegeben durch

$$z_1 = -\left(f'(x_0)^T f'(x_0)\right)^{-1} f'(x_0)^T f(x_0)$$

Satz 2.2. *Sei $f \in C^1(D, \mathbb{R}^n)$, $D \subseteq \mathbb{R}^m$ offen, $n \geq m$,*

$$\varphi(x) := \frac{1}{2} |f(x)|_2^2, \quad x \in D.$$

Sei $x \in D$ fest, der Rang von $f'(x)$ gleich m, $f'(x)^T f(x) \neq 0$.
Dann ist die Gauß-Newton-Richtung

Gauß-Newton-Richtung

$$\boxed{z_{GN} := -\left(f'(x)^T f'(x)\right)^{-1} f'(x)^T f(x)}$$

eine Abstiegsrichtung (s. Seite 19) für φ in x.

Beweis. φ ist C^1, $x \in D$ fest, $z \in \mathbb{R}^m$ beliebig,

$$\begin{aligned} \varphi'(x) z &= \left\langle f'(x)^T f(x), z \right\rangle \\ \nabla\varphi(x) &= f'(x)^T f(x). \end{aligned}$$

Dann ist

$$\underbrace{\varphi(x + \alpha z) - \varphi(x)}_{<0} = \underbrace{\alpha \cdot \varphi'(x) z + o(\alpha)}_{\varphi'(x)z<0 \Leftrightarrow z \text{ Abstiegsrichtung}}, \quad \alpha > 0$$

$$\langle \nabla\varphi(x), z_{GN} \rangle = -\left\langle f'(x)^T f(x), \underbrace{\left(f'(x)^T f'(x)\right)^{-1}}_{=:M=RR^T} f'(x)^T f(x) \right\rangle, \; M \text{ symm., pos.def.}$$

$$= -\left| R^T \underbrace{f'(x)^T f(x)}_{\neq 0} \right|_2^2 < 0 \qquad \text{da } R \text{ regulär ist}$$

□

Schrittweitenwahl nach Armijo: Sei $\gamma \in \left(0, \frac{1}{2}\right)$ fest

Armijo-Schritt-weitenwahl

x_j, $\varphi(x_j)$ seien bekannt, z_{j+1} sei Gauß-Newton-Richtung, $\varphi'(x_j) z_{j+1}$ sei berechnet.

Betrachte zwei Geraden:

$$\begin{aligned} g_1 &: \varphi(x_j) + \gamma\alpha\varphi'(x_j) z_{j+1} \\ g_2 &: \varphi(x_j) + \alpha\varphi'(x_j) z_{j+1} \end{aligned}$$

Test, ob für $\alpha = 1$ gilt, daß $\varphi(x_j + \alpha z_{j+1}) \leq \varphi(x_j) + \gamma\alpha\varphi'(x_j) z_{j+1}$.

Wenn ja, so $\alpha_{j+1} := 1$.

Wenn nein, so $\alpha := \frac{\alpha}{2}$, neuer Test usw.

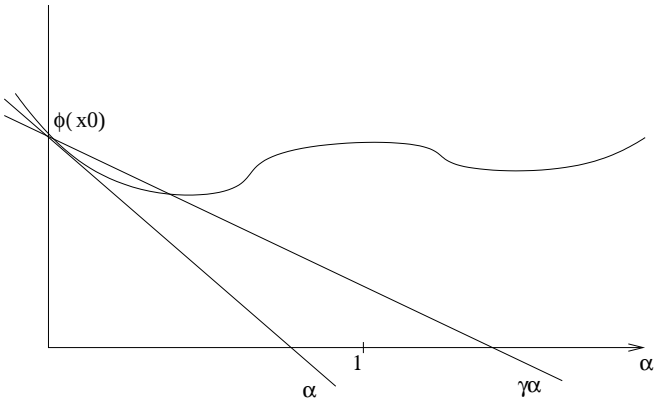


ABBILDUNG 2.2. Schrittweitenwahl nach Armijo

$\rightsquigarrow \alpha_{j+1} \in (0,1]$ wird so bestimmt, daß

$$\boxed{\varphi(x_j + \alpha_{j+1} z_{j+1}) \leq \varphi(x_j) + \gamma \alpha_{j+1} \cdot \varphi'(x_j) z_{j+1}}$$

gedämpftes Gauß-Newton-Verfahren

Definition. Gedämpftes Gauß-Newton-Verfahren mit Startwert x_0:

$$\boxed{\begin{array}{rcll} x_{j+1} & = & x_j + \alpha_{j+1} z_{j+1} & \\ z_{j+1} \text{ aus } f'(x_j) z_{j+1} & = & -f(x_j) \quad \text{Kleinste-Quadrate-Lösung} & j \geq 0 \\ \alpha_{j+1} \text{ nach Armijo} & & & \end{array}}$$

Konvergenzsatz

Satz 2.3. *Sei* $f \in C^1(\mathbb{R}^m, \mathbb{R}^n)$, $n \geq m$, $\varphi(x) := \frac{1}{2}|f(x)|_2^2$, $x \in \mathbb{R}^m$. *Sei* $x_0 \in \mathbb{R}^m$ *fest und*

$$N_0 := \{x \in \mathbb{R}^m \,|\, |f(x)|_2 \leq |f(x_0)|_2\}$$

Für alle $x \in N_0$ *sei der Rang von* $f'(x)$ *gleich* m *und es gelte mit Konstanten* $K, c > 0$*:*

$$\begin{array}{rcl} \|f'(x)\|_2 & \leq & K \\ |f'(x) z|_2 & \geq & c \cdot |z|_2, \quad \forall z \in \mathbb{R}^m. \end{array}$$

$f'(x)$ *sei Lipschitz-stetig auf* N_0 *(d.h.* $\|f'(x) - f'(\bar{x})\|_2 \leq L \cdot |x - \bar{x}|_2$*). Dann gilt:*

(1) *Das gedämpfte Gauß-Newton-Verfahren mit Startwert* x_0 *ist durchführbar und liefert eine monoton fallende Folge* $\{\varphi(x_j)\}_{j \geq 0}$*. Es gilt* $\varphi'(x_j) \xrightarrow{j \to \infty} 0$*.*

(2) *Ist* N_0 *kompakt und hat* γ *auf* N_0 *genau einen stationären Punkt* $x_* \in N_0$*, so endet das Verfahren entweder wegen* $\varphi'(x_{k_*}) = 0$ *in* $x_{k_*} = x_*$ *oder es ist unendlich fortsetzbar und* $x_j \xrightarrow{j \to \infty} x_*$*.*

(3) *Gilt in* (2) *noch* $f(x_*) = 0$*, so gibt es einen Index* $j_* \in \mathbb{N}$ *mit* $\alpha_j = 1$ *für* $j \geq j_*$ *und quadratischer Konvergenz ab* j_**.*

Beweis. [S79, §10.2, Satz 10.2.5, Algorithmus 10.2.1] □

Bemerkung 2.4. Anstelle des Newton-Verfahren kann auch ein Quasi-Newton- oder Sekantenverfahren benutzt werden. Zum Beispiel Sekanten-Matrix Aufdatierung:

$$g_j(x) := f(x_j) + \underbrace{f'(x_j)}_{A_j}(x - x_j)$$

3. Minimierungsaufgaben (Optimierungsaufgaben)

Wir betrachten eine Funktion $\varphi \in C^2(\mathbb{R}^m, \mathbb{R})$, $U \subseteq \mathbb{R}^m$. Aufgabe ist es, $\varphi(x)$ für $x \in U$ zu minimieren.

Bemerkung. Wir bezeichnen dabei φ als Zielfunktion und U als die Menge der zulässigen Punkte.

Wir betrachten zwei Typen:

(1) $U = \mathbb{R}^m$ oder $U \subseteq \mathbb{R}^m$ offen und wir betrachten $\varphi'(x) = 0$. Minimierungsprobleme dieses Typs bezeichnen wir als freie bzw. unrestriktive Minimierungsaufgaben.

(2) $U \subset \mathbb{R}^m$ nichtleer und nicht offen, U wird häufig durch (Un)Gleichungen beschrieben:

$$U = \{ x \in \mathbb{R}^m \mid g_i(x) \leq 0, i = 1, \ldots, p;\ g_{p+j}(x) = 0, j = 1, \ldots, q \},\ g_j \in C^1(\mathbb{R}^m, \mathbb{R}),$$

zum Beispiel

$$U = \left\{ x \in \mathbb{R}^2 \,\middle|\, x_1^2 + y_1^2 - 1 = 0,\ -x^2 \leq 0 \right\}$$

bezeichnen wir als Minimierungsprobleme mit Nebenbedingungen bzw. Restriktionen.

3.1. Freie Minimierungssprobleme.

$$\boxed{\min \{ \varphi(x) \mid x \in \mathbb{R}^m \}} \tag{3.1}$$

Definition. $x_* \in \mathbb{R}^m$ heißt globale Lösung von (3.1), falls $\varphi(x_*) \leq \varphi(x)$ für alle $x \in \mathbb{R}^m$.

$x_* \in \mathbb{R}^m$ heißt lokale Lösung von (3.1), falls $\varphi(x_*) \leq \varphi(x)$ für alle $x \in U(x_*)$.

Definition. $z \in \mathbb{R}^m$ heißt Abstiegsrichtung zu φ in $x \in \mathbb{R}^m$, falls $\varphi(x + \alpha z) < \varphi(x)$ für alle hinreichend kleinen $\alpha > 0$.

Bemerkung. z ist Abstiegsrichtung für φ in $x \Leftrightarrow \varphi'(x) z < 0$.

Definition. $x_* \in \mathbb{R}^m$ heißt stationärer Punkt (Extremalpunkt), falls

$$\varphi'(x_*) z \geq 0 \quad \forall z \in \mathbb{R}^m$$

$(\Leftrightarrow \varphi'(x_*) = 0)$

Bemerkung. Dabei bezeichnen wir $\varphi'(x)^T = \nabla\varphi(x)$ als Gradient und $\nabla^2\varphi(x)$ als Hessesche Matrix (bei $\varphi \in C^2(\mathbb{R}^m, \mathbb{R})$)

Die steilste Abstiegsrichtung zu φ in x ist — Gradientenverfahren

$$\boxed{z_G := -\nabla\varphi(x)}$$

(„Gradientenverfahren" bis circa 1965).

Lemma 3.1. *Sei $\varphi \in C^1(\mathbb{R}^m, \mathbb{R})$, $x \in \mathbb{R}^m$ fest, $\nabla\varphi(x) \neq 0$. Sei $H \in L(\mathbb{R}^m)$ symmetrisch und positiv definit. Dann ist*

$$z := -H \cdot \nabla\varphi(x)$$

Abstiegsrichtung zu φ in x.

Beweis. $H = RR^T$,

$$\varphi'(x) z = -\langle \nabla\varphi(x), H\nabla\varphi(x) \rangle = -\left\langle R^T\nabla\varphi(x), R^T\nabla\varphi(x) \right\rangle = -\left| R^T\nabla\varphi(x) \right|_2^2 < 0$$

□

Folgerung. Die Newton-Richtung — Newton-Richtung

$$\boxed{z_N := -\left(\nabla^2\varphi(x)\right)^{-1} \cdot \nabla\varphi(x)}$$

für φ in x, $\varphi \in C^2(\mathbb{R}^m, \mathbb{R})$, $x \in \mathbb{R}^m$, $\nabla\varphi(x) \neq 0$ ist eine Abstiegsrichtung, falls die Hessematrix $\nabla^2\varphi(x)$ positiv definit ist.

Beispiel. $m = 2$

$$\begin{aligned}
\varphi(x_1, x_2) &:= -x_1^2 + x_2^2 + 1 \\
\nabla\varphi(x) &= \begin{pmatrix} -2x_1 \\ 2x_2 \end{pmatrix} \\
\nabla^2\varphi(x) &= \begin{pmatrix} -2 & 0 \\ 0 & 2 \end{pmatrix} \\
z_G &:= -\nabla\varphi(x) = \begin{pmatrix} 2x_1 \\ -2x_2 \end{pmatrix} \\
z_N &= -\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\
\langle \nabla\varphi(x), z_N \rangle &= 2\left(x_1^2 - x_2^2\right)
\end{aligned}$$

Lemma 3.2. *Sei $c \in \mathbb{R}^m$ und $W \in L(\mathbb{R}^m)$ symmetrisch und positiv definit. Dann hat*

$$\psi(x) := c^T x + \frac{1}{2} x^T W x = \langle c, x \rangle + \frac{1}{2} \langle Wx, x \rangle, \quad x \in \mathbb{R}^m,$$

auf $\mathbb{R}^m$ genau einen stationären Punkt x_* und es gilt

$$\boxed{x_* = -W^{-1} \cdot c}$$

x_* ist globales Minimum: $\psi(x_*) < \psi(x)\ \forall x \in \mathbb{R}^m$, $x \neq x_*$.

Beweis. $\psi'(x) = c^T + (Wx)^T$, $\nabla\psi(x) = c + Wx$

$\nabla\psi(x_*) = 0 \Leftrightarrow c + Wx_* = 0$ (inhomogene lineare Gleichung)

$\nabla^2\psi(x_*) = W$ ist positiv definit, d.h. x_* ist Minimum.

□

Bemerkung. ψ heißt quadratische Zielfunktion.

Ziel war/ist die Minimierung von $\varphi(x)$, $x \in \mathbb{R}^m$ (3.1). Seien nun x_j, $\nabla\varphi(x_j)$ bereits berechnet, $\varphi \in C^2(\mathbb{R}^m, \mathbb{R})$

$$\begin{aligned}
\rho_j(x) &:= \varphi(x_j) + \langle \nabla\varphi(x_j), x - x_j \rangle + \frac{1}{2} \left\langle \nabla^2\varphi(x_j)(x - x_j), x - x_j \right\rangle \\
&= \varphi(x_j) + \varphi'(x_j) \cdot (x - x_j) + \frac{1}{2} \cdot (x - x_j)^T \cdot \nabla^2\varphi(x_j) \cdot (x - x_j) \\
\rho_j(x_j) &= \varphi(x_j) \\
\nabla\rho_j(x) &= \nabla\varphi(x_j) + \nabla^2\varphi(x_j)(x - x_j) \\
\nabla\rho_j(x_j) &= \nabla\varphi(x_j)
\end{aligned}$$

Falls die Hessematrix $\nabla^2\varphi(x_j)$ positiv definit ist, können wir Lemma 3.2 anwenden und erhalten ein x_*, welches das Minimum dieser Funktionen ist:

$$x_* - x_j = \underbrace{\left(-\nabla^2\varphi(x_j)\right)^{-1} \cdot \nabla\varphi(x_j)}_{\text{Newton-Richtung}}$$

Wir betrachten nun den Fall, daß x_j, $\nabla\varphi(x_j)$, x_{j-1} und $\nabla\varphi(x_{j-1})$ bereits berechnet sind.

$$\rho_{j-1}(x) := \varphi(x_{j-1}) + \langle \nabla\varphi(x_{j-1}), x - x_{j-1} \rangle + \frac{1}{2} \langle H_{j-1} \cdot (x - x_{j-1}), x - x_{j-1} \rangle$$

ρ_{j-1} sei verwendet worden, um x_j zu berechnen. Wir setzen desweiteren voraus, daß H_{j-1} symmetrisch und positiv definit ist.

$$\begin{aligned}
x_j &= x_{j-1} + \alpha_j z_j \\
z_j \text{ aus } H_j z_j &= -\nabla\varphi(x_{j-1}) \text{ bestimmt}
\end{aligned}$$

Sei

$$\varphi_j(x) := \varphi(x_j) + \langle \nabla\varphi(x_j), x - x_j\rangle + \frac{1}{2}\langle H_j(x - x_j), x - x_j\rangle$$

Idee: Aufdatierung von H_j aus H_{j-1} so, daß H_j symmetrisch und positiv definit ist.

Es gilt:

$$\begin{aligned} \varphi_j(x_j) &= \varphi(x_j) \\ \nabla\varphi_j(x) &= \nabla\varphi(x_j) + H_j(x - x_j) \\ \nabla\varphi_j(x_j) &= \nabla\varphi(x_j) \end{aligned}$$

Quasi-Newton-Bedingung:

$$\begin{aligned} H_j(x_j - x_{j-1}) &= \nabla\varphi(x_j) - \nabla\varphi(x_{j-1}) \\ &\Updownarrow \\ \nabla\varphi_j(x_{j-1}) &= \nabla\varphi(x_{j-1}) \end{aligned}$$

[Aufdatierungsformeln: siehe Jochen Werner, "Numerische Mathematik 2", Vieweg 1992]

Man spricht hier auch von Quasi-Newton-Richtungen: Quasi-Newton-Richtung

$$\boxed{z_{j+1} := -H_j^{-1} \cdot \nabla\varphi(x_j) \quad H_j \text{ aus Aufdatierungsformel}}$$

Diese Verfahren werden BFGS-Verfahren genannt: $\approx$ 1970 Broyden, Fletcher, Goldberg, Shanno. Es sind die aktuell besten Verfahren für glatte Minimierung ohne Nebenbedingung mit moderatem m.

Schrittweitenwahl ausgehend von x_j in Richtung z_{j+1} (Abstiegsrichtung): Schritt-weitenwahl

$$\psi(\alpha) := \varphi(x_j + \alpha z_{j+1})$$

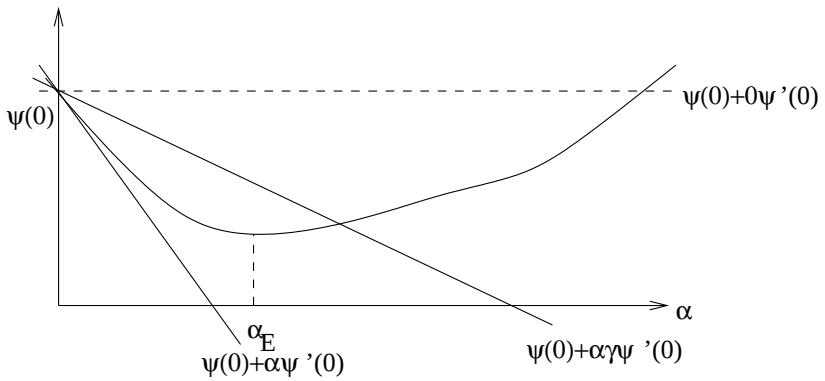


ABBILDUNG 3.1. „Exakte" Schrittweite und Armijo-Schrittweitenwahl

(1) „Exakte" Schrittweite: kleinstes positives $\alpha_E > 0$ mit $\psi'(\alpha_E) = 0$

$$\begin{aligned} \psi(0) &= \varphi(x_j) \\ \psi'(\alpha) &= \varphi(x_j + \alpha z_{j+1}) \cdot z_{j+1} \\ \psi'(0) &= \varphi(x_j) \cdot z_{j+1} < 0 \end{aligned}$$

(2) Armijo-Schrittweite, $\gamma \in \left(0, \frac{1}{2}\right)$

$\alpha_A > 0$ mit $\psi(\alpha_A) \le \psi(0) + \alpha_A \cdot \gamma \cdot \psi'(0)$ (Armijo-Bedingung (A))

$$\varphi(x_{j+1}) \le \varphi(x_j) + \alpha_A \cdot \gamma \cdot \varphi'(x_j) \cdot z_{j+1}$$

(3) Powell-Schrittweite (Wolfe-Schrittweite)

$\alpha_{PA} > 0$ und Bedingung (A) und zusätzlich wird verlangt, daß

$$\psi'(\alpha_{PA}) \geq \beta \cdot \psi'(0)$$

gilt mit einem Parameter $\beta \in (\gamma, 1)$

Überlineare Konvergenz kann für BFGS-Verfahren unter entsprechenden Voraussetzungen bewiesen werden.

