Numerik für Informatiker und Bioinformatiker

Daniel Weiß

SS 2010

Folgende Literatur bildet die Grundlage dieser Vorlesung:

- P. Deuflhard, A. Hohmann, *Numerische Mathematik 1, Eine algorithmisch orientierte Einführung*, de Gruyter, 2002^3.
- R.W. Freund, R.H.W. Hoppe, *Stoer/Bulirsch: Numerische Mathematik 1*, Springer, 2007^{10}.
- F. Locher, *Numerische Mathematik für Informatiker*, Springer-Verlag, 1991.
- Ch. Lubich, Vorlesung zur Numerik.
- J. Schropp, Vorlesung zur Algorithmischen Mathematik.

Kapitel 1

Lineare Gleichungssysteme

Problem: Für vorgegebene $(n \times n)$-Matrix A und rechte Seite $b \in \mathbb{R}^n$ finde Lösungsvektor $x \in \mathbb{R}^n$ mit

$$Ax = b \tag{1.1}$$

oder ausführlich

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \dots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \dots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \dots & + & a_{nn}x_n & = & b_n, \end{array} \tag{1.2}$$

wobei wir mit a_{ij} das Element von A in der i-ten Zeile und j-ten Spalte bezeichnen.

Fragen: Wann hat das Problem (1.1) eine (eindeutige) Lösung? Wenn eine Lösung existiert, wie berechne ich diese?

Beispiel 1.
(i) Der einfache Fall $n = 1$:

$$ax = b$$

mit Lösung $x = \frac{b}{a}$ für $a \neq 0$.

(ii) Das gestaffelte lineare Gleichungssystem

$$\begin{array}{ccccccccc} r_{11}x_1 & + & r_{12}x_2 & + & \dots & + & r_{1n}x_n & = & c_1 \\ & & r_{22}x_2 & + & \dots & + & r_{2n}x_n & = & c_2 \\ & & & \ddots & & & \vdots & & \vdots \\ & & & & r_{n-1,n-1}x_{n-1} & + & r_{n-1,n}x_n & = & c_{n-1} \\ & & & & & & r_{nn}x_n & = & c_n \end{array} \tag{1.3}$$

lässt sich im Fall $r_{ii} \neq 0$ für $i = 1, \dots, n$ durch die so genannte Rückwärtssubstitution lösen: Lösen der letzten Gleichung ergibt:

$$x_n = \frac{c_n}{r_{nn}}.$$

Lösen der vorletzten Gleichung:

$$x_{n-1} = \left(c_{n-1} - r_{n-1,n}x_n\right)/r_{n-1,n-1}.$$

Allgemein gilt:

$$x_i = \left(c_i - \sum_{j=i+1}^{n} r_{ij}x_j\right)/r_{ii} \tag{1.4}$$

für $i = n, n-1, ..., 1$.

Satz 1. *(Existenz einer eindeutigen Lösung) Das Problem* (1.1) *besitzt genau dann eine eindeutige Lösung* x^**, wenn* A *invertierbar ist. Die Lösung ist in diesem Fall gegeben durch*

$$x^* = A^{-1}b.$$

Wiederholung: Eine Matrix quadratische A heißt invertierbar, falls A^{-1} existiert mit

$$A \cdot A^{-1} = A^{-1} \cdot A = I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Beispiel 2.

(i) Das Gleichungssystem

$$\begin{pmatrix} 1 & -3 \\ 4 & 2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

besitzt die eindeutige Lösung

$$x^* = \frac{1}{14}\begin{pmatrix} 5 \\ -3 \end{pmatrix}.$$

Die Inverse ist

$$\frac{1}{14}\begin{pmatrix} 2 & 3 \\ -4 & 1 \end{pmatrix}.$$

(ii) Das Gleichungssystem

$$\begin{pmatrix} 1 & -3 & 2 \\ 4 & 2 & 1 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix}$$

besitzt offenbar keine Lösung.

(iii) Das Gleichungssystem

$$\begin{pmatrix} 1 & -3 & 2 \\ 4 & 2 & 1 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

besitzt unendlich viele Lösungen:

$$\frac{1}{14}\begin{pmatrix} 5 \\ -3 \\ 0 \end{pmatrix} + \lambda \begin{pmatrix} 1 \\ -1 \\ -2 \end{pmatrix},\ \lambda \in \mathbb{R}.$$

1.1 Gaußsches Eliminationsverfahren

Problem: Löse $Ax = b$, wobei A invertierbar ist.

Motivation des Gaußschen Eliminationsverfahrens an einem Beispiel:

$$\underbrace{\begin{pmatrix} 1 & 4 & -1 \\ 3 & 0 & 5 \\ 2 & 2 & 1 \end{pmatrix}}_{=:A} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.$$

Erstes Ziel: Elimination der Variablen x_1 aus den unteren beiden Gleichungen. Diese "Zeilenoperationen" können durch Multiplikation des Gleichungssystems von Links mit der Matrix

$$L_1 := \begin{pmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix}$$

realisiert werden:

$$L_1 A = \begin{pmatrix} 1 & 4 & -1 \\ 0 & -12 & 8 \\ 0 & -6 & 3 \end{pmatrix}, \qquad L_1 b = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}.$$

Da die Matrix L_1 invertierbar ist, sind die Lösungen des Gleichungssystems $Ax = b$ genau die Lösungen von $L_1 Ax = L_1 b$.
Nächstes Ziel: Elimination der Variablen x_2 aus der unteren Gleichung. Diese "Zeilenoperation" kann durch Multiplikation des bereits modifizierten Gleichungssystems $L_1 Ax = L_1 b$ von Links mit der Matrix

$$L_2 := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{1}{2} & 1 \end{pmatrix}$$

realisiert werden. Wir finden insgesamt:

$$L_2 L_1 A = \underbrace{\begin{pmatrix} 1 & 4 & -1 \\ 0 & -12 & 8 \\ 0 & 0 & -1 \end{pmatrix}}_{=:R} \qquad L_2 L_1 b = \underbrace{\begin{pmatrix} 1 \\ -1 \\ \frac{3}{2} \end{pmatrix}}_{=:c}.$$

Das Gleichungssystem $Rx = c$ kann nun durch Rückwärtssubstitution gelöst werden (vgl. System (1.4)).

Idee des Gaußschen Eliminationsverfahrens

Wir wollen das Gleichungssystem (1.2) in ein gestaffeltes System der Form (1.3) umformen. Erster Schritt: Wir lassen die erste Zeile unverändert und eliminieren die Variable x_1 in den restlichen Zeilen, d.h. wir ersetzen die Zeile i durch

$$(\text{Zeile } i) - \frac{a_{i1}}{a_{11}} \cdot (\text{Zeile } 1), \qquad a_{11} \neq 0,$$

für $i = 2, ..., n$. Wegen der Invertierbarkeit von A lässt sich nach einem eventuellen Zeilentausch $a_{11} \neq 0$ immer garantieren. Das Element a_{11} heißt *Pivotelement*. Wir erhalten

$$\begin{array}{ccccccccc} a_{11}^{(1)}x_1 & + & a_{12}^{(1)}x_2 & + & \dots & + & a_{1n}^{(1)}x_n & = & b_1^{(1)} \\ & & a_{22}^{(1)}x_2 & + & \dots & + & a_{2n}^{(1)}x_n & = & b_2^{(1)} \\ & & \vdots & & & & \vdots & & \vdots \\ & & a_{n2}^{(1)}x_2 & + & \dots & + & a_{nn}^{(1)}x_n & = & b_n^{(1)} \end{array} \tag{1.5}$$

mit $b_1^{(1)} = b_1$ und $a_{1j}^{(1)} = a_{1j}$ für $j = 1, ..., n$ (die erste Zeile bleibt unverändert) und

$$a_{ij}^{(1)} = a_{ij} - \frac{a_{i1}}{a_{11}} a_{1j}$$
$$b_i^{(1)} = b_i - \frac{a_{i1}}{a_{11}} b_1$$

für $i, j = 2, ..., n$. Kennen wir nun die Lösung $(x_2, \dots, x_n)^T$ des reduzierten Systems (System (1.5) ohne die erste Gleichung), so lässt sich x_1 mit Hilfe der ersten Gleichung in (1.5) bestimmen.
Wir wenden dasselbe Verfahren auf das reduzierte System an und erhalten so rekursiv ein gestaffeltes System:

$$(\underbrace{A^{(0)}}_{:=A}, \underbrace{b^{(0)}}_{:=b}) \to (A^{(1)}, b^{(1)}) \to (A^{(2)}, b^{(2)}) \to \dots \to (\underbrace{A^{(n-1)}}_{=:R}, \underbrace{b^{(n-1)}}_{=:c})$$

Konkret gilt:

$$A^{(k)} \;=\; L_k P_k A^{(k-1)}, \qquad b^{(k)} \;=\; L_k P_k b^{(k-1)}.$$

Hier ist P_k eine *Permuationsmatrix*, welche im Fall $a_{kk}^{(k)} = 0$ (oder bei bestimmter Pivotwahl, s.u.) zwei Zeilen vertauscht und L_k die *Frobenius-Matrix*

$$L_k = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & -l_{k+1,k} & 1 & & \\ & & \vdots & & \ddots & \\ & & -l_{n,k} & & & 1 \end{pmatrix} \tag{1.6}$$

(alle unbestimmten Einträge sind 0) mit $l_{ik} := \frac{\bar{a}_{ik}^{(k)}}{\bar{a}_{kk}^{(k)}}$, wobei $\bar{a}_{ij}^{(k)}$ die Elemente von $P_k A^{(k-1)}$ sind.

Permutationsmatrix: Die Spalten einer Permutationsmatrix bestehen aus Einheitsvektoren $e_i = (0, \dots, 0, \underbrace{1}_{\text{Stelle } i}, 0, \dots, 0)^T$, wobei jeder Enheitsvektor genau einmal auftritt. z.B.

$$P \;=\; \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Daher folgt $|\det P| = 1$ und somit auch die Invertierbarkeit der Permutationsmatrizen.
Gilt $P = (e_1, \dots, e_{j-1}, \underbrace{e_l}_{\text{Stelle } j}, e_{j+1}, e_j, \dots, e_{l-1}, \underbrace{e_j}_{\text{Stelle } l}, e_{l+1}, \dots, e_n)$ so bewirkt die Multiplikation mit P von links eine Vertauschung der j-ten und l-ten Zeilen:

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_2 \\ x_1 \\ x_3 \end{pmatrix}$$

Das Gaußsche Eliminationsverfahren liefert:

Satz 2. *(über die LR-Zerlegung) Für jede invertierbare Matrix A existiert eine Permutationsmatrix P derart, dass eine Dreieckszerlegung*

$$PA = LR$$

möglich ist, wobei R eine obere Dreiecksmatrix und

$$L = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ l_{21} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ l_{n1} & \cdots & l_{n,n-1} & 1 \end{pmatrix}$$

eine untere unipotente Dreiecksmatrix ist. Eine Dreiecksmatrix heißt unipotent, falls die Elemente auf der Hauptdiagonalen alle geich 1 sind.

Wiederholung: Eine quadratische Matrix A heißt untere (obere) Dreiecksmatrix, falls

$$a_{ij} = 0 \text{ für } i < j \quad (i > j)$$

gilt.

Beweis: Das Gaußsche Eleminationsverfahren liefert

$$R \;=\; A^{(n-1)} \;=\; L_{n-1}P_{n-1}A^{(n-2)} \;=\; L_{n-1}P_{n-1}\cdot\ldots\cdot L_1P_1A.$$

Wir wollen nun die unipotenten und die Permutationsmatrizen "trennen" und fügen hierzu Identitäten der Form $I = P^{-1}P$ ein

$$\begin{aligned} R = L_{n-1}P_{n-1}L_{n-2}\underbrace{P_{n-1}^{-1}P_{n-1}}_{=I}P_{n-2}L_{n-3}\underbrace{(P_{n-1}P_{n-2})^{-1}P_{n-1}P_{n-2}}_{=I}P_{n-3}\cdot\ldots \\ \cdot L_2\underbrace{(P_{n-1}\cdot\ldots\cdot P_3)^{-1}P_{n-1}\cdot\ldots\cdot P_3}_{=I}P_2L_1\underbrace{(P_{n-1}\cdot\ldots\cdot P_2)^{-1}P_{n-1}\cdot\ldots\cdot P_2}_{=I}P_1A \\ = \hat{L}_{n-1}\hat{L}_{n-2}\cdot\ldots\cdot\hat{L}_1\underbrace{P_{n-1}\cdot\ldots\cdot P_1}_{=:P}A \end{aligned}$$

mit $\hat{L}_{n-1} := L_{n-1}$ und $\hat{L}_k := P_{n-1}\cdot\ldots\cdot P_{k+1}L_k(P_{n-1}\cdot\ldots\cdot P_{k+1})^{-1}$ für $k = n-2,\ldots,1$. Die Matrizen $\hat{L}_k$ sind wiederum Frobenius-Matrizen der Form (1.6), wobei die Einträge $\hat{l}_{jk}$ bis auf Permutation genau den l_{jk} entsprechen.

□

Algorithmus:

(i) Bestimme Matrizen P, L und R gemäß Satz 2 mit $PA = LR$ (Dreieckszerlegung)

(ii) Löse $Lc = Pb$ (Vorwärtssubstitution, vgl. Ü)

(iii) Löse $Rx = c$ (Rückwärtssubstitution)

Rechenaufwand der LR-Zerlegung:
$A \to A^{(1)}$: $n-1$ Divisionen, $(n-1)^2$ Multiplikationen und Additionen
$A \to L, R$: Also insgesamt $\sum_{j=1}^{n-1}(j^2+j) = \frac{n^3}{3} - \frac{n}{3}$ Multiplikationen und Divisionen
Hauptarbeit des Algorithmus liegt somit in der Berechnung der LR-Zerlegung.

Beachte:

$$\sum_{j=1}^{n-1} j^2 + \sum_{j=1}^{n-1} j = \frac{(n-1)n(2n-1)}{6} + \frac{n(n-1)}{2}$$
$$= \frac{n^3}{3} - \frac{n^2}{6} - \frac{2n^2}{6} + \frac{n}{6} + \frac{n^2}{2} - \frac{n}{2}.$$

Speicherplatz: Da Elemente mit Werten 0 und 1 nicht notwendigerweise gespeichert werden müssen, lässt sich das Gaußsche Eliminationsverfahren bei Speicherung der Permutationsmatrix mit $n(n+2)$ Speicherplätzen realisieren. Die relevanten Einträge der Frobenius-Matrizen können im Array der Matrix A bzw. $A^{(k)}$ gespeichert werden. Die Projektionsmatrix P kann durch weitere n Speicherplätze repräsentiert werden.

Spaltenpivotwahl: Selbst wenn $a_{11} \neq 0$ bzw. im k-ten Schritt $a_{kk}^{(k-1)} \neq 0$ gilt, kann eine Zeilenvertauschung sinnvoll sein. Bei der Spaltenpivotwahl wählt man als Pivotelement im k-ten Schritt das Element $a_{jk}^{(k-1)}$ mit

$$|a_{jk}^{k-1}| = \max_{k \leq i \leq n} |a_{i,k}^{(k-1)}|.$$

Dies führt zu

$$|l_{ij}| \leq 1 \text{ für alle } i, j$$

und somit zu einer besseren Stabilität des Verfahrens (zur Stabilität: siehe unten).

Zwei Beispiele, welche die den Rest des Kapitels motivieren:

Beispiel 3. *(zur Kondition des Problems) Betrachte das Gleichungssystem*

$$\begin{pmatrix} 1 & 1 \\ 1 & 1-\epsilon \end{pmatrix} x = \begin{pmatrix} 4 \\ 4-\epsilon \end{pmatrix}.$$

Die Lösung ist offenbar

$$x = \begin{pmatrix} 3 \\ 1 \end{pmatrix}.$$

Ersetzen wir die rechte Seite durch

$$\bar{b} = \begin{pmatrix} 4+\epsilon \\ 4-2\epsilon \end{pmatrix},$$

wobei $0 < \epsilon << 1$ *sehr klein sein kann, so erhalten wir die Lösung*

$$\bar{x} = \begin{pmatrix} 1+\epsilon \\ 3 \end{pmatrix}.$$

Das Beispiel macht deutlich, dass "kleine" Störungen der Eingabedaten zu "großen" Änderungen in der Lösung führen können. Aber wie klein ist "klein" und wie groß ist "groß"? Um den Einfluss von diesen Störungen auf die Lösung messen zu können, beschäfftigen wir uns weiter unten mit Normen.

Wichtige Frage: Wie wirken sich Störungen der Eingabegrößen (hier A und b) auf die Lösung unabhängig vom gewählten Algorithmus aus? (Kondition des Problems)

Beispiel 4. *(zur Stabilität der Gauß-Elimination) Wir lösen das Gleichungssystem*

$$\begin{pmatrix} 5\cdot 10^{-3} & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 1 \end{pmatrix}$$

in zweistelliger Gleitpunktrechnung, wobei wir als Pivotelement

a) das Element $a_{11} = 5\cdot 10^{-3}$ wählen. Nach einem Schritt des Gaußschen Eliminationsverfahrens erhalten wir das System

$$\begin{pmatrix} 5\cdot 10^{-3} & 1 \\ 0 & -200 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0.5 \\ -99 \end{pmatrix}$$

mit Lösung

$$x = \begin{pmatrix} 0 \\ 0.5 \end{pmatrix}.$$

b) das Element $a_{21} = 1$ wählen. Wir erhalten nun nach Vertauschung der Zeilen und der Gauß-Elimination

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0.5 \end{pmatrix}$$

mit Lösung

$$x = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}.$$

Beachte, dass für die exakte Lösung des Gleichungssystem gilt:

$$x = \begin{pmatrix} \frac{100}{199} \\ \frac{99}{199} \end{pmatrix} \approx \begin{pmatrix} 0.503 \\ 0.497 \end{pmatrix}.$$

Erklärung: Falls $|l_{21}|$ "groß" ist (hier $2\cdot 10^2$), gilt gemäß der Gauß-Elimination

$$\begin{aligned} a_{22}^{(1)} &= a_{22} - l_{21}a_{12} \approx -l_{21}a_{12} \\ b_2^{(1)} &= b_2 - l_{21}b_1 \quad \approx -l_{21}b_1 \end{aligned}$$

und somit auch

$$x_2 = \frac{b_2^{(1)}}{a_{22}^{(1)}} \approx \frac{b_1}{a_{12}}.$$

Bei der Berechnung von x_1 kommt es jedoch zur Stellenauslöschung (vgl. Ü):

$$x_1 = \frac{b_1 - a_{12}x_2}{a_{11}}.$$

Der Ausweg hier ist ein Zeilentausch, d.h. die Anwendung der Spaltenpivotwahl. Wir können bei dieser Wahl $|l_{21}| \leq 1$ bzw. allgemeiner $|l_{ij}| \leq 1$ für alle i, j garantieren. Tatsächlich kann aber auch bei der Gauß-Elimination mit Spaltenpivotwahl die ungünstige oben beschriebene Situation auftreten.

Wichtige Frage: Wie wirken sich Rundungsfehler, welche während der Durchführung eines bestimmten Algorithmus entstehen, auf die Berechnung der Lösung aus? (Stabilität des Algorithmus)

1.2 Einschub: Gleitpunktrechnung, Matrixnormen

1.2.1 Gleitpunktrechnung

Die Menge der im Computer darstellbaren reellen Zahlen ist offenbar endlich. Bei der heute üblichen normalisierten Gleitpunktdarstellung (engl. floating point representation) wird eine Zahl dargestellt als

$$z = a \cdot d^e,$$

wobei die Basis d eine Zweierpotenz ist (in der Regel 2,8,16) und der Exponent e eine ganze Zahl mit

$$e_{min} \leq e \leq e_{max}.$$

Die so genannte *Mantisse* a ist entweder 0 oder eine Zahl mit $d^{-1} \leq |a| < 1$ der Form

$$a = v \sum_{i=1}^{l} a_i d^{-i},$$

wobei $a_i \in \{0, \ldots, d-1\}$, $a_1 \neq 0$ und v das Vorzeichen und l die Mantissenlänge bezeichnet. Die Mantissenlänge ist für die relative Genauigkeit der Darstellung verantwortlich. Jede Zahl $x \neq 0$ mit

$$d^{e_{min}-1} \leq |x| \leq d^{e_{max}}(1 - d^{-l})$$

lässt sich nach Rundung durch eine Gleitpunktzahl $\mathrm{rd}(x)$ darstellen: Sei

$$x = v \underbrace{(0.a_1 a_2 \ldots a_l a_{l+1} \ldots)}_{=a} d^e$$

mit $a_1 \neq 0$ und $a_i \in \{0, \ldots, d-1\}$. Wir definieren

$$\mathrm{rd}(x) = v \cdot a' d^e$$

mit

$$a' := \begin{cases} 0.a_1 a_2 \ldots a_l, & \text{falls } 0 \leq a_{l+1} < \frac{d}{2} \\ 0.a_1 a_2 \ldots a_l + d^{-l}, & \text{falls } a_{l+1} \geq \frac{d}{2}. \end{cases}$$

Offensichtlich gilt

$$|x - \mathrm{rd}(x)| \leq |x - g|$$

für alle anderen durch den Computer darstellbaren Zahlen (Maschinenzahlen) g.
Für den relativen Fehler von $\mathrm{rd}(x)$ gilt:

$$\begin{aligned} \frac{|x - \mathrm{rd}(x)|}{|x|} &\leq \frac{d}{2} \frac{d^{-(l+1)}}{|a|} \\ &\leq \frac{d}{2} d^{-l} \\ &= \frac{d^{(1-l)}}{2}. \end{aligned} \tag{1.7}$$

Wir bezeichnen die Zahl $eps := \frac{d^{(1-l)}}{2}$ als die (relative) Maschinengenauigkeit.
Gleichung (1.7) ist äquivalent zu

$$\mathrm{rd}(x) = x(1+\epsilon) \qquad \text{mit } |\epsilon| \leq eps. \tag{1.8}$$

Falls $|x|$ kleiner als die betragsmäßig kleinste Maschinenzahl $d^{e_{min}-1}$ ist, spricht man von Exponentenunterlauf (engl. underflow), im Fall $|x| > d^{e_{max}}(1-d^{-l})$ von Exponentenüberlauf (engl. overflow).

Das Resultat einer arithmetischen Operation $x \pm y,\ x \cdot y, x/y$ muss keine Maschinenzahl sein selbst wenn es x und y sind. Wir definieren die so genannte Gleitpunktoperationen für zwei Maschinenzahlen x und y durch

$$\begin{aligned} x\hat{+}y &:= \mathrm{rd}(x+y) \\ x\hat{-}y &:= \mathrm{rd}(x-y) \\ x\hat{\cdot}y &:= \mathrm{rd}(x \cdot y) \\ x\hat{/}y &:= \mathrm{rd}(x/y). \end{aligned}$$

Offenbar gilt mit (1.8) ebenso

$$\begin{aligned} x\hat{+}y &= (x+y)(1+\epsilon_1) \\ x\hat{-}y &= (x-y)(1+\epsilon_2) \\ x\hat{\cdot}y &= (x \cdot y)(1+\delta_1) \\ x\hat{/}y &= (x/y)(1+\delta_2) \qquad\qquad |\epsilon_i|, |\delta_i| \leq eps. \end{aligned}$$

Bemerkung 1.
(i) Die Gleitpunkt-Realisierung von $\circ \in \{+,-,\cdot,/\}$ ist im Allgemeinen nicht assoziativ, d.h. es kommt auf die Reihenfolge der auszuführenden Operationen an.
(ii) Für zwei Maschinenzahlen x, y gilt:

$$x\hat{+}y = x, \text{ falls } |y| \leq \frac{eps}{d}|x|.$$

1.2.2 Matrixnormen

Ziel: Wir wollen Fehler und Abweichungen von Vektoren und Matrizen "messen", d.h. die "Größe" eines Vektors oder einer Matrix durch eine Zahl beschreiben.

Definition 1. *Eine Abbildung $\|.\| : V \to \mathbb{R}$, V ein Vektorraum, heißt eine Norm auf V, wenn gilt:*
(i) $\|v\| \geq 0$ *und* $(\|v\| = 0 \Leftrightarrow v = 0)$, *(positive Definitheit)*

(ii) $\|\alpha v\| = |\alpha|\|v\|$, *(Homogenität)*

(iii) $\|v_1 + v_2\| \leq \|v_1\| + \|v_2\|$ *(Dreiecksungleichung)*

für alle Vektoren $v, v_i \in V$ und $\alpha \in \mathbb{R}$.

Beispiel 5. *Wichtige Beispiele im $\mathbb{R}^n$ sind*
(i) $\|x\|_1 = \sum_{i=1}^n |x_i|$

(ii) $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$

(iii) $\|x\|_\infty = \max_{i=1,\ldots,n} |x_i|$

Jede Norm auf $\mathbb{R}^{n\times n}$ heißt Matrixnorm. Von besonderem Interesse sind Matrixnormen, die zu einer gegebenen Vektornorm passen, d.h. es gilt

$$\|Ax\| \leq \|A\|\|x\| \tag{1.9}$$

für alle $x \in \mathbb{R}^n$ und $A \in \mathbb{R}^{n\times n}$. Solche Normen sind hilfreich zur Herleitung von Abschätzungen.

Definition 2. *Sei $\|.\|$ eine beliebige Norm auf $\mathbb{R}^n$. Dann definieren wir die zugehörige Matrixnorm auf dem Raum der quadratischen $(n \times n)$-Matrizen durch*

$$\|A\| := \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|} \text{ für } A \in \mathbb{R}^{n\times n}.$$

Offenbar gilt für Matrixnormen der Definition 2 die Ungleichung (1.9), wobei $\|A\|$ die kleinste Zahl mit dieser Eigenschaft ist. Des Weiteren ist durch diese Matrixnorm tatsächlich eine Norm im Sinne von Definition 1 gegeben, d.h. es gelten die Eigenschaften (i)-(iii). Zusätzlich gilt

$$\|I\| = 1$$
$$\|A \cdot B\| \leq \|A\| \cdot \|B\|.$$

Die Abschätzung wird die *Submultiplikativität* der Matrixnorm genannt. Eine wichtige Beobachtung ist, dass die Matrixnorm aus Definition 2 von der speziellen Wahl der Norm auf $\mathbb{R}^n$ abhängt:

Satz 3. *Sei A eine quadratische $(n \times n)$-Matrix. Es gilt:*
(i) $\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^n |a_{ij}|$ *(Spaltensummennorm)*

(ii) $\|A\|_2 = \sqrt{\text{größter EW von } A^T A}$ *(Spektralnorm)*

(iii) $\|A\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^n |a_{ij}|$ *(Zeilensummennorm)*

Ende des Einschubs.

1.3 Kondition linearer Gleichungssysteme

Wir wollen nun Normen benutzen, um bei einem linearen Gleichungssystem

$$Ax = b$$

den Einfluss von Abweichungen (Störungen) der Eingabegrößen A und b auf die Lösung x abzuschätzen.

Störungen der rechten Seite: Sei $\bar{x}$ die Lösung des Systems

$$Ax = \bar{b},$$

so gilt

$$x - \bar{x} \;=\; A^{-1}b - A^{-1}\bar{b} \;=\; A^{-1}(b - \bar{b})$$

und somit die Abschätzung der absoluten Abweichung

$$\underbrace{\|x - \bar{x}\|}_{\substack{\text{absolute Abweichung von}\\ \bar{x} \text{ zu } x \text{ gemessen in der}\\ \text{Norm } \|.\|.}} \;=\; \|A^{-1}(b - \bar{b})\| \;\leq\; \|A^{-1}\|\|b - \bar{b}\|. \tag{1.10}$$

Eine weitere aussagekräftige Größe ist die relative Abweichung von $\bar{x}$ zu x. Mit Abschätzung (1.12) folgt:

$$\underbrace{\frac{\|x-\bar{x}\|}{\|x\|}}_{\substack{\text{relative Abweichung von}\\ \bar{x} \text{ zu } x \text{ gemessen in der}\\ \text{Norm } \|.\|.}} \leq \frac{\|b\|\|A^{-1}\|}{\|x\|} \underbrace{\frac{\|b-\bar{b}\|}{\|b\|}}_{\substack{\text{relative Störung der}\\ \text{rechten Seite.}}}.$$

Mit $\|b\| = \|Ax\| \leq \|A\|\|x\|$ gilt:

$$\frac{\|x-\bar{x}\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|b-\bar{b}\|}{\|b\|}. \tag{1.11}$$

Definition 3. *Wir nennen*

$$\text{cond}(A) := \|A\|\|A^{-1}\|$$

die Konditionszahl der Matrix A.

Wir wollen nun Normen benutzen, um bei einem linearen Gleichungssystem

$$Ax = b$$

den Einfluss von Abweichungen (Störungen) der Eingabegrößen A und b auf die Lösung x abzuschätzen.

