5 Nichtlineare Gleichungen

In diesem Kapitel betrachten wir numerische Verfahren zur approximativen Lösung *nichtlinearer* Gleichungen oder Gleichungssystemе bzw. zur Bestimmung von Nullstellen skalarer oder auch vektor-wertiger Funktionen.

Es sei f eine (reellwertige) stetige Funktion auf einem Intervall $I = [a, b]$. Das einfachste Verfahren zur Bestimmung von Nullstellen von f beruht auf der folgenden Konsequenz des Zwischenwertsatzes für stetige Funktionen: *Existiert ein Teilintervall* $I_0 = [a_0, b_0] \subset I$ *mit* $f(a_0)f(b_0) < 0$, *so hat* f *in* I_0 *mindestens eine Nullstelle.* Die sog. „Intervallschachtelung“ erzeugt nun ausgehend von einem solchen I_0 eine Folge von Intervallen $I_t = [a_t, b_t]$, $t = 1, 2, \ldots$, welche jeweils mindestens eine Nullstelle von f enthalten, durch die Iteration

$$x_t := \tfrac{1}{2}(a_t + b_t)\,, \qquad (\, f(x_t) = 0 \;\Rightarrow\; \text{STOP}\,),$$

mit der Auswahlvorschrift

$$\begin{aligned} f(a_t)f(x_t) < 0 \quad &\Rightarrow \quad a_{t+1} := a_t\,, \quad b_{t+1} := x_t\,, \\ f(a_t)f(x_t) > 0 \quad &\Rightarrow \quad a_{t+1} := x_t\,, \quad b_{t+1} := b_t. \end{aligned}$$

Offenbar ist dann $a_t \le a_{t+1} \le b_{t+1} \le b_t$ und

$$|b_{t+1} - a_{t+1}| = \tfrac{1}{2}|b_t - a_t| = 2^{-t-1}|b_0 - a_0|. \tag{5.0.1}$$

Die monotonen Zahlenfolgen $(a_t)_{t\in\mathbb{N}}$, $(b_t)_{t\in\mathbb{N}}$ konvergieren gegen ein $z \in I_0$, welches wegen $f(z)^2 = \lim_{t\to\infty} f(a_t)f(b_t) \le 0$ notwendig Nullstelle von f ist. Dieses Verfahren ist numerisch sehr stabil, aber auch sehr langsam; für $b_0 - a_0 = 1$ erhält man z. B. aus der obigen a priori Abschätzung ($2^{-10} \le 10^{-3}$):

$$|x_9 - z| < 10^{-3}\,, \quad |x_{19} - z| < 10^{-6}\,, \quad |x_{29} - z| < 10^{-9}.$$

Die Intervallschachtelung für stetige Funktionen liefert stets eine Nullstelle, sofern für das Startintervall ein Vorzeichenwechsel vorliegt. Dieses Vorgehen ist naturgemäß auf *reelle* Funktionen beschränkt. Die im Folgenden betrachteten Verfahren sind dagegen teilweise auch für komplexwertige Funktionen anwendbar.

5.1 Das Newton-Verfahren im $\mathbb{R}^1$

Ist die gegebene Funktion f auf dem Intervall $[a, b]$ stetig differenzierbar, so kann diese Zusatzinformation zur effizienteren Berechnung einer Nullstelle verwendet werden. Das (klassische) Newton-Verfahren" (auch Newton-Raphson[1]-Verfahren" genannt) ist moti-

[1]Joseph Raphson (1648–1715): Englischer Mathematiker; an der Universität Cambride; sein Buch Änalysis Aequationum Universalis" (1660) enthält bereits die Newton-Methode (50 Jahre vor Newton selbst); übersetzte einige Werke Newtons (von Latein nach Englisch); eigene Beitrge zur Analysis.

viert durch die folgende grafische Überlegung (s. Abb. 5.1):

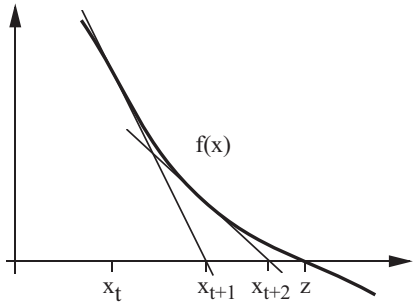


Abbildung 5.1: *Geometrische Interpretation des Newton-Verfahrens*

Im Punkt x_t wird die Tangente an $f(x)$ berechnet und deren Schnittpunkt mit der x-Achse als neue Näherung x_{t+1} für die Nullstelle z von f genommen. Die Tangente ist gegeben durch die Gleichung

$$T(x) = f'(x_t)(x - x_t) + f(x_t).$$

Ihre Nullstelle x_{t+1} ist bestimmt durch

$$x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)}. \tag{5.1.2}$$

Diese Iteration ist offenbar möglich, wenn die Ableitungswerte $f'(x_t)$ nicht zu klein werden. In dieser Form gestattet das Newton-Verfahren es also, *einfache* Nullstellen zu approximieren.

Satz 5.1 (Newton-Verfahren): *Die Funktion $f \in C^2[a,b]$ habe im Innern des Intervalls $[a,b]$ eine Nullstelle z, und es sei*

$$m := \min_{a \le x \le b} |f'(x)| > 0\,, \quad M := \max_{a \le x \le b} |f''(x)|.$$

Sei $\rho > 0$ so gewählt, dass

$$q := \frac{M}{2m}\,\rho < 1, \qquad K_\rho(z) := \{x \in \mathbb{R} |\, |x - z| \le \rho\} \subset [a,b]. \tag{5.1.3}$$

Dann sind für jeden Startpunkt $x_0 \in K_\rho(z)$ die Newton-Iterierten $x_t \in K_\rho(z)$ definiert und konvergieren gegen die Nullstelle z. Dabei gelten die a priori Fehlerabschätzung

$$|x_t - z| \le \frac{2m}{M}\, q^{(2^t)}\,, \quad t \in \mathbb{N}, \tag{5.1.4}$$

und die a posteriori Fehlerabschätzung

$$|x_t - z| \le \frac{1}{m}\,|f(x_t)| \le \frac{M}{2m}\,|x_t - x_{t-1}|^2, \quad t \in \mathbb{N}. \tag{5.1.5}$$

Beweis: Der Beweis erfordert einige Vorbereitungen. Für Punkte $x, y \in [a,b]$, $x \neq y$, gilt aufgrund des Mittelwertsatzes der Differentialrechnung mit einem $\zeta \in [x,y]$:

$$\left|\frac{f(x)-f(y)}{x-y}\right| = |f'(\zeta)| \ge m,$$

und folglich

$$|x-y| \le \frac{1}{m}\,|f(x)-f(y)|.$$

(Die Nullstelle z von f ist also die einzige in $[a,b]$.) Weiter gilt die Taylor-Formel mit Restglied zweiter Ordnung:

$$f(y) = f(x) + (y-x)f'(x) + \underbrace{(y-x)^2 \int_0^1 f''(x+s(y-x))(1-s)\,ds}_{=:\,R\,(y;\,x)}.$$

Mit Hilfe der Voraussetzung erhalten wir

$$|R(y;x)| \le M\,|y-x|^2 \int_0^1 (1-s)\,ds = \frac{M}{2}|y-x|^2.$$

Für $x \in K_\rho(z)$ setzen wir $g(x) := x - f'(x)^{-1}f(x)$ und finden

$$g(x) - z = x - \frac{f(x)}{f'(x)} - z = -\frac{1}{f'(x)}\,\underbrace{\{f(x) + (z-x)f'(x)\}}_{=\,-R\,(z;\,x)}.$$

Also ist

$$|g(x) - z| \le \frac{M}{2m}\,|x-z|^2 \le \frac{M}{2m}\,\rho^2 < \rho, \tag{5.1.6}$$

d. h.: $g(x) \in K_\rho(z)$. Die Abbildung g bildet die Menge $K_\rho(z)$ in sich ab. Für $x_0 \in K_\rho(z)$ bleiben also alle Newton-Iterierten in $K_\rho(z)$. Setzt man

$$\rho_t := \frac{M}{2m}\,|x_t - z|,$$

so impliziert (5.1.6), dass

$$\rho_t \le \rho_{t-1}^2 \le \ldots \le \rho_0^{2^t}, \quad |x_t - z| \le \frac{2m}{M}\,\rho_0^{2^t}.$$

Für $\rho_0 = \frac{M}{2m}\,|x_0 - z| \le \frac{M}{2m}\,\rho < 1$ liegt also die Konvergenz $x_t \to z (t \to \infty)$ vor mit der behaupteten a priori Fehlerabschätzung. Zum Beweis der a posteriori Fehlerabschätzung setzt man in der Taylor-Formel $y = x_t,\ x = x_{t-1}$, und erhält

$$f(x_t) = \underbrace{f(x_{t-1}) + (x_t - x_{t-1}) f'(x_{t-1})}_{=\,0} + R(x_t; x_{t-1})$$

bzw.

$$|x_t - z| \le \frac{1}{m}\,|f(x_t) - \underbrace{f(z)}_{=\,0}| \le \frac{M}{2m}\,|x_t - x_{t-1}|^2.$$

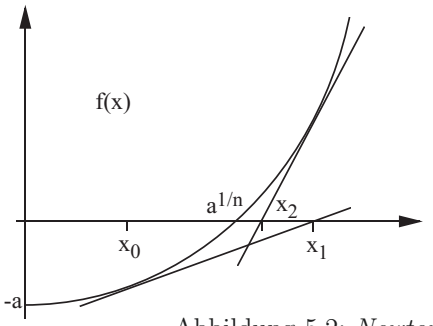
Dies vervollständigt den Beweis. Q.E.D.

Für eine zweimal stetig differenzierbare Funktion f existiert zu jeder einfachen Nullstelle z $(f(z) = 0,\ f'(z) \neq 0)$ stets eine (möglicherweise sehr kleine) Umgebung $K_\rho(z)$, für welche die Voraussetzungen von Satz 5.1 erfüllt sind. Das Problem beim Newton-Verfahren ist also die Bestimmung eines im Ëinzugsbereich" der Nullstelle z gelegenen Startpunktes x_0. Ist ein solcher einmal gefunden, so konvergiert das Newton-Verfahren enorm schnell gegen die Nullstelle z: Im Fall $q \le \frac{1}{2}$ gilt z.B. nach nur 10 Iterationsschritten bereits $(2^{10} > 1.000)$

$$|x_{10} - z| \le \frac{2m}{M}\,q^{1.000} \sim \frac{2m}{M}\,10^{-300}.$$

Beispiel 5.1: Newton-Verfahren zur Wurzelberechnung:
Die n-te Wurzel einer Zahl $a > 0$ ist Nullstelle der Funktion $f(x) = x^n - a$. Das Newton-Verfahren zur Berechnung von $z = \sqrt[n]{a} > 0$ hat die Gestalt

$$x_{t+1} = x_t - \frac{x_t^n - a}{n x_t^{n-1}} = \frac{1}{n}\left\{(n-1)\,x_t + \frac{a}{x_t^{n-1}}\right\}. \tag{5.1.7}$$



$$x_0 > 0 \;\Rightarrow\; \left\{ \begin{array}{l} x_t > \sqrt[n]{a}\,, \quad t \in \mathbb{N}\,. \\ \sqrt[n]{a} < x_{t+1} < x_t \end{array} \right.$$

Abbildung 5.2: *Newton-Iteration zur Wurzelberechnung*

Aufgrund von Satz 5.1 konvergiert $x_t \to z\,(t \to \infty)$, wenn nur x_0 nahe genug bei z gewählt wird. Bei diesem einfachen Beispiel kann aber mit Hilfe der folgenden geometrischen Betrachtung die Konvergenz für jeden Startpunkt $x_0 > 0$ gesichert werden (s. Abb. 5.1) Die monoton fallende Folge $(x_t)_{t\in\mathbb{N}}$ konvergiert notwendig gegen $\sqrt[n]{a}$. Für hinreichend großes t ist dann x_t im Einzugsbereich der Nullstelle z, und die Fehlerabschätzung von Satz 5.1 gelten mit diesem x_t als Startpunkt. Auf diese Weise wird auf vielen Rechenern die Wurzel $\sqrt[n]{a}$ berechnet.