3.2. Elementare Ansätze für Minimierung mit Nebenbedingungen.

(3.2) $$\boxed{\min\{\varphi(x) \mid x \in U\}}$$

$U \subseteq \mathbb{R}^m$ sei nicht offen

- Globale Lösung: $x_* \in U$ mit $\varphi(x_*) \leq \varphi(x)$, $x \in U$.
- Lokale Lösung: $x_* \in U$ mit $\varphi(x_*) \leq \varphi(x)$ für $x \in U \cap \mathcal{U}(x_*)$, $\mathcal{U}(x_*) \subseteq \mathbb{R}^m$ Umgebung von x_*

(1) U sei abgeschlossen und konvex, z.B. $U := \{x \in \mathbb{R}^m \mid x_i \geq 0\}$

Sei $x \in U$, $z = y - x$

$$\varphi(x + \alpha(y - x)) - \varphi(x) = \alpha \cdot \langle \nabla\varphi(x), y - x\rangle + o(\alpha)$$

$y - x$ ist Abstiegsrichtung für φ in $x \Leftrightarrow \langle \nabla\varphi(x), y - x\rangle < 0$

$y \in U$ definiert dann eine „zulässige Abstiegsrichtung".

projizierte Gradienten

Extremalbedingung:

$$\boxed{\forall y \in U : \langle \nabla\varphi(x_*), y - x_*\rangle \geq 0}$$

Dies definiert das Verfahren der projizierten Gradienten.

(2) U ist durch Gleichungen gegeben, z.B. $U := \{x \in \mathbb{R}^m \mid g(x) = 0\}$, $g \in C^1(\mathbb{R}^m, \mathbb{R})$

Lagrange-Multiplikatoren

Methode der Lagrange-Multiplikatoren

Ersatzfunktion:

$$\boxed{\psi(x, \lambda) := \varphi(x) + \lambda g(x), \quad x \in \mathbb{R}^m, \lambda \in \mathbb{R}}$$

Minimiere $\psi(x, \lambda)$:

(3.3) $$\boxed{\min\{\psi(x, \lambda) \mid x \in \mathbb{R}^m, \lambda \in \mathbb{R}\}}$$

$$\nabla\psi(x, \lambda) = \begin{pmatrix} \nabla\varphi(x) + \lambda \cdot \nabla g(x) \\ g(x) \end{pmatrix}$$

(x_*, λ_*) ist stationäre Lösung von (3.3) $\rightsquigarrow g(x_*) = 0$, d.h. $x_* \in U$.

Sei (x_*, λ_*) lokale Lösung von (3.3):

$$\forall x \in \mathcal{U}(x_*) \cap U : \varphi(x_*) = \psi(x_*, \lambda_*) \leq \psi(x, \lambda_*) = \varphi(x)$$

Strafterme

(3) Einführung von Straftermen: $U = \{x \in \mathbb{R}^m \mid g(x) = 0, h(x) \leq 0\}$

$$\boxed{\varphi_\varepsilon(x, t) := \varphi(x) + \frac{1}{\varepsilon}\left\{|g(x)|_2^2 + |h(x) - t|_2^2\right\}, \quad x \in \mathbb{R}^m, t \in \mathbb{R}}$$

Dann ist $\varphi_\varepsilon(x, t) \geq \varphi(x)$ und $\varphi_\varepsilon(x, t) = \varphi(x)$ für $x \in U$, $t = h(x)$.

Wir betrachten dann

$$\boxed{\min\{\varphi_\varepsilon(x, t) \mid x \in \mathbb{R}^m, t \leq 0\}}$$

4. Interpolation

Ziel der Interpolation ist die Darstellung von Funktionen zur Vereinfachung oder zur Auswertung von Tabellen.

4.1. Interpolationspolynome von Lagrange und Newton.

Gegeben seien paarweise voneinander verschiedene Stützstellen $t_1, \ldots, t_n \in \mathbb{R}$. $f(t_1), \ldots, f(t_n) \in \mathbb{R}$ seien die Funktionswerte einer Funktion f.

Gesucht ist nun das Polynom P mit $\mathrm{grad} P \leq n-1$, welches den Interpolationsbedingungen

$$P(t_i) = f(t_i), \quad i = 1, \ldots, n,$$

genügt. Wir nennen P das Interpolationspolynom zu f und den Stützstellen $t_1, \ldots, t_n$.

$$\begin{aligned} \omega(t) &:= (t-t_1)\cdots(t-t_n) \quad \text{(Stützstellenpolynom)} \\ p_i(t) &:= \prod_{j=1, j\neq i}^{n} \frac{t-t_j}{t_i-t_j} = \frac{\omega(t)}{(t-t_i)\,\omega'(t_i)} \\ \rightsquigarrow\ p_i(t_k) &= \delta_{i,k} = \begin{cases} 1 & k=i \\ 0 & k\neq i \end{cases}, \quad \mathrm{grad} p_i = n-1 \end{aligned}$$

$$\boxed{L(t) := \sum_{i=1}^{n} p_i(t)\, f(t_i)}$$

Interpolationspolynom Lagrange

Dann ist $L(t_k) = f(t_k)$, $k = 1, \ldots, n$ und $\mathrm{grad} L \leq n-1$. L ist das Interpolationspolynom von Lagrange.

Differenzenquotienten (dividierte Differenzen) zu f und $t_1, \ldots, t_n$:

$$f[t_{i_1}, t_{i_2}] := \frac{f(t_{i_1}) - f(t_{i_2})}{t_{i_1} - t_{i_2}} \qquad \text{Differenzenquotient 1. Ordnung}$$

$$f[t_{i_1}, t_{i_2}, t_{i_3}] := \frac{f[t_{i_1}, t_{i_2}] - f[t_{i_2}, t_{i_3}]}{t_{i_1} - t_{i_3}} \qquad \text{Differenzenquotient 2. Ordnung}$$

$$f[t_{i_1}, \ldots, t_{i_k}] := \frac{f[t_{i_1}, \ldots, t_{i_{k-1}}] - f[t_{i_2}, \ldots, t_{i_k}]}{t_{i_1} - t_{i_k}} \qquad \text{Differenzenquotient } k-1 \text{. Ordnung}$$

Eigenschaften:

(1) Bildung der Differenzenquotienten ist eine lineare Operation, denn zu Funktionen f, g und $\alpha, \beta \in \mathbb{R}$, $h := \alpha f + \beta g$ gilt:

$$h[\cdots] := \alpha f[\cdots] + \beta g[\cdots]$$

(2)

$$f[t_{i_1}, t_{i_2}] = \frac{f(t_{i_1})}{t_{i_1} - t_{i_2}} + \frac{f(t_{i_2})}{t_{i_2} - t_{i_1}}$$

Durch Induktion kann gezeigt werden, daß

$$f[t_{i_1}, \ldots, t_{i_k}] = \sum_{l=1}^{k} \frac{f(t_{i_l})}{\prod_{s=1, s\neq l}^{k} (t_{i_l} - t_{i_s})}$$

(3) Die Reihenfolge der Argumente in den Differenzenquotienten ist vertauschbar.

Sei $t \notin \{t_1, \ldots, t_n\}$:

$$\begin{aligned} f(t) &= f(t_1) + (t-t_1) \cdot \underbrace{f[t_1, t]}_{= f[t_1,t_2] + (t-t_2) f[t_1,t_2,t]} \\ &= f(t_1) + (t-t_1) f[t_1,t_2] + (t-t_1)(t-t_2) f[t_1,t_2,t_3] + \ldots + \prod_{i=1}^{n-1} (t-t_i) f[t_1,\ldots,t_n] \\ &\quad + \prod_{i=1}^{n} (t-t_i) f[t_1,\ldots,t_n,t] \\ f(t_k) &= f(t_1) + (t_k-t_1) f[t_1,t_2] + \ldots + (t_k-t_1)\cdots(t_k-t_{k-1}) f[t_1,\ldots,t_k], \quad k = 1,\ldots,n \end{aligned}$$

Interpolationspolynom Newton

$$\boxed{N(t) := f(t_1) + (t-t_1) f[t_1,t_2] + (t-t_1)(t-t_2) f[t_1,t_2,t_3] + \ldots + \prod_{i=1}^{n-1} (t-t_i) f[t_1,\ldots,t_n]}$$

ist ein Polynom mit $\operatorname{grad} N \leq n-1$ und wir bezeichnen es als Newtonsches Interpolationspolynom. Es gilt $f(t_k) = N(t_k)$, $k = 1,\ldots,n$, und $N(t) \equiv L(t)$.

Das Interpolationspolynom von Lagrange hat den Nachteil, daß man bei ungenügender Genauigkeit alle bisherigen Rechenschritte erneut durchführen muß. Beim Interpolationspolynom von Newton ist dies nicht nötig, da bei zusätzlichen Stützstellen nur zusätzliche Additionen ausgeführt werden müssen.

$$\begin{array}{llllll} t_1 & f(t_1) & & & & \\ & & f[t_1,t_2] & & & \\ t_2 & f(t_2) & & f[t_1,t_2,t_3] & & \\ & & f[t_2,t_3] & & \ddots & \\ \vdots & \vdots & & & & f[t_1,t_2,\ldots,t_n] \\ & & & & & \qquad\qquad f[t_1,\ldots,t_{n+1}] \\ \vdots & \vdots & & f[t_{n-2},t_{n-1},t_n] & & f[t_2,t_3,\ldots,t_{n+1}] \\ & & f[t_{n-1},t_n] & & & \\ t_n & f(t_n) & & f[t_{n-1},t_n,t_{n+1}] & & \\ & & f[t_n,t_{n+1}] & & & \\ t_{n+1} & f(t_{n+1}) & & & & \end{array}$$

Restglied der Interpolation:

$$f(t) - N(t) = \begin{cases} 0 & \text{für } t \in \{t_1,\ldots,t_n\} \\ \omega(t) f[t_1,\ldots,t_n,t] & \text{sonst} \end{cases}$$

Lemma 4.1. *Seien $t_1,\ldots,t_k \in [a,b]$ paarweise verschiedene Stützstellen, $f \in C^{k-1}([a,b])$. Dann existiert ein $\theta \in [a,b]$ mit*

$$\boxed{f[t_1,\ldots,t_k] = \frac{1}{(k-1)!} \cdot f^{(k-1)}(\theta)}$$

Bemerkung. Für $k = 2$ ist dies der Mittelwertsatz.

Beweis. $\tilde{N}$ sei das Interpolationspolynom zu den Stüztstellen $t_1,\ldots,t_k$ und f. Sei

$$\tilde{R} := f - \tilde{N}$$

Dann ist $\tilde{R}(t_i) = 0$, $i = 1,\ldots,k$, und $\tilde{R} \in C^{k-1}([a,b])$.

$\tilde{R}$ hat mindestens k voneinander verschiedene Nullstellen, $\tilde{R}'$ hat mindestens $k-1$ voneinander verschiedene Nullstellen (Satz von Rolle), usw. Dann hat $\tilde{R}^{(k-1)}$ mindestens eine Nullstelle θ und

es gilt:

$$\begin{aligned} 0 = \tilde{R}^{(k-1)}(\theta) &= f^{(k-1)}(\theta) - \tilde{N}^{(k-1)}(\theta) \\ &= f^{(k-1)}(\theta) - \left(\prod_{i=1}^{k} (t - t_i) f[t_1, \ldots, t_k] \right)^{(k-1)} \\ &= f^{(k-1)}(\theta) - (k-1)! \cdot f[t_1, \ldots, t_k] \end{aligned}$$

□

Satz 4.2. *Sei $f \in C^n([a,b])$ und seien $t_1, \ldots, t_n \in [a,b]$ paarweise voneinander verschieden. Sei P das zugehörige Interpolationspolynom. Dann gilt* Restgliedabschätzung

$$\boxed{\|f - P\|_\infty \leq (b-a)^n \cdot \frac{1}{n!} \cdot \left\| f^{(n)} \right\|_\infty}$$

Satz 4.3. *Sei $f \in C^\infty([a,b])$, $t_1, \ldots, t_n \in [a,b]$ paarweise voneinander verschieden und P_n sei das Interpolationspolynom zu $t_1, \ldots, t_n$. Weiterhin gelte $\left\| f^{(n)} \right\|_\infty \leq K \cdot M^n$, $n \in \mathbb{N}$. Dann gilt* Konvergenzsatz

$$\boxed{\|f - P_n\|_\infty \xrightarrow{n \to \infty} 0}$$

Bemerkung. Die Klasse der Funktionen, die diesen Voraussetzungen genügt, ist sehr klein.

Beispiel.

- Sei
$$f(t) = \begin{cases} e^{-\frac{1}{t^2}} & t \in (0,1] \\ 0 & t \in [-1,1] \end{cases},$$
$f \in C^\infty$. Wir betrachten die Stützstellen $t_1, \ldots, t_n \in [-1,0]$ mit $f(t_i) = 0$.
Dann gilt: $P_n(t) \equiv 0$, $n \in \mathbb{N}$ und
$$|f(t) - P_n(t)| = f(t) \not\to 0 \ (n \to \infty), \quad t \in (0,1]$$
- Die Runge-Funktion
$$f(t) = \frac{1}{1 + 25t^2}, \quad t \in [-1,1],$$
erfüllt die Voraussetung $\left\| f^{(n)} \right\|_\infty \leq K \cdot M^n$ nicht.

4.2. Interpolierende kubische Splinefunktionen.

Definition. Eine Funktion $S \in C^2([a,b])$ heißt kubische Splinefunktion (oder kubischer Spline) zur Zerlegung $\pi : a = t_0 < t_1 < \ldots < t_n = b$, falls $S|_{(t_{j-1}, t_j)}$, $j = 1, \ldots, n$, Polynome höchstens 3. Grades sind. kubischer Spline

Wie kann man nun bei gegebenen Stützstellen einer Funktion f die Splinefunktion so konstruieren, daß $f(t_i) = S(t_i)$ für $i = 0, \ldots, n$ gilt?

Wir haben $n \cdot 4$ Parameter zur Verfügung (n Teilintervalle mit je 4 Parametern) und $(n-1) \cdot 3 + (n+1) = 4n - 2$ Bedingungen (Stetigkeit der 0., 1. und 2. Ableitung an den $n-1$ inneren Punkten, an denen die Teilintervalle zusammenkommen und $n+1$ Interpolationsbedingungen).

Die Zerlegung ist gegeben durch $\pi : a = t_0 < t_1 < \ldots < t_n = b$ und wir definieren

$$h_j := t_j - t_{j-1}, \quad j = 1, \ldots, n$$

j sei fixiert, $t \in (t_{j-1}, t_j)$:

$$\begin{aligned} S''(t) &:= M_j \cdot \frac{t - t_{j-1}}{h_j} + M_{j-1} \cdot \frac{t_j - t}{h_j} \\ S'(t) &= \frac{1}{2} M_j \cdot \frac{(t - t_{j-1})^2}{h_j} - \frac{1}{2} M_{j-1} \cdot \frac{(t_j - t)^2}{h_j} + c_j \\ S(t) &= \frac{1}{6} M_j \cdot \frac{(t - t_{j-1})^3}{h_j} + \frac{1}{6} M_{j-1} \cdot \frac{(t_j - t)^3}{h_j} + c_j (t - t_{j-1}) + d_j \end{aligned}$$

Bestimmen c_j, d_j aus $S(t_i) = f(t_i)$, $i = j-1, j$:

$$\begin{aligned} f(t_{j-1}) = S(t_{j-1}) &= \frac{1}{6} M_{j-1} h_j^2 + d_j \\ f(t_j) = S(t_j) &= \frac{1}{6} M_j h_j^2 + c_j h_j + d_j \\ \rightsquigarrow d_j &= f(t_{j-1}) - \frac{1}{6} h_j^2 \cdot M_{j-1} \\ c_j &= \frac{1}{h_j} \left\{ f(t_j) - f(t_{j-1}) + \frac{1}{6} h_j^2 \cdot M_{j-1} - \frac{1}{6} M_j h_j^2 \right\} \\ &= f[t_{j-1}, t_j] - \frac{1}{6} h_j \cdot (M_j - M_{j-1}) \end{aligned}$$

Stetigkeitsbedingungen für S' auf $[a, b]$: $S'(t_i^-) = S'(t_i^+)$, $i = 1, \ldots, n-1$

$$\begin{aligned} S'(t_j^-) &= S'(t_j^+) \\ \frac{1}{2} M_j \cdot h_j + c_j &= -\frac{1}{2} M_j \cdot h_{j+1} + c_{j+1} \\ \frac{1}{2} M_j \cdot h_j + f[t_{j-1}, t_j] - \frac{1}{6} h_j (M_j - M_{j-1}) &= -\frac{1}{2} M_j \cdot h_{j+1} + f[t_j, t_{j+1}] - \frac{1}{6} h_{j+1} \cdot (M_{j+1} - M_j) \\ \frac{1}{6} h_j \cdot M_{j-1} + \left(\frac{1}{3} h_j + \frac{1}{3} h_{j+1} \right) \cdot M_j + \frac{1}{6} h_{j+1} \cdot M_{j+1} &= f[t_j, t_{j+1}] - f[t_{j-1}, t_j], \quad j = 1, \ldots, n-1 \\ \lambda_j &:= \frac{h_j}{h_j + h_{j+1}} \\ \mu_j &:= \frac{h_{j+1}}{h_j + h_{j+1}} \\ \rightsquigarrow \mu_j + \lambda_j &= 1 \\ h_j + h_{j+1} &= t_{j+1} - t_{j-1} \\ \rightsquigarrow \frac{1}{6} \lambda_j \cdot M_{j-1} + \frac{1}{3} \cdot M_j + \frac{1}{6} \mu_j \cdot M_{j+1} &= f[t_{j-1}, t_j, t_{j+1}], \quad j = 1, \ldots, n-1 \end{aligned}$$

M_0, M_n seien gegeben. Dann sind die übrigen Parameter des kubischen Splines durch folgendes Gleichungssystem bestimmt:

$$\underbrace{\begin{pmatrix} \frac{1}{3} & \frac{1}{6}\mu_1 & 0 & \cdots & 0 \\ \frac{1}{6}\lambda_2 & \frac{1}{3} & \frac{1}{6}\mu_2 & \ddots & \vdots \\ 0 & \frac{1}{6}\lambda_3 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \frac{1}{3} & \frac{1}{6}\mu_{n-2} \\ 0 & \cdots & 0 & \frac{1}{6}\lambda_{n-1} & \frac{1}{3} \end{pmatrix}}_{A_{n-1}} \cdot \begin{pmatrix} M_1 \\ \vdots \\ \vdots \\ \vdots \\ M_{n-1} \end{pmatrix} = \begin{pmatrix} f[t_0, t_1, t_2] \\ \vdots \\ \vdots \\ \vdots \\ f[t_{n-2}, t_{n-1} t_n] \end{pmatrix} - \frac{1}{6} \begin{pmatrix} \lambda_1 M_0 \\ 0 \\ \vdots \\ 0 \\ \mu_{n-1} M_n \end{pmatrix}$$

Lemma 4.4. *A_{n-1} ist regulär und $\left\| A_{n-1}^{-1} \right\| \leq 6$.*

Bemerkung. D.h. die Norm der Inversen ist beschränkt für alle Zerlegungen. Dabei ist nicht die Beschränkung durch genau 6 wichtig, sondern nur die Existenz einer oberen Schranke.

