Lothar Banz

Angewandte Mathematik

Analysis und Numerik für gewöhnliche Differentialgleichungen, Newton-Verfahren

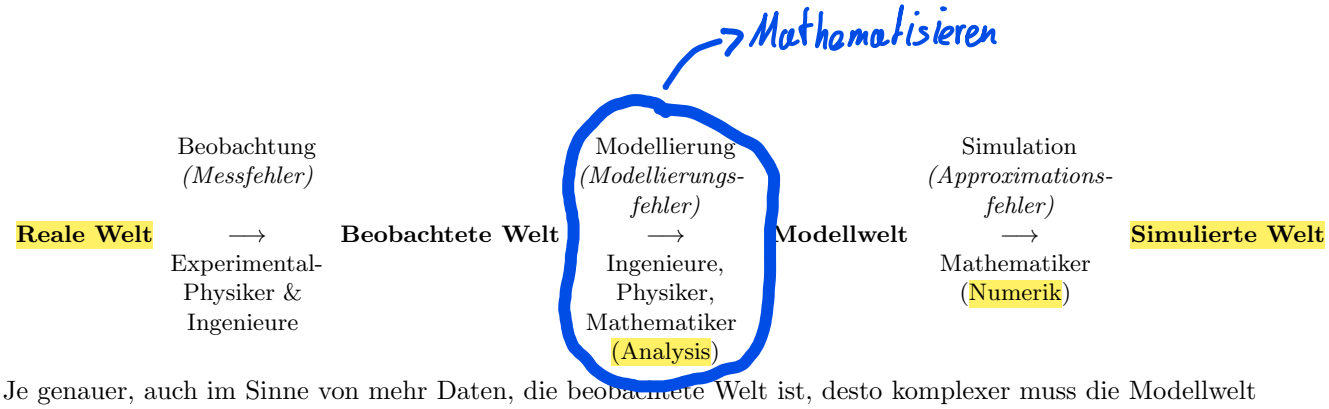
8. August 2025

Vorlesungsskript

Inhaltsverzeichnis

Einführung

wir brauchen:
- geringe Fehler (Mess.)
- genaue Modelle



Je genauer, auch im Sinne von mehr Daten, die beobachtete Welt ist, desto komplexer muss die Modellwelt werden um nicht zu viele der beobachteten Daten zu ignorieren. Dies führt in der Regel zu komplizierteren Modellen, welche in der Regel nur noch sehr schwer lösbar sind und gegebenenfalls zu einem großen Simulationsfehler führen.
Typischerweise lassen sich Veränderungen, also Ableitungen, leichter beobachten, weswegen viele Modelle gewöhnliche oder partielle Ableitungen beinhalten. Beim Lösen gilt es die Aktion des Inversens dieser Differentialoperatoren auf eine problembezogene Funktion anzuwenden.

Beispiel 1.1 (Freier Fall)
Beobachtung: Erdbeschleunigung $g \approx 9,81 m/s^2$.
Modellierung: Die Masse $m > 0$ sei in einem Punkt konzentriert und es gibt keinen Luftwiderstand also keine Reibung. Auf die Masse wirkt also nur die Schwerkraft. Das Koordinatensystem ist so gewählt, dass $y(0) = 0$ und $y(t) \geq 0$, wobei $y(t)$ die Position in Abhängigkeit von der Zeit t beschreibt. Der Körper wird aus der Ruhelage losgelassen, d.h. $\frac{d}{dt}y(0) = \dot{y}(0) = 0$. Aus dem Newton'schen Kräftegleichgewicht folgt

Geschw. = 0

$$m\frac{d^2}{dt^2}y(t) = m\ddot{y}(t) = F_S = mg.$$

Wir erhalten die gewöhnliche Differentialgleichung (DGL) mit Anfangsbedingungen

$$\ddot{y}(t) = g, \quad y(0) = 0, \ \dot{y}(0) = 0.$$

Simulation: Zum Lösen dieser DGL (siehe Kapitel 2.3) sei $v(t) = \dot{y}(t)$. Folglich ist $\dot{v}(t) = g$. Integrieren liefert $v(t) = gt + C_1$ mit beliebiger Integrationskonstante $C_1 \in \mathbb{R}$. Aus der Anfangsbedingung folgt $\dot{y}(0) = v(0) = 0 = g \cdot 0 + C_1 = C_1$. Integrieren der Substitutiongleichung $gt = v(t) = \dot{y}(t)$ liefert $y(t) = \frac{1}{2}gt^2 + C_2$ mit $C_2 \in \mathbb{R}$. Die zweite Anfangsbedingung ergibt $y(0) = 0 = \frac{1}{2}g \cdot 0^2 + C_2 = C_2$. Somit ist $y(t) = \frac{1}{2}gt^2$ die Simulation eines frei fallenden Objektes.

$= \int gt \, dt$

Beispiel 1.2 (Mathematisches Pendel)
Beobachtung: Erdbeschleunigung $g \approx 9,81 m/s^2$.
Modellierung: Die Masse $m > 0$ sei in einem Punkt konzentriert. Der Faden mit Länge l sei masselos und es gibt keine Reibung bei der Aufhängung und mit der Luft. Auf den Massepunkt wirkt nur die Schwerkraft. Das Superpositionierungsprinzip liefert: Schwerkraft $F_G = F_R + F_S$, siehe Bild 1.1. Die Rückstellkraft F_R ist dabei tangential zur Kreisbahn. Die Fadenspannkraft F_S ist orthogonal zu F_R und die Schwerkraft wirkt nach unten. Also ist

$$F_G = \begin{pmatrix} 0 \\ -mg \end{pmatrix}.$$

Schwerkraft nur nach unten

Das Kräftegleichgewicht lässt sich in einer Formel schreiben als

$$m \cdot a_{tan}(t) = -mg \cdot \sin(\phi(t))$$

mit Tangentialbeschleunigung $a_{tan}(t)$ und Winkel $\phi(t)$. Für die Winkelbeschleunigung gilt $a_{tan}(t) = l \cdot \ddot{\phi}(t)$. Somit erhalten wir die Differentialgleichung

$$ml\ddot{\phi}(t) = -mg\sin(\phi(t)) \quad \Leftrightarrow \quad \ddot{\phi}(t) + \frac{g}{l}\sin(\phi(t)) = 0.$$

Die Anfangsauslenkung und -geschwindigkeit seien $\phi(0) = \phi_0$ und $\dot{\phi}(0) = \phi_1$. Wegen dem $\sin(\phi(t))$ können wir zur Zeit obiges Problem nicht exakt und explizit lösen. Wir müssten auf numerische Verfahren zum Lösen zurückgreifen oder die Modellierung anpassen. Wenn wir also zusätzlich annehmen, dass ϕ klein ist, so ist $\sin(\phi) \approx \phi$ und das Problem wird dadurch Linearisiert:

$$\ddot{\phi}(t) + \frac{g}{l}\phi(t) = 0 \quad \text{mit } \phi(0) = \phi_0 \text{ und } \dot{\phi} = \phi_1.$$

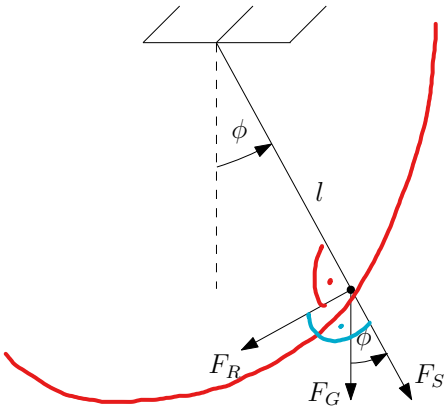


Abb. 1.1: Skizze des mathematischen Pendels.

Beispiel 1.3 (Volterra Modell)

Sei die Population von Jägern $J(t) \geq 0$. Ebenso sei die Population von Beute $B(t) \geq 0$. B vermehrt sich mit konstanter Rate $\beta \geq 0$ und wird von J gejagt, so dass

mit $\alpha \geq 0$. Für J soll analog gelten

$$B'(t) = \beta B(t) - \alpha B(t)J(t)$$
$$J'(t) = -\gamma J(t) + \alpha B(t)J(t)$$

mit $\gamma \geq 0$. Sei die Anfangspopulation $J(0) = J_0$ und $B(0) = B_0$, so erhalten wir das System von Differentialgleichungen

$$\frac{d}{dt}\begin{pmatrix} J(t) \\ B(t) \end{pmatrix} = \begin{pmatrix} J \\ B \end{pmatrix}'(t) = \begin{pmatrix} -\gamma J(t) + \alpha B(t)J(t) \\ \beta B(t) - \alpha B(t)J(t) \end{pmatrix}, \quad \begin{pmatrix} J \\ B \end{pmatrix}(0) = \begin{pmatrix} J_0 \\ B_0 \end{pmatrix}.$$

2

Kochrezepte für elementare DGLs

Wichtig für Prüfung!

Dieses Kapitel ist dem exakten Lösen von gewöhnlichen Differentialgleichungen (DGLs), engl. ordinary differential equations (ODEs), gewidmet.

