Übungen zur Vorlesung Numerik I
Sommersemester 2010

PD Dr. Thorsten Hüls
Dipl.-Math. Denny Otten

Übungsblatt 2
22.04.2010

Abgabe: Donnerstag, 29.04.2010, 10:00 Uhr in das Postfach des jeweiligen Tutors.
Mo.-Tutorium: Paul Voigt, paulvoigt@web.de, Postfach 195 in V3-128
Di.-Tutorium: Denny Otten, dotten@math.uni-bielefeld.de, Postfach 44 in V3-128
Mi.-Tutorium: Ingwar Petersen, ipeterse@math.uni-bielefeld.de, Postfach 227 in V3-128

Aufgabe 4: (Rundungsfehler der Gleitkommaarithmetik)
Sei $b \in \mathbb{N}$, $b \geqslant 2$ gerade, $G(\tau, b)$ die Menge der Gleitkommazahlen zur Basis b mit Mantissenlänge τ und sei $\mathrm{rd} : \mathbb{R} \to G(\tau, b)$ die übliche Rundungsvorschrift.
Man zeige für alle $t \in \mathbb{R}$:

$$|t - \mathrm{rd}(t)| \leqslant |t - g| \quad \forall\, g \in G(\tau, b). \tag{1}$$

Hinweis: Es empfiehlt sich, folgendermaßen vorzugehen:

(i) Zeigen Sie: Gilt (1) für $t > 0$, so auch für $t < 0$.

(ii) Wählen Sie für $t \in \mathbb{R}$ eine *geeignete* Darstellung der Form $t = b^n \sum_{j=1}^{\infty} \beta_j b^{-j}$.

(iii) Zeigen Sie: Zu jedem $g \in G(\tau, b)$ gibt es ein $g' \in G(\tau, b) \cap [b^{n-1}, \infty)$, so dass $|g' - t| \leq |g - t|$.

(iv) Zeigen Sie für $g, g' \in G(\tau, b) \cap [b^{n-1}, \infty)$ mit $g \neq g'$, dass $|g - g'| \geq b^{n-\tau}$.

(v) Beweisen Sie nun (1) mit Hilfe von Skript, Satz 2.3 für alle $g \in G(\tau, b) \cap [b^{n-1}, \infty)$ mit $g \neq \mathrm{rd}(t)$.

(6 Punkte)

Aufgabe 5: (Berechnung von Konditionszahlen)
(a) Berechnen Sie die relative komponentenweise Konditionszahl $\widehat{\kappa}(F, x)$ für die folgenden Auswertungsprobleme:

(i) $F : \mathbb{R} \to \mathbb{R}$ mit $F(x) = ax^n\sqrt{bx}$, $x > 0$, $a, b \in \mathbb{R}$ mit $b > 0$, $n \in \mathbb{N}_0$.

(ii) $F : \mathbb{R}^2 \to \mathbb{R}$ mit $F\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{ax_1^n}{bx_2^m}$, $x_2 \neq 0$, $a, b \in \mathbb{R}$ mit $b \neq 0$, $n, m \in \mathbb{N}$.

(iii) $F : \mathbb{R} \to \mathbb{R}^2$ mit $F(x) = \begin{pmatrix} \mathrm{sech}(x) \\ \tanh(x) \end{pmatrix}$

(b) Betrachten Sie die aus der Aufgabe 2 bekannte Iteration

$$x^{(n)} = f_{a,b}\left(x^{(n-1)}\right), \quad n = 1, 2, \ldots$$

mit

$$f_{a,b} : \mathbb{R}^2 \to \mathbb{R}^2 \quad \text{mit} \quad \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto f_{a,b}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} := \begin{pmatrix} ax_1 \\ -(b-a^2)x_1^2 + bx_2 \end{pmatrix}$$

und dem Anfangswert $x^{(0)} = \begin{pmatrix} c \\ c^2 \end{pmatrix}$.

(1) Bestimmen Sie die relative komponentenweise Konditionszahl $\widehat{\kappa}$ des Auswertungsproblems $\left(f_{a,b}, \begin{pmatrix} c \\ c^2 \end{pmatrix}\right)$. Setzen Sie nun $a = \frac{1}{2}$, $b = 2$, $c = 2$ und berechnen Sie die

(2) Kondition $\widetilde{\kappa}(0)$ der Gesamtiteration, also des Auswertungsproblems

$$\left(\underbrace{f \circ \cdots \circ f}_{n-mal}, \begin{pmatrix} c \\ c^2 \end{pmatrix}\right)$$

Hinweis: Vgl. Skript, Formel (2.22) sowie die explizite Darstellung der Folge $\left(x^{(n)}\right)_{n\in\mathbb{N}_0}$ aus Aufgabe 2.
Wie verhält sich für $n \to \infty$?

(6 Punkte)

Aufgabe 6: (Programmieraufgabe, Kondition von Auswertungsproblemen)
Man betrachte noch einmal das Beispiel zur Berechnung der Lösung $F(p,q) = -p + \sqrt{p^2+q}$ der quadratischen Gleichung $t^2 + 2pt - q = 0$ für $q > 0, p \in \mathbb{R}$.
Zwei Algorithmen sind gegeben durch

Algorithmus 1: $F^1(p,q) = (p, \sqrt{p^2+q})$, $F^2(y_1, y_2) = y_2 - y_1$.

Algorithmus 2: $F^1(p,q) = (p, q, \sqrt{p^2+q})$, $F^2(y_1, y_2, y_3) = \frac{y_2}{y_1+y_3}$.

In der Vorlesung wurden die Konditionszahlen bereits im Fall $p > 0$ untersucht. Bestimmen Sie nun für $p < 0 < q$

(1) (a) die relative Kondition des Auswertungsproblems $\widehat{\kappa}(F, (p,q))$.

(1) (b) die relativen Konditionszahlen $\widehat{\kappa}_1 = \widehat{\kappa}(F^1, (p,q))$ und $\widehat{\kappa}_2 = \widehat{\kappa}(F^2, F^1(p,q))$ zum Algorithmus 1.

(1) (c) die relativen Konditionszahlen $\widehat{\kappa}_1$ und $\widehat{\kappa}_2$ zum Algorithmus 2.

(1) Welcher Algorithmus ist in diesem Fall gutartig?

(1) (d) Berechnen Sie mit Hilfe der Rundungsfehler GUI die Werte der beiden Algorithmen zur Mantissenlänge 8 für $q = 1$, $p = -10, -100, -1000$. Interpretieren Sie die Ergebnisse.

(1)

(6 Punkte)

Aufgabe 4: (Rundungsfehler der Gleitkommaarithmetik)

Sei $b \in \mathbb{N}$ gerade mit $b \geq 2$, $G(\tau, b)$ die Menge der Gleitkommazahlen zur Basis b mit Mantissenlänge τ und $rd: \mathbb{R} \to G(\tau, b)$ die übliche Rundungsvorschrift. Dann gilt:

$$\forall t \in \mathbb{R}\ \forall g \in G(\tau, b): \quad |t - rd(t)| \leq |t - g| \tag{4.1}$$

Beweis:

1. **1. Fall:** $(t = 0)$. Trivialerweise gilt:

$$|t - rd(t)| \underset{\substack{\uparrow \\ t=0 \\ rd(0)=0}}{=} 0 \leq |g| \underset{\substack{\uparrow \\ t=0}}{=} |t - g| \quad \forall g \in G(\tau, b)$$

