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

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

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 $i = n, n-1, ..., 1$.

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:

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

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*

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

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

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

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.

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

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.

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:

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

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

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

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

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

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.

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:*

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

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

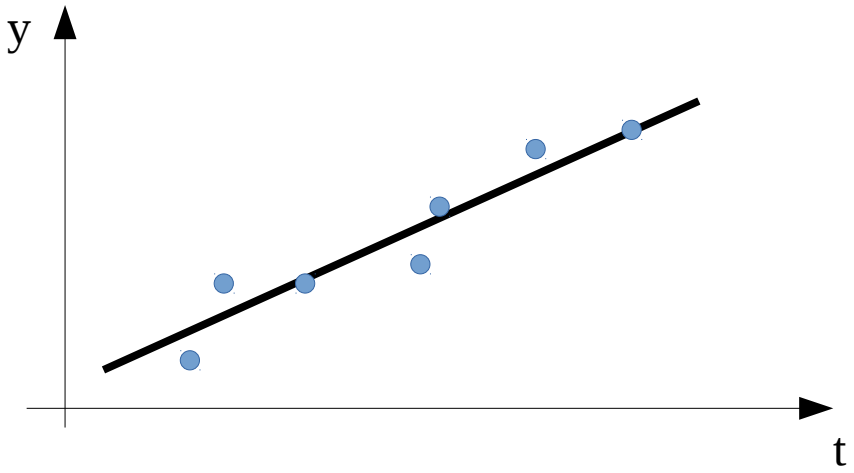
$$\begin{aligned}
|y_k - \bar{y}_k| &\leq \frac{n}{1 - eps} \cdot eps|y_k| + \mathcal{O}(eps^2) \\
&= n \cdot eps|y_k| + \mathcal{O}(eps^2) \text{ für } k = 1, \ldots, n-1.
\end{aligned}$$

Insgesamt folgt

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

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

Beispiel 10.



geg.: Meßdaten (t_j, y_j), $j = 1, ..., m$

suchen Funktion $y = f(t) = \sum_{i=1}^{n} x_i q_i(t)$

so, dass $y_i \approx f(t_j)$ für alle j

$q_1, ..., q_n$ geg. Funktionen, $x_1, ..., x_n$ unbekannt

genauer:

$$\left\| \begin{pmatrix} e_1 \\ : \\ e_m \end{pmatrix} \right\| = \text{min!} \quad \text{wobei } e_j = y_j - f(t_j)$$

euklidische Norm: $\sum_{j=1}^{m} e_j^2 = \text{min!}$

Methode der kleinsten Fehlerquadrate (Gauß)

Wiederholung: Gesamtalgorithmus mit LR-Zerlegung

(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)

Gesamtalgorithmus mit Cholesky-Verfahren

(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)

Gesamtalgorithmus mit QR-Zerlegung

(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)

Gesamtalgorithmus lineares Ausgleichsproblem

(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)

Beispiel 11.

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

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.

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

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

Berechnung der Iterierten in double precision liefert:

F_1			F_2		
x_0	=	2	x_0	=	2
x_1	=	$\underline{1.7}5$	x_1	=	$\underline{1.7}5$
x_2	=	$\underline{1.732}1$	x_2	=	$\underline{1.73}4$
x_3	=	$\underline{1.7320508}1$	x_3	=	$\underline{1.732}4$
x_4	=	$\underline{1.732050807568877}$	x_4	=	$\underline{1.73209}2$
x_5	=	$\underline{1.732050807568877}$	x_5	=	$\underline{1.732056}$