Für den Spezialfall $n = 2$ wollen wir den Einzugsbereich der quadratischen Konvergenz des Newton-Verfahrens bestimmen. Es gilt

$$x_{t+1} - \sqrt{a} = \frac{1}{2}\left\{x_t + \frac{a}{x_t}\right\} - \sqrt{a} = \frac{1}{2x_t}\left\{x_t^2 + a - 2x_t\sqrt{a}\right\} = \frac{1}{2x_t}\,(x_t - \sqrt{a})^2,$$

also für $t \geq 1$ (wegen $x_t > \sqrt{a}$)

$$|x_{t+1} - \sqrt{a}| \leq \frac{1}{2\sqrt{a}}\,|x_t - \sqrt{a}|^2.$$

Quadratische Konvergenz liegt vor für Startwerte x_0 mit der Eigenschaft

$$\frac{1}{2\sqrt{a}}\,|x_0 - \sqrt{a}| < 1 \quad \text{bzw.} \quad |x_0 - \sqrt{a}| < 2\sqrt{a}.$$

Aus der Ungleichung $\sqrt{a} \leq x_t$, $t \in \mathbb{N}$, ergibt sich noch die Beziehung

$$\frac{a}{x_t} \leq \sqrt{a} \leq x_t,$$

was für ein Abbruchkriterium verwendet werden kann:

$$0 \leq e_t := x_t - \frac{a}{x_t} \leq \varepsilon \quad \Longrightarrow \quad \text{STOP}.$$

Die folgende Tabelle zeigt das Konvergenzverhalten der Newton-Iteration zur Berechnung von $x = \sqrt{2} = 1.414213562373095\ldots$ (16-stellige Rechnung). In jedem Iterationsschritt verdoppelt sich die Anzahl der richtigen Dezimalstellen:

$$\begin{aligned}
x_0 &= 2 \\
x_1 &= \underline{1}.5 \\
x_2 &= \underline{1.416}\,, && e_2 \leq 5 \cdot 10^{-3} \\
x_3 &= \underline{1.41421}568627451\,, && e_3 \leq 5 \cdot 10^{-6} \\
x_4 &= \underline{1.41421356137}469\,, && e_4 \leq 5 \cdot 10^{-12}.
\end{aligned}$$

Bemerkung 5.1: Die Bedingungen von Satz 5.1 lassen sich so modifizieren, dass auf die Voraussetzung der Existenz einer Nullstelle verzichtet werden kann, und, ähnlich wie beim Banachschen Fixpunktsatz, die Konvergenz der Newton-Folge gegen eine (lokal eindeutige) Nullstelle folgt. Diese Variante von Satz 5.1, der sog. „Satz von Newton-Kantorowitsch“, wird im Rahmen der Diskussion des Newton-Verfahrens im $\mathbb{R}^n$ bewiesen.

Bemerkung 5.2: Das Hauptproblem bei der Durchführung des Newton-Verfahrens ist die Bestimmung eines geeigneten Startwertes x_0, da der Einzugsbereich der quadratischen Konvergenz in der Praxis häufig sehr klein ist. Deshalb arbeitet man meist mit dem sog. „gedämpften Newton-Verfahren"

$$x_{t+1} = x_t - \lambda_t \frac{f(x_t)}{f'(x_t)}, \tag{5.1.8}$$

mit einem „Dämpfungsparameter" $\lambda_t \in (0, 1]$. Die geeignete Wahl dieses Parameters ist eine „Wissenschaft" für sich. Sie wird später im Zusammenhang mit dem Newton-Verfahren im $\mathbb{R}^n$ diskutiert werden.

Mehrfache Nullstellen

Wir betrachten nun den kritischen Fall, dass mit dem Newton-Verfahren eine mehrfache Nullstelle berechnet werden soll. Sei dazu zunächst z eine zweifache Nullstelle der Funktion f, d. h.: $f(z) = f'(z) = 0,\ f''(z) \neq 0$. Für die Newton-Iteration gilt dann

$$x_{t+1} = x_t - \frac{f(x_t) - f(z)}{f'(x_t) - f'(z)} = x_t - \frac{f'(\zeta_t)}{f''(\eta_t)}$$

mit Zwischenpunkten $\zeta_t, \eta_t \in [x_t, z]$. Der Quotient $f(x_t)/f'(x_t)$ bleibt also für $x_t \to z$ wohl definiert. Sei nun allgemein z eine p-fache Nullstelle der Funktion $f \in C^{p+1}[a, b]$:

$$f(z) = \ldots = f^{(p-1)}(z) = 0\,, \quad f^{(p)}(z) \neq 0.$$

Aus der Taylor-Formel um z

$$f(x) = \underbrace{f(z)}_{=0} + \ldots + \frac{1}{(p-1)!}(x-z)^{p-1} \underbrace{f^{(p-1)}(z)}_{=0} + (x-z)^p \underbrace{\frac{1}{p!} f^{(p)}(\zeta_x)}_{=:\, Q(z;x)}$$

folgt durch Ableiten

$$f'(x) = Q'(z;x)(x-z)^p + p\, Q(z;x)(x-z)^{p-1}.$$

Also ist für $f'(x) \neq 0$:

$$\begin{aligned} \frac{f(x)}{f'(x)} &= \frac{(x-z)Q(z;x)}{Q'(z;x)(x-z) + p\, Q(z;x)} \\ &= \frac{x-z}{p} - \frac{1}{p}(x-z)^2 \frac{Q'(z;x)}{Q'(z;x)(x-z) + p\, Q(z;x)}. \end{aligned}$$

Für den Iterationsansatz

$$x_{t+1} = x_t - \alpha\, \frac{f(x_t)}{f'(x_t)} \tag{5.1.9}$$

folgt dann

$$\begin{aligned} x_{t+1} - z &= x_t - z - \alpha \frac{f(x_t)}{f'(x_t)} \\ &= (x_t - z)\Big(1 - \frac{\alpha}{p}\Big) + (x_t - z)^2 \frac{\alpha Q'(z; x_t)}{pQ'(z; x_t)(x_t - z) + p^2 Q(z; x_t)}. \end{aligned}$$

Bei der Wahl von $\alpha = p$ erhält man für das so modifizierte Newton–Verfahren

$$x_{t+1} = x_t - p\frac{f(x_t)}{f'(x_t)}. \tag{5.1.10}$$

ein analoges „quadratisches“ Konvergenzverhalten wie im Fall einer einfachen Nullstelle.

Vereinfachtes Newton-Verfahren

Ist z Nullstelle einer stetig differenzierbaren Funktion f, so konvergiert die Newton-Iteration

$$x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)} \to z \quad (t \to \infty),$$

wenn x_0 hinreichend nahe bei z gewählt war. Jeder Iterationsschritt erfordert die Auswertung der Ableitung $f'(x_t)$, was bei komplizierten (möglicherweise auch nur implizit definierten) Funktionen f unter Umständen zuviel Aufwand erfordert. In solchen Fällen geht man zum sog. „vereinfachten Newton-Verfahren“ über

$$x_{t+1} = x_t - \frac{f(x_t)}{f'(c)} \tag{5.1.11}$$

mit einem festen, geeignet gewählten Punkt c. Diese Iteration ist Spezialfall der allgemeineren „Fixpunktiteration“

$$x_{t+1} = x_t + \sigma\, f(x_t) \tag{5.1.12}$$

mit einer geeigneten Zahl $\sigma \in \mathbb{R}$, $\sigma \neq 0$, zur Berechnung einer Nullstelle von f. Konvergiert hier $x_t \to z\,(t \to \infty)$, so gilt im Limes

$$\begin{array}{ccccccc} x_{t+1} & = & x_t & + & \sigma f(x_t) & & \\ \downarrow & & \downarrow & & \downarrow & & (t \to \infty) \\ z & = & z & + & \sigma f(z) & & \end{array}$$

d. h.: z ist „Fixpunkt“ der Abbildung $g(x) = x + \sigma f(x)$ und wegen $\sigma \neq 0$ notwendig Nullstelle von f. Der Vorteil der obigen Fixpunktiteration besteht in ihrer ableitungsfreien Form. Kriterien für die Konvergenz einer Fixpunktiteration $x_{t+1} = g(x_t)$, $t = 0, 1, 2, \ldots$, werden wir später in einem etwas allgemeineren Rahmen herleiten.

Wir wollen nun noch das Newton-Verfahren zur Berechnung von Nullstellen von Polynomen

$$p(x) = a_0 + a_1 x + \ldots + a_n x^n\,, \quad a_n \neq 0,$$

spezialisieren. Zunächst verschafft man sich etwa (im Reellen) mit der Intervallschachtelung einen groben Überblick über die Lage der Nullstellen. Nach Vorgabe einer Fehlertoleranz $\varepsilon \gg \text{eps}$ lautet der Newton-Algorithmus dann wie folgt:

1. Wahl eines Startwertes x;

2. Auswertung von $p(x)$ und $p'(x)$ mit dem Horner Schema:

$$\begin{aligned} i = n, n-1, \ldots, 0: \quad & \alpha_i = a_i + \alpha_{i+1} x\,, \quad \beta_i = \alpha_i + \beta_{i+1} x \\ & (\alpha_{n+1} = \beta_{n+1} = 0) \\ & p(x) = \alpha_0\,, \quad p'(x) = \beta_1\,, \\ (\beta_1 = 0: \quad & \text{Startwert ändern}) \quad \text{sei} \quad \beta_1 \neq 0\,; \end{aligned}$$

3. Newton-Korrektur $q = \dfrac{\alpha_0}{\beta_1}\,, \quad |q| \leq \varepsilon \begin{cases} \text{ja:} & x \quad \text{wird akzeptiert;} \\ \text{nein:} & \text{Iterationsschritt;} \end{cases}$

4. Iterationsschritt $x := x - q\,,$ weiter mit (2).

5.2 Das Konvergenzverhalten iterativer Verfahren

Das Newton-Verfahren besitzt lokal in der Umgebung einer Nullstelle die charakteristische Konvergenzeigenschaft

$$|x_t - z| \;\leq\; c|x_{t-1} - z|^2. \tag{5.2.13}$$

Man nennt es daher „quadratisch konvergent“ oder auch “von 2-ter Ordnung“.

Definition 5.1: *Allgemein spricht man bei einem Iterationsverfahren zur Berechnung einer Größe z von Konvergenz mit der „Ordnung“ α, $\alpha \geq 1$, wenn gilt*

$$|x_t - z| \;\leq\; c|x_{t-1} - z|^{\alpha}\,, \tag{5.2.14}$$

mit einer festen Konstante $c > 0$. Im Fall $\alpha = 1$, d. h. „linearer“ Konvergenz, nennt man die „beste“ Konstante c „lineare Konvergenzrate“. Gilt die Abschätzung

$$|x_t - z| \leq c_t |x_{t-1} - z| \tag{5.2.15}$$

mit einer Nullfolge $c_t \to 0$ $(t \to \infty)$, so spricht man von „superlinear“ Konvergenz.

Im Fall $\alpha > 1$ impliziert die Beziehung (5.2.14) wiederum Konvergenz $x_t \to z\,(t \to \infty)$, wenn der Startwert x_0 hinreichend nahe bei z liegt:

$$c^{\frac{1}{\alpha-1}}|x_t - z| \;\le\; \left[c^{\frac{1}{\alpha-1}}|x_{t-1} - z|\right]^{\alpha} \;\le\; \ldots \;\le\; \underbrace{\left[c^{\frac{1}{\alpha-1}}|x_0 - z|\right]^{\alpha^t}}_{<1\,!} \;\to\; 0.$$

Im Fall $\alpha = 1$ folgt Konvergenz für $c < 1$:

$$|x_t - z| \le c|x_{t-1} - z| \le \ldots \le c^t|x_0 - z| \to 0 \quad (t \to \infty).$$

Bei Fixpunktiterationen $x_{t+1} = g(x_t)$ mit stetig differenzierbarer Abbildung g gilt

$$\left|\frac{x_{t+1} - z}{x_t - z}\right| = \left|\frac{g(x_t) - g(z)}{x_t - z}\right| \;\to\; |g'(z)| \quad (t \to \infty),$$

d. h.: Die lineare Konvergenzrate ist asymptotisch (für $t \to \infty$) gerade gleich $|g'(z)|$. Im Falle $g'(z) = 0$ liegt also (mindestens) superlineare Konvergenz der Fixpunktiteration vor.