2. Fall: $(t < 0)$. Angenommen die Aussage (4.1) sei für $t > 0$ bereits gezeigt, d.h. es gilt

$$\forall t \in \mathbb{R} \text{ mit } t > 0\ \forall g \in G(\tau, b): \quad |t - rd(t)| \leq |t - g|, \tag{4.2}$$

dann erhalten wir aus (4.2), „$g \in G(\tau, b) \Rightarrow -g \in G(\tau, b)$" und $rd(t) = -rd(-t)$, dann

$$rd(t) = \begin{cases} b^n (0, \beta_1 \beta_2 \cdots \beta_\tau) & , \text{falls } 0 \leq \beta_{\tau+1} \leq \frac{b}{2} - 1 \\ b^n \left[(0, \beta_1 \beta_2 \cdots \beta_\tau) + b^{-\tau} \right] & , \text{falls } \frac{b}{2} \leq \beta_{\tau+1} \leq b - 1 \end{cases}$$

$$= - \begin{cases} -b^n (0, \beta_1 \beta_2 \cdots \beta_\tau) & , \text{falls } 0 \leq \beta_{\tau+1} \leq \frac{b}{2} - 1 \\ -b^n \left[(0, \beta_1 \beta_2 \cdots \beta_\tau) + b^{-\tau} \right] & , \text{falls } \frac{b}{2} \leq \beta_{\tau+1} \leq b - 1 \end{cases}$$

$$= -rd(-t) \qquad \forall t \in \mathbb{R}$$

die Behauptung

$$|t - rd(t)| \underset{\substack{\uparrow \\ rd(t) = -rd(-t)}}{=} |t + rd(-t)| \underset{\substack{\uparrow \\ |x| = |-x|}}{=} |(-t) - rd(-t)|$$

$$\underset{\substack{\uparrow \\ t<0 \Rightarrow -t>0 \\ \& (4.2) \& \text{„} g \in G(\tau,b) \Rightarrow -g \in G(\tau,b)\text{"}}}{\leq} |(-t) - (-g)| = |g - t| \underset{\substack{\uparrow \\ |x| = |-x|}}{=} |t - g| \qquad \forall t \in \mathbb{R} \text{ mit } t < 0\ \forall g \in G(\tau, b)$$

Damit ist die Aufgabe gelöst, insofern wir (4.2) gezeigt haben. (d.h. wir behandeln nun den Fall $t > 0$).

2. Sei $t \in \mathbb{R}$ mit $t > 0$ beliebig, dann gilt zunächst

$$\exists_1\, n \in \mathbb{Z}: \quad t \in [b^{n-1}, b^n[. \tag{4.3}$$

Weiter können wir t darstellen als (siehe Skript (2.2))

$$t = b^n \sum_{j=1}^{\infty} \beta_j b^{-j}, \quad \beta_j \in \{0, \ldots, b-1\},\ \beta_1 \neq 0,\ j \in \mathbb{N} \tag{4.4}$$

Für die Eindeutigkeit der Gleitkommadarstellung (4.4) genügt es, wenn wir $\beta_j \neq b-1$ für unendlich viele $j \in \mathbb{N}$ fordern. Fortsetzung des Beweises folgt in Schritt 5.

3. **Hilfsaussage 1:** Sei $t \in \mathbb{R}$ mit $t > 0$ und $t = b^n \sum_{j=1}^{\infty} \beta_j b^{-j}$ wie in (4.4). Dann gilt

$$\forall g \in G(\tau, b)\ \exists g' \in G(\tau, b) \cap [b^{n-1}, \infty[: \quad |g' - t| \leq |g - t|.$$

Beweis:

1. Fall: $(g \geq b^{n-1})$. Wähle $g' := g$, dann gilt wegen $g \geq b^{n-1}$, dass $g' \in G(\tau, b) \cap [b^{n-1}, \infty[$ und es folgt

$$|g - t| \underset{\substack{\uparrow \\ \text{Def. von } g'}}{=} |g' - t|$$

2. Fall: $(g < b^{n-1})$. Wähle $g' := b^{n-1}$, dann gilt $g' \in G(\tau, b) \cap [b^{n-1}, \infty[$. Daraus sowie aus der Darstellung von t erhalten wir

$$g < b^{n-1} = g' \leq t \quad \Rightarrow g - t < g' - t \leq 0$$

$$\Rightarrow |g - t| > |g' - t|$$

■

4. HILFSAUSSAGE 2:

$$\forall g, g' \in G(r,b) \cap [b^{n-1}, \infty[\text{ mit } g \neq g' : |g - g'| \geq b^{n-r}$$

Beweis:

Seien $g, g' \in G(r,b) \cap [b^{n-1}, \infty[$ mit $g \neq g'$, d.h.

$$g = b^{m_1} \sum_{j=1}^{r} \alpha_j^1 b^{-j}, \quad m_1 \geq n, \ \alpha_j^1 \in \{0, \dots, b-1\}, \ \alpha_1^1 \neq 0, \ j = 1, \dots, r$$

$$g' = b^{m_2} \sum_{j=1}^{r} \alpha_j^2 b^{-j}, \quad m_2 \geq n, \ \alpha_j^2 \in \{0, \dots, b-1\}, \ \alpha_1^2 \neq 0, \ j = 1, \dots, r$$

Offensichtlich gilt nun $g \cdot b^{r-n} \in \mathbb{Z}$ und $g' \cdot b^{r-n} \in \mathbb{Z}$, denn

$$g \cdot b^{r-n} = \underbrace{b^{m_1 - n}}_{\in \mathbb{N}} \cdot \sum_{j=1}^{r} \underbrace{\alpha_j^1}_{\in \{0,\dots,b-1\} \subset \mathbb{N}} \underbrace{b^{r-j}}_{\in \mathbb{N}}$$

($b^{m_1-n} \in \mathbb{N}$: denn $b \in \mathbb{N}$ mit $b \geq 2$ & $m_1 - n \in \mathbb{Z}$ mit $m_1 - n \geq 0$; $b^{r-j} \in \mathbb{N}$: denn $b \in \mathbb{N}$ mit $b \geq 2$ & $r - j \in \mathbb{N}_0$, $j = 1, \dots, r$)

$$\Rightarrow g \cdot b^{r-n} \in \mathbb{Z} \quad (\text{sogar } \mathbb{N}_0)$$

($\mathbb{N}$ abgeschlossen bzgl. $+$ und $\cdot$)

$g' \cdot b^{r-n}$: analog

und daraus erhalten wir

$$|g - g'| \cdot b^{r-n} \geq 1.$$

(↑ Zeile zuvor & $g \neq g'$)

Multiplikation mit b^{n-r} liefert die Behauptung

$$|g - g'| \geq b^{n-r}. \quad \blacksquare$$

5. Für die Darstellung von t in (4.4) gilt nach der Definition der Rundungsvorschrift rd

$$rd(t) \in G(r,b) \cap [b^{n-1}, \infty[. \qquad (4.5)$$

Sei nun $g \in G(r,b) \cap [b^{n-1}, \infty[$ mit $g \neq rd(t)$, dann sind wegen (4.5) die Voraussetzungen von Hilfsaussage 2 erfüllt und es gilt

$$|g - rd(t)| \geq b^{n-r}. \qquad (4.6)$$

Andererseits liefert Satz 2.3 aus der Vorlesung

$$|t - rd(t)| \leq \frac{1}{2} \cdot b^{n-r} \qquad (4.7)$$

(↑ Beweis v. Satz 2.3)

Aus (4.6) und (4.7) erhalten wir insgesamt

$$b^{n-r} \underset{(4.6)}{\leq} |g - rd(t)| \underset{\text{Ergänzen \& } \Delta\text{'s-Ungl.}}{\leq} |g - t| + |t - rd(t)| \underset{(4.7)}{\leq} |g - t| + \frac{1}{2} \cdot b^{n-r} \qquad (4.8)$$

$$\Rightarrow |t - rd(t)| \underset{(4.7)}{\leq} \frac{1}{2} \cdot b^{n-r} \underset{(4.6)}{\leq} \frac{1}{2} |g - rd(t)| \underset{\text{Ergänzen \& } \Delta\text{'s-Ungl.}}{\leq} \frac{1}{2} \left(|g - t| + |t - rd(t)| \right) = \frac{1}{2} |g - t| + \frac{1}{2} |t - rd(t)|$$

$(\cdot 2)$, $-\frac{1}{2}|t - rd(t)|$

$$\Rightarrow |t - rd(t)| \leq |g - t| \quad \forall g \in G(r,b) \cap [b^{n-1}, \infty[\qquad (4.9)$$

Sei jetzt $g \in G(r,b)$ beliebig, dann gilt nach Hilfsaussage 1

$$\exists g' \in G(r,b) \cap [b^{n-1}, \infty[: |g' - t| \leq |g - t| \qquad (4.10)$$

Daraus und aus (4.9) mit g' erhalten wir

$$|t - rd(t)| \underset{(4.9)}{\leq} |g' - t| \underset{(4.10)}{\leq} |g - t|$$

und es folgt (4.2) und somit (4.1). $\blacksquare$

Aufgabe 5: zu (a):

