Kapitel 3

Nichtlineare Gleichungen

3.1 Existenz von Lösungen

Sei $f : K^n \to K^n$ eine Abbildung. Gesucht ist ein $\overline{x} \in K^n$ mit $f(\overline{x}) = 0$.

Beispiele:

1) $f(x) = x^2 - 2px - q$ für $K = \mathbb{C}$ oder $K = \mathbb{R}$. Sei $d = p^2 + q$. Im komplexen Fall gibt es zwei Lösungen für $d \neq 0$, eine Lösung für $d = 0$. Im reellen Fall gibt es für $d > 0$ zwei Lösungen, für $d = 0$ eine, für $d < 0$ überhaupt keine. Die Berechnung der Lösung $\overline{x}$ kann durch die Formel

$$x_{1,2} = p \pm \sqrt{d}$$

erfolgen. Die Auswertung dieser Formel ist aber im Hinblick auf Rundungsfehler keineswegs harmlos. Ist etwa $p > 0$, $|q| \ll p$, so ist $\sqrt{d} \sim p$, und bei der Berechnung von x_2 tritt Auslöschung auf. In diesem Fall ist es besser, x_2 nach der Formel $x_2 = -q/x_1$ zu berechnen.

2) $f(x) = x^3 + 3px - 2q$ für $K = \mathbb{C}$. Die Berechnung der - maximal drei - Lösungen kann durch die Cordani'schen Formeln erfolgen:

$$\begin{array}{rclcrclcrcll} x_1 & = & u + v & , & x_2 & = & \varepsilon_1 u + \varepsilon_2 v & , & x_3 & = & \varepsilon_2 u + \varepsilon_1 v & , \\ u & = & (q + \sqrt{d})^{1/3} & , & v & = & (q - \sqrt{d})^{1/3} & , & d & = & p^3 + q^2 & , \end{array}$$

$$\varepsilon_{1,2} = -\frac{1}{2}(1 \pm i\sqrt{3}) \, .$$

Auch hier tritt unter Umständen, nämlich für $|p| \ll |q|$, Auslöschung auf.

3) $f(x) = x - \tan x$, $K = \mathbb{R}$. Ein Blick auf den Graphen von $\tan x$ zeigt, daß es in jedem Intervall $((k - \frac{1}{2}), (k + \frac{1}{2})\pi)$, $k \in \mathbb{Z}$, genau eine Lösung gibt.

Ein primitives Verfahren zur Lösung von Gleichungen in $\mathbb{R}^1$ ist die Intervallhalbierung. Sei $f : \mathbb{R} \to \mathbb{R}$ stetig und $f(a)f(b) < 0$, $a < b$. Dann liegt in (a, b) sicher eine Nullstelle von f. Zu ihrer Berechnung verwenden wir den Algorithmus

```
f_a = f(a)  ;   f_b = f(b);
    while       (b - a >= ε)
            {   c = (b + a)/2;
                f_c = f(c);
                if (f_a f_c < 0)  {b = c; f_b = f_c;}
                else {a = c; f_a = f_c;}
            }
```

Nach $n + 2$ Funktionsauswertungen hat er das Intervall (a, b), in dem eine Lösung liegt, um den Faktor 2^n verkürzt. Für einfache Funktionen f ist dies akzeptabel. Wir werden natürlich effizientere Methoden kennenlernen.

Das theoretische Hilfsmittel zur Lösung nichtlinearer Gleichungen sind Fixpunktsätze. $x \in \mathbb{R}^n$ heißt Fixpunkt einer Abbildung $g : \mathbb{R}^n \to \mathbb{R}^n$, falls $g(x) = x$. Jede Aufgabe $f(x) = 0$ läßt sich in der Form $g(x) = x$ schreiben. Man braucht ja nur $g(x) = f(x) + x$ zu setzen.

