Kapitel 6
Problem und Kondition

Wir haben ein Problem. Die Eingabedaten sind gewisse reelle Zahlen. Da nicht einmal die reellen Zahlen exakt dargestellt werden können, wird die vom Computer ausgegebene Lösung im allgemeinen wohl auch nicht richtig sein. Wenn daran schon nichts zu ändern ist, wollen wir wenigstens wissen, wie groß der Ausgabefehler höchstens werden kann, am besten schon vor der Rechnung. Dazu entwickeln wir nun für verschiedene Probleme jeweils ein mathematische Modell, welches den Ausgabefehler zwar nur näherungsweise beschreibt, dafür aber leicht auszuwerten ist. Im Mittelpunkt steht dabei immer die Kondition des Problems.

6.1 Kondition der Grundrechenarten

Wir beginnen mit einem sehr einfachen Problem, der

$$\text{Auswertung von Summe, Differenz, Produkt oder Quotient von } x,\, y \in \mathbb{R} \setminus \{0\}\ ,$$

und wollen feststellen, wie sich relative Fehler in den Eingabedaten x und y, auf das Ergebnis auswirken. Seien also

$$\tilde{x} = x(1+\varepsilon_x)\ , \qquad \tilde{y} = y(1+\varepsilon_y) \tag{6.1}$$

Näherungen von $x, y \neq 0$. Dann gilt

$$\frac{|x-\tilde{x}|}{|x|} = |\varepsilon_x|\ , \quad \frac{|y-\tilde{y}|}{|y|} = |\varepsilon_y|\ ,$$

und wir bezeichnen daher den *relativen Eingabefehler* mit

$$\varepsilon = \max\left\{|\varepsilon_x|, |\varepsilon_y|\right\}\ .$$

Handelt es sich um die gerundeten Eingabedaten $\tilde{x} = \mathrm{rd}(x), \tilde{y} = \mathrm{rd}(y) \in \mathbb{G}(q,\ell)$, so ist nach Satz 5.5 der relative Eingabefehler durch die Maschinengenauigkeit beschränkt, also $\varepsilon \leq eps(q,\ell)$.

Definition 6.1 (Relative Kondition der Grundrechenarten). Es bezeichne $*$ eine der vier Grundrechenarten $+, -, \cdot, /$ und es gelte $x * y \neq 0$. Die kleinste Zahl κ_{rel} mit der Eigenschaft

$$\frac{|x*y - \tilde{x}*\tilde{y}|}{|x*y|} \leq \kappa_{\mathrm{rel}}\varepsilon + o(\varepsilon) \tag{6.2}$$

für alle genügend kleinen ε heisst dann *relative Kondition* von $*$ an der Stelle x, y.

Wir werden weiter unten sehen, dass wir für die Grundrechenarten tatsächlich eine Konstante κ_{rel} findet, für die eine Folge $\varepsilon_x, \varepsilon_y \to 0$ existiert, so dass in (6.2) Gleichheit gilt. Das Landau-Symbol $o(\varepsilon)$ bedeutet dabei nach Definition (Einzelheiten und Rechenregeln finden sich in Abschnitt A.6 des Anhangs)

$$\left(\frac{|x*y - \tilde{x}*\tilde{y}|}{|x*y|} - \kappa_{\mathrm{rel}}\varepsilon\right)\varepsilon^{-1} \to 0 \quad \text{für } \varepsilon \to 0\ .$$

Damit ist klar, daß für genügend kleine ε der lineare Anteil $\kappa_{\text{rel}}\varepsilon$ in der Darstellung (6.2) des relativen Ausgabefehlers dominiert. Ignoriert man den Term höherer Ordnung $o(\varepsilon)$, so erhält man mit $\kappa_{\text{rel}}\varepsilon$ eine leicht auszuwertende und für $\varepsilon \to 0$ beliebig genaue obere Schranke für den maximalen Ausgabefehler.

Die Kondition eines Problems beschreibt die maximale Fehlerverstärkung
bis auf Terme höherer Ordnung.

Diese Art von näherungsweiser Beschreibung der Wirklichkeit ist typisch für mathematische Modelle wie wir sie vor allem aus der Physik kennen: Das Modell $v(t) = \frac{1}{2}gt^2$ für die Geschwindigkeit v eines fallenden Steins unter der Erdbeschleunigung g vernachlässigt Terme „höherer Ordnung“ für Drall, Luftreibung und vieles mehr, ist dafür aber einfach auszuwerten und für hinreichend kleine Geschwindigkeiten beliebig genau.

Damit die Auswertung auch wirklich einfach ist, brauchen wir natürlich die Kondition κ_{rel}. Diese wollen wir jetzt für die Grundrechenarten ausrechnen. Wir beginnen mit der Addition.

Satz 6.2 (Addition). *Unter der Voraussetzung $x,y > 0$ gilt*

$$\frac{|(x+y)-(\tilde{x}+\tilde{y})|}{|x+y|} \leq \varepsilon \,. \tag{6.3}$$

Die Kondition der Addition ist also $\kappa_+ = 1$.

Beweis. Mit der Dreiecksungleichung erhalten wir aus (6.1) die Abschätzung

$$\begin{aligned} |(x+y)-(\tilde{x}+\tilde{y})| &= |(x-\tilde{x})+(y-\tilde{y})| = |x\varepsilon_x + y\varepsilon_y| \\ &\leq |x||\varepsilon_x| + |y||\varepsilon_y| \leq (|x|+|y|)\varepsilon = |x+y|\varepsilon \,. \end{aligned} \tag{6.4}$$

Im letzten Schritt haben wir $x,y > 0$ ausgenutzt. Im Falle $\varepsilon_x = \varepsilon_y = \varepsilon > 0$ gilt sogar $|(x+y)-(\tilde{x}+\tilde{y})| = |x+y|\varepsilon$. Die Abschätzung (6.4) kann also nicht verbessert werden. □

Offenbar tauchen in diesem Fall überhaupt keine Terme höherer Ordnung auf. Außerdem ist die Addition positiver Zahlen ausgesprochen gutmütig: Schlimmstenfalls übertragen sich Eingabefehler in den Summanden verstärkungsfrei auf die Summe.

Satz 6.3 (Subtraktion). *Unter den Voraussetzungen $x,y > 0$ und $x \neq y$ gilt*

$$\frac{|(x-y)-(\tilde{x}-\tilde{y})|}{|x-y|} \leq \left(\frac{|x|+|y|}{|x-y|}\right)\varepsilon \,. \tag{6.5}$$

Die Kondition der Subtraktion ist also

$$\kappa_- = \frac{|x|+|y|}{|x-y|} \,.$$

Beweis. Wir gehen genauso vor wie im Beweis zu Satz 6.2 und erhalten anstelle von (6.4) diesmal

$$|(x-y)-(\tilde{x}-\tilde{y})| = |(x-\tilde{x})+(\tilde{y}-y)| = |x\varepsilon_x - y\varepsilon_y| \leq |x||\varepsilon_x| + |y||\varepsilon_y| \leq (|x|+|y|)\varepsilon \,.$$

Auch diese Abschätzung lässt sich nicht verbessern. □

Wieder kommen keine Terme höherer Ordnung vor. Im Gegensatz zu κ_+ hängt die Kondition κ_- der Subtraktion positiver Zahlen aber von x und y ab. Insbesondere kann κ_- beliebig groß werden, wenn sich x und y wenig unterscheiden. Beispielsweise gilt mit beliebigem $m \in \mathbb{N}$

$$x = 1 \approx y = 1 + 10^{-m} \quad \Longrightarrow \quad \frac{|x|+|y|}{|x-y|} > 2 \cdot 10^m \,.$$

Die relative Kondition der Subtraktion wird für $x \approx y$ beliebig groß .

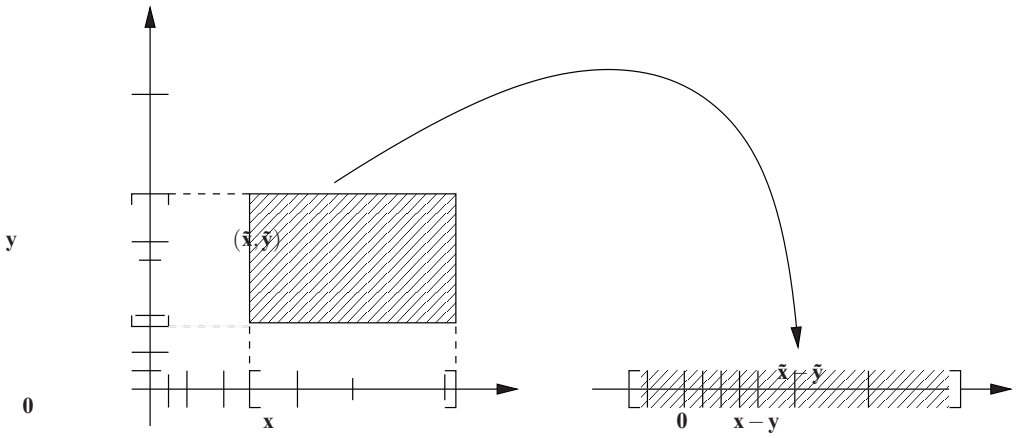

Abb. 6.1 Fehlerverstärkung durch Auslöschung.

Die praktischen Auswirkungen wollen wir anhand eines einfachen Zahlenbeispiels illustrieren. Wir gehen dabei von $\ell = 4$ gültigen Stellen im Dezimalsystem aus. Im Falle $x = \frac{1}{3}$ und $y = 0,3332$ ist dann $\tilde{x} = 0.3333$ und $\tilde{y} = y$. Wir erhalten das Ergebnis

$$\tilde{x} - \tilde{y} = 0.0001$$

und nach *rechnerinterner Normalisierung* die Gleitkommadarstellung

$$\tilde{x} - \tilde{y} \mathrel{\hat{=}} 0.1XYZ \cdot 10^{-3} \ .$$

Das ist anstelle der *vier richtigen Stellen* in den Eingabedaten nur noch *eine richtige Stelle* im Ergebnis! Dieses Phänomen heißt *Auslöschung*. Anstelle der Symbole X, Y, Z können dabei irgenwelche Zahlen stehen. Welche Zahlen das sind, ist rechnerabhängig. Man sollte nicht $X = Y = Z = 0$ erwarten. In unserem Beispiel liegt die Verstärkung des relativen Fehlers bei $\kappa_- = 4999$.

Wir kommen zur Multiplikation.

Satz 6.4 (Multiplikation). *Unter den Voraussetzungen $x \neq 0$ und $y \neq 0$ gilt*

$$\frac{|(x \cdot y) - (\tilde{x} \cdot \tilde{y})|}{|x \cdot y|} \leq 2\,\varepsilon + o(\varepsilon) \ . \tag{6.6}$$

Die relative Kondition der Multiplikation ist also $\kappa_\bullet = 2$

Beweis. Einsetzen von (6.1) nebst Dreiecksungleichung liefert die Abschätzung

$$|x \cdot y - \tilde{x} \cdot \tilde{y}| = |xy||\varepsilon_x + \varepsilon_y + \varepsilon_x\varepsilon_y| \leq |xy|(2\varepsilon + \varepsilon^2) \ ,$$

und offenbar ist $\varepsilon^2 = o(\varepsilon)$. Wieder gilt Gleichheit für $\varepsilon_x = \varepsilon_y = \varepsilon > 0$. □

Diesmal taucht ein Term höherer Ordnung auf: $o(\varepsilon) = \varepsilon^2$. Den exakten maximalen Ausgabefehler

$$\max_{|\varepsilon_x|,|\varepsilon_y| \leq \varepsilon} \frac{|(x \cdot y) - (\tilde{x} \cdot \tilde{y})|}{|x \cdot y|} = \kappa_\bullet \varepsilon + \varepsilon^2$$

können wir für dieses elementare Problem zwar auch noch einfach berechnen, aber schon bei einfacher Genauigkeit $\varepsilon = \text{eps} = 5,96 \cdot 10^{-8}$ liefert der lineare Anteil

$$\kappa_\bullet \varepsilon = 11,92 \cdot 10^{-8} \approx 11,920000355216 \cdot 10^{-8} = \kappa_\bullet \varepsilon + \varepsilon^2$$

eine ziemlich gute Näherung.