Störungen der rechten Seite: Sei $\bar{x}$ die Lösung des Systems

$$A\bar{x} = \bar{b},$$

so gilt

$$x - \bar{x} \;=\; A^{-1}b - A^{-1}\bar{b} \;=\; A^{-1}(b-\bar{b})$$

und somit die Abschätzung der absoluten Abweichung

$$\underbrace{\|x-\bar{x}\|}_{\substack{\text{absolute Abweichung von}\\ \bar{x} \text{ zu } x \text{ gemessen in der}\\ \text{Norm } \|.\|.}} \;=\; \|A^{-1}(b-\bar{b})\| \;\leq\; \|A^{-1}\|\|b-\bar{b}\|. \tag{1.12}$$

Eine weitere aussagekräftige Größe ist die relative Abweichung von $\bar{x}$ zu x. Mit Abschätzung (1.12) folgt:

$$\underbrace{\frac{\|x-\bar{x}\|}{\|x\|}}_{\substack{\text{relative Abweichung von}\\ \bar{x} \text{ zu } x \text{ gemessen in der}\\ \text{Norm } \|.\|.}} \leq \frac{\|b\|\|A^{-1}\|}{\|x\|} \underbrace{\frac{\|b-\bar{b}\|}{\|b\|}}_{\substack{\text{relative Störung der}\\ \text{rechten Seite.}}}.$$

Mit $\|b\| = \|Ax\| \leq \|A\|\|x\|$ gilt:

$$\frac{\|x-\bar{x}\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|b-\bar{b}\|}{\|b\|}. \tag{1.13}$$

Definition 4. *Wir nennen*

$$\text{cond}(A) := \|A\|\|A^{-1}\|$$

die Konditionszahl der Matrix A.

Bemerkung 2. *Ungleichung* (1.13) *macht deutlich: Die Konditionszahl von A ist ein Maß der Sensitivität des relativen Fehlers gegenüber relativen Störungen der rechten Seite b. Diese Sensitivität scheint umso geringer desto kleiner* $\mathrm{cond}(A)$ *ist. Jedoch ist die Konditionszahl der Matrix A nur eine obere Schranke dieser Sensitivität und es gilt:*

$$\begin{aligned} 1 = \|I\| &= \|AA^{-1}\| \\ &\leq \|A\|\|A^{-1}\| = \mathrm{cond}(A). \end{aligned}$$

Für reelles $A = a$ *ist die Konditionszahl minimal gleich 1.*

Eigenschaften der Konditionszahl:

$$\begin{aligned} &\mathrm{cond}(A) = \mathrm{cond}(\alpha A),\ \alpha \in \mathbb{R}\backslash\{0\} \\ &\mathrm{cond}(A) = \frac{\max_{\|y\|=1} \|Ay\|}{\min_{\|z\|=1} \|Az\|} \end{aligned} \tag{1.14}$$

Mit Gleichung (1.14) lässt sich die Kondition auch für nicht quadratische Matrizen formulieren.

Beispiel 6. *Wir betrachten das Gleichungssystem aus Beispiel 3 mit*

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

Offenbar ist die Inverse gegeben durch

$$A^{-1} = -\frac{1}{\epsilon} \begin{pmatrix} 1-\epsilon & -1 \\ -1 & 1 \end{pmatrix}.$$

Für die Zeilensummennorm finden wir daher $\|A\|_\infty = 2$ *bzw.* $\|A^{-1}\|_\infty = \frac{2}{\epsilon}$ *und somit*

$$\mathrm{cond}_\infty(A) = \frac{4}{\epsilon}.$$

Für $b = (4,\ 4-\epsilon)^T$ *und der Lösung* $x = (3,\ 1)^T$ *gilt zudem*

$$\frac{\|b\|_\infty \|A^{-1}\|_\infty}{\|x\|_\infty} = \frac{8}{3\epsilon},$$

was die schlechte Konditionierung des Gleichungssystems in Beispiel 3 erklärt.

Störungen der Eingabegrößen A **und** b**:**

Satz 4. *Sei A eine invertierbare Matrix und*

$$Ax = b, \qquad \bar{A}\bar{x} = \bar{b}.$$

Seien weiter die relativen Abweichungen der Matrix $\bar{A}$ *zu A und der rechten Seite* $\bar{b}$ *zu b beschränkt:*

$$\frac{\|A-\bar{A}\|}{\|A\|} \leq \epsilon_A, \qquad \frac{\|b-\bar{b}\|}{\|b\|} \leq \epsilon_b.$$

Dann gilt die Abschätzung:

$$\frac{\|x-\bar{x}\|}{\|x\|} \leq \frac{\mathrm{cond}(A)}{1-\epsilon_A \cdot \mathrm{cond}(A)}(\epsilon_A + \epsilon_b)$$

falls $\epsilon_A \cdot \mathrm{cond}(A) < 1$.

Bemerkung 3. *Mit $\epsilon_A, \epsilon_b \leq \epsilon$ erhalten wir*

$$\frac{\|x - \bar{x}\|}{\|x\|} \leq 2\epsilon \cdot \mathrm{cond}(A) + \mathcal{O}(\epsilon^2).$$

Dabei bezeichnet $\mathcal{O}(\epsilon^2)$ eine Funktion, die selbst bei Division durch ϵ^2 im Grenzfall $\epsilon \to 0$ beschränkt bleibt.

Beweis: Offenbar gilt

$$\begin{aligned} b - \bar{b} &= Ax - \bar{A}\bar{x} \\ &= A(x - \bar{x}) + (A - \bar{A})\bar{x}. \end{aligned}$$

Nach Multiplikation mit A^{-1} erhalten wir entsprechend umgeformt

$$x - \bar{x} = A^{-1}\Big(b - \bar{b} - (A - \bar{A})\bar{x})\Big)$$

und somit die Abschätzung

$$\|x - \bar{x}\| \leq \|A^{-1}\|\Big(\underbrace{\|A - \bar{A}\|}_{\leq \epsilon_A \|A\|}\overbrace{\|\bar{x}\|}^{\leq \|x\| + \|x - \bar{x}\|} + \underbrace{\|b - \bar{b}\|}_{\epsilon_b \|b\|}\Big).$$

Mit $\|b\| \leq \|A\|\|x\|$ erhalten wir nach algebraischen Umformungen:

$$(1 - \epsilon_A \cdot \mathrm{cond}(A))\|x - \bar{x}\| \leq \mathrm{cond}(A)\|x\|(\epsilon_A + \epsilon_b).$$

□

Bemerkung 4. *Nach Satz 4 mißt die Konditionszahl die relative Störempfindlichkeit der Lösung x von $Ax = b$ gegenüber relativen Abweichungen der Matrix A und der rechten Seite b. Sie ist aber nur eine obere Schranke dieser Störempfindlichkeit. Trotzdem ist die Abschätzung des Satzes optimal im folgenden Sinn: Für vorgegebene Matrix A lassen sich $\bar{A}$ und b finden, so dass Gleichheit gilt. Da wir aber nicht immer an beliebigen rechten Seiten interessiert sind und auch nicht beliebige Störungen zulassen, ist die Abschätzung des Satzes oft zu pessimistisch.*

Beispiel 7. *(Lubich) Betrachte das Gleichungssystem*

$$\begin{pmatrix} 1 & 1 \\ 0 & 10^{-8} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}.$$

Es gilt

$$\begin{aligned} \mathrm{cond}_\infty(A) &= \|A\|_\infty \|A^{-1}\|_\infty \\ &= 2 \cdot 10^8. \qquad \text{(sehr groß)} \end{aligned}$$

Gestörtes System:

$$\begin{pmatrix} 1 + \epsilon_1 & 1 + \epsilon_2 \\ 0 & 10^{-8}(1 + \epsilon_3) \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \underbrace{\begin{pmatrix} (1 + \epsilon_4)b_1 \\ (1 + \epsilon_5)b_2 \end{pmatrix}}_{=:\bar{b}},$$

wobei $0 \leq |\epsilon_i| \leq eps$ mit der Maschinengenauigkeit eps. Wir untersuchen jetzt die Abhängigkeit der einzelnen Komponenten von den ϵ_i. Sei dazu x die Lösung des Ausgangssystems und $\bar{x}$ die des gestörten Systems.

2.Komponente: Offenbar gilt:

$$\bar{x}_2 \;=\; \underbrace{10^8 b_2}_{=x_2} \frac{1+\epsilon_5}{1+\epsilon_3} \;=\; x_2(1+\frac{\epsilon_5-\epsilon_3}{1+\epsilon_3})$$

und somit die Gleichheit

$$|x_2 - \bar{x}_2| = |x_2| \frac{|\epsilon_5 - \epsilon_3|}{|1+\epsilon_3|}.$$

Umgeformt:

$$\frac{|x_2-\bar{x}_2|}{|x_2|} = \frac{|\epsilon_5-\epsilon_3|}{|1+\epsilon_3|} \;\leq\; 2eps + \mathcal{O}(eps^2)$$

1.Komponente: Für $\bar{x}_1$ finden wir

$$\begin{aligned}
\bar{x}_1 &= \big[(1+\epsilon_4)b_1 - (1+\epsilon_2)\bar{x}_2\big]/(1+\epsilon_1) \\
&= \big[\underbrace{b_1 - x_2}_{=x_1} + \epsilon_4 b_1 - \epsilon_2 \bar{x}_2 - x_2 \frac{\epsilon_5-\epsilon_3}{1+\epsilon_3}\big]/(1+\epsilon_1) \\
&= x_1 + \big[\epsilon_4 b_1 - \epsilon_1 x_1 - \epsilon_2 \bar{x}_2 - x_2 \frac{\epsilon_5-\epsilon_3}{1+\epsilon_3}\big]/(1+\epsilon_1).
\end{aligned}$$

Mit $b_1 = x_1 + x_2$ und $\bar{x}_2 = x_2(1+\frac{\epsilon_5-\epsilon_3}{1+\epsilon_3})$ erhalten wir die Darstellung

$$\frac{|x_1-\bar{x}_1|}{|x_i|} = \frac{1}{|x_i|}\big[(\epsilon_4-\epsilon_1)x_1 + (\epsilon_4 - \epsilon_2 - \frac{\epsilon_5-\epsilon_3}{1+\epsilon_3}(1+\epsilon_2))x_2\big]/(1+\epsilon_1)$$

und somit auch die Abschätzung

$$\frac{|x_1-\bar{x}_1|}{|x_i|} \leq \big(2\frac{|x_1|}{|x_i|} + 4\frac{|x_2|}{|x_i|}\big)eps + \mathcal{O}(eps^2).$$

Insgesamt

$$\begin{aligned}
\frac{|x_1-\bar{x}_1|}{\|x\|_\infty} &\leq 6eps + \mathcal{O}(eps^2) \\
\frac{|x_2-\bar{x}_2|}{\|x\|_\infty} &\leq 2eps + \mathcal{O}(eps^2).
\end{aligned}$$

Wir betrachten allgemeiner die Situation:

$$\begin{aligned}
\bar{a}_{ij} &:= a_{ij}(1+\epsilon_{ij}), & |\epsilon_{ij}| &\leq eps, \\
\bar{b}_i &:= b_i(1+\epsilon_i), & |\epsilon_i| &\leq eps.
\end{aligned}$$

Offenbar gilt

$$\begin{aligned}
\|A - \bar{A}\|_\infty &\leq \|A\|_\infty eps, \\
\|b - \bar{b}\|_\infty &\leq \|b\|_\infty eps.
\end{aligned}$$

Betrachte nun alternativ das Gleichungssystem

$$DAx = Db \quad \text{mit } D = \text{diag}(d_1, \ldots, d_n),\ d_i \neq 0,$$

und das gestörte System

$$D\bar{A}x = D\bar{b}.$$

Seien x und $\bar{x}$ wieder die Lösungen der entsprechenden Systeme. Die Multiplikation des Systems mit einer invertierbaren Matrix von links ändert natürlich den Zusammenhang von x und $\bar{x}$ nicht, liefert aber bessere Abschätzung: Da

$$\begin{aligned} \|DA - D\bar{A}\|_\infty &\leq \|DA\|_\infty eps \\ \|Db - D\bar{b}\|_\infty &\leq \|Db\|_\infty eps \end{aligned}$$

folgt mit Satz 4

$$\begin{aligned} \frac{\|x - \bar{x}\|_\infty}{\|x\|_\infty} &\leq \frac{\text{cond}_\infty(DA)}{1 - eps \cdot \text{cond}_\infty(DA)} 2eps \\ &= 2eps \cdot \text{cond}_\infty(DA) + \mathcal{O}(eps^2). \end{aligned}$$

Wähle für obiges Beispiel konkret

$$D = \begin{pmatrix} 1 & 0 \\ 0 & 10^8 \end{pmatrix}.$$

Damit

$$DA = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

und $\text{cond}_\infty(DA) = 4$.

Beispiel 8. *(Deuflhard) Die Lösung des Gleichungssystems $Ax = b$ mit einer Diagonalmatrix*

$$A = \begin{pmatrix} 1 & 0 \\ 0 & \epsilon \end{pmatrix}$$

ist offensichtlich ein gut konditioniertes Problem, da die Gleichungen entkoppelt sind (zwei unabhängige skalare Gleichungen). Andererseits ist aber

$$\text{cond}_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty = \frac{1}{|\epsilon|}.$$

Die Konditionszahl gemessen in der Maximumsnorm $\|.\|_\infty$ wird daher beliebig groß für kleine $0 < |\epsilon| << 1$. Sie ist ein Maß der Sensitivität der Lösung gegenüber beliebigen Störungen, auch Störungen außerhalb der Hauptdiagonalen.

Matrizen mit kleiner Kondition:

(i) I, $\text{cond}(\alpha I) = 1$

(ii) Orthogonale Matrizen $U^T U = I$. Denn es gilt:

$$\begin{aligned} \|Ux\|_2^2 &= x^T U^T \cdot Ux \\ &= x^T x \;=\; \|x\|_2^2 \end{aligned}$$

und somit für die zugehörige Matrixnorm

$$\|U\|_2 = 1.$$

Da die Inverse $U^{-1} = U^T$ offenbar wieder orthogonal ist, gilt insgesamt

$$\text{cond}_2(U) = 1.$$

(iii) Spline-Interpolation (später) führt auf Matrizen

$$A = \frac{1}{h} \begin{pmatrix} 4 & 1 & & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ & & & 1 & 4 \end{pmatrix}.$$

Da $\text{cond}(A) = \text{cond}(hA)$ gilt, gehen wir ohne Einschränkung von $h = 1$ aus. Es gilt weiterhin $\|A\|_\infty = 6$. Zur Bestimmung der Inversen von A schreiben wir

$$A = 4(I + N) \text{ mit } N \;=\; \begin{pmatrix} 0 & \frac{1}{4} & & & \\ \frac{1}{4} & 0 & \frac{1}{4} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{1}{4} & 0 & \frac{1}{4} \\ & & & \frac{1}{4} & 0 \end{pmatrix}.$$

Nach dem Satz über die Neumann-Reihe gilt:

$$(I + N)^{-1} = \sum_{i=0}^{\infty} (-N)^i.$$

Denn: $(I + N) \cdot \sum_{i=0}^{\infty} (-N)^i \;=\; \sum_{i=0}^{\infty} (-N)^i - \sum_{i=0}^{\infty} (-N)^{i+1} \;=\; I$ und $\|N\|_\infty = \frac{1}{2} < 1$.

Damit folgt:

$$\begin{aligned} \|A^{-1}\|_\infty &= \frac{1}{4} \|(I + N)^{-1}\|_\infty \\ &\leq \frac{1}{4} (\|I\|_\infty + \|N\|_\infty + \|N\|_\infty^2 + \ldots) \\ &= \frac{1}{4} \sum_{i=0}^{\infty} \left(\frac{1}{2}\right)^i \;=\; \frac{1}{2} \end{aligned}$$

und für die Konditionszahl von A

$$\text{cond}_\infty(A) \leq 3.$$

Die Matrix ist also unabhängig von h und n gut konditioniert.

Matrizen mit großer Kondition:

(i) Hilbertmatrizen $A = (\frac{1}{i+j-1})_{i,j=1,\ldots,n}$ also

$$A = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & & \\ \frac{1}{3} & \frac{1}{4} & & & \\ \frac{1}{4} & & & & \\ \vdots & & & & \end{pmatrix}$$

Es gilt:

n	cond(A)
1	1
2	27
3	740
4	2300
$\vdots$	
10	$3.5{\cdot}10^{13}$

(ii) Zu Beispiel 8:

$$A = \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_n \end{pmatrix}$$

mit $\max_j |a_j| >> \min_k |a_k|$. Dann gilt:

$$\text{cond}_2 = \frac{\max_j |a_j|}{\min_k |a_k|} >> 1.$$

1.4 Stabilität der Gauß-Elimination

Bezeichne x die exakte Lösung von $Ax = b$ bzw. $\hat{x}$ die mit einem (zunächst beliebigen) Algorithmus berechnete Näherungslösung (inklusive aller Rundungsfehler).

Definition 5. *Der Algorithmus heißt numerisch stabil*

(i) im Sinne der Vorwärtsanalyse, falls

$$\frac{\|x - \hat{x}\|}{\|x\|} \leq C \cdot \text{cond}(A) \cdot eps$$

mit nicht allzu großem C gilt, d.h. der Einfluss von Rundungsfehlern während der Rechnung ist nicht viel größer als der Einfluss von Rundungsfehlern (relative Abweichung der Größenordnung eps) in den Daten.

(ii) im Sinne der Rückwärtsanalyse, falls das numerische Ergebnis $\hat{x}$ als exakte Lösung einer Gleichung $\bar{A}\hat{x} = \bar{b}$ interpretiert werden kann mit

$$\frac{\|A - \bar{A}\|}{\|A\|} \leq C \cdot eps, \qquad \frac{\|b - \bar{b}\|}{\|b\|} \leq C \cdot eps$$

mit nicht allzu großem C.

Bemerkung 5.

(i) Mit der numerischen Stabilität im Sinne der Rückwärtsanalyse folgt die Stabilität der Vorwärtsanalyse aus Satz 4:

$$\frac{\|x - \hat{x}\|}{\|x\|} \leq 2C \cdot \text{cond}(A) \cdot eps + \mathcal{O}(eps^2).$$

(ii) Für die Stabilität der Rückwärtsanalyse ist die Kenntnis der Konditionszahl von A nicht nötig.

(iii) (Deuflhard:) Die Idee der von J.H. Wilkinson eingeführten Rückwärtsanalyse besteht darin, die durch den Algorithmus verursachten Fehler auf die Eingabegröße zurückzuspielen und so als zusätzliche Eingabefehler zu interpretieren. Dazu fassen wir die fehlerbehafteten Resultate als exakte Ergebnisse zu gestörten Eingabegrößen auf.

Bezeichnungen: Im Folgenden interpretieren wir den Vergleich und den Betrag von Matrizen komponentenweise:

$$A \leq B :\Leftrightarrow \ a_{ij} \leq b_{ij} \ \ \forall_{ij}$$
$$|A| := (|a_{ij}|)_{i,j=1,\ldots,n}$$

Beispiel 9. *(Rückwärtsanalyse des Skalarprodukts)*
Das Skalarprodukt $< y, z >$, für $y, z \in \mathbb{R}^n$ lässt sich rekursiv berechnen durch

$$< y, z > = y_n z_n + < y^{n-1}, z^{n-1} >, \tag{1.15}$$

wobei $y^{n-1} := (y_1, \ldots, y_{n-1})^T$ und $z^{n-1} := (z_1, \ldots, z_{n-1})^T$.
Die Gleitpunktrealisierung des Skalarprodukts gemäß (1.15) berechnet für Gleitpunktzahlen y, z den Wert

$$< y, z >_{fl} = < \bar{y}, z >$$

für ein $\bar{y} \in \mathbb{R}^n$ mit

$$|y - \bar{y}| \leq n \cdot eps|y| + \mathcal{O}(eps^2).$$

Beweis durch Induktion: Für $n = 1$ erhalten wir

$$< y, z >_{fl} = y \hat{\cdot} z \ = \ y \cdot z(1 + \delta),$$

wobei δ mit $|\delta| \leq eps$ den relativen Fehler der Multiplikation beschreibt. Setze $\bar{y} := y(1+\delta)$. Dann gilt offenbar

$$< y, z, >_{fl} =< \bar{y}, z >$$

und

$$|y - \bar{y}| = |y \cdot \delta| = |\delta||y| \leq eps|y|.$$

Sei $n > 1$ und die Behauptung für $n-1$ bereits bewiesen. Für die Gleitpunktrealisierung der Rekursion (1.15) *gilt:*

$$\begin{aligned} < y, z >_{fl} &= y_n \hat{z}_n \hat{+} < y^{n-1}, z^{n-1} >_{fl} \\ &= \big(y_n z_n (1+\delta) + < y^{n-1}, z^{n-1} >_{fl}\big)(1+\epsilon), \end{aligned}$$

wobei diesmal δ und ϵ mit $|\epsilon|, |\delta| \leq eps$ die relativen Fehler der Multiplikation bzw. der Addition charakterisieren. Nach Induktionsvoraussetzung gilt ferner

$$< y^{n-1}, z^{n-1} >_{fl} =< c, z^{n-1} >$$

für ein $c \in \mathbb{R}^{n-1}$ mit

$$|y^{n-1} - c| \leq (n-1)eps|y^{n-1}| + \mathcal{O}(eps^2).$$

Wir setzen $\bar{y}_n := y_n(1+\delta)(1+\epsilon)$ und $\bar{y}_k := c_k(1+\epsilon)$ für $k = 1, \ldots, n-1$. Damit folgt:

$$\begin{aligned} < y, z >_{fl} &= y_n z_n (1+\delta)(1+\epsilon) + < y^{n-1}, z^{n-1} >_{fl} (1+\epsilon) \\ &= \bar{y}_n z_n + < \underbrace{c \cdot (1+\epsilon)}_{=\bar{y}^{n-1}}, z^{n-1} > \\ &=< \bar{y}, z > \end{aligned}$$

und

$$\begin{aligned} |y_n - \bar{y}_n| &\leq 2eps|y_n| + eps^2|y_n| \\ |y_k - \bar{y}_k| &\leq |y_k - c_k| + |c_k - \bar{y}_k| \\ &\leq (n-1)eps|y_k| + eps|\bar{y}_k| + \mathcal{O}(eps^2) \\ &\leq n \cdot eps|y_k| + eps|y_k - \bar{y}_k| + \mathcal{O}(eps^2). \end{aligned}$$

Somit gilt auch

$$(1 - eps)|y_k - \bar{y}_k| \leq n \cdot eps|y_k| + \mathcal{O}(eps^2)$$

also

$$\begin{aligned} |y_k - \bar{y}_k| &\leq \frac{n}{1 - eps} \cdot eps|y_k| + \mathcal{O}(eps^2) \\ &= n \cdot eps|y_k| + \mathcal{O}(eps^2) \text{ für } k = 1, \ldots, n-1. \end{aligned}$$

Insgesamt folgt

$$|y - \bar{y}| \leq n \cdot eps|y| + \mathcal{O}(eps^2).$$

Insbesondere ist das Skalarprodukt im Sinne der Rückwärtsanalyse stabil mit $C = n$.

Satz 5. *(Rückwärtsanalyse der Vorwärtssubstitution)*
Sei $L \in \mathbb{R}^{n\times n}$ eine untere Dreiecksmatrix und $b \in \mathbb{R}^n$ ein Vektor jeweils aus Gleitpunktzahlen bestehend. Die Gleitpunktrealisierung der Vorwärtssubstitution zur Lösung eines gestaffelten Gleichungssystems $Lx = b$ berechnet eine Lösung $\hat{x}$, welche exakte Lösung eines Systems $\bar{L}x = b$ ist, mit $\bar{L}$ untere Dreiecksmatrix und

$$|L - \bar{L}| \leq n \cdot eps|L| + \mathcal{O}(eps^2),$$

d.h. die Vorwärtssubstitution ist stabil im Sinne der Rückwärtsanalyse mit $C = n$.

Beweis: Wir betrachten zunächst den einfachen Fall $n = 1$, d.h. die skalare Gleichung $lx = b$. Sei $\hat{x}$ die Lösung von

$$l\hat{\cdot}\hat{x} = b.$$

Es gilt $l\hat{\cdot}\hat{x} = lx(1 + \delta)$, wobei δ mit $|\delta| \leq eps$ den relativen Fehler der Multiplikation beschreibt. Mit $\bar{l} := l(1 + \delta)$ ist somit die Behauptung des Satzes erfüllt.

Im Fall $n > 1$ ist die k-te Komponente des Lösungsvektors $x = (x_1, \ldots, x_n)^T$ bestimmt durch

$$\begin{aligned} l_{kk}x_k &= b_k - (l_{k1}x_1 + \ldots + l_{k,k-1}x_{k-1}) \\ &= b_k - < l^{k-1}, x^{k-1} >,\ k = 1, \ldots, n, \end{aligned}$$

wobei wir wieder die abkürzenden Schreibweisen $l^{k-1} := (l_{k1}, \ldots, l_{k,k-1})^T$ und $x^{k-1} := (x_1, \ldots, x_{k-1})^T$ verwendet haben. Bezeichne $\hat{x}$ die Lösung der Realisierung in Gleitpunkt-Arithmetik

$$l_{kk}\hat{\cdot}\hat{x}_k = b_k \hat{-} < l^{k-1}, \hat{x}^{k-1} >_{fl} .$$

Offenbar gilt dann auch

$$l_{kk}\hat{x}_k(1 + \delta_k) = \big(b_k - < l^{k-1}, \hat{x}^{k-1} >_{fl}\big)(1 + \epsilon_k)$$

für $k = 1, \ldots, n$, wobei δ_k und ϵ_k die relativen Fehler der Multiplikation bzw. der Addition beschreiben mit $|\epsilon_k|, |\delta_k| \leq eps$.
Nach Beispiel 9 wissen wir bereits, dass

$$< l^{k-1}, \hat{x}^{k-1} >_{fl} = < \bar{l}^{k-1}, \hat{x}^{k-1} >$$

für einen Vektor $\bar{l}^{k-1} = (\bar{l}_{k1}, \ldots, \bar{l}_{k,k-1})^T$ mit

$$|l^{k-1} - \bar{l}^{k-1}| \leq (k - 1)eps|l^{k-1}| + \mathcal{O}(eps^2).$$

Setzen wir $\bar{l}_{kk} := l_{kk}(1 + \delta_k)/(1 + \epsilon_k)$, so ist $\bar{L}$ definiert und es gilt die Behauptung des Satzes.

□

Der folgende Satz liefert eine Aussage zur Stabilität der LR-Zerlegung im Sinne der Rückwärtsanalyse.