Sei $F: \mathbb{R}^n \to \mathbb{R}^m$. Wir berechnen in dieser Aufgabe die relative komponentenweise Kondition des Auswertungsproblems (F, x), die durch (siehe: Skript (2.17))

$$\hat{\kappa}(F,x) := \lim_{\delta \to 0} \|L(\delta)\|_\infty = \max_{i=1,\dots,m} \sum_{j=1}^{n} \left| \frac{\partial F_i}{\partial x_j}(x) \right| \cdot \frac{|x_j|}{|F_i(x)|} \tag{5.1}$$

gegeben ist.

zu (i): Betrachte

$$F: \mathbb{R}_+^* \to \mathbb{R} \text{ mit } F(x) = a \cdot x^n \cdot \sqrt{b \cdot x} = a \cdot b^{\frac{1}{2}} \cdot x^{n+\frac{1}{2}}, \quad \mathbb{R}_+^* := \{x \in \mathbb{R} \mid x > 0\}$$

wobei $a, b \in \mathbb{R}$ mit $b > 0$ und $n \in \mathbb{N}_0$. Wegen

$$\frac{\partial F}{\partial x}(x) = a \cdot b^{\frac{1}{2}} \cdot (n+\tfrac{1}{2}) \cdot x^{n-\frac{1}{2}} \tag{5.2}$$

erhalten wir aus (5.1) die Kondition

$$\hat{\kappa}(F,x) \underset{(5.1)}{=} \left| \frac{\partial F}{\partial x}(x) \right| \cdot \frac{|x|}{|F(x)|} \underset{(5.2)}{=} \left| a \cdot b^{\frac{1}{2}} \cdot (n+\tfrac{1}{2}) \cdot x^{n-\frac{1}{2}} \right| \cdot \frac{|x|}{|a \cdot b^{\frac{1}{2}} \cdot x^{n+\frac{1}{2}}|}$$

$$= |n+\tfrac{1}{2}| = n+\tfrac{1}{2}.$$

zu (ii): Betrachte

$$F: \mathbb{R}^2 \to \mathbb{R} \text{ mit } F\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{a \cdot x_1^n}{b \cdot x_2^m} = a \cdot b^{-1} \cdot x_1^n \cdot x_2^{-m}, \quad x_2 \neq 0$$

wobei $a, b \in \mathbb{R}$ mit $b \neq 0$ und $n, m \in \mathbb{N}$. Wegen

$$\frac{\partial F}{\partial x_1}(x) = a \cdot b^{-1} \cdot n \cdot x_1^{n-1} \cdot x_2^{-m} \tag{5.3}$$

$$\frac{\partial F}{\partial x_2}(x) = a \cdot b^{-1} \cdot (-m) \cdot x_1^n \cdot x_2^{-(m+1)}$$

erhalten wir aus (5.1) die Kondition

$$\hat{\kappa}(F,x) \underset{(5.1)}{=} \sum_{j=1}^{2} \left| \frac{\partial F}{\partial x_j}(x) \right| \cdot \frac{|x_j|}{|F(x)|}$$

$$\underset{(5.3)}{=} \left| a \cdot b^{-1} \cdot n \cdot x_1^{n-1} \cdot x_2^{-m} \right| \cdot \frac{|x_1|}{|a \cdot b^{-1} \cdot x_1^n \cdot x_2^{-m}|} + \left| a \cdot b^{-1} \cdot (-m) \cdot x_1^n \cdot x_2^{-(m+1)} \right| \cdot \frac{|x_2|}{|a \cdot b^{-1} \cdot x_1^n \cdot x_2^{-m}|}$$

$$= |n| + |-m| = |n| + |m| = n + m.$$

zu (b): Betrachte die Funktion $F := f_{a,b}$ aus Aufgabe 2

$$F: \mathbb{R}^2 \to \mathbb{R}^2 \text{ mit } F\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a \cdot x_1 \\ -(b-a^2)x_1^2 + b x_2 \end{pmatrix} =: \begin{pmatrix} F_1\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ F_2\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \end{pmatrix}$$

wobei $a, b \in \mathbb{R}$. Wegen

$$\frac{\partial F_1}{\partial x_1}(x) = a \qquad \frac{\partial F_2}{\partial x_1}(x) = -2(b-a^2)x_1 \tag{5.5}$$

$$\frac{\partial F_1}{\partial x_2}(x) = 0 \qquad \frac{\partial F_2}{\partial x_2}(x) = b$$

erhalten wir aus (5.1) die Kondition

$$\hat{\kappa}(F,x) \underset{(5.1)\,(5.5)}{=} \max\left\{ |a| \cdot \frac{|x_1|}{|a \cdot x_1|} + |0| \cdot \frac{|x_2|}{|a \cdot x_1|}, \right.$$

$$\left. |-2(b-a^2)x_1| \cdot \frac{|x_1|}{|-(b-a^2)x_1^2 + b x_2|} + |b| \cdot \frac{|x_2|}{|-(b-a^2)x_1^2 + b x_2|} \right\}$$

$$= \max\left\{1, \frac{|-2(b-a^2)x_1^2| + |b x_2|}{|-(b-a^2)x_1^2 + b x_2|}\right\} = \frac{|-2(b-a^2)x_1^2| + |b x_2|}{|-(b-a^2)x_1^2 + b x_2|}.$$

$$|-(b-a^2)x_1^2 + b x_2| \leq |-2(b-a^2)x_1^2| + |b x_2| \quad (\Delta\text{'s Ungl.})$$

$$\Rightarrow 1 \leq \frac{|-2(b-a^2)x_1^2| + |b x_2|}{|-(b-a^2)x_1^2 + b x_2|} \tag{5.6}$$

→ Fortsetzung folgt nach (a)(iii)

Zu (iii): Betrachte die Funktion

$$F : \mathbb{R} \to \mathbb{R}^2 \quad \text{mit} \quad F(x) = \begin{pmatrix} \operatorname{sech}(x) \\ \tanh(x) \end{pmatrix} =: \begin{pmatrix} F_1(x) \\ F_2(x) \end{pmatrix}$$

Wegen

$$\frac{\partial F_1}{\partial x}(x) = -\operatorname{sech}(x)\cdot\tanh(x)$$

$$\frac{\partial F_2}{\partial x}(x) = 1-\tanh^2(x) \qquad (5.4)$$

erhalten wir aus (5.1) die Kondition

$$\hat{\kappa}(F,x) \underset{(5.1)}{=} \max\left\{ \left|\frac{\partial F_1}{\partial x}(x)\right| \cdot \frac{|x|}{|F_1(x)|}, \left|\frac{\partial F_2}{\partial x}(x)\right| \cdot \frac{|x|}{|F_2(x)|} \right\}$$

$$\underset{(5.4)}{=} \max\left\{ \frac{|1-\operatorname{sech}(x)\cdot\tanh(x)|\cdot|x|}{|\operatorname{sech}(x)|}, \frac{|1-\tanh^2(x)|\cdot|x|}{|\tanh(x)|} \right\}$$

$$= \max\left\{ |x\cdot\tanh(x)|, \left|\frac{x(1-\tanh^2(x))}{\tanh(x)}\right| \right\}$$

$$= \begin{cases} |x\cdot\tanh(x)| & , x \in]-\infty, -\operatorname{arctanh}(\frac{\sqrt{2}}{2})] \cup [\operatorname{arctanh}(\frac{\sqrt{2}}{2}), +\infty[\\ \left|\frac{x(1-\tanh^2(x))}{\tanh(x)}\right| & , x \in [\operatorname{arctanh}(\frac{\sqrt{2}}{2}), \operatorname{arctanh}(\frac{\sqrt{2}}{2})] \end{cases}$$

Fortsetzung zu (b):

Aus der Kondition des Auswertungsproblems (F,x), die in (5.6) berechnet wurde, erhalten wir die Kondition des Auswertungsproblems $(F, \binom{c}{c^2})$, indem $\binom{x_1}{x_2} = \binom{c}{c^2}$ gesetzt wird:

$$\hat{\kappa}\left(F_1, \binom{c}{c^2}\right) \underset{(5.6)}{=} \frac{|(b-a^2)c^2| + |bc^2|}{|-(b-a^2)c^2 + bc^2|}$$

$$= \frac{|(b-a^2)c^2| + |bc^2|}{|a^2c^2|} = \frac{|b-a^2| + |b|}{|a^2|} \qquad (5.7)$$