Zum Abschluss betrachten wir noch die Division.

Satz 6.5 (Division). *Unter den Voraussetzungen $x \neq 0$, $y \neq 0$ und $\varepsilon < 1$ gilt*

$$\frac{|x/y - \tilde{x}/\tilde{y}|}{|x/y|} \leq 2\,\varepsilon + o(\varepsilon)\,. \tag{6.7}$$

Die relative Kondition der Division ist also $\kappa_/ = 2$.

Beweis. Einsetzen von (6.1) ergibt

$$\begin{aligned} |x/y - \tilde{x}/\tilde{y}| &= \left|1 - \frac{1+\varepsilon_x}{1+\varepsilon_y}\right| |x/y| = \left|\frac{\varepsilon_y - \varepsilon_x}{1+\varepsilon_y}\right| |x/y| \\ &\leq \frac{2\varepsilon}{1-\varepsilon}\,|x/y| = \left(2\varepsilon + \frac{2\varepsilon^2}{1-\varepsilon}\right)|x/y|\,, \end{aligned}$$

und offenbar ist $2\varepsilon^2/(1-\varepsilon) = o(\varepsilon)$. Auch diese Abschätzung ist scharf, kann also nicht verbessert werden. □

Man beachte, daß wir zum ersten Mal voraussetzen mussten, daß ε genügend klein ist.

6.2 Kondition von Funktionsauswertungen

Gegeben sei ein Intervall $I \subset \mathbb{R}$, eine Funktion $f : I \to \mathbb{R}$ und $x_0 \in I$. Wir betrachten das *Problem* der

$$\text{Auswertung von } f \text{ an der Stelle } x_0\,. \tag{6.8}$$

Dabei kann die Funktion f explizit gegeben sein oder nicht (siehe zum Beispiel Abschnitt 6.4). Wie schon im vorigen Abschnitt wollen wir für die Auswirkungen von Eingabefehlern x_0 auf die Ausgabe $f(x_0)$ eine obere Schranke angeben, die einfach zu berechnen und bis auf Terme höherer Ordnung richtig ist. Dabei wird es keine Rolle spielen, ob es sich um Rundungsfehler handelt oder nicht. Wir müssen aber Ein- und Ausgabefehler quantifizieren können, das heißt, wir müssen uns für ein Fehlermaß entscheiden. Wir beginnen mit dem absoluten Fehler.

6.2.1 Absolute Kondition

Zum Aufwärmen betrachten wir die lineare Funktion

$$f(x) = ax + b$$

mit $a, b \in \mathbb{R}$. Offenbar gilt

$$|f(x_0) - f(x)| = |a||x_0 - x|\,.$$

Der Eingabefehler $|x_0 - x|$ wird also durch Auswertung von f um den Faktor $|a|$ verstärkt oder, falls $|a| < 1$ ist, gedämpft. Von der absoluten Kondition verlangen wir nun, daß sie dieses lineare Verhalten bis auf Terme höherer Ordnung reproduziert.

Definition 6.6 (Absolute Kondition). Die *absolute Kondition* κ_{abs} von (6.8) ist die kleinste Zahl mit der Eigenschaft

$$|f(x_0) - f(x)| \leq \kappa_{\text{abs}}|x_0 - x| + o(x_0 - x) \tag{6.9}$$

für genügend kleine Störungen $|x_0 - x|$. Liegt (6.9) für keine reelle Zahl κ_{abs} vor, so wird $\kappa_{\text{abs}} = \infty$ gesetzt.

Die Einschränkung auf „genügend kleine Störungen“ bedeutet, daß es ein $\rho > 0$ geben muß, so daß (6.9) für alle x mit $|x_0 - x| \leq \rho$ richtig ist. Es kann sein, daß es kein $\rho > 0$ gibt, so daß f für alle x mit $|x_0 - x| \leq \rho$ auch nur auswertbar ist. Dann ist $\kappa_{\text{abs}} = \infty$.

Auf elegantere Weise kann man die absolute Kondition auch wie folgt charakterisieren (vergleiche Definition A.7 im Anhang)

$$\kappa_{\text{abs}} = \limsup_{x \to x_0} \frac{|f(x_0) - f(x)|}{|x_0 - x|}\,. \tag{6.10}$$

Die Kondition beschreibt die maximale Verstärkung des Eingabefehlers in x_0 durch Auswertung von f bis auf Terme höherer Ordnung.

Im Gegensatz zum exakten Ausgabefehler

$$\max_{|x_0-x|\le\varepsilon} |f(x_0)-f(x)| \tag{6.11}$$

kann man die für $\varepsilon \to 0$ beliebig genaue obere Schranke $\kappa_{\text{rel}}\varepsilon$ leicht berechnen. Dazu muß man naürlich erst einmal κ_{rel} kennen. Wir betrachten einige Beispiele.

Die absolute Kondition von $f(x) = ax+b$ ist offenbar $\kappa_{\text{abs}} = |a|$.

Ist die Funktion f unstetig in x_0, so gilt $\kappa_{\text{abs}} = \infty$. Der Einfachheit halber zeigen wir das nur für die Auswertung der speziellen Funktion f,

$$f(x) = \begin{cases} 0 & \text{falls } x < 0\,, \\ 1 & \text{falls } x \ge 0\,, \end{cases}$$

an der Stelle $x_0 = 0$: Auch für beliebig kleine Eingabefehler ist im Falle $x < x_0$ immer $|f(x_0)-f(x)| = 1$.

Aber auch bei stetigen Funktionen kann $\kappa_{\text{abs}} = \infty$ vorliegen. Sei nämlich

$$f(x) = \begin{cases} -\sqrt{-x} & \text{falls } x < 0\,, \\ \sqrt{x} & \text{falls } x \ge 0\,, \end{cases} \qquad x \in \mathbb{R}$$

Mit der binomischen Formel folgt dann für $x, x_0 > 0$

$$\begin{aligned} f(x_0)-f(x) &= \frac{1}{2\sqrt{x_0}}(x_0-x) + \frac{\sqrt{x_0}-\sqrt{x}}{2(x_0+\sqrt{x_0 x})}(x_0-x) \\ &= \frac{1}{2\sqrt{x_0}}(x_0-x) + o(x_0-x) \end{aligned}$$

und somit $\kappa_{\text{abs}} = \frac{1}{2\sqrt{x_0}}$. Nun sei $x_0 = 0$. Im Widerspruch zu $\kappa_{\text{abs}} = \infty$ sei $\kappa \in \mathbb{R}$ eine Zahl mit der Eigenschaft

$$\sqrt{x} = |f(0)-f(x)| \le \kappa|x_0-x| + o(x_0-x) = \kappa x + o(x) \qquad \forall x > 0\,.$$

Dann ergibt Division durch $\sqrt{x}$ die Abschätzung

$$1 \le \kappa\sqrt{x} + \frac{o(x)}{x}\sqrt{x}\,.$$

Grenzübergang $x \to 0$ führt auf den Widerspruch $1 \le 0$. Damit ist $\kappa_{\text{abs}} = \infty$ die absolute Kondition der Auswertung von f an der Stelle $x_0 = 0$.

Für differenzierbare Funktionen f lässt sich κ_{abs} leicht berechnen.

Satz 6.7. *Ist f differenzierbar in x_0, so gilt $\kappa_{\text{abs}} = |f'(x_0)|$.*

Beweis. Nach Definition der Ableitung gilt

$$\lim_{x\to x_0} \frac{f(x_0)-f(x)}{x_0-x} = f'(x_0)$$

oder gleichbedeutend

$$\lim_{x\to x_0} \frac{f(x_0)-f(x)-f'(x_0)(x_0-x)}{x_0-x} = 0\,.$$

Daraus erhält man unmittelbar

$$f(x_0)-f(x) = f'(x_0)(x_0-x) + o(x_0-x)\,. \tag{6.12}$$

Nimmt man nun auf beiden Seiten den Betrag und verwendet die Rechenregel $|a+o(\varepsilon)| = |a|+o(\varepsilon)$, so folgt die Behauptung. □

Zusammen mit den Rechenregeln der Differentialrechnung ist Satz 6.7 ein äußerst wertvolles Hilfsmittel zur Berechnung von κ_{rel}. Im Gegensatz zur Stetigkeit ist die Differenzierbarkeit aber nicht notwendig für $\kappa_{\text{abs}} < \infty$.

Beispielsweise ist $\kappa_{\text{abs}} = 1$ die absolute Kondition von $f(x) = |x|$, denn aus der Dreiecksungleichung $|a+b| \le |a| + |b|$ erhält man

$$|f(x_0) - f(x)| = \big||x_0| - |x|\big| \le |x_0 - x| \,.$$

Die folgende Definition verallgemeinert diesen Sachverhalt.

Definition 6.8 (Lipschitz-Stetigkeit). Die Funktion $f : I \to \mathbb{R}$ heißt *Lipschitz-stetig* mit *Lipschitz-Konstante L*, falls

$$|f(x) - f(y)| \le L|x - y| \qquad \forall x, y \in I \,.$$

Beispielsweise sind $f(x) = x$, $f(x) = |x|$ und $f(x) = \sin(x)$ Lipschitz-stetig mit Lipschitz-Konstante $L = 1$.

Aus der Definition folgt unmittelbar:

Satz 6.9. *Ist $f : I \to \mathbb{R}$ Lipschitz-stetig mit Lipschitz-Konstante L, so genügt die absolute Kondition κ_{abs} von* (6.8) *der Abschätzung*

$$\kappa_{\text{abs}} \le L \,.$$

Wir betrachten noch die Kondition von geschachtelten Funktionen.

Satz 6.10. *Es sei $f(x) = g \circ h(x) = g(h(x))$. Es bezeichne $\kappa_{\text{abs}}(h, x_0)$ die absolute Kondition der Auswertung von h an der Stelle x_0 und $\kappa_{\text{abs}}(g, y_0)$ die absolute Kondition der Auswertung von g an der Stelle $y_0 = h(x_0)$. Dann ist*

$$\kappa_{\text{abs}} \le \kappa_{\text{abs}}(g, y_0)\ \kappa_{\text{abs}}(h, x_0) \,. \tag{6.13}$$

Ist h differenzierbar in x_0 und g differenzierbar in y_0, so liegt in (6.13) *Gleichheit vor.*

Beweis. Wir brauchen nur den Fall $\kappa_{\text{abs}}(h, x_0) < \infty$ und $\kappa_{\text{abs}}(g, y_0) < \infty$ zu betrachten. Nach Definition gilt

$$\begin{aligned} |f(x_0) - f(x)| &\le \kappa_{\text{abs}}(g, y_0)|h(x_0) - h(x)| + o(h(x_0) - h(x)) \\ &\le \kappa_{\text{abs}}(g, y_0)\kappa_{\text{abs}}(h, x_0)|x_0 - x| + \kappa_{\text{abs}}(g, y_0)o(x_0 - x) + \\ &\qquad + o\left(\kappa_{\text{abs}}(h, x_0)|x_0 - x| + o(x_0 - x)\right) \\ &= \kappa_{\text{abs}}(g, y_0)\kappa_{\text{abs}}(h, x_0)|x_0 - x| + o(x_0 - x) \,. \end{aligned}$$

Der zweite Teil der Behauptung folgt aus Satz 6.7 und der Kettenregel. □

Als Beispiel betrachten wir die geschachtelte Funktion

$$f(x) = d|\sin(x)| \,, \qquad x_0 = 0 \,.$$

Setzt man $g(y) = d|y|$ und $h(x) = \sin(x)$, so folgt aus Satz 6.10 nebst Satz 6.7 sofort $\kappa_{\text{abs}} \le d$. Obwohl g nicht differenzierbar ist, kann man in diesem Fall sogar $\kappa_{\text{abs}} = d$ zeigen. Die Fehlerverstärkung wächst also linear mit d.