Satz 6. *(Rückwärtsanalyse der LR-Zerlegung durch Gauß-Elimination)*
Sei $A \in \mathbb{R}^{n\times n}$ eine Matrix von Gleitpunktzahlen, die eine LR-Zerlegung besitzt. Dann berechnet das durch Gleitpunkt-Arithmetik realisierte Gaußsche Eliminationsverfahren Matrizen $\hat{L}$ und $\hat{R}$ mit:

$$|A - \hat{L}\hat{R}| \leq (n + 3)eps|\hat{L}||\hat{R}| + \mathcal{O}(eps^2). \tag{1.16}$$

Beweis: Durch Induktion: $n = 1$ ist klar. Sei $n > 1$ und die Behauptung für $n-1$ bereits gezeigt.
Sei nun A eine $(n \times n)$-Gleitpunktmatrix. Wir schreiben

$$A = \begin{pmatrix} \alpha & w^T \\ v & C \end{pmatrix}$$

mit $\alpha \in \mathbb{R},\ v, w \in \mathbb{R}^{n-1}$ und $C \in \mathbb{R}^{(n-1)\times(n-1)}$.
Die Gauß-Elimination berechnet $z = \frac{v}{\alpha}$ und damit $C^{(1)} = C - zw^T$. Seien $\hat{z}$ und $\hat{C}^{(1)}$ in der entsprechenden Gleitpunktrealisierung berechnet, d.h.

$$\begin{aligned} \hat{z} &= v \hat{/} \alpha \\ \hat{C}^{(1)} &= C \hat{-} \hat{z} \hat{\cdot} w^T. \end{aligned}$$

Dann gilt

$$\begin{aligned} \hat{z}_i &= \frac{v_i}{\alpha}(1+\delta_i) \\ \hat{c}_{ij}^{(1)} &= (c_{ij} - \hat{z}_i w_j (1+\delta_{ij}))(1+\epsilon_{ij}) \end{aligned}$$

mit $|\delta_i|, |\delta_{ij}|, |\epsilon_{ij}| \leq eps$. Damit gilt:

$$|z - \hat{z}| \leq eps|z|.$$

Weiter folgt:

$$\begin{aligned} |\hat{c}_{ij}^{(1)} - c_{ij}^{(1)}| &= |\epsilon_{ij}||c_{ij}| + |\hat{z}_i w_j \underbrace{(1+\delta_{ij})(1+\epsilon_{ij})}_{1+\delta_{ij}+\epsilon_{ij}+\mathcal{O}(eps^2)} - z_i w_j| \\ &\leq eps|c_{ij}| + 2eps|z_i w_j| + |(\hat{z}_i - z_i)w_j| + \mathcal{O}(eps^2) \\ &\leq eps|c_{ij}| + 2eps|z_i w_j| + eps|z_i||w_j| + \mathcal{O}(eps^2) \\ &\leq \big(|c_{ij}| + 3|z_i||w_j|\big)eps + \mathcal{O}(eps^2) \end{aligned}$$

bzw.

$$\begin{aligned} |\hat{C}^{(1)} - C^{(1)}| &\leq eps\big(|\underbrace{C}| + 3|z||w|^T\big) + \mathcal{O}(eps^2) \\ &= C^{(1)} + zw^T \\ &\leq eps\big(|C^{(1)}| + 4|z||w|^T\big). \end{aligned}$$

Der Algorithmus berechnet nun die LR-Zerlegung von $\hat{C}^{(1)}$. Bezeichnen $\hat{L}^{(1)}$ und $\hat{R}^{(1)}$ die durch Gleitpunkt-Arithmetik erhaltenen Matrizen. Nach Induktionsvoraussetzung gilt:

$$|\hat{C}^{(1)} - \hat{L}^{(1)}\hat{R}^{(1)}| \leq (n+2)eps|\hat{L}^{(1)}||\hat{R}^{(1)}| + \mathcal{O}(eps^2).$$

Wir wissen

$$\begin{aligned} \hat{L}\hat{R} = \begin{pmatrix} 1 & 0 \\ \hat{z} & \hat{L}^{(1)} \end{pmatrix}\begin{pmatrix} \alpha & w^T \\ 0 & \hat{R}^{(1)} \end{pmatrix} &= \begin{pmatrix} \alpha & w^T \\ \alpha\hat{z} & \hat{z}w^T + \hat{L}^{(1)}\hat{R}^{(1)} \end{pmatrix} \\ A = LR = \begin{pmatrix} 1 & 0 \\ z & L^{(1)} \end{pmatrix}\begin{pmatrix} \alpha & w^T \\ 0 & R^{(1)} \end{pmatrix} &= \begin{pmatrix} \alpha & w^T \\ \alpha z & zw^T + L^{(1)}R^{(1)} \end{pmatrix}. \end{aligned} \tag{1.17}$$

Somit

$$A - \hat{L}\hat{R} = \begin{pmatrix} 0 & 0 \\ \alpha(z-\hat{z}) & (z-\hat{z})w^T + \underbrace{L^{(1)}R^{(1)}}_{=C^{(1)}} - \hat{L}^{(1)}\hat{R}^{(1)} \end{pmatrix}.$$

Wir schreiben $C^{(1)} = C^{(1)} - \hat{C}^{(1)} + \hat{C}^{(1)}$ und erhalten mit den obigen Abschätzungen

$$|A - \hat{L}\hat{R}| \leq eps \begin{pmatrix} 0 & 0 \\ |\alpha||z| & |z||w|^T + |C^{(1)}| + 4|z||w|^T + (n+2)|\hat{L}^{(1)}||\hat{R}^{(1)}| \end{pmatrix} + \mathcal{O}(eps^2).$$

Mit

$$\begin{aligned} |C^{(1)}| &= |C^{(1)} - \hat{C}^{(1)} + \hat{C}^{(1)} - \hat{L}^{(1)}\hat{R}^{(1)} + \hat{L}^{(1)}\hat{R}^{(1)}| \\ &\leq \underbrace{|C^{(1)} - \hat{C}^{(1)}|}_{=\mathcal{O}(eps)} + \underbrace{|\hat{C}^{(1)} - \hat{L}^{(1)}\hat{R}^{(1)}|}_{=\mathcal{O}(eps)} + |\hat{L}^{(1)}\hat{R}^{(1)}| \\ &= |\hat{L}^{(1)}\hat{R}^{(1)}| + \mathcal{O}(eps) \end{aligned}$$

finden wir

$$\begin{aligned} |A - \hat{L}\hat{R}| &\leq eps \begin{pmatrix} 0 & 0 \\ |\alpha||z| & 5|z||w|^T + (n+3)|\hat{L}^{(1)}||\hat{R}^{(1)}| \end{pmatrix} + \mathcal{O}(eps^2) \\ &\leq \underbrace{(n+3)}_{\geq 5} eps \begin{pmatrix} |\alpha| & |w|^T \\ |\alpha||z| & |z||w|^T + |\hat{L}^{(1)}||\hat{R}^{(1)}| \end{pmatrix} + \mathcal{O}(eps^2). \end{aligned}$$

Investieren wir nun abschließend $|z| = |\hat{z}| + \mathcal{O}(eps)$, so erhalten wir mit (1.17) die Behauptung

$$\begin{aligned} |A - \hat{L}\hat{R}| &\leq (n+3)eps \begin{pmatrix} |\alpha| & |w|^T \\ |\alpha||\hat{z}| & |\hat{z}||w|^T + |\hat{L}^{(1)}||\hat{R}^{(1)}| \end{pmatrix} + \mathcal{O}(eps^2) \\ &\leq (n+3)eps|\hat{L}||\hat{R}| + \mathcal{O}(eps^2). \end{aligned}$$

□

Bemerkung 6. *Wichtige Frage im Zusammenhang der Stabilität: Können $|\hat{L}|$ und $|\hat{R}|$ in Abschätzung* (1.16) *groß gegenüber den Einträgen in A werden?*
Bei Spaltenpivotsuche gilt:

$$|l_{ij}| \leq 1$$

für alle $i, j = 1, \ldots, n$. Für die Elemente der Matrix $\hat{R}$ sieht die Situation jedoch nicht so gut aus. Hier gilt im Allgemeinen:

$$\max_{i,j} |\hat{r}_{ij}| \leq 2^{n-1} \cdot \max_{i,j} |a_{ij}|.$$

Diese Abschätzung ist meist zu pessimistisch kann aber auftreten. Bei zufällig gewählten Matrizen A wird

$$\max_{i,j} |\hat{r}_{ij}| \approx n \cdot \max_{i,j} |a_{ij}|$$

beobachtet.

Satz 7. *(Rückwärtsanalyse der Gauß-Elimination ohne Pivotwahl)*
Seien $A \in \mathbb{R}^{n \times n}$ eine Matrix und $b \in \mathbb{R}^n$ ein Vektor von Gleitpunktzahlen. Des Weiteren besitze A eine LR-Zerlegung und es seien $\hat{L}, \hat{R}$ wie in Satz 6. Das in Gleitpunkt-Arithmetik erhaltene Ergebnis $\hat{x}$ von $\hat{L}\hat{c} = b,\ \hat{R}x = \hat{c}$ erfüllt

$$\bar{A}\hat{x} = b$$

für eine Matrix $\bar{A}$ mit

$$|A - \bar{A}| \leq 3(n+1)eps|\hat{L}||\hat{R}| + \mathcal{O}(eps^2).$$

Beweis: Ohne Rundungsfehler wäre

$$\left.\begin{array}{rcl} A & = & LR \\ Lc & = & b \\ Rx & = & c \end{array}\right\} \Rightarrow Ax \;=\; b.$$

Statt der exakten LR-Zerlegung haben wir $\hat{L}$ und $\hat{R}$. Nach Satz 5 erhalten wir in der Gleitpunkt-Arithmetik $\hat{x}$ als Lösung von

$$\bar{\hat{L}}\hat{c} = b$$
$$\bar{\hat{R}}x = \hat{c}$$

mit

$$|\hat{L} - \bar{\hat{L}}| \leq n \cdot eps|\hat{L}| + \mathcal{O}(eps^2)$$
$$|\hat{R} - \bar{\hat{R}}| \leq n \cdot eps|\hat{R}| + \mathcal{O}(eps^2).$$

Wir setzen $\bar{A} := \bar{\hat{L}}\bar{\hat{R}}$ und erhalten somit

$$\bar{A}\hat{x} = b$$

und

$$\begin{aligned} |A - \bar{A}| &= |A - \hat{L}\hat{R} + \hat{L}\hat{R} - \bar{\hat{L}}\hat{R} + \bar{\hat{L}}\hat{R} - \bar{\hat{L}}\bar{\hat{R}}| \\ &\leq \underbrace{|A - \hat{L}\hat{R}|}_{\leq (n+3)|\hat{L}||\hat{R}|eps + \mathcal{O}(eps^2)} + |\hat{L} - \bar{\hat{L}}||\hat{R}| + \underbrace{|\bar{\hat{L}}|}_{=|\hat{L}| + \mathcal{O}(eps)} |\hat{R} - \bar{\hat{R}}| \\ &\leq 3(n+1)eps|\hat{L}||\hat{R}| + \mathcal{O}(eps^2). \end{aligned}$$

□

Satz 8. *(Rückwärtsanalyse der Gauß-Elimination mit Spaltenpivotwahl)*
Seien $A \in \mathbb{R}^{n\times n}$ eine Matrix und $b \in \mathbb{R}^n$ ein Vektor von Gleitpunktzahlen. Des Weiteren sei die Gauß-Elimination mit Spaltenpivotwahl durchführbar, d.h. $PA = LR$ für eine Permutationsmatrix P und L, R der LR-Zerlegung. Die Gauß-Elimination mit Spaltenpivotwahl für das Gleichungssystem $Ax = b$ in der Gleitpunkt-Arithmetik berechnet ein $\hat{x}$, so dass

$$\bar{A}\hat{x} = b$$

für eine Matrix $\bar{A}$ mit

$$\frac{\|A - \bar{A}\|_\infty}{\|A\|_\infty} \leq 3(n+1)n^2 \frac{\alpha_{max}}{\max_{i,j}|a_{ij}|} eps + \mathcal{O}(eps^2), \tag{1.18}$$

wobei α_{max} der größte Betrag eines Elements ist, welches im Laufe des Verfahrens in den Matrizen $A^{(1)}$ bis $A^{(n-1)}$ auftritt.

Beweis: Das Verfahren liefert in der Gleitpunkt-Arithmetik $\hat{P}, \hat{L}, \hat{R}$ und $\hat{x}$. Dann besitzt $\hat{P}A$ eine LR-Zerlegung und $\hat{L}$ und $\hat{R}$ sind die in der Gleitpunkt-Arithmetik berechneten Dreiecksmatrizen. Nach Satz 7 existiert eine Matrix $\overline{PA}$ mit

$$\overline{PA}\hat{x} = \hat{P}b$$

und

$$|\hat{P}A - \overline{PA}| \leq 3(n+1)eps|\hat{L}||\hat{R}| + \mathcal{O}(eps^2).$$

Wir definieren $\bar{A} := \hat{P}^T\overline{PA}$ und finden mit der Identität $\hat{P}^T\hat{P} = I$ die Abschätzung

$$\begin{aligned}\|A - \bar{A}\|_\infty &= \|\hat{P}^T\hat{P}A - \hat{P}^T\overline{PA}\|_\infty \\ &\leq \underbrace{\|\hat{P}^T\|_\infty}_{=1}\|\hat{P}A - \overline{PA}\|_\infty \\ &\leq 3(n+1)eps\|\hat{L}\|_\infty\|\hat{R}\|_\infty + \mathcal{O}(eps^2).\end{aligned}$$

Die Spaltenpivotwahl sorgt dafür, dass alle Komopnenten von $\hat{L}$ vom Betrag kleiner oder gleich 1 sind, d.h.

$$\|\hat{L}\|_\infty \leq n.$$

Die Norm von $\hat{R}$ können wir abschätzen durch

$$\begin{aligned}\|\hat{R}\|_\infty &\leq n \cdot \max_{i,j}|\hat{r}_{ij}| \\ &\leq n \cdot \alpha_{max}.\end{aligned}$$

Insgesamt folgt also

$$\|A - \bar{A}\|_\infty \leq 3(n+1)n^2\alpha_{max}eps + \mathcal{O}(eps^2). \tag{1.19}$$

Die Behauptung folgt nun leicht aus (1.19) und aus $\max_{i,j}|a_{ij}| \leq \|A\|_\infty$.

□

Bemerkung 7.

(i) Tatsächlich gilt (1.18) *auch mit* $3(n+1)n^2$ *ersetzt durch* $2n^3$ *(siehe Deuflhard).*

(ii) Die Stabilität der Gauß-Elimination mit Spaltenpivotwahl im Sinne der Rückwärtsanalyse wird somit durch die Größe des Faktors

$$\rho_n(A) := \frac{\alpha_{max}}{\max_{ij}|a_{ij}|}$$

bestimmt. Allgemein gilt

$$\rho_n(A) \leq 2^{n-1},$$

wobei die Schranken (in pathologischen Fällen) tatsächlich angenommen wird. Die Gauß-Elimination mit Spaltenpivotwahl ist also über die ganze Menge der invertierbaren Matrizen nicht stabil. Doch für Matrizen mit bestimmten Strukturen ist $\rho_n(A)$ wesentlich kleiner und das Verfahren stabil. Für symmetrische positiv definite Matrizen gilt zum Beispiel $\rho_n(A) = 1$.
Denn nach Satz 7 gilt im Fall einer symmetrisch positiv definiten Matrix

$$|A - \bar{A}| \leq 3(n+1)eps|\hat{L}||\hat{L}^T| + \mathcal{O}(eps^2) \tag{1.20}$$

mit

$$|A - \hat{L}\hat{L}^T| \leq (n+3)eps|\hat{L}||\hat{L}^T| + \mathcal{O}(eps^2) \;=\; \mathcal{O}(eps),$$

also $\hat{L}\hat{L}^T = A + \mathcal{O}(eps)$. *Die Matrix* $|\hat{L}|$ *kann jedoch im Verhältnis zu* $a := \max_{ij} |a_{ij}|$ *nicht groß werden. Denn*

$$a_{ii} + \mathcal{O}(eps) = \sum_{k=1}^{i} \hat{l}_{ij}^2 \;\geq\; \hat{l}_{ij}^2$$

für alle j *und daher*

$$|\hat{l}_{ij}| \leq \sqrt{a} + \mathcal{O}(eps).$$

Mit der Abschätzung $\|\hat{L}\|_\infty \leq n\sqrt{a} + \mathcal{O}(eps)$, *welche so offenbar auch für die Transponierte von* $\hat{L}$ *gilt, folgt mit Ungleichung* (1.20) *die Abschätzung*

$$\frac{\|A - \bar{A}\|_\infty}{\|A\|_\infty} \leq 3(n+1)n^2 eps + \mathcal{O}(eps^2),$$

d.h. der Nachweis für $\rho_n(A) = 1$.
Für tridiagonale Matrizen

$$A = \begin{pmatrix} * & * & & \\ * & \ddots & \ddots & \\ & \ddots & \ddots & * \\ & & * & * \end{pmatrix}$$

gilt $\rho_n(A) \leq 2$ *und für obere Hessenberg-Matrizen*

$$A = \begin{pmatrix} * & \cdots & \cdots & * \\ * & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & * & * \end{pmatrix}$$

gilt $\rho_n(A) \leq n$ *(vgl. Ü).*

1.5 Cholesky-Verfahren für symmetrische, positiv definite Matrizen

Definition 6. *Eine quadratische Matrix* $A \in \mathbb{R}^{n\times n}$ *heißt*
(i) symmetrisch, falls gilt:

$$A = A^T \qquad\qquad (a_{ij} = a_{ji} \;\; \forall i, j = 1, \ldots, n).$$

(ii) positiv definit, falls für alle Vektoren $x \in \mathbb{R}^n \backslash \{0\}$ *gilt:*

$$x^T A x > 0. \tag{1.21}$$

Lemma 1. *Sei* $A \in \mathbb{R}^{n\times n}$ *symmetrisch und positiv definit. Dann ist* A *invertierbar und die Elemente auf der Hauptdiagonalen von* A *sind positiv, d.h.* $a_{ii} > 0$ *für* $i = 1, \ldots, n$. *Des Weiteren gilt*

$$\max_{i,j=1,\ldots,n} |a_{ij}| = \max_{i=1,\ldots,n} a_{ii}, \tag{1.22}$$

d.h. der Wert des betragsmäßig größten Elements der Matrix A ist ein Element der Hauptdiagonalen.

Beweis: Wäre A nicht invertierbar, so gäbe es ein $x \neq 0$ im Kern von A, d.h. $Ax = 0$. Insbesondere wäre dann auch

$$x^T Ax = 0,$$

was im Widerspruch zu (1.21) stünde.

Die Diagonalelemente sind positiv, da nach (1.21) gilt:

$$a_{ii} = e_i^T A e_i > 0$$

für $i = 1, \ldots, n$.

Gleichung (1.22) folgt aus

$$|a_{ij}| \leq \sqrt{a_{ii} a_{jj}} \leq \frac{1}{2}(a_{ii} + a_{jj}) \text{ für } i, j = 1, \ldots, n,$$

was wiederum aus der positiven Definitheit der Matrizen $\begin{pmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{pmatrix}$ folgt. Zusätzlich haben wir investiert, dass die Determinante (Produkt der Eigenwerte) einer positiv definiten Matrix positiv ist. Für die Eigenwerte einer positiven definiten Matrix gilt nämlich (mit Eigenvektor $x \neq 0$):

$$Ax = \lambda x \Rightarrow \underbrace{x^T Ax}_{>0} = \lambda \underbrace{x^T x}_{>0}$$
$$\Rightarrow \lambda \text{ positiv.}$$

□

Satz 9. *Sei $A \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit. Dann kann die Gauß-Elimination ohne Zeilenvertauschung durchgeführt werden und die dadurch erhaltene Restmatrix ist wiederum symmetrisch und positiv definit. Für die Zerlegung $A = LR$ gilt $R = DL^T$, wobei D eine positiv definite Diagonalmatrix ist.*

Beweis: Wir schreiben

$$A = \begin{pmatrix} a_{11} & z^T \\ z & C \end{pmatrix}$$

und wählen $a_{11} > 0$ (siehe Lemma 1) als Pivotelement. Für

$$L_1 A = A^{(1)} = \begin{pmatrix} a_{11} & z^T \\ 0 & C^{(1)} \end{pmatrix}$$

gilt:
a) $C^{(1)}$ ist symmetrisch: $c_{ij}^{(1)} = a_{i+1,j+1} - \frac{a_{i+1,1}}{a_{11}} a_{1,j+1} = a_{j+1,i+1} - \frac{a_{j+1,1}}{a_{11}} a_{1,i+1} = c_{ji}^{(1)}$.

b) $C^{(1)}$ ist positiv definit: Sei $y \in \mathbb{R}^{n-1}\backslash\{0\}$. Wir werden x_1 so definieren, dass

$$y^T C^{(1)} y = \begin{pmatrix} x_1 \\ y \end{pmatrix}^T A \begin{pmatrix} x_1 \\ y \end{pmatrix} > 0, \tag{1.23}$$

gilt, wobei die Ungleichung aus der positiven Definitheit der Matrix A folgt. Aber wie ist x_1 zu definieren? Für beliebiges x_1 gilt

$$\begin{pmatrix} x_1 \\ y \end{pmatrix}^T A \begin{pmatrix} x_1 \\ y \end{pmatrix} = a_{11}x_1^2 + 2x_1 z^T y + y^T C y.$$

Für die Matrix $C^{(1)}$ finden wir gemäß der Gauß-Elimination

$$C^{(1)} = C - \frac{1}{a_{11}} z \cdot z^T \qquad\qquad (z \cdot z^T = (a_{i1}a_{j1})_{i,j=2,\ldots,n}).$$

Wir können somit die Gleichheit in (1.23) garantieren, wenn

$$-\frac{1}{a_{11}}(y^T z)^2 = a_{11}x_1^2 + 2x_1 z^T y$$

gilt. Dies ist erfüllt für $x_1 = -\frac{y^T z}{a_{11}}$.

c) Weiter gilt:

$$L_1 A L_1^T = \left(\begin{array}{c|ccc} a_{11} & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & C^{(1)} & \\ 0 & & & \end{array}\right).$$

Rekursiv folgt:

$$L_{n-1} \cdot \ldots \cdot L_1 A L_1^T \cdot \ldots \cdot L_{n-1}^T = D,$$

wobei D eine positiv definite Diagonalmatrix ist. Mit $L := (L_{n-1} \cdot \ldots \cdot L_1)^{-1}$ gilt

$$A = LDL^T$$

(beachte allgemein $(M^T)^{-1} = (M^{-1})^T$).

□

Bemerkung 8. *Eine Spalten- oder Zeilenpivotwahl sollte nicht durchgeführt werden, da sie die Struktur von A zerstört.*

Da $D = \mathrm{diag}(d_i)$ positiv definit ist, existiert $D^{\frac{1}{2}} = \mathrm{diag}(\sqrt{d_i})$ und daher die *Cholesky-Zerlegung*

$$A = \bar{L}\bar{L}^T$$

mit unterer Dreiecksmatrix $\bar{L} = LD^{\frac{1}{2}}$.

Algorithmus zur Berechnung von $\bar{L} = (l_{ij})_{i,j=1,\ldots,n}$:

$$\begin{pmatrix} l_{11} & & \\ \vdots & \ddots & \\ l_{n1} & \cdots & l_{nn} \end{pmatrix} \begin{pmatrix} l_{11} & \cdots & l_{n1} \\ & \ddots & \vdots \\ & & l_{nn} \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$$

$i = 1: \; a_{11} = l_{11}^2 \quad \Rightarrow \quad l_{11} = \sqrt{a_{11}}$
$i > 1: \; a_{i1} = l_{i1} l_{11} \quad \Rightarrow \quad l_{i1} = \frac{a_{i1}}{l_{11}}$

allgemein:
$i = k: \; a_{kk} = l_{k1}^2 + l_{k2}^2 + \ldots + l_{kk}^2 \quad \Rightarrow \quad l_{kk} = \sqrt{a_{kk} - l_{k1}^2 - \ldots - l_{k,k-1}^2}$
$i > k: \; a_{ik} = l_{i1} l_{k1} + l_{i2} l_{k2} + \ldots + l_{ik} l_{kk} \quad \Rightarrow \quad l_{ik} = \frac{a_{ik} - l_{i1} l_{k1} - \ldots - l_{i,k-1} l_{k,k-1}}{l_{kk}}$

Algorithmus:

for $k = 1, \ldots, n$ do
  $l_{kk} = \sqrt{a_{kk} - l_{k1}^2 - \ldots - l_{k,k-1}^2}$
  for $i = k+1, \ldots, n$ do
    $l_{ik} = (a_{ik} - l_{i1} l_{k1} - \ldots - l_{i,k-1} l_{k,k-1}) / l_{kk}$
  end do
end do

Rechenaufwand der Cholesky-Zerlegung:
n Wurzeln (vernachlässigbar). Multiplikationen oder Divisionen (ebenso viele Additionen):

$$\sum_{k=1}^{n} (k - 1 + \underbrace{n - k + (n-k)(k-1)}_{=(n-k)k}) = \underbrace{\sum_{k=0}^{n-1} k}_{=\frac{n(n-1)}{2}} + \sum_{k=1}^{n} k(n-k)$$

$$\sum_{k=1}^{n} k(n-k) = n^3 \frac{1}{n} \sum_{k=1}^{n} \frac{k}{n} (1 - \frac{k}{n})$$
$$\approx n^3 \int_0^1 x(1-x)dx \; = \; \frac{1}{6} n^3 \qquad \text{(Hälfte der allg. Gauß-Elimination)}$$

Gesamt-Algorithmus:
(i) Bestimme mit dem Cholesky-Verfahren $\bar{L}$
mit $A = \bar{L} \cdot \bar{L}^T$ (Cholesky-Zerlegung)

(ii) Löse $\bar{L} c = b$ (Vorwärtssubstitution)

(iii) Löse $\bar{L}^T x = c$ (Rückwärtssubstitution)

1.6 QR-Zerlegung

Zu einer gegebenen Matrix $A \in \mathbb{R}^{m \times n}$ mit $m \geq n$ konstruieren wir eine Zerlegung

$$A = QR$$

mit orthogonaler Matrix $Q \in \mathbb{R}^{m\times m}$ (d.h. $QQ^T = I$) und

$$R = \begin{pmatrix} \tilde{R} \\ 0 \end{pmatrix} \in \mathbb{R}^{m\times n},\ \tilde{R} \in \mathbb{R}^{n\times n} \text{ obere Dreicksmatrix.}$$

Eine solche Zerlegung kann z.B. mittels Householder-Transformationen konstruiert werden.
Im Fall $m = n$ nutzen wir die Zerlegung zum Lösen des linearen Gleichungssystems $Ax = b$.

Algorithmus:
(i) Bestimme Matrizen Q und R mittels Householder-Transformationen
mit $A = QR$ (QR-Zerlegung)

(ii) Löse $Qc = b$ ($Q^{-1} = Q^T$, also $c = Q^T b$)

(iii) Löse $Rx = c$ (Rückwärtssubstitution)

Dieses Vorgehen liefert einen besonders stabilen Algorithmus, benötigt aber ungefähr doppelt so viele Operationen wie die Gauß-Elimination.

Im Fall linearer Ausgleichsprobleme ($m > n$)

$$\|Ax - b\|_2 = \min$$

finden wir mit der Zerlegung und der Orthogonalität

$$\begin{aligned} \|Ax - b\|_2 &= \|Q^T(Ax - b)\|_2 \\ &= \|Rx - Q^T b\|_2 \ = \ \min, \end{aligned}$$

was sich aufgrund der Eigenschaften von R und Q leicht lösen lässt (vgl. Abschnitt 1.7).

1.7 Lineare Ausgleichsprobleme

Betrachte das überbestimmte Gleichungssystem

$$Ax = b$$

mit $b \in \mathbb{R}^m$ und $A \in \mathbb{R}^{m\times n}$, $m > n$. Ein solches Gleichungssystem besitzt im Allgemeinen keine Lösung.

Beispiel 10. *Betrachte:*

$$\begin{pmatrix} 2 & 1 \\ 1 & 4 \\ 3 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 5 \\ 2 \end{pmatrix}.$$

Die oberen beiden Gleichungen legen x_1 und x_2 fest:

$$x_1 = x_2 = 1.$$

Jedoch ist $3 \neq 2$.

Man sucht alternativ nach einem $x \in \mathbb{R}^n$ mit

$$\|Ax - b\|_2 = min.$$

Satz 10. *(Gauß) Seien $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ mit $m > n$. Der Vektor $x \in \mathbb{R}^n$ ist genau dann eine Lösung des linearen Ausgleichsproblems $\|Ax - b\|_2 = min$, falls er die so genannte Normalengleichung*

$$A^T Ax = A^T b$$

erfüllt. Insbesondere ist das lineare Ausgleichsproblem genau dann eindeutig lösbar, wenn der Rang A maximal ist, d.h. Rang$(A) = n$ gilt.

Bemerkung 9. *Ist der Rang von A maximal, so ist $A^T A$ eine symmetrische positiv definite Matrix.*

Beweis: Wir zeigen zunächst

$$\|Ax - b\|_2 \text{ minimal} \Longleftrightarrow Ax - b \text{ orthogonal auf } V := \{Ax | x \in \mathbb{R}^n\} \subset \mathbb{R}^m.$$

Mit der Definition der euklidischen Norm folgt für beliebiges y:

$$\begin{aligned}\|A(x+y) - b\|_2^2 &= (A(x+y) - b)^T (A(x+y) - b)\\ &= (Ax - b + Ay)^T (Ax - b + Ay)\\ &= (Ax-b)^T(Ax-b) + 2(Ay)^T(Ax-b) + (Ay)^T(Ay)\\ &= \|Ax-b\|_2^2 + 2(Ay)^T(Ax-b) + \|Ay\|_2^2.\end{aligned}$$

Also auch

$$\|A(x+\alpha y) - b\|_2^2 = \|Ax-b\|_2^2 + 2(Ay)^T(Ax-b)\cdot\alpha + \|Ay\|_2^2 \cdot \alpha^2.$$

für jedes $y \in \mathbb{R}^n$ und $\alpha \in \mathbb{R}$. Wir finden daher die Äquivalenz

$$\|Ax - b\|_2 \text{ minimal} \Longleftrightarrow 2(Ay)^T(Ax-b) = 0 \quad \forall y \in \mathbb{R}^n.$$

Beachte: $2(Ay)^T(Ax-b)\cdot\alpha + \|Ay\|_2^2 \cdot \alpha^2$ ist eine quadratische Funktion in α und $(Ay)^T(Ax-b)$ ist dominant für $0 < |\alpha| << 1$.