Definition 5.2: *Ein Fixpunkt z einer stetig differenzierbaren Abbildung g heißt „anziehend", wenn $|g'(z)| < 1$ ist, da dann die Fixpunktiteration (sog. „sukzessive Approximiert") für jeden hinreichend nahe bei z gelegenen Startwert gegen ihn konvergiert. Im Fall $|g'(z)| > 1$ heißt er äbstoßend", da er durch sukzessive Approximation i. Allg. nicht angenähert werden kann.*

Einen Hinweis zur Konstruktion von Verfahren höherer Ordnung gibt der folgende Satz.

Satz 5.2 (Iterative Verfahren): *Die Funktion g sei in einer Umgebung des Fixpunktes z p-mal stetig differenzierbar mit $p \ge 2$. Genau dann hat die Fixpunktiteration $x_{t+1} = g(x_t)$ die genaue Ordnung p, wenn*

$$g'(z) = \ldots = g^{(p-1)}(z) = 0 \quad \textit{und} \quad g^{(p)}(z) \neq 0. \tag{5.2.16}$$

Beweis: (i) Sei $g'(z) = \ldots = g^{(p-1)}(z) = 0$. Die Taylor-Formel mit dem Restglied p-ter Ordnung erhält dann im Punkt z die Form

$$x_{t+1} - z = g(x_t) - g(z) = \sum_{i=1}^{p-1} \frac{(x_t - z)^i}{i!}\, g^{(i)}(z) + \frac{(x_t - z)^p}{p!}\, g^{(p)}(\zeta_t),$$

und folglich

$$|x_{t+1} - z| \le \frac{1}{p!} \max |g^{(p)}|\, |x_t - z|^p.$$

(ii) Sei nun umgekehrt die Iteration von p-ter Ordnung, d. h.: $|x_{t+1} - z| \le c|x_t - z|^p$. Gäbe es ein minimales $m \le p-1$ mit $g^{(m)}(z) \neq 0$, aber $g^{(i)}(z) = 0$, $i = 1, \ldots, m-1$, so

konvergierte jede Iteriertenfolge $(x_t)_{t\in\mathbb{N}}$ mit hinreichend kleinem $|x_0 - z| \neq 0$ notwendig gegen z wie

$$|x_t - z| = |\frac{1}{m!}\, g^{(m)}(\zeta_t)|\, |x_{t-1} - z|^m$$

Dies impliziert aber im Widerspruch zur Annahme:

$$|g^{(m)}(z)| = \lim_{t\to\infty} |g^{(m)}(\zeta_t)| \leq c\, m! \lim_{t\to\infty} |x_t - z|^{p-m} = 0.$$

Hieraus folgt auch, dass im Fall $g'(z) = \ldots = g^{(p-1)}(z) = 0$, aber $g^{(p)}(z) \neq 0$, die Iteration nicht von höherer als p-ter (ganzzahliger) Ordnung sein kann. Q.E.D.

Beispiel 5.2: Beim Newton-Verfahren zur Bestimmung einer einfachen Nullstelle der Funktion f ist mit $g(x) = x - f'(x)^{-1} f(x)$ also

$$g'(z) = 1 - \frac{f'(z)^2 - f(z)f''(z)}{f'(z)^2} = 0,$$

und i. Allg. $g''(z) \neq 0$. Die Newton-Iteration ist also, wie wir schon gesehen haben, von 2-ter Ordnung.

Beispiel 5.3: Bei einer Fixpunktiteration von mindestens 3-ter Ordnung muss $g'(z) = g''(z) = 0$ gelten. Zur Konstruktion eines solchen Verfahrens zur Nullstellenbestimmung machen wir den Ansatz

$$g(x) = x - r(x) + s(x)r(x)^2 \quad \text{mit} \quad r(x) = \frac{f(x)}{f'(x)}.$$

Wegen $r(z) = 0$ und $r'(z) = 1$ ist hier automatisch $g'(z) = 0$. Die zusätzliche Forderung $g''(z) = 0$ wird z. B. erfüllt für

$$s(x) = \frac{r''(x)}{2r'(x)^2}.$$

Dieses Verfahren erfordert also die Auswertung der Ableitungen bis zur Ordnung 3 der Funktion f.

Zur Klärung der numerischen Bedeutung des Ordnungsbegriffes definieren wir für eine Iterationsfolge $(x_t)_{t\in\mathbb{N}}$

$$e_t := x_t - z \quad (\textit{absoluter} \text{ Fehler}), \quad \bar{e}_t := \frac{e_t}{z} \quad (\textit{relativer} \text{ Fehler für } z \neq 0).$$

Haben x_t und z die dezimalen Gleitpunktdarstellungen (mit gemeinsamen Exponenten und m gleichen Mantissenstellen)

$$\begin{aligned} z &= a_m \ldots a_1 . a_{-1} \ldots \cdot 10^s, \quad a_m \neq 0, \\ x_t &= a_m \ldots a_1 . \tilde{a}_{-1} \ldots \cdot 10^s, \end{aligned}$$

so gilt

$$|\bar{e}_t| = \left|\frac{x_t - z}{z}\right| \le 10^{-m},$$

d. h.: Die Größe

$$\rho_t := -\log_{10}|\bar{e}_t| = m$$

gibt ungefähr die Anzahl der richtigen Mantissendezimalen von x_t an. Wegen

$$|\bar{e}_{t+1}| = \left|\frac{x_{t+1} - z}{z}\right| = |g'(\zeta_t)| \left|\frac{x_t - z}{z}\right|, \quad \zeta_t \in [x_t, x_{t+1}],$$

gilt

$$\rho_{t+1} = -\log_{10}|\bar{e}_{t+1}| = -\log_{10}|g'(\zeta_t)| \underbrace{-\log_{10}|\bar{e}_t|}_{\rho_t}$$

und im Limes

$$\rho_{t+1} - \rho_t \to -\log_{10}|g'(z)| \quad (t \to \infty). \tag{5.2.17}$$

(i) Die numerische Bedeutung der „asymptotischen“ *linearen* Konvergenzrate $|g'(z)|$ einer Fixpunktiteration ist also, dass sich in jedem Iterationsschritt (für große t) die Anzahl der richtigen Mantissendezimalen um $-\log_{10}|g'(z)|$ erhöht (für $|g'(z)| \neq 0$).

(ii) Für eine Iteration p-ter Ordnung mit $p \ge 2$ gilt

$$|x_{t+1} - z| = \frac{1}{p!}|g^{(p)}(\zeta_t)|\,|x_t - z|^p, \quad t \ge 1,$$

mit $\zeta_t \to z\ (t \to \infty)$. Also ist in diesem Fall

$$|\bar{e}_{t+1}| = \left|\frac{x_{t+1} - z}{z}\right| = \underbrace{\left|\frac{1}{p!}\,g^{(p)}(\zeta_t)\right|\,|z|^{p-1}}_{=:\,\sigma_t} \underbrace{\left|\frac{x_t - z}{z}\right|^p}_{=|\bar{e}_t|^p}$$

und

$$\sigma_t \to \left|\frac{1}{p!}\,g^{(p)}(z)\right|\,|z|^{p-1} \quad (t \to \infty).$$

Es folgt die Beziehung $(\rho_t = -\log_{10}|\bar{e}_t|)$

$$\rho_{t+1} = p\,\rho_t - \log_{10}\sigma_t,$$

und hieraus wegen $\rho_t \to \infty\ (t \to \infty)$

$$\lim_{t\to\infty} \frac{\rho_{t+1}}{\rho_t} = p. \tag{5.2.18}$$

Dies lässt sich so interpretieren, dass sich bei einer Iteration p-ter Ordnung (für große t) die Anzahl der richtigen Mantissendezimalen in jedem Schritt etwa ver-p-facht. Dies

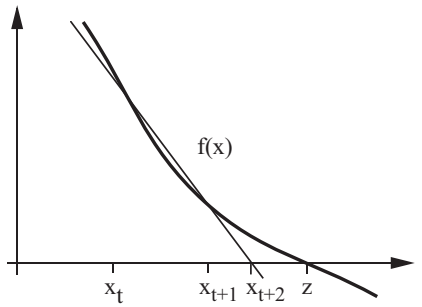
wird durch unser obiges Beispiel beim Newton-Verfahren bestätigt. Wir fassen die bisher abgeleiteten Ergebnisse zusammen:

Regel 5.1: *Bei einem „linear“ konvergenten Iterationsverfahren erhöht sich in jedem Schritt die Anzahl von „exakten“ Dezimalstellen in der Näherung in etwa um den Summanden* $|\log_{10}(g'(z))|$*; bei einer Iteration der Ordnung* $p > 1$ *ver-p-facht sich in jedem Schritt die Anzahl der „exakten“ Dezimalstellen.*

5.3 Interpolationsmethoden

Das Ziel ist die iterative Berechnung von Nullstellen ohne Auswertung von Ableitungen, aber effizienter als mit Intervallschachtelung oder einfacher sukzessiver Approximation. Dabei werden wir auf ein Iterationsverfahren mit nicht ganzzahliger Ordnung geführt.

Die „Sekantenmethode“ berechnet ausgehend von einem Paar von Werten x_{t-1}, x_t die neue Iterierte x_{t+1} als Nullstelle der Geraden (Sekante) durch die Punkte $(x_{t-1}, f(x_{t-1}))$, $(x_t, f(x_t))$ (s. Abb. 5.3):



$$s(x) = f(x_t) + (x - x_t)\frac{f(x_t) - f(x_{t-1})}{x_t - x_{t-1}}$$

Iteration:

$$x_{t+1} = x_t - f(x_t)\frac{x_t - x_{t-1}}{f(x_t) - f(x_{t-1})}\,.$$

Abbildung 5.3: *Geometrische Interpretation des Sekanten-Verfahrens*

Im Gegensatz zu den bisher betrachteten Verfahren handelt es sich hierbei um ein sog.„Zweischrittverfahren“, d. h.: Die Iterierte x_{t+1} wird jeweils aus den *beiden* vorausgehenden Iterierten x_t, x_{t-1} berechnet. Zum Starten der Sekantenmethode sind zwei Anfangsschätzungen x_0, x_1 für die Nullstelle erforderlich. Analog zum Konvergenzsatz 5.1 für das Newton-Verfahren haben wir auch eine Konvergenzaussage für die Sekantenmethode. Dabei spielen die durch die Vorschrift

$$\gamma_0 = \gamma_1 = 1\,, \quad \gamma_{t+1} = \gamma_t + \gamma_{t-1}\,, \quad t \in \mathbb{N},$$

definierten sog. Fibonacci[2]-Zahlen” γ_t eine wichtige Rolle.

[2]Leonardo Pisano (aus Pisa), genannt Fibonacci (um 1170 – um 1250): „erster“ bedeutender Mathematiker des Abendlandes; gehörte zum Gelehrtenkreis um Kaiser Friedrich II; brachte von ausgedehnten Reisen eine systematische Einführung in das indisch-arabische Zahlensystem nach Europa; in seinem Rechenbuch „Liber abacci“ untersuchte er u. a. die nach ihm benannte Folge als einfaches Modell für das Wachstum von Populationen.