2.1 Grundlegende Definitionen

Definition 2.1
Sei $\emptyset \neq D \subset \mathbb{R}^{n+1}$ offen und $f : D \to \mathbb{R}$ stetig. Dann heißt

$$y^{(n)} := \frac{d^n y}{dx^n} = f(x, y, y', y'', \ldots, y^{(n-1)}) \qquad (2.1)$$

$y^n = \ldots\ldots$ Form

explizite gewöhnliche Differentialgleichung n-ter Ordnung. Allgemeine Differentialgleichungen der Form

nur 1 Variable

$$F(x, y, y', \ldots, y^{(n)}) = 0 \qquad \ldots\ldots = 0 \quad \text{Form}$$

heißen implizit.

Definition 2.2
Sei $I \subset \mathbb{R}$ ein echtes Intervall, d.h. ein Intervall mit mindestens zwei Punkten. Eine n-mal differenzierbare Funktion $\phi : I \to \mathbb{R}$ heißt Lösung von (2.1), wenn für alle $x \in I$ gilt

1. $(x, \phi(x), \phi'(x), \ldots, \phi^{(n-1)}(x)) \in D$, → alle Werte im Def. Bereich
2. $f(x, \phi(x), \phi'(x), \ldots, \phi^{(n-1)}(x)) = \phi^{(n)}(x)$. → Eine Kombination (ggf. mit Konstanten) löst die expl. DGL.

Definition 2.3
Sei $(\xi, \eta_1, \eta_2, \ldots, \eta_n) \in D$. Dann heißt

$$y(\xi) = \eta_1, \; y'(\xi) = \eta_2, \; \ldots, \; y^{(n-1)}(\xi) = \eta_n \qquad (2.2)$$

Anfangsbedingung für (2.1). Beides zusammen heißt Anfangswertproblem (AWP).

Definition 2.4
Ist $\phi : I \to \mathbb{R}$ eine Lösung von (2.1) und gilt

$\phi()$ erfüllt alle Anfangswerte

$$\xi \in I \quad \text{sowie} \quad \phi(\xi) = \eta_1, \; \ldots, \; \phi^{(n-1)}(\xi) = \eta_n. \qquad (2.3)$$

So ist ϕ eine Lösung des Anfangswertproblems (2.1)-(2.2).

Definition 2.5
Das Anfangswertproblem (2.1)-(2.2) heißt lösbar, wenn es ein echtes Intervall $I \subset \mathbb{R}$ mit $\xi \in I$ und eine Lösung $\phi : I \to \mathbb{R}$ des AWP gibt.

DGL n-ter Ordnung sind in einem gewissen Sinne äquivalent zu Systemen von DGL 1. Ordnung.

Definition 2.6
Sei $\emptyset \neq D \subset \mathbb{R}^{n+1}$ offen und $\vec{f} : D \to \mathbb{R}^n$ stetig, dann heißt

$$\begin{pmatrix} y_1' \\ \vdots \\ y_n' \end{pmatrix} = \vec{y}' = \vec{f}(x, \vec{y}) = \vec{f}(x, y_1, y_2, \dots, y_n) = \begin{pmatrix} f_1(x, y_1(x), \dots y_n(x)) \\ f_2(\dots) \\ \vdots \\ f_n(x, y_1(x), \dots, y_n(x)) \end{pmatrix} \quad (2.4)$$

wie Beute-Jäger-System

ein explizites System von n Differentialgleichungen 1. Ordnung oder auch System n-ter Ordnung.

Definition 2.7
Ist $(\xi, \vec{\eta}) = (\xi, \eta_1, \dots, \eta_n) \in D$ so heißt

$$\vec{y}(\xi) = \vec{\eta} \quad \text{bzw.} \quad y_1(\xi) = \eta_1, \dots, y_n(\xi) = \eta_n \quad (2.5)$$

z.B. $\xi = 0$, $\begin{pmatrix} y_1(0) \\ \vdots \\ y_n(0) \end{pmatrix} = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix}$

Anfangswert für das System (2.4) und (2.4)-(2.5) heißt ebenfalls Anfangswertproblem.

Definition 2.8
Sei $I \subset \mathbb{R}$ ein echtes Intervall und $\vec{\phi} : I \to \mathbb{R}^n$ eine differenzierbare Funktion, dann heißt $\vec{\phi}$ Lösung von (2.4), falls für alle $x \in I$ gilt

1. $(x, \vec{\phi}(x)) \in D$
2. $\vec{\phi}'(x) = \vec{f}(x, \vec{\phi})$

→ Achtung: nicht mischen mit DGL nter Ordnung

Erfüllt $\vec{\phi}$ auch (2.5), so ist es eine Lösung des AWPs.

Satz 2.9
Sei die explizite DGL n-ter Ordnung

$$y^{(n)} = f(x, y, y', \dots, y^{(n-1)}) \quad (2.6)$$

gegeben. Dabei sei $\emptyset \neq D \subset \mathbb{R}^{n+1}$ offen und $f : D \to \mathbb{R}$ stetig. Mit

$$y_1 := y, \quad y_2 := y', \quad \dots, \quad y_n := y^{(n-1)}$$

einsetzen

$y_1' = y' = y_2$

geht (2.6) über in das System

$$\vec{y}' = \begin{pmatrix} y_1' \\ y_2' \\ \vdots \\ y_{n-1}' \\ y_n' \end{pmatrix} = \begin{pmatrix} y_2 \\ y_3 \\ \vdots \\ y_n \\ f(x, y_1, y_2, \dots, y_n) \end{pmatrix} = \begin{pmatrix} y_2 \\ y_3 \\ \vdots \\ y_n \\ f(x, \vec{y}) \end{pmatrix}. \quad (2.7)$$

Substituieren, damit aus Ordnung n → Ordnung 1, dafür statt 1 GL. → n GL.

$(y^{(n-1)})' = y^{(n)}$

Ist $(\xi, \vec{\eta}) = (\xi, \eta_1, \dots, \eta_n) \in D$, so sind

$$y(\xi) = \eta_1, \ \dots, \ y^{(n-1)}(\xi) = \eta_n \quad \text{und} \quad \vec{y}(\xi) = \vec{\eta}$$

Anfangsbedingungen von (2.6) bzw. (2.7).
Die Gleichungen (2.6) und (2.7) sind äquivalent im folgenden Sinne:
Ist ϕ Lösung von (2.6), so ist

$$\vec{\phi} = (\phi, \phi', \dots, \phi^{(n-1)})^\top$$

Lösung von (2.7).
Ist umgekehrt

$$\vec{\phi} = (\phi_1, \dots, \phi_n)^\top$$

Lösung von (2.7), so ist $\phi = \phi_1$ Lösung von (2.6).

Teilen sich Lösung!

⇒ Aussuchen, was besser zu lösen geht.

Beweis. Hausübung. □

Beispiel 2.10
Die DGL 2. Ordnung

$$y'' + y = 0 \quad \Leftrightarrow \quad y'' = -y + 0$$

mit den Anfangsbedingungen $y(0) = 0$, $y'(0) = 1$ ist äquivalent zu dem System

$$\begin{pmatrix} y_1' \\ y_2' \end{pmatrix} = \begin{pmatrix} y_2 \\ -y_1 \end{pmatrix} \quad \text{mit} \quad \begin{pmatrix} y_1(0) \\ y_2(0) \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

y'

wie oben:
$y_1' = y' = y_2$
$y_2' = (y')' = y''$
$y'' + y = 0 \Rightarrow y'' = -y = -y_1$

Die Funktionen $y(x) = \sin(x)$ und

$$\vec{y} = \begin{pmatrix} \sin(x) \\ \cos(x) \end{pmatrix}$$

sind die eindeutigen Lösungen dieser Anfangswertprobleme.

2.2 Explizite Gleichungen 1. Ordnung

Sei $\emptyset \neq D \subset \mathbb{R}^2$ offen, $(x_0, y_0) \in D$ und $f : D \to \mathbb{R}$ stetig. Wir betrachten im Folgenden

$$y' = f(x, y) \text{ ggf. mit der Anfangsbedingung } y(x_0) = y_0. \tag{2.8}$$

2.2.1 Gleichung mit getrennten Variablen

Sei

$$\frac{dy}{dx} = y' = f(x) \cdot g(y)$$

mit $f : I \to \mathbb{R}$ und $g : J \to \mathbb{R}$ stetig für $I, J \subset \mathbb{R}$.

Form muss nicht immer möglich sein

Kochrezept 2.11 (Trennung der Variablen, TdV)

wenn $f(x) \neq 0$ und $g(y) = 0$
dann $y(x) = y_0$ Konstante Lösung

1. Auf stationäre Lösungen $\phi(x) \equiv y_0$ mit $g(y_0) = 0$ überprüfen.
2. Variablen trennen
$$\frac{dy}{g(y)} = f(x)dx.$$
3. Integrieren
$$G(y) := \int \frac{dy}{g(y)} = \int f(x)\,dx =: F(x) + c.$$
4. Bei AWP die Integrationskonstante $c = G(y_0) - F(x_0)$ an die Anfangsbedingung anpassen.
5. Eventuell die implizite Gleichung nach y oder nach x auflösen.

