Kapitel 5

Lösung nichtlinearer Gleichungen

5.1 Das allgemeine Iterationsverfahren, Banachscher Fixpunktsatz

5.2 Nullstellen reeller Funktionen, Newton-Verfahren

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

Wir behandeln in diesem Kapitel die Lösung nichtlinearer Gleichungen und Gleichungssysteme. Dies wird oft erforderlich, wenn analytische Lösungsmethoden nicht verfügbar oder für praktische Zwecke zu kompliziert sind.

Beispiele:

(i) Die Gleichung $x = e^{-x}$ hat genau eine reelle Lösung x^*. Die Lösung x^* liegt im Intervall $[0, 1]$. Dies folgt durch Monotoniebetrachtung und Anwendung des Zwischenwertsatzes. Verschiedene numerische Verfahren liefern die Näherung $x^* = 0.567143$ mit einer Genauigkeit von 10^{-6}.

(ii) Physikalische Grundlage: Der Betrag der Gravitationskraft von zwei Punktmassen m_1 und m_2 (in kg) in der Ebene mit dem Abstand $r = \|P_1 - P_2\|$ (in Meter) ist nach dem Newtonschen Gesetz

$$\|F\| = G\frac{m_1 m_2}{r^2} \qquad \text{mit} \quad G = 6.67 \cdot 10^{-11} Nm^2/kg,$$

der Kraftvektor F hat die Koordinatenform

$$F = \pm G\frac{m_1 m_2}{r^3}(P_1 - P_2) \in \mathbb{R}^2.$$

Aufgabe: Im ebenen Gravitationsfeld mit **drei** gegebenen Punktmassen m_k in den Punkten

$$P_1 = (x_1, 0), \quad P_2 = (x_2, 0), \quad P_3 = (0, y_3).$$

sollen die Koordinaten eines Gleichgewichtspunkts $P = (x, y)$ bestimmt werden. Trägt der Punkt selbst die Masse m, so wirken die drei Kräfte

$$F_k(x, y) = G\frac{m m_k}{\|P_k - P\|^3}\,(P_k - P)\,, \qquad k = 1, 2, 3.$$

Das Gleichgewicht wird also durch die Gleichung

$$F_1(x,y) + F_2(x,y) + F_3(x,y) = \vec{0}$$

beschrieben. Dies ist ein nichtlineares Gleichungssystem mit zwei Unbekannten x und y (die Masse m wird herausgekürzt).

5.1 Das allgemeine Iterationsverfahren, Banachscher Fixpunktsatz

In vielen Anwendungen sind Gleichungssysteme zu lösen, in denen die Unbekannten **nichtlinear** auftreten. Wir behandeln in diesem Abschnitt allgemeine Gleichungen der Form

$$x = F(x), \qquad x \in D \subset V, \tag{5.1}$$

wobei V ein Banachraum (= vollständiger normierter Vektorraum) und $F : D \to V$ eine stetige Funktion ist. Im Standardfall $V = \mathbb{R}^n$ mit der euklidischen Norm oder der Maximumsnorm ist

$$F(x) = (F_1(x), \ldots, F_n(x))^T$$

eine stetige vektorwertige Funktion. Die Gleichung (5.1) ist dann ein nichtlineares Gleichungssystem mit n Veränderlichen $(x_1, \ldots, x_n) \in D$.

5.1.1 Definition: Fixpunkt, Fixpunktverfahren im normierten Raum V
Es sei V ein Banachraum, $D \subseteq V$ nichtleer und $F : D \to V$ eine stetige Funktion.

a) Ein Element $x^* \in D$ mit $x^* = F(x^*)$ heißt *Fixpunkt* von F.

b) Falls $F(x) \in D$ für alle $x \in D$ gilt, so nennt man F eine *Selbstabbildung*.

c) Für eine Selbstabbildung $F : D \to D$ kann zu einem *Startwert* $x^{(0)} \in D$ die Folge

$$x^{(k+1)} = F(x^{(k)}), \qquad k = 0, 1, 2, \ldots$$

gebildet werden. Diese Iterationsvorschrift nennt man *allgemeines Iterationsverfahren*, *Fixpunktverfahren* oder *Methode der sukzessiven Approximation*.

Bemerkung: Manche Autoren sprechen von einer Funktion F von D "in sich" und meinen damit die Eigenschaft der Selbstabbildung. Für eine Funktion $F : [a,b] \to \mathbb{R}$ prüft man diese Eigenschaft mit den üblichen Methoden der Kurvendiskussion: Monotoniebereiche, relative und absolute Maxima/Minima von F.

Unter bestimmten Voraussetzungen an F liefert das Fixpunktverfahren eine *konvergente Folge* von Punkten $x^{(k)} \in D$.

5.1.2 Beispiel: Im ersten Beispiel zur Lösung von $x = e^{-x}$ ergibt sich zum Startwert $x_0 = 0$ die Folge $x_{k+1} = e^{-x_k}$, $k \geq 0$, also

$$x_0 = 0, \quad x_1 = 1, \quad x_2 = 0.36788, \ldots, x_{20} = 0.567135...$$

Diese Folge lässt sich gut veranschaulichen, wenn wir den Graphen von $F(x) = e^{-x}$ und die Winkelhalbierende $y = x$ zeichnen. Der Wert x_0 wird dabei nicht auf der x-Achse, sondern auf der Winkelhalbierenden markiert, von dort geht es vertikal zum Graphen von F, dann horizontal zur Winkelhalbierenden: dies markiert x_1, von hier vertikal zum Graphen und horizontal zur Wh'en zu x_2, usw. Man erkennt den Charakter eines "anziehenden Fixpunkts" im folgenden Bild.

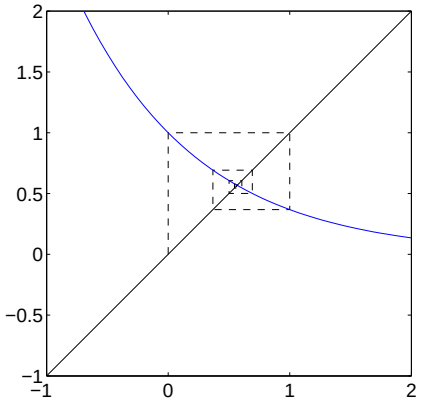

Abbildung 5.1: Fixpunktverfahren für $F(x) = e^{-x}$ zum Startwert $x_0 = 1$.

Eine einfache Aussage klärt zunächst, dass als mögliche Grenzwerte des Fixpunktverfahrens wirklich nur die Fixpunkte von F in Frage kommen. Hier wird noch nichts über die Konvergenz selbst ausgesagt.

5.1.3 Lemma
Es sei V ein Banachraum, $D \subset V$ abgeschlossen, $F : D \to D$ eine stetige Selbstabbildung und $(x^{(k)})_{k\geq 0}$ die zum Startwert $x^{(0)} \in D$ gebildete Folge des Fixpunktverfahrens.

Falls diese Folge konvergiert, so gilt

$$x^* = \lim_{k\to\infty} x^{(k)} \in D \qquad \text{und} \qquad F(x^*) = x^*.$$

Beweis: Die Abgeschlossenheit von D und $x^{(k)} \in D$ für alle k ergibt $x^* \in D$. Die Stetigkeit von F ergibt

$$x^* = \lim_{k\to\infty} x^{(k+1)} = \lim_{k\to\infty} F(x^{(k)}) = F\Big(\lim_{k\to\infty} x^{(k)}\Big) = F(x^*).$$

Der zentrale Begriff bei der Untersuchung der Konvergenz ist die *Kontraktion*.

5.1.4 Definition: kontrahierende Abbildung
Es sei V ein Banachraum, $D \subset V$ nichtleer und $F : D \to V$ eine Funktion.

a) F heißt *Lipschitz-stetig* mit der *Lipschitzkonstanten* $L \geq 0$, falls gilt

$$\|F(x) - F(y)\| \leq L\|x - y\| \qquad \text{für alle} \quad x, y \in D.$$

b) F heißt *stark kontrahierend* (bzgl. der gegebenen Norm), falls F Lipschitz-stetig mit einer Lipschitzkonstanten $L < 1$ ist. In diesem Fall nennt man L auch *Kontraktionszahl* von F.

Bemerkung für $V = \mathbb{R}^n$: Die Lipschitz-Stetigkeit von F ist eine Eigenschaft, die unabhängig von der gegebenen Norm ist, weil alle Normen auf $\mathbb{R}^n$ äquivalent sind. Sie impliziert die Stetigkeit von F. Jedoch hängt die Konstante L von der gegebenen Norm ab: die Kontraktionseigenschaft kann bzgl. einer Norm erfüllt und bezüglich einer anderen Norm verletzt sein.