6.2.2 Relative Kondition

Unter den natürlichen Voraussetzungen $x_0 \neq 0$ und $f(x_0) \neq 0$ definieren wir die *relative Kondition* in vollkommener Analogie zur absoluten Kondition als maximalen Verstärkungsfaktor des relativen Fehlers.

Definition 6.11 (Relative Kondition). Die *relative Kondition* κ_{rel} von (6.8) ist die kleinste Zahl mit der Eigenschaft

$$\frac{|f(x_0) - f(x)|}{|f(x_0)|} \le \kappa_{\text{rel}} \frac{|x_0 - x|}{|x_0|} + o(x_0 - x) \,. \tag{6.14}$$

Ist (6.14) für keine reelle Zahl κ_{rel} richtig, so wird $\kappa_{\text{rel}} = \infty$ gesetzt.

Zwischen relativer und absoluter Kondition besteht ein enger Zusammenhang.

Satz 6.12. *Es gilt*

$$\kappa_{\text{rel}} = \frac{|x_0|}{|f(x_0)|}\,\kappa_{\text{abs}}\,. \tag{6.15}$$

Beweis. Offenbar ist (6.9) äquivalent mit

$$\frac{|f(x_0)-f(x)|}{|f(x_0)|} \le \frac{|x_0|}{|f(x_0)|}\,\kappa_{\text{abs}}\,\frac{|x_0-x|}{|x_0|} + o(x_0-x)\,,$$

und nach Definition ist κ_{abs} die kleinste Zahl, mit der diese Abschätzung gilt. □

Absolute und relative Kondition können völlig unterschiedliche Größenordnungen haben. Ein einfaches Beispiel ist $f(x) = ax$ mit $a, x_0 \neq 0$. Aus Satz 6.7 und Satz 6.12 folgt

$$\kappa_{\text{abs}} = |f'(x_0)| = |a|\,, \qquad \kappa_{\text{rel}} = \frac{|x_0||f'(x_0)|}{|f(x_0)|} = 1\,.$$

In der Mathematik spielen Invarianzen oft eine wichtige Rolle (vgl. etwa die Invarianz des ggT in Lemma 4.12). Im Gegensatz zur absoluten Kondition ist die relative Kondition κ_{rel} *skalierungsinvariant* in folgendem Sinne.

Satz 6.13. *Die relative Kondition* κ_{rel} *von* (6.8) *ist invariant gegenüber der Multiplikation von* f *mit reellen Zahlen* $\lambda \neq 0$.

Wir notieren noch eine einfache, aber wichtige Konsequenz aus unseren Konditionsbetrachtungen: Das Problem

Berechnung von $f(x_0)$ an der Stelle $x_0 \in \mathbb{R}$ bis auf die relative Genauigkeit TOL

im allgemeinen nur dann numerisch lösbar, wenn

$$\kappa_{\text{rel}} \cdot \text{eps} < \text{TOL}$$

vorliegt, andernfalls leider nicht!

6.3 Nullstellen quadratischer Polynome

Unser Problem besteht darin, die Nullstellen eines quadratischen Polynoms zu finden, also die quadratische Gleichung

$$x^2 + px + q = 0\,. \tag{6.16}$$

zu lösen. Unter der Voraussetzung $p^2/4 > q$ gibt es genau zwei verschiedene Nullstellen, die sich explizit als Funktion der Koeffizienten $p, q \in \mathbb{R}$ angeben lassen. Eine der beiden Nullstellen ist

$$x(p,q) = -\frac{p}{2} + \sqrt{\frac{p^2}{4} - q}\,.$$

Störungen von q. Wir halten zunächst p fest und untersuchen die Auswirkungen von Störungen von q auf diese Nullstelle. Für festes p ist $f(q) := x(p,q)$ eine differenzierbare Funktion von $q \in \{z \in \mathbb{R} \mid z < p^2/4\}$ mit der Ableitung

$$\partial_q x(p,q) := f'(q) = -\frac{1}{\sqrt{p^2 - 4q}}\,.$$

Man nennt $\partial_q x(p,q)$ *partielle* (teilweise) Ableitung von $x(p,q)$ nach q. Wir können nun Satz 6.7 anwenden und erhalten

$$|x(p,q) - x(p,\tilde{q})| \le \kappa_{\text{abs},q}|q-\tilde{q}| + o(q-\tilde{q})$$

mit der absolute Kondition

$$\kappa_{\text{abs},q} = |\partial_q x(p,q)|\,.$$

Falls $q \to p^2/4$, wächst $\kappa_{\text{abs},q}$ über alle Grenzen. Man beachte, daß bei Gleichheit $q = p^2/4$ eine doppelte Nullstelle vorliegt und im Falle $q > p^2/4$ keine reelle Lösung von (6.16) existiert.

Störungen von p. Als nächstes wollen wir die Auswirkungen von Störungen des Koeffizienten p auf die Nullstelle $x(p,q)$ untersuchen. Dazu halten wir nun q fest und wenden Satz 6.7 auf die differenzierbare Funktion $g(p) := x(p,q)$ von $p \in \{z \in \mathbb{R} \mid z^2/4 > q\}$ an. Die Ableitung von g ist

$$\partial_p x(p,q) := g'(p) = -\frac{1}{2} + \frac{p}{2\sqrt{p^2-4q}}\,,$$

und Anwendung von Satz 6.7 ergibt

$$|x(p,q) - x(\tilde{p},q)| = \kappa_{\text{abs},p}|p-\tilde{p}| + o(p-\tilde{p})$$

mit der absoluten Kondition

$$\kappa_{\text{abs},p} = |\partial_p x(p,q)|\,.$$

Wie zuvor wird $\kappa_{\text{abs},p}$ beliebig groß, wenn $p^2/4 \to q$, also wenn man sich einer doppelten Nullstelle nähert. Aber es gibt auch gute Nachrichten: Im Falle $q \approx 0$ ist $\kappa_{\text{abs},p} \approx 0$, denn die Nullstelle $x(p,0) = 0$ hängt nicht von p ab. Störungen von p werden im Falle $q \approx 0$ also sogar gedämpft.

Störungen von p und q. Was passiert aber nun, wenn beide Koeffizienten p und q gestört werden? Die Antwort steckt in folgendem Lemma.

Lemma 6.14. *Sei $a < b$, $c < d$ und $h : D = (a,b) \times (c,d) \to \mathbb{R}$ eine Funktion, deren partielle Ableitungen $\partial_x h(x,y)$ und $\partial_y h(x,y)$ für alle $(x,y) \in D$ existieren. Außerdem sei $\partial_y h(x,y)$ für alle $(x,y) \in D$ stetig in x. Dann gilt für alle $(x,y),(\tilde{x},\tilde{y}) \in D$*

$$h(x,y) - h(\tilde{x},\tilde{y}) = \partial_x h(x,y)(x-\tilde{x}) + \partial_y h(x,y)(y-\tilde{y}) + o(|x-\tilde{x}| + |y-\tilde{y}|) \tag{6.17}$$

Beweis. Im Beweis zu Satz 6.7 haben wir aus der Definition der Ableitung die Formel (6.12) gewonnen. Hält man y fest und betrachtet $h(x,y)$ als eine Funktion von x, so liefert (6.12)

$$h(x,y) - h(\tilde{x},y) = \partial_x h(x,y)(x-\tilde{x}) + o(x-\tilde{x})\,. \tag{6.18}$$

Wenn man $\tilde{x}$ fixiert und $h(\tilde{x},y)$ als Funktion von y auffasst, erhält man auf dieselbe Weise

$$h(\tilde{x},y) - h(\tilde{x},\tilde{y}) = \partial_y h(\tilde{x},y)(y-\tilde{y}) + o(y-\tilde{y})\,. \tag{6.19}$$

Aus der Stetigkeit von $\partial_y h(x,y)$ in x folgt

$$\left(\partial_y h(x,y) - \partial_y h(\tilde{x},y)\right)(y-\tilde{y}) = o(|x-\tilde{x}| + |y-\tilde{y}|)\,,$$

denn

$$0 \le \frac{|\left(\partial_y h(x,y) - \partial_y h(\tilde{x},y)\right)(y-\tilde{y})|}{|x-\tilde{x}| + |y-\tilde{y}|} \le |\partial_y h(x,y) - \partial_y h(\tilde{x},y)| \to 0$$

für $|x-\tilde{x}| + |y-\tilde{y}| \to 0$. Zusammen mit (6.19) bedeutet das

$$h(\tilde{x},y) - h(\tilde{x},\tilde{y}) = \partial_y h(x,y)(y-\tilde{y}) + o(|x-\tilde{x}| + |y-\tilde{y}|)\,. \tag{6.20}$$

Aus (6.18) und (6.20) erhält man schließlich

$$\begin{aligned} h(x,y) - h(\tilde{x},\tilde{y}) &= \big(h(x,y) - h(\tilde{x},y)\big) + \big(h(\tilde{x},y) - h(\tilde{x},\tilde{y})\big) \\ &= \partial_x h(x,y)(x-\tilde{x}) + \partial_y h(x,y)(y-\tilde{y}) + o(|x-\tilde{x}| + |y-\tilde{y}|) \end{aligned}$$

und damit die Behauptung. □

Die Eigenschaft (6.17) kann man auch zur Grundlage einer Definition machen: Eine Funktion $h : D \to \mathbb{R}$ heisst *differenzierbar* an der Stelle $(x,y) \in D$, wenn es Zahlen $\partial_x h(x,y), \partial_y h(x,y) \in \mathbb{R}$ gibt, so daß (6.17) vorliegt. Der Zeilenvektor $\nabla h(x,y) = (\partial_x h(x,y), \partial_y h(x,y))$ heisst dann *Ableitung von h* an der Stelle (x,y) (siehe etwa [23, Kapitel 2]). Damit stehen wir an der Schwelle zur Differentialrechnung mit mehreren Veränderlichen.

Anstatt diese Schwelle zu überschreiten, wollen wir Lemma 6.14 auf die Funktion $x(p,q)$ anwenden.

Satz 6.15. *Es sei $p^2/4 > q$. Dann gilt für die Auswirkung genügend kleiner Störungen $p-\tilde{p}$ und $q-\tilde{q}$ auf die Nullstelle $x(p,q)$ die Abschätzung*

$$|x(p,q) - x(\tilde{p},\tilde{q})| \leq \kappa_{\text{abs},p}|p-\tilde{p}| + \kappa_{\text{abs},q}|q-\tilde{q}| + o(|p-\tilde{p}| + |q-\tilde{q}|)\,. \tag{6.21}$$

Beweis. Zunächst können wir wegen $p^2/4 > q$ eine Menge $D = (a,b)\times(c,d)$ finden, so daß $(p,q) \in D$ und $\tilde{p}^2/4 - \tilde{q} > 0\ \forall(\tilde{p},\tilde{q}) \in D$ vorliegt. Auf D existieren die partiellen Ableitungen $\partial_p x$ und $\partial_q x$. Außerdem ist $\partial_q x(p,q)$ stetig in p. Aus Lemma 6.14 folgt daher

$$x(p,q) - x(\tilde{p},\tilde{q}) = \partial_p x(p,q)(p-\tilde{p}) + \partial_q x(p,q)(q-\tilde{q}) + o(|p-\tilde{p}| + |q-\tilde{q}|)$$

für alle $(\tilde{p},\tilde{q}) \in D$. Mit der Dreiecksungleichung erhalten wir daraus die Behauptung. □

Die Auswirkungen von Störungen der Koeffizienten p und q
lassen sich einfach addieren!