Satz 5.3 (Sekanten-Methode): *Die Funktion $f \in C^2[a,b]$ habe im Innern des Intervalls $[a,b]$ eine Nullstelle z, und es sei*

$$m := \min_{a\le x\le b} |f'(x)| > 0\,, \quad M := \max_{a\le x\le b} |f''(x)| < \infty. \tag{5.3.19}$$

Sei ferner $\rho > 0$ so gewählt, dass

$$q \equiv \frac{M}{2m}\rho < 1, \qquad K_\rho(z) = \{x \in \mathbb{R}\,|\,|x-z| \le \rho\} \subset [a,b].$$

Dann sind für jedes Paar von Startwerten $x_0, x_1 \in K_\rho(z)$, $x_0 \ne x_1$, die Iterierten $x_t \in K_\rho(z)$ der Sekantenmethode wohl definiert und konvergieren gegen die Nullstelle z. Dabei gelten die a priori Fehlerabschätzung

$$|x_t - z| \le \frac{2m}{M}\, q^{\gamma_t}\,, \quad t \in \mathbb{N}, \tag{5.3.20}$$

und die a posteriori Fehlerabschätzung

$$|x_t - z| \le \frac{1}{m}\,|f(x_t)| \le \frac{M}{2m}\,|x_t - x_{t-1}|\,|x_t - x_{t-2}|\,, \quad t \in \mathbb{N}. \tag{5.3.21}$$

Beweis: (i) Die Argumentation ist ähnlich wie im Beweis von Satz 5.1 für das Newton-Verfahren. Für je zwei Punkte $x, y \in [a,b]$, $x \ne y$, gilt wieder

$$|x-y| \le \frac{1}{m}\,|f(x) - f(y)|,$$

woraus u. a. die Eindeutigkeit der Nullstelle z folgt.

(ii) Weiter ist

$$\frac{f(x)-f(y)}{x-y} = -\int_0^1 \frac{d}{dr}\, f(x + r(y-x))\, \frac{dr}{x-y} = \int_0^1 f'(x + r(y-x))\, dr.$$

Mit einem dritten Punkt $\zeta \in [a,b]$, $\zeta \ne x$, ergibt sich hiermit

$$\begin{aligned}
\frac{f(x)-f(y)}{x-y} - \frac{f(x)-f(\zeta)}{x-\zeta} &= \int_0^1 \Big\{\, f'(x - r(y-x)) - f'(x + r(\zeta - x))\,\Big\}\, dr \\
&= -\int_0^1 \Big\{ \int_0^r \frac{d}{ds}\, f'(x + r(y-x) + s(\zeta - y))\, ds \Big\}\, dr \\
&= \int_0^1 \Big\{ \int_0^r f''(x + r(y-x) + s(\zeta - y))\, ds \Big\}\, dr\,(y-\zeta),
\end{aligned}$$

bzw.

$$\Big|\frac{f(x)-f(y)}{x-y} - \frac{f(x)-f(\zeta)}{x-\zeta}\Big| \le \frac{M}{2}\,|y-\zeta|.$$

Für Punkte $x, y \in K_\rho(z)$, $x \neq y$, $x \neq z$, $y \neq z$ definieren wir

$$g(x,y) := x - f(x)\,\frac{x-y}{f(x)-f(y)}.$$

Dann gilt

$$\begin{aligned} g(x,y) - z &= x - z - f(x)\frac{x-y}{f(x)-f(y)} \\ &= \frac{x-y}{f(x)-f(y)} \Big\{ (x-z)\frac{f(x)-f(y)}{x-y} - f(x) + \underbrace{f(z)}_{=0} \Big\} \end{aligned}$$

und folglich

$$\begin{aligned} |g(x,y) - z| &\le |x-z|\,\Big|\frac{f(x)-f(y)}{x-y} - \frac{f(x)-f(z)}{x-z}\Big| \\ &\le \frac{M}{2m}\,|x-z|\,|y-z| \;\le\; \frac{M}{2m}\,\rho^2 < \rho. \end{aligned}$$

Die Iterierten x_t der Sekantenmethode bleiben also in der Menge $K_\rho(z)$, und es gilt

$$|x_{t+1} - z| \le \frac{M}{2m}|x_t - z|\,|x_{t-1} - z|.$$

(iii) Setzt man $\rho_t := \frac{M}{2m}|x_t - z|$, so folgt

$$\rho_{t+1} \;\le\; \rho_t \rho_{t-1}\,, \quad t \in \mathbb{N},$$

d. h. mit $\rho_0 \le q$, $\rho_1 \le q$ gilt $\rho_t \le q^{\gamma_t}$, $t \in \mathbb{N}$. Wegen $\gamma_t \to \infty$ $(t \to \infty)$ und $q < 1$ konvergiert also

$$|x_t - z| = \frac{2m}{M}\,\rho_t \;\le\; \frac{2m}{M}\,q^{\gamma_t} \to 0 \quad (t \to \infty).$$

(iv) Zum Nachweis der a posteriori Fehlerabschätzung setzen wir oben $x = x_{t-1}$, $y = x_t$ und $\zeta = x_{t-2}$ $(x_{t-2} \neq x_{t-1}$, da sonst bereits $f(x_{t-1}) = 0)$ und finden

$$\begin{aligned} |x_t - z| &\le \frac{1}{m}\,|f(x_t) - f(z)| \\ &\le \frac{1}{m}\,\Big|f(x_{t-1}) + (x_t - x_{t-1})\frac{f(x_t)-f(x_{t-1})}{x_t - x_{t-1}}\Big| \\ &\le \frac{1}{m}\,|x_t - x_{t-1}|\,\Big|\frac{f(x_t)-f(x_{t-1})}{x_t - x_{t-1}} - \frac{f(x_{t-1})-f(x_{t-2})}{x_{t-1}-x_{t-2}}\Big| \\ &\le \frac{M}{2m}\,|x_t - x_{t-1}|\,|x_t - x_{t-2}|. \end{aligned}$$

Q.E.D.

Zur Beurteilung der Konvergenzgeschwindigkeit der Sekantenmethode benötigen wir Informationen über das Anwachsen der Fibonacci-Zahlen γ_t für $t \to \infty$.

Hilfssatz 5.1: *Die Fibonacci-Zahlen verhalten sich asymptotisch wie*

$$\gamma_t \sim \frac{\lambda_1}{\sqrt{5}}\,\lambda_1^t \sim 0.723 \cdot (1.618)^t, \tag{5.3.22}$$

wobei $\lambda_1 := \frac{1}{2}(1 \pm \sqrt{5})$ *gerade der sog. „goldene Schnitt“ ist.*

Beweis: Die Fibonacci-Zahlen genügen nach Konstruktion der (linearen) homogenen Differenzengleichung

$$\gamma_{t+2} - \gamma_{t+1} - \gamma_t = 0\,, \quad t \geq 0. \tag{5.3.23}$$

Deren Lösungsmenge ist, wie man leicht sieht, ein zweidimensionaler Vektorraum. Zur Konstruktion einer Lösunf machen wir den Ansatz $\gamma_t = \lambda^t$ und erhalten die Gleichung

$$\lambda^t(\lambda^2 - \lambda - 1) = 0$$

zur Bestimmung von λ. Die Wurzeln $\lambda_{1,2} = \frac{1}{2}(1 \pm \sqrt{5})$ der quadratischen Gleichung $\lambda^2 - \lambda - 1 = 0$ ergeben durch

$$\gamma_t = c_1\lambda_1^t + c_2\lambda_2^t\,, \quad c_1, c_2 \quad \text{beliebig}, \tag{5.3.24}$$

die allgemeine Lösung der Differenzengleichung. Durch Berücksichtigung der Anfangsbedingungen $\gamma_0 = \gamma_1 = 1$ werden die Konstanten c_1, c_2 festgelegt:

$$\left.\begin{array}{rcl} c_1 + c_2 & = & 1 \\ c_1\lambda_1 + c_2\lambda_2 & = & 1 \end{array}\right\} \Rightarrow c_1 = \frac{1-\lambda_2}{\lambda_1 - \lambda_2} = \frac{\lambda_1}{\sqrt{5}}\,, \quad c_2 = \frac{\lambda_1 - 1}{\lambda_1 - \lambda_2} = -\frac{\lambda_2}{\sqrt{5}}.$$

Die Fibonacci-Zahlen haben also die Gestalt

$$\gamma_t = \frac{1}{\sqrt{5}}\left\{\lambda_1^{t+1} - \lambda_2^{t+1}\right\}, \quad \lambda_{1,2} = \frac{1}{2}(1 \pm \sqrt{5}). \tag{5.3.25}$$

Asymptotisch für $t \to \infty$ verhält sich γ_t wie

$$\gamma_t \sim \frac{\lambda_1}{\sqrt{5}}\,\lambda_1^t \sim 0.723 \cdot (1.618)^t,$$

was zu zeigen war. Q.E.D.

Die Sekantenmethode konvergiert also asymptotisch mindestens so schnell wie ein Ein-Schrittverfahren der Ordnung $p = 1.6$. In jedem Schritt ist dabei nur eine neue Funktionsauswertung, nämlich die von $f(x_t)$, erforderlich. *Ein* Schritt des Newton-Verfahrens (Auswertung von $f(x_t)$ und $f'(x_t)$) ist also mindestens so aufwendig wie *zwei* Schritte der Sekantenmethode. Fasst man jedoch zwei Schritte der Sekantenmethode zu einem

Makroschritt zusammen, so erhält man wegen

$$|x_{2t} - z| \leq \frac{2m}{M} q^{\gamma_{2t}}, \quad \gamma_{2t} \sim 0.723\,(2.618)^t \quad (\lambda_1^2 = \lambda_1 + 1), \tag{5.3.26}$$

ein Verfahren der Ordnung $p \geq 2.6$. Bei gleichem Arbeitsaufwand konvergiert also die Sekantenmethode asymptotisch (für große t) schneller als das Newton-Verfahren. Dieser theoretische Vorteil wird aber in der Praxis oft durch eine große Rundungsfehleranfälligkeit der Sekantenmethode relativiert. Konvergiert nämlich hier $f(x_t) \to 0$ monoton (mit nicht alternierenden Vorzeichen), so tritt im Sekantenschritt

$$x_{t+1} = x_t - f(x_t)\,\frac{x_t - x_{t-1}}{f(x_t) - f(x_{t-1})}. \tag{5.3.27}$$

Auslöschung auf. Zur Stabilisierung der Methode kombiniert man sie mit der Intervallschachtelungsidee zur sog. „Regula falsi“.

Definition 5.3: *Werden im Sekanten-Verfahren die Intervallendpunkte $a_t < b_t$ so gewählt, dass $f(a_t)f(b_t) < 0$ ist, d. h. dass f eine Nullstelle $z \in (a_t, b_t)$ hat, so spricht man von der „Regula falsi“.*

Beim Sekantenschritt unter berücksichtigung der Regula falsi,

$$x_t := a_t - f(a_t)\,\frac{a_t - b_t}{f(a_t) - f(b_t)}, \tag{5.3.28}$$

tritt dann keine Auslöschung im Term $f(a_t) - f(b_t)$ auf, solange $b_t - a_t \gg \text{eps}$. Offenbar ist $a_t \leq x_t \leq b_t$. Das neue Intervall $[a_{t+1}, b_{t+1}]$ wird bestimmt durch die Vorschrift: ($f(x_t) = 0 \Rightarrow$ STOP)

$$\begin{aligned} f(x_t)f(a_t) > 0 \quad &\Longrightarrow \quad a_{t+1} = x_t\,, \quad b_{t+1} = b_t\,, \\ f(x_t)f(a_t) < 0 \quad &\Longrightarrow \quad a_{t+1} = a_t\,, \quad b_{t+1} = x_t. \end{aligned} \tag{5.3.29}$$

Die Regula falsi ist offensichtlich numerisch stabiler als die ihr zugrunde liegende Sekantenmethode, doch konvergiert sie i. Allg. nur linear. In Extremfällen ist sie sogar langsamer (größere Konvergenzrate) als das einfache Intervallschachtelungsverfahren.