Der Begriff "kontrahierend" wird manchmal für die schwächere Eigenschaft

$$\|F(x) - F(y)\| < \|x - y\| \qquad \text{für alle} \quad x, y \in D$$

verwendet. Daher benutzen wir das Adverb "stark" um anzugeben, dass die Konstante $L < 1$ in der Lipschitz-Bedingung steht. Dies ist eine wesentliche Verschärfung (vgl. etwa die Aussagen zum Quotientenkriterium bei Reihen).

Wir betrachten meistens Funktionen $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$. Die Berechnung von Lipschitz-Konstanten L erfolgt mit Hilfe der Ableitung (im eindimensionalen Fall) bzw. der Funktionalmatrix (im n-dimensionalen Fall).

5.1.5 Satz: Lipschitz-Konstanten differenzierbarer Funktionen auf $\mathbb{R}^n$
Die Menge $D \subset \mathbb{R}^n$ sei *konvex*. $F : D \to \mathbb{R}^n$ mit den Komponentenfunktionen $F_1, \ldots, F_n : D \to \mathbb{R}$ sei stetig differenzierbar, ihre Funktionalmatrix bezeichnen wir mit

$$DF(x) = \begin{pmatrix} \frac{\partial F_1(x)}{\partial x_1} & \cdots & \frac{\partial F_1(x)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial F_n(x)}{\partial x_1} & \cdots & \frac{\partial F_n(x)}{\partial x_n}. \end{pmatrix}$$

Zur gegebenen Norm auf $\mathbb{R}^n$ verwenden wir eine verträgliche Matrixnorm und setzen

$$L := \sup_{x \in D} \|DF(x)\|.$$

Falls $L < \infty$ gilt, ist F Lipschitz-stetig und L ist eine Lipschitzkonstante von F.

Beweis: (siehe Analysis II) Für jede Komponente F_j und alle $x, y \in D$ gilt

$$F_j(y) - F_j(x) = \int_0^1 \nabla F(x + t(y - x)) \cdot (y - x)\, dt,$$

weil D konvex ist. Daraus ergibt sich die Vektorgleichung

$$F(y) - F(x) = \int_0^1 DF(x + t(y - x)) \cdot (y - x)\, dt,$$

wobei auf der rechten Seite ein vektorwertiges Integral steht (Integration jeder Komponente). Die Dreiecksungleichung für das Integral und die Verträglichkeit der Norm erlaubt die Abschätzung

$$\|F(y) - F(x)\| \leq \int_0^1 \|DF(x + t(y - x))\| \cdot \|y - x\| \, dt.$$

Hieraus folgt sofort die Behauptung.

Jetzt haben wir alle Bestandteile gesammelt, um einen der wichtigsten Sätze der Analysis zu beweisen, der von Stefan Banach[1] stammt.

5.1.6 Banachscher Fixpunktsatz

Sei V ein Banachraum, $D \subset V$ abgeschlossen und $F : D \to D$ eine *stark kontrahierende Selbstabbildung* mit der Kontraktionskonstante $0 \leq L < 1$. Dann gilt:

a) F hat genau einen Fixpunkt $x^* \in D$.

b) Für jeden Startwert $x^{(0)} \in D$ konvergiert die Folge $(x^{(k)})_{k\geq 0}$ der Fixpunktiteration gegen diesen Fixpunkt.

c) Für alle $k \geq 1$ gilt

$$\|x^{(k)} - x^*\| \leq L \, \|x^{(k-1)} - x^*\| \qquad \text{(mindestens lineare Konvergenz)}$$

$$\|x^{(k)} - x^*\| \leq \frac{L^k}{1 - L} \, \|x^{(1)} - x^{(0)}\| \qquad \text{(a priori Fehlerabschätzung)}$$

$$\|x^{(k)} - x^*\| \leq \frac{L}{1 - L} \, \|x^{(k)} - x^{(k-1)}\| \qquad \text{(a posteriori Fehlerabschätzung)}$$

Beweis: 1. Es gibt höchstens einen Fixpunkt: $\|x_1^* - x_2^*\| = \|F(x_1^*) - F(x_2^*)\| \leq L\|x_1^* - x_2^*\|$ mit $L < 1$ ergibt notwendig $\|x_1^* - x_2^*\| = 0$, also $x_1^* = x_2^*$.

2. Die Folge $(x^{(k)})_{k\geq 0}$ der Fixpunktiteration $x^{(k)} = F(x^{(k-1)})$ zu einem beliebigen Startwert $x^{(0)} \in D$ ist eine Cauchyfolge in D: hierzu werden für $k \geq 0$ und $m \in \mathbb{N}$ die Abschätzungen

$$\|x^{(k+1)} - x^{(k)}\| \leq L\|x^{(k)} - x^{(k-1)}\| \leq \cdots \leq L^k\|x^{(1)} - x^{(0)}\|,$$

$$\|x^{(k+m)} - x^{(k)}\| \leq \sum_{j=0}^{m-1} \|x^{(k+j+1)} - x^{(k+j)}\| \leq \sum_{j=0}^{m-1} L^{k+j}\|x^{(1)} - x^{(0)}\| \leq \frac{L^k}{1 - L}\|x^{(1)} - x^{(0)}\|$$

hergeleitet. Dabei wird im letzten Schritt die endliche Summe auf die geometrische Reihe vergrößert. Die zweite Abschätzung sichert (wegen $L^k \to 0$ für $k \to \infty$) die Bedingung einer Cauchy-Folge.

3. Die Menge $D \subset V$ ist abgeschlossen, also selbst wieder vollständig. Deshalb konvergiert die Cauchy-Folge $(x^{(k)})_{k\to\infty}$ gegen ein $x^* \in D$. Nach Lemma 5.1.3 ist der Grenzwert x^* ein Fixpunkt von F.

Damit sind die Existenz (Teil 3) und die Eindeutigkeit (Teil 1) in Aussage a) bewiesen, ebenso die Konvergenz in Aussage b). Die erste Ungleichung in c) folgt direkt aus der Kontraktionseigenschaft. Die a-priori Ungleichung in c) folgt aus Beweisteil 2 durch die Grenzwertbetrachtung $m \to \infty$. Für die a-posteriori Ungleichung in c) zeigt man analog zu Beweisteil 2

$$\|x^{(k+m)} - x^{(k)}\| \leq \sum_{j=1}^{m} \|x^{(k+j)} - x^{(k+j-1)}\| \leq \sum_{j=1}^{m} L^j\|x^{(k)} - x^{(k-1)}\| \leq \frac{L}{1 - L}\|x^{(k)} - x^{(k-1)}\|$$

[1] polnischer Mathematiker 1892-1945, Begründer der Funktionalanalysis in seiner Dissertation on 1922

und lässt m gegen unendlich gehen.

Bemerkung: Für die Existenz mindestens eines Fixpunkts von $F : D \to D$ gibt es weitere berühmte Sätze (von Brouwer, Schauder); dabei wird die Konvexität von D gefordert, aber keine Kontraktionsbedingung benötigt.

Im eindimensionalen Fall $D = [a, b]$ folgt die Existenz mindestens eines Fixpunkts ganz einfach: die stetige Funktion $g : [a, b] \to \mathbb{R}$ mit $g(x) = x - F(x)$ hat mindestens eine Nullstelle, weil

$$g(a) = a - F(a) \leq 0 \leq b - F(b) = g(b).$$

Hingegen liefert der Banachsche Fixpunktsatz zusätzlich zur Existenz auch die Eindeutigkeit sowie ein Berechnungsverfahren, nämlich die Fixpunktiteration.

In praktischen Aufgaben ist meist eine Funktion $F : D \to \mathbb{R}^n$ gegeben, die weder eine Selbstabbildung von D noch kontrahierend auf D ist. Man erhält beide Eigenschaften erst durch geschickte **Einschränkung** des Definitionsbereichs auf eine Teilmenge $\widetilde{D} \subset D$. Dies gelingt im eindimensionalen Fall meist durch Betrachtung der Monotoniebereiche von F, im mehrdimensionalen Fall hilft manchmal die folgende Methode.

5.1.7 Methode: Kugelbedingung

Es sei V ein Banachraum, $D \subset V$ nichtleer und abgeschlossen sowie $F : D \to V$ stetig. Weiter seien $\xi_0 \in D$, $r > 0$ sowie $0 \leq L < 1$ mit den folgenden Eigenschaften gegeben:

(1) $K_r(\xi_0) = \{x \in V \mid \|x - \xi_0\| \leq r\} \subset D$.

(2) Es gilt die Kontraktionsbedingung

$$\|F(x) - F(y)\| \leq L\|x - y\| \qquad \text{für alle} \quad x, y \in K_r(\xi_0).$$

(3) Es gilt die *Kugelbedingung* $\|F(\xi_0) - \xi_0\| \leq r(1 - L)$.