Weiter gilt offenbar

$$\begin{aligned}0 = (Ay)^T(Ax-b) = y^T(A^TAx - A^Tb) \quad \forall y \in \mathbb{R}^n\\ \Longleftrightarrow A^TAx = A^Tb.\end{aligned}$$

□

Das Gleichungssystem $A^TAx = A^Tb$ kann für Matrizen A mit maximalem Rang mit dem Cholesky-Verfahren gelöst werden. Man beachte dabei

Lemma 2. *Für eine Matrix $A \in \mathbb{R}^{m\times n}$ mit maximalem Rang $n \le m$ gilt*

$$\text{cond}_2(A^TA) = (\text{cond}_2(A))^2.$$

Beweis: Nach Gleichung (1.14) gilt für die Kondition rechteckiger Matrizen

$$\begin{aligned}(\text{cond}_2(A))^2 &= \frac{\max_{\|x\|_2=1}\|Ax\|_2^2}{\min_{\|x\|_2=1}\|Ax\|_2^2}\\ &= \frac{\max_{\|x\|_2=1} x^TA^TAx}{\min_{\|x\|_2=1} x^TA^TAx}\\ &= \frac{\text{größter EW von } A^TA}{\text{kleinster EW von } A^TA}.\end{aligned}$$

Weiter gilt

$$\begin{aligned}\mathrm{cond}_2(A^TA) &= \frac{\max_{\|x\|_2=1}\|A^TAx\|_2}{\min_{\|x\|_2=1}\|A^TAx\|_2}\\ &= \frac{\sqrt{\text{größter EW von } (A^TA)^2}}{\sqrt{\text{kleinster EW von } (A^TA)^2}}\\ &= \frac{\sqrt{(\text{größter EW von } A^TA)^2}}{\sqrt{(\text{kleinster EW von } A^TA)^2}}\end{aligned}$$

Da A^TA positiv definit ist, sind alle EWe von A^TA echt positiv also

$$\mathrm{cond}_2(A^TA) = (\mathrm{cond}_2(A))^2.$$

□

Satz 11. *(über die Kondition linearer Ausgleichsprobleme)*
Sei A eine rechteckige $m \times n$-Matrix mit maximalem Rang $n \le m$, $b \in \mathbb{R}^m$ und $x \neq 0$ die eindeutige Lösung des linearen Ausgleichsproblems

$$\|Ax - b\|_2 = \min.$$

Bezeichne ϑ den Winkel zwischen b und dem Raum V, d.h.

$$\sin(\vartheta) = \frac{\|Ax-b\|_2}{\|b\|_2}.$$

(i) Ist $\bar{x}$ Lösung des gestörten Ausgleichsproblems

$$\|Ax - \bar{b}\|_2 = \min,$$

so gilt:

$$\frac{\|x-\bar{x}\|_2}{\|x\|_2} \le \frac{\mathrm{cond}_2(A)}{\cos(\vartheta)}\frac{\|b-\bar{b}\|_2}{\|b\|_2}.$$

(ii) Ist $\bar{x}$ Lösung des gestörten Ausgleichsproblems

$$\|\bar{A}x - b\|_2 = \min,$$

so gilt:

$$\frac{\|x-\bar{x}\|_2}{\|x\|_2} \le (\mathrm{cond}_2(A) + (\mathrm{cond}_2(A))^2\tan(\vartheta))\frac{\|A-\bar{A}\|_2}{\|A\|_2}.$$

Bemerkung 10. *Ist das Residuum $r = Ax - b$ im Verhältnis zu b klein, so wird die Kondition des linearen Ausgleichsproblems durch $\mathrm{cond}_2(A)$ beschrieben, während die Kondition der Normalengleichung in etwa durch*

$$cond_2(A^TA) = (\mathrm{cond}_2(A))^2$$

beschrieben wird. In diesem Fall sollte man zur Lösung des linearen Ausgleichsproblems ein direkt auf A basierendes Verfahren verwenden. Dafür spricht ebenfalls die Anzahl von Operationen, die nötig sind um A^TA zu berechnen. Diese Anzahl ist ungefähr $\frac{1}{2}n^2m$ während für die Cholesky-Zerlegung von A^TA nur ca. $\frac{1}{6}n^3$ Operationen nötig sind.

Satz 12. *Seien $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ eine Matrix mit vollem Rang, $b \in \mathbb{R}^m$ und Q und R die Matrizen einer QR-Zerlegung von A, d.h.*

$$Q^T A = R = \begin{pmatrix} \tilde{R} \\ 0 \end{pmatrix}$$

mit invertierbarer Matrix $\tilde{R} \in \mathbb{R}^{n\times n}$.
Dann ist $x = \tilde{R}^{-1}c$ die Lösung des linearen Ausgleichsproblems $\|Ax - b\|_2 = \min$, wobei c definiert ist durch $Q^T b = \begin{pmatrix} c \\ d \end{pmatrix}$.

Beweis: Da Q orthogonal ist, folgt:

$$\begin{aligned} \|Ax - b\|_2^2 &= \|Q^T(Ax - b)\|_2^2 \\ &= \|Rx - \begin{pmatrix} c \\ d \end{pmatrix}\|_2^2 \\ &= \|\tilde{R}x - c\|_2^2 + \|d\|_2^2 \;\geq\; \|d\|_2^2. \end{aligned}$$

Für $x := \tilde{R}^{-1}c$ ist die Minimalität von $\|Ax - b\|_2^2$ und somit auch von $\|Ax - b\|_2$ gewährleistet.

□

Bemerkung 11. *Die Norm des Residuums $r = Ax - b$ ist entsprechend den Abschätzungen des Beweises genau $\|d\|_2$, d.h.*

$$\|r\|_2 = \|d\|_2.$$

Algorithmus:
(i) Bestimme Matrizen Q und R mittels Householder-Transformationen
mit $A = QR$ (QR-Zerlegung)

(ii) Berechne $Q^T b = \begin{pmatrix} c \\ d \end{pmatrix}$

(iii) Löse $\tilde{R}x = c$ (Rückwärtssubstitution)

Kapitel 2

Nichtlineare Gleichungssysteme

Problem: Für vorgegebene Abbildung $f : D \subset \mathbb{R}^n \to \mathbb{R}^n$ finde $x \in \mathbb{R}^n$ mit

$$f(x) = 0 \tag{2.1}$$

oder ausführlicher

$$\begin{aligned} f_1(x_1, \ldots, x_n) &= 0, \\ &\vdots \\ f_n(x_1, \ldots, x_n) &= 0. \end{aligned}$$

Einerseits führt die mathematische Modellierung auf Gleichungssysteme der Form (2.1), andererseits treten bei vielen Anwendungen solche Systeme als Teilprobleme auf. Während es im linearen Fall eine vollständige Lösungstheorie gibt, lässt sich Gleichung (2.1) für nichtlineares f im Allgemeinen nicht ansehen, ob sie eine Lösung besitzt.

Beispiel 11.

keine Lösung:	$f(x) = e^x$
mehrere Lösungen:	$f(x) = x^2 - a$
unendlich viele Lösungen:	$f(x) = x \sin \frac{1}{x}$

Lösungen lassen sich zudem nur in einigen speziellen Situationen explizit angeben und selbst die analytische Lösung kann unter Umständen erst nach dem Lösen eines Problems der Form (2.1) numerisch ausgewertet werden.

Beispiel 12. *Tatsächlich wird die Quadratwurzel einer positiven reellen Zahl a als Nullstelle der Nichtlinearen Gleichung*

$$x^2 - a = 0$$

interpretiert und durch ein Iterationsverfahren näherungsweise bestimmt. Auch das Lösen einer allgemeinen quadratischen Gleichung

$$x^2 + px + q = 0$$

mit analytischer Lösung

$$x_{1,2} = -\frac{p}{2} \pm \frac{1}{2}\sqrt{p^2 - 4q}$$

lässt sich numerisch nur bei Kenntnis der entsprechenden Quadratwurzel durchführen.

Im linearen Fall war es möglich die Lösung "exakt" (bis auf Rundungsfehler) z.B. mit dem Gaußschen Eliminationsverfahren zu berechnen. Für nichtlineares f werden wir uns im Allgemeinen mit einer Näherungslösung zufrieden geben müssen, welche zusätzlich zu den Rundungsfehlern mit Verfahrensfehlern (genauer Abbruchfehlern) behaftet ist.

2.1 Fixpunktiterationen

Problem: Für vorgegebene Abbildung $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$ finde $x \in \mathbb{R}^n$ mit

$$F(x) = x \tag{2.2}$$

Definition 7. *Ein Element $x^* \in D$ heißt Fixpunkt von F, falls (2.2) gilt. Im eindimensionalen Fall sind Fixpunkte genau die Stellen, wo der Graph die Winkelhalbierende (des I und III Quadranten) schneidet.*

Zusammenhang zu Nullstellengleichungen der Form (2.1): Problem (2.2) ist offenbar äquivalent zu

$$-Af(x) = 0,$$

falls die Matrix $A \in \mathbb{R}^{n\times n}$ invertierbar ist. Insbesondere ist somit die Nullstellengleichung

$$f(x) = 0$$

äquivalent zu der Fixpunktgleichung

$$F(x) = x$$

mit $F(x) := x - Af(x)$. Diese Überlegungen bleiben auch für von x abhängiges A richtig, sofern $A(x) \in \mathbb{R}^{n\times n}$ invertierbar ist.

Idee der Fixpunktiteration: Geschicktes Umformen der Nullstellengleichung (2.1) in eine Fixpunktgleichung der Form (2.2) und Berechnung der Folge $\{x_i\}_{i\in\mathbb{N}}$ ausgehend von einem Startwert x_0 gemäß der Vorschrift

$$x_{k+1} = F(x_k),$$

wobei die so definierte Folge gegen einen Fixpunkt x^* konvergiert, der auch Problem (2.1) löst.

Beispiel 13. *Die Nullstellengleichung $x^2 - 3 = 0$ besitzt genau dieselben Lösungen wie die Fixpunktgleichungen*

$$\begin{aligned} x = F_1(x) &:= x - \frac{x^2-3}{2x} \\ x = F_2(x) &:= x - \frac{x^2-3}{4}. \end{aligned}$$

Berechnung der Iterierten in double precision liefert:

F_1			F_2		
x_0	=	2	x_0	=	2
x_1	=	1.75	x_1	=	1.75
x_2	=	1.7321	x_2	=	1.734
x_3	=	1.73205081	x_3	=	1.7324
x_4	=	1.732050807568877	x_4	=	1.732092
x_5	=	1.732050807568877	x_5	=	1.732056

Beispiel 14. *Die Nullstellengleichung $2x - \tan x = 0$, $x \in]-\frac{\pi}{2}, \frac{\pi}{2}[$, besitzt genau dieselben Lösungen wie die Fixpunktgleichungen*

$$\begin{aligned} x = F_1(x) &:= \frac{1}{2}\tan x \\ x = F_2(x) &:= \arctan(2x). \end{aligned}$$

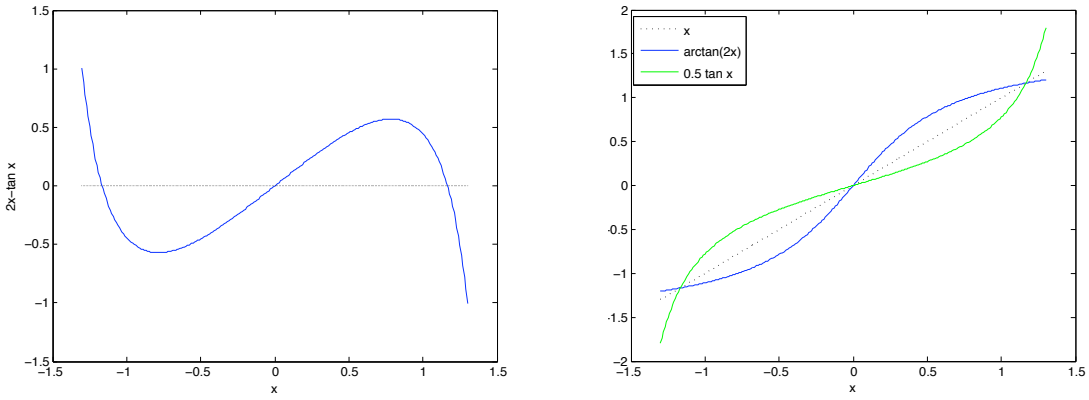


Abbildung 2.1: Links ist der Graph der Funktion $2x - \tan x$ dargestellt, rechts die Graphen von F_i.

Berechnung der Iterierten in double precision liefert:

F_1			F_1			F_2		
x_0	=	1	x_0	=	1.2	x_0	=	1.2
x_1	=	0.78	x_1	=	1.286	x_1	=	1.176
x_2	=	0.49	x_2	=	$1.708 > \pi/2$	x_2	=	1.1688
x_3	=	0.27				x_3	=	1.1666
x_4	=	0.14				x_4	=	1.1659
x_5	=	0.069				x_5	=	1.16566
x_6	=	0.035				x_6	=	1.165591
x_7	=	0.017				x_7	=	1.165571
$\vdots$						$\vdots$		
x_{27}	=	0.000000166				x_{27}	=	1.165561185207212
x_{28}	=	0.0000000083				x_{28}	=	1.165561185207212

Der Banachsche Fixpunktsatz ist einer der zentralen Sätze der angewandten Mathematik. Er liefert nicht nur die Existenz und die Eindeutigkeit eines Fixpunktes, sondern auch ein konstruktives Vorgehen und nützliche Abschätzungen. Um den Satz formulieren zu können, benötigen wir den Begriff der Kontraktion.

Definition 8. *Eine Abbildung $F : D \to D \subset \mathbb{R}$ ist eine Kontraktion auf D, falls ein $0 \leq \theta < 1$ existiert mit*

$$\|F(x) - F(y)\| \leq \theta \|x - y\|$$

für alle $x, y \in D$. Insbesondere ist also der Abstand der Bildpunkte von x und y kleiner als der Abstand von x und y selbst.

Satz 13. *(Banachscher Fixpunktsatz)*
Es sei $F : D \to D$ eine Kontraktion auf D, D abgeschlossene Teilmenge des $\mathbb{R}^n$, mit Kontraktionszahl $0 \leq \theta < 1$.
Dann gilt:
(i) Es existiert genau ein Fixpunkt x^* von F.

(ii) Die durch die Vorschrift

$$x_{k+1} = F(x_k)$$

definierte Folge konvergiert gegen x^* für jeden Startwert $x_0 \in D$.

(iii) Es gelten die Abschätzungen

$$\|x^* - x_k\| \leq \theta \|x^* - x_{k-1}\| \qquad \text{(monotone Abnahme)}$$

$$\|x^* - x_k\| \leq \frac{\theta^k}{1-\theta} \|x_0 - x_1\| \qquad \text{(a priori-Abschätzung)}$$

$$\|x^* - x_k\| \leq \frac{\theta}{1-\theta} \|x_{k-1} - x_k\| \qquad \text{(a posteriori-Abschätzung).}$$

2.2 Das Newton-Verfahren

Satz 14. *(lokal quadratische Konvergenz des Newton-Verfarens)*
Sei $D \subset \mathbb{R}^n$ offen und $f : D \to \mathbb{R}^n$ zweimal stetig differenzierbar. Es existiere ein $x^* \in D$ mit $f(x^*) = 0$. Des Weiteren sei die Jacobi-Matrix $f'(x^*)$ ausgewertet an der Nullstelle invertierbar.
Dann gibt es eine Kugel

$$K := K_\rho(x^*) \;=\; \{x \in \mathbb{R}^n | \|x - x^*\|_\infty \leq \rho\} \subset D,$$

so dass x^* die einzige Nullstelle von f in K ist. Zudem liegen die Folgeglieder

$$x_{k+1} = x_k - f'(x_k)^{-1} f(x_k)$$

für jeden Startwert $x_0 \in K$ ebenfalls in K, und es gilt

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

Weiter existiert eine Konstante $C > 0$ mit

$$\|x^* - x_{k+1}\| = C \|x^* - x_k\|^2 \tag{2.3}$$

für $k \in \mathbb{N}$.

Bemerkung 12.
(i) Formelzeile (2.3) *besagt, dass die quadratische Konvergenz vorliegt.*

(ii) Ein Problem des Newton Verfahrens ist sein möglicherweise kleiner Einzugsbereich, d.h. ρ im Satz 14 ist klein (und natürlich auch unbekannt). Startet man das Newton-Verfahren zu weit von der Nullstelle entfernt, so divergiert es oft.

Anschauung für $n = 1$: siehe Abbildung 2.2

Praktische Durchführung:

Wähle Startwert x_0
while $(\|\Delta x_k)\| > TOL$ do
    Löse $f'(x_k)\Delta x_k = -f(x_k)$ (lineares Gleichungssystem, berechne LR-Zerlegung)
    Berechne $x_{k+1} = x_k + \Delta x_k$
end do.

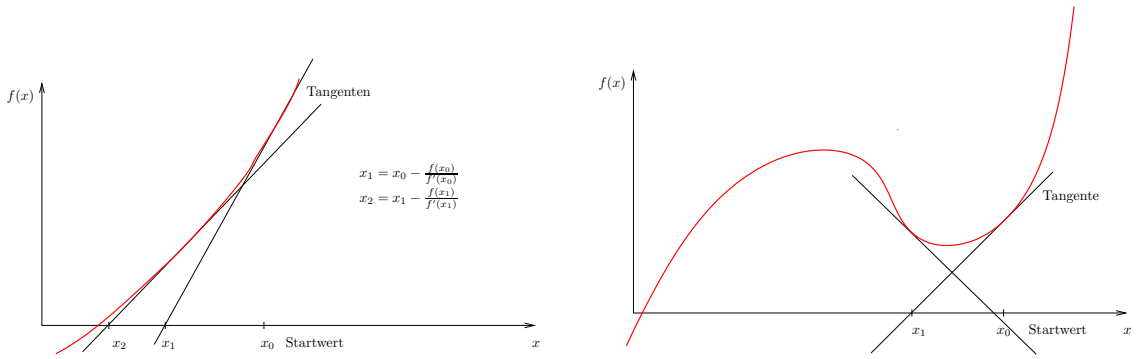


Abbildung 2.2: Links konvergiert das Newton-Verfahren, rechts lässt sich keine Konvergenz beobachten

Bemerkung 13. *Für das Abbruchkriterum sind andere Varianten möglich. Man sollte jedoch nicht* $\|f(x_k)\| \leq TOL$ *zum Abbruchkriterium machen. Denn die Ersetzung von* $f(x) = 0$ *durch* $Af(x) = 0$ *mit invertierbarer Matrix* A *ändert die exakte Lösung und die Iterierten* x_k *des Newton-Verfahrens nicht und sollte daher auch das Abbruchkriterium nicht ändern.*

Vereinfachtes Newton-Verfahren: Beim gewöhnlichen Newton-Verfahren muss pro Iteration die Ableitung von f einmal ausgewertet werden. Zudem wird pro Iteration beim Lösen des linearen Gleichungssystems eine LR-Zerlegung dieser Ableitung bestimmt. Dieses Vorgehen ist im Allgemeinen sehr teuer. Beim vereinfachten Newton-Verfahren ersetzen wir die Ableitung durch eine konstante Matrix

$$A \approx f'(x_0).$$

Es ist somit insgesamt höchstens eine Auswertung und eine Berechnung der LR-Zerlegung nötig. Wir verlieren jedoch die quadratische Konvergenz. Die Konvergenz des vereinfachten Newton-Verfahrens ist linear. Das Verfahren kann als Fixpunktiteration der Abbildung

$$F(x) = x - A^{-1} f(x)$$

aufgefasst werden.

Praktische Durchführung:

Wähle Startwert x_0 und berechne LR-Zerlegung von $A \approx f'(x_0)$
while $(\|\Delta x_k)\| > TOL$ do
    Löse $A\Delta x_k = -f(x_k)$
    Berechne $x_{k+1} = x_k + \Delta x_k$
end do.

Bemerkung 14. *Tatsächlich werden auch lineare Gleichungssysteme durch Iterationsverfahren näherungsweise gelöst.*

Kapitel 3

Interpolation und Approximation

Problem:
a) Suche für Stützpunkte $(x_0, f_0), \ldots, (x_n, f_n)$ ein Polynom $p(x)$ vom Grad $\leq n$ mit

$$p(x_i) = f_i.$$

b) Suche für eine gegebene Funktion $f : [a, b] \to \mathbb{R}$ eine möglichst einfach auszuwertende Funktion $p : [a, b] \to \mathbb{R}$ (Polynome, stückweise Polynome, trigonometrische Funktionen,...), so dass $f - p$ "klein" ist, z.B.

(i) $\int_a^b (f(x) - p(x))^2 dx = \min!$

(ii) $\max_{x \in [a,b]} |f(x) - p(x)| = \min!$

(iii) zusätzlich zu (i) oder (ii): für endlich viele Punkte $f(x) = p(x)$ (Interpolation).

Satz 15. *(Weierstraßscher Approximationssatz) Gegeben sei eine stetige Funktion $f : [a, b] \to \mathbb{R}$. Dann existiert für jedes (noch so kleine) $\epsilon > 0$ eine natürliche Zahl n und ein Polynom vom Grad n mit*

$$\max_{x \in [a,b]} |f(x) - p(x)| \leq \epsilon.$$

3.1 Polynominterpolation

Situation wie in a): Gegeben sind $n + 1$ Stützwerte $f_i := f(x_i)$ einer Funktion f an den Stützstellen $x_0 < x_1 < \ldots < x_n$.

Suche: Polynom $p(x)$ vom Grad $\leq n$ mit

$$p(x_i) = f_i \text{ für } i = 0, \ldots, n.$$

Wir sagen: Das Polynom p interpoliert f an den Stützstellen $x_0, \ldots, x_n$.

Die Polynominterpolation ist wichtig in der Theorie der numerischen Integration und bei der Realisierung von Extrapolationsverfahren (siehe Kapitel 4).

Fragen: Existiert ein solches Polynom? Gibt es mehrere Polynome die f an den Stützstellen $x_0, \ldots, x_n$ interpolieren?

Zur Eindeutigkeit: Wir wissen: Eine Gerade ist durch Vorgabe von zwei verschiedene Punkte eindeutig bestimmt. Eine Parabel ist durch Vorgabe dreier Punkte (an paarweise verschiedenen Stellen) eindeutig bestimmt.

Allgemein: Ein Polynom vom Grad $\leq n$ ist durch Vorgabe von $n+1$ Punkten (an paarweise verschiedenen Stellen) eindeutig bestimmt. Denn: Seien $p(x)$ und $q(x)$ zwei Polynome vom Grad $\leq n$ mit

$$p(x_i) - q(x_i) = 0$$

für paarweise verschiedene $x_0, \ldots, x_n$, so besitzt das Differenzpolynom $p(x) - q(x)$ einen Grad $\leq n$ und $n+1$ verschiedene Nullstellen. Daher folgt nach dem Fundamentalsatz der Algebra $p(x) - q(x) \equiv 0$.

Zur Existenz: Wir wählen die so genannten Lagrange-Polynome L_i zu den Stützstellen $x_0, \ldots, x_n$, welche den Grad n besitzen und folgende Eigenschaft haben:

$$L_i(x_j) = \begin{cases} 0, & \text{falls} \quad i \neq j \\ 1, & \text{falls} \quad i = j. \end{cases} \tag{3.1}$$

Im Grunde lösen wir also zunächst $n+1$ einfache Interpolationsprobleme der Form (3.1). Wir setzen dann

$$p(x) = \sum_{i=0}^{n} f_i L_i(x).$$

Offenbar gilt:

$$L_i(x) = \frac{\overbrace{\prod_{j=0, j\neq i}^{n} (x - x_j)}^{\text{“}Produkt der Nullstellen\text{”}}}{\underbrace{\prod_{j=0, j\neq i}^{n} (x_i - x_j)}_{\text{“Normierungsfaktor”}}}$$

Satz 16. *(Lagrangsche Interpolationsformel) Zu $n+1$ Stützpunkten (x_i, f_i), $i = 0, \ldots, n$ mit paarweisen verschiedenen Stützstellen x_i existiert genau ein Interpolationspolynom $p(x)$ vom Grad $\leq n$, welches gegeben ist durch*

$$p(x) = \sum_{i=0}^{n} f_i L_i(x), \tag{3.2}$$

wobei $L_i(x)$ die Lagrange-Polynom der Stützstellen x_i sind:

$$L_i(x) = \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j}.$$

Bemerkung 15. *Interpretieren wir $P_n := \{p \text{ Polynom} | \deg p \leq n\}$ als Vektorraum, so sind die Lagrange-Polynome L_i bezüglich des Skalarprodukts*

$$< p, q > := \sum_{i=0}^{n} p(x_i) q(x_i)$$

eine Orthonormalbasis und (3.2) *eine entsprechende Linearkombination.*

3.1.1 Kondition des Problems

Frage: Wie wirken sich Störungen der Eingabegrößen (hier die Stützwerte f_i) auf die Lösungen der Interpolation aus?

Gegeben: Stützpunkte (x_i, f_i) und gestörte Stützpunkte $(x_i, \bar{f}_i)$.

Wissen:

$$p(x) = \sum_{i=0}^{n} f_i L_i(x)$$
$$\bar{p}(x) = \sum_{i=0}^{n} \bar{f}_i L_i(x).$$

Somit folgt:

$$\begin{aligned} |p(x) - \bar{p}(x)| &\leq \sum_{i=0}^{n} |f_i - \bar{f}_i||L_i(x)| \\ &\leq \max_{i=0,\ldots,n} |f_i - \bar{f}_i| \sum_{i=0}^{n} |L_i(x)|. \end{aligned} \tag{3.3}$$

Definition 9. *Wir nennen*

$$\Lambda_n := \max_{x\in[a,b]} \sum_{i=0}^{n} |L_i(x)|$$

die Lebesgue-Konstante bezüglich der Stützstellen $x_0, \ldots, x_n$ auf dem Intervall $[a, b]$.

Es gilt also:

Satz 17. *Seien $p(x)$ und $\tilde{p}(x)$ die Interpolationspolynome zu den Stützpunkten (x_i, f_i) bzw. $(x_i, \tilde{f}_i)$, $i = 0, \ldots, n$. Dann gilt:*

$$\max_{x\in[a,b]} |p(x) - \tilde{p}(x)| \leq \Lambda_n \cdot \max_{i=0,\ldots,n} |f_i - \tilde{f}_i|,$$

wobei Λ_n die kleinste Zahl mit dieser Eigenschaft ist.

Bemerkung 16. *Die Lebesgue-Konstante ist invariant unter affinen Transformationen und daher nur von der relativen Lage der Stützstellen x_i zueinander abhängig.*

Beispiel 15. *Auf dem Intervall $[-1, 1]$ wählen wir*
a) äquidistante Stützstellen $x_i = -1 + \frac{2i}{n}$

b) Tschebyscheff-Stützstellen $x_i = \cos(\frac{2i+1}{2n+2}\pi)$

n	*a)* Λ_n	*b)* Λ_n
5	≈ 3.10	≈ 2.1
10	≈ 29.9	≈ 2.5
15	≈ 512	≈ 2.7
20	≈ 10987	≈ 2.9

Vorsicht bei Interpolationspolynomen hohen Grades!

Den folgenden beiden Abschnitten liegt die Idee zugrunde, das volle Interpolationsproblem mit $n + 1$ Stützpunkten schrittweise aus den Lösungen für weniger Stützpunkte aufzubauen.

3.1.2 Der Algorithmus von Neville-Aitken

Sind wir tatsächlich nur an einem Wert $p(x^*)$ des Interpolationspolynoms interessiert, so brauchen wir $p(x)$ nicht explizit kennen, sondern berechnen $p(x^*)$ rekursiv. Grundlage für diese rekursive Berechnung bildet das folgende Resultat.

Lemma 3. *(Aitken) Für das Interpolationspolynom $p = p(f|x_0, \ldots, x_n)$ von f in den Stützstellen $x_0, \ldots, x_n$ gilt die Rekursionsformel*

$$p(x) = \frac{(x_0 - x)p(f|x_1, \ldots, x_n) - (x_n - x)p(f|x_0, \ldots, x_{n-1})}{x_0 - x_n}, \tag{3.4}$$

wobei $p(f|x_1, \ldots, x_n)$ und $p(f|x_0, \ldots, x_{n-1})$ die Interpolationspolynome von f in den Stützstellen $x_1, \ldots, x_n$ bzw. $x_0, \ldots, x_{n-1}$ sind.

Beweis: Setze $q(x) :=$r.S. von (3.4). Dann gilt offenbar $q(x_i) = f_i$ für $i = 0, \ldots, n$ und $\deg(q) \leq n$, also $p = q$.