Beispiel 2.12
Sei

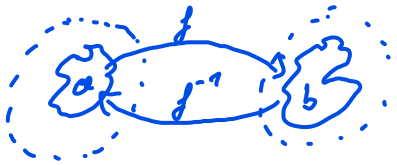
$$y'(x) = -\frac{y}{x} \quad \text{und} \quad y(2) = 1.$$

1. Die einzige mögliche stationäre Lösung ist $y(x) \equiv 0$. Aber diese steht im Widerspruch zu $y(2) = 1$.
2. Trennung der Variablen
$$\frac{dy}{y} = -\frac{1}{x}dx$$
3. Integrieren
$$\int \frac{dy}{y} = \int -\frac{1}{x}\,dx \quad \Leftrightarrow \quad \ln(|y|) = -\ln(|x|) + c$$
4. Integrationskonstante anpassen
$$c = \ln(|y_0|) + \ln(|x_0|) = \ln(1) + \ln(2) = \ln(2)$$
5. Nach y auflösen
$$y = \pm \exp(\ln(2)) \cdot \frac{1}{|x|} = \frac{2}{x}, \quad \Leftarrow \ln(|y|) = -\ln(|x|) + \ln(2)$$
da x und y nach Anfangsbedingung positiv sind.

Gradient als Jacobi Matrix

Satz 2.13 (Satz der lokalen Umkehrbarkeit)
Sei $D \subset \mathbb{R}^n$ offen, $a \in D$, $b := f(a)$ und $f : D \to \mathbb{R}^n$ stetig differenzierbar mit $\nabla f(a)$ invertierbar. Dann gibt es offene Umgebungen $U = U(a) \subset D \subset \mathbb{R}^n$ um a und $V = V(b) \subset \mathbb{R}^n$ um b, so dass $f : U \to V$ bijektiv ist. Die Umkehrfunktion $f^{-1} : V \to U$ ist stetig differenzierbar mit

Gesucht $(f^{-1})'$: $(\nabla f^{-1}) \circ f(x) = (\nabla f(x))^{-1}$. $(\cdot)^{-1}$



Ist f k-mal stetig differenzierbar, so ist auch f^{-1} k-mal stetig differenzierbar.

Satz 2.14 (Eindeutigkeitssatz für Gleichungen mit getrennten Variablen)
Ist $(x_0, y_0) \in I \times J$ mit $I, J \subset \mathbb{R}$ offene Intervalle und $g(y_0) \neq 0$, so ist das AWP

$$y'(x) = f(x)g(y) \quad \text{mit} \quad y(x_0) = y_0 \tag{2.9}$$

und $f : I \to \mathbb{R}$, $g : J \to \mathbb{R}$ stetig eindeutig lösbar.

Beweis. "Existenz": $G(y) := \int_{y_0}^{y} \frac{1}{g(t)}\,dt$ ist eine Stammfunktion von $\frac{1}{g}$ mit $G(y_0) = 0$. G ist wegen $G'(y_0) = \frac{1}{g(y_0)} \neq 0$ und der Stetigkeit von $G' = \frac{1}{g}$ in einer Umgebung von y_0 umkehrbar. Die Funktion f ist nach Voraussetzung ebenfalls stetig. Also ist $F(x) = \int_{x_0}^{x} f(t)\,dt$ eine Stammfunktion von f mit $F(x_0) = 0$. Wegen $G(y_0) = 0$ ist $G^{-1}(0) = y_0$ und eine Lösung des AWP ist gegeben durch

$$y = \phi(x) := G^{-1}(F(x)),$$

denn es gilt

1. $\phi(x_0) = y_0$,
2. ϕ ist in einer Umgebung von x_0 definiert und ist differenzierbar nach dem Satz der lokalen Umkehrbarkeit,
3. $\phi'(x) = (G^{-1})'(F(x)) \cdot F'(x) = \frac{1}{G'(G^{-1}(F(x)))} f(x) = g(\phi)f(x)$.

"Eindeutigkeit": Sei $\psi(x)$ irgendeine Lösung des AWP, d.h. $\psi'(x) = f(x)g(\psi(x))$ in einer Umgebung von x_0. Mit der Substitution $u := \psi(t)$ folgt

$$G(\psi) = \int_{y_0}^{\psi} \frac{1}{g(u)}\,du = \int_{x_0}^{x} \frac{\psi'(t)}{g(\psi(t))}\,dt = \int_{x_0}^{x} f(t)\,dt = F(x) \quad \Rightarrow \quad \psi = G^{-1}(F(x)).$$

Also stimmt jede Lösung mit der konstruierten Lösung aus dem Existenzteil des Beweises überein. □

2.2.2 DGLs vom Typ: $y' = f(ax + by + c)$

Sei $y' = f(ax + by + c)$. Für $b = 0$ ist dies eine Gleichung mit getrennten Variablen. Sei deshalb $b \neq 0$. Die Substitution $z = ax + by + c$ führt das Problem auf eine Gleichung mit getrennten Variablen zurück.

Lemma 2.15
Seien $a, b, c \in \mathbb{R}$, $b \neq 0$ und $f : I \to \mathbb{R}$ stetig in $I \subset \mathbb{R}$. Dann gilt: Ist $\phi(x)$ eine Lösung der DGL $y' = f(ax + by + c)$, so ist $\psi(x) := ax + b\phi(x) + c$ eine Lösung der DGL $z' = a + bf(z)$ vom Typ Trennung der Variablen und umgekehrt.

Beweisskizze. "⇒: $\psi' = a + b\phi'(x) = a + bf(\psi)$
"⇐: $\phi = \frac{1}{b}(\psi - ax - c) \Rightarrow \phi' = \frac{1}{b}(\psi' - a) = \frac{1}{b}(a + bf(\psi) - a) = f(ax + b\phi + c)$ □

Kochrezept 2.16 (für $y' = f(ax + by + c)$)

1. Substituiere
$$z = ax + by + c.$$
2. Dies liefert die DGL
$$z' = a + by' = \Big(a + bf(z)\Big) \cdot 1$$
vom Typ Trennung der Variablen.
3. Die Lösungen sind implizit in
$$x - x_0 = \int \frac{dz}{a + bf(z)}$$
enthalten.

Beispiel 2.17
Sei
$$y' = (x + y)^2,$$
d.h. $a = b = 1$, $c = 0$. Substitution liefert

$$z = x + y \quad \Rightarrow \quad z' = 1 + y' = 1 + f(z) = 1 + z^2.$$

Das Trennung der Variablen Kochrezept ergibt

$$\arctan(z) = \int \frac{1}{1+z^2}\,dz = \int 1\,dx = x + c \quad \Rightarrow \quad z = \tan(x+c).$$

Rücksubstitution liefert

$$y = z - x = \tan(x+c) - x.$$

2.2.3 Ähnlichkeits-DGLs

Definition 2.18
Sei $f : I \to \mathbb{R}$ stetig in $I \subset \mathbb{R} \setminus \{0\}$. Dann heißt

$$y' = f\left(\frac{y}{x}\right) \tag{2.10}$$

eine homogene oder Ähnlichkeits-DGL.

Analog zum vorherigen Abschnitt führen wir solche DGLs mit der Substitution $z = \frac{y}{x}$ auf eine Gleichung mit getrennten Variablen zurück.

Lemma 2.19
Ist $y(x)$ eine Lösung der homogenen DGL (2.10), so ist $z(x) := \frac{y(x)}{x}$ eine Lösung von

$$xz' + z = f(z) \quad \text{bzw.} \quad z' = \frac{f(z) - z}{x} \tag{2.11}$$

und umgekehrt.

Kochrezept 2.20 (für $y' = f(\frac{y}{x})$)

1. Substituiere
$$y = zx.$$
2. Dies liefert die Gleichung
$$y' = xz' + z = f(z).$$
3. Lösen dieser Gleichung mit dem TdV-Kochrezept liefert
$$\int \frac{dz}{f(z) - z} = \int \frac{dx}{x}.$$

4. Rücksubstituieren

Beispiel 2.21
Sei

$$y' = \frac{y+x}{y-x} = \frac{\frac{y}{x}+1}{\frac{y}{x}-1}.$$

Die Substitution $y = xz$ liefert

$$xz' = y' - z = \frac{z+1}{z-1} - \frac{z(z-1)}{z-1} = -\frac{z^2-2z-1}{z-1}.$$

Das Trennung der Variablen Kochrezept ergibt

$$\begin{aligned}
&\int \frac{z-1}{z^2-2z-1}\,dz = -\int \frac{dx}{x} \\
\Leftrightarrow \quad &\frac{1}{2}\int \frac{2z-2}{z^2-2z-1}\,dz = \frac{1}{2}\ln(|z^2-2z-1|) = -\ln(|x|) + c \\
\Leftrightarrow \quad &\ln(|z^2-2z-1|) = \ln\left(\frac{1}{|x|^2}\right) + \hat{c} \\
\Leftrightarrow \quad &|z^2-2z-1| = \frac{1}{x^2}\cdot\tilde{c} \\
\Leftrightarrow \quad &z^2-2z-1 = \bar{c}\cdot\frac{1}{x^2}.
\end{aligned}$$

Rücksubstitution liefert

$$y^2 - 2xy - x^2 = z^2x^2 - 2zx^2 - x^2 = \bar{c}.$$

2.2.4 Lineare Differentialgleichungen 1. Ordnung

Definition 2.22
Sei $I \subset \mathbb{R}$ ein Intervall, $x_0 \in I$, $y_0 \in \mathbb{R}$ und $a, b : I \to \mathbb{R}$ stetig. Eine DGL der Form

$$y' = a(x)y + b(x) \tag{2.12}$$

heißt lineare DGL 1. Ordnung und

$$y' = a(x)y + b(x), \quad y(x_0) = y_0 \tag{2.13}$$

heißt lineares AWP 1. Ordnung. Die Funktion $b(x)$ heißt Störglied oder Inhomogenität. Die Gleichung heißt homogen, falls $b \equiv 0$ ansonsten inhomogen.

$$y' = a(x)y \tag{2.14}$$

heißt die zu (2.12) gehörige homogene Gleichung.