Beweis. $A_{n-1} = L + D + R$, $D = \frac{1}{3}I$, $D^{-1} = 3I$

$$\begin{aligned} A_{n-1} &= D\left(I + \underbrace{D^{-1}(L+R)}_{=:-B}\right) \\ &= D \cdot (I - B) \\ -B &= \frac{1}{2}\begin{pmatrix} 0 & \mu_1 & 0 & \cdots \\ \lambda_2 & 0 & \mu_2 & \ddots \\ 0 & \lambda_3 & \ddots & \ddots \\ \vdots & \ddots & \ddots & 0 \end{pmatrix} \\ \|B\|_\infty &\leq \frac{1}{2} \end{aligned}$$

Dann ist $I - B$ regulär (L. 1.1) und es ist

$$\begin{aligned} \left\|(I-B)^{-1}\right\|_\infty &\leq \frac{1}{1 - \|B\|_\infty} \leq 2 \\ \rightsquigarrow \left\|A_{n-1}^{-1}\right\| &\leq \left\|D^{-1}\right\| \cdot \left\|(I-B)^{-1}\right\| \leq 3 \cdot 2 \end{aligned}$$

□

Satz 4.5. *Seien zur Zerlegung $\pi : a = t_0 < t_1 < \ldots < t_n = b$ die Funktionswerte $f(t_i) \in \mathbb{R}$, $i = 0, \ldots, n$, gegeben. Dann gibt es zu beliebig gegebenen $M_0, M_n \in \mathbb{R}$ genau eine kubische Splinefunktion S mit $S(t_i) = f(t_i)$, $i = 0, \ldots, n$.*

Es gilt dann $S''(a) = M_0$, $S''(b) = M_n$.

Definition. Diese Funktion heißt dann interpolierende kubische Splinefunktion.

Falls $M_0 = M_n = 0$, so spricht man von der natürlichen interpolierenden kubischen Splinefunktion.

Bemerkung. Eine andere Auswahl von M_0, M_n, z.B. bei $f(a) = f(b)$

$$\begin{aligned} S'(a) &= S'(b) \\ S''(a) &= S''(b) \end{aligned}$$

liefert einen periodischen Spline.

Satz 4.6. *Sei $f \in C^2([a,b])$, $\pi : a = t_0 < t_1 < \ldots < t_n = b$, $h = \max\limits_{j=1,\ldots,n} h_j$. Dann gilt für die natürliche interpolierende Splinefunktion S zu π und f:* Restgliedabschätzung

$$\boxed{\begin{aligned} \|S - f\|_\infty &\leq \tfrac{3}{2} \cdot \|f''\|_\infty \cdot h^2 \\ \|S' - f'\|_\infty &\leq 3 \cdot \|f''\|_\infty \cdot h \end{aligned}}$$

Beweis. Sei $f \in C^2([a,b])$ und $t \in (t_{j-1}, t_j)$. Es ist

$$\begin{aligned} S(t) - f(t) &= \frac{1}{6} \cdot M_j \cdot \frac{(t-t_{j-1})^3}{h_j} + \frac{1}{6} \cdot M_{j-1} \cdot \frac{(t_j - t)^3}{h_j} + \underbrace{f[t_{j-1}, t_j](t-t_{j-1}) + f(t_{j-1}) - f(t)}_{(t-t_{j-1})(t-t_j) f[t_{j-1}, t_j, t]} \\ &\quad -\frac{1}{6} h_j \cdot (M_j - M_{j-1})(t-t_{j-1}) - \frac{1}{6} h_j^2 \cdot M_{j-1},\ t \in [t_{j-1}, t_j] \\ &= (t-t_{j-1})(t-t_j) f[t_{j-1}, t_j, t] \\ &\quad + M_{j-1} \cdot \underbrace{\left\{(t_j - t)^2 - h_j^2\right\}}_{\leq h_j^2} \cdot \frac{1}{6} \cdot \frac{t_j - t}{h_j} + M_j \cdot \underbrace{\left\{(t-t_{j-1})^2 - h_j^2\right\}}_{\leq h_j^2} \cdot \frac{1}{6} \cdot \frac{t-t_{j-1}}{h_j} \end{aligned}$$

Für die natürliche interpolierende Spline-Funktion ist $M_0 = 0$, $M_n = 0$, so daß

$$\begin{aligned} \begin{pmatrix} M_1 \\ \vdots \\ M_{n-1} \end{pmatrix} &= A_{n-1}^{-1} \cdot \begin{pmatrix} f[t_0, t_1, t_2] \\ \vdots \\ f[t_{n-2}, t_{n-1}, t_n] \end{pmatrix} \\ \max_{i=1,\ldots,n-1} |M_i| &\leq 6 \cdot \max_{i=1,\ldots,n-1} |f[t_{i-1}, t_i, t_{i+1}]| \\ &\overset{\text{(L. 4.1)}}{\leq} 6 \cdot \frac{1}{2} \cdot \|f''\|_\infty, \quad i = 1, \ldots, n-1 \end{aligned}$$

Dies liefert dann

$$\begin{aligned} |S(t) - f(t)| &\leq h_j^2 \cdot \frac{1}{2} \cdot \|f''\|_\infty + 2 \cdot h_j^2 \cdot \frac{1}{6} \cdot 6 \cdot \frac{1}{2} \cdot \|f''\|_\infty \\ &= h_j^2 \cdot \frac{3}{2} \cdot \|f''\|_\infty \\ &\leq \frac{3}{2} h^2 \cdot \|f''\|_\infty \end{aligned}$$

$$\begin{aligned} S'(t) &= \frac{M_j}{2} \cdot \frac{(t-t_{j-1})^2}{h_j} - \frac{M_{j-1}}{2} \cdot \frac{(t_j - t)^2}{h_j} + f[t_{j-1}, t_j] - \frac{1}{6} h_j (M_j - M_{j-1}) \\ S'(t) - f'(t) &= f[t_{j-1}, t_j] - f'(t) + M_j h_j \left(\frac{1}{2} \frac{(t-t_{j-1})^2}{h_j^2} - \frac{1}{6}\right) + M_{j-1} h_j \left(-\frac{1}{2} \frac{(t_j - t)^2}{h_j} + \frac{1}{6}\right) \\ \max_{i=1,\ldots,n-1} |M_i| &\leq 6 \cdot \max_{i=1,\ldots,n-1} |f[t_{i-1}, t_i, t_{i+1}]| \\ \rightsquigarrow |M_i| &\leq 6 \cdot \frac{1}{2} \|f''\|_\infty, \quad i = 1, \ldots, n-1 \\ |f[t_{j-1}, t_j] - f'(t)| &\overset{\text{L. 4.1}}{=} |f'(\theta_j) - f'(t)|, \quad \theta_j \in (t_{j-1}, t_j) \\ &\leq \|f''\|_\infty \cdot |\theta_j - t| \\ &\leq \|f''\|_\infty \cdot h_j \\ \rightsquigarrow |S'(t) - f'(t)| &\leq \|f''\|_\infty \cdot h + \frac{6}{2} \cdot \|f''\|_\infty h \cdot \left(\frac{1}{2} - \frac{1}{6}\right) + \frac{6}{2} \cdot \|f''\|_\infty h \cdot \left|-\frac{1}{2} + \frac{1}{6}\right| \\ &\leq \|f''\|_\infty \cdot h \cdot \left(1 + 3 \cdot \frac{1}{3} + 3 \cdot \frac{1}{3}\right) \\ &= 3 \cdot \|f''\|_\infty \cdot h \end{aligned}$$

□

Bemerkung. Für $f \in C^1([a,b])$ ist

$$\begin{aligned}
\max_{i=1,\ldots,n-1} |f[t_{i-1},t_i,t_{i+1}]| &= \max_{i=1,\ldots,n-1} \left| \frac{f[t_{i-1},t_i] - f[t_i,t_{i+1}]}{t_{i+1}-t_{i-1}} \right| \\
&\leq \max_{i=1,\ldots,n-1} \frac{1}{h_i + h_{i-1}} \left(|f[t_{i-1},t_i]| + |f[t_i,t_{i+1}]| \right) \\
&\leq \frac{1}{2 \cdot h_{\min}} \cdot 2 \cdot \|f'\|_\infty \\
\rightsquigarrow \|S-f\|_\infty &\leq \|f'\|_\infty \cdot \text{Konstante} \cdot \frac{h^2}{h_{\min}}
\end{aligned}$$

Satz 4.7. *Für $f \in C([a,b])$ konvergieren die natürlichen interpolierenden kubischen Splinefunktionen zu äquidistanten Zerlegungen gleichmäßig gegen f, wenn die maximale Schrittweite h gegen 0 konvergiert.* Konvergenzsatz

Extremaleigenschaft: $\pi : a = t_0 < t_1 < \ldots < t_n = b$, $f(t_i)$, $i = 0,\ldots,n$, seien fixiert

$$\mathcal{F} := \left\{ g \in C^2([a,b]) \,\middle|\, g(t_i) = f(t_i),\ i = 0,\ldots,n \right\}$$

Die interpolierenden kubischen Splinefunktionen zu π und f gehören - ebenso wie die Interpolationspolynome - zu $\mathcal{F}$.

Satz 4.8. *Ist S die natürliche interpolierende Splinefunktion, so gilt* Extremaleigenschaft

$$\boxed{\int_a^b (g''(t))^2\,dt > \int_a^b (S''(t))^2\,dt, \quad g \in \mathcal{F}, g \neq S}$$

Beweis. Sei $g \in \mathcal{F}$.

$$\int_a^b (g''(t))^2\,dt - \int_a^b (S''(t))^2\,dt = \underbrace{\int_a^b (g''(t) - S''(t))^2\,dt}_{\geq 0} + \underbrace{2\int_a^b S''(t)\,g''(t)\,dt - 2\int_a^b (S''(t))^2\,dt}_{=:I}$$

$$\begin{aligned}
\frac{1}{2}I &= \int_a^b S''(t)\,(g''(t) - S''(t))\,dt = \sum_{j=1}^n \int_{t_{j-1}}^{t_j} S''(t)\,(g''(t) - S''(t))\,dt \\
&= \sum_{j=1}^n \left(S''(t)\,(g'(t) - S'(t))\big|_{t_{j-1}}^{t_j} - \int_{t_{j-1}}^{t_j} S'''(t)\,(g'(t) - S'(t))\,dt \right) \\
&= \underbrace{S''(t_n)}_{=0}\,(g'(t_n) - S'(t_n)) - \underbrace{S''(t_0)}_{=0}\,(g'(t_0) - S'(t_0)) - \sum_{j=1}^n \int_{t_{j-1}}^{t_j} \frac{1}{h_j}\,(M_j - M_{j-1})\,(g'(t) - S'(t))\,dt \\
&= -\sum_{j=1}^n \frac{1}{h_j}\,(M_j - M_{j-1}) \cdot \underbrace{(g(t) - S(t))\big|_{t_{j-1}}^{t_j}}_{0 \text{ wg. Interpolationseigenschaft } g \in \mathcal{F}} = 0
\end{aligned}$$

Für ein $g \in \mathcal{F}$ gelte

$$\begin{aligned}
\int_a^b (g''(t) - S''(t))^2\,dt &= 0 \\
\rightsquigarrow g''(t) &\equiv S''(t) \\
g(t) &\equiv S(t) + \alpha(t - t_0) + \beta \\
f(t_0) = g(t_0) &= S(t_0) + \beta = f(t_0) + \beta \quad \rightsquigarrow \beta = 0 \\
f(t_n) = g(t_n) &= S(t_n) + \alpha(t_n - t_0) = f(t_n) + \alpha(t_n - t_0) \quad \rightsquigarrow \alpha = 0 \\
\rightsquigarrow g &\equiv S
\end{aligned}$$

□

5. Numerische Berechnung von Integralen

Sei $f \in C([a,b])$. $\int_a^b f(t)\,dt = ?$

5.1. Interpolationsformeln.

5.1.1. *Gaußsche Quadraturformeln.*

Unter Benutzung des Lagrangschen Interpolationspolynoms können wir die Gaußsche Quadraturformel (Gauß-Lagrange-Formel) definieren:

$$\int_a^b f(t)\,dt = \int_a^b \sum_{j=1}^{n} p_j(t) \cdot f(t_j)\,dt + \underbrace{\int_a^b \omega(t) \cdot f[t_1,\ldots,t_n,t]}_{=:R(f)}$$

mit $t_1,\ldots,t_n \in [1,b]$ paarweise verschieden, $\omega(t) = \prod_{j=1}^{n}(t-t_j)$ und $p_j(t)$ Lagrange-Polynome (siehe §4.1).

Gaußsche Quadraturformel

(5.1)
$$\boxed{\begin{aligned} A_j &:= \int_a^b p_j(t)\,dt, \quad j = 1,\ldots,n \\ \int_a^b f(t)\,dt &\approx \sum_{j=1}^{n} A_j f(t_j) \end{aligned}}$$

$$\begin{aligned} R(f) &:= \int_a^b f(t)\,dt - \sum_{j=1}^{n} A_j f(t_j) \\ &= \int_a^b \omega(t) \cdot f[t_1,\ldots,t_n,t]\,dt \end{aligned}$$

Definition. Die Formel (5.1) ist genau $\Leftrightarrow$ statt „$\approx$" steht „$=$", d.h. $R(f) = 0$.

algebraischer Genauigkeitsgrad

Definition. Der algebraische Genauigkeitsgrad μ einer Interpolationsformel ist der maximale Grad, den ein Polynom haben kann, damit die Formel genau ist.

Für die Gaußsche Quadraturformel (5.1) gilt nach Konstruktion $\mu \geq n-1$. Der maximale Genauigkeitsgrad beträgt $2n-1$.

Beispiel. Wir betrachten als Beispiel den Fall $n = 1$: Rechteckregel:

$$\begin{aligned} \int_a^b f(t)\,dt &\approx A_1 \cdot f(t_1) \\ A_1 &= \int_a^b dt = b-a \end{aligned}$$

Rechteckregel

$$\boxed{\int_a^b f(t)\,dt \approx (b-a) \cdot f(t)}$$

Es ist $\mu \geq 0$. Ansatz: $f(t) = t - t_1$, $f[t,t_1] \equiv 1$.

$$\begin{aligned} R(f) &= \int_a^b (t-t_1) \cdot f[t_1,t]\,dt \\ &= \int_a^b (t-t_1)\,dt \\ &= \begin{cases} 0 & \text{falls } t_1 = \frac{a+b}{2} \\ \neq 0 & \text{sonst} \end{cases} \end{aligned}$$

Falls $t_1 = \frac{a+b}{2}$, so gilt $\mu = 1$, sonst $\mu = 0$.

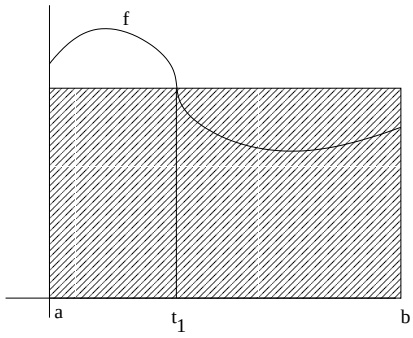


ABBILDUNG 5.1. Interpoliertes Integral im Fall $n = 1$ (Rechteckregel)

Alle Interpolationsformeln sind genau für Polynome 0. Grades und es gilt dann:

$$\int_a^b dt = \sum_{j=1}^{n} A_j = b - a$$

Für die Gaußsche Quadraturformel gilt $\mu \leq 2n-1$, denn für $f(t) = (\omega(t))^2$, $\omega(t) = \prod_{i=1}^{n}(t - t_i)$, ist $\text{grad} f = 2n$ und

$$\int_a^b f(t)\,dt > 0, \text{ aber } \sum_{j=1}^{n} A_j \underbrace{f(t_j)}_{=0} = 0$$

Satz 5.1. *Die Interpolationsformel mit n paarweise voneinander verschiedenen Knoten* $t_1, \ldots, t_n \in [a,b]$ *hat den algebraischen Genauigkeitsgrad* $\mu = 2n-1$ *genau dann, wenn*

$$\int_a^b \omega(t)\,p(t)\,dt = 0$$

für alle Polynome p mit $\text{grad}(p) \leq n-1$ *gilt.*

Bemerkung. Polynome ω, p mit $\langle \omega, p \rangle_{L_2(a,b)} := \int_a^b \omega(t)\,p(t)\,dt = 0$ heißen orthogonale Polynome.

Beweis.

(⇒): Sei $\mu = 2n-1$, $\text{grad}(p) \leq n-1$. Wir bilden $f := \omega \cdot p$, $\text{grad}(f) \leq 2n-1$. Dann ist

$$\int_a^b \omega(t)\,p(t)\,dt = \sum_{j=1}^{n} A_j f(t_j) = 0, \qquad \text{da } f(t_j) = \underbrace{\omega(t_j)}_{=0} \cdot p(t_j) = 0$$

(⇐): Sei f ein Polynom mit $n \leq \text{grad}(f) \leq 2n-1$. Dann existieren Polynome p und r, so daß $f = \omega p + r$ und $\text{grad}(p), \text{grad}(r) \leq n-1$. Dann ist wegen $\omega(t_j) = 0$:

$$f(t_j) = r(t_j), \quad j = 1, \ldots, n.$$

$$\int_a^b f(t)\,dt = \underbrace{\int_a^b \omega(t)\,p(t)\,dt}_{=0} + \int_a^b r(t)\,dt = \sum_{j=1}^{n} A_j r(t_j) = \sum_{j=1}^{n} A_j f(t_j)$$

□

Lemma 5.2. *Es gelte für ein Polynom* $\tilde{\omega}$, $\text{grad}\tilde{\omega} = n$, *die Beziehung*

$$\int_a^b \tilde{\omega}(t)\,p(t)\,dt = 0$$

für alle Polynome p mit $\text{grad}(p) \leq n-1$.

Dann hat $\tilde{\omega}$ *in* (a,b) *genau n paarweise voneinander verschiedene Nullstellen.*

Beweis. G. Opfer: Numerische Mathematik für Anfänger, Vieweg 1993, S. 103 □

Definition. Knoten zum maximalen Genauigkeitsgrad $\mu = 2n-1$ heißen Gaußsche Knoten.

5.1.2. *Newton-Cotes-Formeln.*

Newton-Cotes-Formeln

Bei den Newton-Cotes-Formeln handelt es sich um Quadraturformeln auf äquidistanten Unterteilungen des Intervalls [a,b]. D.h. wir betrachten (5.1) mit $t_1 = a$, $t_2 = a+h$,...,$t_n = a+(n-1)\,h = b$.

$$\rightsquigarrow\ h = \frac{b-a}{n-1}, \quad n \geq 2$$

Beispiel. Fall $n = 2$: Trapezregel

$$\begin{aligned} \int_a^b f(t)\,dt &\approx A_1 f(a) + A_2 f(b) \\ A_1 &= \int_a^b \frac{t-b}{a-b} dt = \frac{1}{2}(b-a) \\ A_2 &= \int_a^b \frac{t-a}{b-a} dt = \frac{1}{2}(b-a) \end{aligned}$$

Trapezregel

$$\boxed{\int_a^b f(t)\,dt \approx \frac{b-a}{2}\left(f(a)+f(b)\right)}$$

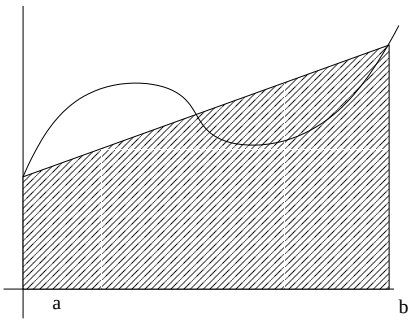


ABBILDUNG 5.2. Newton-Cotes-Formel für $n = 2$: Trapezregel

Es ist $\mu \geq 1$. Für $f(t) = (t-a)(b-t)$ ist $\int_a^b f(t)\,dt > 0$, aber $\frac{b-a}{2}(f(a)+f(b)) = 0$. $\rightsquigarrow \mu = 1$.