□

Die Interpolationspolynome für nur einen Stützpunkt sind die konstanten Polynome

$$p(f|x_i) = f_i \text{ für } i = 0, \ldots, n.$$

Mit der vereinfachenden Notation

$$T_{ik} := p(f|x_{i-k}, \ldots, x_i)(x^*),\ i \geq k$$

für ein festes x^* lässt sich der Funktionswert

$$p(x^*) = p(f|x_0, \ldots, x_n)(x^*) = T_{nn}$$

gemäß der Vorschrift

$$\begin{aligned} T_{i0} &:= f_i \text{ für } i = 0, \ldots, n \\ T_{ik} &= T_{i,k-1} + \frac{x^* - x_i}{x_i - x_{i-k}}(T_{i.k-1} - T_{i-1,k-1}) \text{ für } i \geq k \end{aligned}$$

berechnen. Diese Berechnung lässt sich durch das Schema von Neville darstellen:

$$\begin{array}{c|cccccccccc} x_0 & f_0 & = & T_{00} & & & & & & & \\ & & & & \searrow & & & & & & \\ x_1 & f_1 & = & T_{10} & \to & T_{11} & & & & & \\ & & & & \searrow & & \searrow & & & & \\ x_2 & f_2 & = & T_{20} & \to & T_{21} & \to & T_{22} & & & \\ \vdots & \vdots & & & & & & & & & \\ & & & & \searrow & & & \searrow & & & \\ x_{n-1} & f_{n-1} & = & T_{n-1,0} & \to & T_{n-1,1} & \to & \ldots & \to & T_{n-1,n-1} & \\ & & & & \searrow & & & & & & \searrow \\ x_n & f_n & = & T_{n0} & \to & T_{n1} & \to & \ldots & \to & T_{n,n-1} & \to & T_{nn} \end{array}$$

3.1.3 Newtonsche Interpolationsformel / Dividierte Differenzen

Das Verfahren von Neville ist unpraktisch, wenn man das Polynom selbst sucht oder das Polynom an mehreren Stellen auswerten will. Für diese Fälle eignet sich der Newton-Algorithmus. Wir schreiben:

$$p(f|x_0,\ldots,x_n) = a_0 + a_1(x-x_0) + a_2(x-x_0)(x-x_1) + \ldots + a_n(x-x_0)\cdot\ldots\cdot(x-x_{n-1}). \tag{3.5}$$

Beobachtungen:

(i) Darstellungen eines Polynoms $p(x)$ vom Grad $\leq n$:

a) $p(x) = c_0 + c_1 x + \ldots + c_{n-1}x^{n-1} + c_n x^n$ zur Basis der Monome $\{1, x, x^2, \ldots, x^n\}$.

b) $p(x) = b_0 L_0(x) + b_1 L_1(x) + \ldots + b_n L_n(x)$ zur Basis der Lagrange-Polynome $\{L_0(x), \ldots, L_n(x)\}$.

c) $p(x) = a_0 + a_1 w_1(x) + \ldots + a_n w_n(x)$ zur Newton-Basis $\{w_0(x), \ldots, w_n(x)\}$ mit

$$w_i(x) = \prod_{j=0}^{i-1}(x - x_j).$$

Einfache Folgerung $a_n = c_n$ und

$$p(f|x_0,\ldots,x_n) = p(f|x_0,\ldots,x_{n-1}) + a_n w_n(x), \tag{3.6}$$

da $w_n(x_i) = 0$ für $i = 0,\ldots,n-1$.

(ii) Das Polynom $p(x)$ in Darstellung (3.5) bzw. (i) c) lässt sich (wie auch die Darstellung in (i) a)) durch das so genannte Horner-Schema auswerten:

$$p(\xi) = a_0 + (\xi - x_0)\cdot\Big(a_1 + (\xi - x_1)\Big(a_2 + \ldots(\xi - x_{n-2})\Big(a_{n-1} + (\xi - x_{n-1})a_n\Big)\ldots\Big)\Big),$$

wobei die Koeffizienten a_i nacheinander aus den Beziehungen

$$\begin{aligned} f_0 &= p(x_0) = a_0 \\ f_1 &= p(x_1) = a_0 + (x_1 - x_0)a_1 \\ f_2 &= p(x_2) = a_0 + (x_2 - x_0)a_1 + (x_2 - x_0)(x_2 - x_1)a_2, \text{ usw.} \end{aligned} \tag{3.7}$$

bestimmt werden können.

Aufwand der Koeffizientenbestimmung durch (3.7):
a_1: 2 Additionen, 1 Division
a_2: 4 Additionen, 2 Multiplikationen, 1 Division
a_3: 6 Additionen, 4 Multiplikationen, 1 Division

a_i: $2i$ Additionen, $2(i-1)$ Multiplikationen, 1 Division

Insgesamt:
n Divisionen
$n(n-1)$ Multiplikationen
$n(n+1)$ Additionen

Definition 10. *Wir nennen den Koeffizienten a_n in (3.6) die n-te dividierte Differenz von f zu den Stützstellen $x_0,\ldots,x_n$, und wir schreiben*

$$f[x_0,\ldots,x_n] := a_n.$$

Wir nennen die Koeffizienten bezüglich der Newton-Basis (die a_i in obiger Beobachtung (i) c)) die dividierten Differenzen von f zu den Stützstellen $x_0, \ldots, x_n$.

Frage: Lassen sich die dividierten Differenzen billiger bestimmen als durch Darstellung (3.7)?

1. Definiere jeweils die 0-te Differenz von f zu der Stützstelle x_i durch

$$f[x_i] := f_i.$$

Wir finden mit Formel (3.4)

$$\begin{aligned} \underbrace{p(f|x_i, x_{i+1})}_{=f[x_i]+(x-x_i)f[x_i,x_{i+1}]} &= \frac{(x_i - x)p(f|x_{i+1}) - (x_{i+1} - x)p(f|x_i)}{x_i - x_{i+1}} \\ &= \frac{(x_i - x)f[x_{i+1}] - (x_{i+1} - x)f[x_i]}{x_i - x_{i+1}} \end{aligned}$$

und somit

$$\begin{aligned} f[x_i, x_{i+1}] &= \frac{(x_i - x)f[x_{i+1}] - (x_i - x)f[x_i]}{x_i - x_{i+1}} \cdot \frac{1}{x - x_i} \\ &= \frac{f[x_i] - f[x_{i+1}]}{x_i - x_{i+1}}, \end{aligned}$$

d.h. die 1-te dividierte Differenz zweier (benachbarter) Stellen lassen sich leicht aus den entsprechenden Stützwerten durch "dividierte Differenzen" berechnen.

2. Wir gehen nun davon aus, dass die $(n-1)$-ten dividierten Differenzen $f[x_1, \ldots, x_n]$ und $f[x_0, \ldots, x_{n-1}]$ bekannt sind. Wiederum mit Formel (3.4) und (3.6) finden wir

$$\begin{aligned} p(f|x_0, \ldots, x_n) &= p(f|x_0, \ldots, x_{n-1}) + f[x_0, \ldots, x_n]w_n(x) \\ &= \frac{(x_0 - x)p(f|x_1, \ldots, x_n) - (x_n - x)p(f|x_0, \ldots, x_{n-1})}{x_0 - x_n}. \end{aligned}$$

Nach Koeffizientenvergleich des Faktors x^n erhalten wir

$$\begin{aligned} f[x_0, \ldots, x_n] &= -\frac{f[x_1, \ldots, x_n] - f[x_0, \ldots, x_{n-1}]}{x_0 - x_n} \\ &= \frac{f[x_0, \ldots, x_{n-1}] - f[x_1, \ldots, x_n]}{x_0 - x_n}. \end{aligned}$$

Anordnung der dividierten Differenzen im so genannten Differenzenschema:

$$\begin{array}{lllllllllll} f_0 & = & f[x_0] & & & & & & & & \\ & & & \searrow & & & & & & & \\ f_1 & = & f[x_1] & \to & f[x_0, x_1] & & & & & & \\ & & & \searrow & & \searrow & & & & & \\ f_2 & = & f[x_2] & \to & f[x_1, x_2] & \to & f[x_0, x_1, x_2] & & & & \\ & \vdots & & & & & & & & & \\ & & & \searrow & & & & \searrow & & & \\ f_{n-1} & = & f[x_{n-1}] & \to & f[x_{n-2}, x_{n-1}] & \to & \cdots & \to & f[x_0, \ldots, x_{n-1}] & & \\ & & & \searrow & & & & & & \searrow & \\ f_n & = & f[x_n] & \to & f[x_{n-1}, x_n] & \to & \cdots & \to & f[x_1, \ldots, x_n] & \to & f[x_0, \ldots, x_n] \end{array}$$

Die Hauptdiagonale liefert die Koeffizienten von $p(f|x_0, \ldots, x_n)$.

Aufwand:
2-te Spalte: $2n$ Additionen, n Divisionen
3-te Spalte: $2(n-1)$ Additionen, $n-1$ Divisionen

Insgesamt:
$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$ Divisionen
$2\sum_{i=1}^{n} i = n(n+1)$ Additionen
Billiger als Koeffizientenbestimmung durch (3.7).

Satz 18. *(Newtonsche Interpolationsformel) Zu $n+1$ Stützpunkten (x_i, f_i), $i = 0, \ldots, n$ mit paarweise verschiedenen Stützstellen x_i existiert genau ein Interpolationspolynom $p(x)$ vom Grad $\leq n$, welches gegeben ist durch*

$$p(x) = f[x_0] + (x - x_0)f[x_0, x_1] + \ldots + (x - x_0) \cdot \ldots \cdot (x - x_{n-1})f[x_0, \ldots, x_n],$$

wobei die dividierten Differenzen gegeben sind durch

$$f[x_i] := f_i,$$
$$f[x_i, \ldots, x_{i+k}] = \frac{f[x_i, \ldots, x_{i+k-1}] - f[x_{i+1}, \ldots, x_{i+k}]}{x_i - x_{i+k}}$$

für $1 \leq k \leq n - i$.

3.1.4 Das Restglied der Polynominterpolation

Wir untersuchen nun die Approximationseigenschaft des Interpolationspolynoms $p(x)$ von f in den Stützstellen $x_0, \ldots, x_n$, d.h. den Fehler

$$f(x) - p(x).$$

Satz 19. *Sei $f : [a, b] \to \mathbb{R}$ mindestens $(n+1)$-mal stetig differenzierbar und $p(x)$ das Interpolationspolynom von f in den Stützstellen $x_0, \ldots, x_n \in [a, b]$ vom Grad $\leq n$. Dann existiert zu jedem $x \in [a, b]$ eine Zwischenstelle $\xi = \xi(x) \in (a, b)$ mit*

$$f(x) - p(x) = w_{n+1}(x) \cdot \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

Beweis: Wir setzen

$$F(x) = f(x) - p(x) - K \cdot w_{n+1}(x)$$

und bestimmen für ein $\bar{x} \neq x_i$, $i = 0, \ldots, n$, die Konstante K so, dass $F(\bar{x}) = 0$ gilt. Dies ist möglich da $w_{n+1}(\bar{x}) \neq 0$.
Insgesamt besitzt F somit $n+2$ Nullstellen und nach dem Satz von Rolle die Ableitung F' noch $n+1$ Nullstellen usw. Schließlich besitzt $F^{(n+1)} = f^{(n+1)}(x) - K(n+1)!$ eine Nullstelle $\xi = \xi(\bar{x})$. Daher gilt

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

und mit Auflösung nach K die Behauptung des Satzes.

□

Mit Darstellung (3.6) und dem vorangegangenen Beweis gilt

$$f[x_0, \ldots, x_{i-1}, \bar{x}, x_i, \ldots, x_n] = \frac{f^{(n+1)}(\xi)}{(n+1)!},$$

falls $x_{i-1} < \bar{x} < x_i$.

Betrachten wir die Funktionsklasse

$$\mathcal{F} = \{f \in C^{n+1}([a,b]) | \max_{\tau \in [a,b]} |f^{(n+1)}(\tau)| \leq M(n+1)!\}$$

für eine Konstante $M > 0$, so hängt der Approximationsfehler offenbar entscheidend von der Wahl der Stützstellen $x_0, \ldots, x_n$ in Form von

$$w_{n+1}(x) = (x - x_0) \cdot \ldots \cdot (x - x_n)$$

ab. In der Tat ist die Approximationseigenschaft von Interpolationspolynomen im Allgemeinen nicht so gut, wie der Weierstraßsche Approximationssatz Satz 15 vermuten lässt. Im nächsten Abschnitt werden wir jedoch zeigen wie sich

$$\max_{x \in [a,b]} |w_{n+1}(x)|$$

bei entsprechender Wahl der Stützstellen minimieren lässt.

3.1.5 Tschebyscheff-Interpolation

Ziel: Approximation von $f : [a,b] \to \mathbb{R}$ durch Interpolationspolynome mit möglichst "günstigen" Stützstellen (gute Kondition, optimale Approximation von $f \in \mathcal{F}$).

Ohne Einschränkung sei $[a,b] = [-1,1]$. Denn mit der affine Transformation

$$\begin{array}{rcl} [-1,1] & \longleftrightarrow & [a,b] \\ x & \longmapsto & \frac{a+b}{2} + \frac{b-a}{2} x = y \\ \frac{2}{b-a} y - \frac{a+b}{b-a} & \longleftarrow & y \end{array}$$

lässt sich das Intervall $[a,b]$ in das Intervall $[-1,1]$ überführen ohne die Interpolations- und Approximationseigenschaften zu verändern.
Wir definieren rekursiv die Tschebyscheff-Polynome

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_{n+1}(x) &= 2x \cdot T_n(x) - T_{n-1}(x). \end{aligned}$$

Das Polynom T_n vom Grad n ist ebenfalls gegeben durch:

$$T_n(x) = \cos(n \cdot \arccos x).$$

Beweis durch Induktion: $n = 0$ und $n = 1$ klar. Sei die Behauptung für n gezeigt. Mit der Definition der Teschbyscheff-Polynome gilt:

$$\begin{aligned} T_{n+1}(x) &= 2x \cdot T_n(x) - T_{n-1}(x) \\ &= 2x \cdot \cos(n \cdot \arccos x) - \cos((n-1) \arccos x) \\ &= \underbrace{2 \underbrace{\cos(\arccos x)}_{=x} \cdot \cos(n \cdot \arccos x)}_{=\cos((n+1)\varphi)+\cos((n-1)\varphi)} - \cos((n-1) \underbrace{\arccos x}_{=:\varphi}) \\ &= \cos((n+1)\varphi). \end{aligned}$$

Beachte: Nach dem Additionstheorem des Cosinus gilt:

$$\cos(n\varphi + \varphi) + \cos(n\varphi - \varphi) = 2\cos(n\varphi)\cos(\varphi).$$

Folgerungen:
(i) Die Nullstellen von T_n sind $\cos\left(\frac{2k+1}{2n}\pi\right)$, $k = 0, \ldots, n-1$.

(ii) $T_n(\cos\frac{k\pi}{n}) = (-1)^k$ für $k = 0, \ldots, n$

(iii) $|T_n(x)| \leq 1$ für $|x| \leq 1$

(iv) Der Koeffizient von x^n ist 2^{n-1}

Beispiel 16.

$$T_2(x) = 2x^2 - 1$$
$$T_3(x) = 2^2x^3 - 3x$$
$$T_4(x) = 2^3x^4 - 8x^2 + 1$$

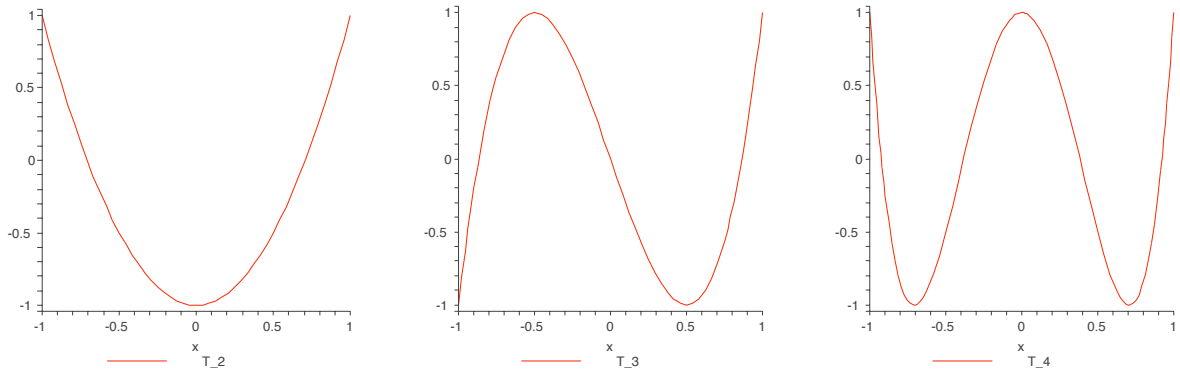


Abbildung 3.1: Tschebyscheff-Polynome T_2, T_3 und T_4.

Satz 20. *Unter allen* $(x_0, \ldots, x_n)^T \in \mathbb{R}^{n+1}$ *wird* $\max_{x\in[-1,1]} |w_{n+1}(x)|$ *minimal, wenn die* x_i *genau die Nullstellen des* $(n+1)$*-ten Tschebyscheff-Polynoms* T_{n+1} *sind, d.h. wenn*

$$x_k = \cos\left(\frac{2k+1}{2n+2}\pi\right) \text{ für } k = 0, \ldots, n$$

gilt. Der minimale Wert ist 2^{-n}.

Zum Beweis des Satzes benutzen wir folgendes Resultat:

Lemma 4. *Sei* $q(x) = 2^{n-1}x^n + \ldots$ *ein Polynom vom Grad* $\leq n$ *ungleich des* n*-ten Tschebyscheff-Polynoms* T_n. *Dann gilt:*

$$\max_{x\in[-1,1]} |q(x)| > 1 = \max_{x\in[-1,1]} |T_n(x)|.$$

Beweis: Wir nehmen an $|q(x)| \leq 1$ für alle $x \in [-1,1]$ und führen diese Annahme zu einem Widerspruch.

Es gilt nach Folgerung (ii)

$$T_n(1) = 1$$
$$T_n(\cos\frac{\pi}{n}) = -1.$$

Wir betrachten die Differenz $T_n(x) - q(x)$ auf dem Intervall $[\cos\frac{\pi}{n}, 1]$. Da nach Voraussetzung T_n und q den selben Koeffizienten vor x^n besitzen, nämlich 2^{n-1} gilt

$$\deg T_n - q \leq n - 1.$$

Nach dem Zwischenwertsatz besitzt $T_n(x)-q(x)$ mindestens eine Nullstelle in $[\cos\frac{\pi}{n}, 1]$. Beachte dazu: Entweder besitzt $T_n(x)-q(x)$ bereits eine Nullstelle in einem Randpunkt oder es gilt $T_n(\cos\frac{\pi}{n})-q(\cos\frac{\pi}{n}) < 0$ und $T_n(1) - q(1) > 0$.
Entsprechend folgt:

$$\begin{aligned}
T_n(x) - q(x) &\quad \text{hat (mindestens) eine Nullstelle in} \quad [\cos\tfrac{2}{n}\pi, \cos\tfrac{\pi}{n}] \\
T_n(x) - q(x) &\quad \text{hat (mindestens) eine Nullstelle in} \quad [\cos\tfrac{3}{n}\pi, \cos\tfrac{2}{n}\pi] \\
&\vdots \\
T_n(x) - q(x) &\quad \text{hat (mindestens) eine Nullstelle in} \quad [-1, \cos\tfrac{n-1}{n}\pi],
\end{aligned}$$

also besitzt $T_n(x) - q(x)$ insgesamt n Nullstellen in $[-1,1]$. Beachte wiederum: Fallen zwei Nullstellen in einem Randpunkt der einzelnen Intervalle zusammen, so handelt es sich um eine doppelte Nullstelle, da T_n und q dort ein Extremum besitzen.
Da aber das Polynom $T_n - q$ höchstens den Grad $n-1$ besitzt, handelt es sich um das Nullpolynom:

$$T_n = q.$$

Dies ist aber ein Widerspruch zur Annahme, die daher falsch sein muss.

□

Beweis:(zu Satz 20) Es gilt:

$$\max_{x\in[-1,1]} |w_{n+1}(x)| = \frac{1}{2^n} \max_{x\in[-1,1]} |\underbrace{2^n w_{n+1}(x)}_{=2^n x^{n+1}+\ldots}|.$$

Die Behauptung des Satzes folgt nun aus dem vorangehenden Lemma (vgl. Ü).

□

Satz 21. *Für die Lebesgue-Konstanten zu den Tschebyscheff-Stützstellen gilt:*

$$\Lambda_n \leq 3 \text{ für } n \leq 20$$
$$\Lambda_n \leq 4 \text{ für } n \leq 100$$
$$\Lambda_n \approx \frac{2}{\pi}\log n \text{ für } n \to \infty.$$

Vergleiche mit den Lebesgue-Konstanten bei äquidistanten Stützstellen!

Wir wissen, dass die Tschebyscheff-Polynome $T_0, \ldots, T_n$ eine Basis des Vektorraums P_n bilden. Sie sind bezüglich des Skalarprodukts

$$< p, q > := \sum_{i=0}^{n} p(x_i)q(x_i)$$

orthogonal, wobei x_i die Nullstellen von T_{n+1} sind. Tatsächlich gilt (ohne Beweis)

$$< T_k, T_j > = \begin{cases} 0, & \text{falls} \quad k \neq j \text{ (Orthogonalität)} \\ \frac{1}{2}(n+1), & \text{falls} \quad k = j > 0 \\ (n+1), & \text{falls} \quad k = j = 0 \end{cases}$$

für $k, j \leq n$.

Mit der Orthogonalität der Tschebyscheff-Polynome folgt

$$p = p(f|x_0, \ldots, x_n) = \sum_{i=0}^{n} \underbrace{\frac{< p, T_i >}{< T_i, T_i >}}_{=:c_i \text{ bzw. } \frac{c_0}{2} \text{ für i=0}} T_i$$

und mit der Definition des Skalarproduktes oben

$$\begin{aligned} < p, T_k > &= \sum_{i=0}^{n} p(x_i)T_k(x_i) \\ &= \sum_{i=0}^{n} f(x_i)T_k(x_i) \\ &= \sum_{i=0}^{n} f_i \cos\left(k\frac{2i+1}{2n+2}\pi\right). \end{aligned}$$

Insgesamt erhalten wir damit den folgenden Satz.

Satz 22. *(Tschebyscheffsche Interpolationsformel) Zu $n+1$ Stützpunkten (x_i, f_i), $i = 0, \ldots, n$, wobei die Stützstellen genau den Nullstellen des Tschebyscheff-Polynoms T_{n+1} entsprechen, lässt sich das eindeutige Interpolationspolynom $p(x) = p(f|x_0, \ldots, x_n)$ vom Grad $\leq n$ darstellen durch*

$$p(x) = \frac{1}{2}c_0 + c_1T_1(x) + \ldots + c_nT_n(x) \tag{3.8}$$

mit

$$c_k = \frac{2}{n+1}\sum_{i=0}^{n} f_i \cos\left(k\frac{2i+1}{2n+2}\pi\right)$$

für $k \geq 0$.

Zu den speziellen Stützstellen $x_i = \cos(\frac{2i+1}{2n+2}\pi)$ steht und somit neben der Lagrangeschen und der Newtonschen eine weitere Interpolationsformel zur Verfügung.

Fragen: Wie effizient lassen sich die Koeffizienten c_k berechnen? Lässt sich $p(x)$ in der Form (3.8) leicht auswerten?

a) Die direkte Berechnung der c_k erfordert $(n+1)^2$ Multiplikationen. Die schnelle Fourier-Transformation (FFF) benötigt $\mathcal{O}(n \log n)$ Multiplikationen. Zum Vergleich: Berechnung der dividierten Differenzen der Newtonschen Interpolationsformal benötigt $\frac{n(n+1)}{2}$ Divisionen. Für hinreichend große n ist es daher zweckmäßig die Koeffizienten mit FFT zu berechnen. Dabei ist es günstig, wenn $n+1 = 2^m$ eine 2-er Potenz ist.

b) Das Polynom $p(x)$ lässt sich bei bekannten Koeffizienten leicht berechnen:

Satz 23. *(Clenshaw-Algorithmus) Sei $p(x)$ ein Polynom mit*

$$p(x) = c_0 + c_1 T_1(x) + \ldots + c_n T_n(x).$$

Sei weiter $d_{n+2} = d_{n+1} = 0$ und

$$d_k = c_k + 2x \cdot d_{k+1} - d_{k+2} \text{ für } k = n, n-1, \ldots, 1.$$

Dann gilt:

$$p(x) = c_0 + x d_1 - d_2.$$

Bemerkung 17. *Der Aufwand zur Berechnung von $p(x)$ ist somit $n+1$ Multiplikationen und $2(n+1)$ Additionen.*

Beweis: Zunächst gilt $d_n = c_n$. Mit der Rekursionsformel

$$T_k(x) = 2x \cdot T_{k-1}(x) - T_{k-2}(x)$$

folgt:

$$\begin{aligned}
p(x) &= c_0 + c_1 T_1(x) + \ldots + c_{n-3} T_{n-3}(x) + (c_{n-2} - d_n) T_{n-2}(x) + (\underbrace{c_{n-1} + 2x d_n}_{=d_{n-1}}) T_{n-1}(x) \\
&= c_0 + c_1 T_1(x) + \ldots + c_{n-4} T_{n-4}(x) + (c_{n-3} - d_{n-1}) T_{n-3}(x) + (\underbrace{c_{n-2} - d_n + 2x d_{n-1}}_{=d_{n-2}}) T_{n-2}(x) \\
&= \ldots \text{ (induktiv)} \\
&= c_0 + (c_1 - d_3) \underbrace{T_1(x)}_{=x} + d_2 \underbrace{T_2(x)}_{=2x^2-1} \\
&= c_0 + x(c_1 + 2x d_2 - d_3) - d_2 \\
&= c_0 + x d_1 - d_2.
\end{aligned}$$

□

Allgemein ist bei Verwendung von Rekursionen wichtig, wie sich Fehler (z.B. Rundungsfehler) fortpflanzen, also die Stabilität der Rekursion. Mit anderen Worten: "Kleine" Fehler am Beginn der Rechnung, sollen keine "großen" Auswirkungen auf die spätere Rechnung haben.

Beispiel 17. *(einer instabilen Rekursion)*
Gegeben sei die Rekursion $x_{n+1} = 10x_n - 9$ mit Startwert
a) $x_1 = 1$. Dann gilt $x_n = 1$ für alle n.

b) $\tilde{x}_1 = 1 + \epsilon$. Dann gilt $\tilde{x}_n = 1 + 10^{n-1}\epsilon$ für alle n.

Satz 24. *(über die Stabilität des Clenshaw-Algorithmus) Sei $p(x)$ wie in Satz 23 und $\tilde{d}_k$ durch folgende Rekursion berechnet:*

$$0 = \tilde{d}_{n+2} = \tilde{d}_{n+1}$$
$$\tilde{d}_k = c_k 2x \cdot \tilde{d}_{k+1} - \tilde{d}_{k+2} + \epsilon_k$$

für $k = n, n, \ldots, 1$, wobei ϵ_k zum Beispiel Rundungsfehler im k-ten Schritt sind. Dann gilt:

$$|\underbrace{c_0 + x\tilde{d}_1 - \tilde{d}_2}_{=:\tilde{p}(x)} - p(x)| \leq \sum_{i=0}^{n} |\epsilon_i|$$

für $|x| \leq 1$.

Beweis: Setze $e_k := \tilde{d}_k - d_k$. Offenbar gilt

$$e_{n+2} = e_{n+1} = 0$$
$$e_k = \epsilon_k + 2x \cdot e_{k+1} - e_{k+2} \text{ für } k = n, n-1, \ldots, 1.$$

Somit nach Satz 23

$$\underbrace{\epsilon_0 + xe_1 - e_2}_{=\tilde{p}(x)-p(x)} = \epsilon_0 + \epsilon_1 T_1(x) + \ldots + \epsilon_n T_n(x).$$

Wegen $|T_j(x)| \leq 1$ für $|x| \leq 1$ folgt

$$|\tilde{p}(x) - p(x)| \leq \sum_{i=0}^{n} |\epsilon_i|.$$

□

Bemerkung 18. *Approximationen durch Summe von Tschebyscheff-Polynomen werden im Rechner zur Berechnung von Funktionen wie* $\log, \exp, \sin, \cos, \ldots$ *verwendet.*

Beispiel 18. *Berechnung von $\log(x)$ für $0 < x_{min} \leq x \leq x_{max}$, wobei x_{min} und x_{max} die kleinste bzw. die größte darstellbare Zahl im Rechner sind.*
Gleitpunktdarstellung (mit $d = 2$):

$$x = a \cdot 2^{N+1}$$

mit $a = \sum_{i=1}^{l} a_i 2^{-i}$ und $a_i \in \{0, 1\}$, $a_1 = 1$. Also existiert ein $t \in [0, 1)$ mit

$$x = (1 + t) \cdot 2^N.$$

Mit dem Additionstheorem des Logarithmus erhalten wir

$$\log x = \log(1 + t) + N \log 2.$$

Wir Approximieren $\log(1 + t)$ auf $[0, 1]$ bzw. $\log\left(1 + \frac{1+s}{2}\right)$ auf dem Intervall $[-1, 1]$ durch Tschebyscheff-Interpolation. Für den Approximationsfehler gilt nach Satz 19

$$|\log\left(1 + \frac{1+x}{2}\right) - p(x)| \leq \frac{|w_{n+1}(x)|}{(n+1)!} |\Big(\log\left(1 + \frac{1+\xi}{2}\right)\Big)^{(n+1)}|.$$

Beachte:

$$\Big(\log\Big(1+\frac{1+x}{2}\Big)\Big)' = \frac{1}{1+\frac{1+x}{2}}\frac{1}{2},$$

also auch

$$\begin{aligned}\Big(\log\Big(1+\frac{1+x}{2}\Big)\Big)^{(n+1)} &= (\frac{1}{1+\frac{1+x}{2}})^{(n)}\frac{1}{2}\\ &= \Big(\frac{1}{2}\Big)^{n+1}\frac{(-1)^n n!}{(1+\frac{1+x}{2})^{n+1}}.\end{aligned}$$

Somit gilt für den Interpolationsfehler die Abschätzung

$$\begin{aligned}|\log\Big(1+\frac{1+x}{2}\Big) - p(x)| &\le \frac{1}{2^n(n+1)!}\frac{n!}{2^{n+1}|1+\frac{1+x}{2}|^{n+1}}\\ &\le \frac{1}{2(n+1)4^n}.\end{aligned}$$

Zum Beispiel gilt für $n = 16$

$$|\log\Big(1+\frac{1+x}{2}\Big) - p(x)| \le 10^{-11}.$$

Kapitel 4

Numerische Integration

Problem: Berechne für gegebene Funktion $f : [a, b] \to \mathbb{R}$ das Riemann-Integral

$$I(f) := \int_a^b f(x)dx.$$

Oft ist nur eine numerische Näherung möglich.