Satz 2.23 (Struktursatz für lineare Gleichungen 1. Ordnung)
Alle Lösungen der inhomogenen DGL (2.12) haben die Form

$$y(x) = y_s(x) + y_h(x) = y_s(x) + cy_1(x)$$

mit einem beliebigen $c \in \mathbb{R}$. Dabei ist y_s eine spezielle Lösung der inhomogenen DGL (2.12) sowie y_h eine beliebige und y_1 eine Basislösung der zugehörigen homogenen DGL (2.14). Die Lösungsschar der inhomogenen Gleichung bildet also einen 1-dimensionalen affinen Funktionenraum. Die Lösungsschar der homogenen linearen Gleichung 1. Ordnung bildet einen 1-dimensionalen Vektorraum.

Beweisskizze. Die Funktion $y = y_s + cy_1$ mit $c \in \mathbb{R}$ ist eine Lösung von (2.12), denn

$$y' = y_s' + cy_1' = a(x)y_s + b(x) + ca(x)y_1 = a(x)(y_s + cy_1) + b(x) = a(x)y + b(x).$$

Seien $y, \tilde{y}$ beides Lösungen von (2.12). Dann gilt für die Differenz $z = y - \tilde{y}$

$$z' = y' - \tilde{y}' = a(x)y + b(x) - a(x)\tilde{y} - b(x) = a(x)z.$$

Die Differenz zweier Lösungen ist gerade eine Lösung von (2.14) also ist $z = y - \tilde{y} = cy_1$. □

Zum Lösen von (2.12) mit Hilfe des Struktursatzes 2.23 ist zuerst mit Trennung der Variablen die Basislösung

$$y_1 = \exp\left(\int_{x_0}^{x} a(t)\,dt\right)$$

zu berechnen. Für die spezielle Lösung y_s machen wir jetzt den Ansatz $y_s(x) = c(x)y_1(x)$ mit einer unbekannten Funktion $c(x)$. Damit y_s eine spezielle Lösung ist muss gelten

$$y_s' = c'(x)y_1 + c(x)y_1' = c'(x)y_1 + c(x)a(x)y_1(x) \overset{!}{=} a(x)y_s(x) + b(x) = a(x)c(x)y_1(x) + b(x).$$

Produktregel — weil L. und r. beides Lsg.

Folglich hat

$$c'(x)y_1(x) = b(x) \quad \Rightarrow \quad c(x) = \int_{x_0}^{x} \frac{b(t)}{y_1(t)}\,dt$$

zu gelten. Alles ineinander einsetzten liefert die Lösung

$$y(x) = \left(y_0 + \int_{x_0}^{x} b(t) \cdot \exp\left(-\int_{x_0}^{t} a(s)\,ds\right) dt\right) \cdot \exp\left(\int_{x_0}^{x} a(t)\,dt\right)$$

für (2.12) bzw. (2.13). Diese Argumentation führt auf folgendes Kochrezept.

Kochrezept 2.24 (Variation der Konstanten, VdK)

1. Bestimme eine Stammfunktion
$$A(x) = \int a(x)\,dx.$$
2. $y_1 := e^{A(x)}$ ist eine Basislösung der homogenen Gleichung $y' = a(x)y$.
3. Bestimme eine spezielle Lösung y_s der inhomogenen Gleichung durch den Ansatz
$$y_s = c(x)y_1(x).$$
4. Für $c(x)$ erhalten wir die direkt zu integrierende Gleichung
$$c'(x)y_1(x) = b(x) \quad \Leftrightarrow \quad c'(x) = \frac{b(x)}{y_1(x)} \quad \Leftrightarrow \quad c(x) = \int \frac{b(x)}{y_1(x)}\,dx.$$
5. Integrieren und Einsetzen liefert y_s.
6. Die allgemeine Lösung der inhomogenen DGL ist
$$y = y_s + \tilde{c}y_1$$
mit $\tilde{c} \in \mathbb{R}$.
7. Gegebenenfalls $\tilde{c}$ an die Anfangsbedingung anpassen.

Beispiel 2.25
Sei
$$y' = 2y + e^x \quad \text{mit} \quad y(3) = -2.$$
Es ist $a(x) = 2$ und $b(x) = e^x$. Ausführen des Kochrezepts VdK erfolgt in folgenden Schritten

1. $A(x) = \int 2\,dx = 2x$
2. $y_1(x) = e^{2x}$
3. $y_s(x) = c(x)e^{2x}$
4. $c'(x) = e^x e^{-2x} = e^{-x}$
5. $c(x) = \int e^{-x}\,dx = -e^{-x} \Rightarrow y_s(x) = -e^{-x}e^{2x} = -e^x$
6. $y(x) = -e^x + \tilde{c}e^{2x}$
7. $y(3) = -2 \Rightarrow -2 = -e^3 + \tilde{c}e^6 \Leftrightarrow \tilde{c} = \frac{-2+e^3}{e^6} = e^{-3} - 2e^{-6}$

2.2.5 Bernoulli- und Ricatti-DGL

Definition 2.26
Sei $\alpha \in \mathbb{R}$, $I \subset \mathbb{R}$ offen und $a, b : I \to \mathbb{R}$ stetig. Dann heißt
$$y' = a(x)y + b(x)y^\alpha \tag{2.15}$$
Bernoulli-DGL mit Exponent α.

Für $\alpha = 0$ und $\alpha = 1$ ist dies eine lineare DGL 1. Ordnung. Für $a \equiv 0$ oder $b \equiv 0$ ist dies eine DGL vom Typ getrennter Variablen.

Kochrezept 2.27 (für $y' = a(x)y + b(x)y^\alpha$, Bernoulli-DGL)

1. Multiplikation von (2.15) mit $y^{-\alpha}$ liefert
$$y'y^{-\alpha} = a(x)y^{1-\alpha} + b(x).$$
2. Die Substitution $z = y^{1-\alpha}$, also $y = z^{1/(1-\alpha)}$, liefert
$$z' = (1-\alpha)y^{-\alpha}y'.$$
3. Dies liefert die mit VdK zu lösende lineare DGL 1. Ordnung
$$z' = (1-\alpha)a(x)z + (1-\alpha)b(x). \tag{2.16}$$

4. Rücksubstitution ergibt y.

Bemerkung 2.28 1. Ist $y : I_0 \to \mathbb{R}_{>0}$ eine positive Lösung von (2.15) auf dem Teilintervall $I_0 \subset I$, so ist $z = y^{1-\alpha}$ eine positive Lösung der linearen DGL (2.16) und umgekehrt.

2. Ist $\alpha \notin \mathbb{Z}$, so ist y^α nur für $y > 0$ definiert. Also können auch nur positive Funktionen y Lösungen von (2.15) sein.

3. Ist $\alpha \geq 0$, so ist $y \equiv 0$ eine spezielle Lösung. Es kann sein, dass weitere Lösungen in die x-Achse einmünden z.B. bei $y' = \sqrt{|y|}$.

4. Ist $\alpha \in \mathbb{Z}$, so kann es auch negative Lösungen für (2.15) geben.

5. Ist $\alpha \in \mathbb{Z}$ ungerade, so ist mit y auch $-y$ eine Lösung von (2.15). Aus den positiven Lösungen $z(x)$ von (2.16) erhalten wir bis auf $y \equiv 0$ alle Lösungen von (2.15) in der Form $y = \pm z^{1/(1-\alpha)}$.

6. Ist $\alpha \in \mathbb{Z}$ gerade und y eine negative Lösung von (2.15), so ist $u(x) = -y(x)$ eine Lösung der geänderten Bernoulli-DGL $y' = a(x)y - b(x)y^\alpha$. Gegenüber (2.15) wurde hier $b(x)$ durch $-b(x)$ ersetzt. Also ist $z = -u^{1-\alpha}$ eine Lösung der ursprünglichen linearen DGL (2.16). Aus der Lösung $z(x)$ der Gleichung (2.16) erhalten wir, bis auf $y \equiv 0$, alle Lösungen von (2.15) in der Form $y = \operatorname{sgn}(z) \cdot |z(x)|^{1/(1-\alpha)}$.

Beispiel 2.29
Sei

$$y' + xy = xy^3 \quad \Leftrightarrow \quad y' = -xy + xy^3.$$

Es ist $\alpha = 3$. $y \equiv 0$ und $y \equiv \pm 1$ sind offensichtliche Lösungen. Multiplizieren mit y^{-3} liefert

$$y'y^{-3} + xy^{-2} = x.$$

Die Substitution $z = y^{-2}$ liefert jetzt

$$z' = -2y^{-3}y' = -\frac{2}{y^3}(-xy + xy^3) \quad \Leftrightarrow \quad z' - 2xz = -2x.$$

Die homogene Gleichung $z' = 2xz$ hat die allgemeine Lösung $z_h = ce^{x^2}$ mit $c \in \mathbb{R}$ beliebig. Wie leicht gesehen werden kann ist $z_s \equiv 1$ eine spezielle Lösung. Somit ist $z = 1 + ce^{x^2}$ und $y = \pm 1/\sqrt{1 + ce^{x^2}}$.