Dann bildet F die Menge $K_r(\xi_0)$ in sich ab, d.h. die Einschränkung $F\,|_{K_r(\xi_0)}$ ist eine stark kontrahierende Selbstabbildung.

Beweis: Mit der Dreiecksungleichung, Kontraktions- und Kugelbedingung folgt

$$\|F(x) - \xi_0\| \leq \|F(x) - F(\xi_0)\| + \|F(\xi_0) - \xi_0\| \leq L\|x - \xi_0\| + (1 - L)r \leq r.$$

5.1.8 Veranschaulichung von Fixpunktverfahren in $\mathbb{R}$:

x^* sei Fixpunkt der stetig differenzierbaren Funktion $F : [a, b] \to [a, b]$. Mögliche Szenarien für die Iterationsfolge $(x^{(k)})$ werden durch das matlab-file `fixpointplot.m` dargestellt. Dabei werden drei Szenarien mit Hilfe der Ableitung von F im Fixpunkt unterschieden:

(i) $|F'(x^*)| > 1$: abstoßender Fixpunkt, z.B. $F(x) = \cosh(x)/2$ bei $x \approx 2.1$, keine Konvergenz

(ii) $0 < F'(x^*) < 1$: anziehender Fixpunkt, z.B. $F(x) = \cosh(x)/2$ bei $x \approx 0.6$ mit monotoner Konvergenz,

(iii) $-1 < F'(x^*) < 0$: anziehender Fixpunkt, z.B. $F(x) = \cos x$ bei $x \approx 0.7$ mit alternierender Konvergenz.

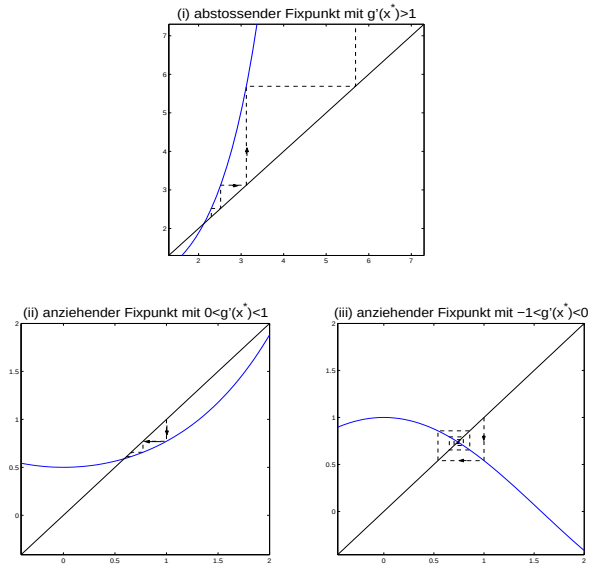

Abbildung 5.2: Drei unterschiedliche Typen von Fixpunkten: abstoßend (oben) und anziehend (unten) mit monotoner (links) bzw. alternierender (rechts) Konvergenz.

5.1.9 Bemerkung: Nullstelle vs. Fixpunkt

Ein nichtlineares Gleichungssystem

$$G(x) = 0$$

zu gegebener Funktion $G : \mathbb{R}^n \to \mathbb{R}^n$ (z.B. zur Lösung einer Extremwertaufgabe) lässt sich in vielfältiger Weise äquivalent umwandeln in die *Fixpunkt-Form*

$$F(x) = x.$$

Allg. Beispiel: mit einer invertierbaren Matrix $C \in \mathbb{R}^{n\times n}$ setze

$$F(x) = x + CG(x).$$

5.1.10 Beispiel in $\mathbb{R}$: Die Gleichung $G(x) = x^6 - x - 1 = 0$ besitzt eine positive Lösung im Intervall $[0, 2]$, und zwar $x^* \approx 1.13472413840152$.

1. Die Funktion $F_1(x) = (1 + x)^{1/6}$ hat als Fixpunkte alle Nullstellen von G. Sie ist eine stark kontrahierende Selbstabbildung auf $D = [0, 2]$. Die Fixpunktiteration zu F_1 mit Startwert $x_0 = 2$ konvergiert monoton gegen x^*, weil $0 < F_1'(x) < 1/6$ in D gilt.

   Zur Konvergenzgeschwindigkeit: wie in Satz 5.1.6(c) angegeben verkleinert sich der Fehler $|x_k - x^*|$ pro Schritt um einen Faktor $L \leq 1/6$, nach anfänglichem "Einpendeln" ist der Abnahmefaktor circa $F_1'(x^*) = 0.089$. Man nennt dies *lineare Konvergenz*.

2. Die Funktion $F_2(x) = x - \frac{x^6 - x - 1}{6x^5 - 1}$ (vgl. Newton-Verfahren in der Einleitung zu Beginn des Semesters und Abschnitt 5.2) besitzt als Fixpunkte ebenfalls die Nullstellen von G. Die Fixpunktiteration zu F_2 mit Startwert $x_0 = 2$ beginnt mit linearer Fehlerabnahme (bis x_4), ist dann im sog. Einzugsbereich von x^* (vgl. Abschnitt 5.2) und "schaltet um" auf quadratische Konvergenz (= Verdopplung genauer Stellen pro Schritt). Diese Beschleunigung der Konvergenz wird genauer begründet, man beachte hierzu $F_2'(x^*) = 0$.

F_1	F_2
2.00000000000000	2.00000000000000
1.20093695517600	1.68062827225131
1.14051569756263	1.43073898823906
1.13523664844046	1.25497095610944
1.13476953843976	1.16153843277331
1.13472816047178	1.13635327417051
1.13472449472677	1.13473052834363
1.13472416996929	1.13472413850022

Tabelle 5.1: Zwei Fixpunktiterationen zur Lösung von $x^6 - x - 1 = 0$ in der Nähe von $x = 2$

Wir wollen nun die **Konvergenz-Geschwindigkeit** einer konvergenten Folge $(x^{(k)})_{k\geq 0}$ mit

$$\lim_{k\to\infty} x^{(k)} = x^*$$

genauer beschreiben.

5.1.11 Definition: Konvergenzordnung
Es sei V ein Banachraum und $(x^{(k)})_{k\geq 0}$ eine konvergente Folge in V mit $\lim_{k\to\infty} x^{(k)} = x^*$.

a) Die Folge $(x^{(k)})_{k\geq 0}$ konvergiert *linear* (oder hat die *Konvergenzordnung* $p = 1$), falls es $k_0 \in \mathbb{N}$ und $0 < c < 1$ gibt, so dass gilt

$$\frac{\|x^{(k+1)} - x^*\|}{\|x^{(k)} - x^*\|} \leq c \quad \text{für alle} \quad k \geq k_0.$$

b) Die Folge $(x^{(k)})_{k\geq 0}$ konvergiert *superlinear*, falls

$$\limsup_{k\to\infty} \frac{\|x^{(k+1)} - x^*\|}{\|x^{(k)} - x^*\|} = 0.$$

c) Die Folge $(x^{(k)})_{k\geq 0}$ hat die *Konvergenzordnung* $p > 1$, falls gilt

$$\limsup_{k\to\infty} \frac{\|x^{(k+1)} - x^*\|}{\|x^{(k)} - x^*\|^p} < \infty.$$

Bemerkung:

- $p = 1$**:** Für $k > k_0$ gelten die a-priori und a-posteriori Fehlerabschätzungen des Banachschen Fixpunktsatzes:

$$\begin{aligned} \|x^{(k)} - x^*\| &\leq \frac{c^{k-k_0}}{1-c} \|x^{(k_0+1)} - x^{(k_0)}\|, \\ \|x^{(k)} - x^*\| &\leq \frac{c}{1-c} \|x^{(k)} - x^{(k-1)}\|. \end{aligned}$$

- $p = 2$**:** Man spricht von *quadratischer Konvergenz*, wie z.B. bei F_2 in Beispiel 5.1.10. Zur Erläuterung wählen wir $c > 0$ und $k_0 \in \mathbb{N}$ mit

$$\frac{\|x^{(k+1)} - x^*\|}{\|x^{(k)} - x^*\|^2} \leq c \quad \text{für alle} \quad k \geq k_0$$

und betrachten die Nullfolge $\epsilon_k = c\|x^{(k)} - x^*\|$. Dann gilt für $k \geq k_0$

$$\epsilon_{k+1} \leq c^2 \|x^{(k)} - x^*\|^2 = \epsilon_k^2,$$

also per Induktion

$$\epsilon_k \leq \epsilon_{k_0}^{2^{k-k_0}}.$$

Sobald man einmal $\epsilon_{k_0} < 1$ erreicht hat, tritt also sehr schnelle Konvergenz ein: grob gesprochen verdoppelt sich die Anzahl der exakten Dezimalstellen von $x^{(k)}$ zu $x^{(k+1)}$.