Mit Blick auf Lemma 6.14 gilt das auch allgemein: Wenn mehr als ein Parameter gestört wird, können wir unsere Resultate aus Abschnitt 6.2 auf jeden dieser Parameter einzeln anwenden. Von dieser Einsicht werden wir auch später noch Gebrauch machen.

Die obigen Resultate lassen sich auf Nullstellen von Polynomen n-ter Ordnung, also auf Lösungen der Gleichung

$$\sum_{i=0}^{n} a_i x^i = 0$$

erweitern. Da für $n \geq 5$ im allgemeinen aber keine geschlossene Darstellung von Lösungen x als Funktion $x(a_0,\ldots,a_{n-1})$ der Koeffizienten $a_0,\ldots,a_{n-1}$ existiert, wird das aber etwas komplizierter [14].

6.4 Lösung nichtlinearer Gleichungen

Wir betrachten das folgende Problem: Berechne die Lösung x^* der nichtlinearen Gleichung

$$x^* \in (a,b): \qquad g(x^*) = y^* \tag{6.22}$$

zu einer gegebenen Funktion $g : (a,b) \to \mathbb{R}$ und rechter Seite $y^* \in \mathbb{R}$. Dabei setzen wir voraus, daß das Problem (6.22) sinnvoll gestellt ist, also eine eindeutig bestimmte Lösung x^* besitzt. Wir wollen die Auswirkung von Störungen der rechten Seite y^* auf die Lösung x^* studieren. Dazu definieren wir zunächst die absoluten Kondition κ_{abs} von (6.22).

Definition 6.16 (Absolute Kondition nichtlinearer Gleichungen). Die *absolute Kondition* κ_{abs} von (6.22) ist die kleinste Zahl mit der Eigenschaft

$$|x^* - x| \leq \kappa_{\text{abs}}|y^* - y| + o(y^* - y) \tag{6.23}$$

für alle rechten Seiten $y \neq y^*$ mit genügend kleinem Abstand $|y^* - y| > 0$ zu $y^* = g(x^*)$ und den zugehörigen Lösungen x des gestörten Problems

$$x \in (a,b): \qquad g(x) = y\,. \tag{6.24}$$

Ist (6.26) für keine relle Zahl κ_{abs} richtig, so wird $\kappa_{\text{abs}} = \infty$ gesetzt.

Existenz einer differenzierbaren Inversen von g. Obwohl (6.22) eindeutig lösbar ist, kann das gestörte Problem (6.24) für alle $y \neq y^*$ unlösbar sein. Dann ist $\kappa_{\text{abs}} = \infty$. Das ist beispielsweise für $g(x) = x^2$ und $y^* = 0$ der Fall: Für $y \neq y^*$ hat $x^2 = y$ entweder zwei reelle Lösungen oder gar keine.

Hinreichende Bedingungen für die Existenz und Eindeutigkeit von Lösungen des gestörten Problems (6.24) liefert das folgende Lemma.

Lemma 6.17. *Es sei $g \in C^1(a,b)$, $x^* \in (a,b)$ und es gelte $g(x^*) = y^*$ sowie $g'(x^*) \neq 0$. Dann gibt es α, $\beta \in \mathbb{R}$ mit $a < \alpha < x^* < \beta < b$, so daß das gestörte Problem* (6.24) *für jedes $y \in V = [g(\alpha), g(\beta)]$ eine eindeutig bestimmte Lösung $x \in U = [\alpha, \beta]$ besitzt.*

Beweis. Ohne Beschränkung der Allgemeinheit sei $g'(x^*) > 0$. Dann gibt es nach Definition der Stetigkeit von g' (vgl. Anhang A.3) zu $\delta = \frac{g'(x^*)}{2} > 0$ ein $\varepsilon > 0$, so daß $[x^* - \varepsilon, x^* + \varepsilon] \subset (a,b)$ und

$$|g'(x^*) - g'(x)| \leq \tfrac{1}{2} g'(x^*) \qquad \forall x \in [x^* - \varepsilon, x^* + \varepsilon]$$

vorliegt. Wir setzen $\alpha = x^* - \varepsilon$, $\beta = x^* + \varepsilon$ und $U = [\alpha, \beta]$. Mit $|g'(x^*) - g'(x)| \geq g'(x^*) - g'(x)$ folgt dann

$$g'(x) \geq \tfrac{1}{2} g'(x^*) > 0 \qquad \forall x \in U \, .$$

Damit ist die Funktion g streng monoton wachsend auf U, das heisst, es gilt

$$x_1 < x_2 \quad \Longleftrightarrow \quad g(x_1) < g(x_2) \qquad \forall x_1, x_2 \in U \, ,$$

denn aus dem Mittelwertsatz (vgl. z.B. Königsberger [22, Abschnitt 9.3]) folgt ja

$$\frac{g(x_1) - g(x_2)}{x_1 - x_2} = g'(x) \geq \tfrac{1}{2} g'(x^*) > 0$$

mit einer Zwischenstelle $x \in (x_1, x_2) \subset U$.

Nach dem Zwischenwertsatz (vgl. z.B. Königsberger [22, Abschnitt 7.4]) wird deshalb jeder Wert $y \in V = [g(\alpha), g(\beta)]$ genau einmal angenommen. Das ist gerade die Behauptung. □

Nach Lemma 6.17 ist g in einer genügend kleinen Umgebung U von x^* umkehrbar. Die Umkehrfunktion $g^{-1} : V \to U$ von g ordnet jeder rechten Seite $y \in V$ die Lösung $g^{-1}(y) = x \in U$ von (6.24) zu. Nach Definition gilt dann

$$g(g^{-1}(y)) = y \quad \forall y \in V \, .$$

Als Beispiel betrachten wir $g(x) = \sin(x)$, $(a,b) = \mathbb{R}$ und $y^* = 0$. In diesem Fall können wir $U = [\alpha, \beta] = [-\pi/2, \pi/2]$ und $V = [-1, 1]$ wählen, denn für $y \in [-1,1]$ hat die Gleichung $\sin(x) = y$ die eindeutig bestimmte Lösung $x = \arcsin(y)$.

Das Problem (6.22) ist also äquivalent zur

$$\text{Auswertung von } g^{-1} : V \to \mathbb{R} \text{ an der Stelle } y^* \in V \, . \tag{6.25}$$

Die absolute Kondition von (6.22) stimmt mit der in Definition 6.6 eingeführten absoluten Kondition von (6.25) überein. Das kann man durch Einsetzen von $x^* = g^{-1}(y^*)$, $x = g^{-1}(y)$ in (6.23) bestätigen. Zur Bestimmung der Kondition von (6.25) können wir daher unsere Resultate aus Abschnitt 6.2 anwenden, insbesondere Satz 6.7 mit $f = g^{-1}$. Vorher haben wir allerdings noch die Differenzierbarkeit von g^{-1} nachzuweisen.

Lemma 6.18. *Unter den Voraussetzungen von Lemma 6.17 ist $g^{-1} \in C^1(V)$ und es gilt*

$$(g^{-1})'(y) = \frac{1}{g'(g^{-1}(y))} \quad \forall y \in V \, .$$

Beweis. Wir zeigen zunächst die Stetigkeit von g^{-1} auf V. Im Widerspruch zur Stetigkeit in $y \in V$ nehmen wir an, daß es ein $\delta > 0$ gibt, so daß für jedes $\varepsilon = \frac{1}{n}$ und jedes $n \in \mathbb{N}$, ein $y_n \in V$ gibt, so daß $|y - y_n| \leq \varepsilon = \frac{1}{n}$ und $|g^{-1}(y) - g^{-1}(y_n)| > \delta$ vorliegt. Setzt man $x_n = g^{-1}(y_n)$ und $x = g^{-1}(y)$, bedeutet das $g(x_n) \to g(x)$ aber $|x - x_n| > \delta$. Nun ist offenbar $(x_n) \in U = [\alpha, \beta]$ und U ist ein abgeschlossenes, beschränktes Intervall. Nach dem Satz von Bolzano-Weierstrass (siehe z.B. Königsberger [22, Abschnitt 5.5]) gibt es daher ein $\bar{x} \in U$, gegen das eine Teilfolge $(x_{n_k})_{k \in \mathbb{N}}$ konvergiert. Aus der Stetigkeit von g folgt $g(x_{n_k}) \to g(\bar{x})$ und wegen $g(x_{n_k}) \to g(x)$ muss wegen der Eindeutigkeit des Grenzwert $g(\bar{x}) = g(x)$ sein. Andererseits folgt aus $|x - x_n| > \delta \; \forall n \in \mathbb{N}$ unmittelbar $x \neq \bar{x}$. Das ist ein Widerspruch zur Umkehrbarkeit von g auf U. Also ist g^{-1} stetig auf V.

Nun zeigen wir die Differenzierbarkeit von g^{-1} auf V. Sei $y \in V$ und $(y_n) \subset V$ eine gegen y konvergente Folge. Dann ist $x_n = g^{-1}(y_n) \in U$ und wegen der Stetigkeit von g^{-1} konvergiert $x_n \to x = g^{-1}(y) \in U$. Aus der Differenzierbarkeit von g folgt nun

$$\lim_{n\to\infty}\frac{g^{-1}(y)-g^{-1}(y_n)}{y-y_n}=\lim_{n\to\infty}\frac{x-x_n}{g(x)-g(x_n)}=\lim_{n\to\infty}\frac{1}{\frac{g(x)-g(x_n)}{x-x_n}}=\frac{1}{g'(x)}\ .$$

□

Absolute und relative Kondition nichtlinearer Gleichungen.

Satz 6.19. *Es sei $g\in C^1(a,b)$, $x^*\in(a,b)$ und es gelte $g(x^*)=y^*$ sowie $g'(x^*)\neq 0$. Dann ist die absolute Kondition κ_{abs} der Lösung der nichtlinearen Gleichung* (6.22) *bei Störung der rechten Seite y^* gegeben durch*

$$\kappa_{\text{abs}}=\frac{1}{|g'(x^*)|}\ .$$

Beweis. Die Behauptung folgt mit Lemma 6.17 und 6.18 aus Satz 6.7. □

Nichtlineare Gleichungen mit $g'(x^*)\approx 0$ sind also beliebig schlecht konditioniert. Bei Gleichheit $g'(x^*)=0$ spricht man von einer *doppelten Nullstelle*. Dann gilt $\kappa_{\text{abs}}=\infty$.

Als erstes Beispiel betrachten wir $g(x)=x^2$ und $y^*=0$. Dann ist $x_0=0$ eine doppelte Nullstelle und daher $\kappa_{\text{abs}}=\infty$.

Ähnlich sieht es aus für $g(x)=x^3$ und $y^*=0$. Auch wenn in diesem Fall für alle $y^*\in\mathbb{R}$ eine eindeutig bestimmte Lösung existiert, ist die Tangente an g in (x^*,y^*) parallel zur x-Achse (schleifender Schnitt). Die Kondition der Auswertung der Umkehrabbildung $g^{-1}(y)=\sqrt[3]{y}$ in $y^*=0$ ist $\kappa_{\text{abs}}=\infty$.

Im Falle $g(x)=\frac{1}{x}$ und $y^*>0$ erhält man $x^*=g^{-1}(y^*)=\frac{1}{y^*}$ als Lösung von $g(x^*)=y^*$. Für verschwindendes y^* ist die Tangente an g in (x^*,y^*) und die x-Achse fast parallel. Dementsprechnd wächst $\kappa_{\text{abs}}=\frac{1}{|g'(x^*)|}=(x^*)^2=\frac{1}{(y^*)^2}\to\infty$ für $y^*\to 0$. Für $y^*=0$ gibt es keine Lösung.

Oft ist man an der Anzahl von gültigen Stellen der Lösung von (6.22) interessiert. Unter der Annahme $y^*\neq 0$ und $x^*\neq 0$ betrachten wir deshalb auch die *relative Kondition*.