Beispiel 19.
(i) Rechteckregel: Wir approximieren $I(f)$ durch das Rechteck

$$I(f) \approx (b-a)f(a).$$

(ii) Mittelpunktregel: Wir werten die Funktion im Unterschied zu (i) im Mittelpunkt $\frac{a+b}{2}$ aus:

$$I(f) \approx (b-a)f(\frac{a+b}{2}).$$

(iii) Trapezregel: Bei der Rechteck- und der Mittelpunktregel haben wir die Funktion f durch eine konstante Funktion approximiert. Bei der Trapezregel wählen wir die lineare Funktion, welche durch die Punkte $(a, f(a))$ und $(b, f(b))$ verläuft:

$$I(f) \approx (b-a)\frac{f(a)+f(b)}{2}.$$

(iv) Simpsonregel: Wir legen eine Parabel durch die drei Punkte $(a, f(a))$, $(\frac{a+b}{2}, f(\frac{a+b}{2}))$ und $(b, f(b))$ und berechnen die Fläche unter der Parabel:

$$I(f) \approx \frac{b-a}{6}\big(f(a) + 4f(\frac{a+b}{2}) + f(b)\big).$$

Bevor wir uns etwas konkreter mit numerischen Verfahren zur Berechnung einer Näherungslösung beschäftigen betrachten wir Eigenschaften des bestimmten Integrals $I(f)$ und die Kondition des Problems.

Eigenschaften des bestimmten Integrals:

(i) Das Integral $\int_a^b f(x)dx$ existiert für stückweise stetige Funktionen. Ohne Einschränkung sei f im Folgenden stetig (siehe (ii)), d.h.

$$I: \ C[a,b] \to \mathbb{R}, \quad f \mapsto I(f)$$

auf dem Raum der stetigen Funktionen auf $[a,b]$.

(ii) Für jedes $c \in [a,b]$ gilt:

$$\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx$$

d.h. das Integral ist additiv bezüglich einer Zerlegung des Integrationsintervalls.

(iii) I ist linear, d.h. für alle stetigen Funktionen f, g und alle reellen Zahlen $\lambda, \mu \in \mathbb{R}$ gilt

$$I(\lambda f + \mu g) = \lambda I(f) + \mu I(g).$$

(iv) I ist monoton, d.h. falls $f \geq g$ auf $[a,b]$, dann auch

$$\int_a^b f(x)dx \geq \int_a^b g(x)dx.$$

Aus der Monotonie folgt

$$|\int_a^b f(x)dx| \leq \int_a^b |f(x)|dx, \ \forall_{f \in C[a,b]}. \tag{4.1}$$

Tatsächlich ist diese Aussage eine Charakterisierung, d.h. eine äquivalente Definition, der Monotonie.

Um Störungen der Eingabedaten (hier $f \in C[a,b]$) messen zu können, führen wir folgende Norm ein:

$$\|f\|_1 := \int_a^b |f(x)|dx = I(|f|).$$

In dieser Norm finden wir die Kondition des Problems:

Lemma 5. *Die (relative) Kondition der Integralberechnung $\int_a^b f(x)dx$ bezüglich der Norm $\|.\|_1$ ist*

$$\text{cond}_1 = \frac{I(|f|)}{|I(f)|},$$

d.h. es gilt

$$\frac{\left|\int_a^b f(x)dx - \int_a^b \hat{f}(x)dx\right|}{\left|\int_a^b f(x)dx\right|} \leq \text{cond}_1 \frac{\|f - \hat{f}\|_1}{\|f\|_1}.$$

Beweis: Folgt unmittelbar aus der Linearität und der Monotonie des Integrals:

$$\left|\int_a^b f(x)dx - \int_a^b \hat{f}(x)dx\right| \leq \int_a^b |f(x) - \hat{f}(x)|dx.$$

□

Das Problem ist somit schlecht konditioniert, wenn das Integral über den Betrag der Funktion im Verhältnis zum Betrag des Integrals sehr groß ist. Interpretieren wir das Integral als unendliche Summe, so wird die Analogie zur Auslöschung bei der Addition deutlich. Insbesondere bei stark oszillierenden Integranden, wo sich "die Flächen gegenseitig auslöschen" ist die Kondition des Problems schlecht. Solche Integranden treten in zahlreichen Anwendungen auf.

4.1 Quadratur-Formel

Die allgemeine Form einer Quadratur-Formel ist gegeben durch:

$$\int_a^b f(x)dx \approx (b-a)\overbrace{\sum_{i=1}^{s} b_i f(\underbrace{a+c_i(b-a)}_{\text{Stützstelle}})}^{\text{gewichtetes Mittel der Funktionswerte an den Stützstellen}}.$$

Dabei bezeichnen wir die b_i als Gewichte und die c_i als die Knoten der Quadraturformel. Tatsächlich ist eine Quadraturformel durch die Gewichte und Knoten eindeutig bestimmt. Wir schreiben daher kurz $(b_i, c_i)_{i=1,\ldots,s}$.

Für die im Beispiel 19 erwähnten Quadraturformeln gilt:

Rechteckregel:	$s=1$	$b_1=1$	$c_1=0$
Mittelpunktregel:	$s=1$	$b_1=1$	$c_1=\frac{1}{2}$
Trapezregel:	$s=2$	$b_1=b_2=\frac{1}{2}$	$c_1=0, c_2=1$
Simpsonregel:	$s=3$	$b_1=b_2=\frac{1}{6}$, $b_2=\frac{4}{6}$	$c_1=0$, $c_2=\frac{1}{2}$, $c_3=1$

Bemerkung 19. *Die Quadraturformel ist ebenfalls linear in f und monoton für $b_i \geq 0$, $i=1,\ldots,n$.*

Mit der Anzahl s von Knoten und Gewichten steigt der Aufwand der Quadraturformel gemessen in Funktionsauswertungen von f. Bei größerem Aufwand erwarten wir eine bessere Näherungslösung des Integrals. Die Approximationsgüte einer Quadraturformel wird durch die so genannte Ordnung charakterisiert.

Ordnung einer Quadraturformel: Jede Quadraturformel sollte zumindest Integrale mit konstantem Integranden K exakt berechnen können, d.h.

$$\int_a^b K dx = (b-a)K = (b-a)\sum_{i=1}^{s} b_i K.$$

Diese Mindestanforderung führt auf die Bedingung

$$\sum_{i=1}^{s} b_i = 1.$$

Um entsprechende Bedingungen für lineare, quadratische, kubische,... Integranden herzuleiten, gehen wir

ohne Einschränkung von $a = 0$ und $b = 1$ aus.

$$\frac{1}{2} = \int_0^1 x dx = \sum_{i=1}^{s} b_i c_i$$
$$\frac{1}{3} = \int_0^1 x^2 dx = \sum_{i=1}^{s} b_i c_i^2$$

bzw. allgemein:

$$\frac{1}{p} = \int_0^1 x^{p-1} dx = \sum_{i=1}^{s} b_i c_i^{p-1}.$$

Definition 11. *Eine Quadraturformel $(b_i, c_i)_{i=1,\ldots,s}$ hat die Ordnung p, falls sie exakte Lösungen für alle Polynome vom Grad $\leq p-1$ liefert.*

Nach den Überlegungen oben ist dies äquivalent zu der Bedingung

$$\sum_{i=1}^{s} b_i c_i^{q-1} = \frac{1}{q} \text{ für } q = 1, \ldots, p. \tag{4.2}$$

Nachtrag zur ohne Einschränkung gemachten Annahme $a = 0,\ b = 1$. Für ein Polynom $f(x)$ vom Grad $q \leq p-1$ gilt nach der Substitutionsregel:

$$\int_a^b f(x)dx = (b-a) \int_0^1 \underbrace{f(a + \tau(b-a))}_{\substack{\text{ebenfalls ein Polynom} \\ \text{vom Grad } q}} d\tau$$
$$= (b-a) \sum_{i=1}^{s} b_i f(a + c_i(b-a)).$$

Beispiel 20. *Die Ordnungen der im Beispiel 19 angegebenen Quadraturformeln sind*

Rechteckregel:	$p = 1$	$(s = 1)$	
Mittelpunktregel:	$p = 2$!	$(s = 1)$	
Trapezregel:	$p = 2$	$(s = 2)$	
Simpsonregel:	$p = 4$!	$(s = 3)$	$(q = 5 : \ \frac{5}{24} \neq \frac{1}{5})$

Warum ist die Mittelpunktregel auch exakt für lineare Funktionen und die Simpsonregel auch für Polynome vom Grad 3?

Definition 12. *Eine Quadraturformel heißt symmetrisch, falls gilt:*

$$c_i = 1 - c_{s+1-i}$$
$$b_i = b_{s+1-i},$$

d.h. die Knoten sind symmetrisch zum Punkt $\frac{1}{2}$ verteilt und der Gewichtsvektor liest sich von oben nach unten oder von unten nach oben identisch.

Satz 25. *Die Ordnung einer symmetrischen Quadraturformel ist gerade.*

Beweis: Wir nehmen an, die Ordnung sei ungerade, und führen dies zu einem Widerspruch. Konkret nehmen wir an, die Quadraturformel sei exakt für Polynome vom Grad $\leq 2m-2$, und zeigen, dass sie tatsächlich exakt für Polynome bis zum Grad $\leq 2m-1$ ist.

Sei $f(x)$ ein Polynom vom Grad $2m-1$. Dann lässt sich f darstellen als

$$f(x) = K(x-\frac{1}{2})^{2m-1} + g(x),$$

wobei $g(x)$ maximal den Grad $2m-2$ besitzt. Somit gilt aufgrund der Linearität des Integrals

$$\int_0^1 f(x)dx = K\int_0^1 (x-\frac{1}{2})^{2m-1}dx + \underbrace{\int_0^1 g(x)dx}_{\substack{\text{wird exakt durch die}\\ \text{Quadraturformel be-}\\ \text{rechnet}}}.$$

Wir betrachten den ersten Summanden genauer:

$$\int_0^1 (x-\frac{1}{2})^{2m-1}dx = \int_{-\frac{1}{2}}^{\frac{1}{2}} x^{2m-1}dx = 0.$$

Für die entsprechende Quadraturformel gilt

$$\begin{aligned}\sum_{i=1}^{s} b_i(\underbrace{c_i-\frac{1}{2}}_{\frac{1}{2}-c_{s+1-i}})^{2m-1} &= \sum_{i=1}^{s} b_{s+1-i}(\frac{1}{2}-c_{s+1-i})^{2m-1}\\ &= -\sum_{j=1}^{s} b_j(c_j-\frac{1}{2})^{2m-1}\end{aligned}$$

und daher auch

$$\sum_{i=1}^{s} b_i(c_i-\frac{1}{2})^{2m-1} = 0.$$

Insgesamt erhalten wir

$$\begin{aligned}\int_0^1 f(x)dx &= \int_0^1 g(x)dx\\ &= \sum_{i=0}^{s} b_i g(c_i) = \sum_{i=0}^{s} b_i f(c_i).\end{aligned}$$

□

Im folgenden Satz wird deutlich, dass bei vorgegebenen Knoten $c_1 < \ldots < c_s$ die Quadraturformel schon eindeutig bestimmt ist, wenn wir mindestens die Ordnung s fordern. Die Gewicht $b_1, \ldots, b_s$ lassen sich dann eindeutig aus den Ordnungsbedingungen (4.2) (p ersetzt durch s) berechnen. Dies ist leicht einzusehen. Denn (4.2) ist in diesem Fall äquivalent zu

$$\begin{pmatrix} c_1^0 & c_2^0 & \ldots & c_s^0\\ c_1^1 & c_2^1 & \ldots & c_s^1\\ \vdots & \vdots & & \vdots\\ c_1^{s-1} & c_2^{s-1} & \ldots & c_s^{s-1}\end{pmatrix}\begin{pmatrix} b_1\\ b_2\\ \vdots\\ b_s\end{pmatrix} = \begin{pmatrix} 1\\ \frac{1}{2}\\ \vdots\\ \frac{1}{s}\end{pmatrix}$$

und die Vandermonde-Matrix C ist genau dann invertierbar, wenn die Knoten c_i paarweise verschieden sind. Insbesondere gilt mit $\tau = (1, \frac{1}{2}, \ldots, \frac{1}{s})^T$ die Darstellung

$$b = C^{-1}\tau.$$

Alternativ gilt für das i-te Lagrange-Polynom $L_i(x)$ zu den paarweise verschiedenen Knoten $c_1, .., c_s$ die Gleichung

$$b_i = \sum_{j=1}^{s} b_j L_i(c_j) = \int_0^1 L_i(x)dx.$$

Das i-te Lagrange-Polynom der Knoten $c_1 < \ldots < c_s$ ist das eindeutig bestimmte Polynom vom Grad $= s-1$, welches in allen Knoten c_j, $j \neq i$, verschwindet und in c_i den Wert 1 annimmt:

$$\deg L_i = s-1, \qquad L_i(c_j) = \begin{cases} 0, & \text{falls} \quad i \neq j \\ 1, & \text{falls} \quad i = j. \end{cases}$$

Satz 26. *Seien Knoten $c_1 < \ldots < c_s$ vorgegeben. Verlangen wir von einer Quadraturformel $(b_i, c_i)_{i=1,\ldots,s}$ mindestens die Ordnung s, so sind die Gewichte eindeutig bestimmt durch*

$$b = C^{-1}\tau$$

bzw.

$$b_i = \int_0^1 L_i(x)dx,$$

wobei

$$L_i(x) = \prod_{j=1, j\neq i}^{s} \frac{x - c_j}{c_i - c_j}$$

das i-te Lagrange-Polynom bezüglich der Knoten c_j ist.

4.1.1 Quadraturformeln mit erhöhter Ordnung

Satz 26 macht deutlich: Sind die Knoten $c_1 < \ldots < c_s$ erst einmal gewählt, so sind die Gewichte einer Quadraturformel mit Ordnung $p \geq s$ und somit die Formel insgesamt bereits festgelegt. Eine Frage, die sich nun stellt ist, wie die Knoten gewählt werden sollten, um die Ordnung $p \geq s$ zu maximieren. Wie groß kann die Ordnung überhaupt sein?

Wir suchen Quadraturformeln mit Ordnung $p = s + m$, $m \geq 1$, d.h. Polynome vom Grad $\leq s + m - 1$ sollen exakt integriert werden. Um entsprechende Bedingungen an die Knoten herzuleiten, benutzen wir das Polynom

$$M(x) = (x - c_1)(x - c_2) \cdot \ldots \cdot (x - c_s).$$

Offenbar ist der Grad von $M(x)$ gleich s und für jedes Polynom $f(x)$ vom Grad $\leq s + m - 1$ finden wir

$$f(x) = M(x)g(x) + r(x),$$

wobei $g(x)$ und $r(x)$ Polynome vom Grad $\leq m - 1$ bzw. $\leq s - 1$ sind. Damit gilt

$$\int_0^1 f(x)dx = \int_0^1 M(x)g(x)dx + \int_0^1 r(x)dx$$
$$\sum_{i=1}^{s} b_i f(c_i) = \sum_{i=1}^{s} b_i \underbrace{M(c_i)}_{=0} g(c_i) + \sum_{i=0}^{s} b_i r(c_i),$$

wobei jeweils die letzen Summanden gleich sind. Wir erhalten somit:

Satz 27. *Sei $(b_i, c_i)_{i=1,\dots,s}$ eine Quadraturformel der Ordnung $p \geq s$. Die Ordnung ist genau dann $s+m$, falls*

$$\int_0^1 M(x)g(x)dx = 0 \tag{4.3}$$

für alle Polynome g vom Grad $\leq m-1$ gilt.

Bemerkung 20.
(i) Wir definieren das Skalarprodukt

$$< f, g > = \int_0^1 f(x)g(x)dx. \tag{4.4}$$

Bedingung (4.3) *besagt daher, dass $M(x)$ orthogonal zum Raum aller Polynome vom Grad $\leq m-1$ bezüglich des in* (4.4) *definierten Skalarprodukts steht.*

(ii) Gleichung (4.3) *stellt nur Bedingungen an die Knoten, was nach Satz 26 keine Überraschung darstellt.*

Satz 28. *Die maximale Ordnung einer Quadraturformel ist $2s$.*

Beweis: Die Gauß-Quadraturformeln besitzen die Ordnung $2s$ (siehe Satz 29). Eine höhere Ordnung ist nicht möglich, da

$$< M, M > = \int_0^1 M(x)^2 dx > 0,$$

d.h. (4.3) ist für $g = M$ nicht erfüllt. Beachte: $\deg(M) = s$.

□

Satz 29. *(Gauß 1814) Es existiert eine eindeutige Quadraturformel der Ordnung 2s. Sie ist gegeben durch*

$$c_i = \frac{1}{2}(1 + \gamma_i)$$

für $i = 1, \dots, s$, wobei $\gamma_1, \dots, \gamma_s$ die Nullstellen des Legendre-Polynoms vom Grad s sind. Die Gewichte b_i sind gemäß Satz 26 eindeutig bestimmt.

Beweis: Nach Satz 27 gilt:

$$\begin{aligned}
\text{Ordnung } p = 2s &\iff \int_0^1 M(x)g(x)dx = 0 \quad \forall_{g,\deg(g)\leq s-1} \\
&\iff \int_{-1}^1 M\Big(\frac{1}{2}x + \frac{1}{2}\Big) f(x)dx = 0 \quad \forall_{f,\deg(f)\leq s-1} \\
&\iff M\Big(\frac{1}{2}x + \frac{1}{2}\Big) = \underbrace{K \cdot (\text{Legendre-Polynom vom Grad } s)}_{=\prod_{i=0}^{s} \frac{1}{2}(x-\gamma_i)} \\
&\qquad \uparrow \text{ Theorie orthogonaler Polynome} \\
&\iff c_i - \frac{1}{2} = \frac{\gamma_i}{2}.
\end{aligned}$$

□

Nachtrag:

(i) Die Gewichte der Gaußschen Quadraturformel sind positiv. Denn: Wir wissen nach Satz 26 gilt $b_i = \int_0^1 L_i(x)dx$ mit

$$L_i(c_j) = \begin{cases} 0, & \text{falls} \quad i \neq j \\ 1, & \text{falls} \quad i = j \end{cases}$$

und $\deg L_i = s-1$. Da die Ordnung der Gauß-Quadraturformel $p = 2s$ ist und $\deg L_i^2 = 2s-2 \leq 2s-1$, folgt

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

(ii) Da die Nullstellen der Legendre-Polynome symmetrisch zum Punkt 0 im Intervall $[-1, 1]$ liegen, gilt dies auch für die Knoten bezogen auf das Intervall $[0, 1]$ und den Punkt $\frac{1}{2}$. Zudem gilt $b_i = b_{s+1-i}$. Denn:

$$\begin{aligned} L_{s+1-i}(x) &= \prod_{j=1, j\neq i}^{s} \frac{x - c_{s+1-j}}{c_{s+1-i} - c_{s+1-j}} \\ &\underset{c_i = 1-c_{s+1-i}}{=} \prod_{j=1, j\neq i}^{s} \frac{1 - x - c_j}{c_i - c_j} = L_i(1-x) \end{aligned}$$

und somit

$$\begin{aligned} b_{s+1-i} = \int_0^1 L_{s+1-i}(x)dx &= \int_0^1 L_i(1-x)dx \\ &= -\int_1^0 L_i(y)dy = b_i. \end{aligned}$$

Also sind die Gaußschen Quadraturformeln symmetrisch.

Beispiel 21. *Bezeichne P_i das Legendre-Polynom vom Grad i.*

s=1: *Es gilt $P_1(x) = x$ und somit $\gamma_1 = 0$. Wir erhalten also gemäß Satz 29 $c_1 = \frac{1}{2}$ und $b_1 = 1$. Dies ist genau die oben bereits eingeführte Mittelpunktregel mit Ordnung $p = 2$.*

s=2: *Es gilt $P_2(x) = \frac{3}{2}x^2 - \frac{1}{2}$ und somit $\gamma_{1,2} = \pm\frac{\sqrt{3}}{3}$. Wir erhalten wieder mit Satz 29 die Parameter*

$$\begin{aligned} c_1 &= \frac{1}{2} - \frac{\sqrt{3}}{6} \\ c_2 &= \frac{1}{2} + \frac{\sqrt{3}}{6} \\ b_1 = b_2 &= \frac{1}{2} \end{aligned}$$

und somit die Quadraturformel

$$\int_0^1 f(x)dx \approx \frac{1}{2} f\Big(\frac{1}{2} - \frac{\sqrt{3}}{6}\Big) + \frac{1}{2} f\Big(\frac{1}{2} + \frac{\sqrt{3}}{6}\Big)$$

der Ordnung 4.

s=3: *Es gilt* $P_3(x) = \frac{5}{2}x^3 - \frac{3}{2}x$ *und somit* $\gamma_2 = 0, \gamma_{1,3} = \pm\frac{\sqrt{15}}{5}$. *Für die Knoten finden wir gemäß Satz 29*

$$\begin{aligned} c_1 &= \frac{1}{2} - \frac{\sqrt{15}}{10} \\ c_2 &= 0 \\ c_3 &= \frac{1}{2} + \frac{\sqrt{15}}{10}. \end{aligned}$$

Wegen der Symmetrie gilt $b_1 = b_3$. *Weiter gilt aufgrund Bedingung* (4.2) *für* $q = 1$

$$2b_1 + b_2 = 1$$

und gemäß Satz 26 auch

$$\begin{aligned} b_2 = \int_0^1 l_2(x)dx &= \int_0^1 \frac{(x - \frac{1}{2} + \frac{\sqrt{15}}{10})(x - \frac{1}{2} - \frac{\sqrt{15}}{10})}{-\frac{\sqrt{15}}{10}\frac{\sqrt{15}}{10}}dx \\ &= -\int_0^1 \frac{100}{15}\big((x - \frac{1}{2})^2 - \frac{15}{100}\big)dx \\ &= -\frac{100}{15}\underbrace{\int_0^1 (x - \frac{1}{2})^2 dx}_{=\frac{1}{12}} + 1 = \frac{8}{18}. \end{aligned}$$

Für die Gewichte erhalten wir also insgesamt:

$$\begin{aligned} b_1 = b_3 &= \frac{5}{18} \\ b_2 &= \frac{4}{9}. \end{aligned}$$

Die Gaußsche Quadraturformel der Ordnung 6 lautet also

$$\int_0^1 f(x)dx \approx \frac{5}{18}f\Big(\frac{1}{2} - \frac{\sqrt{15}}{10}\Big) + \frac{4}{9}f\Big(\frac{1}{2}\Big) + \frac{5}{18}f\Big(\frac{1}{2} + \frac{\sqrt{15}}{10}\Big).$$

4.1.2 Untersuchung des Quadraturfehlers

Der Fehler des näherungsweise berechneten Integrals ist

$$\begin{aligned} \int_a^b f(x)dx - (b-a)\sum_{i=1}^s b_i f(a + c_i(b-a)) &= (b-a)\int_0^1 f(a + t(b-a))dt - (b-a)\sum_{i=1}^s b_i f(a + c_i(b-a)) \\ &= (b-a)\Big[\int_0^1 g(\tau)d\tau - \sum_{i=1}^s b_i g(c_i)\Big] \end{aligned} \tag{4.5}$$

mit $g(x) := f(a + x(b-a))$. Um den Fehler zu untersuchen, betrachten wir das lineare Funktional

$$R(g) = \int_0^1 g(\tau)d\tau - \sum_{i=1}^s b_i g(c_i), \tag{4.6}$$

welches jeder Funktion $g : [0,1] \to \mathbb{R}$ den Quadraturfehler auf dem normierten Intervall $[0,1]$ zu ordnet. Das Funktional ist linear in g, da

$$R(\lambda g + \mu f) = \lambda R(g) + \mu R(f).$$

Allgemein kann R auf Vektorräumen von Funktionen operieren, für die das Integral $\int_0^1 g(\tau)d\tau$ definiert ist. Aber wie lässt sich $R(g)$ alternativ zu (4.6) bestimmen? Um diese Frage beantworten zu können beschränken wir uns auf p-mal stetig differenzierbare Funktionen, wobei p die Ordnung der Quadraturformel ist. Die folgende Integraldarstellung von $R(g)$ für $g \in C^p([0,1])$ geht auf Peano zurück.

Satz 30. *(Peano, 1913-1918) Die Quadraturformel habe die Ordnung p und $g : [0,1] \to \mathbb{R}$ sei p-mal stetig differenzierbar. Dann gilt*

$$R(g) = \int_0^1 K_p(t) g^{(p)}(t) dt,$$

wobei der so genannte Peano-Kern definiert ist durch

$$K_p(t) = \frac{1}{(p-1)!} R(x \mapsto (x-t)_+^{p-1}).$$

Klärung der Schreibweise:

a) Die Funktion $._+ : \mathbb{R} \to \mathbb{R}^{\geq 0}$ ist definiert durch

$$x_+ = \begin{cases} x, & \text{falls } x > 0, \\ 0, & \text{falls } x \leq 0 \end{cases}$$

und $(. - t)_+^n : \mathbb{R} \to \mathbb{R}^{\geq 0}$ ist definiert durch

$$(x-t)_+^n = \begin{cases} (x-t)^n, & \text{falls } x > t, \\ 0, & \text{falls } x \leq t. \end{cases}$$

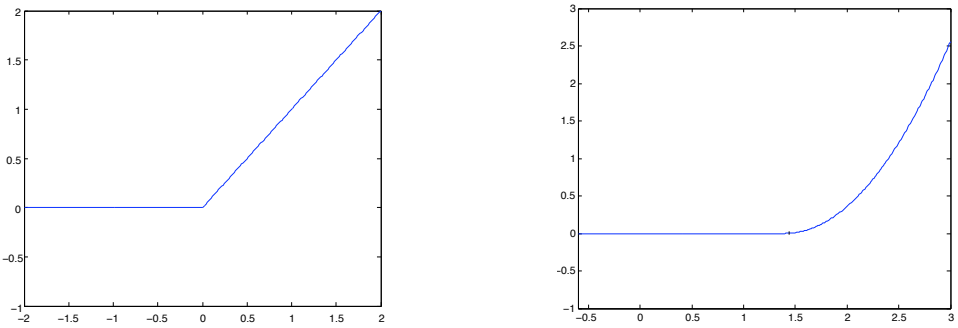


Abbildung 4.1: Graph der Funktionen $._+ : \mathbb{R} \to \mathbb{R}^{\geq 0}$ (links) und $(. - 1.4)_+^2 : \mathbb{R} \to \mathbb{R}^{\geq 0}$ (rechts).

b) Die Schreibweise $R(x \mapsto (x-t)_+^{p-1})$ bedeutet, dass wir das lineare Funktional R auf die von x abhängige Funktion $h(x) = (x-t)_+^{p-1}$ anwenden.