Setze nun $a=\frac{1}{2}$, $b=2$ und $c=2$. Zur Berechnung der Kondition $\tilde{\kappa}(0)$ der Gesamtiteration wissen wir aus Aufgabe 2, Blatt 1, dass sich $x^{(n)}$ explizit darstellen lässt durch

$$x^{(n)} = f(x^{(n-1)}) = (f\circ\cdots\circ f)(x^{(0)}) \underset{x^{(0)}=\binom{c}{c^2}}{\underset{\text{Aufgabe 2, Blatt 1}}{=}} \begin{pmatrix} a^n\cdot c \\ a^{2n}\cdot c^2 \end{pmatrix} \underset{c=2}{\overset{a=\frac{1}{2}}{=}} \begin{pmatrix} (\frac{1}{2})^n\cdot 2 \\ (\frac{1}{2})^{2n}\cdot 2^2 \end{pmatrix}$$

$$= \begin{pmatrix} (\frac{1}{2})^{n-1} \\ (\frac{1}{2})^{2(n-1)} \end{pmatrix}, \quad n \in \mathbb{N} \qquad (5.8)$$

Nach der Konditionsformel (2.22) im Skript, gilt für die Kondition der Gesamtiteration

$$\tilde{\kappa}(0) \underset{\text{Skript (2.22)}}{=} 1 + \sum_{j=1}^{n} \prod_{\nu=j}^{n} \hat{\kappa}_\nu$$

mit

$$\hat{\kappa}_n := \hat{\kappa}(F, x^{(n)}) \underset{(5.8)}{=} \hat{\kappa}\left(F, \begin{pmatrix} (\frac{1}{2})^{n-1} \\ (\frac{1}{2})^{2(n-1)} \end{pmatrix}\right)$$

$$\underset{\substack{(5.7)\\ a=\frac{1}{2}\\ b=2}}{=} \frac{\left|2(\frac{7}{4})\cdot(\frac{1}{2})^{2(n-1)}\right| + \left|(\frac{1}{2})^{2(n-1)}\cdot 2\right|}{\left|(\frac{1}{2})^2\cdot(\frac{1}{2})^{2(n-1)}\right|} = \frac{2\cdot\frac{7}{4}+2}{\frac{1}{4}} = \frac{\frac{22}{4}}{\frac{1}{4}} = 22 \quad n\in\mathbb{N}$$

und demzufolge

$$\tilde{\kappa}(0) = 1 + \sum_{j=1}^{n} \prod_{\nu=j}^{n} 22 \xrightarrow{n\to\infty} \infty$$

$$= 1 + \sum_{j=1}^{n} 22^j = \frac{22^{n+1}-1}{22-1}$$

Aufgabe 6:

Zu (a): Betrachte

$$F : \mathbb{R}^2 \longrightarrow \mathbb{R} \quad \text{mit} \quad F\begin{pmatrix} p \\ q \end{pmatrix} = -p + \sqrt{p^2+q} \tag{6.1}$$

wobei $p, q \in \mathbb{R}$ mit $p < 0 < q$. Wegen

$$\frac{\partial F}{\partial p}(p,q) \underset{(6.1)}{=} -1 + \tfrac{1}{2}(p^2+q)^{-\frac{1}{2}} \cdot 2p = -1 + \frac{p}{\sqrt{p^2+q}} = \frac{p - \sqrt{p^2+q}}{\sqrt{p^2+q}}$$

$$\frac{\partial F}{\partial q}(p,q) \underset{(6.1)}{=} \tfrac{1}{2}(p^2+q)^{-\frac{1}{2}} = \frac{1}{2\sqrt{p^2+q}} \tag{6.2}$$

erhalten wir aus (5.1) $\left(\text{mit } \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} p \\ q \end{pmatrix}\right)$ die relative komponentenweise Kondition

$$\hat{\kappa}(F,x) \underset{(5.1)}{=} \left|\frac{\partial F}{\partial p}(p,q)\right| \cdot \frac{|p|}{|F(p,q)|} + \left|\frac{\partial F}{\partial q}(p,q)\right| \cdot \frac{|q|}{|F(p,q)|}$$

$$\underset{(6.2)}{=} \left|\frac{1}{\sqrt{p^2+q}}\right| \cdot \left[\frac{|p - \sqrt{p^2+q}| \cdot |p|}{|-p + \sqrt{p^2+q}|} + \left|\tfrac{1}{2}\right| \cdot \frac{|q|}{|-p+\sqrt{p^2+q}|}\right]$$

$$\underset{\substack{p<0 \\ q>0}}{=} \frac{1}{\sqrt{p^2+q}} \cdot \left[\frac{(-p+\sqrt{p^2+q}) \cdot (-p) \cdot 2 \; + \; q}{2 \cdot (-p+\sqrt{p^2+q})}\right]$$

$$= \frac{1}{\sqrt{p^2+q}} \cdot \left[-p + \frac{q}{2(-p+\sqrt{p^2+q})}\right]$$

$$= \frac{1}{\sqrt{p^2+q}} \cdot \left[-p + \frac{q \cdot (p+\sqrt{p^2+q})}{2 \cdot q}\right] \qquad \text{(denn: } (a+b)(a-b) = a^2 - b^2 \text{ 3. binomische Formel)}$$

$$\cdot \underbrace{\left(\frac{p+\sqrt{p^2+q}}{p+\sqrt{p^2+q}}\right)}_{=1}$$

$$= \frac{1}{2} \cdot \underbrace{\frac{-p}{\sqrt{p^2+q}}}_{\leq 1} + \frac{1}{2} \leq \frac{1}{2} + \frac{1}{2} = 1$$

Begründung: $p < 0 \Rightarrow p^2 > 0$ und $-p > 0$

$q > 0 \Rightarrow p^2 + q \geq p^2 \geq 0$ (↑ $q>0$, ↑ Zeile zuvor)

$\sqrt{\cdot}$ streng mon. wachsend $\Rightarrow \sqrt{p^2+q} > \sqrt{p^2} = -p > 0$

$$\Rightarrow 1 = \frac{\sqrt{p^2+q}}{\sqrt{p^2+q}} > \frac{-p}{\sqrt{p^2+q}} > 0$$

Bemerkung: Mit der Kondition für $p > 0$, die bereits in der Vorlesung gezeigt wurde erhalten wir insgesamt:

$$\hat{\kappa}(F,x) \begin{cases} \leq 1, & p \leq 0 \\ \leq 2, & p > 0 \end{cases}$$

Zu (b): Betrachte

$$F^1 : \mathbb{R}^2 \longrightarrow \mathbb{R}^2 \quad \text{mit} \quad F^1\begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} p \\ \sqrt{p^2+q} \end{pmatrix} =: \begin{pmatrix} F_1^1(p,q) \\ F_2^1(p,q) \end{pmatrix}$$

$$F^2 : \mathbb{R}^2 \longrightarrow \mathbb{R} \quad \text{mit} \quad F^2\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = y_2 - y_1$$

wobei $p, q \in \mathbb{R}$ mit $p < 0 < q$. Wegen

$$\frac{\partial F_1^1}{\partial p}(p,q) = 1, \quad \frac{\partial F_1^1}{\partial q}(p,q) = 0, \quad \frac{\partial F_2^1}{\partial p}(p,q) = \frac{p}{\sqrt{p^2+q}}, \quad \frac{\partial F_2^1}{\partial q}(p,q) = \frac{1}{2\sqrt{p^2+q}} \tag{6.3}$$

erhalten wir aus (5.1)