Die Newton-Cotes-Formeln haben einen geringeren Genauigkeitsgrad, sind aber andererseits praktischer zu handhaben.

Lemma 5.3. *Sei $f \in C^2([a,b])$. Dann gibt es ein $\theta \in [a,b]$ mit*

$$R(f) := \int_a^b f(t)\,dt - \frac{1}{2}(b-a)\left(f(a)+f(b)\right) = -\frac{1}{12}(b-a)^3 f''(\theta)$$

Beweis.

$$-R(f) = -\int_a^b (t-a)(t-b)\,f[a,b,t]\,dt \overset{\text{(L. 4.1)}}{=} \int_a^b \underbrace{(t-a)(t-b)}_{=:g(t)} \cdot \frac{1}{2} f''(\theta_t)\,dt$$

Es ist $g \geq 0$ und $\int_a^b g = \frac{1}{6}(b-a)^3 > 0$. Nach dem ersten Mittelwertsatz existiert dann ein $\theta \in [a,b]$, so daß

$$-R(f) = \int_a^b g(t)\,dt \cdot \frac{1}{2} f''(\theta) = \frac{(b-a)^3}{12} \cdot f''(\theta)$$

□

5.2. **Zusammengesetzte Formeln.**

Bisher hatten wir Quadraturformeln kennengelernt, welche durch Integration eines Interpolationspolynoms entstanden sind. Bei zusammengesetzten Quadraturformeln betrachten wir dagegen Funktionen, welche stückweise aus Polynomen zusammengesetzt sind (z.B. kubische Splines).

Zusammengesetzte (große / iterierte) Trapezregel: $a = t_0 < \ldots < t_n = b, t_j = a + jh, h = \frac{1}{n}(b-a)$ — Trapezregel, zusammengesetzt

$$\boxed{\int_a^b f(t)\,dt \approx S_T(h) := \sum_{j=1}^{n} \frac{h}{2}\left(f(t_{j-1}) + f(t_j)\right) = \frac{b-a}{2n}\left(f(a) + 2\sum_{j=1}^{n-1} f(a+jh) + f(b)\right)}$$

Satz 5.4. *Sei $f \in C^2([a,b])$. Dann ist*

$$\int_a^b f(t)\,dt - S_T(h) = -\frac{b-a}{12} f''(\eta(h))\,h^2$$

Beweis.

$$\begin{aligned}\int_a^b f(t)\,dt - S_T(h) &= \sum_{j=1}^{n}\left(\int_{t_{j-1}}^{t_j} f(t)dt - \frac{h}{2}\left(f(t_{j-1}) + f(t_j)\right)\right)\\ &\overset{\text{(L. 5.3)}}{=} \sum_{j=1}^{n}\left\{-\frac{1}{12}h^3 f''(\eta_j)\right\}\\ &= -\frac{1}{12}h^2\frac{b-a}{n}\sum_{j=1}^{n} f''(\eta_j)\end{aligned}$$

Da $f \in C^2$, folgt mit dem Satz von Weierstraß: $K_{\min} \le f''(t) \le K_{\max}$, $t \in [a,b]$, und

$$K_{\min} = \frac{1}{n}\sum_{j=1}^{n} K_{\min} \le \underbrace{\frac{1}{n}\sum_{j=1}^{n} f''(\eta_j)}_{=f''(\eta)} \le \frac{1}{n}\sum_{j=1}^{n} K_{\max} = K_{\max}, \quad \eta_j = \eta_j(h), \eta = \eta(h).$$

und damit folgt:

$$\int_a^b f(t)\,dt - S_T(h) = -\frac{b-a}{12} f''(\eta(h)) \cdot h^2$$

□

Folgerung. Sei $f \in C^2([a,b])$. Dann ist — Restgliedabschätzung

$$\left|\int_a^b f(t)\,dt - S_T(h)\right| \le \frac{b-a}{12} \cdot \|f''\|_\infty \cdot h^2$$

Satz 5.5. *(Asymptotische Entwicklung in Potenzen von h). Sei $f \in C^{2m+2}([a,b])$. Dann ist*

$$S_T(h) = \int_a^b f(t)\,dt + a_2 h^2 + \ldots + a_{2m}h^{2m} + a_{2m+2}(h)\,h^{2m+2}, \quad h = \frac{b-a}{n}, n \in \mathbb{N}$$

mit Koeffizienten

$$\begin{aligned}a_{2k} &:= \frac{B_{2k}}{(2k)!} \cdot \left(f^{(2k-1)}(b) - f^{(2k-1)}(a)\right), \quad k = 1, \ldots, m\\ a_{2m+2}(h) &:= \frac{B_{2m+2}}{(2m+2)!} \cdot (b-a) \cdot f^{(2m+2)}(\eta(h))\end{aligned}$$

(B_i sind Bernoulli-Zahlen).

Bemerkung. Satz 5.4 ist ein Spezialfall von Satz 5.5.

5.3. Extrapolation und Romberg-Integration.

Gegeben sei eine Funktion

$$S(h) := a_0 + a_2 h^2 + \ldots + a_{2m} h^{2m} + a_{2m+2}(h)\, h^{2m+2}, \quad h \in D, D \subseteq \mathbb{R}$$

mit Konstanten a_{2i}, $i = 0, \ldots, m$, und $a_{2m+2}(\cdot)$ sei beschränkt.

Dabei sei D derart, daß für eine Konstante $q \in (0,1)$ aus $h \in D$ sofort $qh \in D$ folgt.

Beispiel. $S_T(h)$, $D := \left\{ \frac{b-a}{n} \,\middle|\, n \in \mathbb{N} \right\}$, $q = \frac{1}{2}$

$S(h)$ repräsentiert das Ergebnis eines numerischen Verfahrens für a_0 mit Schrittweite h. Wenn $a_2 \neq 0$, so ist die (Approximations-)Ordnung 2 $\left(h^2\right)$.

Richardson-Extrapolation:

$$\begin{aligned}
S(qh) &= a_0 + a_2 q^2 h^2 + \ldots + a_{2m} q^{2m} h^{2m} + a_{2m+2}(qh)\, q^{2m+2} h^{2m+2} \\
q^2 S(h) &= q^2 a_0 + a_2 q^2 h^2 + \ldots \\
\underbrace{\frac{S(qh) - q^2 S(h)}{1-q^2}}_{S^{[1]}(h)} &= a_0 + \underbrace{a_4 \frac{q^4 - q^2}{1-q^2}}_{a_4^{[1]}} h^4 + \ldots + \underbrace{a_{2m} \frac{q^{2m} - q^2}{1-q^2}}_{a_{2m}^{[1]}} h^{2m} + \underbrace{\frac{a_{2m+2}(qh)\, q^{2m+2} - a_{2m+2}(h)\, q^2}{1-q^2}}_{a_{2m+2}^{[1]}(h)} h^{2m+2}
\end{aligned}$$

$$S^{[2]}(h) := \frac{S^{[1]}(qh) - q^4 S^{[1]}(h)}{1-q^4} = a_0 + a_6^{[2]} h^6 + \ldots \quad \text{hat Ordnung } O\left(h^6\right), \text{ falls } a_6^{[2]} \neq 0$$

Extrapolation

Sind die Ergebnisse $S(h_i)$ mit Schrittweiten $h_1, h_2 = qh_1, \ldots, h_{m+1} = qh_m$, $i = 1, \ldots, m+1$, bekannt, so läßt sich ein Ergebniss der Ordnung $O\left(h^{2m+2}\right)$ mittels folgender Formel extrapolieren:

$$\boxed{S^{[m]}(h) := \frac{S^{[m-1]}(qh) - q^{2m} S^{[m-1]}(h)}{1-q^{2m}}}$$

h_1	$S(h_1)$				
		$S^{[1]}(h_1)$			
h_2	$S(h_2)$		$S^{[2]}(h_1)$		
		$S^{[1]}(h_2)$		$\ddots$	
$\vdots$	$\vdots$				$S^{[m]}(h_1)$
				$\cdot$	
$\vdots$	$\vdots$		$S^{[2]}(h_{m-1})$		
		$S^{[1]}(h_m)$			
h_{m+1}	$S(h_{m+1})$				
Ordnung:	h^2	h^4	h^6		h^{2m+2}

Die Romberg-Integration (1955) ist die Verbindung der zusammengesetzten Trapezregel mit der Extrapolation $q = \frac{1}{2}$.

Beispiel.

(1) $\int_0^1 \frac{4}{1+t^2} dt = \pi = 3.141592\ldots$

$h_1 = \frac{1}{2}$ $\quad S_T(h_1) = 3.1$

$\quad \rightsquigarrow S_T^{[1]}(h_1) = 3.141568$

$h_2 = \frac{1}{4}$ $\quad S_T(h_2) = 3.131$

Erst $S_T(h)$ mit $h < 0.01$ würde eine zu $S_T^{[1]}$ vergleichbare Genauigkeit liefern.

(2) Sei $f \in C^1(I)$, $t \in I$, fest. Aufgabe ist es, $f'(t)$ zu berechnen.
$S_1(h) := \frac{f(t+h) - f(t)}{h}$, $S_2(h) := \frac{f(t+h) - f(t-h)}{2h}$

6. Numerische Lösung gewöhnlicher Anfangswertaufgaben (AWA)

Gegeben sei $f : \mathbb{R}^m \times I \to \mathbb{R}^m$, $I \subseteq \mathbb{R}$ ein Intervall, und f sei stetig und besitze eine stetige partielle Ableitung f_x. Sei $t_0 \in I$, $x_0 \in \mathbb{R}^m$. Gesucht ist eine Lösung von

$$\boxed{\begin{array}{rcl} x'(t) & = & f(x(t), t) \\ x(t_0) & = & x_0 \end{array}}$$

Speziell:

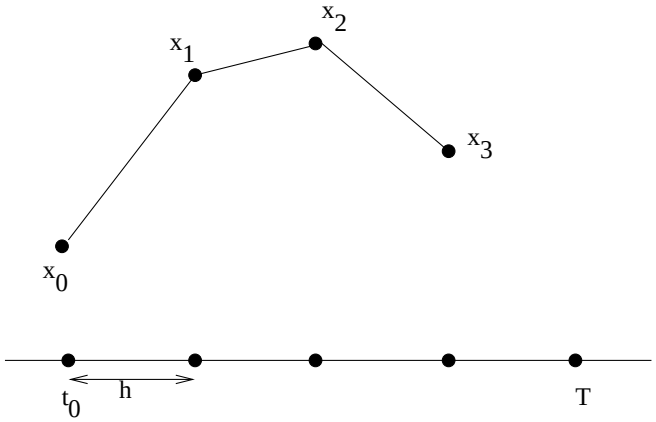
$$\begin{array}{rcl} f(x,t) & := & Bx, \quad x \in \mathbb{R}^m, \quad t \in (-\infty, \infty), t_0 = 0 \\ x'(t) & = & Bx(t) \\ x(t) & = & e^{tB} x_0 \quad \text{Lösung der AWA} \end{array}$$

Für $m = 1$ ergibt dies die skalare DGL: $x'(t) = bx(t)$ mit der Lösung $x(t) = e^{tb} x_0$.

Herkunft:

- mechanische Bewegung (Roboter, ...)
- elektrische Schaltungen
- Reaktionskinetik
- Populationsdynamik
- Wirtschaft:
  - 1972 „Grenzen des Wachstum"
  - Daten ab 1900
  - Prognosen bis 2100

Polygonzugverfahren von Euler ($m = 1$) (1740)



$$\begin{array}{rcl} t_j & := & t_0 + jh \\ x_1 & := & x_0 + hf(x_0, t_0) \\ \rightsquigarrow p(t) & := & x_0 + (t - t_0) f(x_0, t_0), \quad t \in [t_0, t_1] \\ x_i & := & x_{i-1} + hf(x_{i-1}, t_{i-1}) \\ \rightsquigarrow p(t) & := & x_{i-1} + (t - t_{i-1}) f(x_{i-1}, t_{i-1}), \quad t \in [t_{i-1}, t_i] \end{array}$$

Cauchy 1820:

$$\|p(\cdot) - x(\cdot)\|_\infty \xrightarrow{h \to 0} 0$$

Peano 1890:

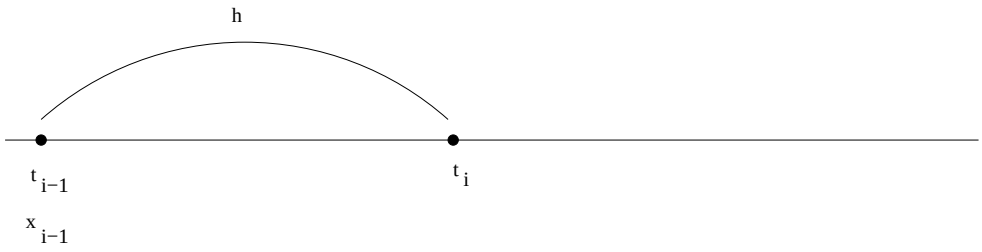
$$\max_{i=1,\ldots,n} |x_i - x(t_i)| \to 0, \quad n = \frac{T - t_0}{h}$$

Schreibweise:

(1) statt $t_i^{(h)}, x_i^{(h)}, n^{(h)}$ schreiben wir t_i, x_i, n.

(2) Zur Funktion $x(\cdot)$ bezeichnen wir mit $x(t_i) \in \mathbb{R}^m$ die Funktionswert und mit $x_i \in \mathbb{R}^m$ die Approximation zu $x(t_i)$.

Berechnung von x_i (mit Einschrittverfahren):

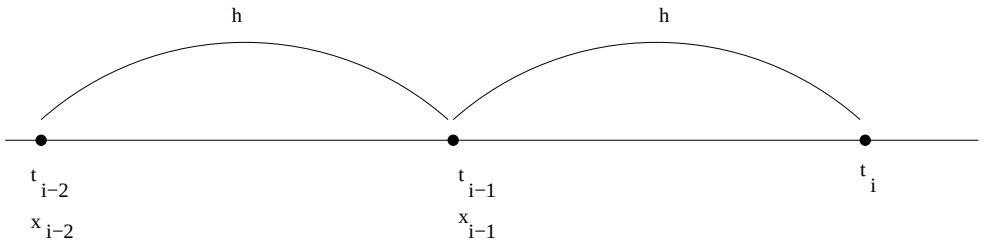


$$\begin{aligned} q(t) &:= x_{i-1} \cdot \frac{t_i - t}{h} + x_i \cdot \frac{t - t_{i-1}}{h} \\ q'(t) &= \frac{1}{h}(x_i - x_{i-1}) \end{aligned}$$

Bedingung:

$$\begin{aligned} q'(t_{i-1}) &= f(q(t_{i-1}), t_{i-1}) \\ \frac{1}{h}(x_i - x_{i-1}) &= f(x_{i-1}, t_{i-1}) \\ x_i &= x_{i+1} + hf(x_{i-1}, t_{i-1}) \end{aligned}$$

Verwendung von zwei bereits bekannten Werten x_{i-2}, x_{i-1} zur Berechnung von x_i:



Ansatz mittels Lagrange-Interpolationspolynom:

$$\begin{aligned} q(t) &= x_i \cdot \frac{t - t_{i-1}}{t_i - t_{i-1}} \cdot \frac{t - t_{i-2}}{t_i - t_{i-2}} + x_{i-1} \cdot \frac{t - t_i}{t_{i-1} - t_i} \cdot \frac{t - t_{i-2}}{t_{i-1} - t_{i-2}} + x_{i-2} \cdot \frac{t - t_i}{t_{i-2} - t_i} \cdot \frac{t - t_{i-1}}{t_{i-2} - t_{i-1}} \\ q'(t) &= \dots \end{aligned}$$

(1) Die Bedingung $q'(t_{i-2}) = f(q(t_{i-2}), t_{i-2})$ liefert

$$\frac{1}{h}\left(-\frac{1}{2}x_i + 2x_{i-1} - \frac{3}{2}x_{i-2}\right) = f(x_{i-2}, t_{i-2}) \tag{6.1}$$

explizite Bedingung für x_i

(2) Die Bedingung $q'(t_i) = f(q(t_i), t_i)$ liefert

$$\frac{1}{h}\left(\frac{3}{2}x_i - 2x_{i-1} + \frac{1}{2}x_{i-2}\right) = f(x_i, t_i) \tag{6.2}$$

implizite Bedingung für x_i

Beispiel. $m = 1$, $x'(t) = -3x(t)$, $t_0 = 0$, $x_0 = 1 \rightsquigarrow x(t) = e^{-3t}$, $t \in [0,1]$

$x_0 = 1$, $x_1 := e^{-3h}$, $n = \frac{1}{h}$:

n:	10	10^2	10^3	10^4
$\lvert x(1) - x_n \rvert$ mittels (6.1)	$3 \cdot 10^2$	$6 \cdot 10^{42}$		
$\lvert x(1) - x_n \rvert$ mittels (6.2)	$5 \cdot 10^{-3}$	$5 \cdot 10^{-5}$	$5 \cdot 10^{-7}$	10^{-8}

Formal sind beide Verfahren gleichwertig, in der Praxis ist (6.1) aber fehleranfälliger.

6.1. Ansatz linearer Mehrschrittverfahren und Konsistenz.

Gegeben sei das Anfangswertproblem $x'(t) - f(x(t), t) = 0$ mit $x(t_0) = x_0$.

Die Näherungsformel

Mehrschrittformel

$$\boxed{\frac{1}{h}\sum_{j=0}^{k} \alpha_j x_{l-j} - \sum_{j=0}^{k} \beta_j f(x_{l-j}, t_{l-j}) = 0, \quad l \geq k} \tag{6.3}$$

heißt k-schrittige Formel (Mehrschrittformel) mit Startphase ($l < k$) und Laufphase ($l \geq k$). Dabei ist $\alpha_0 \neq 0$.

Für $k = 1$ spricht man von einem Einschrittverfahren, für $k > 1$ von einem Mehrschrittverfahren (Zweischritt-, Dreischritt-, ...)

Ist $\beta_0 = 0$ so handelt es sich um ein explizites Verfahren, anderenfalls um ein implizites (in diesem Fall ist pro Schritt ein Gleichungssystem zu lösen).