Satz 3.1.1 *(Fixpunktsatz von Brouwer): Sei $D \subseteq \mathbb{R}^n$ konvex und kompakt, $g : D \to D$ stetig. Dann besitzt g in D einen Fixpunkt.*

Beweis: Der Beweis für $n \geq 1$ findet sich in E. Burger, Einführung in der Theorie der Spiele, 2. Auflage, S. 162-165. Für $n = 1$ ist der Satz trivial: Sei $D = [a, b]$. Ist $g(a) = a$ oder $g(b) = b$, so besitzt g einen Fixpunkt. Andernfalls ist $g(a) > a$, $g(b) < b$. Die Funktion $f(x) = x - g(x)$ hat dann in $[a, b]$ einen Zeichenwechsel und damit eine Nullstelle, und diese ist Fixpunkt von g.

□

Beispiele:

1) In $\mathbb{R}^2$ betrachten wir

$$g\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \sin(x_2 + e^{x_1}) \\ \cos(x_1 - e^{x_2}) \end{pmatrix}$$

Die Menge $D = \{x \in \mathbb{R}^2 : \|x\|_\infty \leq 1\}$ ist konvex und kompakt, und g ist dort stetig. Offenbar ist $g(D) \subseteq D$. Also hat g in D einen Fixpunkt.

2) Der Brouwer'sche Satz wird häufig auf Probleme der Volkswirtschaft angewendet. Ein Markt bestehe aus n Gütern mit Preisen $x_1, \ldots, x_n \geq 0$ und aus m Teilnehmern (Produzenten oder Konsumenten). Bei den Preisen $x = (x_1, \ldots, x_n)$ kauft oder produziert der ℓ-te Teilnehmer die Menge $f_i^\ell(x)$ von Gut i. Die totale Nachfrage bzw. Produktion von Gut i ist dann $f_i(x) = \sum_{\ell=1}^m f_i^\ell(x)$. Man sagt, der Markt sei im Gleichgewicht, wenn

$$f_i(x) \leq 0 \ , \quad i = 1, \ldots, n, \quad f_i(x) = 0 \quad \text{falls } \ x_i > 0 \ .$$

Wir machen folgende Voraussetzungen:

1) Die Nachfrage hängt nur von den relativen Preisen ab, d.h.

$$f_i^\ell(tx) = f_i^\ell(x) \quad , \quad t > 0 \ .$$

2) Der Markt ist abgeschlossen, d.h.

$$\sum_{i=1}^n x_i f_i^\ell(x) = 0 \quad , \quad \ell = 1, \ldots, m$$

3) Die f_i^ℓ sind stetig.

Dann gibt es einen Preisvektor x, bei welchem der Markt im Gleichgewicht ist.

Zum Beweis setzen wir $D = \{x \in \mathbb{R}^n : x \geq 0, \ \sum_{i=1}^n x_i = 1\}$ und auf D

$$g_i(x) = \frac{x_i + \ \mathrm{Max}\ (0, f_i(x))}{1 + \sum_{k=1}^n \ \mathrm{Max}\ (0, f_k(x))} \ , \quad i = 1, \ldots, n \ .$$

Die Abbildung $g(x) = (g_1(x), \ldots, g_n(x))$ ist dann stetig auf der konvexen kompakten Menge D und bildet D in sich ab. Nach dem Satz gibt es ein $x \in D$ mit $g(x) = x$, also

$$x_i = \frac{x_i + \text{ Max } (0, f_i(x))}{1 + \sum\limits_{k=1}^{n} \text{ Max } (0, f_k(x))} \quad , \quad i = 1, \ldots, n \; .$$

Hieraus folgt, daß x Gleichgewichtspunkt ist (Aufgabe 26).

3.2 Iterationsverfahren

Iterationsverfahren gehören zu den wichtigsten Hilfsmitteln der Numerischen Mathematik. Sie berechnen mittels einer - oft sehr einfachen - Rekursionsformel eine Folge von Näherungen, welche gegen die gesuchte Lösung konvergiert. Grundlage vieler Iterationsverfahren ist der Fixpunktsatz für kontrahierende Abbildungen.