Definition 6.20 (Relative Kondition nichtlinearer Gleichungen). Die *relative Kondition* κ_{rel} von (6.22) ist die kleinste Zahl mit der Eigenschaft

$$\frac{|x^*-x|}{|x^*|}\le\kappa_{\text{rel}}\frac{|y^*-y|}{|y^*|}+o(y^*-y)\tag{6.26}$$

für alle rechten Seiten $y\neq y^*$ mit genügend kleinem Abstand $|y^*-y|>0$ zu y^* und den zugehörigen Lösungen x des gestörten Problems (6.24). Ist (6.26) für keine relle Zahl richtig, so wird $\kappa_{\text{rel}}=\infty$ gesetzt.

Im Gegensatz zur absoluten Kondition ist die relative Kondition κ_{rel} skalierungsinvariant (vergleiche Satz 6.13 und Aufgabe 6.8).

Durch Anwendung von Satz 6.12 erhalten wir unmittelbar das folgende Analogon von Satz 6.19.

Satz 6.21. *Es sei $g\in C^1(a,b)$, $x^*\in(a,b)$ und es gelte $g(x^*)=y^*$ sowie $g'(x^*)\neq 0$. Dann ist die relative Kondition κ_{rel} der Lösung der nichtlinearen Gleichung* (6.22) *bei Störung der rechten Seite y^* gegeben durch*

$$\kappa_{\text{rel}}=\frac{|g(x^*)|}{|x^*||g'(x^*)|}\ .$$

Offenbar kann die absolute oder relative Kondition aus Satz 6.19 und 6.21 nur berechnet werden, wenn man die exakte Lösung x^* kennt. Im allgemeinen ist das natürlich nicht der Fall. Auf der anderen Seite müssen zur Lösung von (6.22) typischerweise iterative Verfahren, eingesetzt werden, die jeweils eine gegen die Lösung x^* konvergente Folge (x_n) produzieren. Dabei können Näherungen $\kappa_{\text{abs},n}$ und $\kappa_{\text{rel},n}$ für κ_{abs} und κ_{rel} gleich mitberechnet werden, indem man in Satz 6.19 und 6.21 die exakte Lösung x^* durch die Näherung x_n ersetzt.

Relative Kondition und maximal erreichbare Genauigkeit der Lösung. Wir zeigen nun, daß die relative Kondition κ_{rel} die maximal erreichbare Genauigkeit bei der iterativen Lösung von (6.22) limitiert. Wegen der Skalierungsinvarianz von κ_{rel} können wir dabei ohne Beschränkung der Allgemeinheit $y^*=1$ annehmen.

Satz 6.22. *Es sei $y^*=1$, und für eine mit einem beliebigen Iterationsverfahren berechnete Iterierte x_{n^*} mit der relativen Genauigkeit*

$$\frac{|x^*-x_{n^*}|}{|x^*|}\le\tfrac{1}{2}\kappa_{\text{rel}}\ \text{eps}\tag{6.27}$$

gelte

$$\frac{|x^* - x_{n^*}|}{|x^*|} = \kappa_{\text{rel}} \frac{|y^* - y_{n^*}|}{|y^*|} \ . \tag{6.28}$$

mit $y_{n^*} = g(x_{n^*})$. *Dann folgt* $\text{rd}(g(x_{n^*})) = y^*$.

Beweis. Aus (6.27), (6.28) und $y^* = 1$ folgt

$$\tfrac{1}{2}\kappa_{\text{rel}}\text{eps} \geq \frac{|x^* - x_{n^*}|}{|x^*|} = \kappa_{\text{rel}} \frac{|y^* - y_{n^*}|}{|y^*|} = \kappa_{\text{rel}} |1 - y_{n^*}|$$

und damit

$$\tfrac{1}{2}\text{eps} \geq |1 - y_{n^*}| \ .$$

Mit Satz 5.8 erhält man daraus $\text{rd}(y_{n^*}) = 1$ oder, gleichbedeutend, $\text{rd}(g(x_{n^*})) = y^*$. □

Die Abschätzung in Satz 6.21 ist scharf. Die Voraussetzung (6.28) bedeutet gerade, daß der schlechtest mögliche Fall eintritt. Das ist beispielsweise für die affine Funktion $g(x) = (x-1)/\gamma + 1$ mit $\kappa_{\text{rel}} = \gamma > 0$ der Fall.

Aus $\text{rd}(g(x_{n^*})) = y^*$ folgt, daß der Rechner die exakte Lösung x^* nicht von der Näherung x_{n^*} unterscheiden kann. Sobald die Genauigkeitsschranke $\frac{1}{2}\kappa_{\text{rel}}\,\text{eps}$ unterschritten ist, lässt sich der relative Iterationsfehler daher nicht weiter reduzieren. Unser Iterationsverfahren bleibt stecken. Das hat weitreichende Folgen:

Gleichungen können im allgemeinen nur bis auf eine relative Genauigkeit von $\frac{1}{2}\kappa_{\text{rel}}\,\text{eps}$ numerisch gelöst werden.

Alles was über diese Genauigkeitsschranke hinausgeht, ist Glücksache.

Als Beispiel betrachten wir die nichtlineare Gleichung

$$x^* \in (0, \gamma): \qquad g(x^*) = \exp(\tan((x^* - 1)/\gamma)) = 1 = y^*$$

mit der Lösung $x^* = 1$. Die relative Kondition ist $\kappa_{\text{rel}} = \gamma$. Das linke Bild in Abbildung 6.2 zeigt $g(x)$ für $\gamma = 10^9$. Zur iterativen Lösung dieser Gleichung verwenden wir das Newton-Verfahren

$$x_{n+1} = x_n - \frac{g(x_n) - 1}{g'(x_n)} \ , \qquad n = 0, 1, \ldots ,$$

mit dem Startwert $x_0 = \gamma$. Für Einzelheiten zum Newton-Verfahren verweisen wir auf Deuflhard und Hohmann [10, Kapitel 4].

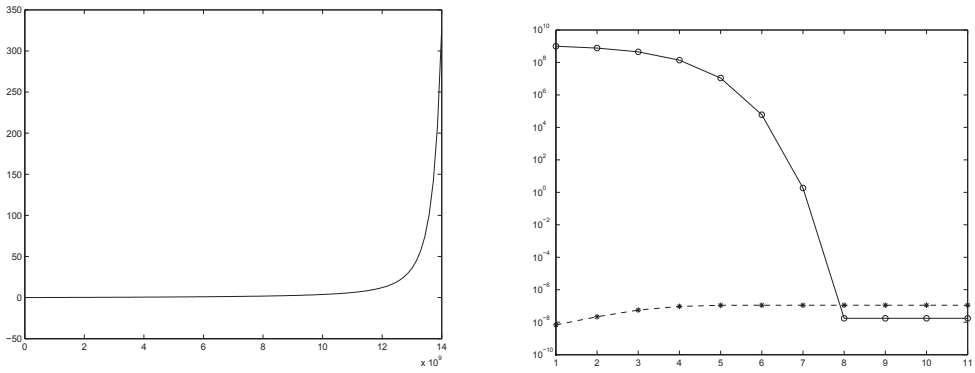

Abb. 6.2 Iterationsgeschichte und approximative Genauigkeitsschranke $\kappa_{\text{rel},n}\text{eps}$ für $\gamma = 10^9$

Das rechte Bild von Abbildung 6.2 zeigt, ebenfalls für $\gamma = 10^9$, den Iterationsfehler $|x^* - x_n|$ (durchgezogene Linie) und die Näherung $\frac{1}{2}\kappa_{\text{rel},n}\text{eps} = \frac{1}{2}\frac{|g(x_n)|}{|x_n||g'(x_n)|}\text{eps}$ für die Genauigkeitsschranke $\frac{1}{2}\kappa_{\text{rel}}\,\text{eps}$ (gestrichelte Linie) über der Anzahl n der Iterationen. Offenbar bleibt die anfänglich schnelle Konvergenz

nach $n^* = 7$ Schritten stecken. Dann ist die Genauigkeitkeitschranke $\frac{1}{2}\kappa_{\text{rel}}\,\text{eps} \approx \kappa_{\text{rel},n}\text{eps}$ unterschritten und die maximale Genauigkeit erreicht. Daran ändern auch andere Startwerte nichts. Beginnt man beispielsweise mit dem besseren Startwert $x_0 = \gamma/2$, so werden zwar nur $n^* = 5$ Iterationsschritte benötigt, um mit einem relativen Fehler von $1.03 \cdot 10^{-7}$ die Genauigkeitkeitschranke $\frac{1}{2}\kappa_{\text{rel}}\,\text{eps} = 1.11 \cdot 10^{-7}$ knapp zu unterschreiten, aber dann ist auch Schluß. Wie weit man jeweils im n^*-ten Schritt unter die Genauigkeitkeitschranke kommt ist Glückssache.

γ	10	10^6	10^9	10^{14}
$\kappa_{\text{rel}}\text{eps}$	$1.11 \cdot 10^{-15}$	$1.11 \cdot 10^{-10}$	$1.11 \cdot 10^{-7}$	$1.11 \cdot 10^{-2}$
$\frac{\lvert x^* - x_{n^*}\rvert}{\lvert x^*\rvert}$	$4.44 \cdot 10^{-16}$	$1.00 \cdot 10^{-10}$	$1.74 \cdot 10^{-8}$	$5.52 \cdot 10^{-3}$

Tabelle 6.1 Genauigkeitsschranke und erreichte Genauigkeit in Abhängigkeit von $\kappa_{\text{rel}} = \gamma$

In Tabelle 6.1 sind entsprechenden Resultate für verschiedene $\kappa_{\text{rel}} = \gamma$ jeweils für den Startwert $x_0 = \gamma$ zusammengestellt. Die Konvergenz stagniert jeweils nach Unterschreiten der Genauigkeitsschranke $\frac{1}{2}\kappa_{\text{rel}}$ eps. Mehr ist nicht drin.

In der Praxis muss neben der Kondition natürlich auch der Iterationsfehler $\frac{|x^* - x_n|}{|x^*|}$ approximiert werden, denn die Lösung x^* ist ja nicht bekannt. Numerisch brechenbare Abbruchkriterien finden sich bei Deuflhard [9, Chapter 2]

Die lokale Umkehrbarkeit von g aus Lemma 6.17 und 6.18 lässt sich auf Systeme nichtlinearer Gleichungen verallgemeinern [23, 3.3]. Auf diese Weise kann man analog zu den Sätzen 6.19, 6.21 und 6.22 auch Aussagen über die Kondition und die Grenzen der Genauigkeit der numerischen Lösung von *Systemen* nichtlinearer Gleichungen gewinnen. Wir gehen darauf hier nicht weiter ein, sondern verweisen auf die einschlägige Literatur [9, 10]. In Kapitel 9 kommen wir allerdings auf den wichtigen Spezialfall *linearer Gleichungssysteme* zurück.

6.5 Drei-Term-Rekursionen

Nehmen wir einmal an, daß ein ausgewachsenes Kaninchenpaar in jedem Monat ein neues Kaninchenpaar hervorbringt, daß Kaninchenpaare zwei Monate nach ihrer Geburt ausgewachsen sind und dass in jedem Monat ein Anteil α der ausgewachsenen Paare stirbt. Wie groß ist die Anzahl F_k der ausgewachsene Paare nach k Monaten? Nun, das hängt sicher davon ab, wie viele Paare am Anfang da waren. Beginnen wir mit einem neugeborenen Paar, also mit $F_0 = 0$, so ist dieses Paar nach einem Monat ausgewachsen, also $F_1 = 1$. Nun sei $k \geq 1$ beliebig aber fest gewählt und F_{k-1}, F_k bekannt. Innerhalb eines Monats kommen dann zu den F_k ausgewachsenen Paaren gerade die F_{k-1} vor einem Monat geborenen und mittlerweile auch ausgewachsenen hinzu und es sterben αF_k der ausgewachsenen Paare. Es gilt also

$$F_{k+1} = (1-\alpha)\cdot F_k + F_{k-1}\,, \qquad k = 1,2,3,\ldots\,. \tag{6.29}$$

Mit dieser Rekursionsformel lässt sich nun die Entwicklung der Anzahl der Kaninchenpaare F_k nach $k = 2,3,\ldots$ Monaten aus den Anfanggswerten F_0, F_1 berechnen, wobei wir uns nicht davor irritieren lassen, dass F_k in diesem Modell nicht unbedingt eine natürliche Zahl ist.