5.4 Methode der sukzessiven Approximation im $\mathbb{R}^n$

Im Folgenden betrachten wir iterative Verfahren zur Lösung nichtlinearer Gleichungssysteme im $\mathbb{R}^n$

$$f_i(x_1, \ldots, x_n) = 0\,, \quad i = 1, \ldots, n, \tag{5.4.30}$$

bzw. $f(x) = 0$ mit $f = (f_1, \ldots, f_n)^T$ und $x = (x_1, \ldots, x_n)^T$. Zur Berechnung einer Lösung z verwendet die sog. „Methode der sukzessiven Approximation“ die Iteration (in

Anlehnung an das vereinfachte Newton-Verfahren in einer Dimension und unter Hochstellung des Iterationsindexes bei vektor- oder matrix-wertigen Größen):

$$x^{t+1} = x^t + C^{-1} f(x^t)\,, \quad t = 0, 1, 2, \ldots\,, \tag{5.4.31}$$

mit einer geeigneten regulären Matrix $C \in \mathbb{R}^{n\times n}$. Konvergiert dann $x^t \to z\,(t \to \infty)$, so ist für stetiges f im Limes $z = z + C^{-1} f(z)$ bzw. $f(z) = 0$. Die Lösung z des Gleichungssystems ist also ein sog. „Fixpunkt"' der Abbildung $g(x) := x + C^{-1} f(x)$.

Wir wollen nun die Konvergenz von Fixpunktiterationen der Form

$$x^{t+1} = g(x^t)\,, \quad t = 0, 1, 2, \ldots\,, \tag{5.4.32}$$

untersuchen. Dazu sei im Folgenden $\|\cdot\|$ eine beliebige Vektornorm auf $\mathbb{R}^n$ und mit derselben Bezeichnung $\|\cdot\|$ die zugehörige natürliche Matrizennorm.

Definition 5.4: *Sei $G \subset \mathbb{R}^n$ eine (nichtleere) abgeschlossene Menge. Eine Abbildung $g : G \to \mathbb{R}^n$ heißt „Lipschitz[3]-stetig" (kurz „L-stetig"), wenn mit einem $q > 0$ gilt:*

$$\|g(x) - g(y)\| \le q\,\|x - y\|, \quad x, y \in G. \tag{5.4.33}$$

Ist die sog. „Lipschitz-Konstante" $q < 1$, so nennt man g eine „Kontraktion" auf G.

Beispiel 5.4: a) Die Funktion $f(x) = |x|$ ist L-stetig auf ganz $\mathbb{R}$, wegen

$$|\,|x| - |y|\,| \le |x - y|.$$

b) Die Funktion $f(x) = \sqrt{|x|}$ ist nicht L-stetig bei $x = 0$ aber (lokal) L-stetig sonst:

$$|\,|x|^{1/2} - |0|^{1/2}\,| = |x|^{1/2} \ge |x|^{-1/2}|x - 0|.$$

Der folgende fundamentale „Banachsche[4] Fixpunktsatz" sichert die Existenz von Fixpunkten von Kontraktionen.

Satz 5.4 (Sukzessive Approximation): *Sei $G \subset \mathbb{R}^n$ eine nichtleere, abgeschlossene Punktmenge und $g : G \to G$ eine Kontraktion. Dann existiert genau ein Fixpunkt $z \in G$ von g, und für jeden Startpunkt $x^0 \in G$ konvergiert die Folge der durch (5.4.32) erzeugten sukzessiven Approximationen $x^t \to z\,(t \to \infty)$. Es gelten die a posteriori und a priori Fehlerabschätzungen*

$$\|x^t - z\| \le \frac{q}{1-q}\,\|x^t - x^{t-1}\| \le \frac{q^t}{1-q}\,\|x^1 - x^0\|. \tag{5.4.34}$$

[3]Rudolf O. S. Lipschitz (1832–1903): Deutscher Mathematiker aus Königsberg; seit 1864 Professor in Bonn; arbeitete auf verschiedenen Gebieten der Mathematik.

[4]Stefan Banach (1892-1945): Polnischer Mathematiker; Professor in Lvov; begründete die Funktionalanalysis.

Beweis: Da g die Menge G in sich abbildet, sind für $x^0 \in G$ die Iterierten $x^t = g(x^{t-1}) = \ldots = g^t(x^0)$ definiert, und es gilt

$$\begin{aligned}\|x^{t+1} - x^t\| &= \|g(x^t) - g(x^{t-1})\| \\ &\leq q\,\|x^t - x^{t-1}\| \leq \ldots \leq q^t\,\|x^1 - x^0\|.\end{aligned}$$

Wir wollen zeigen, dass $(x^t)_{t\in\mathbb{N}}$ eine Cauchy-Folge ist. Seien dazu $\varepsilon > 0$ und $m \geq 1$ beliebig vorgegeben:

$$\begin{aligned}\|x^{t+m} - x^t\| &\leq \|x^{t+m} - x^{t+m-1}\| + \ldots + \|x^{t+1} - x^t\| \\ &\leq \underbrace{\{q^{t+m-1} + \ldots + q^t\}}\,\|x^1 - x^0\| \\ &= q^t \sum_{i=0}^{m-1} q^i = q^t\,\frac{1-q^m}{1-q} \leq \varepsilon \quad \text{für} t \geq t(\varepsilon).\end{aligned}$$

Also existiert $z = \lim_{t\to\infty} x^t \in G$ (wegen der Abgeschlossenheit von G) mit $z = g(z)$. Die Eindeutigkeit des Fixpunktes z folgt sofort aus der Kontraktionseigenschaft von g. Zum Nachweis von (5.4.34) schreiben wir

$$\begin{aligned}\|x^{t+m} - x^t\| &\leq \|x^{t+m} - x^{t+m-1}\| + \ldots + \|x^{t+1} - x^t\| \\ &\leq \underbrace{\{q^m + \ldots + q\}}\,\|x^t - x^{t-1}\|\,, \quad m \geq 1. \\ &\qquad \leq q/(1-q)\end{aligned}$$

Durch Grenzübergang $m \to \infty$ folgt daraus

$$\|z - x^t\| \leq \frac{q}{1-q}\,\|x^t - x^{t-1}\| \leq \frac{q^t}{1-q}\,\|x^1 - x^0\|.$$

Q.E.D.

Zur Anwendung des Banachschen Fixpunktsatzes auf eine Abbildung $g : G \to \mathbb{R}^n$ muss gezeigt werden, dass es eine abgeschlossene (nichtleere) Teilmenge von G gibt, die von g in sich abgebildet wird, und auf der g eine Kontraktion ist. Sei g eine Kontraktion auf der Kugel

$$K_\rho(c) \equiv \{\, x \in \mathbb{R}^n \mid \|x - c\| \leq \rho \,\}\,, \quad \rho > 0,$$

um einen Punkt $c \in \mathbb{R}^n$ mit Lipschitz-Konstante $q < 1$. Für $x \in K_\rho(c)$ gilt dann

$$\|g(x) - c\| \leq \underbrace{\|g(x) - g(c)\|}_{\leq q\rho} + \|g(c) - c\|.$$

Unter der Bedingung

$$\|g(c) - c\| \leq (1-q)\rho \tag{5.4.35}$$

bildet dann g die Menge $K_\rho(c)$ in sich ab. Ist g differenzierbar, so wird die Matrix

$$g'(x) \equiv \Big(\frac{\partial g_i}{\partial x_j}\Big)_{i,j=1,\dots,n} \in \mathbb{R}^{n\times n}$$

der partiellen Ableitungen die „Jacobi[5]-Matrix“ genannt.

Hilfssatz 5.2 (L-Stetigkeit): *Die Abbildung $g : G \to \mathbb{R}^n$ sei stetig differenzierbar, und die Menge G sei konvex. Dann gilt*

$$\|g(x) - g(y)\| \le \sup_{\zeta\in G} \|g'(\zeta)\|\, \|x - y\|, \quad x, y \in G, \tag{5.4.36}$$

d. h.: Im Falle $\sup_{\zeta\in G} \|g'(\zeta)\| < 1$ ist g eine Kontraktion auf G.

Beweis: Seien $x, y \in G$. Wir setzen für $i = 1, \dots, n$:

$$\varphi_i(s) := g_i(x + s(y - x)), \quad 0 \le s \le 1,$$

und haben damit

$$g_i(y) - g_i(x) = \varphi_i(1) - \varphi_i(0) = \int_0^1 \varphi_i'(s)\, ds.$$

Wegen

$$\varphi_i'(s) = \sum_{j=1}^{n} \frac{\partial g_i}{\partial x_j}(x + s(y - x))(y - x)_j$$

und den Stetigkeitseigenschaften der Vektornorm folgt

$$\begin{aligned}\|g(y) - g(x)\| &= \Big\| \int_0^1 g'(x + s(y - x)) \cdot (y - x)\, ds\Big\| \\ &\le \int_0^1 \|g'(x + s(y - x))\|\, ds\, \|y - x\| \le \sup_{\zeta\in G} \|g'(\zeta)\|\, \|y - x\|.\end{aligned}$$

Dies impliziert die Bahauptung. Q.E.D.

Korollar 5.1: *Mit Hilfe der Abschätzung aus Hilfssatz 5.2 und (5.4.35) ergibt sich, dass es zu jedem Fixpunkt $z \in G$ von g, in dem $\|g'(z)\| < 1$ gilt, eine Umgebung*

$$K_\rho(z) = \{x \in \mathbb{R}^n |\, \|x - z\| \le \rho\} \subset G$$

gibt, so dass g eine Kontraktion von $K_\rho(z)$ in sich ist.

[5]Carl Gustav Jakob Jacobi (1804–1851): Deutscher Mathematiker; schon als Kind hochbegabt; wirkte in Königsberg und Berlin; Beiträge zu vielen Bereichen der Mathematik: Zahlentheorie, elliptische Funktionen, partielle Differentialgleichungen, theoretische Mechanik.

Wir betrachten nun wieder die Lösung der Gleichung $f(x) = 0$ mit Hilfe der sukzessiven Approximation

$$x^{t+1} = x^t + C^{-1} f(x^t)\,, \quad t = 0, 1, 2, \ldots\,. \tag{5.4.37}$$

Nach den obigen Überlegungen ist die Konvergenz dieser Iteration z. B. gesichert, wenn f auf einer geeigneten Kugel $K_\rho(c) \subset \mathbb{R}^n$ stetig differenzierbar ist, und wenn dort gilt

$$\sup_{\zeta \in K_\rho(c)} \|I + C^{-1} f'(\zeta)\| =: q < 1\,, \quad \|C^{-1} f(c)\| \le (1-q)\rho\,. \tag{5.4.38}$$

Beispiel 5.5: Seien $A \in \mathbb{R}^{n\times n}$ und $b \in \mathbb{R}^n$ gegeben. Das lineare Gleichungssystem $Ax = b$ ist äquivalent zur Nullstellenaufgabe $f(x) := b - Ax = 0$. Zu deren iterativen Lösung betrachten wir mit einer regulären Matrix $C \in \mathbb{R}^{n\times n}$ die Fixpunktaufgabe

$$x = g(x) := x + C^{-1} f(x) = x + C^{-1}(b - Ax) = \underbrace{(I - C^{-1}A)}_{=B}\, x + \underbrace{C^{-1}b}_{=c}\,.$$

Die Matrix $B := I - C^{-1}A$ wird die „Iterationsmatrix“ der zugehörigen Fixpunktiteration („sukzessive Approximation“) genannt:

$$x^{t+1} = Bx^T + c, \quad t = 1, 2, \ldots\,.$$

Die Abbildung g ist wegen

$$\|g(x) - g(y)\| = \|B(x-y)\| \le \|B\|\,\|x-y\|$$

für $\|B\| < 1$ eine Kontraktion auf ganz $\mathbb{R}^n$. Dabei ist $\|\cdot\|$ eine geeignete (natürliche) Matrizennorm. Nach dem Banachschen Fixpunktsatz konvergiert daher die sukzessive Approximation gegen den (eindeutig bestimmten) Fixpunkt der Abbildung g bzw. die Lösung des Gleichungssystems $Ax = b$.