- Analog für $p > 1$ reell: ab einem Index k_0 erhöht sich die Anzahl der exakten Dezimalstellen um den Faktor p.

Wir betrachten nun den speziellen Fall der Fixpunkt-Iteration zu einer reellen Funktion $F : [a, b] \to \mathbb{R}$. Falls die Folge $(x_k)_{k\geq 0}$ mit $x_k = F(x_{k-1})$ gegen einen Fixpunkt x^* konvergiert, so lässt sich die Konvergenzordnung oft mit Methoden der Differentialrechnung bestimmen.

5.1.12 Satz: Konvergenzordnung der Fixpunkt-Iteration in $\mathbb{R}$

Es sei $p \in \mathbb{N}$. Die Funktion $F : [a, b] \to [a, b]$ sei p-mal stetig differenzierbar und besitze den Fixpunkt x^*. Weiter sei $x_0 \in [a, b]$ so, dass die Folge $(x_k)_{k\geq 0}$ des Fixpunktverfahrens $x_{k+1} = F(x_k)$ gegen x^* konvergiert und $x_k \neq x^*$ für alle $k \geq 0$ erfüllt.

a) Falls $p \geq 2$ und

$$F'(x^*) = \cdots = F^{(p-1)}(x^*) = 0, \qquad F^{(p)}(x^*) \neq 0,$$

gilt, so hat die Folge (x_k) die exakte Konvergenzordnung p.

b) Falls $p = 1$ und $0 < |F'(x^*)| < 1$ gilt, so konvergiert die Folge (x_k) linear, aber nicht superlinear.

Beweis: Man zeigt mit der Taylorformel und dem Mittelwertsatz für Integrale

$$|x_{k+1} - x^*| = |F(x_k) - F(x^*)| = \left| F^{(p)}(\xi_k) \, \frac{(x_k - x^*)^p}{p!} \right|$$

mit einer Stelle ξ_k zwischen x^* und x_k. Wegen $\lim_{k\to\infty} x_k = x^*$ und der Stetigkeit von $F^{(p)}$ folgt

$$\lim_{k\to\infty} \frac{\|x^{(k+1)} - x^*\|}{\|x^{(k)} - x^*\|^p} = \lim_{k\to\infty} \frac{|F^{(p)}(\xi_k)|}{p!} = \frac{|F^{(p)}(x^*)|}{p!}.$$

Für $p \geq 2$ ist die Bedingung der Konvergenzordnung p erfüllt. Außerdem ist für jedes $q > p$ die Folge

$$\frac{\|x^{(k+1)} - x^*\|}{\|x^{(k)} - x^*\|^q} = \|x^{(k)} - x^*\|^{p-q} \frac{\|x^{(k+1)} - x^*\|}{\|x^{(k)} - x^*\|^p}$$

unbeschränkt, also ist p die exakte Konvergenzordnung.

Für $p = 1$ folgt aus $0 < |F'(x^*)| < 1$ die lineare Konvergenz in Definition 5.1.11 mit der Konstanten $c = (1 + |F'(x^*)|)/2 < 1$. Außerdem ist die Bedinung der superlinearen Konvergenz nicht erfüllt.

5.1.13 Beispiel: Die Funktionen F_1 und F_2 in Beispiel 5.1.10 erfüllen

$$0 < |F_1'(x^*)| = \frac{1}{6}(1 + x^*)^{-5/6} < 1, \qquad F_2'(x^*) = 0, \quad F_2''(x^*) \neq 0.$$

Bei F_1 liegt lineare Konvergenz, bei F_2 quadratische Konvergenz vor. Dies wurde anhand der Zahlenwerte bereits in Beispiel 5.1.10 erläutert.

5.2 Nullstellen reeller Funktionen, Newton-Verfahren

Wir behandeln in diesem Abschnitt die Bestimmung von Nullstellen einer stetigen Funktion $f[a, b] \to \mathbb{R}$. Die verwendeten numerischen Verfahren fallen in zwei Kategorien:

- **Typ 1:** Iterationsverfahren ähnlich zu Abschnitt 5.1 (z.B. Newton-Verfahren, Sekantenverfahren):
  berechne eine *Iterationsfolge* $(x_k)_{k\in\mathbb{N}_0}$ mit

  $$\lim_{k\to\infty} x_k = z, \qquad z \text{ Nullstelle von } f.$$

- **Typ 2:** Einschließungsverfahren (z.B. Bisektion, Regula falsi):
  berechne eine Folge von Intervallen $[x_k, y_k] \subseteq [a, b]$, $k \geq 0$, mit

  $$\lim_{k\to\infty} x_k = z \quad \text{und/oder} \quad \lim_{k\to\infty} y_k = z, \qquad z \text{ Nullstelle von } f,$$

  und mit Vorzeichenwechsel $f(x_k)f(y_k) \leq 0$ zwischen den Intervallenden. Nach dem Zwischenwertsatz enthält dann jedes Intervall $[x_k, y_k]$ eine Nullstelle von f.

5.2.1 Algorithmus: Newton-Verfahren
Es sei $f \in C^1[a,b]$ und es gelte $f'(x) \neq 0$ für alle $x \in (a,b)$; f besitze eine Nullstelle $z \in [a,b]$. Das Newton-Verfahren lautet:
Wähle einen Startwert x_0 in $[a,b]$ und berechne für $k = 1, 2, \ldots$

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$

Das Verfahren bricht ab, wenn $x_k \notin [a,b]$ oder eine maximale Iterationszahl überschritten wird.

Ansonsten endet das Verfahren mit dem Ergebnis x_k, wenn $f(x_k) \approx 0$ oder eine Fehlerabschätzung vom Typ $|x_k - z| \leq \epsilon$ erreicht wurde (siehe Satz 5.2.4).

Geometrische Deutung: Beim Newton-Verfahren ist x_{k+1} die Nullstelle der Tangente an den Graphen von f im Punkt $(x_k, f(x_k))$.

Für die Praxis ist das Newton-Verfahren das wichtigste Verfahren zur Nullstellenberechnung, da es bei "guter Wahl" des Startwerts sehr schnelle Konvergenz von x_k gegen eine Nullstelle von f liefert. Falls f mehrere Nullstellen besitzt, versucht man diese durch unterschiedliche Startwerte zu berechnen.

5.2.2 Beispiel: Das Polynom $f(x) = x^3 - x^2 - x - 1$ hat zwei Nullstellen im Intervall $[-1, 0]$ und eine weitere Nullstelle im Intervall $[1, 2]$ (Begründung graphisch oder mit dem Zwischenwertsatz durch Einsetzen von $x = -1, -0.5, 0, 1, 2$). Die größte Nullstelle $z_3 \in [1, 2]$ erhalten wir durch Wahl des Startwerts $x_0 = 2$:

$$x_0 = 2, \quad x_1 = 1.7286, \quad x_2 = 1.65127, \quad x_3 = 1.64494046, \quad x_4 = 1.64489912.$$

Für die kleinste Nullstelle $z_1 \in [-1, -0.5]$ versuchen wir es mit Startwert $x_0 = -1$:

$$x_0 = -1, \quad x_1 = -0.725, \quad x_2 = -0.586, \quad x_3 = -0.537, \quad x_4 = -0.53038, \quad x_5 = -0.530247.$$

Dass hier jeweils monotone Konvergenz vorliegt, wird in den Übungen behandelt: f ist monoton und konkav in $(-\infty, z_1)$, monoton und konvex in $[z_3, \infty)$. Für die mittlere Nullstelle nehmen wir $x_0 = 0$:

$$x_0 = 0, \quad x_1 = -0.1, \quad x_2 = -0.1143, \quad x_3 = -0.1146520, \quad x_4 = -0.114652261114.$$

Das sieht alles sehr einfach aus, ABER: andere Startwerte können ganz anderes Verhalten der Folge x_k ergeben, z.B.

$$x_0 = 1, \quad f'(1) = 0, \text{ also bricht das Verfahren sofort ab.}$$

$$x_0 = -0.32, \; x_1 = 1.2867, \; x_2 = 1.9413, \; x_3 = 1.70684, \; x_4 = 1.64850, \; x_5 = 1.644912, \; \ldots,$$

also ergibt sich $x_k \to z_3$, obwohl x_0 viel näher bei z_1 und z_2 liegt. All dies lässt sich anhand des Graphen von f sehr gut erklären.

5.2.3 Bemerkung zur Durchführung des Newton-Verfahrens: Die Berechnung von $f(x_k)$ und $f'(x_k)$ für ein Polynom f erfolgt über das Hornerschema:

	1	-1	-1	-0.1	
$x_0 = 2$		2	2	2	
	1	1	1	**1.9**	$= f(2)$
$x_0 = 2$		2	6	$-$	
	1	3	**7**		$= f'(2)$

Also: zu $x_0 = 2$ liefert das Newton-Verfahren $x_1 = 2 - \frac{1.9}{7} \approx 1.7286$.