Für $\alpha = 0$, d.h. unter der Annahme, dass keine Kaninchen sterben, ergibt sich die Rekursionsformel $F_{k+1} = F_k + F_{k-1}$. Die sich daraus ergebenden F_k, $k = 0,1,2\ldots$ sind natürliche Zahlen und heissen *Fibonacci-Zahlen*.[1] Wir haben sie schon in Abschnitt 4.3 im einem ganz anderen Zusammenhang kennengelernt. Die Berechnungsvorschrift (6.29) ist ein Beispiel für eine homogene *Drei-Term-Rekursion*

$$x_{k+1} + a_k x_k + b_k x_{k-1} + c_k = 0\,, \qquad k = 1,2,3,\ldots\,, \tag{6.30}$$

[1] Leonardo da Pisa (≈1180-1241), genannt Fibonacci, verfasste das um 1200 erschienene *Liber abbaci* (Rechenbuch), in dessen zwölftem Kapitel unter anderem unser Kaninchenproblem behandelt wird.

mit gegebenen Koeffizienten $a_k, b_k, c_k \in \mathbb{R}$ und Anfangswerten $x_0, x_1 \in \mathbb{R}$. Drei-Term-Rekursionen spielen in Analysis und Numerischer Mathematik eine wichtige Rolle. Beispielsweise ist $v_k(x) = cos(kx)$ die Lösung der Drei-Term-Rekursion

$$v_0(x) = 1\,, \quad v_1(x) = \cos(x)\,, \qquad v_{k+1}(x) - 2\cos(x)v_k(x) + v_{k-1}(x) = 0\,, \tag{6.31}$$

und die sogenannten *Tschebyscheff-Polynome*[2] $T_k(x) = \cos(n \arccos(x))$ genügen der Drei-Term-Rekursion

$$T_0(x) = 1\,, \quad T_1(x) = x\,, \qquad T_{k+1}(x) - 2xT_k(x) + T_{k-1}(x) = 0\,.$$

Wir wollen die Auswirkungen von Störungen der Anfangswerte x_0, x_1 auf die Folgenglieder x_k, $k = 2, 3, \ldots$, untersuchen und beschränken uns dabei auf homogene Drei-Term-Rekursionen

$$x_{k+1} + ax_k + bx_{k-1} = 0\,, \qquad k = 1, 2, 3, \ldots\,, \tag{6.32}$$

mit konstanten Koeffizienten $a, b \in \mathbb{R}$. Zur Vorbereitung verschaffen wir uns eine geschlossene Darstellung der Folgenglieder x_k durch die Parameter a, b und die Anfangswerte x_0, x_1.

Lemma 6.23. *Das* charakteristische Polynom

$$p(\lambda) = \lambda^2 + a\lambda + b \tag{6.33}$$

habe die zwei verschiedenen, reellen Nullstellen λ_2, λ_1, *und es sei* $|\lambda_2| \geq |\lambda_1| > 0$. *Dann gilt*

$$x_k = \alpha(x_0, x_1)\lambda_1^k + \beta(x_0, x_1)\lambda_2^k \tag{6.34}$$

mit den Koeffizienten

$$\alpha(x_0, x_1) = \frac{\lambda_2 x_0 - x_1}{\lambda_2 - \lambda_1}\,, \quad \beta(x_0, x_1) = \frac{x_1 - \lambda_1 x_0}{\lambda_2 - \lambda_1}\,. \tag{6.35}$$

Beweis. Wir bestätigen (6.34) mit vollständiger Induktion. Die Fälle $k = 0$ und $k = 1$ erledigt man durch Einsetzen. Die Behauptung gelte nun für $k, k-1$ mit $k \geq 1$. Dann folgt mit den Abkürzungen $\alpha = \alpha(x_0, x_1)$ und $\beta = \beta(x_0, x_1)$

$$\alpha\lambda_1^{k+1} + \beta\lambda_2^{k+1} + a(\alpha\lambda_1^k + \beta\lambda_2^k) + b(\alpha\lambda_1^{k-1} + \beta\lambda_2^{k-1}) =$$
$$\alpha\lambda_1^{k-1}(\underbrace{\lambda_1^2 + a\lambda_1 + b}_{=0}) + \beta\lambda_2^{k-1}(\underbrace{\lambda_2^2 + a\lambda_2 + b}_{=0}) = 0$$

und damit die Behauptung. □

Die Voraussetzung $|\lambda_1| > 0$ ist äquivalent zu $b \neq 0$ und sichert damit, daß (6.32) nicht zu einer Zwei-Term-Rekursion degeneriert. Demgegenüber bedeutet die Existenz zweier verschiedener, reeller Nullstellen λ_2, λ_1 durchaus eine Einschränkung. Für die Drei-Term-Rekursion (6.31) ist diese Bedingung beispielsweise nicht erfüllt.

Ein Beweis, der veranschaulicht wie man auf die Formel (6.34) kommt und der sich noch dazu auf n-Term-Rekursionen verallgemeinern lässt, findet sich im Abschnitt A.9 des Anhangs.

Störungen von x_0. Nun kommen wir auf die Auswirkungen von Störungen der Anfangswerte x_0, x_1 auf x_k zurück und wollen zunächst nur Störungen von x_0 betrachten. Wir halten also x_1 fest. Dann besteht unser Problem darin, die Funktion

$$f_k(x) := \alpha(x, x_1)\lambda_1^k + \beta(x, x_1)\lambda_2^k\,, \qquad x \in \mathbb{R}\,,$$

an der Stelle $x = x_0$ auszuwerten, denn nach Definition von f_k gilt ja $x_k = f_k(x_0)$. Legen wir das relative Fehlermaß zugrunde, so werden die Auswirkungen von Störungen von x_0 auf $x_k = f_k(x_0)$ durch die relative Kondition κ_{rel}^k beschrieben. Zur Berechnung von κ_{rel}^k brauchen wir nur die Sätze 6.12 und 6.7 anzuwenden.

[2] Pafnuti Lwowitsch Tschebyscheff (1821-1894) leistete wichtige Beiträge in der Zahlentheorie. Die von ihm erfundenen Tschebyscheff-Polynome spielen eine zentrale Rolle bei der Approximation stetiger Funktionen durch Polynome.

Satz 6.24. *Es sei $x_0 \neq 0$ und $f_k(x_0) \neq 0$. Unter den Voraussetzungen von Lemma 6.23 ist dann für $k = 2,3,\ldots$ die relative Kondition κ_{rel}^k der Auswertung von f_k an der Stelle x_0 gegeben durch*

$$\kappa_{\text{rel}}^k = \frac{|x_0|}{|x_0 - R_k x_1/\lambda_1|}\,, \qquad R_k = \frac{\lambda_2^k - \lambda_1^k}{\lambda_2^k - \lambda_2\lambda_1^{k-1}}\,. \tag{6.36}$$

Beweis. Unter Verwendung der Ableitungen $\partial_1\alpha(x,x_1)$ und $\partial_1\beta(x,x_1)$ nach den jeweils ersten Funktionsargument x gilt

$$f_k'(x_0) = \partial_1\alpha(x_0,x_1)\lambda_1^k + \partial_1\beta(x_0,x_1)\lambda_2^k = \frac{\lambda_2\lambda_1^k - \lambda_1\lambda_2^k}{\lambda_2 - \lambda_1}\,.$$

Durch Ausschreiben von (6.34) nebst elementaren Umformungen bestätigt man

$$f_k(x_0) = \frac{1}{\lambda_2 - \lambda_1}\left((\lambda_2 x_0 - x_1)\lambda_1^k - (\lambda_1 x_0 - x_1)\lambda_2^k\right) = f_k'(x_0)\,(x_0 - R_k x_1/\lambda_1)\,,$$

und Einsetzen in (6.15) in Satz 6.12 liefert schließlich

$$\kappa_{\text{rel}}^k = \frac{|x_0||f_k'(x_0)|}{|f_k(x_0)|} = \frac{|x_0|}{|x_0 - R_k x_1/\lambda_1|}\,.$$

□

Wir wollen untersuchen, wie sich Eingabefehler in x_0 für große k auf die Folgenglieder $x_k = f_k(x_0)$ auswirken und konzentrieren uns dabei auf den Fall $|\lambda_2| > |\lambda_1|$. Unter dieser Voraussetzung gilt.

$$\lim_{k\to\infty} R_k = 1\,.$$

Daher ist

$$\lim_{k\to\infty} \kappa_{\text{rel}}^k = \begin{cases} \dfrac{|x_0|}{|x_0 - x_1/\lambda_1|}\,, & \text{falls } x_0 \neq x_1/\lambda_1\,, \\ \infty\,, & \text{falls } x_0 = x_1/\lambda_1\,. \end{cases} \tag{6.37}$$

Im Falle $|\lambda_2| > |\lambda_1|$ und $x_0 = x_1/\lambda_1$ werden Störungen von x_0 durch die Drei-Term-Rekursion (6.32) mit wachsendem k beliebig stark vergrößert.

Wir wollen zunächst den *generischen Fall* $x_0 \neq x_1/\lambda_1$ anhand eines Zahlenbeispiels illustrieren. Dazu betrachten wir die Drei-Term-Rekursion (6.29) von Fibonacci, also

$$x_{k+1} - x_k - x_{k-1} = 0\,, \quad k = 1,2,3\ldots\,. \tag{6.38}$$

Da der Anfangswert $x_0 = 0$ der Fibonacci-Zahlen nicht zum relativen Fehlerkonzept passt, wählen wir $x_0 = 1$ und $x_1 = 2$. Nach Lemma 6.23 sind die Folgenglieder x_k gegeben durch

$$x_k = f_k(x_0) = \tfrac{1}{10}\left((5 - 3\sqrt{5})\lambda_1^k + (5 + 3\sqrt{5})\lambda_2^k\right)\,, \qquad \lambda_1 = \tfrac{1-\sqrt{5}}{2}\,,\ \lambda_2 = \tfrac{1+\sqrt{5}}{2}\,,$$

und es gilt $\kappa_{\text{rel}}^k \to \frac{1}{2+\sqrt{5}}$ für $k \to \infty$. Mit wachsendem k werden relative Eingabefehler in x_0 also sogar gedämpft. Die folgende Tabelle bestätigt dies auch numerisch: Die Auswirkungen der Störung $\tilde{x}_0 = x_0(1 + 10^{-5})$ in der 6. Stelle bleiben durchweg auf die 6. Stelle beschränkt und werden mit wachsendem k sogar immer kleiner.

Im Falle $x_0 = x_1/\lambda_1$ wird die Fibonacci-Drei-Term-Rekursion (6.38) plötzlich bösartig. Zur Illustration wählen wir $x_1 = \lambda_1$ und $x_0 = x_1/\lambda_1 = 1$. Die folgende Tabelle zeigt, wie schnell nun die 10 gültigen Stellen von $\tilde{x}_0 = x_0(1 + 10^{-10})$ verloren gehen.