Definition 2.30
Seien $a, b, c : I \to \mathbb{R}$ mit $I \subset \mathbb{R}$. Dann heißt

$$y' = a(x) + b(x)y + c(x)y^2 \tag{2.17}$$

Ricatti-Gleichung.

Für $c \equiv 0$ ist dies eine lineare DGL 1. Ordnung und für $a \equiv 0$ eine Bernoulli-DGL.

Kochrezept 2.31 (für $y' = a(x) + b(x)y + c(x)y^2$, Ricatti-DGL)

1. Versuche durch sinnvolle Rate-Ansätze eine spezielle Lösung y_1 von (2.17) zu bestimmen.
2. Die Substitution $z = (y - y_1)^{-1}$ liefert für z die lineare DGL

$$z' = -(b(x) + 2c(x)y_1)z - c(x).$$

3. Löse diese lineare DGL.
4. Eventuell die Anfangsbedingungen einarbeiten.
5. $y = \frac{1}{z} + y_1$ und y_1 sind Lösungen der Ricatti-DGL (2.17).

Beispiel 2.32
Sei

$$y' + (2x + 1)y - y^2 = 1 + x + x^2 \quad \Leftrightarrow \quad y' = (x - y) + (x - y)^2 + 1.$$

Offensichtlich ist $y_1 = x$ eine spezielle Lösung. Die Substitution $z = \frac{1}{y-x}$, mit $\frac{-z'}{z} = y' - 1$ liefert für z die lineare DGL

$$z' = z - 1.$$

Mit dem Kochrezept Variation der Konstanten folgt

$$z = 1 + Ce^x$$

mit $C \in \mathbb{R}$ beliebig. Somit hat die ursprüngliche Ricatti-Gleichung die Lösungen

$$y = x + \frac{1}{1 + Ce^x} \quad \text{und} \quad y = x.$$

2.3 DGL höherer Ordnung

2.3.1 DGLs vom Typ: y kommt nicht vor

Sei

$$y^{(n)} = f(x, y', y'', \ldots, y^{(n-1)}).$$

Durch die Substitution $z = y'$ lässt sich die Ordnung der DGL um eins reduzieren. So ist

$$z^{(n-1)} = f(x, z, z', \ldots, z^{(n-2)}).$$

Beispiel 2.33
Sei

$$y'' = 2xy' \qquad \Leftrightarrow \qquad z = y' \quad \text{und} \quad z' = 2xz.$$

Dafür ist $z = Ce^{(x^2)}$ mit $C \in \mathbb{R}$ und somit $y = C \int e^{(x^2)}\, dx + \hat{C}$.

2.3.2 DGLs vom Typ: x kommt nicht vor

Sei

$$y'' = f(y, y').$$

Differentialgleichungen dieser Art heißen autonom (selbstständig, unabhängig), da die unabhängige Variable x nicht auftritt.

Kochrezept 2.34 (für $y'' = f(y, y')$, autonome DGL)

1. Lösungen von $f(\eta, 0) = 0$ liefern die stationäre Lösungen $y \equiv \eta$ (Ruhelage).
2. Mit der Substitution

$$y'(x) = p(y) = p(y(x)) \quad \Rightarrow \quad y'' = p'(y) \cdot y' = p'(y) \cdot p = p'p.$$

3. Wir erhalten für $p(y)$ die DGL 1. Ordnung

$$pp' = f(y, y') = f(y, p).$$

4. Bestimme die allgemeine Lösung $p = p(y)$ dieser Gleichung (1. Integration).
5. Eventuell die Anfangsbedingung $y'(x_0) = p(y_0) = v_0$ einarbeiten.
6. Löse $y' = p(y)$ mit Trennung der Variablen (2. Integration)

$$x = \int \frac{dy}{p(y)}.$$

7. Eventuell die Anfangsbedingung $y(x_0) = y_0$ einarbeiten.

Beispiel 2.35
Sei

$$5yy'' + (y')^2 = 0, \quad y(0) = y'(0) = 1.$$

So ist

$$y'' = \frac{-(y')^2}{5y} = f(y, y').$$

Wegen $y'(0) = 1$ gibt es keine stationäre Lösung des AWPs. Die Substitution $y' = p(y)$ liefert

$$5ypp' = -p^2 \quad \Leftrightarrow \quad \frac{p'}{p} = -\frac{1}{5y}.$$

Trennung der Variablen liefert wiederum

$$p = Cy^{-1/5}.$$

Die Anfangsbedingungen $y_0 = y(0) = 1$ und $v_0 = y'(0) = 1$ liefern

$$1 = C \cdot 1^{-1/5},$$

also $C = 1$ und somit $p = y^{-1/5}$. Trennung der Variablen liefert jetzt

$$x = \int y^{1/5}\, dy = \frac{5}{6} y^{6/5} + \hat{C}.$$

Aus der Anfangsbedingung $y(0) = 1$ folgt

$$0 = \frac{5}{6} \cdot 1^{6/5} + \hat{C},$$

also $\hat{C} = -\frac{5}{6}$, und somit ist

$$y(x) = \left(\frac{6}{5}x + 1\right)^{5/6}$$

die Lösung.

2.3.3 DGLs vom Typ: $y'' = g(y)$ (kein x oder y')

Sei

$$y'' = g(y).$$

Multiplizieren dieser Gleichung mit y' liefert

$$y'y'' - y'g(y) = 0.$$

Deshalb gilt für die Energiefunktion

$$\mathrm{E}(y, y')(x) := \underbrace{\frac{1}{2}(y'(x))^2}_{\text{kinetische Energie}} \underbrace{- \int_{x_0}^{x} y'(\xi) g(y(\xi))\, d\xi}_{\text{potentielle Energie}} = C = \text{ const.}$$

Beachte, dass $\frac{d}{dx}\mathrm{E}(y, y')(x) = y'y'' - y'g(y) = 0$. Die Substitution $\eta = y(\xi)$ liefert

$$\mathrm{E}(y, y') = \frac{1}{2}(y')^2 - \int_{y_0}^{y} g(\eta)\, d\eta = C.$$

Kochrezept 2.36 (für $y'' = g(y)$, Energiemethode)

1. Die Lösungen (Nullstellen) von $g(\eta) = 0$ liefern die stationären Lösungen $y \equiv \eta$ (Ruhelage).

2. Bestimme die Stammfunktion
$$G(y) = \int g(y)\,dy.$$
3. Auflösen der Energiegleichung $(y')^2 = 2G(y) + C_1$ nach y' liefert
$$y' = \pm\sqrt{2G(y) + C_1}.$$
4. Vorzeichen und 1. Integrationskonstante C_1 an die Anfangsbedingung anpassen.
5. Obige Gleichung für y' durch Trennung der Variablen lösen, also
$$x = \int \frac{dy}{\pm\sqrt{2G(y) + C_1}} + C_2.$$
6. Eventuell nach y auflösen und C_2 an die Anfangsbedingung anpassen.

Beispiel 2.37 (Pendelgleichung, siehe Beispiel 1.2)
Sei
$$y'' + \gamma \sin(y) = 0$$
mit $\gamma > 0$. So sind $y = k\pi$ mit $k \in \mathbb{Z}$ die Ruhelagen. Die Energiemethode liefert
$$y' = \pm\sqrt{C_1 + 2\gamma\cos(y)} = \pm\sqrt{\underbrace{C_1 + 2\gamma}_{=C_2} - 4\gamma\sin^2(\frac{y}{2})}.$$
Seien die Anfangsbedingungen $y(0) = 0$ und $y'(0) = v_0 > 0$ gegeben. So ist
$$y'(x) = \sqrt{v_0^2 - 4\gamma\sin^2(\frac{y}{2})}$$
solange der Radikand ≥ 0 ist, also für $|\sin(\frac{y}{2})| \leq \frac{v_0}{2\sqrt{\gamma}}$. Jetzt gilt
$$x = x(y) = \int_0^y \frac{dv}{\sqrt{v_0^2 - 4\gamma\sin^2(\frac{v}{2})}},$$
was sich nicht trivial nach $y(x)$ auflösen lässt.

2.4 Potenzreihenansatz

Satz 2.38
Sei $D \subset \mathbb{R}^{n+1}$ offen und $(x_0, y_0) \in D$. Die Koordinatenfunktionen f_k von $f : D \to \mathbb{R}^n$ seien um (x_0, y_0) in Potenzreihen entwickelbar. Dann ist das AWP
$$y' = f(x, y), \quad y(x_0) = y_0$$
eindeutig lösbar und die Lösung lässt sich um x_0 in eine vektorwertige Potenzreihe
$$y(x) = \sum_{k=0}^{\infty} a_k (x - x_0)^k$$
entwickeln.

Der Potenzreihenansatz wird immer an den Typ der DGL angepasst, läuft aber immer nach dem gleichen Muster ab. Wir zeigen das Vorgehen hier nur exemplarisch für einen Typ.