$$\hat{\kappa}_1 = \hat{\kappa}(\mathcal{F}^1, \binom{p}{q}) \underset{(5.1)}{=} \max\left\{ \left|\frac{\partial \mathcal{F}_1^1}{\partial p}(p,q)\right| \cdot \frac{|p|}{|\mathcal{F}_1^1(p,q)|} + \left|\frac{\partial \mathcal{F}_1^1}{\partial q}(p,q)\right| \cdot \frac{|q|}{|\mathcal{F}_1^1(p,q)|}, \right.$$
$$\left. \left|\frac{\partial \mathcal{F}_2^1}{\partial p}(p,q)\right| \cdot \frac{|p|}{|\mathcal{F}_2^1(p,q)|} + \left|\frac{\partial \mathcal{F}_2^1}{\partial q}(p,q)\right| \cdot \frac{|q|}{|\mathcal{F}_2^1(p,q)|} \right\}$$
$$\underset{(6.3)}{=} \max\left\{ \underbrace{|1| \cdot \frac{|p|}{|p|}}_{=1} + \underbrace{|0| \cdot \frac{|q|}{|p|}}_{=0}, \left|\frac{p}{\sqrt{p^2+q}}\right| \cdot \frac{|p|}{|\sqrt{p^2+q}|} + \left|\frac{1}{2\sqrt{p^2+q}}\right| \cdot \frac{|q|}{|\sqrt{p^2+q}|} \right\}$$
$$= \max\left\{ 1, \frac{p^2}{p^2+q} + \frac{q}{2(p^2+q)} \right\}$$
$$= \max\left\{ 1, \underbrace{\frac{2p^2+q}{2p^2+2q}}_{<1} \right\} = 1$$

(denn: $q \leq 2q$, da $q > 0$)

Weiter erhalten wir wegen

$$\frac{\partial \mathcal{F}^2}{\partial y_1}(y_1,y_2) = -1, \quad \frac{\partial \mathcal{F}^2}{\partial y_2}(y_1,y_2) = 1 \tag{6.4}$$

die Konditionszahl

$$\hat{\kappa}(\mathcal{F}^2, \binom{y_1}{y_2}) \underset{(5.1)}{=} \left|\frac{\partial \mathcal{F}^2}{\partial y_1}(y_1,y_2)\right| \cdot \frac{|y_1|}{|\mathcal{F}^2(y_1,y_2)|} + \left|\frac{\partial \mathcal{F}^2}{\partial y_2}(y_1,y_2)\right| \cdot \frac{|y_2|}{|\mathcal{F}^2(y_1,y_2)|}$$
$$\underset{(6.4)}{=} |-1| \cdot \frac{|y_1|}{|y_2-y_1|} + |1| \cdot \frac{|y_2|}{|y_2-y_1|}$$
$$= \frac{|y_1|+|y_2|}{|y_2-y_1|} \tag{6.5}$$

und somit

$$\hat{\kappa}_2 := \hat{\kappa}(\mathcal{F}^2, \mathcal{F}^1(\binom{p}{q})) \underset{\uparrow}{=} \frac{|p|+|\sqrt{p^2+q}|}{|-p+\sqrt{p^2+q}|} \underset{\uparrow}{=} \frac{-p+\sqrt{p^2+q}}{-p+\sqrt{p^2+q}} = 1$$

(6.5) mit $\binom{y_1}{y_2} = \mathcal{F}^1(\binom{p}{q}) = \binom{p}{\sqrt{p^2+q}}$ $\quad p<0 \Rightarrow -p>0 \Rightarrow |p| = -p$

Bemerkung:

$$\hat{\kappa}_1 = \hat{\kappa}(\mathcal{F}^1, \binom{p}{q}) = \begin{cases} 1, & p \leq 0 \quad \text{(siehe Aufgabe)} \\ 1, & p \geq 0 \quad \text{(siehe Vorlesung)} \end{cases}$$

$$\hat{\kappa}_2 = \hat{\kappa}(\mathcal{F}^2, \mathcal{F}^1(\binom{p}{q})) = \begin{cases} 1, & p \leq 0 \quad \text{(siehe Aufgabe)} \\ \geq \frac{4p^2}{q}, & p \geq 0 \quad \text{(siehe Vorlesung)} \end{cases}$$

Zu (c): Betrachte

$$\mathcal{F}^1: \mathbb{R}^2 \longrightarrow \mathbb{R}^3 \quad \text{mit} \quad \mathcal{F}^1\binom{p}{q} = \begin{pmatrix} p \\ q \\ \sqrt{p^2+q} \end{pmatrix}$$

$$\mathcal{F}^2: \mathbb{R}^3 \longrightarrow \mathbb{R} \quad \text{mit} \quad \mathcal{F}^2\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \frac{y_2}{y_1+y_3}$$

wobei $p,q \in \mathbb{R}$ mit $p<0<q$. Wegen

$$\frac{\partial \mathcal{F}_1^1}{\partial p}(p,q) = 1, \quad \frac{\partial \mathcal{F}_2^1}{\partial p}(p,q) = 0, \quad \frac{\partial \mathcal{F}_3^1}{\partial p}(p,q) = \frac{p}{\sqrt{p^2+q}}$$
$$\frac{\partial \mathcal{F}_1^1}{\partial q}(p,q) = 0, \quad \frac{\partial \mathcal{F}_2^1}{\partial q}(p,q) = 1, \quad \frac{\partial \mathcal{F}_3^1}{\partial q}(p,q) = \frac{1}{2\sqrt{p^2+q}} \tag{6.6}$$

erhalten wir aus (5.1)

$$\hat{K}_1 := \hat{K}(F^1, \binom{p}{q}) \underset{(5.1)}{=} \max\Bigg\{ \left|\frac{\partial F_1^1}{\partial p}(p,q)\right| \cdot \frac{|p|}{|F_1^1(p,q)|} + \left|\frac{\partial F_1^1}{\partial q}(p,q)\right| \cdot \frac{|q|}{|F_1^1(p,q)|} ,$$
$$\left|\frac{\partial F_2^1}{\partial p}(p,q)\right| \cdot \frac{|p|}{|F_2^1(p,q)|} + \left|\frac{\partial F_2^1}{\partial q}(p,q)\right| \cdot \frac{|q|}{|F_2^1(p,q)|} ,$$
$$\left|\frac{\partial F_3^1}{\partial p}(p,q)\right| \cdot \frac{|p|}{|F_3^1(p,q)|} + \left|\frac{\partial F_3^1}{\partial q}(p,q)\right| \cdot \frac{|q|}{|F_3^1(p,q)|} \Bigg\}$$

$$\underset{(6.8)}{=} \max\Bigg\{ \underbrace{|1| \cdot \frac{|p|}{|p|}}_{=1} + \underbrace{|0| \cdot \frac{|q|}{|p|}}_{=0} , \underbrace{|0| \cdot \frac{|p|}{|q|}}_{=0} + \underbrace{|1| \cdot \frac{|q|}{|q|}}_{=1} ,$$
$$\left|\frac{p}{\sqrt{p^2+q}}\right| \cdot \frac{|p|}{|\sqrt{p^2+q}|} + \left|\frac{1}{2\sqrt{p^2+q}}\right| \cdot \frac{|q|}{|\sqrt{p^2+q}|} \Bigg\}$$

$$= \max\Big\{1, 1, \underbrace{\frac{2p^2+q}{2p^2+2q}}_{<1 \text{ (denn: } q \overset{q>0}{<} 2q)} \Big\} = 1$$