Wir setzen den exakten Lösungswert des AWA in die Näherungsformel ein und betrachten den Defekt („lokaler Fehler"):

lokaler Fehler

$$\boxed{\tau_l := \frac{1}{h}\sum_{j=0}^{k} \alpha_j \cdot x(t_{l-j}) - \sum_{j=0}^{k} \beta_j \cdot f(x(t_{l-j}), t_{l-j}), \quad l \geq k}$$

Da $f(x(t_{l-j}), t_{l-j}) = x'(t_{l-j})$ können wir auch schreiben:

$$\tau_l = \frac{1}{h}\sum_{j=0}^{k} (\alpha_j x(t_{l-j}) - h\beta_j x'(t_{l-j}))$$

Unter der Voraussetzung, daß die exakte Lösung $x(\cdot) \in C^{p+1}$ ist, gilt (Taylorentwicklung um t_l):

$$\begin{aligned} x(t_{l-j}) &= \sum_{i=0}^{p+1} \frac{(-jh)^i}{i!} x^{(i)}(t_l) + o\left(h^{p+1}\right) \\ x'(t_{l-j}) &= \sum_{i=0}^{p} \frac{(-jh)^i}{i!} x^{(i+1)}(t_l) + o\left(h^{p}\right) \end{aligned}$$

Eingesetzt erhalten wir

$$\begin{aligned} \tau_l = {} & \frac{1}{h}\sum_{j=0}^{k} \alpha_j \cdot x(t_l) + h^0 \cdot \sum_{j=0}^{k} (-\alpha_j \cdot j - \beta_j) x'(t_l) + h^1 \cdot \sum_{j=0}^{k} \left(\alpha_j \cdot \frac{(-j)^2}{2} - \beta_j \cdot (-1)\right) x''(t_l) + \ldots + \\ & + h^p \cdot \sum_{j=0}^{k} \left(\alpha_j \frac{(-j)^{p+1}}{p+1} - \beta_j (-j)^p\right) \frac{1}{p!} x^{(p+1)}(t_l) + o(h^p) \end{aligned}$$

Satz 6.1. *Sind für das lineare k-Schrittverfahren (6.3) die Bedingungen*

$$\sum_{j=0}^{k} \alpha_j = 0, \qquad \sum_{j=0}^{k} (\alpha_j j + \beta_j) = 0, \tag{6.4}$$

$$\sum_{j=0}^{k} \left(\alpha_j \frac{j^{i+1}}{i+1} + \beta_j j^i \right) = 0, \quad i = 1, \ldots, p-1, \tag{6.5}$$

erfüllt (d.h., bis auf den letzten Summenterm in der obigen Gleichung sind alle anderen Summenterme 0), so gilt bei ausreichend glatten Lösungen $x(\cdot)$, daß

$$\boxed{\tau_l = c_p x^{(p+1)}(t_l)\, h^p + o(h^p)}$$

mit der Fehlerkonstanten

$$c_p = \frac{(-1)^{p+1}}{p!} \sum_{j=0}^{k} \left(\alpha_j \frac{j^{p+1}}{p+1} + \beta_j j^p \right)$$

Normierung der Verfahren: $\sum_{j=0}^{k} \beta_j = 1$.

Konsistenz

Definition. Ein Verfahren (6.3) mit (6.4) heißt konsistent. Gilt zusätzlich (6.5), so heißt das Verfahren konsistent mit Ordnung p.

Wichtige Verfahren sind:

(1) Explizites Euler-Verfahren

$$\frac{1}{h}(x_l - x_{l-1}) - f(x_{l-1}, t_{l-1}) = 0$$

$\alpha_0 = 1$, $\alpha_1 = -1$, $\beta_0 = 0$ und $\beta_1 = 1 \rightsquigarrow c_p = \frac{1}{2}$, Konsistenzordnung ist $p = 1$

(2) Implizites Eulerverfahren

$$\frac{1}{h}(x_l - x_{l-1}) - f(x_l, t_l) = 0$$

$\alpha_0 = 1$, $\alpha_1 = -1$, $\beta_0 = 1$ und $\beta_1 = 0 \rightsquigarrow c_p = \frac{1}{2}$, Konsistenzordnung ist $p = 1$

(3) Trapezregel

$$\frac{1}{h}(x_l - x_{l-1}) - \frac{1}{2}(f(x_l, t_l) + f(x_{l-1}, t_{l-1})) = 0$$

$\alpha_0 = 1$, $\alpha_1 = -1$ und $\beta_0 = \beta_1 = \frac{1}{2} \rightsquigarrow c_p = \frac{1}{24}$, Konsistenzordnung ist $p = 2$

(4) BDF (Backward Differentation Formula)

$$\frac{1}{h} \sum_{j=0}^{k} \alpha_j x_{l-j} - f(x_l, t_l) = 0$$

$\beta_0 = 1$, $\beta_j = 0$ für $j = 1, \ldots, k$. Die $\alpha_0, \ldots, \alpha_k$ sind so bestimmt, daß (6.4) erfüllt ist und die Konsistenzordnung $p = k$ erreicht wird.

(Das Beispiel (6.2) ist $k = 2$, das implizite Eulerverfahren ist $k = 1$.)

(5) Adams-Bashforth-Verfahren (Verallgemeinerungen des expliziten Euler-Verfahrens)

$$\frac{1}{h}(x_l - x_{l-1}) - \sum_{j=1}^{k} \beta_j f(x_{l-j}, t_{l-j}) = 0$$

$\alpha_0 = 1$, $\alpha_1 = -1$, $\alpha_j = 0$ für $j = 2, \ldots, k$, $\beta_0 = 0$, $\beta_1, \ldots, \beta_k$ sind so gewählt, daß die Konsistenzordnung $p = k$ erreicht wird.

(6) Adams-Moulton-Verfahren

$$\frac{1}{h}(x_l - x_{l-1}) - \sum_{j=0}^{k} \beta_j f(x_{l-j}, t_{l-j}) = 0$$

$\alpha_0 = 1$, $\alpha_1 = -1$, $\alpha_j = 0$ für $j = 2, \ldots, k$, $\beta_0, \ldots, \beta_k$ sind so gewählt, daß die Konsistenzordnung $p = k + 1$ erreicht wird.

Das Verfahren (6.1) hatte die Werte $k = 2$, $\alpha_0 = -\frac{1}{2}$, $\alpha_1 = 2$, $\alpha_2 = -\frac{3}{2}$, $\beta_0 = 0$, $\beta_1 = 0$ und $\beta_2 = 1$. Damit hat das Verfahren die Konsistenzordnung 2. Diese Eigenschaft hat also nichts mit der schlechten Qualität dieses Verfahrens zu tun.

charakteristisches Polynom

Definition. Das Polynom

$$\boxed{\rho(\lambda) := \sum_{j=0}^{k} \alpha_j \lambda^{k-j}}$$

heißt charakteristisches Polynom des Verfahrens (6.3).

Bemerkung 6.2. $\rho(1) = \sum_{j=0}^{k} \alpha_j = 0$ bei konsistenten Verfahren - Die 1 ist die Konsistenznullstelle des charakteristischen Polynoms.

Betrachten wir erneut die Beispielsverfahren (6.1) und (6.2), so sehen wir, daß für (6.1) das charakteristische Polynom $\rho(\lambda) = -\frac{1}{2}\lambda^2 + 2\lambda - \frac{3}{2}$ die Nullstellen $\lambda_1 = 1$ und $\lambda_2 = 3$ besitzt. Für (6.2) haben wir $\rho(\lambda) = \frac{3}{2}\lambda^2 - 2\lambda + \frac{1}{2}$ mit den Nullstellen $\lambda_1 = 1$ und $\lambda_2 = \frac{1}{3}$. Wir sehen, daß die Nullstellen des charakteristischen Polynoms des „guten" Verfahrens (6.2) im Einheitskreis liegen, während sie beim „schlechten" Verfahren (6.1) nicht vollständig im Einheitskreis liegen. Dies führt dazu, daß sich der Fehler aufschaukelt.

6.2. Stabilität und Konvergenz.

Dahlquistsches Wurzelkriterium stabil

Definition. (Dahlquistsches Wurzelkriterium, 1956) Ein lineares k-Schritt-Verfahren heißt stabil ($h \to 0$ stabil), falls die Wurzeln des charakteristischen Polynoms entweder im Innern des komplexen Einheitskreises liegen oder sie liegen auf dem Einheitskreisbogen und sind einfach.

Es ist

$$\max_{l=1,\ldots,n} |x_l - x(t_l)| \leq S \cdot \max_{l=1,\ldots,n} |\tau(l)|$$

Voraussetzung für DGL:

$$|f_x(x,t)| \leq L, \quad x \in \mathbb{R}^m, t \in [t_0, T]$$

Umformen der k-Schritt-Gleichung:

$$x_l = -\sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0} x_{l-j} + h \sum_{j=0}^{k} \frac{\beta_i}{\alpha_0} f(x_{l-j}, t_{l-j})$$

Dies ist nur ein theoretischer Wert, in der Praxis treten Rundungsfehler auf:

$$\tilde{x}_l = -\sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0} \tilde{x}_{l-j} + h \sum_{j=0}^{k} \frac{\beta_j}{\alpha_0} f(\tilde{x}_{l-j}, t_{l-j}) + \delta_l, \quad l \geq k \tag{6.6}$$

δ_l enthält Rundungsfehler und eventuelle Abbruchfehler.

Startwerte $\tilde{x}_0 = x_0, \tilde{x}_1, \ldots, \tilde{x}_{k-1}$

Es ist

$$x(t_l) = -\sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0} x(t_{l-j}) + h \sum_{j=0}^{k} \frac{\beta_j}{\alpha_0} f(x(t_{l-j}), t_{l-j}) + h\tau_l$$

Mit $\varepsilon_i := x(t_i) - \tilde{x}_i$, $i \geq 0$, wird

$$\varepsilon_l = -\sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0}\varepsilon_{l-j} + h\sum_{j=0}^{k} \frac{\beta_j}{\alpha_0} \underbrace{\{f(x(t_{l-j}), t_{l-j}) - f(\tilde{x}_{l-j}, t_{l-j})\}}_{\text{L. 1.2}: \underbrace{\int_0^1 f_x(\ldots)\, ds}_{=: B_{l,j}} \cdot \varepsilon_{l-j}} + h\tau_l - \delta_l, \quad l \geq k$$

Es gilt $\|B_{l,j}\| \leq L$ für $l \geq k$, $j = 0, \ldots, k$.

globaler Fehler

Damit entsteht folgende Rekursion:

$$\boxed{\varepsilon_l = -\sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0}\varepsilon_{l-j} + h\sum_{j=0}^{k} \frac{\beta_j}{\alpha_0} B_{l,j}\varepsilon_{l-j} + h\tau_l - \delta_l} \tag{6.7}$$

Wunsch:

$$|\varepsilon_l| \leq S \cdot \left\{ \max_{i=k,\ldots,n} \left| \tau_i - \frac{1}{h}\delta_i \right| + \max_{i=1,\ldots,k-1} |\varepsilon_i| \right\}$$

Konvergenzsatz

Satz 6.3. *Sei $|f_x(x,t)| \leq L$, $x \in \mathbb{R}^m$, $t \in [t_0, T]$. Das lineare k-Schrittverfahren (6.3) sei konsistent und stabil. Es gelte $|\beta_0| h_* < \frac{|\alpha_0|}{L}$. Dann gilt:*

(1) *Bei Verwendung von Startwerten $x_i = x(t_i)$, d.h. $\varepsilon_i = 0$, $\delta_i = 0$, $i = 0, \ldots, k-1$, gilt*

$$\boxed{\max_{l=k,\ldots,n} |x(t_l) - x_l| \leq S \cdot \max_{l=k,\ldots,n} |\tau_l|}$$

mit einer Konstanten S gleichmäßig für alle Zerlegungen $t_0 < \ldots < t_n = T$, $h = \frac{T-t_0}{n} \leq h_*$

(2) *Es gilt die Fehlerabschätzung*

$$\boxed{|\varepsilon_l| \leq K \cdot e^{\tilde{L}(t_l - t_0)} \left\{ \max_{i=1,\ldots,k-1} |\varepsilon_i| + \max_{i=k,\ldots,l} \left| \tau_i - \frac{\delta_i}{h} \right| \right\}}$$

mit Konstanten K, $\tilde{L}$ gleichmäßig für alle Zerlegungen mit $h \leq h_*$.

Bemerkung.

(1) Aus (1) folgt Konvergenz für $h \to 0$ mit der Ordnung p

(2) $S := K \cdot e^{\tilde{L}(t_l - t_0)}$

Zum Beweis des Satzes 6.3 benötigen wir zunächst folgendes Lemma:

Lemma 6.4.

(1) *Zu $F \in L(\mathbb{C}^m)$ und $\varepsilon > 0$ gibt es eine $\mathbb{C}^m$-Norm $|.|_{F,\varepsilon}$ mit der Eigenschaft*

$$\|F\|_{F,\varepsilon} := \max_{|z|_{F,\varepsilon}=1} |Fz|_{F,\varepsilon} \leq \rho(F) + \varepsilon$$

(ρ sei der Spektralradius).

(2) *Haben alle betragsmaximalen Eigenwerte von F einfache Struktur (algebraische Vielfachheit = geometrische Vielfachheit), so gibt es eine $\mathbb{C}^m$-Norm $|.|_*$ mit*

$$\|F\|_* = \rho(F).$$

Beweis.

(1) Sei $F = TJT^{-1}$ mit

$$J = \begin{pmatrix} \lambda_1 & \delta_1 & 0 & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \delta_{m-1} \\ 0 & 0 & 0 & \lambda_m \end{pmatrix}$$

δ_i sind 0 oder 1, $|\lambda_1| \geq \ldots \geq |\lambda_m|$.

$$J_\varepsilon := \begin{pmatrix} \lambda_1 & \varepsilon\delta_1 & 0 & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \varepsilon\delta_{m-1} \\ 0 & 0 & 0 & \lambda_m \end{pmatrix}$$

Es gilt $J_\varepsilon = D_\varepsilon^{-1} J D_\varepsilon$ mit

$$D_\varepsilon := \begin{pmatrix} \varepsilon^0 & 0 & \cdots & 0 \\ 0 & \varepsilon^1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \varepsilon^{m-1} \end{pmatrix}$$

$$F = T \cdot D_\varepsilon \cdot J_\varepsilon \cdot D_\varepsilon^{-1} \cdot T^{-1} = (TD_\varepsilon) \cdot J_\varepsilon \cdot (TD_\varepsilon)^{-1}$$

Wir definieren eine Norm auf $\mathbb{C}^m$:

$$|z|_{F,\varepsilon} := \left|(TD_\varepsilon)^{-1} z\right|_1, \; z \in \mathbb{C}^m$$

Induzierte Norm auf $M \in \mathbb{C}^m$:

$$\begin{aligned} \|M\|_{F,\varepsilon} &:= \max_{|z|_{F,\varepsilon}=1} |Mz|_{F,\varepsilon} \\ &= \max_{|(TD_\varepsilon)^{-1}z|_1=1} \left|(TD_\varepsilon)^{-1} M (TD_\varepsilon) \underbrace{(TD_\varepsilon)^{-1} z}_{w}\right|_1 \\ &= \max_{|w|_1=1} \left|(TD_\varepsilon)^{-1} M (TD_\varepsilon) w\right|_1 \\ &= \left\|(TD_\varepsilon)^{-1} M (TD_\varepsilon)\right\|_1 \end{aligned}$$

Speziell gilt:

$$\|F\|_{F,\varepsilon} = \left\|(TD_\varepsilon)^{-1} F (TD_\varepsilon)\right\|_1 = \|J_\varepsilon\|_1 \leq \max_{i=1,\ldots,m} |\lambda_i| + \varepsilon = \rho(F) + \varepsilon$$

(2) $s-1$ sei die Anzahl der betragsmaximalen Eigenwerte. Dann ist $\delta_i = 0$ für $i = 1, \ldots, s-1$:

$$J = \begin{pmatrix} \lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_{s-1} & \delta_{s-1} & & \\ & & & \lambda_s & \ddots & \\ & & & & \ddots & \delta_{m-1} \\ & & & & & \lambda_m \end{pmatrix}$$

Wir wählen ε so, daß $|\lambda_i| + \varepsilon \leq \rho(F)$, $i = s, \ldots, m$ und wenden (1) an.

□

Explizites Euler-Verfahren: $\alpha_0 = 1$, $\alpha_1 = -1$, $\beta_0 = 0$, $\beta_1 = 1$, $\|B_{l,j}\| \le L$ für $l \ge k = 1$, $j = 0, 1$.

$$\begin{aligned}
\varepsilon_l &= \varepsilon_{l-1} + hB_{l,1}\varepsilon_{l-1} + h\left(\tau_l - \frac{\delta_l}{h}\right), \quad \varepsilon_0 = 0 \\
|\varepsilon_l| &\le (1+hL)\cdot|\varepsilon_{l-1}| + h\cdot \underbrace{\max_{i=1,\dots,l}\left|\tau_i - \frac{\delta_i}{h}\right|}_{=:\omega_l} \\
&\le (1+hL)^l\cdot|\varepsilon_0| + \sum_{i=0}^{l-1}(1+hL)^i\,\omega_l h \\
&= (1+hL)^l\cdot|\varepsilon_0| + \omega_l h\frac{(1+hL)^l - 1}{1+hL-1} \\
&\le (1+hL)^l\cdot\left(|\varepsilon_0| + \frac{1}{L}\omega_l\right), \quad (1+hL)^l \le e^{hLl} = e^{L(t_l - t_0)} \\
&\le \max\left(1, \frac{1}{L}\right)\cdot e^{L(t_l-t_0)}\cdot(|\varepsilon_0| + \omega_l)
\end{aligned}$$

Beweis. (von Satz 6.3):

(1) Fall $\beta_0 = 0$: Wir führen die k-Schritt-Rekursion auf eine 1-Schritt-Rekursion zurück:

$$\begin{aligned}
\varepsilon_l &= -\sum_{j=1}^{k}\frac{\alpha_j}{\alpha_0}\varepsilon_{l-j} + h\sum_{j=1}^{k}\frac{\beta_j}{\alpha_0}B_{l,j}\varepsilon_{l-j} + h\left(\tau_l - \frac{\delta_l}{h}\right) \quad l \ge k \\
\varepsilon_{l-1} &= \varepsilon_{l-1} \\
&\vdots \\
\varepsilon_{l-k+1} &= \varepsilon_{l-k+1} \\
\rightsquigarrow\ \mathcal{E}_l &= \mathcal{A}\cdot\mathcal{E}_{l-1} + h\cdot\mathcal{B}_l\cdot\mathcal{E}_{l-1} + h\cdot\Omega_l \\
\mathcal{A} &= \begin{bmatrix} -\frac{\alpha_1}{\alpha_0}I & \cdots & \cdots & -\frac{\alpha_k}{\alpha_0}I \\ I & 0 & \cdots & 0 \\ & \ddots & \ddots & \vdots \\ 0 & & I & 0 \end{bmatrix} \\
\mathcal{B} &= \begin{bmatrix} \frac{\beta_1}{\alpha_0}B_{l,1} & \cdots & \frac{\beta_k}{\alpha_0}B_{l,k} \\ 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \\
\Omega_l &= \begin{bmatrix} \tau_l - \frac{\delta_l}{h} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\
\mathcal{E}_l &:= \begin{pmatrix} \varepsilon_l \\ \vdots \\ \varepsilon_{l-k+1} \end{pmatrix} \in \mathbb{R}^{mk}
\end{aligned}$$

$\mathcal{A}$ ist Frobenius-Blockmatrix, $\mathcal{A} = \begin{pmatrix} -\frac{\alpha_1}{\alpha_0} & \cdots \\ \vdots & \ddots \end{pmatrix} \otimes I$

Nach dem Dahlquistischen Wurzelkriterium gilt für die Eigenwerte von $\mathcal{A}$: $|\lambda_{\mathcal{A}}| < 1$ oder $|\lambda_{\mathcal{A}}| = 1$ mit algebraischer Vielfachheit = geometrischer Vielfachheit.

Nach Lemma 6.4 existieren dann $|.|_*$ und $\|.\|_*$ mit $\|\mathcal{A}\|_* = 1$.