Kochrezept 2.39 (Koeffizientenvergleich, Potenzreihenansatz)

1. Sei
$$y^{(n)} = f(x, y, \ldots, y^{(n-1)}) \tag{2.18}$$
mit den Anfangsbedingungen
$$y(x_0) = y_0,\ y'(x_0) = y_1,\ \ldots,\ y^{(n-1)}(x_0) = y_{n-1}.$$
2. Wir machen den Ansatz
$$y = \sum_{k=0}^{\infty} a_k (x - x_0)^k.$$
Damit sind
$$\begin{aligned}
y &= \sum_{k=0}^{\infty} a_k (x - x_0)^k \\
y' &= \sum_{k=1}^{\infty} k a_k (x - x_0)^{k-1} = \sum_{k=0}^{\infty} (k+1) a_{k+1} (x - x_0)^k \\
y'' &= \sum_{k=1}^{\infty} k(k+1) a_{k+1} (x - x_0)^{k-1} = \sum_{k=0}^{\infty} (k+1)(k+2) a_{k+2} (x - x_0)^k \\
&\vdots \\
y^{(n)} &= \sum_{k=0}^{\infty} (k+1) \cdot \ldots \cdot (k+n) a_{k+n} (x - x_0)^k
\end{aligned}$$
3. Die Anfangsbedingungen liefern $a_0 = y(x_0) = y_0$, $a_1 = y'(x_0) = y_1$, $\cdots$, $(n-1)!\, a_{n-1} = y^{(n-1)}(x_0) = y_{n-1}$.
4. Den Potenzreihenansatz für y und die Potenzreihen von f in die DGL (2.18) einsetzen und die restlichen Koeffizienten durch Vergleichen bestimmen.

Beispiel 2.40
Sei
$$y' = x^2 + y^2, \quad y(0) = 1$$
und
$$y = \sum_{k=0}^{\infty} a_k x^k.$$
Somit ist
$$y' = \sum_{k=0}^{\infty} (k+1) a_{k+1} x^k$$
und aus der Anfangsbedingung $y(0) = 1$ folgt $a_0 = 1$. Einsetzen in die DGL liefert mit dem Cauchy-Produkt für Reihen
$$\sum_{k=0}^{\infty} (k+1) a_{k+1} x^k = 1 \cdot x^2 + \left(\sum_{k=0}^{\infty} a_k x^k \right)^2 = x^2 + \sum_{k=0}^{\infty} \left(\sum_{j=0}^{k} a_j a_{k-j} \right) x^k.$$
Koeffizientenvergleich ergibt:
$$\begin{aligned}
&\text{Zu } x^0: && a_1 = a_0 \cdot a_0 = 1 && \Rightarrow && a_1 = 1 \\
&\text{Zu } x^1: && 2a_2 = a_0 a_1 + a_1 a_0 = 1 + 1 = 2 && \Rightarrow && a_2 = 1 \\
&\text{Zu } x^2: && 3a_3 = 1 + (a_0 a_2 + a_1 a_1 + a_2 a_0) = 1 + 1 + 1 + 1 = 4 && \Rightarrow && a_3 = \frac{4}{3}.
\end{aligned}$$
Für $k \geq 3$ finden wir leicht die Rekursionsformel
$$(k+1) a_{k+1} = \sum_{j=0}^{k} a_j a_{k-j}.$$
Diese Riccati-Gleichung ist nicht "elementar" lösbar. Zeichnen der ersten Potenzreihenapproximation liefert uns jedoch ein Gefühl für die Lösung.

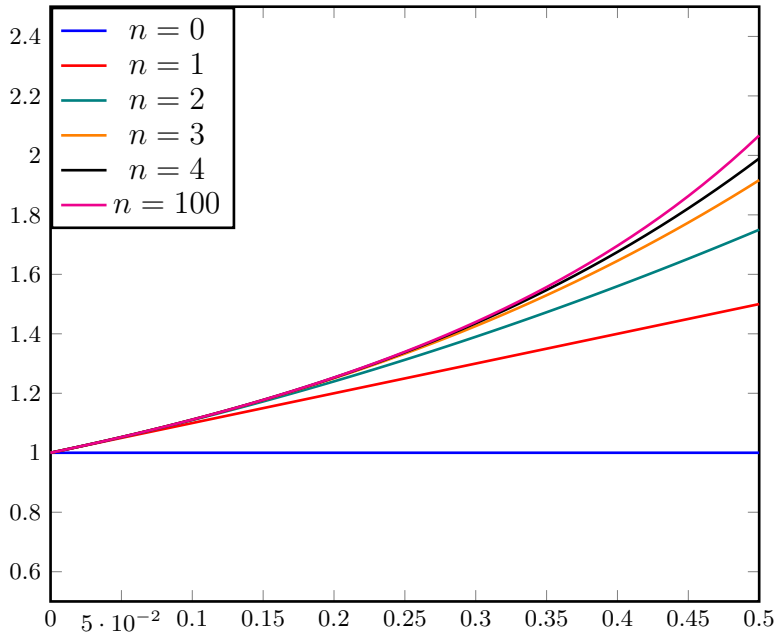


Abb. 2.1: Darstellung der Potenzreihenapproximation $y \approx \sum_{k=0}^{n} a_k x^k$ für unterschiedliche n.

Anstelle des Koeffizientenvergleichs gibt es noch die Methode der fortgesetzte Differentiation.

Kochrezept 2.41 (Fortgesetzte Differentiation, Potenzreihenansatz)

1. Sei
$$y'(x) = f(x, y(x)), \quad y(x_0) = y_0.$$
2. Wir machen den Ansatz
$$y(x) = \sum_{k=0}^{\infty} a_k (x - x_0)^k,$$
und somit ist $n!\, a_n = y^{(n)}(x_0)$.
3. Es gilt
$$0!\, a_0 = a_0 = y(x_0) = y_0$$
$$1!\, a_1 = a_1 = y'(x_0) = f(x_0, y(x_0)) = f(x_0, y_0) = f(x_0, a_0).$$
Für a_2 leiten wir die DGL $y'(x) = f(x, y(x))$ nach x ab. Dies liefert
$$y'' = \frac{d}{dx} f(x, y(x)) = f_x(x, y(x)) + f_y(x, y(x)) \cdot y'(x)$$
und damit
$$2!\, a_2 = y''(x_0) = f_x(x_0, a_0) + f_y(x_0, a_0) \cdot a_1.$$
Erneutes Ableiten liefert
$$\begin{aligned} y^{(3)} &= f_{xx}(x, y(x)) + f_{xy}(x, y(x)) \cdot y'(x) + f_{xy}(x, y(x)) \cdot y'(x) + f_{yy}(x, y(x)) \cdot y'(x) \cdot y'(x) + f_y(x, y(x)) \cdot y''(x) \\ &= f_{xx}(x, y(x)) + 2y'(x) f_{xy}(x, y(x)) + (y'(x))^2 \cdot f_{yy}(x, y(x)) + y''(x) \cdot f_y(x, y(x)). \end{aligned}$$
und damit
$$3!\, a_3 = f_{xx}(x_0, a_0) + 2a_1 f_{xy}(x_0, a_0) + a_1^2 f_{yy}(x_0, a_0) + 2!\, a_2 f_y(x_0, a_0).$$

Die a_n mit $n \geq 4$ werden analog bestimmt.

Beispiel 2.42 (Vgl. Beispiel 2.40)
Sei

$$y' = x^2 + y^2, \quad y(0) = 1.$$

Mit der Methode der fortgesetzten Differentiation erhalten wir

$$\begin{aligned}
&0!\, a_0 = a_0 = y_0 = 1 && \Rightarrow a_0 = 1\\
&1!\, a_1 = a_1 = y'(x_0) = y'(0) = (y(0))^2 = 1 && \Rightarrow a_1 = 1\\
&2!\, a_2 = 2a_2 = y''(x_0) = (2x + 2yy')\big|_{x=0} = 2y(0)\cdot y'(0) = 2\cdot 1\cdot 1 = 2 && \Rightarrow a_2 = 1\\
&3!\, a_3 = 6a_3 = y^{(3)}(x_0) = (2 + 2(y')^2 + 2yy'')\big|_{x=0} = 2(1 + 1 + 1\cdot 2) = 8 && \Rightarrow a_3 = \frac{8}{6}\\
&4!\, a_4 = 24a_4 = y^{(4)}(x_0) = (4y'y'' + 2yy'' + 2yy^{(3)})\big|_{x=0} = 6\cdot 1\cdot 2 + 2\cdot 1\cdot 8 = 12 + 16 = 28 && \Rightarrow a_4 = \frac{28}{24}\\
&\quad\vdots
\end{aligned}$$

2.5 Lineare Systeme

Wie wir am Anfang dies Kapitels gesehen haben, lassen sich DGLs n-ter Ordnung in ein System von n DGLs 1. Ordnung umschreiben. Im Fall von linearen Systemen mit konstanten Koeffizienten lässt sich wieder ein Kochrezept zum exakten Lösen dieser formulieren.
Sei $A = A(x) \in \mathbb{R}^{m\times n}$ eine reelle Matrix. Die Matrixnorm $\|\cdot\|$ wird durch die Vektornorm $\|\cdot\|$ induziert. Insbesondere gelte

$$\|A\cdot B\| \le \|A\|\cdot\|B\| \quad \text{und} \quad \|Ay\| \le \|A\|\cdot\|y\|.$$