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

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

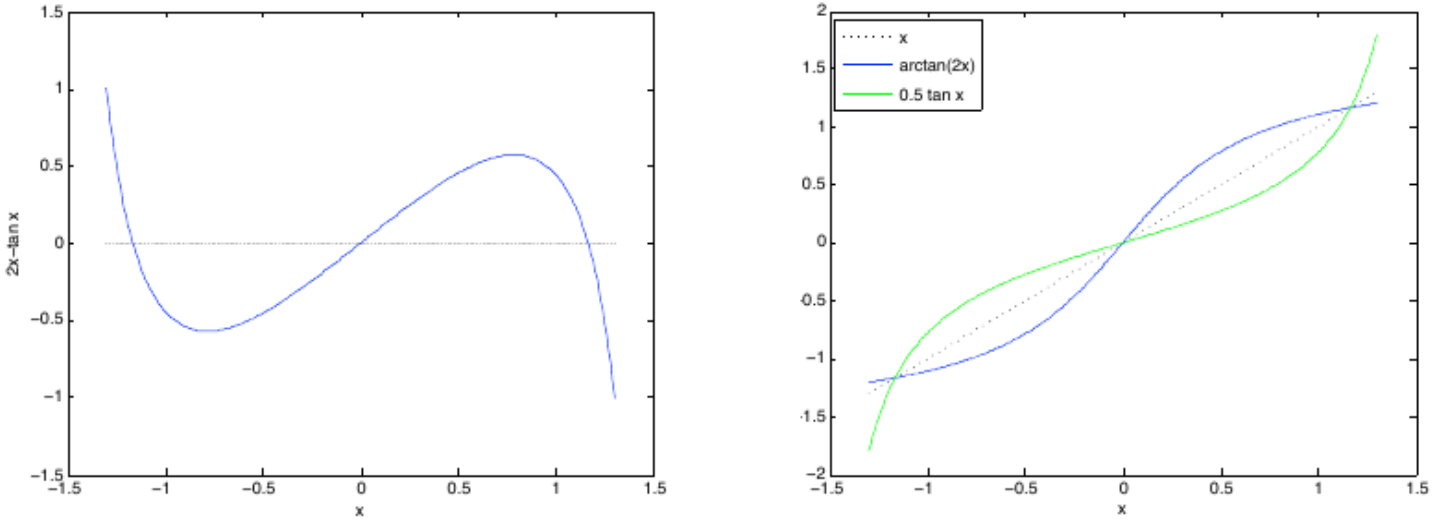


Abbildung 2.1: Links ist der Graph der Funktion $2x - \tan x$ dargestellt, rechts die Graphen von F_i.

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	$=$	$\underline{0}.78$	x_1	$=$	1.286	x_1	$=$	$\underline{1.1}76$
x_2	$=$	$\underline{0}.49$	x_2	$=$	$1.708 > \pi/2$	x_2	$=$	$\underline{1.16}88$
x_3	$=$	$\underline{0}.27$				x_3	$=$	$\underline{1.16}66$
x_4	$=$	$\underline{0}.14$				x_4	$=$	$\underline{1.165}9$
x_5	$=$	$\underline{0.0}69$				x_5	$=$	$\underline{1.165}66$
x_6	$=$	$\underline{0.0}35$				x_6	$=$	$\underline{1.165}591$
x_7	$=$	$\underline{0.0}17$				x_7	$=$	$\underline{1.165}571$
$\vdots$						$\vdots$		
x_{27}	$=$	$\underline{0.000000}166$				x_{27}	$=$	$\underline{1.165561185207212}$
x_{28}	$=$	$\underline{0.0000000}083$				x_{28}	$=$	$\underline{1.165561185207212}$

Zu Bemerkung 12.