Definition 3.2.1 *Sei $D \subseteq K^n$ und $g : D \to K^n$ eine Abbildung. g heißt kontrahierend in D (bezüglich der Norm $\|\cdot\|$ in K^n), wenn es eine Konstante $q < 1$ gibt mit*

$$\|g(x) - g(y)\| \leq q\|x - y\|$$

für alle x, $y \in D$. q heißt Lipschitzkonstante von g in D.

Satz 3.2.1 *Sei $D \subseteq \mathbb{R}^n$ konvex und $g : D \to \mathbb{R}^n$ differenzierbar. Sei*

$$q = \sup_{x \in D} \|g'(x)\| < 1 \ .$$

Dann ist g in D (bezüglich $\|\cdot\|$) kontrahierend. Dabei ist g' die Jacobi-Matrix zu g, also

$$g = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & , \dots , & \frac{\partial g_1}{\partial x_n} \\ \vdots & & \\ \frac{\partial g_n}{\partial x_1} & , \dots , & \frac{\partial g_n}{\partial x_n} \end{pmatrix} \quad \textit{für} \quad g = \begin{pmatrix} g_1 \\ \vdots \\ g_n \end{pmatrix} .$$

Beweis: Sei $f(t) = g(tx + (1-t)y)$, $0 \leq t \leq 1$. Nach der Kettenregel ist

$$f'(t) = g'(tx + (1-t)y)(x - y) \ ,$$

also

$$\begin{aligned} \|g(x) - g(y)\| &= \|f(1) - f(0)\| = \|\int_0^1 f'(t)dt\| \\ &\leq \sup_{0 \leq t \leq 1} \|f'(t)\| \\ &= \sup_{0 \leq t \leq 1} \|g'(tx + (1-t)y)(x - y)\| \\ &\leq q\|x - y\| \ , \end{aligned}$$

weil D konvex ist.

$\square$

Satz 3.2.2 *(Kontraktionssatz, Fixpunktsatz von Banach): Sei $D \subseteq K^n$ abgeschlossen und $g : D \to D$ kontrahierend. Dann hat g in D genau einen Fixpunkt $\overline{x}$. Das Iterationsverfahren*

$$x^{k+1} = g(x^k)\ , \quad k = 0, 1, \ldots$$

konvergiert für jede Wahl von $x^0 \in D$ gegen $\overline{x}$, und es gilt mit der Lipschitz-Konstante q von g in D

$$\|x^k - \overline{x}\| \leq \frac{q^k}{1-q}\|x^1 - x^0\|\ .$$

Beweis:

1) Existenz von $\overline{x}$.

Es ist $x^k \in D$ falls $x^0 \in D$, und

$$\|x^{k+1} - x^k\| \leq q\|x^k - x^{k-1}\| \leq q^2\|x^{k-1} - x^{k-2}\| \leq \ldots \leq q^k\|x^1 - x^0\|\ ,$$

also für $\ell > j$

$$\begin{aligned}
\|x^\ell - x^j\| &= \|\sum_{k=j}^{\ell-1}(x^{k+1} - x^k)\| \\
&\leq \sum_{k=j}^{\ell-1}\|x^{k+1} - x^k\| \\
&\leq \sum_{k-j}^{\ell-1} q^k\|x^1 - x^0\| \\
&= q^j(1 + \cdots + q^{\ell-1-j})\|x^1 - x^0\| \\
&\leq \frac{q^j}{1-q}\,\|x^1 - x^0\| \qquad (2.1)
\end{aligned}$$

wegen der Formel für die geometrische Reihe. Also gilt $x^\ell - x^j \to 0$ für ℓ, $j \to \infty$, d.h. (x^k) ist eine Cauchy-Folge und damit konvergent gegen ein $\overline{x} \in \mathbb{R}^n$. Da D abgeschlossen ist, ist sogar $\overline{x} \in D$, und wegen der Stetigkeit von g ist

$$\overline{x} = \lim_{k\to\infty} x^k = \lim_{k\to\infty} g(x^{k-1}) = g(\overline{x}) \ ,$$

also $\overline{x}$ Fixpunkt von g.