Dass manchmal Konvergenz vorliegt, manchmal nicht, und dass solche "Sprünge" von einem Monotoniebereich von f in einen anderen stattfinden können, wird durch den Begriff "lokal quadratische Konvergenz" ausgedrückt. Damit ist gemeint: **Falls** der Startwert nahe genug an der zu berechnenden Nullstelle liegt, dann erhalten wir sogar quadratische Konvergenz!

5.2.4 Satz: Lokale Konvergenz des Newton-Verfahrens im skalaren Fall
Die Funktion $f : [a, b] \to \mathbb{R}$ sei zweimal stetig differenzierbar und besitze eine Nullstelle $z \in (a, b)$ mit $f'(z) \neq 0$. Weiter gelte

$$m := \min_{x\in[a,b]} |f'(x)| > 0 \qquad \text{(also strikte Monotonie von } f \text{ in } [a, b]).$$

Es sei $M := \max_{x\in[a,b]} |f''(x)|$ und

$$0 < r < \frac{2m}{M} \quad \text{so, dass} \quad K_r(z) = [z - r, z + r] \subset [a, b].$$

Dann konvergiert die Folge (x_k) des Newton-Verfahrens für jeden Startwert $x_0 \in K_r(z)$ gegen z.

An die Stelle der Abschätzungen im Banachschen Fixpunktsatz treten die folgenden a-priori und a-posteriori Fehlerabschätzungen der **quadratischen** Konvergenz mit der Konstanten $L = \frac{Mr}{2m} < 1$:

$$|x_k - z| \leq \frac{M}{2m}|x_{k-1} - z|^2 \leq \frac{2m}{M}\, L^{(2^k)}, \tag{5.2}$$

$$|x_k - z| \leq \frac{1}{m}\,|f(x_k)| \leq \frac{M}{2m}\,|x_k - x_{k-1}|^2. \tag{5.3}$$

Wesentliche Beweisschritte:

1. Wegen $m > 0$ gibt es keine weitere Nullstelle in $[a, b]$ (Satz von Rolle).
2. Für beliebige $x, y \in [a, b]$, $x \neq y$, folgt aus dem Mittelwertsatz

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

Damit ist die erste Ungleichung in (5.3) gezeigt.

3. Die Taylorentwicklung um die Stelle $x \in [a,b]$ ergibt

$$0 = f(z) = f(x) + (z-x)f'(x) + \int_x^z f''(\xi)(z-\xi)\,d\xi.$$

Daraus erhalten wir

$$|f(x) - (x-z)f'(x)| \le \frac{M}{2}(x-z)^2$$

und weiter mit der Iterationsfunktion $\phi(x) = x - \frac{f(x)}{f'(x)}$ des Newton-Verfahrens

$$|\phi(x) - z| = \left|(x-z) - \frac{f(x)}{f'(x)}\right| \le \frac{M}{2m}(x-z)^2.$$

Wegen $\phi(x_{k-1}) = x_k$ ist die erste Ungleichung in (5.2) bewiesen.

4. Wegen $r < \frac{2m}{M}$ folgt für alle $x \in K_r(z)$

$$|\phi(x) - z| \le \frac{M}{2m}(x-z)^2 \le \underbrace{\frac{Mr}{2m}}_{<1}|x-z| < r,$$

also bildet ϕ das Intervall $K_r(z)$ in sich ab. Alle Folgenglieder x_k (bei beliebigem $x_0 \in K_r(z)$) liegen in $K_r(z)$.

5. Nun sind alle Vorbereitungen getroffen. Wir führen die Bezeichnung $\epsilon_k = \frac{M}{2m}|x_k - z|$ ein. Aus 3. folgt nach Multiplikation beider Seiten mit $\frac{M}{2m}$

$$\epsilon_k \le \epsilon_{k-1}^2 \le \cdots \le \epsilon_0^{(2^k)}.$$

Wegen

$$\epsilon_0 = \frac{M}{2m}|x_0 - z| \le \frac{Mr}{2m} = L < 1$$

folgen sowohl die Konvergenz $\epsilon_k \to 0$, also $x_k \to z$, als auch die zweite Ungleichung in (5.2).

6. Zum Beweis der zweiten Ungleichung in (5.3) führen wir analog zu 3. die Taylorentwicklung um die Stelle x_{k-1} durch,

$$|f(x_k) - f(x_{k-1}) - (x_k - x_{k-1})f'(x_{k-1})| \le \frac{M}{2}(x_k - x_{k-1})^2.$$

Als Nullstelle der Tangente erfüllt x_k die Beziehung $f(x_{k-1}) + (x_k - x_{k-1})f'(x_{k-1}) = 0$, also haben wir

$$|f(x_k)| \le \frac{M}{2}(x_k - x_{k-1})^2.$$

5.2.5 Bemerkung:

- Die Umgebung $K_r(z)$ in Satz 5.2.4 gehört zum *Einzugsbereich* der einfachen Nullstelle z der Funktion f : für einen Startwert x_0 in dieser Umgebung erfolgt sehr rasche Konvergenz. Z.B. für $L = 1/2$ erzielt man mit 5 bzw. 10 Iterationsschritten bereits

$$\begin{aligned} |x_5 - z| &\le \frac{2m}{M} \cdot 2^{-32} \approx \frac{4.6m}{M} \cdot 10^{-10}, \\ |x_{10} - z| &\le \frac{2m}{M} \cdot 2^{-1024} \approx \frac{10m}{M} \cdot 10^{-309}. \end{aligned}$$

Die wirkliche Schwierigkeit besteht häufig darin, mit dem Startwert überhaupt eine solche Umgebung zu treffen, da der Radius r sehr klein sein kann.

- In der Praxis findet man durch Ausprobieren einen Startwert x_0, für den das Newton-Verfahren konvergiert. Dabei ist x_0 häufig noch nicht im Einzugsbereich der Nullstelle z. Deshalb kann man zunächst eine langsame (lineare) Annäherung der ersten Iterierten $x_1, \ldots, x_{k_0}$ an die Nullstelle beobachten. Sobald einmal x_{k_0} im Einzugsbereich der quadratischen Konvergenz liegt, setzt die quadratische Konvergenz ein und nur noch 4-5 weitere Schritte liefern sehr hohe Genauigkeit.

- Achtung: Oft wird man gar keine Konvergenz erzielen, wenn der Startwert x_0 nicht nahe genug bei z gewählt wird.

In bestimmten Situationen ist das Newton-Verfahren sehr robust, d.h. Konvergenz wird auch bei weit entferntem Startwert erzielt.

5.2.6 Satz: hinreichende Voraussetzungen für Konvergenz des Newton-Verfahrens
$f : [a, b] \to \mathbb{R}$ sei zweimal stetig differenzierbar und es gelte

- $f(a)f(b) < 0$,
- $f'(x) \neq 0$ und $f''(x) \neq 0$ für alle $x \in [a, b]$.

Weiterhin gelte für den Startwert $x_0 \in [a, b]$ die Beziehung $f(x_0)f''(x_0) > 0$.

Dann hat f genau eine Nullstelle $z \in (a, b)$, die Iterierten x_k des Newton-Verfahrens zum Startwert x_0 liegen in $[a, b]$, und die Folge $(x_k)_{k\geq 0}$ konvergiert **monoton** gegen die Nullstelle z.

Beweis: Spezialfall $f' > 0$ und $f'' > 0$ als Übungsaufgabe.

5.2.7 Bemerkung: Abschließend soll noch ein kurzes Argument für die quadratische Konvergenz des Newton-Verfahrens angegeben werden. Dazu verwenden wir Satz 5.1.12. Die Iterationsfunktion des Newton-Verfahrens lautet

$$F(x) = x - \frac{f(x)}{f'(x)}.$$

Wir erhalten mit der Quotientenregel

$$F'(x) = 1 - \frac{(f'(x))^2 - f(x)f''(x)}{(f'(x))^2} = \frac{f(x)f''(x)}{(f'(x))^2}.$$

Wegen $f(z) = 0$ und $f'(z) \neq 0$ ist also $F'(z) = 0$. Deshalb liegt (mindestens) quadratische Konvergenz vor. Dies ist ein qualitatives Resultat. Unser Satz 5.2.4 gibt dazu eine viel genauere quantitative Aussage.

5.2.8 Ergänzung: modifiziertes Newton-Verfahren bei mehrfacher Nullstelle

Es sei $p \in \mathbb{N}$, $p \geq 2$. Die Funktion $f \in C^{p+1}[a,b]$ besitze eine p-fache Nullstelle $z \in (a,b)$, d.h.

$$f(z) = f'(z) = \cdots = f^{(p-1)}(z) = 0, \qquad f^{(p)}(z) \neq 0,$$

aber es sei $f'(x) \neq 0$ für $x \neq z$. Dann liefert das **modifizierte Newtonverfahren**

$$x_{k+1} = x_k - \mathbf{p}\ \frac{f(x_k)}{f'(x_k)}, \qquad k = 0, 1, 2, \ldots$$

lokal quadratische Konvergenz.