Äquivalenz der Normen auf $\mathbb{R}^{mk}$: $Z = \begin{pmatrix} z_1 \\ \vdots \\ z_k \end{pmatrix}$, $z_i \in \mathbb{R}^m$, $|Z|_\infty := \max\limits_{i=1,\dots,k} |z_i|$

$$\gamma_1 |Z|_\infty \leq |Z|_* \leq \gamma_2 |Z|_\infty$$

$$\begin{aligned}
\|\mathcal{B}_l\|_* &= \max_{Z\neq 0} \frac{|\mathcal{B}_l Z|_*}{|Z|_*} \\
&\leq \max_{Z\neq 0} \frac{\gamma_2 |\mathcal{B}_l Z|_\infty}{\gamma_1 |Z|_\infty} \\
&= \frac{\gamma_2}{\gamma_1} \cdot \|\mathcal{B}_l\|_\infty \\
&= \frac{\gamma_2}{\gamma_1} \cdot \sum_{j=1}^{k} \left|\frac{\beta_j}{\alpha_0}\right| \cdot \|B_{l,j}\| \\
&\leq \underbrace{\frac{\gamma_2}{\gamma_1} \cdot \sum_{j=1}^{k} \left|\frac{\beta_j}{\alpha_0}\right|}_{=:\tilde{L}} \cdot L \\
|\Omega_l|_* &\leq \gamma_2 \cdot |\Omega_l|_\infty \\
&= \gamma_2 \cdot \left|\tau_l - \frac{\delta_l}{h}\right| \\
&\leq \gamma_2 \cdot \underbrace{\max_{i=k,\dots,l} \left|\tau_i - \frac{\delta_i}{h}\right|}_{=\omega_l} \\
\rightsquigarrow |\mathcal{E}_l|_* &\leq |\mathcal{E}_{l-1}|_* + h \cdot \tilde{L} |\mathcal{E}_{l-1}|_* + h \cdot \gamma_2 \omega_l \\
&= \left(1 - h\tilde{L}\right) \cdot |\mathcal{E}_{l-1}| + h\gamma_2\omega_l \\
&\leq \left(1 + h\tilde{L}\right)^{l-k+1} \cdot |\mathcal{E}_{k-1}|_* + \sum_{i=0}^{l-k} \left(1 + h\tilde{L}\right)^i \cdot \gamma_2\omega_l \cdot h \\
&\leq \left(1 + h\tilde{L}\right)^{l-k+1} \cdot \left\{ |\mathcal{E}_{k-1}|_* + \frac{1}{\tilde{L}}\gamma_2\omega_l \right\} \\
&\leq \left(1 + h\tilde{L}\right)^{l} \cdot \left\{ \gamma_2 \cdot \max_{i=1,\dots,k-1} |\varepsilon_i| + \gamma_2 \frac{1}{\tilde{L}}\omega_l \right\} \quad (\varepsilon_0 = 0) \\
&\leq e^{\tilde{L}hl} \cdot \left\{ \gamma_2 \cdot \max_{i=1,\dots,k-1} |\varepsilon_i| + \gamma_2 \frac{1}{\tilde{L}}\omega_l \right\} \\
\rightsquigarrow |\varepsilon_l| &\leq |\mathcal{E}_l|_\infty \leq \frac{1}{\gamma_1} |\mathcal{E}_l|_* \\
&\leq \frac{\gamma_2}{\gamma_1} e^{\tilde{L}hl} \cdot \left\{ \max_{i=1,\dots,k-1} |\varepsilon_i| + \frac{1}{\tilde{L}}\omega_l \right\} \\
K &:= \frac{\gamma_2}{\gamma_1} \max\left(1, \frac{1}{\tilde{L}}\right)
\end{aligned}$$

liefert dann die Behauptung (2).

(2) Fall $\beta_0 \neq 0$: Rückführung auf Fall $\beta_0 = 0$: (6.7) liefert

$$\underbrace{\left(I - h\frac{\beta_0}{\alpha_0}B_{l,0}\right)}_{H_l}\varepsilon_l = -\sum_{j=1}^{k}\frac{\alpha_j}{\alpha_0}\varepsilon_{l-j} + h\sum_{j=1}^{k}\frac{\beta_j}{\alpha_0}B_{l,j}\varepsilon_{l-j} + h\left(\tau_l - \frac{\delta_l}{h}\right), \quad l \geq k$$

Dann ist

$$\|H_l - I\| = h \cdot \left|\frac{\beta_0}{\alpha_0}\right| \cdot \|B_{l,0}\| \leq h \cdot \left|\frac{\beta_0}{\alpha_0}\right| \cdot L \leq h_* \cdot \left|\frac{\beta_0}{\alpha_0}\right| \cdot L < 1$$

Nach dem Störungslemma (L. 1.1) ist dann H_l regulär und

$$\left\|H_l^{-1}\right\| \leq \frac{1}{1 - h_*\left|\frac{\beta_0}{\alpha_0}\right| L}$$

$$H_l^{-1} = I + h \cdot \frac{\beta_0}{\alpha_0}H_l^{-1}B_{l,0}$$

Zusammenfassung.

$\varepsilon_l, \ldots, \varepsilon_{l-k}$ k-Schritt-Rekursion der Dimension m

$\Downarrow$ Aufblasen

$\mathcal{E}_l, \mathcal{E}_{l-1}$ 1-Schritt-Rekursion der Dimension $m \cdot k$

$\Downarrow$ $\mathcal{A}$, Lemma 6.4, Wurzelkriterium

Exponentialabschätzung

$\Downarrow$

Rückrechnung

□

Satz 6.5.

(1) *Alle Adams-Verfahren sind stabil.*

(2) *Die BDF ist für $k \leq 6$ stabil.*

Beweis.

(1) $\alpha_0 = 1$, $\alpha_1 = -1$, $\alpha_i = 0$, $i = 2, \ldots, k$

$\rho(\lambda) = \lambda^k - \lambda^{k-1}$, Nullstellen $\lambda_1 = 1$, $\lambda_i = 0$, $i = 2, \ldots, k$

(2) $k = 1$: $\lambda_1 = 1$

$k = 2$: $\lambda_1 = 1$, $\lambda_2 = \frac{1}{3}$

□

$$|\varepsilon_l| \leq K \cdot \varepsilon^{\tilde{L}\cdot(t_l - t_0)} \cdot \max\left(1, \frac{1}{\tilde{L}}\right) \cdot \left\{\max_{i=1,\ldots,k}|\varepsilon_i| + \max_{i=k,\ldots,l}\left|\tau_i + \frac{\delta_i}{h}\right|\right\}$$

Überschlagsrechnung: Toleranz= 10^{-10}.

Sei $\varepsilon_l \approx \tau_l + \frac{\delta_l}{h}$, $\tau_l \approx h^p$, $\delta_l \approx 10^{-15}$.

$p = 1$: wählen $h = 10^{-10}$: $\varepsilon_l \approx 10^{-10} + \frac{10^{-15}}{10^{-10}} \approx 10^{-5}$

$p = 2$: wählen $h = 10^{-5}$: $\varepsilon_l \approx 10^{-10} + \frac{10^{-15}}{10^{-5}} \approx 10^{-10}$

6.3. **Reflexion des qualitativen asymptotischen Lösungsverhalten auf $[t_0, \infty)$.**

Wir betrachten nun die lineare DGL

$$\begin{aligned} &x'(t) = Bx(t), \quad t \in [0,\infty), \quad t_0 = 0,\ x(0) = x_0 \\ &\delta_1, \ldots, \delta_m \text{ seien die Eigenwerte von } B, \quad \mathrm{Re}\delta_i < 0, \quad i = 1, \ldots, m \end{aligned} \tag{6.8}$$

Lösungen der AWA-en für $x_0 \in \mathbb{R}^m$ sind

$$x(t) = e^{tB} \cdot x_0 \xrightarrow{t\to\infty} 0.$$

$t_i := t_0 + ih$, $i \in \mathbb{N}$, $h > 0$. $x(t_l) \xrightarrow{l\to\infty} 0$ für $x_0 \in \mathbb{R}^m$ und alle $h > 0$.

Numerisches Verfahren: $\{x_l\}_{l\geq 0}$. Frage: $x_l \overset{?}{\to} 0$ für $l \to \infty$.

Beispiel. $m = 1$, $x'(t) = -10^3 \cdot x(t) \rightsquigarrow x(t_l) = e^{-1000h\cdot l} x_0$

(1) explizites Euler-Verfahren:

$$x_l = x_{l-1} - h \cdot 10^3 x_{l-1} = \left(1 - h \cdot 10^3\right) x_{l-1} = \left(1 - h \cdot 10^3\right)^l x_0$$

Bedingung für Nullfolge: $\left|1 - h \cdot 10^3\right| < 1 \rightsquigarrow$ starke Einschränkung der Schrittweite

(2) implizites Euler-Verfahren:

$$\begin{aligned} x_l &= x_{l-1} - h \cdot 10^3 x_l \\ x_l &= \frac{1}{1 + h \cdot 10^3} \cdot x_{l-1} = \left(\frac{1}{1 + h \cdot 10^3}\right)^l \cdot x_0 \xrightarrow{l\to\infty} 0 \end{aligned}$$

für beliebige $h > 0$, d.h. unabhängig von der Schrittweite konvergiert x_l.

6.3.1. *Allgemeines lineares Einschrittverfahren* ($k = 1$).

$\alpha_0 = 1$, $\alpha_1 = -1$, $\beta_0 + \beta_1 = 1$, $\beta_0 \geq 0$, angewandt auf (6.8)

$$\begin{aligned} \frac{1}{h}(x_l - x_{l-1}) &= \beta_0 B x_l + \beta_1 B x_{l-1} \\ (I - \beta_0 hB) x_l &= (I + \beta_1 hB) x_{l-1} \end{aligned} \tag{6.9}$$

$I - \beta_0 hB$ ist regulär, da für die Eigenwerte $\mathrm{Re}(1 - \beta_0 h\delta_i) > 1$ gilt, also kein Eigenwert 0 ist.

$$\begin{aligned} x_l &= \mathcal{A}(hB) \cdot x_{l-1} \text{ mit } \mathcal{A}(hB) := (I - \beta_0 hB)^{-1}(I + \beta_1 hB) \text{ (Übergangsfkt.)} \\ x_l &= (\mathcal{A}(hB))^l \cdot x_0 \end{aligned}$$

$(\mathcal{A}(hB))^l \xrightarrow{l\to\infty} 0 \Leftrightarrow$ alle Eigenwerte von $\mathcal{A}(hB)$ liegen im Innern des komplexen Einheitskreises.

Die Eigenwerte von $\mathcal{A}(hB)$ sind $\frac{1+\beta_1 h\delta_i}{1-\beta_0 h\delta_i}$, $i = 1, \ldots, m$. Wir definieren:

$$\lambda(z) := \frac{1 + \beta_1 z}{1 - \beta_0 z}, \quad z \in \mathbb{C}$$

Definition. Die Menge

$$\boxed{G := \{z \in \mathbb{C} \,|\, |\lambda(z)| < 1\}}$$

heißt Stabilitätsbereich des Verfahrens. Stabilitäts-bereich

Satz 6.6. *Bei Anwendung des Einschrittverfahrens mit $h > 0$ auf die DGL (6.8) werden für alle $x_0 \in \mathbb{R}^m$ genau dann Nullfolgen $\{x_l\}_{l\geq 0}$ erzeugt, wenn*

$$h\sigma_i \in G, \quad i = 1, \ldots, m.$$

Definition. Ein Verfahren heißt A-stabil (absolut stabil), falls A-stabil

$$\boxed{\mathbb{C}^- := \{z \in \mathbb{C} \,|\, \mathrm{Re}(z) < 0\} \subseteq G}$$

Beispiel.

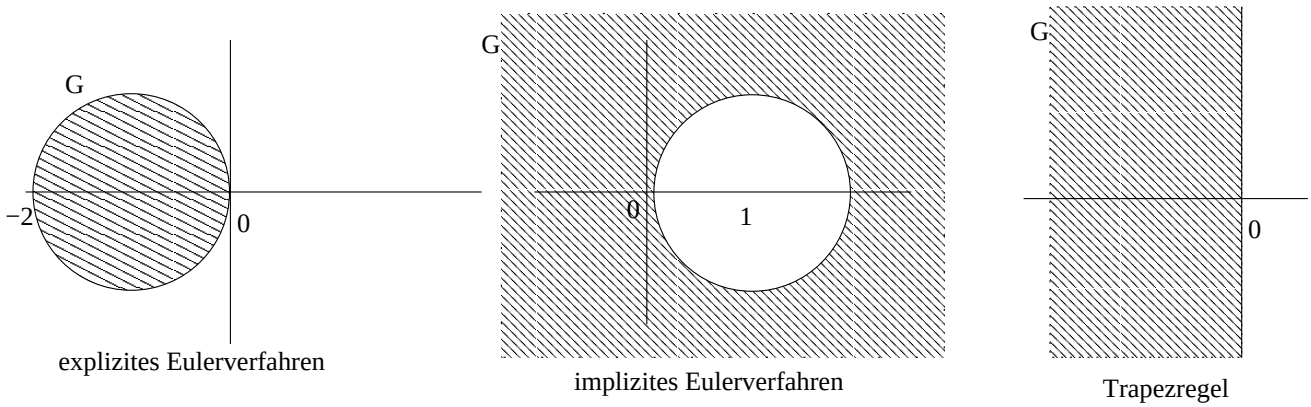


ABBILDUNG 6.1. Stabilitätsbereiche für verschiedene lineare Einschrittverfahren

- expliziter Euler: $\lambda(z) = 1 + z$
- impliziter Euler: $\lambda(z) = \frac{1}{1-z}$
- Trapezregel: $\lambda(z) = \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}$, $G = \mathbb{C}^{-1}$

6.3.2. *Allgemeines lineares k-Schrittverfahren.*

Das betrachtete k-Schrittverfahren sei stabil und konsistent, $\alpha_0 \neq 0$, $\beta_0 \geq 0$. Angewandt auf (6.8) ergibt dies:

$$\begin{aligned} \frac{1}{h}\sum_{j=0}^{k} \alpha_j x_{l-j} &= \sum_{j=0}^{k} \beta_j B x_{l-j} \\ \sum_{j=0}^{k} \left(\alpha_j I - \beta_j \underbrace{hB}_{z} \right) x_{l-j} &= 0 \end{aligned}$$

Konstruiere begleitendes Polynom

$$\boxed{p(\lambda, z) := \sum_{j=0}^{k} (\alpha_j - \beta_j z)\, \lambda^{k-j}, \quad z \in \mathbb{C},\ \lambda \in \mathbb{C}}$$

Stabilitätsbereich

Definition. Seien $\lambda_1(z), \ldots, \lambda_k(z)$ die Nullstellen des Polynoms $p(\cdot, z)$. Dann heißt

$$\boxed{G := \{z \in \mathbb{C} \,|\, |\lambda_i(z)| < 1, \quad i = 1, \ldots, k\}}$$

Stabilitätsbereich.

Beispiel. $k = 1$, $\alpha_0 = 1$, $\alpha_1 = -1$:

$$\begin{aligned} p(\lambda, z) &= (1 - \beta_0 z)\lambda + (-1 - \beta_1 z) \\ \lambda_1(z) &= \frac{1 + \beta_1 z}{1 - \beta_0 z} \end{aligned}$$

Für ein lineares k-Schrittverfahren ist

$$\underbrace{\begin{pmatrix} x_l \\ \vdots \\ x_{l-k+1} \end{pmatrix}}_{X_l} = \mathbb{A}(hB) \underbrace{\begin{pmatrix} x_{l-1} \\ \vdots \\ x_{l-k} \end{pmatrix}}_{X_{l-1}}$$

mit

$$\mathbb{A}(hb) = \begin{pmatrix} -(\alpha_0 I - \beta_0 hB)^{-1}(\alpha_1 I - \beta_1 hB) & \dots & \dots & -(\alpha_0 I - \beta_0 hB)^{-1}(\alpha_k I - \beta_k hB) \\ I & 0 & & \\ & \ddots & \ddots & \\ 0 & & I & 0 \end{pmatrix}$$

und $\mathbb{A}(hb) \in L\left(\mathbb{R}^{mk}\right)$ (Block-Frobenius) hat die Eigenwerte $\lambda_i(h\delta_j)$, $j = 1, \dots, m$, $i = 1, \dots, k$.

Satz 6.7. *Es gilt:*

(1) *Durch ein k-Schrittverfahren ($\alpha_0 > 0$, $\beta_0 \geq 0$) werden bei Anwendung mit der Schrittweite $h > 0$ auf eine stabile Differentialgleichung (6.8) Nullfolgen erzeugt, falls*

$$h\delta_i \in G, \quad i = 1, \dots, m$$

(2) *Ist das Verfahren A-stabil, so werden bei Anwendung auf (6.8) Nullfolgen unabhängig von der Größe der Schrittweite h erzeugt.*

Beispiel für A-stabile Verfahren sind die Trapezregel, das implizite Euler-Verfahren (BDF mit $k = 1$), BDF mit $k = 2$. Die BDF für $k = 3, 4, 5, 6$ sind „fast" A-stabil.

Satz 6.8. *Kein explizites lineares stabiles und konsistentes Mehrschrittverfahren ist A-stabil.*

Beweis. Angenommen, ein explizites Verfahren ($\beta_0 = 0$) sei A-stabil, das heißt, $\mathbb{C}^- \subset G$. Das heißt, für $z \in \mathbb{C}^-$ gilt für die Nullstellen des begleitenden Polynoms:

$$|\lambda_i(z)| < 1, \quad i = 1, \dots, k$$

Als Konsistenzbedingungen haben wir:

$$\sum_{j=0}^{k} \alpha_j = 0$$

$$\sum_{j=0}^{k} (\alpha_j j + \beta_j) = 0$$

und desweiteren gilt das Dahlquistsche Wurzelkriterium.

Wir erhalten als begleitendes Polynom:

$$p(\lambda, z) = \alpha_0 \left(\lambda^k + \left(\frac{\alpha_1}{\alpha_0} - \frac{\beta_1}{\alpha_0} z \right) \lambda^{k-1} + \dots + \left(\frac{\alpha_k}{\alpha_0} - \frac{\beta_k}{\alpha_0} z \right) \lambda^0 \right)$$

Nach dem Satz von Vieta gilt für alle $z \in \mathbb{C}^-$:

$$\left| \frac{\alpha_k}{\alpha_0} - \frac{\beta_k}{\alpha_0} z \right| \leq |\lambda_1(z)| \cdots |\lambda_k(z)| \leq 1$$

und hieraus folgt sofort $\beta_k = 0$, da kein Polynom ersten Grades beschränkt sein kann. Wenden wir den gleichen Satz auf die anderen Koeffizienten an, so erhalten wir auch

$$\left| \frac{\alpha_{k-1}}{\alpha_0} - \frac{\beta_{k-1}}{\alpha_0} z \right| \leq c$$

etc und es gilt $\beta_1 = \dots = \beta_{k-1} = 0$.

Aus der Konsistenzbedingung folgt $\rho(1) = \sum \alpha_j = 0$ und $\sum \alpha_j j = 0$. Daraus folgt $\rho'(1) = 0$, d.h. 1 ist mehrfache Nullstelle des charakteristischen Polynoms. Dies ist ein Widerspruch zur Stabilität (Dahlquistsches Wurzelkriterium). ↯

□

Bezeichnung: Steife Differentialgleichungen sind Differentialgleichungen mit verschieden schnell abklingenden (d.h. für $t \to \infty$ gegen 0 konvergierende) Komponenten. („mit verschiedenen Zeitkonstanten").