2) Eindeutigkeit

Wäre $\tilde{x}$ ein weiterer Fixpunkt von g in D, so hätten wir

$$\|\overline{x} - \tilde{x}\| = \|g(\overline{x}) - g(\tilde{x})\| \leq q\|\overline{x} - \tilde{x}\| \ .$$

Wegen $q < 1$ folgte $\|\overline{x} - \tilde{x}\| = 0$.

3) Fehlerabschätzung

Lassen wir in (2.1) $\ell \to \infty$ streben, so folgt

$$\|\overline{x} - x^j\| \leq \frac{q^j}{1-q}\|x^1 - x^0\| \ .$$

□

Beispiel: Wir wollen die Lösung von $x = \tan x$, in $[\frac{\pi}{2}, \frac{3\pi}{2}]$ durch Iteration berechnen und versuchen zunächst

$$x^{k+1} = \tan x^k \ .$$

Es ist $g(x) = \tan x$ und damit $g'(x) = 1/\cos^2 x \geq 1$. g ist also nicht kontrahierend. Wir müssen unsere Gleichung erst in geeignete Form bringen. Dazu schreiben wir für $x \in [\frac{\pi}{2}, \frac{3\pi}{2}]$ für $x = \tan x$

$$x = \tan(x - \pi) \quad \text{oder} \quad \arctan x = x - \pi \quad \text{oder} \quad x = \pi + \arctan x$$

und setzen $g(x) = \pi + \arctan x$. Dann ist $g'(x) = 1/(1 + x^2)$. In $D = [\frac{\pi}{2}, \frac{3\pi}{2}]$ ist dann $|g'(x)| \leq 1/(1 + \pi^2/4) < 1$, und g bildet D in sich ab. Nach Satz

3.2.1 ist g in D kontrahierend mit $q_1 = 1/(1+\pi^2/4) = 0.2884$. Also hat g in D genau einen Fixpunkt, und das Iterationsverfahren

$$x^{k+1} = \pi + \arctan x^k$$

konvergiert gegen diesen. Mit $x^0 = \pi$ erhalten wir

k	x^k	$\frac{q_1^k}{1-q_1}\lvert x^1 - x^0\rvert$	$\frac{q_2^k}{1-q_2}\lvert x^1 - x^0\rvert$	$\overline{x} - x^k$
0	3.1416	–	–	1.3518
1	4.4042	0.5117	0.1279	0.0892
2	4.4891	0.1476	0.0118	0.0043
3	4.4932	0.0426	0.0011	0.0002
4	4.4934	0.0122	0.0001	0.0000

Wir sehen, daß die Fehlerabschätzung viel zu pessimistisch ist. Man kann auch $D = [\pi, \frac{3\pi}{2}]$ wählen mit $q_2 = 1/(1+\pi^2) = 0.092$ und bekommt dann bessere Abschätzungen.

Welche Fixpunkte von g sind durch das Iterationsverfahren $x^{k+1} = g(x^k)$ berechenbar? - Durch Skizzen im $\mathbb{R}^1$ kommt man zu der Vermutung, daß dies die "anziehenden" Fixpunkt sind, d.h. diejenigen mit $|g'(\overline{x})| < 1$. Dies ist der Inhalt des nächsten Satzes.