Ableitungen und Integrale werden immer komponentenweise verstanden. So ist

$$\int A(x)\,dx = \left(\int a_{ij}(x)\,dx\right)_{i,j}.$$

Im folgenden benötigen wir die Matrix-Exponentialfunktion, d.h. die für alle Matrizen $A \in \mathbb{R}^{n\times n}$ konvergente Matrixreihe

$$e^A = \exp A := \sum_{k=0}^{\infty} \frac{A^k}{k!} = I + A + \frac{1}{2}A^2 + \frac{1}{6}A^3 + \dots.$$

Die Matrix-Exponentialreihe konvergiert bezüglich jeder Norm auf $\mathbb{R}^{n\times n}$ absolut und gleichmäßig auf Kompakta. Für submultiplikative Normen gilt

$$\|e^A\| = \left\|\sum_{k=0}^{\infty}\frac{A^k}{k!}\right\| \le \sum_{k=0}^{\infty}\frac{1}{k!}\|A^k\| \le \sum_{k=0}^{\infty}\frac{1}{k!}\|A\|^k = e^{\|A\|}.$$

Wann immer die Matrixmultiplikation wohldefiniert ist gelten

1. $(AB)' = A'B + AB'$
2. $\left\|\int_a^b A(t)\,dt\right\| \le \int_a^b \|A(t)\|\,dt$
3. Für stetig differenzierbare Matrixfunktionen gilt

$$A\cdot A' = A'\cdot A \quad \Rightarrow \quad (e^A)' = A'\cdot e^A = e^A\cdot A'$$

4. Für konstante Matrixfunktionen gilt

$$(e^{Ax})' = A\cdot e^{Ax}.$$

5. Aus $AB = BA$ folgt $e^Ae^B = e^{A+B} = e^Be^A$. Insbesondere ist $e^Ae^{-A} = I$ und $(e^A)^{-1} = e^{-A}$.

Beispiel 2.43 1. $\exp\begin{pmatrix} z & 0\\ 0 & w\end{pmatrix} = \sum_{k=0}^{\infty}\frac{1}{k!}\begin{pmatrix} z & 0\\ 0 & w\end{pmatrix}^k = \sum_{k=0}^{\infty}\frac{1}{k!}\begin{pmatrix} z^k & 0\\ 0 & w^k\end{pmatrix} = \begin{pmatrix} e^z & 0\\ 0 & e^w\end{pmatrix}$

2. $\exp\begin{pmatrix} x & -y\\ y & x\end{pmatrix} = e^x\begin{pmatrix}\cos y & -\sin y\\ \sin y & \cos y\end{pmatrix}$

Beweis. Sei $J := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. So ist $A := \begin{pmatrix} x & -y \\ y & x \end{pmatrix} = xI + yJ$ mit xI und yJ vertauschbar. Daher gilt $e^A = e^{xI} \cdot e^{yJ}$. Wegen Teil 1 gilt $e^{xI} = e^x I$, und wegen $J^2 = -I$, $J^3 = -J$, $J^4 = I$, etc. gilt

$$e^{yJ} = \sum_{k=0}^{\infty} \frac{y^k}{k!} J^k = \sum_{k=0}^{\infty} \frac{(-1)^k y^{2k}}{(2k)!} I + \sum_{k=0}^{\infty} \frac{(-1)^k y^{2k+1}}{(2k+1)!} J = \cos(y)I + \sin(y)J = \begin{pmatrix} \cos y & -\sin y \\ \sin y & \cos y \end{pmatrix}.$$

□

Wie für linear DGLs gilt für lineare Systeme

Definition 2.44
Sei $\emptyset \neq I \subset \mathbb{R}$ ein offenes Intervall, $x_0 \in I$, $y_0 \in \mathbb{R}^n$ und $A : I \to \mathbb{R}^{n \times n}$, $b : I \to \mathbb{R}^n$ stetig. Eine DGL der Form

$$y' = A(x)y + b(x) \tag{2.19}$$

heißt reelles lineares DGL-System n-ter Ordnung. Die Funktion $b(x)$ heißt Störglied oder Inhomogenität. Das System heißt homogen, falls $b \equiv 0$, ansonsten inhomogen.

$$y' = A(x)y + b(x), \quad y(x_0) = y_0 \tag{2.20}$$

heißt lineares AWP n-ter Ordnung.

Satz 2.45 (Struktursatz für lineare Systeme)
Alle Lösungen des inhomogenen Systems (2.19) haben die Form

$$y(x) = y_s(x) + y_h(x) = y_s(x) + C_1 y_1(x) + \ldots + C_n y_n(x)$$

mit beliebigen $C_i \in \mathbb{R}$. Dabei ist y_s eine spezielle Lösung des inhomogenen Systems (2.19) sowie y_h eine beliebige Lösung des zugehörigen homogenen Systems. $y_1, \ldots, y_n$ sind Basislösungen des homogenen Systems. Die $\mathbb{R}^n$ wertigen Lösungen eines linearen Systems n-ter Ordnung bilden einen n-dimensionalen affinen Funktionenraum über $\mathbb{R}$. Dieser ist sogar ein Vektorraum wenn $b \equiv 0$.

Definition 2.46
Eine Matrix $Y = (y_1, \ldots, y_m)$ heißt Lösungsmatrix des homogenen Systems, wenn die Spalten y_i Lösungen sind. Für die Lösungsmatrix gilt die Matrix-DGL $Y' = A(x)Y$. Ist $m = n$ so heißt die Lösungsmatrix Wronski-Matrix. Sind die Spalten y_i linear unabhängig, so heißt die Wronski-Matrix auch Fundamentalmatrix. Die Spalten bilden dann ein Fundamentalsystem, also eine Basis des Lösungsraums.

Für denn Fall eines homogenen Systems lässt sich die Lösung explizit angeben

Satz 2.47
Sei $B : I \to \mathbb{R}^{n \times n}$ mit $B \cdot B' = B' \cdot B$, so ist $Y(x) = e^{B(x)}$ eine Fundamentalmatrix des homogenen Systems $y' = B'(x)y$.

Ist A konstant, so ist $B(x) = Ax$ und $Y(x) = e^{Ax}$ die Fundamentalmatrix. Im Fall von Anfangsbedingungen ist die eindeutige Lösung somit $y(x) = y(x_0)e^{A(x-x_0)}$. Eine alternative Möglichkeit zum Bestimmen einer Lösungsbasis ist die Eigenwertmethode

Kochrezept 2.48 (Eigenwertmethode für $y' = Ay$ mit konstanter Matrix A)

1. Bestimme die Eigenwerte λ_ν von A und ihre Vielfachheit m_ν, also die Nullstellen des charakteristischen Polynoms $\chi(\lambda) = \det(A - \lambda I)$.
2. Zu jedem Eigenwert λ_ν der Vielfachheit m_ν sind m_ν lineare unabhängige Lösungen der Form $p_j(x)e^{\lambda_\nu x}$ mit Vektorpolynom $p_j(x)$ vom Grad $\leq j$ $(j = 0, 1, \ldots, m_\nu - 1)$ zu bestimmen
3. Das so bestimmte Fundamentalsystem ist im Allgemeinen komplex. Da die Matrix A reell ist, liefern Real- und Imaginärteil des komplexen Fundamentalsystems ein reelles Fundamentalsystem.

Bemerkung 2.49 1. Über $\mathbb{C}$ zerfällt das charakteristische Polynom in n Linearfaktoren. Es gilt also $\sum m_\nu = n$.

2. Zu jedem Eigenwert λ_ν werden als erstes die zugehörigen Eigenvektoren $v \in \mathbb{C}^n$ bestimmt. Denn, wie leicht gezeigt werden kann ist $y = ve^{\lambda_\nu x}$ mit konstantem $v \in \mathbb{C}^n \setminus \{0\}$ genau dann eine Lösung von $y' = Ay$, wenn v ein Eigenvektor von A zum Eigenwert λ_ν ist.
   Liegen m_ν linear unabhängige Eigenvektoren vor, so haben wir bereits genug Lösungen. Ist hingegen die geometrische Vielfachheit k_ν echt kleiner als die algebraische Vielfachheit m_ν, so werden suggestiv $m_\nu - k_\nu$ weitere Lösungen konstruiert. Als erstes werden Lösungen der Form $p_j(x)e^{\lambda_\nu x}$ mit linearem p_j bestimmt, dann solche mir quadratischem p_j und so weiter. Dafür machen wir ein Ansatz mit unbestimmten Vektorkoeffizienten und setzen diesen in das homogene System ein. Die Koeffizienten werden dann durch Koeffizientenvergleich bestimmt.
3. Ist A reell aber die Eigenwerte sind komplex, so treten diese immer in konjungiert komplexen Paaren mit gleicher Vielfachheit auf. Auch besitzen die Eigenwerte λ und $\bar{\lambda}$ konjugiert komplexe Eigenvektoren und liefern somit konjugiert komplexe Basislösungen.