Wir wollen genauer verstehen, was dahintersteckt. Zunächst bestätigt man durch Einsetzen, daß beliebige Linearkombinationen $x_k = \alpha p_k + \beta q_k$ der Folgen

$$p_k = \lambda_1^k\,, \quad q_k = \lambda_2^k\,, \qquad k = 0,1,2,\ldots\,,$$

k	10	100	1000
$x_k = f_k(x_0)$	$1.440000 \cdot 10^2$	$9.273727 \cdot 10^{20}$	$1.137969 \cdot 10^{209}$
$\tilde{x}_k = f_k(\tilde{x}_0)$	$1.440003 \cdot 10^2$	$9.273748 \cdot 10^{20}$	$1.137971 \cdot 10^{209}$
rel. Fehler	$0.208 \cdot 10^{-5}$	$0.237 \cdot 10^{-5}$	$0.175 \cdot 10^{-5}$

Tabelle 6.2 Fehlerdämpfung durch die Drei-Term-Rekursion (6.38) im Falle $x_0 \neq x_1/\lambda_1$ für $\tilde{x}_0 = x_0(1+10^{-5})$.

k	10	20	30
$x_k = f_k(x_0)$	$8.1306187557 \cdot 10^{-3}$	$6.6106961351 \cdot 10^{-5}$	$5.3749049985 \cdot 10^{-7}$
$\tilde{x}_k = f_k(\tilde{x}_0)$	$8.1306221558 \cdot 10^{-2}$	$6.6525059900 \cdot 10^{-5}$	$5.1960211924 \cdot 10^{-5}$
rel. Fehler	$0.418 \cdot 10^{-6}$	$0.632 \cdot 10^{-2}$	$0.957 \cdot 10^{+2}$

Tabelle 6.3 Exponentiell wachsende Fehlerverstärkung im Falle $x_0 = x_1/\lambda_1$ für $\tilde{x}_0 = x_0(1+10^{-5})$.

Lösungen von (6.32) sind. Die Koeffizienten α, β werden gerade durch die beiden Anfangsbedingungen festgelegt (vgl. Lemma 6.23). Aus $|\lambda_2| > |\lambda_1|$ folgt

$$\lim_{k\to\infty} \frac{p_k}{q_k} = 0\,. \tag{6.39}$$

Also ist $q_k \gg p_k$ für $k \gg 1$. Man nennt daher q_k *dominant* und p_k *rezessiv*. Für die Folgenglieder x_k hat (6.39) die Konsequenz

$$x_k \approx \begin{cases} q_k \gg p_k & \text{falls } \beta \neq 0\,, \\ \\ p_k & \text{falls } \beta = 0\,. \end{cases}$$

Nun ist $x_0 = x_1/\lambda_1$ gleichbedeutend mit $\beta(x_0,x_1) = 0$. In diesem Fall enthält $x_k = \alpha(x_0,x_1)p_k$ also *nur* den rezessiven Lösungsanteil. Jede Approximation $\tilde{x}_0 \neq x_0$ führt aber zu $\beta(\tilde{x}_0,x_1) \neq 0$. Die gestörten Folgenglieder $\tilde{x}_k = \alpha(\tilde{x}_0,x_1)p_k + \beta(\tilde{x}_0,x_1)q_k$ enthalten damit auch den dominanten Anteil q_k, und der überlagert die exakte Lösung mit wachsendem k.

Störungen von x_1. Die Auswirkungen von Störungen des anderen Anfangswerts x_1 kann man berechnen, indem man nun x_0 festhält, x_k als Funktion

$$g_k(x) := \alpha(x_0,x)\lambda_1^k + \beta(x_0,x)\lambda_2^k$$

an der Stelle $x = x_1$ auffasst und wieder die Sätze 6.12 und 6.7 anwendet. Auf diese Weise erhält man die relative Kondition

$$\kappa_{\text{rel},1}^k = \frac{|x_1|}{|x_1 - \lambda_1 x_0/R_k|}\,, \quad R_k = \frac{\lambda_2^k - \lambda_1^k}{\lambda_2^k - \lambda_2\lambda_1^{k-1}}\,, \qquad k = 2,3,\ldots\,. \tag{6.40}$$

Im Falle $|\lambda_2| > |\lambda_1|$ gilt daher

$$\lim_{k\to\infty} \kappa_{\text{rel},1}^k = \begin{cases} \dfrac{|x_1|}{|x_1 - \lambda_1 x_0|}\,, & \text{falls } x_1 \neq \lambda_1 x_0\,, \\ \\ \infty\,, & \text{falls } x_1 = \lambda_1 x_0\,. \end{cases} \tag{6.41}$$

Im rezessiven Fall $x_1 = \lambda_1 x_0$ werden also auch Störungen von x_1 mit wachsendem k beliebig stark vergrößert.

Störungen von x_0 und x_1. Werden sowohl x_0 also auch x_1 gestört, so lassen sich die Auswirkungen auf die Folgenglieder $\tilde{x}_k$, $k = 2,3,\ldots$, einfach addieren. Dies folgt wie in Abschnitt 6.3 aus Lemma 6.14. Man erhält also

$$\frac{|x_k - \tilde{x}_k|}{|x_k|} \leq \kappa_{\text{rel},0}^k \frac{|x_0 - \tilde{x}_0|}{|x_0|} + \kappa_{\text{rel},1}^k \frac{|x_1 - \tilde{x}_1|}{|x_1|} + o(|x_0 - \tilde{x}_0| + |x_1 - \tilde{x}_1|)$$

mit der Kondition $\kappa_{\text{rel},0}^k = \kappa_{\text{rel}}^k$ aus Satz 6.24 und der Kondition $\kappa_{\text{rel},1}^k$ aus (6.40).

Drei-Term-Rekursionen mit variablen Koeffizienten zeigen ein ähnliches Verhalten. Beispielsweise führt die naive Verwendung von Drei-Term-Rekursionen schnell in den Besselschen Irrgarten. Neugierige verweisen wir auf Deuflhard et al. [28] oder das Lehrbuch von Deuflhard und Hohmann [10, Kapitel 6.2].

6.6 Ronaldinhos Kondition

Im Jahre 2005 zeigte der damals weltbeste Fußballspieler Ronaldinho in einem Werbespot ein vorher nie gesehenes Kunststück: Von der Strafraumgrenze aus schoss er den Ball viermal hintereinander an die Latte des Tores, so daß der Ball jedesmal wieder zu ihm zurücksprang[3]. Wir wollen uns mit der Frage beschäftigen, ob so etwas möglich ist oder ob da vielleicht tricktechnisch nachgeholfen wurde.

Dazu verschaffen wir uns zuerst ein mathematisches Modell für die Flugkurve des Balls, der von Ronaldinho mit der Abschussgeschwindigkeit v und dem Abschusswinkel α in Richtung Tor getreten wird. Unter Vernachlässigung von Luftreibung und Drall [35] beschreibt der Ball die Wurfparabel

$$h(t) = vt\sin\alpha - \tfrac{1}{2}gt^2 ,$$

$$x(t) = vt\cos\alpha .$$

Dabei bezeichnen $h(t)$ die Flughöhe des Balles und $x(t)$ die horizontale Entfernung vom Abschussort zur Zeit $t \geq 0$, $t = 0$ den Abschusszeitpunkt und $g = 9,81\ \mathrm{m/s^2}$ die Erdbeschleunigung. Die Torlinie ist genau $x_0 = 16$ m vom Abschussort entfernt. Zuerst lösen wir die zweite Gleichung nach dem Zeitpunkt t_0 auf, zu dem der Ball die Torlinie überschreitet, also die Strecke $x(t_0) = x_0$ zurückgelegt hat. Wir erhalten

$$t_0 = \frac{x_0}{v\cos\alpha} .$$

Nun entscheidet die Flughöhe $h(t_0)$ zum Zeitpunkt t_0 darüber, ob die Latte getroffen wird oder nicht. Durch Einsetzen der Formel für t_0 in die erste Gleichung von (6.42) erhält man die Flughöhe

$$h(t_0) = H(v,\alpha) = x_0\tan\alpha - \frac{1}{2}g\frac{x_0^2}{v^2\cos^2\alpha}$$

als Funktion $H(v,\alpha)$ von Abschussgeschwindigkeit und -winkel. Idealerweise sollte der Ball die Torlatte mittig treffen. Da die untere Kante der Latte in $2,44$ m Höhe liegt, und die Latte $0,12$ m breit ist, muss der Ball dazu die Torlinie in genau $h_0 = 2,50$ m Höhe überqueren. Abschussgeschwindigkeit v_0 und -winkel α_0 sollten daher so aufeinander abgestimmt sein, daß genau $H(v_0,\alpha_0) = h_0$ erreicht wird. Gibt man beispielsweise den Abschusswinkel

$$\alpha_0 = \pi/4 = 45^o$$

vor, so erhält man durch Auflösen von $H(v_0,\alpha_0) = h_0$ nach v_0 die passende Abschussgeschwindigkeit

$$v_0 = \frac{\sqrt{g}}{\sqrt{x_0 - h_0}}x_0 \approx 13,64\ \mathrm{m/s} \approx 49,10\ \mathrm{km/Std} . \tag{6.42}$$

Ronaldinho müsste den Ball also im Winkel α_0 und mit der Geschwindigkeit v_0 schießen, um exakt die Mitte der Latte zu treffen. Nun springt der Ball aber auch dann noch zum Schützen zurück (und nicht etwa ins Tor), wenn

$$H(\tilde{v}_0,\tilde{\alpha}_0) \in [h_0 - \Delta h, h + \Delta h] \quad \text{mit} \quad \Delta h = 0,04\ \mathrm{m} \tag{6.43}$$

vorliegt. Ronaldinho kann sich also kleine Fehler leisten, muss aber dabei die Toleranz (6.43) einhalten. Nun weiss man aus Experimenten, daß kein Spieler und keine Spielerin auf der Welt eine Streuung von

$$\Delta v = 0.28\ \mathrm{m/s} \approx 1\ \mathrm{km/Std} \qquad \text{und} \qquad \Delta\alpha = \pi/180 = 1^o \tag{6.44}$$

um die Werte v_0 und α_0 vermeiden kann, auch Ronaldinho nicht. Wir wollen sehen, ob das reicht.

[3] Den Spot kann man beispielsweise unter der Überschrift „Ronaldinho spielt mit einer Latte Pingpong“ unter http://www.youtube.com/watch?v=zcQvrK7pJZo finden.

Störung der Abschussgeschwindigkeit v_0. Wir erleichtern Ronaldinho die Arbeit (und vereinfachen unser Problem) erst einmal dadurch, daß wir den exakten Abschusswinkel $\alpha_0 = \pi/4$ festhalten und nur Störungen

$$|v_0 - \tilde{v}_0| \leq \Delta v$$

von v_0 zulassen. Um die Auswirkung dieser Störungen auf den Ball zu ermitteln, berechnen wir die absolute Kondition $\kappa_{\text{abs},v}$ der Auswertung der Funktion $f(v) := H(v, \alpha_0)$ an der Stelle $v = v_0$. Dazu verwenden wir Satz 6.7. Wir berechnen also die Ableitung

$$\partial_v H(v, \alpha_0) = 2g\frac{x_0^2}{v^3}$$

von $H(v, \alpha_0)$ nach v, setzen $v = v_0$ ein, verwenden die Formel (6.42) und erhalten schließlich

$$\kappa_{\text{abs},v} = \partial_v H(v_0, \alpha_0) = (x_0 - h_0)^{3/2}\frac{2}{\sqrt{g}x_0} \approx 1.98 \text{ s} . \tag{6.45}$$

Es gilt also bis auf Terme höherer Ordnung

$$|H(v_0, \alpha_0) - H(\tilde{v}_0, \alpha_0)| \leq \kappa_{\text{abs},v}\,|v_0 - \tilde{v}_0| \leq 1{,}98 \text{ s} \cdot 0{,}28 \text{ m/s} \approx 0{,}55 \text{ m} .$$

Die Genauigkeit (6.44) reicht also bei weitem nicht aus, um die Latte ausreichend genau zu treffen.