Satz 3.2.3 *(Lokaler Konvergenzsatz): Die Abbildung $g : K^n \to K^n$ besitze einen Fixpunkt $\overline{x}$, und es gebe eine Umgebung von $\overline{x}$, in der g kontrahierend ist. Dann gibt es eine Umgebung $U(\overline{x})$, so daß das Iterationsverfahren $x^{k+1} = g(x^k)$ für jedes $x^0 \in U(\overline{x})$ gegen $\overline{x}$ konvergiert.*

Beweis: Wir können annehmen, daß g in $D = \{x \in K^n : \|x - \overline{x}\| \leq r\}$ mit einem, $r > 0$ kontrahierend ist, und zwar mit der Lipschitz-Konstanten $q < 1$. Dann ist für $x \in D$

$$\|g(x) - \overline{x}\| = \|g(x) - g(\overline{x})\| \leq q\|x - \overline{x}\| < qr\ ,$$

also auch $g(x) \in D$. g bildet also die abgeschlossene Menge D in sich ab und ist dort kontrahierend. Nach Satz 3.2.2 konvergiert das Iterationsverfahren für $x^0 \in D$.

$\square$

Bemerkungen:

1) Sei $K = \mathbb{R}$. Die Bedingung des Satzes III.2.3 ist erfüllt, wenn g in $\overline{x}$ stetig differenzierbar ist und $\rho(g'(\overline{x})) < 1$ ist. Dann gibt es nämlich nach Satz I.3.2 eine Norm $\| \cdot \|$ in $\mathbb{R}^n$ mit $\|g'(\overline{x})\| < 1$ für $x = \overline{x}$. Wegen der Stetigkeit von g' gilt dies dann auch in einer konvexen Umgebung von $\overline{x}$. Nach Satz III.2.1 ist g in dieser Umgebung kontrahierend.

2) Für die Konvergenzgeschwindigkeit des Iterationsverfahrens ist offenbar die Lipschitz-Konstante q entscheidend:

$$\|x^{k+1} - \overline{x}\| = \|g(x^k) - g(\overline{x})\| \leq q\|x^k - \overline{x}\| \ .$$

Der Fehler vermindert sich also in jedem Schritt (mindestens) um den Faktor $q < 1$. Wir sprechen von linearer Konvergenz. Nach Satz III.2.3 und Bemerkung 1 ist für die Konvergenzgeschwindigkeit die Zahl $\rho(g'(\overline{x}))$ entscheidend. Gilt hingegen für ein Iterationsverfahren

$$\|x^{k+1} - \overline{x}\| \leq C\|x^k - \overline{x}\|^p$$

mit einer Zahl $p > 1$, so sprechen wir von Konvergenz (mindestens) der Ordnung p. Für $p = 2$ sprechen wir von quadratischer, für $p = 3$ von kubischer Konvergenz. Quadratische Konvergenz ist dadurch gekennzeichnet, daß sich die Anzahl der korrekten Dezimalen von x^k bei jedem Schritt verdoppelt.

3.3 Das Newton-Verfahren

Zu lösen sei das nichtlineare System $f(x) = 0$, $f : D \to \mathbb{R}^n$, $D \subseteq \mathbb{R}^n$. Die Konvergenzgeschwindigkeit des Iterationsverfahrens $x^{k+1} = g(x^k)$ wird nach Bemerkung 2 aus dem vorigen Paragraphen durch die Zahl $\rho(g'(\overline{x}))$ bestimmt. Bei der Umwandlung einer Gleichung der Form $f(x) = 0$ in eine Fixpunktgleichung $x = g(x)$ versuchen wir daher $g'(\overline{x}) = 0$ zu erreichen. Wir machen für g den Ansatz

$$g(x) = x + A(x)f(x)$$

mit einer invertierbaren Matrix $A = (a_{i,j})$, d.h.

$$g_i(x) = x_i + \sum_{j=1}^{n} a_{i,j}(x)f_j(x) .$$

Wir berechnen die Jacobi-Matrix $g'(x)$. Es ist

$$\frac{\partial g_i}{\partial x_\ell} = \delta_{i,\ell} + \sum_{j=1}^{n} \frac{\partial a_{i,j}}{\partial x_\ell} f_j + \sum_{j=1}^{n} a_{i,j} \frac{\partial f_j}{\partial x_\ell} .$$