Beispiel 5.6: Die Funktion $f(x) = \cosh(x) - 2x = \frac{1}{2}(e^x + e^{-x}) - 2x$ hat genau zwei Nullstellen $z_1 \sim 0.59$, $z_2 \sim 2.1$. Zu ihrer Approximation machen wir den Ansatz

$$g(x) = x + \tfrac{1}{2}\{\cosh(x) - 2x\} = \tfrac{1}{2}\cosh(x), \quad g'(x) = \tfrac{1}{2}\sinh(x),$$

Offensichtlich bildet g das Intervall $[0, z_2]$ in sich ab. Da die Beziehung

$$\max_{0\le x\le b} |g'(x)| = \tfrac{1}{2}\sinh(b) < 1 \quad (\operatorname{arcsinh}(2) = 1.44\ldots)$$

notwendig $b < 2$ voraussetzt, muss $|g'(z_2)| > 1$ sein. Für alle Startwerte $x^0 \in (z_2, \infty)$ divergieren die Iterierten $x^t \to \infty$ für $t \to \infty$. Tatsächlich konvergiert aber $x^t \to z_1\,(t \to \infty)$ sogar für alle $x^0 \in [0, z_2]$ (geometrische Überlegung). Die Bedingung, dass g überall eine Kontraktion sein muss, ist also nicht notwendig für die Konvergenz der sukzessiven Approximation. Der Fixpunkt z_1 mit $|g'(z_1)| < 1$ in Beispiel 5.4 ist also „anziehend“, der Fixpunkt z_2 dagegen „abstoßend“, da hier wegen $|g'(z_2)| > 1$ in jeder Umgebung

von z_2 Startpunkte x^0 existieren, für die die sukzessive Approximation nicht gegen z_2 konvergiert.

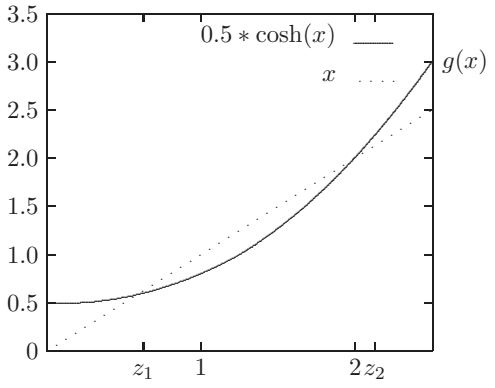


Abbildung 5.4: *Nullstellenproblem*

Nach Wahl einer Fehlertoleranz $\varepsilon >$ eps (z. B.: $\varepsilon = 10^{-4}$) wird ausgehend von $x^0 = 0$ iteriert gemäß $x^{t+1} = \frac{1}{2}\cosh(x^t)(t = 0, 1, 2, \ldots)$ bis das folgende Abbruchkriterium erfüllt ist:

$$\left|\frac{x^{t+1} - x^t}{x^{t+1}}\right| \leq \varepsilon.$$

Wir erhalten:

$$x^1 = 0.\underline{5},\ x^2 = 0.\underline{5}63,\ \ldots,\ x^7 = 0.\underline{589}31,\ \ x^8 = 0.\underline{58936},\ \ldots,\ x^{19} = 0.\underline{5893877633}.$$

$$\left|\frac{x^8 - x^7}{x^8}\right| \leq 0.8532 \cdot 10^{-4}.$$

Auf dem Intervall $[0, 1]$ gilt

$$q = \max_{0 \leq x \leq 1} |g'(x)| = \tfrac{1}{2}\sinh(1) \sim 0.6.$$

Die a priori Fehlerabschätzungen von Satz 5.2 ergibt dann

$$|x^8 - z_1| \leq \frac{0.6^8}{1 - 0.6}|x^1 - x^0| \sim 2 \cdot 10^{-2}.$$

Beispiel 5.7: Zur Bestimmung der Quadratwurzel $A^{1/2} \in \mathbb{R}^{n \times n}$ einer positiv definiten Matrix $A \in \mathbb{R}^{n \times n}$ betrachtet man die Abbildung

$$g(X) = \tfrac{1}{2}(X^2 + B)$$

mit der Matrix $B = I - A$. Dies wird motiviert durch die Äquivalenz

$$Z = g(Z) = \tfrac{1}{2}(Z^2 + B) \quad \Leftrightarrow \quad (I - Z)^2 = I - B = A,$$

bzw. $I - Z = A^{1/2}$. Im Falle $\|B\| = q < 1$ gilt für $X, Y \in K_q(0) = \{C \in \mathbb{R}^{n\times n} \,|\, \|C\| \le q\}$:

$$\|g(X)\| \le \frac{1}{2}(\|X\|^2 + \|B\|) \le \frac{1}{2}(q^2 + q) \le q$$

und

$$\begin{aligned}\|g(X) - g(Y)\| &= \frac{1}{2}\|X^2 - Y^2\| = \frac{1}{2}\|X(X-Y) + (X-Y)Y\| \\ &\le \frac{1}{2}(\|X\| + \|Y\|)\,\|X - Y\| \le q\|X - Y\|,\end{aligned}$$

d. h.: g ist eine Kontraktion der abgeschlossenen Teilmenge $K_q(0) \subset \mathbb{R}^{n\times n}$ in sich. Nach dem Banachschen Fixpunktsatz existiert also genau ein Fixpunkt $Z \in K_q(0)$ von g, und die Folge der sukzessiven Iterierten $X^t = g(X^{t-1})$, $t \in \mathbb{N}$, konvergiert für jeden Startwert $X^0 \in K_q(0)$: $X^t \to Z$ $(t \to \infty)$. Wegen der obigen Äquivalenz ist dann $I - Z = A^{1/2}$. Alle Iterierten X^t und damit auch der Fixpunkt Z sind symmetrisch. Wegen $\|Z\| \le q$ ist daher $I - Z$ auch positiv definit, so daß mit $A^{1/2} := I - Z$ die eindeutug bestimmte, positive" Wurzel von A bestimmt ist.

5.5 Das Newton-Verfahren im $\mathbb{R}^n$

Wir betrachten nun das Newton-Verfahren zur Lösung nichtlinearer Gleichungssysteme mit stetig differenzierbaren Abbildungen $f : D \subset \mathbb{R}^n \to \mathbb{R}^n$. Formal lautet die Newton-Iteration

$$x^{t+1} = x^t - f'(x^t)^{-1} f(x^t), \quad t = 0, 1, 2 \ldots, \tag{5.5.39}$$

mit der Jacobi-Matrix $f'(\cdot)$ von f. In jedem Iterationsschritt ergibt sich ein lineares $(n \times n)$-Gleichungssystem mit $f'(x^t)$ als Koeffizientenmatrix:

$$f'(x^t)x^{t+1} = f'(x^t)x^t - f(x^t), \quad t = 0, 1, 2 \ldots . \tag{5.5.40}$$

Dies macht das Newton-Verfahren wesentlich aufwendiger als die einfache Fixpunktiteration; dafür konvergiert es aber auch sehr viel schneller. Das Newton-Verfahren wird meist in Form einer Defektkorrekturiteration durchgeführt (mit dem "Defekt" $d^t := f(x^t)$):

$$f'(x^t)\delta x^t = f(x^t), \quad x^{t+1} = x^t - \delta x^t, \quad t = 0, 1, 2 \ldots . \tag{5.5.41}$$

Dies spart gegenüber (5.5.40) pro Iterationsschritt eine Matrix-Vektor-Multiplikation.

Im Folgenden geben wir ein Konvergenzresultat für das Newton-Verfahren, welches nebenbei auch die Existenz einer Nullstelle sichert. Mit $\|\cdot\|$ seien die euklidische Vektornorm

und ebenso die zugehörige natürliche Matrizennorm bezeichnet. Sei $f : G \subset \mathbb{R}^n \to \mathbb{R}^n$ eine differenzierbare Abbildung, für die eine Nullstelle z gesucht ist. Die Jacobi-Matrix $f'(\cdot)$ sei auf der Niveaumenge

$$D_* := \{x \in G | \, \|f(x)\| \leq \|f(x^*)\|\}$$

zu einem festen Punkt $x^* \in G$ regulär mit gleichmäßig beschränkter Inverser:

$$\|f'(x)^{-1}\| \leq \beta, \quad x \in D_*.$$

Ferner sei $f'(\cdot)$ auf D_* gleichmäßig L-stetig:

$$\|f'(x) - f'(y)\| \leq \gamma \|x - y\|, \quad x, y \in D_*.$$

Mit diesen Bezeichnungen haben wir den folgenden Satz von Newton-Kantorovich[6]

Satz 5.5 (Newton-Kantorovich): *Unter den vorausgehenden Voraussetzungen sei für den Startpunkt $x^0 \in D_*$ mit $\alpha := \|f'(x^0)^{-1} f(x^0)\|$ die folgende Bedingung erfüllt:*

$$q := \tfrac{1}{2}\alpha\beta\gamma < 1.$$

Dann erzeugt die Newton-Iteration

$$f'(x^t)x^{t+1} = f'(x^t)x^t - f(x^t), \quad t \geq 1,$$

eine Folge $(x^t)_{t\in\mathbb{N}} \subset D$, welche quadratisch gegen eine Nullstelle $z \in D$ von f konvergiert, mit der a priori Fehlerabschätzung

$$\|x^t - z\| \leq \frac{\alpha}{1-q} q^{(2^t-1)}, \quad t \geq 1. \tag{5.5.42}$$

Beweis: Zum Startpunkt $x^0 \in D_*$ gehört die abgeschlossene, nicht leere Niveaumenge

$$D_0 := \{x \in G | \, \|f(x)\| \leq \|f(x^0)\|\} \subset D_*.$$

Wir betrachten die stetige Abbildung $g : D_0 \to \mathbb{R}^d$: $g(x) := x - f'(x)^{-1} f(x)$.

(i) Wir wollen zunächst einige Hilfsresultate ableiten. Für $x \in D_0$ sei

$$x_r := x - r f'(x)^{-1} f(x), \;\; 0 \leq r \leq 1,$$

und $R := \max\{r | \, x_s \in D_0, 0 \leq s \leq r\} = \max\{r | \, \|f(x_s)\| \leq \|f(x^0)\|, \; 0 \leq s \leq r\}$. Für die Vektorfunktion $h(r) := f(x_r)$ gilt

$$h'(r) = -f'(x_r) f'(x)^{-1} f(x), \qquad h'(0) = -h(0).$$

[6]Leonid Vitalyevich Kantorovich (1912–1986): Russischer Mathematiker; Professor an der Universität Leningrad (1934-1960), an der Akademie der Wissenschaften (1961-1971) und an der Universität Moskau (1971-1976); fundamentale Beiträge zur Anwendung der linearen Optimierung in der Ökonomie, zur Funktionalanalysis und Numerik.