Weiter erhalten wir wegen

$$\frac{\partial F^2}{\partial y_1}(y_1,y_2,y_3) = \frac{-y_2}{(y_1+y_3)^2} , \quad \frac{\partial F^2}{\partial y_2}(y_1,y_2,y_3) = \frac{1}{y_1+y_3} , \quad \frac{\partial F^2}{\partial y_3}(y_1,y_2,y_3) = \frac{-y_2}{(y_1+y_3)^2} \quad (6.7)$$

die Konditionszahl

$$\hat{K}(F^2, \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}) \underset{(5.1)}{=} \left|\frac{\partial F^2}{\partial y_1}(y_1,y_2,y_3)\right| \cdot \frac{|y_1|}{|F^2(y_1,y_2,y_3)|} + \left|\frac{\partial F^2}{\partial y_2}(y_1,y_2,y_3)\right| \cdot \frac{|y_2|}{|F^2(y_1,y_2,y_3)|}$$
$$+ \left|\frac{\partial F^2}{\partial y_3}(y_1,y_2,y_3)\right| \cdot \frac{|y_3|}{|F^2(y_1,y_2,y_3)|}$$

$$\underset{(6.7)}{=} \left|\frac{-y_2}{(y_1+y_3)^2}\right| \cdot \frac{|y_1| \cdot |y_1+y_3|}{|y_2|} + \left|\frac{1}{y_1+y_3}\right| \cdot \frac{|y_2| \cdot |y_1+y_3|}{|y_2|}$$
$$+ \left|\frac{-y_2}{(y_1+y_3)^2}\right| \cdot \frac{|y_3| \cdot |y_1+y_3|}{|y_2|}$$

$$= \left|\frac{y_1}{y_1+y_3}\right| + 1 + \left|\frac{y_3}{y_1+y_3}\right| = 1 + \frac{|y_1|+|y_3|}{|y_1+y_3|} \quad (6.8)$$

und somit

$$\hat{K}_2 := \hat{K}(F^2, F^1(\binom{p}{q})) \underset{\uparrow}{=} 1 + \frac{|p| + |\sqrt{p^2+q}|}{|p+\sqrt{p^2+q}|} \underset{\uparrow}{=} 1 + \frac{-p+\sqrt{p^2+q}}{|p+\sqrt{p^2+q}|}$$

(6.8) mit $\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = F^1(\binom{p}{q}) = \begin{pmatrix} p \\ q \\ \sqrt{p^2+q} \end{pmatrix}$ $\qquad p<0$

$$= 1 + \frac{(-p+\sqrt{p^2+q})^2}{|q|} = 1 + \frac{(-p+\sqrt{p^2+q})^2}{q}$$

$\uparrow$ $\cdot \left(\frac{-p+\sqrt{p^2+q}}{-p+\sqrt{p^2+q}}\right)$

$$\geqslant \frac{(-p)^2 + 2\cdot(-p)\cdot\sqrt{p^2+q} + p^2 + q}{q} \underset{\uparrow}{\geqslant} \frac{4p^2}{q}$$

$$(-p)^2 + 2\cdot(-p)\cdot\underbrace{\sqrt{p^2+q}}_{\geq(-p)>0} + p^2 + \underbrace{q}_{>0} > 4(-p)^2 = 4p^2$$

$\uparrow$ $q>0$

Bemerkung:

$$\hat{K}_1 = \hat{K}(F^1, \binom{p}{q}) = \begin{cases} 1 , & p \leq 0 \quad \text{(siehe Aufgabe)} \\ 1 , & p \geqslant 0 \quad \text{(siehe Vorlesung)} \end{cases}$$

$$\hat{K}_2 = \hat{K}(F^2, F^1(\binom{p}{q})) = \begin{cases} \geqslant \frac{4p^2}{q} , & p \leq 0 \quad \text{(siehe Aufgabe)} \\ 2 , & p \geqslant 0 \quad \text{(siehe Vorlesung)} \end{cases}$$

Welcher Algorithmus ist in diesem Fall gutartig?

Der Algorithmus 1 ist

- gutartig $\iff p \leq 0$
- nicht-gutartig $\iff p > 0$.

Der Algorithmus 2 ist

- gutartig $\iff p \geq 0$
- nicht-gutartig $\iff p < 0$.

Zu (d): Starte zunächst die Rundungsfehler GUI

Zu Algorithmus 1:

① : Wähle „Mantissenlänge 8" unter allgemeine Parameter
② : Wähle unter Standardbeispiele die „quad. Gleichung I"
③ : Unter Funktionen gebe im Feld „Anfangswerte" -10 (bzw. -100, -1000) ein.
④ : Wähle den Button Berechnen.

Zu Algorithmus 2:

① : siehe oben
② : Wähle unter Standardbeispiele die „quad. Gleichung II"
③ : siehe oben
④ : siehe oben

Screenshots:

① : Gebe im Terminal „Gimp&" ein.
② : In Gimp wähle in der Menüleiste „File → Acquire → Screenshot" und Kliche auf „Snap"
③ : Wähle nun das Fenster aus, von dem ein Screenshot gemacht werden soll.
④ : Wähle in Gimp nun „drucken" und speichere das Bild.

Interpretation der Ergebnisse:

Wie wir im theoretischen Teil dieser Aufgabe festgestellt haben, ist für $p<0$ der Algorithmus 1 gutartig und der Algorithmus 2 nicht-gutartig. Dieses Ergebnis wird von den numerischen Untersuchungen bestätigt: Dies erkennen wir, wenn wir uns den zugehörigen Fehler ansehen: ($q=1$, Mantissenlänge 8)

	Algorithmus 1	Algorithmus 2
$p=-10$	10^{-7}	10^{-4}
$p=-100$	10^{-7}	10^{-3}
$p=-1000$	10^{-10}	10^{-4}

Größenordnung des Fehlers

Wie wir feststellen ist der Fehler beim Algorithmus 1 erheblich kleiner.

Algorithmus 1:

$p = -10$

Rundungsfehler

allgemeine Parameter

maximale Anzahl an Schritten: 2

Mantissenlaenge: 8

Ausgabe alle 1 Schritte.

iterierte Funktion

Benutzerdef. Fkt.

Standardbeispiele

harmonische Reihe

quad. Gleichung I

quad. Gleichung II

Status

bereit

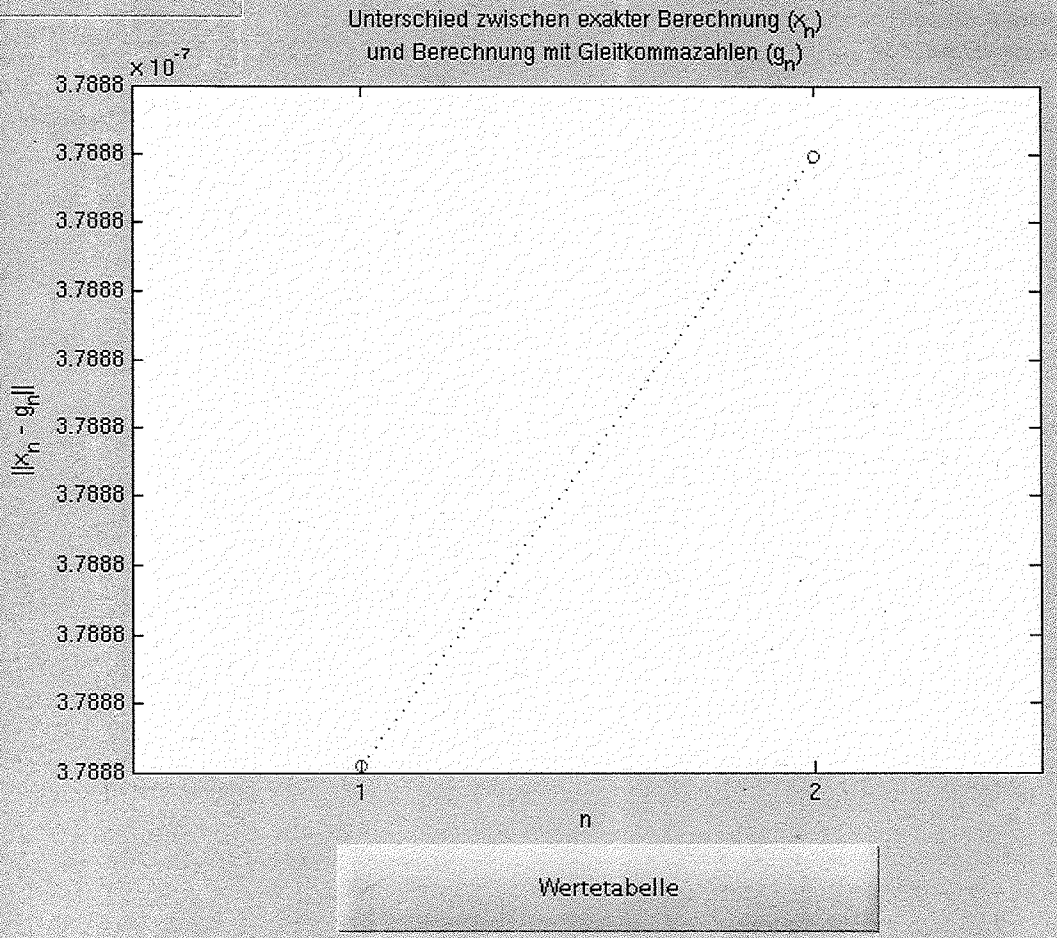
Funktionen

Anfangswerte

-10
1

0

Berechnen



Wertetabelle

Algorithmus 1:

p = 100

Rundungsfehler

allgemeine Parameter

maximale Anzahl an Schritten: 2

Mantissenlaenge: 8

Ausgabe alle 1 Schritte.

iterierte Funktion

Benutzerdef. Fkt.