Für $x = \overline{x}$ ist $f(\overline{x}) = 0$, und wir können diese Gleichungen zusammenfassen zu

$$g'(\overline{x}) = I + A(\overline{x})f'(\overline{x}) .$$

Um $g'(\overline{x}) = 0$ zu erreichen, brauchen wir also nur

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

zu wählen. Damit wird

$$g(x) = x - (f'(x))^{-1}f(x) , \tag{3.1}$$

und das Iterationsverfahren zur Lösung von $f(x) = 0$ lautet

$$x^{k+1} = x^k - (f'(x^k))^{-1}f(x^k) . \tag{3.2}$$

In dieser Form verlangt das Verfahren die Inversion von $f'(x^k)$. Dies ist für große n unzweckmäßig. Man schreibt daher

$$f'(x^k)(x^{k+1} - x^k) + f(x^k) = 0 . \tag{3.3}$$

Dies ist das Newton-Verfahren zur Lösung von $f(x) = 0$. Man hätte es auch einfacher durch Linearisierung herleiten können. Ist x^k eine Näherung für die gesuchte Lösung $\overline{x}$, so hat man für $f \in C^2(D)$

$$f(x) = f(x^k) + f'(x^k)(x - x^k) + \mathrm{O}\,(\|x - x^k\|^2)\ .$$

Man vernachlässigt nun $\mathrm{O}(\|x - x^k\|^2)$ und nimmt als neue Näherung x^{k+1} für $\overline{x}$ die Lösung von

$$0 = f(x^k) + f'(x^k)(x - x^k)\ .$$

Dies ist genau (3.3). Im $\mathbb{R}^1$ ersetzt das Newton-Verfahren also die Kurve $y = f(x)$ durch die Tangente in x^k und berechnet x^{k+1} als Nullstelle der Tangente.

Beispiel: Wir lösen in $\mathbb{R}^1$ die Gleichung $x^2-2 = 0$, berechnen also $\overline{x} = \sqrt{2}$. Es ist

$$f(x) = x^2 - 2\ , \quad q(x) = x - \frac{f(x)}{f'(x)} = x - \frac{x^2-2}{2x} = \frac{1}{2}(x + \frac{2}{x})\ .$$

Mit $x^0 = 1$ erhalten wir

k	x_k	Anzahl der korrekten Dezimalen
0	1	1
1	$1.\underline{5}$	1
2	$1.41\underline{7}$	3
3	$1.41421\underline{6}$	6
4	$1.41421356237\underline{5}$	12

Satz 3.3.1 *Die Abbildung $f : \mathbb{R}^n \to \mathbb{R}^n$ besitze die Nullstelle $\overline{x}$, sei in einer Umgebung von $\overline{x}$ zweimal stetig differenzierbar, und $f'(\overline{x})$ sei invertierbar. Dann gibt es eine Umgebung D von $\overline{x}$, so daß das Newton-Verfahren für alle $x^0 \in D$ quadratisch gegen $\overline{x}$ konvergiert.*

Beweis: Sei $g(x) = x - (f'(x))^{-1}f(x)$. Wegen $g'(\overline{x}) = 0$ gibt es eine Umgebung D von $\overline{x}$ mit $\|g'(x)\| \leq 1/2$ für $x \in D$. Nach Satz III.2.3 konvergiert das Iterationsverfahren $x^{k+1} = g(x^k)$, also das Newton-Verfahren, gegen $\overline{x}$, wenn nur $x^0 \in D$.

Zum Nachweis der quadratischen Konvergenz bilden wir

$$x^{k+1} - \overline{x} = x^k - \overline{x} - (f'(x^k))^{-1}(f(x^k) - f(\overline{x}))\ .$$