Kurze Begründung: Es sei $f \in C^{p+1}[a,b]$ mit p-facher Nullstelle z. Für $x \in [a,b]$ mit $f'(x) \neq 0$ benutzen wir die Taylor-Entwicklung von f und f' um z und erhalten

$$\frac{f(x)}{f'(x)} = \left(f^{(p)}(z)\frac{(x-z)^p}{p!} + r_1(x) \right) \Big/ \left(f^{(p)}(z)\frac{(x-z)^{p-1}}{(p-1)!} + r_2(x) \right)$$

mit den Restgliedern der Ordnung

$$|r_1(x)| = \mathcal{O}(|x-z|^{p+1}), \qquad |r_2(x)| = \mathcal{O}(|x-z|^p) \qquad \text{für } x \to z.$$

Geeignete Umformungen und Abschätzungen des Bruchs ergeben

$$\frac{f(x)}{f'(x)} = \frac{x-z}{p} + \mathcal{O}(|x-z|^2).$$

Mit der Iterationsfunktion $\phi(x) = x - p\frac{f(x)}{f'(x)}$ ergibt sich also

$$|\phi(x) - z| = \left| x - z - p\frac{f(x)}{f'(x)} \right| = \mathcal{O}(|x-z|^2).$$

Dies ersetzt Schritt 3. im Beweis von Satz 5.2.4, die weiteren Argumente folgen wie dort.

Wir wollen im Anschluss noch 3 weitere Verfahren diskutieren: Sekantenverfahren, die Regula falsi und das Bisektionsverfahren.

Falls die Berechnung der Ableitung von f zu mühsam ist, kann man das (langsamere) Sekantenverfahren verwenden. Es konvergiert unter ähnlichen Voraussetzungen wie das Newton-Verfahren. Der Unterschied besteht darin, dass immer zwei alte Werte x_{k-1} und x_k zur Berechnung von x_{k+1} verwendet werden, um anstatt der Ableitung von f eine Sekantensteigung zu berechnen.

5.2.9 Algorithmus: Sekantenverfahren
Es sei $f \in C^1[a,b]$ und es gelte $f'(x) \neq 0$ für alle $x \in (a,b)$; f besitze eine Nullstelle $z \in [a,b]$. Das Sekanten-Verfahren lautet:
Wähle **zwei** Startwerte $x_0 \neq x_1$ in $[a,b]$ und berechne für $k = 1, 2, \ldots$

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}\ f(x_k).$$

Das Verfahren bricht ab, wenn $x_k \notin [a,b]$ oder eine maximale Iterationszahl überschritten wird.

Ansonsten endet das Verfahren mit dem Ergebnis x_k, wenn $f(x_k) \approx 0$ oder eine Fehlerabschätzung vom Typ $|x_k - z| \leq \epsilon$ erreicht wurde (siehe Satz 5.2.4).

Das Sekantenverfahren besitzt wieder "lokale Konvergenz". Zur Angabe der Konvergenzordnung benötigen wir die Fibonacci-Folge

$$F_0 = F_1 = 1, \qquad F_{k+1} = F_k + F_{k-1} \quad \text{für} \quad k \geq 1.$$

Die Zahl $\varphi = \frac{1+\sqrt{5}}{2} \approx 1.618$ nennt man den "Goldenen Schnitt". Sie gibt das Wachstum der Fibonacci-Folge an: es gilt

$$F_n = \frac{1}{\sqrt{5}} \left(\varphi^{n+1} + (-1)^n \frac{1}{\varphi^{n+1}} \right).$$

Der folgende Satz besagt, dass das Sekantenverfahren (unter den angegebenen Voraussetzungen) die Konvergenzordnung $p = \varphi = \frac{1+\sqrt{5}}{2}$ besitzt.

5.2.10 Satz: Lokale Konvergenz des Sekanten-Verfahrens
Die Funktion $f : [a,b] \to \mathbb{R}$ sei zweimal stetig differenzierbar und besitze eine Nullstelle $z \in (a,b)$ mit $f'(z) \neq 0$. Weiter gelte

$$m := \min_{x \in [a,b]} |f'(x)| > 0.$$

Es sei $M := \max_{x \in [a,b]} |f''(x)|$ und $0 < r < \frac{2m}{M}$ so, dass $K_r(z) = [z - r, z + r] \subset [a,b]$ gilt.

Dann konvergiert die Folge (x_k) des Sekanten-Verfahrens für jede Wahl der Startwerte $x_0, x_1 \in K_r(z)$ (mit $x_0 \neq x_1$) gegen z. Es gelten die a-priori und a-posteriori

Fehlerabschätzungen

$$|x_k - z| \leq \frac{2m}{M} L^{F_k} \quad \text{mit} \quad L = \frac{Mr}{2m} < 1, \tag{5.4}$$

$$|x_k - z| \leq \frac{1}{m} |f(x_k)| \leq \frac{M}{2m} |x_k - x_{k-1}||x_k - x_{k-2}|. \tag{5.5}$$

Ein anderer Typ von Verfahren sind die sogenannten Einschließungsverfahren. Dabei wird zur Nullstelle z von $f : [a, b] \to \mathbb{R}$ eine Folge von Intervallen $[x_k, y_k]$ konstruiert mit

$$f(x_k)f(y_k) < 0,$$

so dass wegen des Zwischenwertsatzes immer mindestens eine Nullstelle im Intervall enthalten ist. Beim Bisektionsverfahren konvergiert die Intervalllänge gegen Null, so dass

$$\lim_{k\to\infty} x_k = \lim_{k\to\infty} y_k = z$$

gilt. Hingegen bleibt bei der Regula falsi in den meisten Fällen ein Intervallrand fest und nur eine der Folgen (x_k) oder (y_k) konvergiert gegen die Nullstelle z.

5.2.11 Einschließungsverfahren: Bisektion und Regula falsi
Es sei $f \in C[a, b]$.

1. Wähle **zwei** Startwerte $x_0, y_0 \in [a, b]$ mit $f(x_0)f(y_0) < 0$.
   Dann besitzt f eine Nullstelle im Intervall $I_0 = [x_0, y_0]$ bzw. $I_0 = [y_0, x_0]$.

2. Berechne für $k = 1, 2, \ldots$

$$x_{k+1} = \frac{x_k + y_k}{2} \qquad \text{beim Bisektionsverfahren, oder}$$

$$x_{k+1} = x_k - \frac{x_k - y_k}{f(x_k) - f(y_k)} f(x_k) \qquad \text{bei der Regula Falsi,}$$

$$\text{und setze} \qquad y_{k+1} = \begin{cases} x_k, & \text{falls } f(x_{k+1})f(x_k) \leq 0, \\ y_k, & \text{falls } f(x_{k+1})f(y_k) < 0. \end{cases}$$

   Dann besitzt f eine Nullstelle im Intervall $I_{k+1} = [x_{k+1}, y_{k+1}]$ bzw. $I_{k+1} = [y_{k+1}, x_{k+1}]$.

3. Beende das Verfahren, wenn $f(x_{k+1}) \approx 0$ oder $|y_{k+1} - x_{k+1}|$ die geforderte Genauigkeit ϵ unterschreitet.

5.2.12 Konvergenzbetrachtung

a) Das Bisektionsverfahren konvergiert linear mit $c = 1/2$:

$$|x_k - y_k| = 2^{-k}|x_0 - y_0|, \qquad k \geq 0.$$

Also gilt für eine Nullstelle z von f die *a priori* Abschätzung

$$|x_k - z| \leq 2^{-k}|x_0 - y_0|, \qquad k \geq 0.$$

b) Regula falsi: die Folge $(x_k)_{k\geq 0}$ (der neu berechneten Werte) konvergiert gegen eine Nullstelle z von f. Die Konvergenz ist meist linear.

Beachte: In den meisten Fällen gilt für das Intervallende y_k bei der Regula falsi $y_k = y_{k_0}$ für alle $k \geq k_0$; d.h. ein Intervallende bleibt fest und die Intervall-Länge $|y_k - x_k|$ konvergiert nicht gegen Null. Das andere Intervallende (im Algorithmus die neuen Werte x_{k+1}) konvergieren aber gegen z.

5.2.13 Beispiel: Die Funktion $f : \mathbb{R} \to \mathbb{R}$ mit $f(x) = e^x - x^2 - 3x$ hat eine Nullstelle $z = 0.45677721650410$. Wir wählen $x_0 = 0$ für das Newton-Verfahren, $x_0 = 0$ und $x_1 = 1$ für das Sekantenverfahren und Regula falsi.