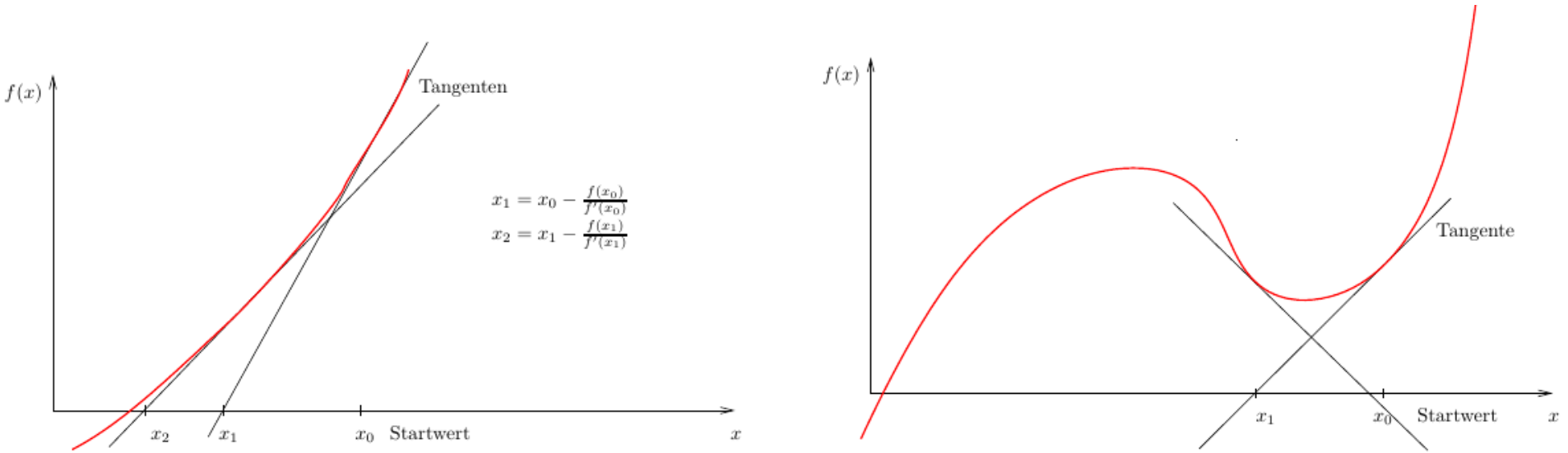


Abbildung 2.2: Links konvergiert das Newton-Verfahren, rechts lässt sich keine Konvergenz beobachten

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!

Beispiel 16.

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

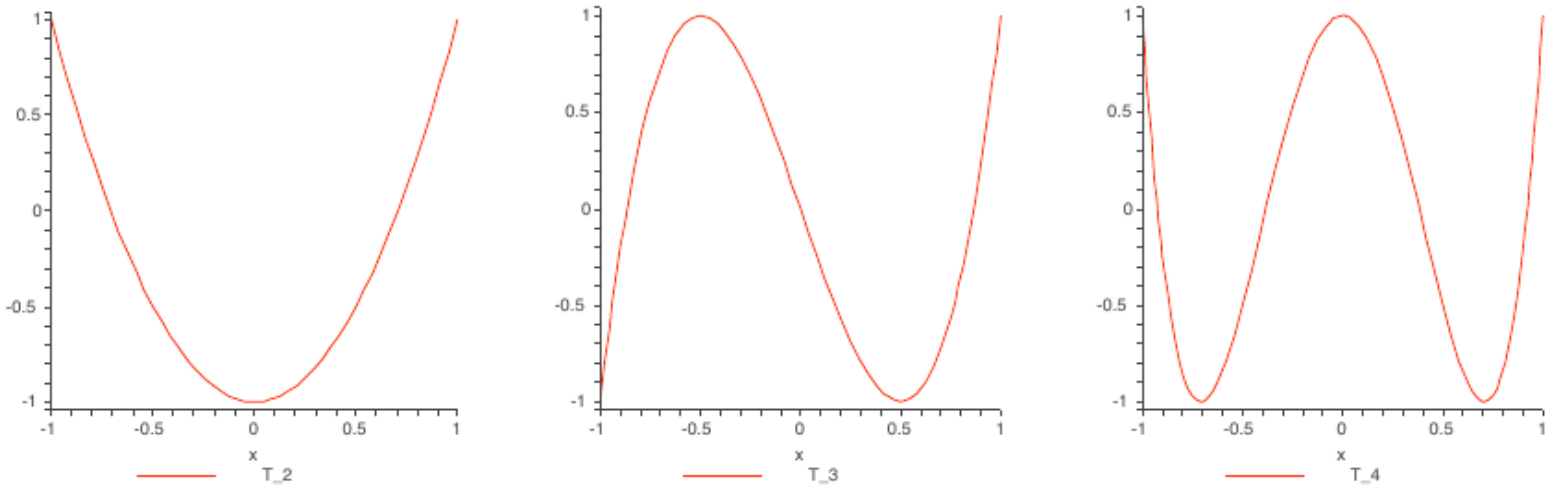


Abbildung 3.1: Tschebyscheff-Polynome T_2, T_3 und T_4.