Der Satz von Taylor liefert

$$f(x^k) - f(\overline{x}) = f'(x^k)(x^k - \overline{x}) + \varepsilon(\overline{x} - x^k)\ ,$$

$$\|\varepsilon(\overline{x} - x^k)\| \leq M\|\overline{x} - x^k\|^2$$

mit einer Konstanten M, welche die zweiten Ableitungen von f enthält. Dazu setzen wir D so klein voraus, daß $f \in C^2(D)$. Es folgt

$$\begin{aligned} x^{k+1} - \overline{x} &= x^k - \overline{x} - (x^k - \overline{x}) - (f'(x^k))^{-1}\varepsilon(\overline{x} - x^k) \\ &= -f'(x^k)^{-1}\varepsilon(\overline{x} - x^k)\ . \end{aligned}$$

Sei nun N eine Konstante mit $\|(f'(x^k))^{-1}\| \leq N$. Dann gilt

$$\|x^{k+1} - \overline{x}\| \leq NM\|x^k - \overline{x}\|^2\ ,$$

und dies bedeutet quadratische Konvergenz.

□

Was passiert, wenn $f'(\overline{x})$ nicht invertierbar ist? - Wir betrachten den Fall $n = 1$. Sei also $f(\overline{x}) = 0,\ \ f \in C^2,\ \ f'(\overline{x}) = 0$, aber $f''(\overline{x}) \neq 0$. Dann ist

$$f(x) = (x - \overline{x})^2 p(x)\ , \quad p(\overline{x}) \neq 0$$

mit $p \in C^2$. Wir berechnen

$$f'(x) = 2(x-\overline{x})p(x)+(x-\overline{x})^2p'(x), f''(x) = 2p(x)+4(x-\overline{x})p'(x)+(x-\overline{x})^2p''(x)$$

und erhalten für $g(x) = x - f(x)/f'(x)$

$$\begin{aligned} g'(x) &= 1 - \frac{f'(x)}{f'(x)} + \frac{f(x)f''(x)}{f'^2(x)} = \frac{(x-\overline{x})^2p(x)(2p(x) + \mathrm{O}(x-\overline{x}))}{4(x-\overline{x})^2(p(x) + \mathrm{O}(x-\overline{x}))^2} \\ &= \frac{1}{2}(1 + \mathrm{O}(x - \overline{x}))\ . \end{aligned}$$

Also ist $g'(\overline{x}) = \frac{1}{2}$, und nach Satz III.2.3 folgt lineare Konvergenz. Für $f'(\overline{x}) = 0$ konvergiert das Newton-Verfahren also immer noch, aber die quadratische Konvergenz geht verloren.

Der Rechenaufwand für das Newton-Verfahren wird im wesentlichen durch die Berechnung von $f'(x)$ bestimmt. Es gibt verschiedene Varianten des Newton-Verfahrens, die diesen Aufwand reduzieren.

1. Das vereinfachte Newton-Verfahren

Hier ersetzt man einfach $\left(f'(x^k)\right)^{-1}$ durch $\left(f'(x^0)\right)^{-1}$, iteriert also gemäß

$$f'(x^0)(x^{k+1} - x^k) = -f(x^k) .$$

Man hat immer noch lokale Konvergenz, aber die quadratische Konvergenz geht verloren.

2. Das Sekanten-Verfahren ("Regula falsi") $(n = 1)$

Hier ersetzt man die Tangente des Newton-Verfahrens durch die Sekante in den Punkten x^k, x^{k-1}, also

$$y = f(x^k) + (x - x^k)\,\frac{f(x^k) - f(x^{k-1})}{x^k - x^{k-1}} ,$$

und nimmt als neue Näherung x^{k+1} die Nullstelle dieser Sekante, also

$$x^{k+1} = x^k - \left(\frac{f(x^k) - f(x^{k-1})}{x^k - x^{k-1}}\right)^{-1} f(x^k) .$$

Mit anderen Worten: Man ersetzt $f'(x^k)$ durch den Differenzenquotienten in x^k, x^{k-1}. Dies ist auch in $\mathbb{C}$ sinnvoll.