Standardbeispiele

harmonische Reihe

quad. Gleichung I

quad. Gleichung II

Status

bereit

Funktionen

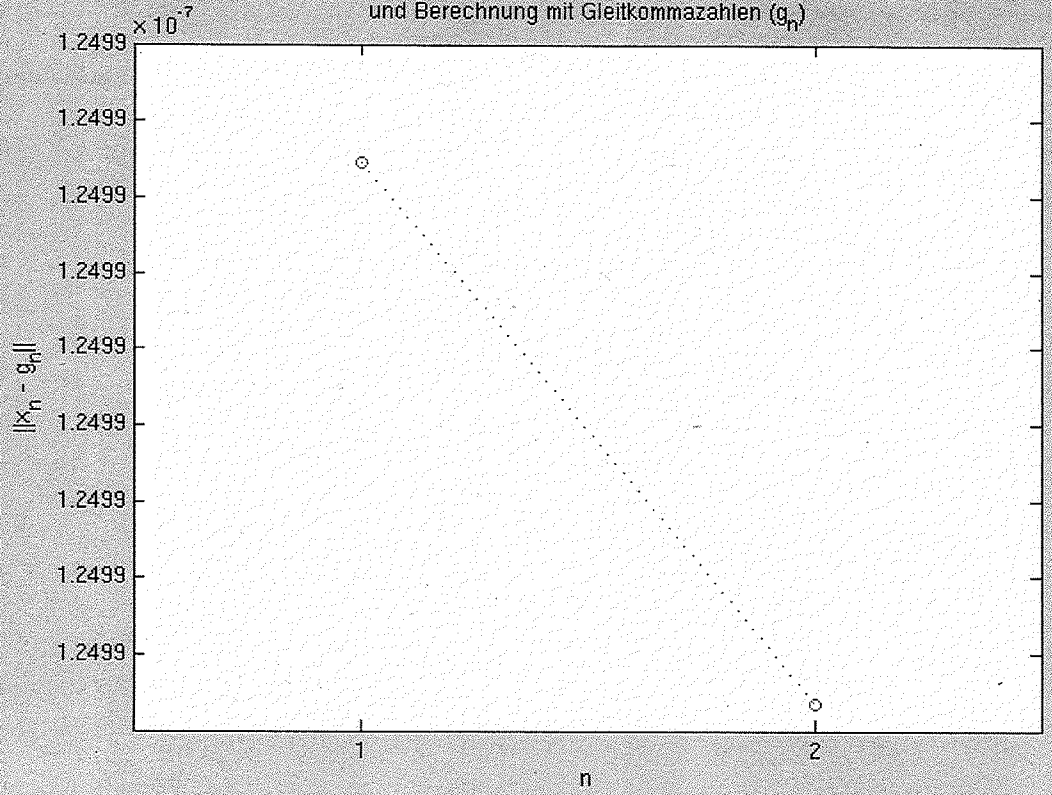
Anfangswerte

-100
1

0

Berechnen

Unterschied zwischen exakter Berechnung (x_n) und Berechnung mit Gleitkommazahlen (g_n)



Wertetabelle

Algorithmus 1:

p = -1000

Rundungsfehler

allgemeine Parameter

maximale Anzahl an Schritten: 2

Mantissenlaenge: 8

Ausgabe alle 1 Schritte.

iterierte Funktion

Benutzerdef. Fkt.

Standardbeispiele

harmonische Reihe

quad. Gleichung I

quad. Gleichung II

Status

bereit

Funktionen

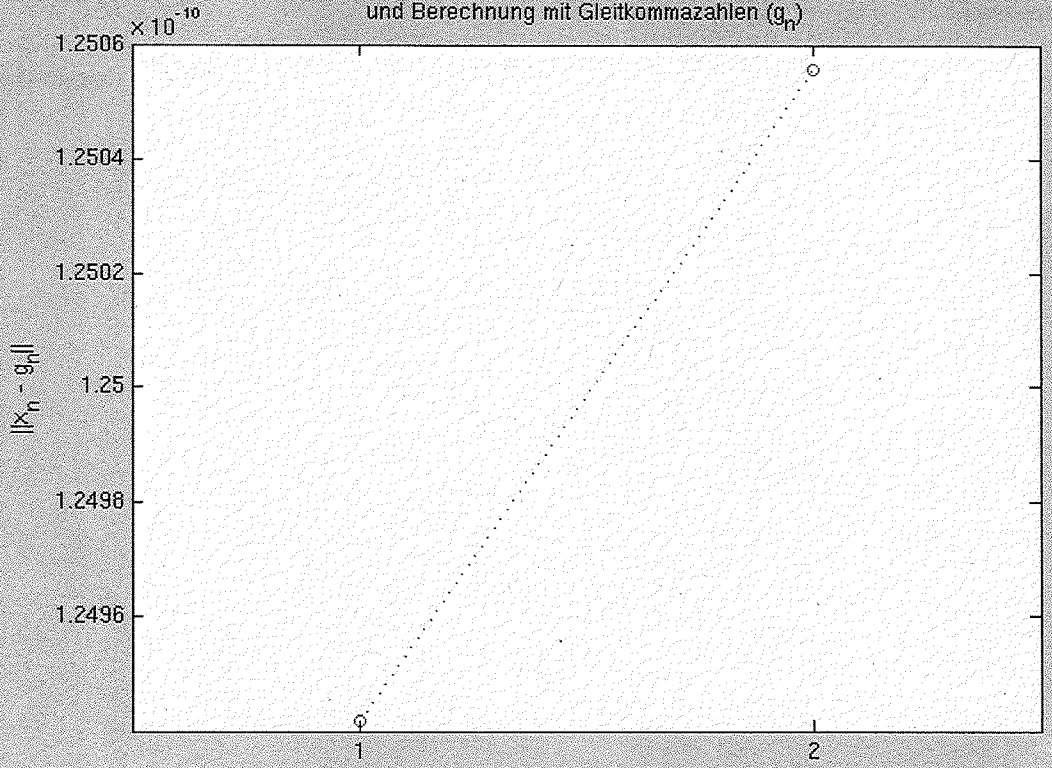
Anfangswerte

-1000
1

0

Berechnen

Unterschied zwischen exakter Berechnung (x_n) und Berechnung mit Gleitkommazahlen (g_n)



Wertetabelle

Algorithmus 2:

$p = -10$

Rundungsfehler

allgemeine Parameter

maximale Anzahl an Schritten: 2

Mantissenlaenge: 8

Ausgabe alle 1 Schritte.

iterierte Funktion

Benutzerdef. Fkt.