Newton:

k	x_k	$f(x_k)$	$f'(x_k)$
0	0	1.0	-2.0
1	0.5	-0.10127872929987	-2.35127872929987
2	0.45692610661688	-0.00034760015192	-2.33464002662699
3	0.45677721850224	-0.00000000466481	-2.33457735881993
4	0.45677721650410	0.00000000000000	-2.33457735797867

Sekantenverfahren:

k	x_k	$f(x_k)$	$\frac{f(x_k)-f(x_{k-1})}{x_k-x_{k-1}}$
0	0	1.0	
1	2.0	-2.61094390106935	-1.80547195053467
2	0.55387180050283	-0.22841253688563	-1.64752431009377
3	0.41523194220986	0.09660863546484	-2.34435591865388
4	0.45644097309810	0.00078496243207	-2.32530760775826
5	0.45677854668597	-0.00000310541286	-2.33450682527118
6	0.45677721646375	0.00000000009420	-2.33457763798244
7	0.45677721650410	-0.00000000000000	-2.33458163537657

Regula falsi: mit $a_k = \min\{x_k, y_k\}$ und $b_k = \max\{x_k, y_k\}$ für das Intervall I_k:

k	a_k	x_{k+1}	b_k	$f(x_{k+1})$	$\frac{f(b_k)-f(a_k)}{b_k-a_k}$
0	0	0.55387180050283	2	-0.22841253688563	-1.80547195053467
1	0	0.45088419718269	0.55387180050283	0.01375034512662	-2.21786437903215
2	0.45088419718269	0.45673197648581	0.55387180050283	0.00010561589150	-2.35137894470151
3	0.45673197648581	0.45677687230257	0.55387180050283	0.00000080356507	-2.35246620106177
4	0.45677687230257	0.45677721388548	0.55387180050283	0.00000000611337	-2.35247447713840
5	0.45677721388548	0.45677721648418	0.55387180050283	0.00000000004651	-2.35247454010149
6	0.45677721648418	0.45677721650395	0.55387180050283	0.00000000000035	-2.35247454058050
7	0.45677721650395		0.55387180050283		

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

Wir behandeln nochmals das Newton-Verfahren, hier aber für nichtlineare Gleichungssysteme im $\mathbb{R}^n$ der Form

$$f(x) = 0 \qquad \text{mit} \qquad x \in D \subset \mathbb{R}^n,$$

wobei $f : D \to \mathbb{R}^n$ mindestens einmal stetig differenzierbar ist. Wir haben die äquivalente Umformung in die Fixpunktform

$$x = x - Cf(x) \qquad \text{mit invertierbarer Matrix } C$$

bereits in Abschnitt 5.1 und in den Programmieraufgaben betrachtet. Wählt man statt einer festen Matrix C nun $C = Df(x)^{-1}$, die Inverse der Jacobi-Matrix von f an der Stelle x, so erhält man das mehrdimensionale Newton-Verfahren.

5.3.1 Newton-Verfahren im $\mathbb{R}^n$
Es sei $D \subseteq \mathbb{R}^n$ ein Gebiet, $f : D \to \mathbb{R}^n$ sei stetig differenzierbar.

Zur Lösung des Gleichungssystems $f(x) = 0$ wählt man einen Startwert $x_0 \in D$ und berechnet

$$x^{(k+1)} = x^{(k)} - [Df(x^{(k)})]^{-1}\ f(x^{(k)}), \qquad k \geq 0.$$

Falls die Jacobi-Matrix $Df(x^{(k)})$ nicht invertierbar ist, bricht das Verfahren mit einer Fehlermeldung ab.

5.3.2 Algorithmus: Newton-Verfahren im $\mathbb{R}^n$

Das Newton-Verfahren im $\mathbb{R}^n$ wird meist so programmiert:

`Gegeben: Startwert` $x^{(0)} \in D$.

`Für` $k = 0, 1, 2, \ldots$:

1. `Berechne` $f(x^{(k)})$, $A^{(k)} := Df(x^{(k)})$,
   (z.B. durch Aufruf von Funktionen `f` und `Df`)

2. `löse das lineare Gleichungssystem` $A^{(k)}s^{(k)} = -f(x^{(k)})$,
   (z.B. mittels LR-Zerlegung von $A^{(k)}$)

3. `Setze` $x^{(k+1)} := x^{(k)} + s^{(k)}$,

`bis` $k \geq k_{\max}$ `oder` $\|(A^{(k)})^{-1} f(x^{(k+1)})\| \leq \text{tol}$.

In jedem Schritt ist also ein lineares Gleichungssystem zu lösen, in dem die *Newton-Korrektur* $s^{(k)}$ berechnet wird. Die Abbruchbedingung ist in Anlehnung an die erste a-posteriori Fehlerabschätzung im skalaren Fall gewählt.

Ähnlich zum eindimensionalen Fall in Satz 5.2.4 gilt die lokal quadratische Konvergenz. Wir formulieren das Ergebnis für differenzierbare Funktionen f mit Lipschitz-stetiger erster Ableitung. Die Definition der Konstanten m und M aus Satz 5.2.4 wird dazu entsprechend angepasst.

5.3.3 Satz: Lokale Konvergenz des Newton-Verfahrens (mehrdimensionaler Fall)

Es sei $D \subset \mathbb{R}^n$ ein konvexes Gebiet und $f : D \to \mathbb{R}^n$ sei stetig differenzierbar. Weiterhin besitze f eine Nullstelle $z \in D$ und die Jacobimatrix $Df(z)$ sei invertierbar.

Mit der gegebenen Norm auf $\mathbb{R}^n$ und einer verträglichen Matrixnorm gelte außerdem:

- Für jedes $x \in D$ ist $Df(x)$ invertierbar und

$$\frac{1}{m} := \sup_{x \in D} \|[Df(x)]^{-1}\| < \infty.$$

- $Df : D \to \mathbb{R}^{n \times n}$ ist Lipschitz-stetig, d.h. mit einer Konstanten $M > 0$ gilt

$$\|Df(x) - Df(y)\| \leq M\|x - y\|, \qquad x, y \in D.$$

Weiter sei $0 < r < \frac{2m}{M}$ so, dass $K_r(z) \subset D$ gilt. Dann konvergiert die Folge $(x^{(k)})_{k \geq 0}$ des Newton-Verfahrens quadratisch für jeden Startwert $x^{(0)} \in K_r(z)$ gegen z. Mit $L = \frac{Mr}{2m} < 1$ gilt

$$\|x^{(k)} - z\| \leq \frac{M}{2m}\|x^{(k-1)} - z\|^2 \leq \frac{2m}{M}\, L^{(2^k)}, \tag{5.6}$$

$$\|x^{(k)} - z\| \leq \frac{1}{m(1-L)}\, \|f(x^{(k)})\| \leq \frac{M}{2m(1-L)}\, \|x^{(k)} - x^{(k-1)}\|^2. \tag{5.7}$$

Wesentliche Beweisschritte im Vergleich mit Satz 5.2.4:

1. Die Eindeutigkeit der Nullstelle $z \in D$ folgt erst aus den anderen Beweisschritten, die Invertierbarkeit der Jacobi-Matrizen reicht noch nicht: den Mittelwertsatz (oder Satz von Rolle) für Vektorfelder gibt es so nicht.
2. entfällt ebenfalls.
3. Wie in Satz 5.1.5 ergibt sich für $x \in D$

$$\begin{aligned} 0 = f(z) &= f(x) + \int_0^1 Df(x + t(z-x))(z-x)\, dt \\ &= f(x) + Df(x)(z-x) + \int_0^1 \left(Df(x + t(z-x)) - Df(x)\right)(z-x)\, dt. \end{aligned}$$

Mit der Konstanten M aus der Lipschitz-Bedingung für Df erhalten wir

$$\|f(x) - Df(x)(x-z)\| \leq \int_0^1 Mt\|z-x\|^2\, dt = \frac{M}{2}\|x-z\|^2$$

und weiter mit der Iterationsfunktion $\phi(x) = x - (Df(x))^{-1}f(x)$ des Newton-Verfahrens

$$\|\phi(x) - z\| = \left\|(x-z) - (Df(x))^{-1}f(x)\right\| \leq \frac{M}{2m}\|x-z\|^2.$$

Wegen $\phi(x^{(k-1)}) = x^{(k)}$ ist die erste Ungleichung in (5.6) bewiesen.

3b. In Ergänzung zu Schritt 3. von Satz 5.2.4 erhalten wir für $x \in K_r(z)$

$$\|x - z\| \leq \left\|(Df(x))^{-1} f(x)\right\| + \frac{M}{2m}(x - z)^2 \leq \frac{1}{m}\|f(x)\| + \frac{Mr}{2m}\|x - z\|.$$

Mit $L = \frac{Mr}{2m} < 1$ folgt hieraus

$$(1 - L)\|x - z\| \leq \frac{1}{m}\|f(x)\|,$$

dies ist die erste Ungleichung in (5.7).