Satz 3.3.2 *Sei $f : \mathbb{R} \to \mathbb{R}$ in einer Umgebung der Nullstelle $\overline{x}$ zweimal stetig differenzierbar, und sei $f'(\overline{x}) \neq 0$. Dann gibt es eine Umgebung D von $\overline{x}$, so daß das Sekanten-Verfahren für jede Wahl von x^0, x^1 in D gegen $\overline{x}$ konvergiert, und zwar gilt $\|\overline{x} - x^k\| \leq c_k$, wobei $c_k \to 0$ mit der Konvergenzordnung $(1 + \sqrt{5})/2 = 1.618$.*

Beweis:

(a) Es gibt eine Konstante C, so daß

$$|x^{k+1} - \overline{x}| \leq C|x^k - \overline{x}||x^{k-1} - \overline{x}| \; , \quad k = 1, 2, \dots \; .$$

Hierzu schreiben wir

$$\begin{aligned} x^{k+1} - \overline{x} &= x^k - \overline{x} - \frac{x^k - x^{k-1}}{f(x^k) - f(x^{k-1})} f(x^k) \\ &= (x^k - \overline{x}) \frac{\dfrac{f(x^k) - f(x^{k-1})}{x^k - x^{k-1}} - \dfrac{f(x^k) - f(\overline{x})}{x^k - \overline{x}}}{\dfrac{f(x^k) - f(x^{k-1})}{x^k - x^{k-1}}} \\ &= (x^k - \overline{x}) \frac{\int_0^1 \{f'(x^{k-1} + t(x^k - x^{k-1})) - f'(\overline{x} + t(x^k - \overline{x}))\} dt}{f'(\xi^k)} \end{aligned}$$

mit ξ^k zwischen x^k und x^{k-1}. In einer Umgebung D von $\overline{x}$ ist

$$|f'(x)| \geq m > 0 \; , \quad |f''(x)| \leq M$$

und daher

$$|x^{k+1} - \overline{x}| \leq |x^k - \overline{x}| \, \frac{M|x^{k-1} - \overline{x}|}{m} \; ,$$

falls x_k, $x_{k-1} \in D$. Ist $D = [\overline{x} - \varepsilon, \overline{x} + \varepsilon]$ und x^0, $x^1 \in D$, so folgt also $|x_2 - \overline{x}| \leq \varepsilon^2 \frac{M}{m}$, und für $\varepsilon \frac{M}{m} \leq 1$ ist auch $x_2 \in D$. Damit bleiben alle x^k in D, und (a) ist gezeigt.

(b) Wir setzen $e^{f_k} = C|x^k - \overline{x}|$. Dann wird

$$f_{k+1} \leq f_k + f_{k-1} \; .$$

Sei $F_0 = F_1 = 1$ und $F_{k+1} = F_k + F_{k-1}$. Ist $|f_0|$, $|f_1| \leq 1$, so gilt offenbar $f_k \leq F_k$, $k = 0, 1, \dots$. Die F_k kann man leicht berechnen. Mit $\tau = \frac{1}{2}(1 + \sqrt{5})$ ist

$$F_k = c_1 \tau^k + c_2 (-\tau)^{-k} \; ,$$

$$c_1 = \frac{\tau^2}{1 + \tau^2} \quad , \quad c_2 = \frac{1}{1 + \tau^2} \; .$$

Wir setzen nun $\varepsilon_k = e^{F_k}$. Dann ist

$$\frac{\varepsilon_k}{\varepsilon_{k-1}^{\tau}} = e^{F_k - \tau F_{k-1}} = e^{c_2((-\tau)^{-k} - \tau(-\tau)^{-k+1})}$$

und dies strebt wegen $\tau > 1$ für $k \to \infty$ gegen 1. Also gilt $\varepsilon_k \leq c\varepsilon_{k-1}^{\tau}$ für ein geeignetes $c > 0$, und

$$|x^k - \overline{x}| \leq \frac{1}{c} e^{f_k} \leq \frac{1}{c} e^{F_k} = \frac{1}{c} \varepsilon_k \ .$$

Die Behauptung des Satzes folgt mit $c_k = \varepsilon_k / c$.

□