Für $0 \le r \le R$ ergibt dies

$$\begin{aligned}\|f(x_r)\| - (1-r)\|f(x)\| &\le \|f(x_r) - (1-r)f(x)\| = \|h(r) - (1-r)h(0)\| \\ &= \Big\| \int_0^r h'(s)\,ds + rh(0) \Big\| = \Big\| \int_0^r \{h'(s) - h'(0)\}\,ds \Big\| \\ &\le \int_0^r \|h'(s) - h'(0)\|\,ds,\end{aligned}$$

und ferner wegen $x_s - x = -sf'(x)^{-1}f(x)$:

$$\begin{aligned}\|h'(s) - h'(0)\| &= \|\{f'(x_s) - f'(x)\}f'(x)^{-1}f(x)\| \\ &\le \gamma\|x_s - x\|\|f'(x)^{-1}f(x)\| \le \gamma s\|f'(x)^{-1}f(x)\|^2.\end{aligned}$$

Dies ergibt

$$\|f(x_r)\| - (1-r)\|f(x)\| \;\le\; \tfrac{1}{2}r^2\gamma\|f'(x)^{-1}f(x)\|^2. \tag{5.5.43}$$

Mit der Größe $\alpha_x := \|f'(x)^{-1}f(x)\|$ und $\|f'(x)^{-1}\| \le \beta$ folgt

$$\|f(x_r)\| \le (1 - r + \tfrac{1}{2}r^2\alpha_x\beta\gamma)\|f(x)\|.$$

Im Falle $\alpha_x \le \alpha$ gilt dann wegen der Voraussetzung $\frac{1}{2}\alpha\beta\gamma < 1$:

$$\|f(x_r)\| \le (1 - r + r^2)\|f(x)\|.$$

Folglich ist in diesem Fall $R = 1$, d. h.: $g(x) \in D_0$. Für solche $x \in D_0$ gilt weiter

$$\|g(x) - g^2(x)\| = \|g(x) - g(x) + f'(g(x))^{-1}f(g(x))\| \le \beta\|f(g(x))\|.$$

Mit Hilfe der Abschätzung (5.5.43) für $r = 1$ folgt bei Beachtung von $g(x) = x_1$:

$$\|g(x) - g^2(x)\| \le \tfrac{1}{2}\beta\gamma\|f'(x)^{-1}f(x)\|^2 = \tfrac{1}{2}\beta\gamma\|x - g(x)\|^2. \tag{5.5.44}$$

(ii) Nach diesen Vorbereitungen kommen wir nun zum Beweis des Satzes. Zunächst wollen wir zeigen, dass die Newton-Iterierten $(x^t)_{t\in\mathbb{N}}$ in D_0 existieren und die Ungleichung

$$\|x^t - g(x^t)\| = \|f'(x^t)^{-1}f(x^t)\| \le \alpha$$

erfüllen. Dies erfolgt durch vollständige Induktion. Für $t = 0$ ist die Aussage trivialerweise richtig; insbesondere ist wegen $\alpha_{x^0} = \alpha$ nach dem oben gezeigten $g(x^0) \in D_0$. Sei nun $x^t \in D_0$ eine Iterierte mit $g(x^t) \in D_0$ und $\|x^t - g(x^t)\| \le \alpha$. Dann folgt

$$\|x^{t+1} - g(x^{t+1})\| = \|g(x^t) - g^2(x^t)\| \le \tfrac{1}{2}\beta\gamma\|x^t - g(x^t)\|^2 \le \tfrac{1}{2}\alpha^2\beta\gamma \le \alpha$$

und somit nach dem oben Gezeigten $g(x^{t+1}) \in D_0$. Also existiert $(x^t)_{t\in\mathbb{N}} \subset D_0$. Als nächstes zeigen wir, dass diese Folge Cauchy-Folge ist. Mit Hilfe von (5.5.44) ergibt sich

$$\|x^{t+1} - x^t\| = \|g^2(x^{t-1}) - g(x^{t-1})\| \le \tfrac{1}{2}\beta\gamma\|g(x^{t-1}) - x^{t-1}\|^2 = \tfrac{1}{2}\beta\gamma\|x^t - x^{t-1}\|^2,$$

und bei Iteration dieser Abschätzung:

$$\begin{aligned}\|x^{t+1}-x^t\| &\le \tfrac{1}{2}\beta\gamma\big(\tfrac{1}{2}\beta\gamma\|x^{t-1}-x^{t-2}\|^2\big)^2 \le (\tfrac{1}{2}\beta\gamma)^{(2^2-1)}\|x^{t-1}-x^{t-2}\|^{(2^2)}\\ &\le (\tfrac{1}{2}\beta\gamma)^{(2^2-1)}\big(\tfrac{1}{2}\beta\gamma\|x^{t-2}-x^{t-3}\|^2\big)^{(2^2)} = (\tfrac{1}{2}\beta\gamma)^{(2^3-1)}\|x^{t-2}-x^{t-3}\|^{(2^3)}.\end{aligned}$$

Fortsetzung der Iteration bis $t=0$ ergibt mit $q=\frac{1}{2}\alpha\beta\gamma$:

$$\|x^{t+1}-x^t\| \le (\tfrac{1}{2}\beta\gamma)^{(2^t-1)}\|x^1-x^0\|^{(2^t)} \le (\tfrac{1}{2}\beta\gamma)^{(2^t-1)}\alpha^{(2^t)} \le \alpha q^{(2^t-1)}.$$

Für beliebiges $m\in\mathbb{N}$ folgt damit wegen $q<1$:

$$\begin{aligned}\|x^{t+m}-x^t\| &\le \|x^{t+m}-x^{t+m-1}\|+\cdots+\|x^{t+2}-x^{t+1}\|+\|x^{t+1}-x^t\|\\ &\le \alpha q^{(2^{t+m-1}-1)}+\cdots+\alpha q^{(2^{t+1}-1)}+\alpha q^{(2^t-1)}\\ &\le \alpha q^{(2^t-1)}\big\{(q^{(2^t)})^{(2^{m-1}-1)}+\cdots+q^{(2^t)}+1\big\}\\ &\le \alpha q^{(2^t-1)}\sum_{j=0}^{\infty}(q^{(2^t)})^j \;\le\; \frac{\alpha q^{(2^t-1)}}{1-q^{(2^t)}}.\end{aligned}$$

Dies besagt, dass $(x^t)_{t\in\mathbb{N}}\subset D_0$ Cauchy-Folge ist. Deren Limes $z\in D_0$ ist dann notwendig ein Fixpunkt von g bzw. Nullstelle von f:

$$z=\lim_{t\to\infty}x^t=\lim_{t\to\infty}g(x^{t-1})=g(z).$$

Durch Grenzübergang $m\to\infty$ erhalten wir auch die Fehlerabschätzung

$$\|z-x^t\|\le\frac{\alpha}{1-q}q^{(2^t-1)},$$

was den Beweis vervollständigt. Q.E.D.

Bemerkung 5.3: Unter der Annahme, dass eine Nullstelle $z\in G$ von f existiert kann die Aussage von Satz 5.1 für das Newton-Verfahren im $\mathbb{R}^1$ sinngemäß auf den $\mathbb{R}^n$ mit der Maximumnorm $\|\cdot\|_\infty$ verallgemeinert werden. Dabei sind die auftretenden Konstanten gemäß $m=1/\beta, M=\gamma$ zu identifizieren. Inbesondere gilt neben der a priori Fehlerabschätzung (5.5.42) auch die folgende a posteriori Fehlerabschätzung (Übungsaufgabe):

$$\|x^t-z\|_\infty\le\frac{1}{m}\,\|f(x^t)\|_\infty\le\frac{M}{2m}\,\|x^t-x^{t-1}\|_\infty^2,\quad t\in\mathbb{N}. \tag{5.5.45}$$

Beispiel 5.8: Zur Bestimmung der Inversen $Z=A^{-1}$ einer regulären Matrix $A\in\mathbb{R}^{n\times n}$ wird gesetzt

$$f(X):=X^{-1}-A,$$

für $X\in\mathbb{R}^{n\times n}$ regulär. Eine Nullstelle dieser Abbildung $f(\cdot):\mathbb{R}^{n\times n}\to\mathbb{R}^{n\times n}$ ist gerade die Inverse $Z=A^{-1}$. Diese soll mit dem Newton-Verfahren berechnet werden. Dazu ist zunächst eine Umgebung von A bzw. von A^{-1} zu bestimmen, auf der $f(\cdot)$ definiert und

differenzierbar ist. Für $X \in K_\rho(A)$ mit $\rho < \|A^{-1}\|^{-1}$ folgt aus $X = A - A + X = A(I - A^{-1}(A - X))$ die Beziehung

$$\|A^{-1}(A - X)\| \le \|A^{-1}\| \, \|A - X\| \le \rho \|A^{-1}\| < 1,$$

d. h.: $I - A^{-1}(A - X)$ und damit auch X sind regulär. Als nächstes ist die Jacobi-Matrix $f'(\cdot)$ von $f(\cdot)$ als Abbildung von $\mathbb{R}^{n\times n}$ in sich zu bestimmen. Für die Durchführung des Newton-Verfahrens genügt es offensichtlich, die Wirkung von $f'(\cdot)$ auf Matrizen $Y \in \mathbb{R}^{n\times n}$ zu bestimmen. Wir wollen zeigen, dass

$$f'(X)Y = -X^{-1}YX^{-1}, \quad Y \in \mathbb{R}^{n\times n}.$$

Dies sieht man wie folgt: Aus $f(X) = X^{-1} - A$ folgt $Xf(X) = I - XA$. Für die Jacobi-Matrizen der rechten und linken Seite gilt

$$\begin{aligned}
\big([Xf(X)]'Y\big)_{j,k} &= \sum_{pq} \frac{\partial}{\partial x_{pq}} \sum_l x_{jl} f_{lk}(X)\, y_{pq} \\
&= \sum_{p,q} \sum_l \Big\{ \underbrace{\frac{\partial x_{jl}}{\partial x_{pq}}}_{\delta_{jp}\cdot\delta_{lq}} f_{lk}(X) + x_{jl} \frac{\partial f_{lk}}{\partial x_{pq}}(X) \Big\} y_{pq} \\
&= \sum_q f_{qk}(X) y_{jq} + \sum_{p,q} \sum_l x_{jl} \frac{\partial f_{lk}}{\partial x_{pq}}(X) y_{pq} \\
&= \big(Yf(X) + Xf'(X)Y \big)_{jk}.
\end{aligned}$$

Analog finden wir

$$[I - XA]'Y = -YA\,.$$

Also ist

$$-YA = Yf(X) + Xf'(X)Y = YX^{-1} - YA - Xf'(X)Y$$

bzw.

$$f'(X)Y = -X^{-1}YX^{-1}.$$

Das Newton-Verfahren

$$f'(X^t)X^{t+1} = f'(X^t)X^t - f(X^t)$$

erhält in diesem Fall also die Gestalt

$$-X^{t^{-1}}X^{t+1}X^{t^{-1}} = -X^{t^{-1}} \underbrace{X^t X^{t^{-1}}}_{=I} - X^{t^{-1}} + A$$

bzw.

$$X^{t+1} = 2X^t - X^t A X^t = X^t\{2I - AX^t\}. \tag{5.5.46}$$

Diese Iteration ist das mehrdimensionale Analogen der Iteration $x_{t+1} = x_t(2 - ax_t)$ im skalaren Fall zur divisionsfreien Berechnung des Kehrwertes $1/a$ einer Zahl $a \neq 0$. Über

die Identität

$$X^{t+1} - Z = 2X^t - X^t A X^t - Z = -(X^t - Z)A(X^t - Z) \tag{5.5.47}$$

gewinnt man die Fehlerabschätzung

$$\|X^{t+1} - Z\| \leq \|A\| \, \|X^t - Z\|^2. \tag{5.5.48}$$

Der Einzugsbereich der quadratischen Konvergenz für das Newton-Verfahren ist in diesem Fall also die Menge

$$\{X \in \mathbb{R}^{n\times n} | \, \|X - Z\| < \|A\|^{-1}\}.$$

5.5.1 Gedämpftes Newton-Verfahren

Bei der Durchführung des Newton-Verfahrens zur Lösung nichtlinearer Gleichungssysteme treten zwei Hauptschwierigkeiten auf:

(i) hoher Aufwand pro Iterationsschritt,

(ii) hinreichend „guter“ Startpunkt x^0 erforderlich.

Zur Überwindung dieser Probleme verwendet man gegebenenfalls das sog. „vereinfachte Newton-Verfahren“

$$f'(c)\delta x^t = f(x^t), \quad x^{t+1} = x^t - \delta x^t, \tag{5.5.49}$$

mit einem geeigneten $c \in \mathbb{R}^n$, etwa $c = x^{(0)}$, welches nahe bei der Nullstelle z liegt. Dabei haben alle zu lösenden Gleichungssysteme dieselbe Koeffizientenmatrix und können mit Hilfe einer einmal berechneten LR-Zerlegung von $f'(c)$ effizient gelöst werden. Andererseits führt man zur Vergrößerung des Konvergenzbereiches des Newton-Verfahrens eine „Dämpfung“ ein,

$$f'(x^t)\delta x^t = f(x^t), \quad x^{t+1} = x^t - \lambda_t \delta x^t, \tag{5.5.50}$$

wobei der Parameter $\lambda_t \in (0, 1]$ zu Beginn klein gewählt wird und dann nach endlich vielen Schritten gemäß einer geeigneten Dämpfungsstrategie $\lambda_t = 1$ gesetzt wird. Der folgende Satz gibt ein konstruktives Kriterium für die a posteriori Wahl des Dämpfungsparameters λ_t.