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 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:

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

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 $|x| \leq 1$.

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:

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

also auch

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

Somit gilt für den Interpolationsfehler die Abschätzung

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

Zum Beispiel gilt für $n = 16$

$$|\log\left(1+\frac{1+x}{2}\right) - p(x)| \leq 10^{-11}.$$

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:

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

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:	$p = 4$!	$(s = 3)$	$(q = 5 : \ \frac{5}{24} \neq \frac{1}{5})$

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:

$$\begin{aligned} b_1 = b_3 &= \frac{5}{18} \\ b_2 &= \frac{4}{9}. \end{aligned}$$

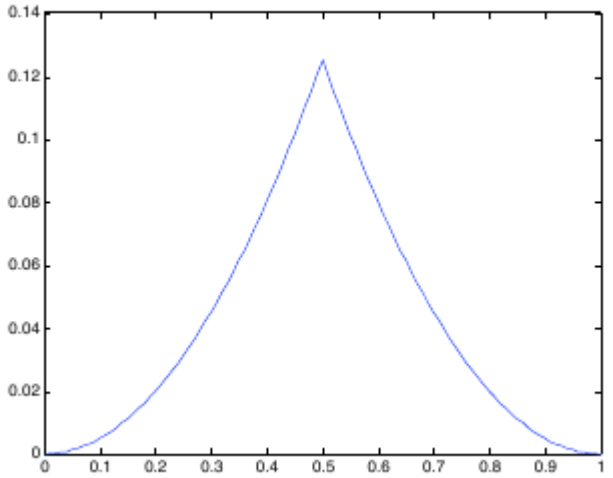
Die Gaußsche Quadraturformel der Ordnung 6 lautet also

$$\int_0^1 f(x)dx \approx \frac{5}{18} f\Big(\frac{1}{2} - \frac{\sqrt{15}}{10}\Big) + \frac{4}{9} f\Big(\frac{1}{2}\Big) + \frac{5}{18} f\Big(\frac{1}{2} + \frac{\sqrt{15}}{10}\Big).$$

Beispiel 22.

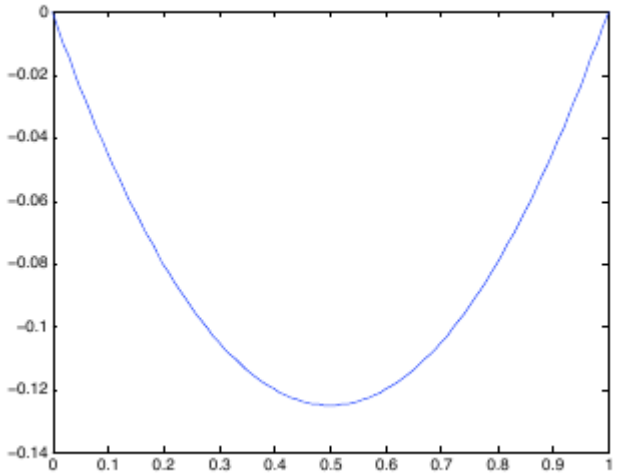
(i) Mittelpunktregel (Ordnung 2, $c_1 = \frac{1}{2}$, $b_1 = 1$)

$$K_2(t) = \frac{(1-t)^2}{2} - (\frac{1}{2} - t)_+ = \begin{cases} \frac{t^2}{2}, & \text{falls } 0 \leq t \leq \frac{1}{2} \\ \frac{(1-t)^2}{2}, & \text{falls } \frac{1}{2} < t \leq 1 \end{cases}$$