Standardbeispiele

harmonische Reihe

quad. Gleichung I

quad. Gleichung II

Status

bereit

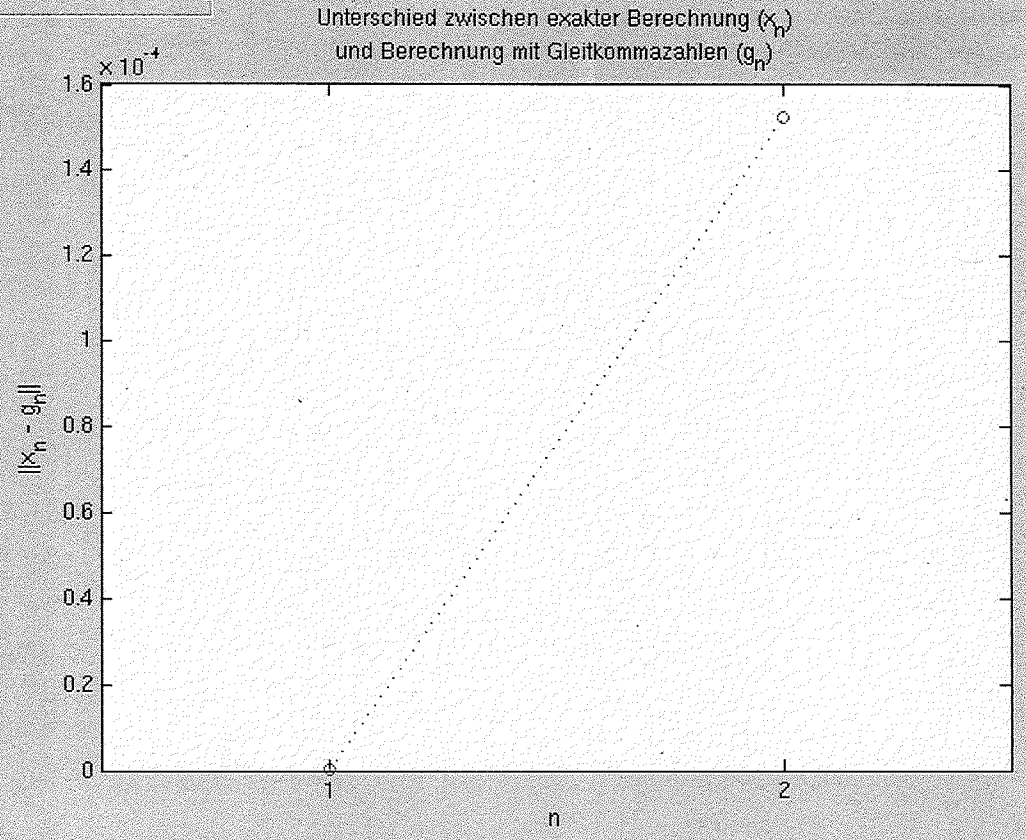
Funktionen

Anfangswerte

-10
1

0

Berechnen



Wertetabelle

Algorithmus 2:

$p = -100$

Rundungsfehler

allgemeine Parameter

maximale Anzahl an Schritten: 2

Mantissenlaenge: 8

Ausgabe alle 1 Schritte.

iterierte Funktion

Benutzerdef. Fkt.

Standardbeispiele

harmonische Reihe

quad. Gleichung I

quad. Gleichung II

Status

bereit

Funktionen

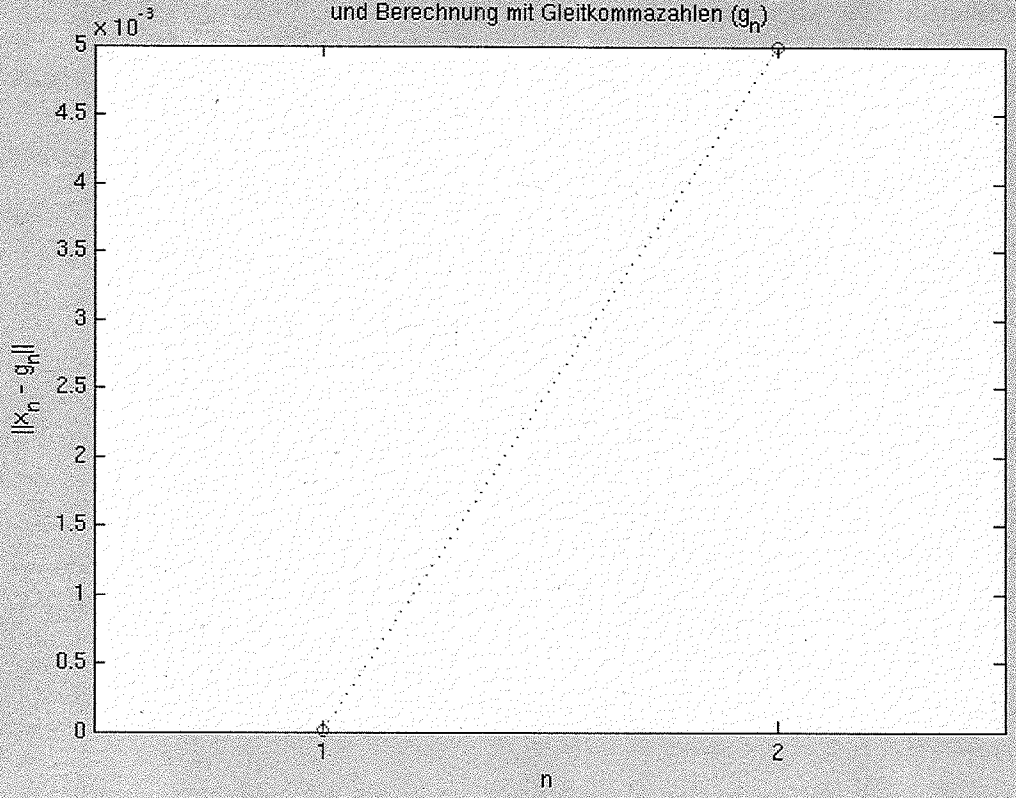
Anfangswerte

-100
1

0

Berechnen

Unterschied zwischen exakter Berechnung (x_n) und Berechnung mit Gleitkommazahlen (g_n)



Wertetabelle

Algorithmus 2:

$p = -1000$

Rundungsfehler

allgemeine Parameter

maximale Anzahl an Schritten: 2

Mantissenlaenge: 8

Ausgabe alle 1 Schritte.

iterierte Funktion

Benutzerdef. Fkt.

Standardbeispiele

harmonische Reihe

quad. Gleichung I

quad. Gleichung II

Status

bereit

Funktionen

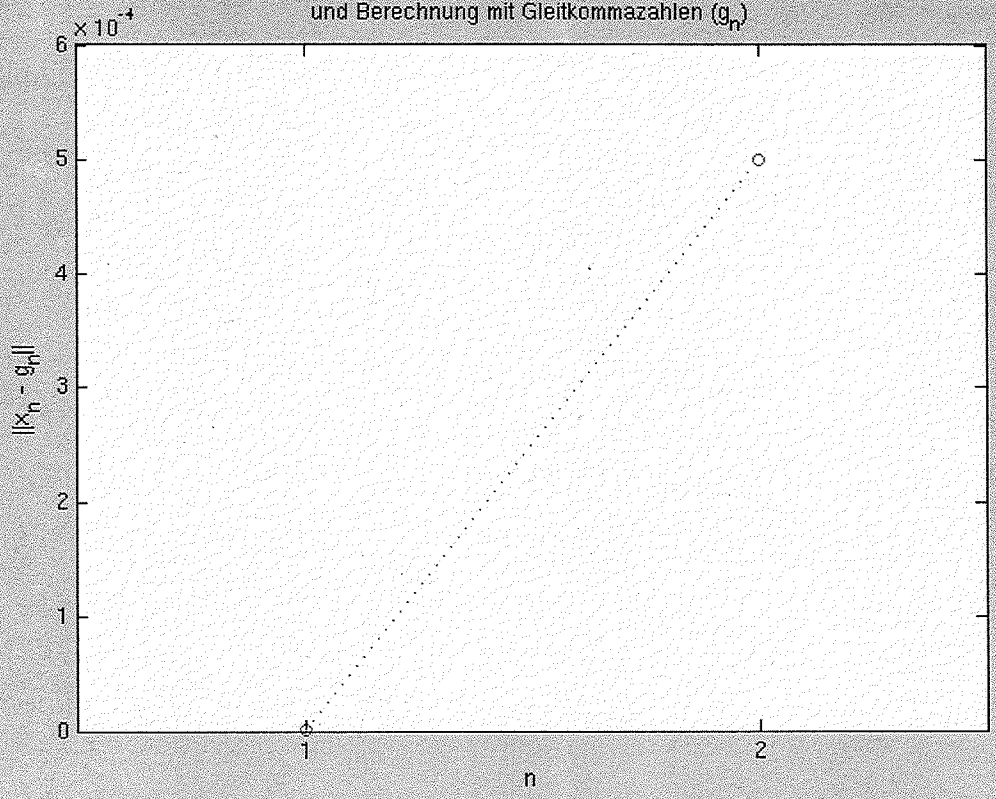
Anfangswerte

-1000
1

0

Berechnen

Unterschied zwischen exakter Berechnung (x_n) und Berechnung mit Gleitkommazahlen (g_n)



Wertetabelle