Beispiel. Sei $m = 2$, $x'(t) = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} x(t)$ wobei σ_1, σ_2 die Eigenwerte seien. Dann gilt:

$$e^{hB} = \begin{pmatrix} e^{h\sigma_1} & 0 \\ 0 & e^{h\sigma_2} \end{pmatrix}$$

Wir betrachten ein Einschrittverfahren: $k = 1$, $\alpha_0 = 1$, $\alpha_1 = -1$:

$$\mathbb{A}(hB) = (I - \beta_0 hB)^{-1} (I + \beta_1 hB) = \begin{pmatrix} \lambda_1(h\sigma_1) & 0 \\ 0 & \lambda_1(h\sigma_2) \end{pmatrix}$$

Sei h moderat, $h \approx 0,1$ und $\sigma_1 \approx -0,1$. Dann ist $\lambda_1(h\sigma_1)$ für alle Verfahren eine „gute" Approximation für $e^{h\sigma_1}$. Sei $\sigma_2 = -10^6$:

- Trapezregel $\left(\beta_0 = \beta_1 = \frac{1}{2}\right)$:

$$\lambda_1(h\sigma_2) = \frac{1 - \frac{1}{2}10^6 \cdot 0.1}{1 + \frac{1}{2}10^6 \cdot 0.1} \approx -1$$

- implizites Euler-Verfahren $(\beta_0 = 1, \beta_1 = 0)$:

$$\lambda_1(h\sigma_2) = \frac{1}{1 + 10^6 \cdot 0.1} \approx 10^{-5}$$

Definition. Ein lineares Mehrschrittverfahren heißt L-stabil, falls es A-stabil ist und zusätzlich

$$\lim_{\mathrm{Re}(z) \to -\infty} \max_{i=1,\ldots,k} |\lambda_i(z)| = 0.$$

Das implizite Eulerverfahren $(\lambda_1(z) = \frac{1}{1-z})$ ist L-stabil. Alle BDF sind „fast" L-stabil.

6.4. **Zur praktischen Realisierung.**

(1) Wie kann man x_l beim impliziten Verfahren berechnen, wenn $\tilde{x}_{l-k}, \ldots, \tilde{x}_{l-1}$ bereits (als Näherung) berechnet sind? Als Gleichung für x_l haben wir bereits:

$$\gamma_l := x_l - h\frac{\beta_0}{\alpha_0} f(x_l, t_l) = \sum_{j=1}^{k} \left(-\frac{\alpha_j}{\alpha_0} \tilde{x}_{l-j} + \frac{\beta_j}{\alpha_0} h f(\tilde{x}_{l-j}, t_{l-j}) \right)$$

Möglichkeiten:

- Newton-Verfahren mit Startwert $x_{l,[0]} := \tilde{x}_{l-1}$ oder aus explizitem Verfahren
- Fixpunkt-Iteration:

$$\begin{aligned} \mathbb{B}_l(x) &:= h\frac{\beta_0}{\alpha_0} f(x, t_l) + \gamma_l, \\ x_{l,[i+1]} &:= \mathbb{B}_l\left(x_{l,[i]}\right), \quad i \geq 0. \end{aligned}$$

$$|\mathbb{B}_l(x) - \mathbb{B}_l(\bar{x})| \leq h \left|\frac{\beta_0}{\alpha_0}\right| L |x - \bar{x}|, \quad x, \bar{x} \in \mathbb{R}^m$$

Wegen (6.8) erhalten wir

$$h \left|\frac{\beta_0}{\alpha_0}\right| \|B\| < 1$$

als Konvergenzbedingung für die Iteration. Für $\alpha_0 = \beta_0 = 0$ muß damit $h|\sigma_i| \leq h\|B\| < 1$ gelten.

Die Fixpunkt-Iteration, obwohl sie weniger Rechnungen benötigt, ist in der Praxis nicht ernsthaft anwendbar, da wieder eine Schrittweitenbeschränkung existiert.

(2) Schrittweitenänderung

Neue Schrittweiten h_j, $t_{l-j} = t_l - jh_j$, führen zu neuen Stützstellen $\bar{t}_{l+1-j} := t_{l+1} - jh_{l+1}$.

Die $\bar{x}_{l+1-j}$ können durch Interpolation bereitgestellt werden.

Eine andere Möglichkeit besteht darin, Formeln für variable Schrittweiten (Adams, BDF) zu benutzen.

(3) Fehlerabschätzung, Fehlerkontrolle

Der exakte Fehler $x(t_l) - \tilde{x}_l$ ist nicht verfügbar. Man muß sich also mit einer Abschätzung $\hat{x}_l - \tilde{x}_l$ begnügen, welche mit einer zweiten Näherung $\hat{x}_l$ ($\hat{p} \geq p$, bzw. $\hat{x}$ ist „genauer") bestimmt wird. Dabei soll $\hat{x}_l$ mit wenig Aufwand berechnet sein und eine genauere Abschätzung als $\tilde{x}_l$ sein (z.B. indem es mit einem Verfahren höherer Konsistenzordnung als $\tilde{x}_l$ berechnet wurde).

Gilt $|\hat{x}_l - \tilde{x}_l| \leq TOL$, so wird $\tilde{x}_l$ als Näherung akzeptiert. Falls nicht, so wiederholen wir den letzten Schritt mit kleinerer Schrittweite.

(4) Schrittweitensteuerung (Ordnungssteuerung)

Um mit der größtmöglichen Schrittweite für eine Toleranz zu arbeiten, benutzt man Prognosen für die Schrittweite h_{l+1}:

$$\tau_l = \underbrace{c_p x^{(p+1)}(t_l)\, h_l^p}_{=:T_l} + o\left(h_l^{p+1}\right), \quad \tau_{l+1} = \underbrace{c_p x^{(p+1)}(t_{l+1})\, h_{l+1}^p}_{=T_{l+1}} + o\left(h_l^{p+1}\right)$$

(wir bezeichnen T_l als den Hauptteil des Fehlers).

Annahme: $x^{(p+1)}(t_l) = x^{(p+1)}(t_{l+1})$

$$\begin{aligned} \rightsquigarrow \frac{T_l}{h_l^p} &= \frac{T_{l+1}}{h_{l+1}^p} \\ \rightsquigarrow h_{l+1} &= h_l \cdot \sqrt[p]{\left|\frac{T_{l+1}}{T_l}\right|} \end{aligned}$$

Als Ansatz setzen wir $|T_{l+1}| = \mathrm{TOL}$ und berechnen eine Approximation $\tilde{T}_l$ von T_l:

$$\boxed{h_{l+1} := h_l \cdot \sqrt[p]{\frac{\mathrm{TOL}}{\left|\tilde{T}_l\right|}}}$$

Literatur: [SB90, 7. Kapitel]

7. Eigenwertaufgaben

Sei A eine $m \times m$-Matrix. Gesucht sind die Eigenwerte λ von A: $Az = \lambda z$, $z \neq 0$

Eine Möglichkeit, die Eigenwerte zu berechnen ist, $p(\lambda)$, das charakteristisches Polynom, zu benutzen mit $p(\lambda) = \det(A - \lambda I)$.

Beispiel. $A = \begin{pmatrix} 5 & 7 & 6 & 5 \\ 7 & 10 & 8 & 7 \\ 6 & 8 & 10 & 9 \\ 5 & 7 & 9 & 10 \end{pmatrix}$, $\Delta A = \begin{pmatrix} 0,1 & & & \\ & & & \\ & & & 0 \end{pmatrix}$

$$\begin{aligned} p_A(\lambda) &= \lambda^4 - 35\lambda^3 + 146\lambda^2 - 100\lambda + 1 \\ p_{A+\Delta A}(\lambda) &= \lambda^4 - 35,1\lambda^3 + 149\lambda^2 - 110,6\lambda + 7,8 \end{aligned}$$

Durch kleinere Störungen der Matrix A verändert sich das charakteristische Polynom nicht unwesentlich. Dieses Verfahren hätte also eine schlechte Kondition. Die mit diesem Verfahren berechneten Eigenwerte für normale Matrizen weichen aber nur ungefähr mit der Störung ab:

Eigenwerte von A	Eigenwerte von $A + \Delta A$
0,010	0,079
0,843	0,844
3,858	3,874
30,289	30,303

Für normale Matrizen $\left(A^T A = AA^T\right)$ liegen die Fehler in den Eigenwerten in der Größenordnung der Fehler der gestörten Matrix.

7.1. Einfache und inverse Vektoriteration zur Bestimmung spezieller Eigenwerte.

7.1.1. *Einfache Vektoriteration (von Mises-Iteration).*

Vektoriteration

Sei A eine $m \times m$-Matrix und diagonalisierbar, $\lambda_1, \ldots, \lambda_m$ Eigenwerte, $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_m|$, $z_1, \ldots, z_m$ Eigenvektor-Basis in $\mathbb{C}^m$.

$$\boxed{x^{(0)} \in \mathbb{C}^m, \qquad x^{(k)} := Ax^{(k-1)} = A^k x^{(0)}, \quad k \geq 1}$$

$$\begin{aligned} x^{(0)} &= \alpha_1 z_1 + \ldots + \alpha_m z_m \\ \rightsquigarrow\ x^{(k)} &= \alpha_1 \lambda_1^k z_1 + \alpha_2 \lambda_2^k z_2 + \ldots + \alpha_m \lambda_m^k z_m \\ &= \lambda_1^k \left\{ \alpha_1 z_1 + \alpha_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k z_2 + \ldots + \alpha_m \left(\frac{\lambda_m}{\lambda_1}\right)^k z_m \right\} \\ \frac{1}{\lambda_1^k} x^{(k)} &\xrightarrow{k\to\infty} \alpha_1 z_1, \text{ wenn } |\lambda_1| > |\lambda_2| \end{aligned}$$

Falls $|\lambda_1| = |\lambda_2| = \ldots = |\lambda_s| > |\lambda_{s+1}|$ und $\lambda_1 = \lambda_2 = \ldots = \lambda_s$, so folgt

$$\frac{1}{\lambda_1^k} x^{(k)} \xrightarrow{k\to\infty} \alpha_1 z_1 + \ldots + \alpha_s z_s,$$

wenn $(\alpha_1 z_1 + \ldots + \alpha_s z_s)_i \neq 0$, so gilt

$$\frac{\frac{1}{\lambda_1^k} x_i^{(k)}}{\frac{1}{\lambda_1^{k-1}} x_i^{(k-1)}} \xrightarrow{k\to\infty} 1 \qquad \rightsquigarrow \qquad \boxed{\frac{x_i^{(k)}}{x_i^{(k-1)}} \xrightarrow{k\to\infty} \lambda_1}$$

Dieses Verfahren bezeichnen wir als Vektoriteration zur Bestimmung des betragsgrößten Eigenwertes einer Matrix. Es wird auch als von Mises Verfahren bezeichnet.

7.1.2. *Inverse Vektoriteration (nach Wielandt).*

Sei A mit Eigenwerten $\lambda_1, \ldots, \lambda_m$ gegeben. Wir wählen

$$\mu \notin \{\lambda \in \mathbb{C} | \det(A - \lambda I) = 0\}.$$

Die Matrix $(A - \mu I)$ ist regulär.

Es gelte:

$$|\lambda_{j_*} - \mu| < |\lambda_i - \mu| \quad \text{für } \lambda_i \neq \lambda_{j_*}$$

(also sind auch mehrfache Eigenwerte erlaubt). D.h., μ liegt zu einem Eigenwert näher als zu allen anderen.

$(A - \mu I)^{-1}$ hat die Eigenwerte $\frac{1}{\lambda_i - \mu}$, $i = 1, \ldots, m$, und es folgt

$$\left|\frac{1}{\lambda_{j_*} - \mu}\right| > \left|\frac{1}{\lambda_i - \mu}\right| \quad \text{für } \lambda_i \neq \lambda_{j_*}$$

Die von Mises-Iteration für $(A - \mu I)^{-1}$ liefert für $k \geq 1$: inverse Vektor-iteration

$$x^{(k)} = (A - \mu I)^{-1} x^{(k-1)}$$

Dies führt zu

$$\boxed{x^{(0)} \in \mathbb{C}^m, \qquad (A - \mu I)\, x^{(k)} = x^{(k-1)}, \quad k \geq 1}$$

und damit

$$\boxed{\frac{x_i^{(k)}}{x_i^{(k-1)}} \xrightarrow{k \to \infty} \frac{1}{|\lambda_{j_*} - \mu|}}$$

7.2. **QR-Verfahren zur Bestimmung aller Eigenwerte einer reellen Matrix.**

Sei $A \in L(\mathbb{R}^m)$, $A = (a_{i,j})$.

Vorbemerkung: Zu $G \in L(\mathbb{R}^m)$ gibt es ein $Q \in L(\mathbb{R}^m)$, $Q^T = Q^{-1}$, und eine obere Dreiecksmatrix $R \in L(\mathbb{R}^m)$ mit $G = QR$.

Mit der Vorgabe $r_{i,i} \geq 0$, $i = 1, \ldots, m$, ist die Faktorisierung eindeutig.

Vereinbarung: Wir setzen $r_{i,i} \geq 0$, $i = 1, \ldots, m$.

QR-Verfahren: QR-Verfahren

$$\boxed{A_0 := A, \qquad A_k = Q_k R_k, \quad A_{k+1} := R_k Q_k, \quad k \geq 0}$$

$$A_{k+1} = Q_k^{-1} A_k Q_k = Q_k^{-1} \cdots Q_0^{-1} A_0 Q_0 \cdots Q_k = (Q_0 \cdots Q_k)^{-1} A (Q_0 \cdots Q_k),$$

also ist A_k ähnlich zu A für alle $k \geq 0$.

Satz 7.1. *(Satz von Schur): Zu $A \in L(\mathbb{R}^m)$ existiert eine orthogonale Matrix $U \in L(\mathbb{R}^m)$ mit*

$$U^{-1} A U = \begin{bmatrix} R_{1,1} & \cdots & R_{1,s} \\ & \ddots & \vdots \\ 0 & & R_{s,s} \end{bmatrix}$$

mit Blöcken $R_{i,i}$, die entweder die Größe 1×1 haben, oder die Größe 2×2. Die 2×2 Blöcke haben konjugiert komplexe Eigenwerte.

Beweis. [S93, S. 275-276] □

A hat Hessenberg-Form, falls $a_{i,j} = 0$ für $i \geq j + 2$.

Vorteil der Hessenberg-Form: Falls eines der Subdiagonalelemente verschwindet, zerfällt das Eigenwertproblem in solche mit kleinerer Dimension:

$$A = \underbrace{\begin{pmatrix} * & \cdots & \cdots & * \\ * & \ddots & * & \vdots \\ & 0 & \ddots & \vdots \\ 0 & & * & * \end{pmatrix}}_{m} = \begin{pmatrix} A_{1,1} & A_{1,2} \\ \underbrace{0}_{m_1} & \underbrace{A_{2,2}}_{m_2} \end{pmatrix}, \quad A_{1,1}, A_{1,2} \text{ haben Hessenberg-Form}$$

$$\det(A - \lambda I_m) = \det(A_{1,1} - \lambda I_{m_1}) \cdot \det(A_{2,2} - \lambda I_{m_2})$$

Transformation einer Matrix A in Hessenberg-Form:

$$A = \begin{bmatrix} a_{1,1} & \\ a_{2,1} & \cdots \\ \vdots & \\ a_{m,1} & \end{bmatrix}$$

Sei $\tilde{Q}_1 \in L\left(\mathbb{R}^{m-1}\right)$ mit

$$\tilde{Q}_1 \begin{bmatrix} a_{2,1} \\ \vdots \\ a_{m,1} \end{bmatrix} = \begin{bmatrix} \gamma_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\bar{Q}_1 := \begin{bmatrix} 1 & \\ & \tilde{Q}_1 \end{bmatrix}$$

$$A^{(1)} := \bar{Q}_1 A \bar{Q}_1^T = \begin{bmatrix} a_{1,1} & & \\ \gamma_1 & a_{2,2}^{(1)} & \cdots \\ 0 & a_{3,2}^{(1)} & \\ \vdots & \vdots & \\ 0 & a_{m,2}^{(1)} & \end{bmatrix}$$

Sei $\tilde{Q}_2 \in L\left(\mathbb{R}^{m-2}\right)$ mit

$$\tilde{Q}_2 \begin{bmatrix} a_{3,2}^{(1)} \\ \vdots \\ a_{m,2}^{(1)} \end{bmatrix} = \begin{bmatrix} \gamma_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\bar{Q}_2 := \begin{bmatrix} 1 & & \\ & 1 & \\ & & \tilde{Q}_2 \end{bmatrix}$$

$$A^{(2)} := \bar{Q}_2 A^{(1)} \bar{Q}_2^T = \begin{bmatrix} a_{1,1} & & \\ \gamma_1 & a_{2,2}^{(1)} & \cdots \\ 0 & \gamma_2 & \\ \vdots & 0 & \\ \vdots & \vdots & \\ 0 & 0 & \end{bmatrix}$$

usw. $A^{(m-2)}$ hat Hessenberg-Form.

Lemma. *Die Hessenberg-Form bleibt unter QR-Transformationen erhalten. Ist* $A^T = A$, *so ist auch* $A^{(k)^T} = A^{(k)}$.

Satz 7.2. *Sei* $A \in L(\mathbb{R}^m)$ *diagonalisierbar,* $A = TDT^{-1}$, $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$, $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_m| > 0$. T^{-1} *besitze eine LR-Zerlegung.*

Für den QR-Algorithmus gilt:

$$\boxed{Q_k \xrightarrow{k\to\infty} \operatorname{diag}\left(\frac{\lambda_1}{|\lambda_1|}, \ldots, \frac{\lambda_m}{|\lambda_m|}\right)}$$

und

$$\boxed{a_{k,i,i} \xrightarrow{k\to\infty} \lambda_i, \quad i = 1, \ldots, m}$$

$$\begin{bmatrix} \ddots & * & \cdots & * \\ \ddots & a_{k,m-2,m-2} & * & * \\ 0 & a_{k,m-1,m-2} & a_{k,m-1,m-1} & * \\ & 0 & a_{k,m,m-1} & a_{k,m,m} \end{bmatrix}$$

Wenn $a_{k,m,m-1} \approx 0$, so ist $a_{k,m,m} \approx \lambda_m \rightsquigarrow$ Abspaltung

Wenn vorher $a_{k,m-1,m-2} \approx 0$, so werden die Eigenwerte des Blockes $\begin{bmatrix} a_{k,m-1,m-1} & * \\ a_{k,m,m-1} & a_{k,m,m} \end{bmatrix}$ als Näherung für λ_{m-1}, λ_m berechnet, danach Abspaltung.

Bemerkung. QR-Algorithmus mit Spektralverschiebung (Shift)

$$\begin{aligned} & & A_0 &:= A \\ k \geq 0: \quad & A_k - \mu_k I &= Q_k R_k \\ & A_{k+1} &:= \mu_k I + R_k Q_k, \end{aligned}$$

z.B. $\mu_k = a_{k,m,m}$

Lemma 7.3. $A = A^T$ *habe Hessenberg-Form, A sei nicht zerfallend* ($a_{i+1,i} \neq 0$, $i = 1, \ldots, m$). *Dann hat A nur einfache Eigenwerte.*

Beweis.

$$A - \lambda I = \begin{bmatrix} a_{11} - \lambda & * & & \\ a_{21} & \ddots & \ddots & \\ & \ddots & \ddots & \\ & & a_{m,m-1} & a_{mm} - \lambda \end{bmatrix},$$

($m-1$ lin. unabh. Spalten)

d.h. $\operatorname{Rang}(A - \lambda I) \geq m - 1$ für alle λ.

Falls λ Eigenwert zu A ist, so ist $\operatorname{Rang}(A - \lambda I) = m - 1$ und $\dim(\ker(A - \lambda I)) = 1$.

Wegen $A^T = A$ ist λ einfacher Eigenwert.