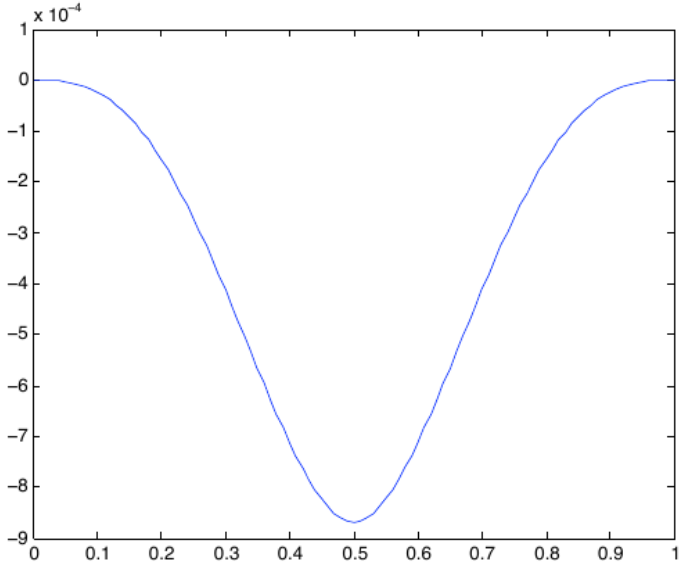
(ii) Trapezregel (Ordnung 2, $c_1 = 0, c_2 = 1,\ b_1 = b_2 = \frac{1}{2}$*)*

$$K_2(t) = -(1-t)\frac{t}{2}$$



(iii) Simpson-Regel (Ordnung 4, $c_1 = 0, c_2 = \frac{1}{2}, c_3 = 1,\ b_1 = b_3 = \frac{1}{6}, b_2 = \frac{2}{3}$*)*

$$K_4(t) = \begin{cases} \frac{t^4}{24} - \frac{t^3}{36}, & \text{falls } 0 \le t \le \frac{1}{2} \\ \frac{t^4}{24} - \frac{5}{36}t^3 + \frac{t^2}{6} - \frac{t}{12} + \frac{1}{72}, & \text{falls } \frac{1}{2} < t \le 1 \end{cases}$$



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:

$$\int_0^1 K_4(t)dt = -\frac{1}{2880}$$

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:*

$$\begin{array}{lcllcl} T(h_1) = T_{11} \approx \underline{2}.86 & & & & \\ & \searrow & & & \\ T(h_2) = T_{21} \approx \underline{2.7}54 & \rightarrow & T_{22} \approx \underline{2.718}9 & & \\ & \searrow & & \searrow & \\ T(h_3) = T_{31} \approx \underline{2.7}27 & \rightarrow & T_{32} \approx \underline{2.718}3 & \rightarrow & T_{33} \approx \underline{2.7182827} \end{array}$$

Für die Approximationsfehler gilt

$$\begin{array}{lcllcl} \varepsilon_{11} \approx 0.14086 & & & & \\ & \searrow & & & \\ \varepsilon_{21} \approx 0.03565 & \rightarrow & \varepsilon_{22} \approx 0.000579 & & \\ & \searrow & & \searrow & \\ \varepsilon_{31} \approx 0.00894 & \rightarrow & \varepsilon_{32} \approx 0.000037 & \rightarrow & \varepsilon_{33} \approx 8.599 \cdot 10^{-7} \end{array}$$

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*

$$
\begin{array}{lllllll}
\varepsilon_{11} \approx 0.54 & & & & & & \\
 & \searrow & & & & & \\
\varepsilon_{21} \approx 0.273 & \rightarrow & \varepsilon_{22} \approx 0.1835 & & & & \\
 & \searrow & & \searrow & & & \\
\varepsilon_{31} \approx 0.0776 & \rightarrow & \varepsilon_{32} \approx 0.0124 & \rightarrow & \varepsilon_{33} \approx 0.00100 & & \\
 & \searrow & & \searrow & & \searrow & \\
\varepsilon_{41} \approx 0.0199 & \rightarrow & \varepsilon_{42} \approx 0.000698 & \rightarrow & \varepsilon_{43} \approx 0.000083 & \rightarrow & \varepsilon_{44} \approx 0.0001
\end{array}
$$

Der "große" Fehler in T_{11} *bewirkt hier, dass die Näherung* T_{43} *besser ist als die Näherung* T_{44}.

Beispiel 26. *Das klassische Runge–Kutta Verfahren der Ordnung $p = 4$ lautet*

$$
\begin{array}{c|cccc}
0 & & & & \\
\frac{1}{2} & \frac{1}{2} & & & \\
\frac{1}{2} & 0 & \frac{1}{2} & & \\
1 & 0 & 0 & 1 & \\
\hline
 & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6}
\end{array}
$$