Satz 5.6 (gedämpftes Newton-Verfahren): *Unter den Voraussetzungen von Satz 5.5 erzeugt für jeden Startpunkt $x^0 \in D_*$ die gedämpfte Newton-Iteration (5.5.50) mit*

$$\lambda_t := \min\Big\{1, \frac{1}{\alpha_t \beta \gamma}\Big\}, \qquad \alpha_t := \|f'(x^t)^{-1} f(x^t)\|, \tag{5.5.51}$$

eine Folge $(x^t)_{t\in\mathbb{N}}$, *für welche nach* t_* *Schritten* $q_* := \frac{1}{2}\alpha_{t_*}\beta\gamma < 1$ *erfüllt ist, so dass ab dann* x^t *quadratisch konvergiert, mit der a priori Fehlerabschätzung*

$$\|x^t - z\| \le \frac{\alpha}{1-q_*}\, q_*^{(2^t-1)}, \quad t \ge t_*. \tag{5.5.52}$$

Beweis: Wir verwenden wieder die Bezeichnungen aus dem Beweis von Satz 5.5. Für ein $x \in D_0$ gilt mit $x_r := x - rf'(x)^{-1}f(x)$, $0 \le r \le 1$, und $\alpha_x := \|f'(x)^{-1}f(x)\|$ die Abschätzung

$$\|f(x_r)\| \le (1 - r + \tfrac{1}{2}r^2\alpha_x\beta\gamma)\|f(x)\|, \quad 0 \le r \le R = \max\{r|\, x_s \in D_0, 0 \le s \le r \le 1\}.$$

Der Vorfaktor wird minimal für

$$r_* = \min\Big\{1, \frac{1}{\alpha_x\beta\gamma}\Big\} > 0 : \quad 1 - r_* + \tfrac{1}{2}r_*^2\alpha_x\beta\gamma \;\le\; 1 - \frac{1}{2\alpha_x\beta\gamma} < 1.$$

Bei Wahl von

$$r_t := \min\Big\{1, \frac{1}{\alpha_t\beta\gamma}\Big\}$$

ist also $(x^t)_{t\in\mathbb{N}} \subset D_0$, und die Norm $\|g(x^t)\|$ fällt streng monoton, d. h.:

$$\|f(x^{t+1})\| \le \Big(1 - \frac{1}{2\alpha_t\beta\gamma}\Big)\|f(x^t)\|.$$

Nach endlich vielen, $t_* \ge 1$, Iterationsschritten ist dann $\frac{1}{2}\alpha_{t_*}\beta\gamma < 1$, und die quadratische Konvergenz der weiteren Folge $(x^t)_{t\ge t_*}$ folgt aus Satz 5.5. Q.E.D.

5.6 Übungsaufgaben

Übung 5.1: Man berechne mit einem Fehler kleiner 10^{-6} die Nullstelle $z = \pi$ der Funktion $f(x) = \sin(x)$:

a) mit der Intervallschachtelung zum Startintervall $[2, 4]$;

b) mit der Fixpunktiteration $x_t = x_{t-1} + f(x_{t-1})$ zum Startwert $x_0 = 4$;

c) mit dem Newton-Verfahren $x_t = x_{t-1} - f'(x_{t-1})^{-1}f(x_{t-1})$ zum Startwert $x_0 = 4$.

Warum konvergiert in diesem Fall die einfache Fixpunktiteration (b) genauso schnell wie das Newton-Verfahren ?

Übung 5.2: Zur Berechnung der Lösung $z \in [0.5, 0.6]$ der Gleichung $x + \ln(x) = 0$ werden folgende Fixpunktiterationen vorgeschlagen:

a) $x_t = -\ln(x_{t-1})$;

b) $x_t = e^{-x_{t-1}}$;

c) $x_t = \frac{1}{2}(x_{t-1} + e^{-x_{t-1}})$.

Welche dieser Iterationen kann man verwenden, welche sollte man verwenden, und lässt sich vielleicht eine noch „bessere“ Iteration angeben?

Übung 5.3: Es sei $a > 0$ gegeben. Man zeige, dass für beliebigen Startwert $x_0 > 0$ die Fixpunktiteration

$$x_t = \frac{x_{t-1}^3 + 3ax_{t-1}}{3x_{t-1}^2 + a}, \quad t = 1, 2, \ldots,$$

monoton gegen $z = \sqrt{a}$ konvergiert. Wie groß ist die lokale Konvergenzordnung? Man überprüfe durch einen numerischen Test das theoretische Ergebnis.

Übung 5.4: Für zweimal stetig differenzierbare Funktionen f konvergiert das Newton-Verfahren lokal quadratisch gegen eine Nullstelle z. Man zeige, dass es für (nur) stetig differenzierbare Funktionen immer noch „super-linear“ konvergiert,

$$\left|\frac{x_t - z}{x_{t-1} - z}\right| \to 0 \quad (t \to \infty),$$

d. h.: Es ist asymptotisch schneller als die einfache Fixpunktiteration.

Übung 5.5 (Praktische Aufgabe): Man schreibe ein Programm zur Berechnung der Nullstellen eines Polynoms

$$p(x) = a_0 + a_1 x + \ldots + a_n x^n, \quad a_n \neq 0,$$

mit Hilfe des Newton-Verfahrens, wobei zur Auswertung von $p(x)$ und $p'(x)$ das Horner Schema zu verwenden ist. Startwerte sollen etwa durch Intervallschachtelung ermittelt werden. Als Abbruchkriterium frage man ab, ob gilt:

$$\frac{|x_{t+1} - x_t|}{|x_t|} = \left|\frac{p(x_t)}{p'(x_t)x_t}\right| \leq 10^{-8}.$$

Dieses Kriterium wird auch im Fall einer mehrfachen Nullstelle, d. h. $p'(x_t) \to 0$ $(t \to \infty)$, verwendet. Es muß allerdings abgefragt werden, ob $p'(x_t)x_t \neq 0$ ist.

Man berechne mit diesem Programm sämtliche Nullstellen der Legendre- und der Tschebyscheff-Polynome vom Grad $p = 4$ und $p = 5$, d. h. die Stützstellen der entsprechenden Gaußschen Quadraturformeln. Dabei sollen jeweils alle Iterierte inkl. der Startwertberechnung bis zur Erfüllung des Abbruchkriteriums ausgegeben werden.
(Hinweis: Die Polynome $L_k(x)$ und $T_k(x)$ erhält man mit Hilfe der Rekursionsformeln für die Legendre- und die Tschebyscheff-Polynome.)

Übung 5.6: Zur Berechnung der Inversen A^{-1} einer regulären Matrix $A \in \mathbb{R}^{n\times n}$ werden die beiden Fixpunktiterationen

$$\begin{aligned} &a) \quad X_t = X_{t-1}(I - AC) + C, \quad t = 1, 2, \ldots, \quad C \in \mathbb{R}^{n\times n} \text{ regulär}, \\ &b) \quad X_t = X_{t-1}(2I - AX_{t-1}), \quad t = 1, 2, \ldots, \end{aligned}$$

betrachtet. Man gebe hinreichende Kriterien für die Konvergenz dieser Iterationen an. Wie würde in diesem Fall das Newton-Verfahren lauten?

Übung 5.7: Es sollen die Schnittpunkte des durch $x_1^2 + x_2^2 = 2$ gegebenen Kreises und der durch $x_1^2 - x_2^2 = 1$ gegebenen Hyperbel bestimmt werden. Wie lauten die exakten Lösungen?

a) Man schreibe die Aufgabenstellung als Nullstellenproblemn einer geeigneten Abbildung $f = f(x_1, x_2) : \mathbb{R}^2 \to \mathbb{R}^2$ und iteriere ausgehend von dem Startwert $x^{(0)} = (1, 1)^T$ mit dem Newton-Verfahren bis das Inkrement $\|x^{(t)} - x^{(t-1)}\|_\infty$ kleiner als 2×10^{-3} ist.

b) Man bestimme zu der Abbildung f aus (a) eine Matrix $C \in \mathbb{R}^{2\times 2}$ in der Form

$$C = \begin{bmatrix} c & c \\ c & -c \end{bmatrix}, \quad c \neq 0,$$

so dass die Fixpunktiteration

$$x^{(t+1)} = x^{(t)} - Cf(x^{(t)})$$

ausgehend vom Startwert $x^{(0)} = (1, 1)^T$ garantiert gegen die Nullstelle z von f im ersten Quadranten der (x_1, x_2)-Ebene konvergiert. Wie viele Schritte müsste man mit der gewählten Fixpunktiteration machen, damit $\|x^{(t)} - z\|_\infty$ kleiner als 2×10^{-3} ist? (Hinweis: Bei den Abschätzungen verwende man die Maximumnorm.)

Übung 5.8: Die Eigenwertaufgabe $Ax = \lambda x$ einer Matrix $A \in \mathbb{R}^{n\times n}$ ist äquivalent zu dem nichtlinearen Gleichungssystem

$$\begin{aligned} Ax - \lambda x &= 0, \\ \|x\|_2^2 - 1 &= 0, \end{aligned}$$

von $n + 1$ Gleichungen in den $n + 1$ Unbekannten $x_1, \ldots, x_n, \lambda$.

a) Man gebe die Newton-Iteration zur Lösung dieses Gleichungssystems an.

b) Man führe zwei (oder bei Interesse auch mehr) Newton-Schritte durch für die Matrix

$$A = \begin{bmatrix} 4 & 0 \\ -1 & 4 \end{bmatrix}$$

mit den Startwerten $x_1^0 = 0$, $x_2^0 = 1.5$, $\lambda^0 = 3.5$. Man berechne die Eigenwerte und Eigenvektoren dieser Matrix und stelle fest, ob das Newton-Verfahren in diesem Fall quadratisch konvergiert.

Übung 5.9: Man untersuche die Konvergenz der Fixpunktiteration

$$x^t = Bx^{t-1} + c$$

für die Matrizen

$$(i) \quad B = \begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.2 & 0.5 & 0.7 \\ 0.1 & 0.1 & 0.1 \end{bmatrix}, \qquad (ii) \quad B = \begin{bmatrix} 0 & 0.5 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

Was ist der Limes der Folgen im Falle der Konvergenz? (Hinweis: Es sind die Eigenwerte der Matrizen abzuschätzen. Dazu kann eine geeignete Norm oder auch der Zusammenhang zwischen den Eigenwerten und der Determinante einer Matrix dienen.)

Übung 5.10 (Praktische Aufgabe): Man schreibe ein Programm zur Realisierung der Iterationsverfahren aus Aufgabe 5.6 für die (positiv definite) Tridiagonalmatrix

$$A = \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}$$

für $n = 2^k$, $k = 2, \ldots, 10$. Zur Vermeidung zu langer Rechenzeiten sollte eine obere Schranke (etwa $k \leq 10^5$ für die Anzahl der Iterationsschritte gesetzt werden. Mit der Matrix $C = \frac{1}{8} I \in \mathbb{R}^{n \times n}$ ist in diesem Fall nach Aufgabe 11.1 die Konvergenz der Iteration (a) für jeden Startwert garantiert. Man verwende daher versuchsweise für beide Iterationen (a) und (b) die Startmatrix $X_0 = \frac{1}{8} I$. Als Abbruchkriterium wähle man die Größe des Residuums $AX_t - I$ gemäß

$$\|AX_t - I\|_\infty = \max_{i=1,\ldots,n} \Big(\sum_{j=1}^{n} |(AX_t)_{ij} - \delta_{ij}| \Big) \leq 10^{-8}.$$

Man gebe die Anzahl der benötigten Iterationen in Abhängigkeit von n an. Sind diese Verfahren konkurrenzfähig mit der direkten Berechnung der Inversen mit Hilfe der simultanen Gauß-Elimination?