□

7.3. **Abschätzung von Eigenwerten.**

Sei $A = (a_{i,j}) \in \mathbb{R}^{n\times n}$.

$$Ax = \lambda x \;\Leftrightarrow\; (\lambda - a_{ii})\, x_i = \sum_{j=1, j\neq i}^{n} a_{i,j} x_j, \quad 1 \leq i \leq n$$

Sei $k \in \{1, \ldots, n\}$ mit $|x_k| = |x|_\infty$. Dann ist

$$\boxed{|\lambda - a_{kk}| \leq \sum_{j=1, j\neq k}^{n} |a_{k,j}|}$$

7.3.1. *Satz von Gerschgorin (1931).*

Gersch-gorin-Kreise

Gerschgorin-Kreise

$$\boxed{K_k := \{z \in \mathbb{C} \,|\, |z - a_{kk}| \leq r_k\} \text{ mit Radius } r_k = \sum_{j=1, j\neq k}^{n} |a_{k,j}|\,, \quad k = 1, \ldots, n}$$

Dann gilt

$$\boxed{\sigma(A) \subseteq \bigcup_{k=1}^{n} K_k}$$

Beispiel. Sei

$$A = \begin{pmatrix} -1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 2 \end{pmatrix}$$

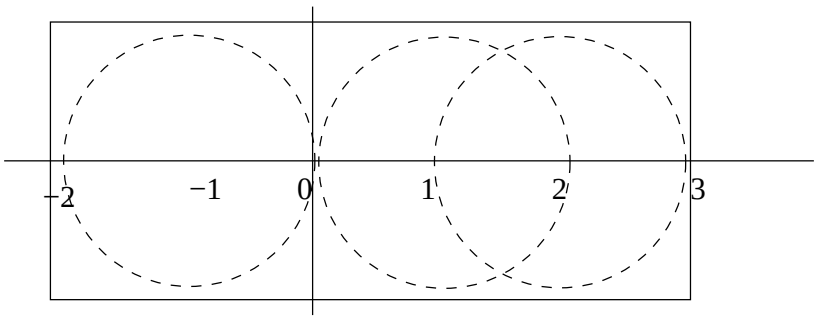


ABBILDUNG 7.1. Gerschorgin-Kreise von A

Eigenwerte: 2.33685, −1.16402, 0.827167

8. Iterations-Verfahren zur Lösung grosser linearer Gleichungssysteme

Gesucht ist eine Lösung des linearen Gleichungssystems

$$\boxed{Ax = b}$$

mit regulärer Matrix $A \in L(\mathbb{R}^m)$.

- A hat spezielle Struktur, bedingt durch seine Herkunft.
- Man ist nur an bestimmter Lösungsgenauigkeit interessiert.

Beispiel. Für gegebenes $f \in C^1$ ist eine eine Funktion u mit

$$-\Delta u = f(x,y) \qquad (x,y) \in \Omega = (0,1) \times (0,1)$$

gesucht. Δ ist der Laplace-Operator: $\Delta u = \frac{\partial^2 u}{\partial x_1^2} + \ldots + \frac{\partial^2 u}{\partial x_n^2}$. Also

$$-u_{xx}(x,y) - u_{yy}(x,y) = f(x,y)$$

Zudem soll gelten:

$$u(x,y)|_{\partial\Omega} = 0$$

Wir wählen ein Gitter

$$x_i = i \cdot h, \quad y_j = j \cdot h, \qquad i,j = 0,1,\ldots,m+1, \quad (m+1)\cdot h = 1$$

$u_{i,j}$ soll $u(x_i, y_j)$ approximieren

$$\begin{aligned} u_{xx}(x,y) &\approx \frac{1}{h^2}\left(u(x+h,y) - 2u(x,y) + u(x-h,y)\right) \\ u_{yy}(x,y) &\approx \frac{1}{h^2}\left(u(x,y+h) - 2u(x,y) + u(x,y-h)\right) \end{aligned}$$

Ersetze die Differentialgleichung in jedem „inneren" Gitterpunkt (x_i, y_j) durch die Differenzengleichung $f_{i,j}$:

$$-\frac{1}{h^2}\left(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}\right) - \frac{1}{h^2}\left(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}\right) = f_{i,j}, \quad i,j = 1,\ldots,m$$

ist Bestimmungsgleichung für m^2 Unbekannte $u_{1,1}, u_{2,1}, \ldots, u_{m,1}, \ldots, u_{m,m}$.

Rand: $u_{0,j} = 0$, $u_{m+1,j} = 0$, $u_{i,0} = 0$, $u_{i,m+1} = 0$

Man kann zeigen, daß $u_{i,j} - u(x_i, y_j) = O(h^2)$. Damit reicht die Lösungsgenauigkeit $O(h^2)$ aus!

$A_{(m)} u_{(m)} = h^2 f_{(m)}$ hat „sparse" Struktur:

$$u_{(m)} = \begin{pmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{mm} \end{pmatrix} \qquad A_{(m)} = \begin{pmatrix} 4 & -1 & & 0 & -1 & & & & \\ -1 & \ddots & \ddots & & & \ddots & & & 0 \\ & \ddots & \ddots & -1 & & & \ddots & & \\ 0 & & -1 & 4 & & & & -1 & \\ -1 & & & & 4 & -1 & & 0 & \\ & \ddots & & & -1 & \ddots & \ddots & & 0 \\ & & \ddots & & & \ddots & \ddots & -1 & \\ & & & -1 & 0 & & -1 & 4 & \\ & 0 & & & & 0 & & & \ddots \end{pmatrix}$$

typisch: Bandstruktur, Symmetrie, Dominanz der Diagonal-Elemente

Bei Gauß-Verfahren würde „fill in" passieren. (Durch LU-Zerlegung oder Householder können dann vollbesetzte Martizen entstehen, wodurch die „sparse"-Struktur verloren geht.)

$A \in L(\mathbb{R}^{2n})$ hat die Eigenwerte $\lambda_{k,l} = 2\left(2 - \cos\frac{k\pi}{n+1} - \cos\frac{l\pi}{n+1}\right)$, $k,l = 1,\ldots,n$. Wegen $\lambda_{\min} = 8\sin^2\left(\frac{\pi}{2(n+1)}\right)$ ist A dann positiv definit.

8.1. **Gesamtschritt-, Einzelschritt-, Relaxationsverfahren.**

Wir formen $Ax = b$ äquivalent in Fixpunktschreibweise um:

$$x = Bx + d$$

Iteration:

$$\boxed{x_0 \in \mathbb{R}^m, \qquad x_j := Bx_{j-1} + d, \quad j \in \mathbb{N}}$$

Satz 8.1. *Sei $B \in L(\mathbb{R}^m)$, $\rho(B) < 1$ und $d \in \mathbb{R}^m$. Dann gilt:*

(1) *$(I - B)$ ist regulär.*

(2) *Für beliebige $x_0 \in \mathbb{R}^m$ gilt*

$$\boxed{x_j \xrightarrow{j\to\infty} (I-B)^{-1} d = A^{-1} b}$$

Beweis.

(1) Wäre $I - B$ singulär, so gäbe es ein $z \in \mathbb{R}^m$, $z \neq 0$, mit $(I - B)z = 0$, d.h. 1 wäre Eigenwert von B. $\lightning$ zu $\rho(B) < 1$.

(2)

$$\begin{aligned} x_j &= Bx_{j-1} + d \\ &= B^2 x_{j-2} + Bd + d \\ &= B^j x_0 + \sum_{i=0}^{j-1} B^i d \\ (I-B) \cdot \sum_{i=0}^{j-1} B^i &= \sum_{i=0}^{j-1} B^i - \sum_{i=1}^{j} B^i = I - B^j \\ \text{also} \quad \sum_{i=0}^{j-1} B^i &= (I-B)^{-1}(I-B) \cdot \sum_{i=0}^{j-1} B^i = (I-B)^{-1} \cdot \left(I - B^j\right) \end{aligned}$$

Nach Lemma 6.4 existiert eine Norm $\|\cdot\|_*$ mit $\|B\|_* \leq \rho(B) + \varepsilon < 1$, also

$$\left\|B^j\right\|_* \leq \|B\|_*^j \xrightarrow{j\to\infty} 0$$

und

$$\left\| \sum_{i=0}^{j-1} B^i - (I-B)^{-1} \right\|_* \leq \left\|(I-B)^{-1}\right\|_* \cdot \|B\|_*^j \xrightarrow{j\to\infty} 0,$$

also $B^j \to 0$,

$$\begin{aligned} \sum_{i=0}^{j-1} B^i d &\xrightarrow{j\to\infty} (I-B)^{-1} d \\ \rightsquigarrow x_j &\xrightarrow{j\to\infty} (I-B)^{-1} d \end{aligned}$$

(D.h. Banachscher Fixpunktsatz für lineare Abbildungen mit $\rho < 1$.)

□

8.1.1. *Gesamtschrittverfahren (Jacobi).*

Voraussetzung: $a_{i,i} \neq 0$, $i = 1, \ldots, m$

$D := \operatorname{diag}(a_{1,1}, \ldots, a_{m,m})$

$$Ax = b \quad \sim \quad Dx = -(A-D)\,x + b \quad \sim \quad x = \underbrace{-D^{-1}(A-D)}_{B}\,x + \underbrace{D^{-1}b}_{d}$$

Gesamtschritt-verfahren

$$\boxed{x_i^{(j)} = \frac{1}{a_{i,i}} \cdot \left(b_i - \sum_{k=1, k\neq i}^{n} a_{i,k} x_k^{(j-1)} \right), \quad 1 \leq i \leq m}$$

8.1.2. *Einschrittverfahren (Gauß-Seidel).*

Voraussetzung: $a_{i,j} \neq 0$, $i = 1, \ldots, m$

Einzelschritt-verfahren

$$\boxed{x_i^{(j)} = \frac{1}{a_{i,i}} \left(b_i - \sum_{k=1}^{i-1} a_{i,k} x_k^{(j)} - \sum_{k=i+1}^{m} a_{i,k} x_k^{(j-1)} \right)}$$

$$A = \underbrace{L}_{\text{untere } \triangle\text{-Matrix}} + \underbrace{D}_{\text{Diagonalmatrix}} + \underbrace{R}_{\text{obere } \triangle\text{-Matrix}}$$

$$\begin{aligned} x_j &= -D^{-1}Lx_j - D^{-1}Rx_{j-1} + D^{-1}b \\ (D+L)\,x_j &= -Rx_{j-1} + b \\ \sim\ x_j &= \underbrace{-(D+L)^{-1}R}_{B} x_{j-1} + \underbrace{(D+L)^{-1}b}_{d} \end{aligned}$$

Satz 8.2. *Sei A strikt diagonaldominant, d.h.* $\displaystyle\sum_{j=1, j\neq i}^{m} |a_{i,j}| < |a_{i,i}|$.

Für das Gesamt- und das Einzelschrittverfahren gilt dann $\|B\|_\infty < 1$.

Beweis. Gesamtschrittverfahren: $B = -D^{-1}(L+R)$

$$\|B\|_\infty = \max_i \sum_{j=1, j\neq i}^{m} \frac{|a_{i,j}|}{|a_{i,i}|} < 1$$

Einzelschritt: $B = -(D+L)^{-1}R$

$$\|B\|_\infty = \max_{|x|_\infty = 1} |Bx|_\infty$$

$|x|_\infty = 1$, $y = Bx \rightsquigarrow |y|_\infty < 1$.

Siehe auch [SB90, Kapitel 8]

□

8.1.3. *Relaxations-Verfahren.*

Parameter ω (Relaxationsparameter zur Konvergenz-Beschleunigung)

Relaxations-Verfahren

$$\boxed{x_j = \omega\left(-D^{-1}Lx_j - D^{-1}Rx_{j-1} + D^{-1}b\right) + (1-\omega)\,x_{j-1}}$$

Idee: Young 1950, „$\rho(B(\omega)) \to \min$"

$$\begin{aligned} B(\omega) &= -(D+\omega L)^{-1}\{\omega R + (\omega-1)D\}, \quad d = (D+\omega L)^{-1}\omega b \\ &= -\left(\frac{1}{\omega}D + L\right)^{-1}\left\{R + D - \frac{1}{\omega}D\right\} \end{aligned}$$

Satz 8.3. *Sei A symmetrisch und positiv definit.*

$$\omega \in (0,2) \quad \rightsquigarrow \quad \rho\left(B\left(\omega\right)\right) < 1$$

Beweis. $a_{i,i} = \langle Ae_i, e_i\rangle > 0$, $A^T = A \rightsquigarrow R = L^T$.

Sei $Bz = \lambda z$, $z \in \mathbb{C}^m$, $\lambda \in \mathbb{C}$ und $|z|^2 = 1$.

Sei $d = \langle Dz, z\rangle$, $l = \langle Lz, z\rangle$ und $a = \langle Az, z\rangle > 0$

$$\rightsquigarrow \lambda = \frac{(2-\omega)\, d - \omega a + \omega\left(l - \bar{l}\right)}{(2-\omega)\, d + \omega a + \omega\left(l - \bar{l}\right)} \rightsquigarrow |\lambda| = 1$$

□

gebräuchlich: $\omega \in [1,2)$ SOR-Verfahren

8.2. **Mehrgitter-Verfahren.**

$$\boxed{A_{(h)} x_h = b_h}$$

feines Gitter h: U_h

grobes Gitter H: U_H

$$I_h^H : U_h \to U_H \qquad \text{Restriktions-Operator}$$
$$I_H^h : U_H \to U_h \qquad \text{Fortsetzungs-Operator}$$

k. Schritt eines Zwei-Gitter-Verfahrens:

x_h^k sei bekannt, bestimme x_h^{k+1}

$$\begin{aligned}
u_{h,0}^k &:= x_h^k \\
u_{h,j}^k &:= B_k u_{h,j-1}^k + d_k, \quad j = 1, \ldots, m_1 \text{ (Vorglättung)} \\
\delta_h^k &:= A_h u_{h,m_1}^k - b_h \\
v_h^k \text{ Lösung von } A_H v_H^k &= -I_h^H \delta_h^k \\
w_{h,0}^k &:= u_{h,m_1}^k + I_H^h v_H^k \quad (\text{Grobgitter} - \text{Korrektur}) \\
w_{h,j}^k &:= B_h w_{h,j-1}^k + d_k, \quad j = 1, \ldots, m_2 \\
x_h^{k+1} &:= w_{h,m_2}^k \\
x_h^{k+1} &= S_h x_h^k + t_h \\
S_h &= B_h^{m_2}\left(I - I_H^h A_H^{-1} I_h^H A_h\right) B_h^{m_1}
\end{aligned}$$

$\rho\left(S_h\right) \leq \gamma < 1$ glm. $\forall h$

Index

A-stabil, 45, 47
Abstiegsrichtung, 19
Adams-Bashforth-Verfahren, 38
Adams-Moulton-Verfahren, 39
Adams-Verfahren, 44
Armijo-Schrittweite, 17, 21
Ausgleichsproblem, 15
  linear, 16

Banach-Lemma, 3
BDF, 38, 44, 48
Bereich der quadratischen Konvergenz, 7
BFGS-Verfahren, 21
Broyden, 11

charakteristisches Polynom, 39, 50

Dämpfung, 10
Dahlquistsches Wurzelkriterium, 39
Defekt, 37
Differenzenquotienten, 23
dividierte Differenzen, 23
durchführbar, 5, 8

Eigenwertaufgaben, 50
Einbettung, 12
Einschrittverfahren, 37, 57
  Stabilitätsbereich, 45
Einzugsbereich, 7
Euler-Verfahren
  explizit, 38, 42, 45, 46
  implizit, 38, 45, 46, 48
explizit, 37
Extrapolation, 34
  Richardson, 34
Extremaleigenschaft, 29
Extremalpunkt, 19

Fehler
  global, 40
  lokal, 37
Fehlerabschätzung, 49
Fixpunkt-Iteration, 48

Gauß-Lagrange-Formel, 30
Gauß-Newton-Richtung, 17
Gauß-Newton-Verfahren
  gedämpft, 18
    Konvergenz, 18
Gauß-Transformation, 16
Gaußsche Knoten, 32
Gaußsche Quadraturformel, 30
genau, 30
Genauigkeitsgrad, algebraisch, 30
Gerschgorin-Kreise, 54
Gesamtschrittverfahren, 57
Gradientenverfahren, 19

Hessenberg-Form, 52
Homotopie, 12

implizit, 37
Integralmittelwertsatz, 4
Integration, Romberg, 34
Interpolationspolynom, 23
  Lagrange, 23
  Newton, 24

kleinste-Quadrate-Lösung, 15
kleinste-Quadrate-Problem, 15
Konsistenz, 37, 38
Konsistenznullstelle, 39
Konsistenzordnung, 38
Konvergenz
  überlinear, 7
  linear, 7
  quadratisch, 7
Konvergenzgeschwindigkeit, 7
Konvergenzsatz
  eingebettetes Newton-Verfahren, 13
  Gauß-Newton-Verfahren, 18
  Interpolationspolynom, 25
  kubsicher Spline, 29
  Mehrschrittverfahren, 40
  Newton-Verfahren, lokal, 5
  Quasi-Newton-Verfahren, lokal, 11
  Sekantenverfahren, lokal, 9

L-stabil, 48
Lösung
  global, 19
  lokal, 19
Lagrange-Multiplikatoren, 22

Mehrgitter-Verfahren, 58
Mehrschrittverfahren, 37
  A-stabil, 45, 47
  explizit, 37
  implizit, 37
  Konvergenz, 40
  L-stabil, 48
  stabil, 39
  Stabilitätsbereich, 46
Minimierungsproblem, 19
Moore-Penrose Inverse, 16

Newton-Cotes-Formeln, 32
Newton-Richtung, 19
Newton-Verfahren, 5
  eingebettet, 12
    Konvergenz, 13
  Einzugsbereich, 7
  lokale Konvergenz, 5
  modifiziert, 10
Normalgleichung, 16

orthogonale Polynome, 31

Polygonzugsverfahren, Euler, 35
Polynom
  charakteristisch, 39, 50
  orthogonal, 31
Powell-Schrittweite, 22
Projizierte Gradienten, 22

QR-Verfahren für Eigenwerte, 51
  mit Shift, 53
quadratische Zielfunktion, 20

Quadraturformel
  algebraischer Genauigkeitsgrad, 30
  Gauß, 30
  genau, 30
  Newton-Cotes, 32
  Rechteckregel, 30
  Trapezregel, 32
    zusammengesetzt, 33
  zusammengesetzt, 33
Quasi-Newton-Bedingung, 11
Quasi-Newton-Richtungen, 21
Quasi-Newton-Verfahren, 11
  lokale Konvergenz, 11

Rang-Eins-Aufdatierungsformel, 11
Rechteckregel, 30
Relaxations-Verfahren, 57
Restglied der Interpolation, 24, 27
Romberg-Integration, 34

Schrittweitenänderung, 49
Schrittweitensteuerung, 49
Schrittweitenwahl, 9, 21
  Armijo, 17, 21
  exakt, 21
  Powell, 22
  Wolfe, 22
Schur, Satz, 51
Sekantenverfahren, 8
  durchführbar, 8
  lokale Konvergenz, 9
Splinefunktion, 25
  Extremaleigenschaft, 29
  kubisch, 25
    natürlich, 27
Störungslemma, 3
Stützstellen, 23
Stützstellenpolynom, 23
stabil, 39
  A-, 45, 47
  L-, 48
Stabilitätsbereich, 45, 46
Startwert, 5
stationärer Punkt, 19
steife Differentialgleichungen, 48
Strafterme, 22
strikt diagonaldominant, 57

Trapezregel, 32, 38, 46, 48
  zusammengesetzt (groß, iteriert), 33

Vektoriteration, 50
  invers, 51
von Mises-Iteration, 50

Wielandt, 51
Wolfe-Schrittweite, 22

Zielfunktion, 19