Insgesamt gilt also

$$\begin{aligned} K_p(t) &= \frac{1}{(p-1)!}\Big(\int_t^1 (\tau - t)^{p-1} d\tau - \sum_{i=1}^{s} b_i (c_i - t)_+^{p-1}\Big) \\ &= \frac{(1-t)^p}{p!} - \sum_{i=1}^{s} b_i \frac{(c_i - t)_+^{p-1}}{(p-1)!}. \end{aligned}$$

Beweis: (von Satz 30) Taylorentwicklung von $g(x)$ in $x = 0$ mit Restglied in Integralform liefert

$$g(x) = \underbrace{g(0) + g'(0)x + \ldots + g^{p-1}(0)\frac{x^{p-1}}{(p-1)!}}_{=:q(x)} + \underbrace{\frac{1}{(p-1)!}\int_0^x g^p(t)(x-t)^{p-1} dt}_{=:r_{p-1}(x)}.$$

Mit der oben eingeführten Schreibweise lässt sich das Restglied auch schreiben als

$$r_{p-1}(x) = \frac{1}{(p-1)!}\int_0^1 g^{(p)}(t)(x-t)_+^{p-1} dt.$$

Da die Quadraturformel die Ordnung p besitzt gilt $R(q) = 0$ und somit wegen der Linearität von R auch

$$\begin{aligned} R(g) &= R(r_{p-1}) \\ &= \frac{1}{(p-1)!} R(x \mapsto \int_0^1 g^p(t)(x-t)_+^{p-1} dt) \\ &\underset{\uparrow}{=} \frac{1}{(p-1)!}\int_0^1 g^{(p)}(t) R(x \mapsto (x-t)_+^{p-1}) dt. \end{aligned}$$

Sätze der Analysis über die Vertauschbarkeit von Grenzwerten

□

Bemerkung 21. *Satz 30 gilt natürlich auch, wenn die Ordnung der Quadraturformel tatsächlich größer als p ist.*

Beispiel 22.
(i) Mittelpunktregel (Ordnung 2, $c_1 = \frac{1}{2}$, $b_1 = 1$)

$$K_1(t) = 1 - t - (\frac{1}{2} - t)_+^0 = \begin{cases} -t, & \text{falls } 0 \le t \le \frac{1}{2} \\ 1 - t, & \text{falls } \frac{1}{2} < t \le 1 \end{cases}$$

$$K_2(t) = \frac{(1-t)^2}{2} - (\frac{1}{2} - t)_+ = \begin{cases} \frac{t^2}{2}, & \text{falls } 0 \le t \le \frac{1}{2} \\ \frac{(1-t)^2}{2}, & \text{falls } \frac{1}{2} < t \le 1 \end{cases}$$

(ii) Trapezregel (Ordnung 2, $c_1 = 0, c_2 = 1$, $b_1 = b_2 = \frac{1}{2}$)

$$\begin{aligned} K_1(t) &= \frac{1}{2} - t \\ K_2(t) &= -(1-t)\frac{t}{2} \end{aligned}$$

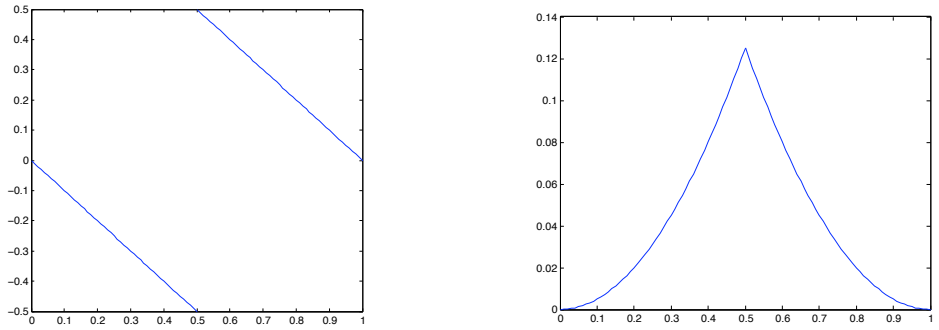


Abbildung 4.2: Graph von K_1 (links) und K_2 (rechts) der Mittelpunktregel.

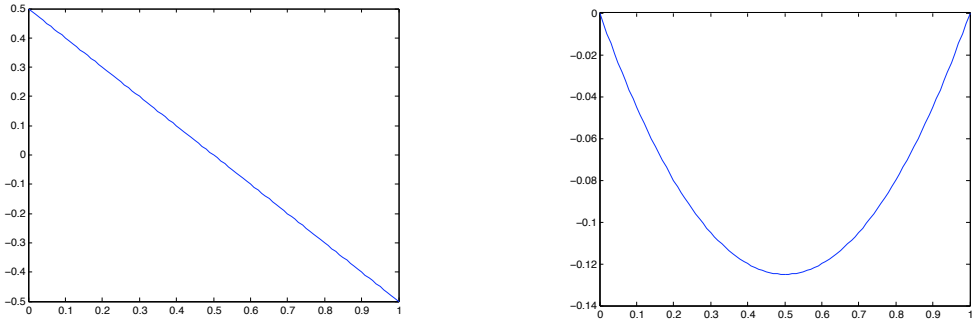


Abbildung 4.3: Graph von K_1 (links) und K_2 (rechts) der Trapezregel.

(iii) Simpson-Regel (Ordnung 4, $c_1 = 0, c_2 = \frac{1}{2}, c_3 = 1$*,* $b_1 = b_3 = \frac{1}{6}, b_2 = \frac{2}{3}$*)*

$$K_1(t) = 1 - t - \frac{2}{3}(\frac{1}{2} - t)_+^0 - \frac{1}{6} \;=\; \begin{cases} \frac{1}{6} - t, & \text{falls } 0 \le t \le \frac{1}{2} \\ \frac{5}{6} - t, & \text{falls } \frac{1}{2} < t \le 1 \end{cases}$$

$$K_2(t) = \frac{(1-t)^2}{2} - \frac{2}{3}(\frac{1}{2} - t)_+ - \frac{1}{6}(1-t) \;=\; \begin{cases} \frac{t^2}{2} - \frac{t}{6}, & \text{falls } 0 \le t \le \frac{1}{2} \\ \frac{t^2}{2} - \frac{5}{6}t + \frac{1}{3}, & \text{falls } \frac{1}{2} < t \le 1 \end{cases}$$

$$K_3(t) = \frac{(1-t)^3}{6} - \frac{2}{6}(\frac{1}{2} - t)_+^2 - \frac{1}{12}(1-t)^2 \;=\; \begin{cases} \frac{t^3}{6} + \frac{t^2}{12}, & \text{falls } 0 \le t \le \frac{1}{2} \\ -\frac{t^3}{6} + \frac{5}{12}t^2 - \frac{t}{3} + \frac{1}{12}, & \text{falls } \frac{1}{2} < t \le 1 \end{cases}$$

$$K_4(t) = \begin{cases} \frac{t^4}{24} - \frac{t^3}{36}, & \text{falls } 0 \le t \le \frac{1}{2} \\ \frac{t^4}{24} - \frac{5}{36}t^3 + \frac{t^2}{6} - \frac{t}{12} + \frac{1}{72}, & \text{falls } \frac{1}{2} < t \le 1 \end{cases}$$

Besitzt der Peano-Kern $K_p(t)$ wie in den Beispiel oben ein konstantes Vorzeichen auf $[0, 1]$, so findet man

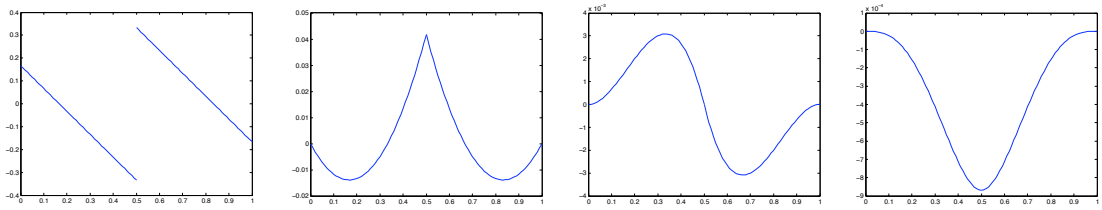


Abbildung 4.4: Graph von K_1, K_2, K_3 und K_4 der Simpson-Regel (von links nach rechts).

mit dem Mittelwertsatz der Integralrechnung die Darstellung

$$R(g) = g^{(p)}(\xi) \int_0^1 K_p(t)dt \tag{4.7}$$

für ein $\xi \in (0,1)$.

Lemma 6. *Es gilt:*

$$\int_0^1 K_p(t) = \frac{1}{p!}\Big[\frac{1}{p+1} - \sum_{i=1}^{s} b_i c_i^p\Big].$$

Beweis: Wir wissen

$$K_p(t) = \frac{(1-t)^p}{p!} - \sum_{i=1}^{s} b_i \frac{(c_i - t)_+^{p-1}}{(p-1)!}.$$

Integrieren liefert somit das Resultat:

$$\begin{aligned}\int_0^1 K_p(t)dt &= \frac{1}{p!}\underbrace{\int_0^1 (1-t)^p dt}_{=\frac{1}{p+1}} - \sum_{i=1}^{s} \frac{b_i}{(p-1)!}\int_0^1 (c_i - t)_+^{p-1} dt \\ &= \frac{1}{(p+1)!} - \sum_{i=1}^{s} \frac{b_i}{(p-1)!}\underbrace{\int_0^{c_i} (c_i - t)^{p-1} dt}_{\frac{c_i^p}{p}}\end{aligned}$$

□

Beispiel 23.
(i) Mittelpunktregel:

$$\int_0^1 K_2(t)dt = \frac{1}{24}$$

(ii) Trapezregel:

$$\int_0^1 K_2(t)dt = -\frac{1}{12}$$

(iii) Simpson-Regel:

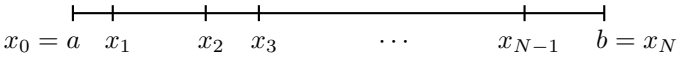
$$\int_0^1 K_4(t)dt = -\frac{1}{2880}$$

Mit (4.5) und Darstellung (4.7) finden wir insgesamt

$$\int_a^b f(x)dx - (b-a)\sum_{i=1}^{s} b_i f(a + c_i(b-a)) = (b-a)^{p+1} f^{(p)}(\xi) \int_0^1 K_p(t)dt$$

für ein $\xi \in (a,b)$. Beachte: $g^{(p)}(x) = (b-a)^p f^{(p)}(a + x(b-a))$. Der Fehler ist somit klein für Intervalle $[a,b]$ mit kleiner Intervalllänge $b - a << 1$.

Idee: Unterteile das Intervall $[a,b]$ in kleine Teilintervalle $[x_j, x_{j+1}]$, $j = 0, \ldots, N-1$, wobei für die Intervalllängen $x_{j+1} - x_j =: h_j << 1$ gilt:



Wende die Quadraturformel auf die Teilintervalle an. Wir finden dann für (p+1)-mal stetig differenzierbares f für den Fehler auf den Teilintervallen $[x_j, x_{j+1}]$ die Darstellung:

$$\begin{aligned}\int_{x_j}^{x_{j+1}} f(x)dx - h_j \sum_{i=1}^{s} b_i f(x_j + c_i h_j) &= h_j^{p+1} \int_0^1 K_p(t) f^{(p)}(x_j + c_i h_j)dt \\ &= \underbrace{h_j^{p+1} f^{(p)}(x_j) \int_0^1 K_p(t)dt}_{\substack{\text{führender / dominierender} \\ \text{Fehlerterm}}} + \mathcal{O}(h_j^{p+2})\end{aligned}$$

bzw.

$$\Big| \int_{x_j}^{x_{j+1}} f(x)dx - h_j \sum_{i=1}^{s} b_i f(x_j + c_i h_j) \Big| \leq h_j^{p+1} \max_{x \in [x_j, x_{j+1}]} |f^{(p)}(x)| \int_0^1 |K_p(t)|dt \; + \; \mathcal{O}(h_j^{p+2})$$

$$=: h_j \cdot err_j.$$

Der Fehler auf dem Gesamtintervall kann somit folgendermaßen abgeschätz werden:

$$\begin{aligned}\Big| \int_a^b f(x)dx - \sum_{j=0}^{N-1} h_j \sum_{i=1}^{s} b_i f(x_j + c_i h_j) \Big| &\leq \sum_{j=0}^{N-1} \Big| \int_{x_j}^{x_{j+1}} f(x)dx - h_j \sum_{i=1}^{s} b_i f(x_j + c_i h_j) \Big| \\ &\leq (b-a) \max_{j=0,\ldots,N-1} err_j.\end{aligned}$$

Bemerkung 22.
(i) Es gilt $\max_j err_j = \mathcal{O}((\max_j h_j)^p)$.

(ii) Ziel: Wähle die Unterteilung von $[a,b]$ *so, dass alle* err_j *möglichst gleich groß sind.*

4.2 Romberg-Integration

4.2.1 Trapezsumme

Wir wählen eine äquidistante Zerlegung des Integrationsintervalls $[a,b]$ mit $h := \frac{b-a}{N}$ wie folgt:

$$x_j := a + jh$$

für $j = 0, \ldots, N$. Auf jedem Intervall $[x_j, x_{j+1}]$ wenden wir die Trapezregel an:

$$\int_{x_j}^{x_{j+1}} f(x)dx \approx \frac{h}{2}[f(x_j) + f(x_{j+1})] \;=:\; T_j.$$

Für das Gesamtintervall $[a,b]$ erhalten wir somit

$$\begin{aligned}\int_a^b f(x)dx &\approx \sum_{j=0}^{N-1} T_j \\ &= h\Big[\frac{f(x_0)}{2} + \sum_{j=1}^{N-1} f(x_j) + \frac{f(x_N)}{2}\Big] \;=:\; T(h). \end{aligned} \tag{4.8}$$

Nach den Überlegungen oben gilt für den Fehler

$$\Big|\int_a^b f(x)dx - T(h)\Big| \le \frac{b-a}{12}h^2 \max_{x\in[a,b]} |f''(x)|. \tag{4.9}$$

Tatsächlich lässt sich noch Genaueres über den Fehler aussagen:

Satz 31. *(ohne Beweis)*
Sei $f \in C^{2m+1}([a,b])$ und $h = \frac{b-a}{N}$. Dann besitzt die Trapezsumme $T(h)$ folgende asymptotische Entwicklung in h^2 (bis zur Ordnung $2m$):

$$T(h) = \underbrace{\int_a^b f(x)dx}_{=:\tau_0} + \tau_2 h^2 + \tau_4 h^4 + \ldots + \tau_{2m} h^{2m} + R_{2m+2}(h)h^{2m+2}$$

mit von h unabhängigen Koeffizienten τ_{2k}. Der Restterm ist gleichmäßig in h beschränkt, d.h. es gibt eine von h unabhängige Konstante $C_{2m+2} > 0$ mit

$$|R_{2m+2}(h)| \le C_{2m+2}|b-a| \tag{4.10}$$

für alle $h = \frac{b-a}{N}$.

Bemerkung 23.
(i) Ist f ein Polynom vom Grad $\le 2m$, so gilt $R_{2m+2} \equiv 0$.

(ii) Die Koeffizienten τ_{2k} hängen vom Intervall $[a,b]$ ab.

4.2.2 Anwendung der Extrapolation

Die Idee der Romberg-Integration: Darstellung (4.9) garantiert

$$\lim_{h\to 0} T(h) = \lim_{N\to\infty} T\Big(\frac{b-a}{N}\Big) \;=\; \int_a^b f(x)dx.$$

Die Berechung von T in $h = 0$ ist natürlich nicht möglich, da in (4.8) die Summe nicht ausgewertet werden kann. Für kleine h können wir jedoch den Restterm der asymptotischen Entwicklung in Satz 31 aufgrund der Abschätzung (4.10) vernachlässigen, so dass sich $T(h)$ für solche h wie ein Polynom in h^2 verhält,

$$p(h^2) = \tau_0 + \tau_2 h^2 + \ldots + \tau_{2m} h^{2m},$$

dessen Wert an der Stele $h = 0$ genau das Integral liefert:

$$p(0) = \int_a^b f(x)dx.$$

Das nächste Problem ist: Wir kennen p nicht. Mögliches Vorgehen, um p bzw. genauer $p(0)$ näherungsweise zu bestimmen: Zu einer Folge von Schrittweiten

$$h_1 = \frac{b-a}{N_1},\ h_2 = \frac{b-a}{N_2}, \ldots, h_l = \frac{b-a}{N_l}$$

mit $0 < N_1 < \ldots < N_l$ bestimmt man die Trapezsummen

$$T_{i1} := T(h_i),\ i = 1, \ldots, l,$$

und mit Hilfe des Neville-Aitken Algorithmus extrapoliert man die Näherungen $p_{ik}(0)$, wobei p_{ik} das Interpolationspolynom zu den Stützstellen

$$(h_{i-k+1}^2, T(h_{i-k+1})), \ldots, (h_i^2, T(h_i))$$

für $1 \leq k \leq i \leq l$ bezeichne. Setzen wir $T_{ik} := p_{ik}(0)$, so lässt sich das Vorgehen wieder gut in einem Tableau, dem so genannten Extrapolationstableau, anordnen:

$$\begin{array}{lccccccc}
T(h_1) = T_{11} & & & & & & & \\
 & \searrow & & & & & & \\
T(h_2) = T_{21} & \rightarrow & T_{22} & & & & & \\
 & \searrow & & \searrow & & & & \\
T(h_3) = T_{31} & \rightarrow & T_{32} & \rightarrow & T_{33} & & & \\
 & \searrow & & \searrow & & \searrow & & \\
T(h_4) = T_{41} & \rightarrow & T_{42} & \rightarrow & T_{43} & \rightarrow & T_{44} & \\
\ldots & & & & & & &
\end{array}$$

Erklärung des Tableaus: Die Pfeile des Tableaus machen deutlich, dass die Näherungen T_{ik} für $2 \leq k \leq i$ aus den Näherungen $T_{i,k-1}$ und $T_{i-1,k-1}$ mit dem Algorithmus von Neville und Aitken berechnet werden. Die Formel dazu lautet entsprechend den Bezeichnungen in diesem Kontext:

$$T_{ik} = T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{\left(\frac{h_{i-k+1}}{h_i}\right)^2 - 1}.$$

Insbesondere brauchen wir die Interpolationspolynome p_{ik} nicht explizit zu kennen, sondern können rekursiv die Funktionswerte bei $h = 0$, also die T_{ik}, berechnen. Die Werte der ersten Spalte des Tableaus sind die Trapezsummen zu den verschiedenen h_i bzw. die "Extrapolationswerte" konstanter Polynome. Die Näherungen der zweiten Spalte sind Extrapolationen linearer Funktionen, die der dritten Spalte von quadratischen Funktionen usw.. Die Pünktchen im Tableau deuten an, dass die Größe von l am Anfang der Integration noch nicht feststeht, sondern erst im Laufe des Algorithmus (siehe unten) bestimmt wird.

Bemerkung 24.

(i) Eine wichtige Entscheidung betrifft die Wahl der h_i. Typischerweise gilt $h_1 = b - a$. Die $h_i, i \geq 2$, werden dann so gewählt, dass der Aufwand der Berechnungen (die Zahl der f-Auswertungen) minimiert wird. Z.B.

$$h_1 = b - a$$
$$h_i = \frac{h_{i-1}}{2},\ i = 2, \ldots$$

oder

$$h_1 = b - a,\ h_2 = \frac{b-a}{2},\ h_3 = \frac{b-a}{3}$$
$$h_i = \frac{h_{i-1}}{2},\ i = 4, \ldots$$

(ii) Man beachte, dass in die Berechnung von T_{ik} mit $k = i - \bar{k}$ die ersten $\bar{k}$ Trapezsummen $T(h_1), \ldots, T(h_{\bar{k}})$ nicht eingehen. Das bedeutet, wir können T_{ik} auch als Element $\bar{T}_{kk}$ eines Extrapolationstableaus interpretieren, welches durch Streichen der ersten $\bar{k}$ Diagonalen des ursprünglichen Tableaus entsteht. Diese Beobachtung ist insofern bedeutsam, dass sich z.B. bei Wahl $h_1 = b - a$ und h_i wie oben mit $a << b$ somit "große" Interpolationsfehler für kleine i nicht negativ auf alle folgenden Näherungen auswirkt (siehe Beispiel 25).

Wir bezeichnen mit

$$\varepsilon_{ik} := |T_{ik} - \tau_0|,\ 1 \leq k \leq i,$$

den Approximationsfehler der durch Extrapolation gewonnenen Näherungen T_{ik} von $\tau_0 = \int_a^b f(x)dx$.

Satz 32. *(ohne Beweis) Für die Approximationsfehler gilt:*

$$\varepsilon_{ik} = |\tau_{2k}| h_{i-k+1}^2 \cdot \ldots \cdot h_i^2 + \sum_{j=i-k+1}^{i} \mathcal{O}(h_j^{2k+2}),\ 1 \leq k \leq i, \tag{4.11}$$

für $h_j \leq h \to 0$.

Abschätzung (4.11) besagt, dass wir pro Spalte 2 Ordnungen gemessen in h gewinnen können.

Beispiel 24. *Berechnung von* $\int_0^1 e^x + 1\,dx = e \approx 2.718281828459046$ *durch Romberg-Integration. Wir wählen die Romberg-Folge $h_1 = b - a, h_i = \frac{h_{i-1}}{2}$ und erhalten das folgende Extrapolationstableau:*

$$\begin{array}{lcllcl} T(h_1) = T_{11} \approx \underline{2}.86 & & & & & \\ & \searrow & & & & \\ T(h_2) = T_{21} \approx \underline{2.7}54 & \to & T_{22} \approx \underline{2.718}9 & & & \\ & \searrow & & \searrow & & \\ T(h_3) = T_{31} \approx \underline{2.7}27 & \to & T_{32} \approx \underline{2.718}3 & \to & T_{33} \approx \underline{2.7182827} \end{array}$$

Für die Approximationsfehler gilt

$$\begin{array}{lcllcl} \varepsilon_{11} \approx 0.14086 & & & & & \\ & \searrow & & & & \\ \varepsilon_{21} \approx 0.03565 & \to & \varepsilon_{22} \approx 0.000579 & & & \\ & \searrow & & \searrow & & \\ \varepsilon_{31} \approx 0.00894 & \to & \varepsilon_{32} \approx 0.000037 & \to & \varepsilon_{33} \approx 8.599 \cdot 10^{-7} \end{array}$$

Die Approximationsfehler der ersten Spalte verbessern sich von Zeile zu Zeile ungefähr mit dem Faktor $\frac{1}{4}$. *Dies ist in der Ordnung* $p = 2$ *und der Halbierung von* h_i *in der Romberg-Folge begründet.*

Beispiel 25. *Für die Approximationsfehler der Romberg-Integration angewendet auf* $\int_0^{\frac{5}{4}\pi} e^{-x} \sin x \, dx = \frac{1}{2} + e^{-\frac{5}{4}\pi} \sin \frac{\pi}{4}$ *mit der Romberg-Folge* $h_1 = b - a, h_i = \frac{h_{i-1}}{2}$ *gilt*

$$\begin{array}{lllllll} \varepsilon_{11} \approx 0.54 & & & & & & \\ & \searrow & & & & & \\ \varepsilon_{21} \approx 0.273 & \to & \varepsilon_{22} \approx 0.1835 & & & & \\ & \searrow & & \searrow & & & \\ \varepsilon_{31} \approx 0.0776 & \to & \varepsilon_{32} \approx 0.0124 & \to & \varepsilon_{33} \approx 0.00100 & & \\ & \searrow & & \searrow & & \searrow & \\ \varepsilon_{41} \approx 0.0199 & \to & \varepsilon_{42} \approx 0.000698 & \to & \varepsilon_{43} \approx 0.000083 & \to & \varepsilon_{44} \approx 0.0001 \end{array}$$

Der "große" Fehler in T_{11} *bewirkt hier, dass die Näherung* T_{43} *besser ist als die Näherung* T_{44}.

Algorithmus der Romberg-Integration (Extrapolation basierend auf der Trapezsumme):

1. Wähle $h_1 > 0$ und setze $i = 1$.
2. Bestimme die Trapezsumme $T_{i1} = T(h_i)$.
3. Berechne T_{ik} für $k = 2, \ldots, i$ durch
$$T_{ik} = T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{\left(\frac{h_{i-k+1}}{h_i}\right)^2 - 1}.$$
Falls T_{ik} genau genug ist, beende den Algorithmus.
4. Falls i zu groß ist, beende den Algorithmus. Ansonsten erhöhe i um 1, wähle $h_i < h_{i-1}$ und gehe zu 2.

4.2.3 Allgemeines zu Extrapolationsverfahren

Lässt sich ein numerisches Verfahren allgemein schreiben als $T(h)$ und gilt

$$\lim_{h \to 0} T(h) = \tau_0,$$

wobei τ_0 die Lösung des eigentlichen Problems ist, so lässt sich die oben beschriebene Extrapolation anwenden, sofern die grundlegende Voraussetzung der Existenz einer asymptotischen Entwicklung des numerischen Verfahrens

$$T(h) = \tau_0 + \tau_p h^p + \tau_{2p} h^{2p} + \ldots + \tau_{mp} h^{mp} + \mathcal{O}(h^{(m+1)p}), \ h \to 0$$

vorliegt.

Beispiel 26. *Die Ableitung einer differenzierbaren Funktion an einer Stelle* x *lässt sich näherungsweise durch den zentralen Differenzenquotienten*

$$T(h) = \frac{f(x+h) - f(x-h)}{2h}$$

bestimmen. Dabei liefert die Differenzierbarkeit der Funktion gerade $\lim_{h\to 0} = f'(x)$*. Setzen wir* f *sogar als* $(2m+3)$*-mal stetig differenzierbare Funktion voraus, so finden wir durch Taylorentwicklungen*

$$T(h) = \frac{1}{2h}\left\{ f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \ldots + \mathcal{O}(h^{2m+3}) \right.$$
$$\left. -f(x) + hf'(x) - \frac{h^2}{2}f''(x) + \ldots + \mathcal{O}(h^{2m+3}) \right\}.$$

Innerhalb der geschweiften Klammer fallen die Terme mit ungeraden Potenzen von h *weg. Nach dem Teile durch* $2h$ *erhalten wir bei entsprechender Definition der Koeffizienten die asymptotische Entwicklung*

$$T(h) = \tau_0 + \tau_2 h^2 + \tau_4 h^4 + \ldots + \tau_{2m} h^{2m} + \mathcal{O}(h^{2m+2}),\ h \to 0.$$

Dagegen besitzt der einseitige Differenzenquotient

$$T(h) = \frac{f(x+h) - f(x)}{h}$$

nur eine Entwicklung der Form

$$T(h) = \tau_0 + \tau_1 h + \tau_2 h^2 + \ldots + \tau_m h^m + \mathcal{O}(h^{m+1}),\ h \to 0.$$

Aus Konvergenzgründen (vgl. Satz 32) ist daher die Approximation der Ableitung an einer Stelle x *durch den zentralen Differenzenquotienten zu realisieren.*

4.3 Adaptive Romberg-Integration

Zunächst einige allgemeine Gedanken zu adaptiven Verfahren basierend auf dem Anfangswertansatz.

Das eigentliche Problem der numerischen Integration lautet: Finde zum Integral $I := \int_a^b f(x)dx$ eine Approximation $\tilde{I}$ innerhalb einer vorgegebenen relativen Genauigkeit tol_{rel}, d.h.

$$|I - \tilde{I}| \leq tol_{rel}|I|.$$

Eine solche Abschätzung kann natürlich nicht explizit überprüft werden, da für eine berechnete Näherung weder $|I|$ noch die Differenz $I - \tilde{I}$ bekannt sind. Es ist jedoch ohne größere Einschräkungen möglich den Betrag von I durch einen Wert I^* zu ersetzen, welcher in der Größenordnung von $|I|$ liegt:

$$|I - \tilde{I}| \leq tol_{rel} I^*. \tag{4.12}$$

Einen solchen Wert könnte der Benutzer vorgeben, wenn er bereits eine gewisse Vorstellung von $|I|$ hat, oder aber I^* wird im Laufe der Integration aus ersten Approximationen gewonnen.

Ähnlich wie beim Übergang der Trapezregel zur Trapezsumme (Unterteilung des Intervalls $[a,b]$ in Teilintervalle, auf welche dann die Trapezregel angewendet wird) wenden wir die Romberg-Integration nicht auf das Gesamtintervall $[a,b]$, sondern zunächst auf das Teilintervall $[a, a+h_1]$ an, wobei die Schrittweite h_1 so gewählt werden soll, dass möglichst

$$|I(a, a+h_1) - \tilde{I}(a, a+h_1)| \approx tol_{abs} \tag{4.13}$$

gilt. Hier bezeichne $I(c,d)$ das exakte Integral und $\tilde{I}(c,d)$ eine Approximation des Integrals auf dem Intervall $[c,d]$. Ist das Teilproblem gelöst, so suchen wir ein entsprechendes h_2 für die Approximation von $I(a+h_1, a+h_1+h_2)$ usw.. Auf jedem Teilintervall wird somit ungefähr der gleiche absolute Fehler gemacht. Der Ansatz mit variablen Schrittweiten das Integral Schritt für Schritt zu berechnen wird als

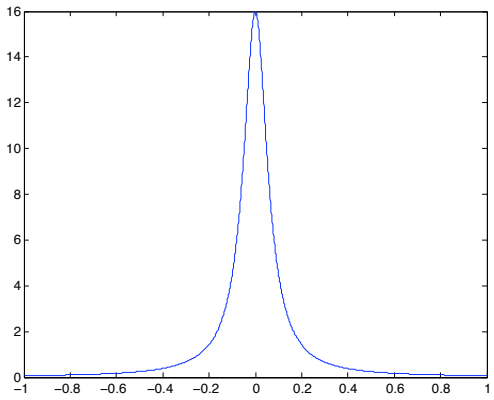


Abbildung 4.5: Graph einer "Nadelfunktion".

Anfangswertansatz bezeichnet.
Die Idee ist eine dem Problem angepasste Zerlegung des Intervalls $[a, b]$. In den Bereichen von $[a, b]$ in denen "viel passiert" sollte eine feinere Zerlegung gewählt werden; da wo "wenig passiert" sind größere Schrittweiten, also längere Teilintervalle, möglich (vgl. Abbildung 4.5).
Wir finden dann für den absoluten Fehler auf dem Gesamtintervall

$$|I - \tilde{I}| \leq m \cdot tol_{abs},$$

wobei m die Zahl der einzelnen Schritte bzw. die Anzahl der Teilintervalle ist. Wir können somit (4.12) garantieren, wenn

$$m \cdot tol_{abs} \leq I^* \cdot tol_{rel}$$

gilt. Man beachte, dass bei Verkleinerung der absoluten Genauigkeit tol_{abs} die Anzahl der Teilintervalle aufgrund der Ordnung des Verfahrens nicht entsprechend schnell steigt. Dies bedeutet: die linke Seite kann bei Vernachlässigung von Rundungsfehlern mit tol_{abs} beliebig klein gemacht werden. In der Praxis stellt sich jedoch das Problem einer guten Wahl von tol_{abs}.