4. ist identisch zum Beweisschritt 4. von Satz 5.2.4: die Iterationsfunktion $\phi(x) = x - (Df(x))^{-1} f(x)$ ist eine Selbstabbildung von $K_r(z)$.

5. ist identisch zum Beweisschritt 5. von Satz 5.2.4: es folgt $x^{(k)} \to z$ und die zweite Ungleichung in (5.6).

6. Analog zum Beweisschritt 6. von Satz 5.2.4 folgt unter Zuhilfenahme der Lipschitz-Konstanten M wie in Schritt 1

$$\|f(x^{(k)})\| = \|f(x^{(k)}) - f(x^{(k-1)}) - Df(x^{(k-1)})(x^{(k)} - x^{(k-1)})\| \leq \frac{M}{2}(x^{(k)} - x^{(k-1)})^2.$$

Damit folgt auch die zweite Ungleichung in (5.7).

5.3.4 Beispiel:

a) `www-aix.gsi.de/~giese/swr` beschreibt als Veranschaulichung zum Einzugsgebiet: Die US-Raumsonde, Start am 20. August 1977 Flug zum Jupiter (Juli 1979), Saturn (August 1981), Uranus (Januar 1986) und Neptun (August 1989) hatte Software-Probleme kurz vor dem Uranus-Vorbeiflug: das Programm zur Flugbahnberechnung lieferte Daten, die nicht mit den gemessenen Daten übereinstimmten. Ursache: Das Programm mußte den Schätzwert der Uranusmasse fortlaufend korrigieren; die Anfangs-Schätzung war aber um 0.3% zu klein, wodurch das Programm gegen ein falsches Maximum konvergierte Nach Verstehen der komplexen Problemlage gelang Korrektur.

b) Wir betrachten $f : \mathbb{R}^2 \to \mathbb{R}^2$ mit $f(x, y) = \begin{pmatrix} x^2 + y^2 - y - 1 \\ x^2 - y^2 + x + 1 \end{pmatrix}$. Die Nullstellenmengen der Komponenten f_1 und f_2 von f sowie den Startwert $(x, y) = (1, 1)$ zeigt Bild 5.3.4. Nach 5 Iterationsschritten sind bereits 12 Stellen genau.

k	x_k	y_k	$f_1(x_k, y_k)$	$f_2(x_k, y_k)$
0	1.000000000000000	1.000000000000000	0	2.000000000000000
1	0.714285714285714	1.571428571428571	0.408163265306122	−0.244897959183673
2	0.636090225563910	1.433082706766917	0.025254112725423	−0.013025043812539
3	0.629994326387653	1.423705708975384	0.000125088074349	−0.000050768100815
4	0.629960525895331	1.423661051983030	0.000000003136721	−0.000000000851774
5	0.629960524947436	1.423661050931536	−0.000000000000001	0.000000000000000

Tabelle 5.2: Newton-Iteration zu $f(x, y) = (x^2 + y^2 - y - 1, x^2 - y^2 + x + 1)$

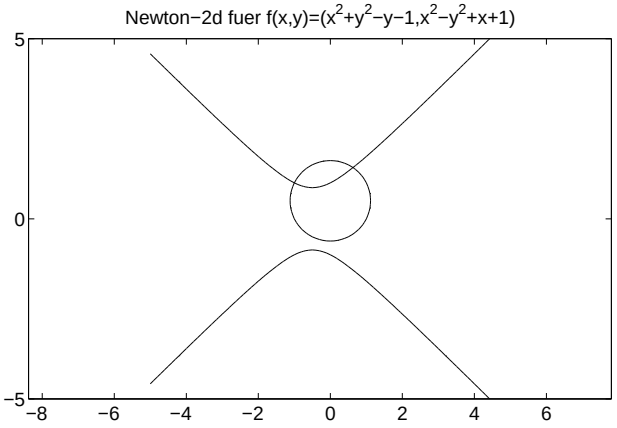

Abbildung 5.3: Nullstellenmengen der Komponenten $f_1(x, y) = x^2 + y^2 - y - 1$ (Kreis) und $f_2(x, y) = x^2 - y^2 + x + 1$ (Hyperbel)

5.3.5 Variante: Vereinfachtes Newton-Verfahren

- Beim Newton-Verfahren im $\mathbb{R}^n$ kostet in jedem Schritt sowohl das Aufstellen der Jacobi-Matrix $A^{(k)} = Df(x^{(k)})$ als auch das Lösen des linearen Gleichungssystems $A^{(k)}s^{(k)} = -f(x^{(k)})$ die meiste Zeit.

- Beim **vereinfachten Newton-Verfahren** hält man die Matrix dieses Gleichungssystems für ℓ Schritte fest. D.h., man rechnet

$$A^{(0)}s^{(k)} = -f(x^{(k)}), \quad x^{(k+1)} = x^{(k)} + s^{(k)}$$

  mit $A^{(0)} = Df(x^{(0)})$ für die ℓ Schritte mit $k = 0, 1, 2, \ldots, \ell - 1$, und stellt dazu nur einmal die LR- oder QR-Zerlegung von $A^{(0)}$ her.

  Erst dann wird $A^{(0)}$ durch $A^{(\ell)} = Df(x^{(\ell)})$ ersetzt und für die nächsten ℓ Schritte $k = \ell, \ell + 1, \ldots, 2\ell - 1$verwendet,

$$A^{(\ell)}s^{(k)} = -f(x^{(k)}), \quad x^{(k+1)} = x^{(k)} + s^{(k)}.$$

  usw.

Das vereinfachte Newton-Verfahren entspricht also in jedem Abschnitt von ℓ Schritten der Fixpunkt-Iteration

$$x^{(k+1)} = x^{(k)} - (A^{(j\ell)})^{-1} f(x^{(k)}), \qquad k = j\ell, \ldots, j\ell + \ell - 1, \quad A^{(j\ell)} = Df(x^{(j\ell)}).$$

Im j-ten Abschnitt der Länge ℓ beobachtet man oft lineare Konvergenz und die zugehörige Konstante $L_j < 1$ wird dabei immer kleiner. Dies ist dann superlineare Konvergenz.

5.3.6 Variante: Gedämpftes Newton-Verfahren

Zur "Globalisierung" der Konvergenz (d.h. auch bei grob gewähltem Startwert $x^{(0)}$) verwendet man die Newton-Korrektur

$$s^{(k)} = -[Df(x^{(k)})]^{-1} f(x^{(k)})$$

nur als **Suchrichtung**, variiert aber ihre Länge so, dass $\|f(x^{(k)})\|$ tatsächlich verkleinert wird. Der folgende Algorithmus erzielt mindestens lineare Konvergenz mit Konstante $L = 1 - \lambda_{\min}/4 < 1$ oder bricht ab.

Gegeben: Startwert $x^{(0)}$.
Für $k = 0, 1, 2, \ldots$:
1. Berechne $f(x^{(k)})$, $A^{(k)} := Df(x^{(k)})$.
2. Löse das lineare Gleichungssystem $A^{(k)} s^{(k)} = -f(x^{(k)})$,
   2.*a* Setze $\lambda = 1$.
   2.*b* **Dämpfung:**
      Setze $x := x^{(k)} + \lambda s^{(k)}$; $C_\lambda := 1 - \lambda/4$;
      Falls $\|(A^{(k)})^{-1} f(x)\| \le C_\lambda \|(A^{(k)})^{-1} f(x^{(k)})\|$, gehe zu 3.
      Sonst:
         Setze $\lambda := \lambda/2$.
         Falls $\lambda \ge \lambda_{\min}$, gehe zu 2.b.
         Sonst ABBRUCH: keine Konvergenz.
3. Setze $x^{(k+1)} := x$,
bis $k \ge k_{\max}$ oder $\|(A^{(k)})^{-1} f(x^{(k+1)})\| \le$ tol.

- Der Algorithmus hat zwei ineinander geschachtelte Schleifen, die durch die Parameter $\lambda_{\min}$ und $k_{\max}$ begrenzt werden, um Endlosschleifen zu vermeiden. Das Erreichen einer dieser Grenzen führt zum Abbruch.

- Die Voraussetzungen an den Startwert $x^{(0)}$ für die Konvergenz gegen die Nullstelle z sind wesentlich schwächer als beim Newton-Verfahren. Daher liegt oft Konvergenz vor, auch wenn $x^{(0)}$ eine grobe Näherung an z ist. Die **anfängliche lineare** Konvergenz führt in den Einzugsbereich der Nullstelle, ab dann wird in 2.b die Bedingung für $\lambda = 1$ erfüllt sein und die **quadratische** Konvergenz setzt ein.