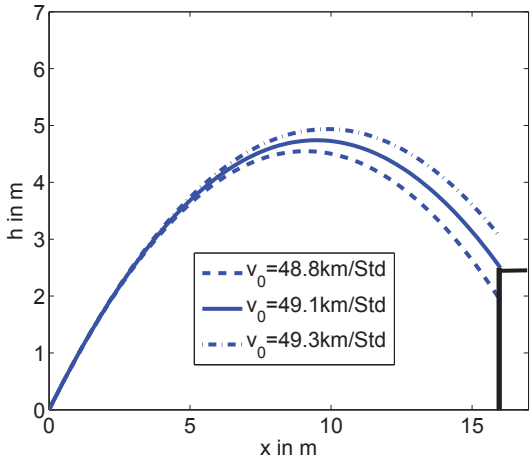

Abb. 6.3 Flugkurven der Bälle mit verschiedenen Abschussgeschwindigkeiten und konstantem Anschusswinkel $\alpha_0 = 45^o$. Der schwarze Kasten am rechten Rand soll das Tor darstellen.

Störung des Abschusswinkels α_0. Nun halten wir die Abschussgeschwindigkeit v_0 fest und erlauben nur Störungen

$$|\alpha_0 - \tilde{\alpha}_0| \leq \Delta\alpha$$

des idealen Abschusswinkels α_0. Deren Auswirkungen auf die Treffergenauigkeit beschreibt die absolute Kondition $\kappa_{\text{abs},\alpha}$ der Auswertung der Funktion $g(\alpha) := H(v_0, \alpha)$ an der Stelle $\alpha = \alpha_0$. Die erhält man nach Satz 6.7 durch Berechnung der Ableitung

$$\partial_\alpha H(v_0, \alpha) = x_0(1 + \tan^2(\alpha)) - g\frac{x_0^2 \sin(\alpha)}{v_0^2 \cos^3(\alpha)} = x_0(1 + \tan^2(\alpha)) - (x_0 - h_0)\frac{\sin(\alpha)}{\cos^3(\alpha)}$$

von $H(v_0, \alpha)$ nach α und Einsetzen von α_0. Das Ergebnis ist

$$\kappa_{\text{abs},\alpha} = 2h_0 = 5 \text{ m} .$$

Es gilt also bis auf Terme höherer Ordnung

$$|H(v_0,\alpha_0)-H(v_0,\tilde{\alpha}_0)| \le \kappa_{\text{abs},\alpha}|\alpha_0-\tilde{\alpha}_0| \le 2h_0\Delta\alpha = 5\cdot\tfrac{\pi}{180}\text{ m} \approx 0{,}087\text{ m}\,.$$

Das ist gar nicht so schlecht, aber auch nicht genau genug.

Störung von Abschussgeschwindigkeit v_0 und -winkel α_0. Mit Blick auf Lemma 6.14 addieren sich die Auswirkungen der Störungen von v_0 und α_0, ähnlich wie zuvor in den Abschnitten 6.3 und 6.5. Bis auf Terme höherer Ordnung erhält man also

$$\begin{aligned}|H(v_0,\alpha_0)-H(\tilde{v}_0,\tilde{\alpha}_0)| &\le \kappa_{\text{abs},v}|v_0-\tilde{v}_0| + \kappa_{\text{abs},\alpha}|\alpha_0-\tilde{\alpha}_0| \\ &\le \kappa_{\text{abs},v}\Delta v + \kappa_{\text{abs},\alpha}\Delta\alpha \approx 0{,}63\text{ m}\,.\end{aligned}$$

Das ist mehr als eine Größenordnung schlechter als nötig. Um dennoch viermal hintereinander zu treffen, hätte Ronaldinho also mehr Glück als Verstand gebraucht. Letztlich hat er auch zugegeben, dass bei dem Spot nachgeholfen wurde.

Ronaldinhos relative Kondition. Aufmerksamen Lesern ist sicher aufgefallen, daß die absolute Kondition $\kappa_{\text{abs},v}$ und $\kappa_{\text{abs},\alpha}$ die Einheiten s und m haben. Der Wert der absoluten Kondition ist also nicht nur von dem gegebenen Problem sondern auch von der Wahl der Einheiten abhängig. Wegen der in Satz 6.13 formulierten Skalierungsinvarianz ist das beim relativen Fehlerkonzept nicht der Fall. Das ist ein großer Vorteil.

Aus Satz 6.15 erhalten wir im vorliegenden Fall

$$\kappa_{\text{rel},v} = \frac{|v_0|}{|H(v_0,\alpha_0)|}\kappa_{\text{abs},v} = 2\frac{x_0-h_0}{h_0} = 10{,}8$$

und

$$\kappa_{\text{rel},\alpha} = \frac{|\alpha_0|}{|H(v_0,\alpha_0)|}\kappa_{\text{abs},\alpha} = 2\alpha_0 = \frac{\pi}{2}\,.$$

Beim Treffen der Latte hilft Ronaldinho allerdings auch das relative Fehlerkonzept nicht. Die erforderliche relative Toleranz

$$\frac{|H(v_0,\alpha_0)-H(\tilde{v}_0,\tilde{\alpha}_0)|}{|H(v_0,\alpha_0)|} \le \frac{\Delta h}{h_0} = 0{,}016$$

wird wegen der bis auf Terme höherer Ordnung gültigen Abschätzung

$$\begin{aligned}\frac{|H(v_0,\alpha_0)-H(\tilde{v}_0,\tilde{\alpha}_0)|}{|H(v_0,\alpha_0)|} &\le \kappa_{\text{rel},v}\frac{|v_0-\tilde{v}_0|}{|v_0|} + \kappa_{\text{rel},\alpha}\frac{|\alpha_0-\tilde{\alpha}_0|}{|\alpha_0|} \\ &\le \kappa_{\text{rel},v}\frac{\Delta v}{|v_0|} + \kappa_{\text{rel},\alpha}\frac{\Delta\alpha}{|\alpha_0|} \approx 0{,}24\end{aligned}$$

wieder um mehr als eine Größenordnung verfehlt.

6.7 Aufgaben

Aufgabe 6.1 *Beweisen Sie für x, $\text{rd}(x) \neq 0$ die Beziehung*

$$\frac{|x-\text{rd}(x)|}{|x|} = \frac{|\text{rd}(x)-x|}{|\text{rd}(x)|} + o(eps)\,,$$

wobei eps die Maschinengenauigkeit bezeichnet.

Aufgabe 6.2 *Definieren und berechnen Sie die absolute Kondition der Grundrechenarten.*

Aufgabe 6.3 1. *Zeigen Sie, daß für die absolute Kondition der Funktion $f = g+h$ die Abschätzung*

$$\kappa_{\text{abs}}(f,x) \le \kappa_{\text{abs}}(h,x) + \kappa_{\text{abs}}(g,x)$$

gilt.

2. *Verwenden Sie dieses Resultat, um die absolute und die relative Kondition der Auswertung von $f(x) = x^5 + |x^3|$ abzuschätzen.*

3. *Berechnen Sie zudem die absolute und die relative Kondition der Auswertung von* $f(x) = \sin^2(x) + \cos^2(x)$ *in* $x = 0$. *Finden Sie ausserdem ein x für das die Abschätzung aus a) nicht scharf ist.*

Aufgabe 6.4 *Finden Sie eine Funktion* $f : \mathbb{R} \to \mathbb{R}$ *und vier verschiedene Werte* x_1, x_2, x_3, x_4 *in der Weise, dass hinsichtlich der Kondition der Funktionsauswertung an der jeweiligen Stelle jede der folgenden vier Möglichkeiten einmal auftritt:*

1. κ_{abs} *groß,* κ_{rel} *klein*
2. κ_{abs} *klein,* κ_{rel} *groß*
3. κ_{abs} *groß,* κ_{rel} *groß*
4. κ_{abs} *klein,* κ_{rel} *klein*

Verwenden Sie nicht die Funktion $f(x) = x^2$ *oder Vielfache davon.*

Aufgabe 6.5 1. *Beweisen Sie folgende Verallgemeinerung von Satz 6.10: Seien* $f_i : D_i \to \mathbb{R}$ *Funktionen mit* $f_i(D_i) \subset D_{i+1}$ *und* $f_i(x_i) = x_{i+1}$ *für alle* $i = 1, \ldots, n-1$. *Bezeichne weiter* $\kappa_{abs}(f_i, x_i)$ *die absolute Kondition von* f_i *an* x_i. *Dann gilt für die absolute Kondition* κ_{abs} *der Funktion* $f(x) = f_n \circ \ldots \circ f_1(x)$ *im Punkt* x_1

$$\kappa_{abs} \leq \kappa_{abs}(f_n, x_n) \cdots \kappa_{abs}(f_1, x_1) \ .$$

Sind die f_i *allesamt differenzierbar, so gilt sogar Gleichheit.*

2. *Berechnen Sie mit Hilfe von Teil 1. eine obere Schranke für die absolute Kondition der beiden Funktionen* $f(x) = x^3 \cos(x)$ *und* $g(x) = e^{|\sin(x)|}$.
3. *Verwenden Sie das Resultat aus Teil 1., um die absolute und die relative Kondition der Auswertung von* $f(x) = x^5 + |x^3|$ *abzuschätzen.*
4. *Berechnen Sie zudem die absolute und die relative Kondition der Auswertung von* $f(x) = \sin^2(x) + \cos^2(x)$ *in* $x = 0$. *Finden Sie ausserdem ein x für das die Abschätzung aus Teil 1. nicht scharf ist.*

Aufgabe 6.6 *Geben Sie eine Funktion an, deren Auswertung mit einfacher Genauigkeit* eps *nicht bis auf eine relative Genauigkeit von* $\text{TOL} = 10^{-2}$ *möglich ist.*

Aufgabe 6.7 *Zeigen Sie die Sätze 6.4 und 6.5 mit Hilfe von Lemma 6.14.*

Aufgabe 6.8 *Berechnen Sie die absolute und relative Kondition der Lösung der nichtlinearen Gleichung* $\lambda g(x^*) = \lambda y^*$ *für beliebiges, reelle Zahlen* $\lambda \neq 0$ *und diskutieren Sie die Ergebnisse.*

Aufgabe 6.9 *Bestimmen Sie die relative Kondition* κ_{rel} *der Lösung der Gleichung*

$$x^* \in (0, \infty) : \qquad \frac{1}{x^*} = y^* \ .$$

Hängt κ_{rel} *von* y^* *ab?*

Aufgabe 6.10 1. *Berechnen Sie unter den Voraussetzungen von Lemma 6.32 die absolute Kondition* κ^k_{abs} *der durch die Drei-Term-Rekursion* (6.32) *erzeugten Folgenglieder* x_k *bei Störung von* x_0.
2. *Unter welcher Bedingung ist* κ^k_{abs} *für* $k \to \infty$ *beschränkt?*

Aufgabe 6.11 1. *Berechnen Sie unter den Voraussetzungen von Lemma 6.32 die absolute Kondition der durch die Drei-Term-Rekursion* (6.32) *erzeugten Folgenglieder* x_k *bei Störung von b.*

Aufgabe 6.12 *a) Auf welchen Wert kann Ronaldinho die absolute Kondition* κ_{abs}, v *durch Wahl eines anderen Abschusswinkels* $\alpha_0 < \pi/4$ *drücken, wenn er über den härtesten Schuss der Welt mit 212 km/Std (Ronny Heberson Furtado de Araújo bei einem Freistoß im Spiel Sporting Lissabon gegen Naval in der Saison 2006/2007) verfügen würde?*
b) Minimieren Sie unter der Bedingung von Teil a) durch Wahl von α_0 *und* v_0 *die Summe* $\kappa_{\text{rel},v} + \kappa_{\text{rel}\alpha}$ *und diskutieren Sie die Konsequenzen für die Trefferquote.*