Somit reduziert sich die Fehlerkontrolle auf die Abschätzung des Fehlers auf einem Teilintervall. Während die Überlegungen oben noch für beliebige numerische Integrationsverfahren formuliert werden können, ist die Schätzung des Approximationsfehlers eng mit der Struktur des Verfahrens verbunden. Konkret bei der Romberg-Integration liegen eine ganze Reihe von Approximationen im Extrapolationstableau vor, so dass eine vernünftige und auch billige Technik zur Fehlerschätzung und zur Schrittweitensteuerung möglich scheint.

4.3.1 Schätzung des Approximationsfehlers

Die Herleitung eines effektiven (effizienten und verlässlichen) Fehlerschätzers ist eine der schwierigsten Aufgaben bei der Entwicklung eines adaptiven Verfahrens. Eine übliche Vorgehensweise ist das Vergleichen von Approximationen verschiedener Ordnung.

Um einen solchen Fehlerschätzer herzuleiten, betrachten wir die Fehler ε_{ik} und das Integral $\int_x^{x+H} f(x)dx$ bezüglich des Intervalls $[x, x + H]$. Satz 32 liefert

$$\begin{aligned}\varepsilon_{ik} &= |\int_x^{x+H} f(x)dx - T_{ik}| \\ &= |\tau_{2k}| h_{i-k+1}^2 \cdot \ldots \cdot h_i^2 + \mathcal{O}(H^{2k+2})\end{aligned}$$

für $h_j \leq H \to 0$. Zudem gilt füer die Koeffizienten der asymptotischen Entwicklung in Satz 31 (ohne Beweis):

$$\tau_{2k} = \bar{\tau}_{2k} H + \mathcal{O}(H^2),$$

wobei $\bar{\tau}_{2k}$ vom Integranden f abhängt. Insgesamt erhalten wir daher

$$\varepsilon_{ik} = |\bar{\tau}_{2k}| \gamma_{ik} H^{2k+1} + \mathcal{O}(H^{2k+2})$$

mit

$$\gamma_{ik} \ := \ \frac{h_{i-k+1}^2 \cdot \ldots \cdot h_i^2}{H^{2k}} \ \leq \ 1.$$

Dies führt auf die Darstellung

$$\begin{aligned} \frac{\varepsilon_{i+1,k}}{\varepsilon_{ik}} &= \frac{\gamma_{i+1,k}}{\gamma_{ik}} + \mathcal{O}(H) \\ &= \Big(\underbrace{\frac{h_{i+1}}{h_{i-k+1}}}_{<<1} \Big)^2 + \mathcal{O}(H). \end{aligned}$$

Somit werden Approximationsfehler innerhalb einer Spalte k mit wachsendem Zeilenindex i weitgehend unabhängig vom Prolem und von H schnell kleiner:

$$\varepsilon_{i+1,k} << \varepsilon_{ik}. \tag{4.14}$$

Beispiel 27. *Quotienten von Beispiel 25:*

$\frac{\varepsilon_{21}}{\varepsilon_{11}} \approx 0,5$

$\frac{\varepsilon_{31}}{\varepsilon_{21}} \approx 0,28 \quad \frac{\varepsilon_{32}}{\varepsilon_{22}} \approx 0.0676$

$\frac{\varepsilon_{41}}{\varepsilon_{31}} \approx 0,26 \quad \frac{\varepsilon_{42}}{\varepsilon_{32}} \approx 0.0562 \quad \frac{\varepsilon_{43}}{\varepsilon_{33}} \approx 0.083$

$\frac{\varepsilon_{51}}{\varepsilon_{41}} \approx 0,252 \quad \frac{\varepsilon_{52}}{\varepsilon_{42}} \approx 0.0602 \quad \frac{\varepsilon_{53}}{\varepsilon_{43}} \approx 0.0208 \quad \frac{\varepsilon_{54}}{\varepsilon_{44}} \approx 0.00437$

$\frac{\varepsilon_{61}}{\varepsilon_{51}} \approx 0,250 \quad \frac{\varepsilon_{62}}{\varepsilon_{52}} \approx 0.0619 \quad \frac{\varepsilon_{63}}{\varepsilon_{53}} \approx 0.01630 \quad \frac{\varepsilon_{64}}{\varepsilon_{54}} \approx 0.00276 \quad \frac{\varepsilon_{65}}{\varepsilon_{55}} \approx 0.0052$

$\frac{1}{4} = 0.25 \quad \frac{1}{16} = 0.0625 \quad \frac{1}{64} = 0.015625 \quad \frac{1}{256} \approx 0.00391$

Es lässt sich gut erkennen, wie sich die Quotienten der Fehler dem Wert $\left(\frac{h_{i+1}}{h_{i-k+1}}\right)^2 = \frac{1}{2^{2k}}$ *annähern.*

Für die Beziehung zwischen den Spalten machen wir folgende Annahme

$$\varepsilon_{i,k+1} << \varepsilon_{ik}. \tag{4.15}$$

Diese Annahme gilt sicher für hinreichend kleine Schrittweite H, muss aber in einem Programm überprüft werden (vgl. Beispiel 25).

Definition 13. *Wir nennen $\bar{\varepsilon}$ einen Schätzer des unzugänglichen Approximationsfehlers*

$$\varepsilon := |I(x, x+H) - \tilde{I}(x, x+H)|,$$

falls positive Konstanten $k_1 \leq 1 \leq k_2$ existieren mit

$$k_1 \varepsilon \ \leq \ \bar{\varepsilon} \ \leq \ k_2 \varepsilon.$$

Nach (4.14) und (4.15) ist das Diagonalelement T_{kk} die genaueste Näherung der ersten k Zeilen des Extrapolationstableaus. Daher wäre ein Schätzer für ε_{kk} wünschenswert. Ein solcher Schätzer lässt sich jedoch aus den bisher berechneten Daten nicht konstruieren. Das folgende Lemma liefert aber einen Schätzer für die bisher zweitbeste Approximation:

Lemma 7. *Unter der Annahme* (4.15) *ist*

$$\bar{\varepsilon}_{k,k-1} := |T_{k,k-1} - T_{kk}|$$

ein Fehlerschätzer für $\varepsilon_{k,k-1}$.

Beweis: Wir schreiben

$$\bar{\varepsilon}_{k,k-1} = |(T_{k,k-1} - I) - (T_{kk} - I)|$$

und finden mit $I = \int_x^{x+H} f(x)dx$ und der Dreiecksungleichung

$$\varepsilon_{k,k-1} - \varepsilon_{kk} \;\leq\; \bar{\varepsilon}_{k,k-1} \;\leq\; \varepsilon_{k,k-1} + \varepsilon_{kk}.$$

Daher gilt

$$\big(1 - \underbrace{\frac{\varepsilon_{kk}}{\varepsilon_{k,k-1}}}_{<<1}\big)\varepsilon_{k,k-1} \;\leq\; \bar{\varepsilon}_{k,k-1} \;\leq\; \big(1 + \underbrace{\frac{\varepsilon_{kk}}{\varepsilon_{k,k-1}}}_{<<1}\big)\varepsilon_{k,k-1}$$

□

Das Diagonalelement T_{kk} wird also genau dann als Lösung akzeptiert, wenn

$$\bar{\varepsilon}_{k,k-1} \leq tol_{abs} \tag{4.16}$$

erfüllt ist. Bedingung (4.16) wird auch subdiagonales Fehlerkriterium genannt.

4.3.2 Schrittweitensteuerung

Wir formulieren hier nur die wichtigsten Aufgaben einer Schrittweitensteuerung ohne auf die tatsächliche Realisierung einzugehen. Für eine genauere Beschreibung gerade im Zusammenhang mit der Romberg-Integration sei auf das Buch *Numerische Mathematik I, Eine algorithmisch orientierte Einführung* von P. Deuflhard und A. Hohmann hingewiesen.

Eine Schrittweitensteuerung umfasst die Bereitstellung einer neuen, angepassten Schrittweite sowohl im Fall der erfolgreichen Berechnung einer Näherung auf dem letzten Teilintervall, als auch im Fall einer zuwiederholenden Berechnung: Wird der aktuelle Schritt akzeptiert, so muss basierend auf dem Fehlerschätzer und den bereits bekannten Daten eine neue optimale Schrittweite h_{opt} für den folgenden Schritt bestimmt werden. Wird der aktuelle Schritt nicht akzeptiert, so muss der aktuelle Schritt mit einer kleineren Schrittweite h_{neu} wiederholt werden. In beiden Fällen liefert eine Schrittweitensteuerung entsprechende Vorschläge.

4.4 Adaptive Mehrgitter-Quadratur

Bei der adaptiven Romberg-Integration, welche auf dem Anfangswertansatz basiert, wird das Integrationsintervall $[a, b]$ in einer willkürlich gewählten Richtung durchlaufen, wobei an Stellen wo "viel passiert"

kleinere Schrittweiten gewählt werden. Die adaptive Mehrgitter-Quadratur basiert auf dem so genannten Randwertansatz, bei welchem ausgehend von dem Gesamtintervall (oder einem groben Gitter des Gesamtintervalls) feinere Gitter und bessere Näherungen des Gesamtintegrals $\int_a^b f(x)dx$ berechnet werden, indem wiederum dort verfeinert werden soll, wo "viel passiert". Bei dem Anfangswertansatz benötigten wir neben einem Fehlerschätzer auch eine Schrittweitensteuerung. Für den Randwertansatz gilt es neben dem Schätzer eine Verfeinerungsregel des Intervalls anzugeben.

Die Herleitung eines Fehlerschätzers ist eng mit der Struktur der Quadraturformel verbunden. Oft wird ein solcher Schätzer durch eine so genannte eingebettete Quadraturformel realisiert. Bezeichne $(b_i, c_i)_{i=1,\dots,s}$ die eigentliche Quadraturformel der Ordnung p. Gesucht ist eine Quadraturformel $(\hat{b}_i, c_i)_{i=1,\dots,s}$ möglichst hoher Ordnung $\hat{p} < p$. Die Knoten der eingebetteten Formel entsprechen der eigentlichen Quadraturformel, um zusätzliche f-Auswertungen zu vermeiden. Wir setzen dann

$$\bar{\varepsilon} := \Big| h \sum_{i=1}^{s} b_i f(x + c_i h) - h \sum_{i=1}^{s} \hat{b}_i f(x + c_i h) \Big|,$$

wobei h die Intervalllänge bezeichne. Dies ist jedoch nur ein Schätzer für die eingebettete, schlechtere Quadraturformel. Betrachten wir zum Beispiel die Gaußsche Quadraturformel der Ordnung $p = 2s$, so hat jede eingebettete Quadraturformel nach Satz 26 höchstens die Ordnung $s - 1$. Insbesondere im Fall der Gauß-Quadratur sind heuristische Verbesserungsvorschläge gemacht worden.

Wir gehen im Folgenden davon aus, dass ein Fehlerschätzer bekannt ist, und gehen näher auf eine mögliche Verfeinerungsregel ein. Für ein Teilintervall $[c, d]$ von $[a, b]$ berechnen wir

$$\tilde{I}(c, d) := (c - d) \sum_{i=1}^{s} b_i f(c + c_i(c - d)).$$

Der Fehlerschätzer liefert eine Näherung

$$\bar{\varepsilon}(c, d) \approx \Big| \int_c^d f(x)dx - \tilde{I}(c, d) \Big|.$$

Mögliches Vorgehen:

(i) Berechne $\tilde{I}_1 := \tilde{I}(a, b)$ und $\varepsilon_1 := \bar{\varepsilon}(a, b)$. Falls

$$\varepsilon_1 \leq I^* \cdot tol_{rel},$$

so akzeptiere die berechnete Näherung. Sonst

(ii) unterteile das Intervall $[a, b]$ und berechne $\tilde{I}$ und $\bar{\varepsilon}$ für die Teilintervalle $\left[a, \frac{a+b}{2}\right]$ und $\left[\frac{a+b}{2}, b\right]$. Falls

$$\varepsilon_1 + \varepsilon_2 \leq I^* \cdot tol_{rel},$$

so akzeptiere $\tilde{I}_1 + \tilde{I}_2$ als Näherung. Sonst

(iii) unterteile jenes Intervall, in dem der Fehler am größten ist. Falls die Summe der Fehler kleiner ist als $I^* \cdot tol_{rel}$, so akzeptiere die Summe der einzelnen Näherung. Sonst wiederhole (iii).

Kapitel 5

Numerik gewöhnlicher Differentialgleichungen

Eine explizite Differentialgleichung besitzt die Form

$$y' = f(t, y) \ \in \mathbb{R}^n, \tag{5.1}$$

wobei $f : [t_0, T] \times \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar ist. Die Variable t ist in vielen Anwendungen die Zeit, y sind die Zustandsvariablen. Streng genommen handelt es sich bei (5.1) um ein System von n gekoppelten Differentialgleichungen

$$\begin{aligned} y_1'(t) &= f_1(t, y_1(t), \ldots, y_n(t)), \\ y_2'(t) &= f_2(t, y_1(t), \ldots, y_n(t)), \\ &\vdots \\ y_n'(t) &= f_n(t, y_1(t), \ldots, y_n(t)). \end{aligned}$$

Falls die Variable t die Zeit repräsentiert, wird für die partielle Ableitung nach t auch oft $\dot{y}$ geschrieben:

$$\dot{y} = f(t, y).$$

Anfangswertproblem (AWP): Finde zu einem vorgegebenen Anfangswert y_0 eine Lösung der Differentialgleichung (5.1) mit

$$y(t_0) = y_0.$$

Wir nennen $y : J \to \mathbb{R}^n$ eine Lösung des AWPs

$$y' = f(t, y), \quad y(t_0) \ = \ y_0, \tag{5.2}$$

falls t_0 im Intervall J liegt und $y(t_0) = y_0$ gilt, y stetig differenzierbar ist und der Differentialgleichung gehorcht.

Die Modellierung durch Differentialgleichungen ist extrem populär in der Physik, Biologie, Chemie, Ingenieurwissenschaften. Auch in anderen Wissenschaften wie der Betriebswirtschaftslehre, der Sportwissenschaften werden dynamische Prozesse durch Differentialgleichungen modelliert.

Satz 33. *(Existenz und Eindeutigkeit) Sei die rechte Seite f des Anfangswertproblems* (5.2) *stetig differenzierbar. Dann besitzt das Problem eine nicht weiter fortsetzbare eindeutige Lösung. Insbesondere existiert ein $t^* > t_0$, so dass die Lösung auf dem Intervall $[t_0, t^*)$ definiert ist.*

Beispiel 28. *Betrachte das AWP*

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

Die Lösung $y(t) = \frac{1}{1-t}$ ist nur für $t \in [0, 1)$ definiert.

5.1 Das explizite Euler-Verfahren

Beim expliziten Euler-Verfahren ersetzen wir die Ableitung der Lösung durch den Differenzenquotienten

$$\frac{y(t) - y(t_0)}{t - t_0} \approx y'(t_0)$$

und berechnen durch die Gleichung

$$\frac{y_1 - y(t_0)}{t_1 - t_0} = f(t_0, y_0)$$

für $t_1 - t_0$ hinreichend klein eine Näherung y_1 von $y(t_1)$. Die Näherung ist offenbar gegeben durch die Gleichung

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

mit $h := t_1 - t_0$. Geometrisch lässt sich dies folgendermaßen beschreiben: Ausgehend von einem Partikel mit der Position y_0 und Geschwindigkeit $v_0 = f(t_0, y_0)$ zum Zeitpunkt t_0 lässt sich die neue Position nach einer kurzen Zeit h approximativ bestimmen durch

$$y_1 = y_0 + hv_0.$$

Dabei ist die vergangene Zeit h so kurz, dass sich der Geschwindigkeitsvektor v_0 nicht stark verändert hat (vgl. Abbildung 5.1). Nach der Zeit h wird dann der Geschwindigkeitsvektor am Punkt (t_1, y_1) neu ausgewertet

$$v_1 := f(t_1, y_1)$$

und eine neue Näherung von $y(t_2)$ mit $t_2 = t_1 + h$ ermittelt:

$$y_2 = y_1 + hv_1.$$

Diese Berechnungen lassen sich entsprechend oft wiederholen, um Näherungen der Lösung auf dem Gesamtintervall $[t_0, T]$ zu erhalten. Das Euler-Verfahren lässt sich kompakt schreiben als

$$\begin{aligned} y_0 &:= y(t_0) \\ y_{n+1} &= y_n + hf(t_n, y_n). \end{aligned}$$

Es stellt sich jetzt natürlich die Frage nach der Qualität der Näherungen: Wie genau ist das Verfahren?

Zunächst betrachten wir den lokalen Fehler: Wir bestimmen den Fehler, welchen wir nach einem einzigen Schritt machen, wobei wir an einem beliebigen Punkt $y(t)$ der Lösung starten. Die Idee besteht darin, die Lösung nach Taylor zu entwickeln, um sie mit der numerischen Lösung, dem Euler-Verfahern, zu vergleichen. Dazu sei bemerkt, dass mit der stetigen Differenzierbarkeit der rechten Seite $f(t, y(t))$

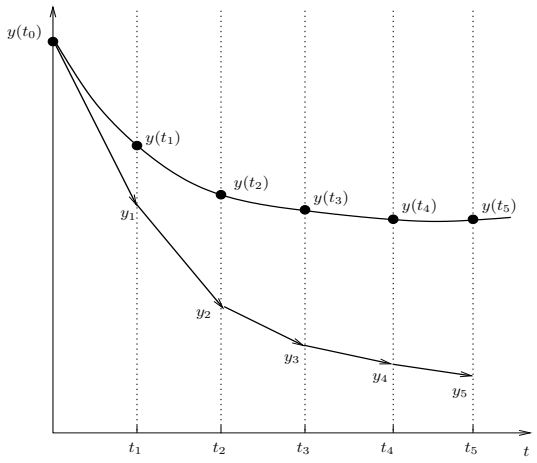


Abbildung 5.1: Explizites Euler-Verfahren

auch die stetige Differenzierbarkeit von $y'(t)$ folgt. Somit ist die Lösung $y(t)$ selbst sogar zweimal stetig differenzierbar. Wir können somit $y(t+h)$ um den Punkt t folgendermaßen entwickeln:

$$y(t+h) = y(t) + h\underbrace{y'(t)}_{=f(t,y(t))} + h^2 \int_0^1 (1-s)y''(t+sh)ds.$$

Für die Differenz der exakten Lösung und der mit dem Euler-Verfahren berechneten Approximation erhalten wir somit

$$y(t+h) - y_1 = h^2 \int_0^1 (1-s)y''(t+sh)ds$$

bzw.

$$\|y(t+h) - y_1\| \leq Ch^2$$

mit $C := \frac{1}{2}\max_{\tau\in[t_0,T]} \|y''(\tau)\|$.

Um den globalen Fehler abschätzen zu können, fordern wir zunächst zusätzlich:

$$\|f(t,y) - f(t,z)\| \leq L\|y-z\| \quad \forall_{y,z\in\mathbb{R}^n}, \forall_{t\in[t_0,T]}, \tag{5.3}$$

d.h. f ist lipschitz-stetig bezüglich des zweiten Arguments geichmäßig in t. Da wir f als stetig differenziebar angenommen haben gilt z.B.

$$L = \sup_{(t,y)\in[t_0,T]\times\mathbb{R}^n} \|\frac{\partial f}{\partial y}(t,y)\| < \infty.$$

Satz 34. *Unter diesen Voraussetzungen gilt für den globalen Fehler des Euler-Verfahrens*

$$\|y(t_n) - y_n\| \leq Mh$$

mit

$$M := \frac{e^{L(T-t_0)} - 1}{L} C$$

Insbesondere gilt

$$\max_{n,t_n\in[t_0,T]} \|y(t_n) - y_n\| \to 0 \text{ für } h \to 0,$$

d.h. die Näherungslösung konvergiert gleichmäßig gegen die exakte Lösung, wenn die Schrittweite gegen 0 *läuft.*

Beweis: Wir werden im Folgenden untersuchen, wie sich Fehler fortpflanzen, und abschließend berechnen, wie sich die Fehler akkumulieren.
Fehlerfortpflanzung: Seien v_{n+1} und w_{n+1} durch einen Schritt des Euler-Verfahrens aus v_n bzw. w_n ermittelt:

$$\begin{aligned} v_{n+1} &= v_n + hf(t_n, v_n), \\ w_{n+1} &= w_n + hf(t_n, w_n). \end{aligned}$$

Dann lässt sich der Abstand der neuen Werte durch den Abstand der Ausgangswerte wie folgt abschätzen:

$$\begin{aligned} \|v_{n+1} - w_{n+1}\| &\le \|v_n - w_n\| + h \underbrace{\|f(t_n, v_n) - f(t_n, w_n)\|}_{\le L\|v_n - w_n\|} \\ &\le (1 + hL)\|v_n - w_n\|. \end{aligned}$$

Fehlerakkumalation: Bezeichne $y_n^k,\ n \ge k$ die Näherung von $y(t_n)$ zum Anfangswert $y(t_k)$ nach $n - k$ Schritten. Insbesondere ist $y_n^0 = y_n$ und $y_n^n = y(t_n)$.

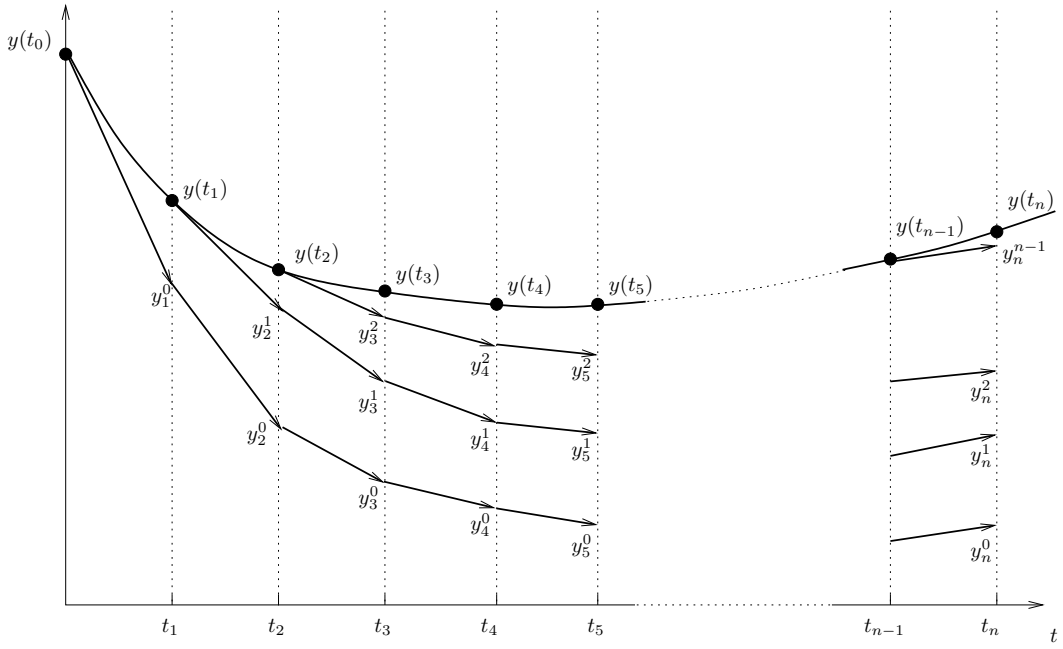


Abbildung 5.2: Lady Windermere's Fächer

Nach den Überlegungen über den lokalen Fehler gilt:

$$\|y_{k+1}^{k+1} - y_{k+1}^k\| \le Ch^2$$

und die Fehlerfortpflanzung liefert

$$\begin{aligned}\|y_n^{k+1} - y_n^k\| &\leq (1+hL)\|y_{n-1}^{k+1} - y_{n-1}^k\| \\ &\leq \dots \\ &\leq (1+hL)^{n-k-1}\|y_{k+1}^{k+1} - y_{k+1}^k\| \\ &\leq (1+hL)^{n-k-1}Ch^2.\end{aligned}$$

Insgesamt ergibt sich daher

$$\begin{aligned}\|y(t_n) - y_n\| &= \|y_n^n - y_n^0\| \\ &\leq \|y_n^n - y_n^{n-1}\| + \|y_n^{n-1} - y_n^{n-2}\| + \dots + \|y_n^2 - y_n^1\| + \|y_n^1 - y_n^0\| \\ &\leq Ch^2 + (1+hL)Ch^2 + \dots + (1+hL)^{n-2}Ch^2 + (1+hL)^{n-1}Ch^2 \\ &\leq Ch^2\big(1 + (1+hL) + \dots + (1+hL)^{n-2} + (1+hL)^{n-1}\big).\end{aligned}$$

Wir nutzen die Formel

$$\sum_{i=0}^{n-1} q^i = \frac{1-q^n}{1-q}$$

und erhalten

$$\begin{aligned}\|y(t_n) - y_n\| &\leq Ch^2\frac{1-(1+hL)^n}{1-(1+hL)} \\ &\leq Ch\frac{(1+hL)^n - 1}{L}.\end{aligned}$$

Für die Exponentialfunktion gilt

$$\exp(x) = 1 + x + \underbrace{\frac{x^2}{2!} + \frac{x^3}{3!} + \dots}_{\geq 0}$$

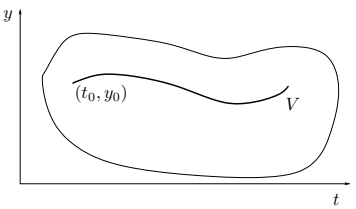
Wir erhalten daher die zu beweisende Abschätzung

$$\begin{aligned}\|y(t_n) - y_n\| &\leq Ch\frac{\exp(hL)^n - 1}{L} \\ &\leq Ch\frac{\exp(nhL) - 1}{L} \\ &\leq Ch\frac{\exp((T-t_0)L) - 1}{L}.\end{aligned}$$

□

Bemerkung 25.

Sei V eine beschränkte Umgebung der Trajektorie $\{(t, y(t)) | t \in [t_0, T]\}$. Satz 34 zeigt: Falls die Schrittweite hinreichend klein ist, so bleibt die numerische Trajektorie $\{(t_n, y_n) | t_n \in [t_0, T]\}$ in V. Insbesondere geht dann nur die Lipschitz-Konstante von $f|_V$ in die Fehlerabschätzung ein.



5.2 Runge–Kutta Verfahren

In diesem Abschnitt suchen wir Verallgemeinerungen des Euler-Verfahrens, um bei gleicher Schrittweite h höhere Genauigkeit zu erhalten:

$$\|y(t_n) - y_n\| = \mathcal{O}(h^p), \quad p > 1.$$

Ein mögliches Vorgehen ist: Wähle nicht nur eine Tangente, sondern das gewichtete Mittel mehrerer Tangenten an verschiedenen Stellen

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i f(t_n + c_i h, Y_i). \tag{5.4}$$

Um die Näherungen Y_i von $y(t_n + c_i h)$ zu berechnen, muss ein im Allgemeinen nichtlineares $(s \cdot n)$-dimensionales Gleichungssystem gelöst werden:

$$Y_i = y_n + h \sum_{j=1}^{s} a_{ij} f(t_n + c_j h, Y_j), \quad i = 1, \ldots, s. \tag{5.5}$$

Man nennt die Y_i die Stufenwerte. Gilt $a_{ij} = 0$ für $j \geq i$, so können die Stufenwerte explizit berechnet werden. Die Verfahrensmatrix A besitzt in diesem Fall die Struktur

$$A = \begin{pmatrix} 0 & \ldots & \ldots & 0 \\ a_{21} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ a_{s1} & \ldots & a_{s,s-1} & 0 \end{pmatrix}.$$

Man nennt das Verfahren, welches durch die Gleichungen (5.4) und (5.5) gegeben ist, Runge–Kutta Verfahren. Es ist durch die Gewichte b_i, die Knoten c_i und die Verfahrensmatrix A eindeutig bestimmt. Runge–Kutta Verfahren werden im so genannten Butcher-Tableau repräsentiert:

$$\begin{array}{c|ccc} c_1 & a_{11} & \ldots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \ldots & a_{ss} \\ \hline & b_1 & \ldots & b_s \end{array}$$

Definition 14. *Ein Runge–Kutta Verfahren besitzt die Ordnung p, falls für jedes Anfangswertproblem*

$$y' = f(t, y), \quad y(t_0) = y_0$$

mit $(p+1)$-mal stetig differenzierbarem f für den lokalen Fehler

$$y(t_0 + h) - y_1 = \mathcal{O}(h^{p+1})$$

gilt.

Satz 35. *Für den globalen Fehler gilt dann*

$$\|y(t_n) - y_n\| \leq M h^p,$$

wobei die Konstante M unabhängig von n und h mit $t_n = t_0 + nh \in [t_0, T]$ ist.

Beispiel 29. *Das klassische Runge–Kutta Verfahren der Ordnung $p = 4$ lautet*

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

Zusammenhang zur numerischen Integration: Das Anfangswertproblem (5.2) ist äquivalent zur Integraldarstellung

$$y(t) = y_0 + \int_{t_0}^{t} f(s, y(s))ds.$$