Satz 2.50 → übersprungen
Sei $A \in \mathbb{R}^{n\times n}$ konstant, $\lambda_1, \dots, \lambda_k$ verschiedene reelle und $\lambda_{k+1} = \alpha_{k+1}+i\beta_{k+1}, \dots, \lambda_s, \overline{\lambda_{k+1}}, \dots, \overline{\lambda_s} = \alpha_s - i\beta_s$ die verschiedenen echt komplexen Eigenwerte von A, jeweils mit der Vielfachheit m_ν. Dann existiert ein reelles Fundamentalsystem zu $y' = Ay$ der Form

$$\begin{aligned} e^{\lambda_\nu x}p_{\nu,j}(x), &\quad (0 \le j < m_\nu,\ 1 \le \nu \le k) \\ \operatorname{Re}\left(e^{\lambda_\mu x}p_{\mu,j}(x)\right),\ \operatorname{Im}\left(e^{\lambda_\mu x}p_{\mu,j}(x)\right), &\quad (0 \le j < m_\mu,\ k < \mu \le s) \end{aligned}$$

Beispiel 2.51
Es sei

$$y' = Ay = \begin{pmatrix} 1 & -1 \\ 4 & -3 \end{pmatrix} y.$$

Die Matrix A hat den doppelten Eigenwert $\lambda = -1$ mit dem Eigenvektor $v = (1,2)^\top$. Die erste Basislösung ist daher $\phi_1(x) = ve^{-x}$. Für die Zweite machen wir den Ansatz $y = (a+bx)e^{-x}$. Einsetzen in die DGL liefert

$$y' = be^{-x} - (a+bx)e^{-x} = Ay = Aae^{-x} + Abxe^{-x} \quad \Rightarrow \quad -b = Ab \text{ und } b - a = Aa.$$

Foglich ist $b = v = (1,2)^\top$ und durch direktes Berechnen $a = (0,-1)^\top$. Somit ist $\phi_2(x) = (x, 2x-1)^\top e^{-x}$. Die Fundamentalmatrix ist

$$Y(x) = e^{-x}\begin{pmatrix} 1 & x \\ 2 & 2x-1 \end{pmatrix}.$$

Die Bestimmung einer Lösung für inhomogene Systeme basiert wie im ein-dimensionalen auf Variation der Konstanten. Sei Y ein Fundamentalsystem des homogenen Systems $y' = A(x)y$, so erhalten wir eine spezielle Lösung durch den Ansatz

$$y_s = Y(x)\vec{c}(x) = c_1(x)y_1(x) + \dots + c_n(x)y_n(x). \tag{2.21}$$

Einsetzen in das inhomogene System $y' = Ay + b$ liefert $Y(x)\vec{c}' = b(x)$ oder auch $\vec{c}' = Y^{-1}b$. Als Fundamentalmatrix ist Y stets invertierbar. Somit ist $\vec{c} = \int Y^{-1}(x)b(x)\,dx$ und die spezielle Lösung somit

$$y_s(x) = Y(x)\int_{x_0}^{x} Y^{-1}(\xi)b(\xi)\,d\xi.$$

nicht besprochen

Ist A konstant, so ist $Y(x) = e^{Ax}$ und

$$y_s(x) = e^{Ax}\int_{x_0}^{x} e^{-A\xi}b(\xi)\,d\xi = \int_{x_0}^{x} e^{A(x-\xi)}b(\xi)\,d\xi.$$

Im Fall einer konstanten Matrix A kann VdK durch einen Rateansatz ersetzt werden. Hat der Störterm die Form

$$b(x) = p(x)e^{\lambda x}$$

mit einem vektorwertigen Polynom p vom Grad k, so existiert eine spezielle Lösung der Form

$$y_s(x) = q(x)e^{\lambda x}$$

mit einem vektorwertigen Polynom q. Ist λ kein Eigenwert von A, so ist der Grad von q kleiner gleich k. Ist hingegen λ ein m-facher Eigenwert, so ist der Grad von q kleiner gleich $m+k$. Ist der Störterm die Summe solcher speziellen Störglieder, so lässt sich das Problem mit Hilfe des Superpositionsprinzips sukzessive Lösen.

→ Vielfachheit → Grad

$y' = Ay + b_1 + b_2$
$y_1' = Ay_1 + b_1$
$y_2' = Ay_2 + b_2$
$\Big\} \; y_1 = y_2 = y$ dann $y' = y_1' + y_2'$
$= A(y) + b_1 + b_2$

Beispiel 2.52
Es sei

$$y' = Ay + pe^{-x} = \begin{pmatrix} 1 & -1 \\ 4 & -3 \end{pmatrix} y + \begin{pmatrix} 1 \\ 0 \end{pmatrix} e^{-x}.$$

Die Matrix A hat den doppelten Eigenwert $\lambda = -1$ mit dem Eigenvektor $v = (1,2)^\top$. Die Fundamentalmatrix ist

$$Y(x) = e^{-x} \begin{pmatrix} 1 & x \\ 2 & 2x-1 \end{pmatrix}.$$

Grad 0+2

Weil p konstant und $\lambda = -1$ ein doppelter Eigenwert ist, liefert der Rateansatz $y_s(x) = q(x)e^{-x} = (a + bx + cx^2)e^{-x}$. Einsetzen in die inhomogene Gleichung liefert

$$y' = \left[(b-a) + (2c-b)x - cx^2\right] e^{-x} = A(a + bx + cx^2)e^{-x} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} e^{-x}.$$

Koeffizientenvergleich ergibt

$$(A+I)c = 0, \quad (A+I)b = 2c, \quad (A+I)a = b - (1,0)^\top$$

mit den Lösungen $c = v = (1,2)^\top$, $b = (1,0)^\top$ und $a = (0,0)^\top$. Die allgemeine Lösung lautet daher

$$y(x) = y_s(x) + C_1 y_1(x) + C_2 y_2(x) = \begin{pmatrix} x^2 + x \\ 2x^2 \end{pmatrix} e^{-x} + C_1 \begin{pmatrix} 1 \\ 2 \end{pmatrix} e^{-x} + C_2 \begin{pmatrix} x \\ 2x-1 \end{pmatrix} e^{-x}.$$

3

Analysis für DGLs

Lipschitz konstante:
- wie stark darf sich der Funktionswert ändern pro Argumentänderung

Ziel dieses Kapitels sind Beweise zu Existenz, Eindeutigkeit und Abhängigkeiten der Lösung von den Daten (rechte Seite und Anfangswert). Die meisten Aussagen werden für reellwertige Lösungen formuliert. Verallgemeinerungen zu vektorwertigen Lösungen sind dem Leser als leichte Übung überlassen. Ebenso lassen sich die rechtsseitigen Streifen- und Quaderversionen zu linksseitigen Versionen leicht umschreiben.

3.1 Grundlegende Resultate

Satz 3.1 (Banachscher Fixpunktsatz)
Sei $D \neq \emptyset$ eine abgeschlossene Teilmenge eines vollständigen und normierten Raumes X und

zusammenziehen

$$f : D \to D$$

ein kontrahierender (Lipschitz-Konstante $L < 1$) Operator, dann existiert genau ein Fixpunkt $x^* \in D$ von f, und die durch

↳ $f(x^*) = x^*$

$$x_{n+1} = f(x_n) \text{ für } n = 0, 1, 2, \ldots$$

gegebene Folge konvergiert für jeden Startwert $x_0 \in D$ gegen x^*. Es gelten zudem die a priori Fehlerabschätzung

max. Fehler

$$\|x_n - x^*\| \leq \frac{L^n}{1-L}\|x_1 - x_0\|$$

und die a posteriori Fehlerabschätzung

wie weit noch weg von x^*, basierend auf letzten Änd.

$$\|x_n - x^*\| \leq \frac{L}{1-L}\|x_n - x_{n-1}\|,$$

wobei $L < 1$ die Kontraktionszahl des Operators repräsentiert.

Beweis. Aus $x_n = f(x_{n-1})$ folgt unmittelbar

$$\|x_{n+1} - x_n\| = \|f(x_n) - f(x_{n-1})\| \leq L\|x_n - x_{n-1}\| \leq L^2\|x_{n-1} - x_{n-2}\| \leq \cdots \leq L^n\|x_1 - x_0\|.$$

Für $p \geq 1$ liefert dies mit der Teleskopsumme, der Dreiecksungleichung und der geometrischen Reihe

$\|x_{n+p} - x_{n+p-1} + x_{n+p-1} - \ldots x_n\|$

$$\|x_{n+p} - x_n\| \leq \sum_{i=1}^{p} \|x_{n+i} - x_{n+i-1}\| \leq \sum_{i=1}^{p} L^{n+i-1}\|x_1 - x_0\| = \|x_1 - x_0\|L^n \sum_{i=0}^{p-1} L^i = \|x_1 - x_0\|L^n \frac{1-L^p}{1-L}.$$

Dug

Folglich ist $\{x_n\} \subset D$ eine Cauchy-Folge, weil die rechte Seite in der obigen Abschätzung für n hinreichend groß beliebig klein wird. Weil D eine abgeschlossen Teilmenge eines normierten Raums ist, ist $\lim_{n\to\infty} x_n = x^* \in D$. Weil jeder kontrahierender Operator trivialerweise stetig ist, ist

$$x^* = \lim_{n\to\infty} x_n = \lim_{n\to\infty} f(x_{n-1}) = f(\lim_{n\to\infty} x_{n-1}) = f(x^*)$$

und x^* somit ein Fixpunkt.
Seien x^* und y^* beides Fixpunkte von f, so gilt

EXISTENZ

Fixpunkte

$$\|x^* - y^*\| = \|f(x^*) - f(y^*)\| \leq L\|x^* - y^*\| < \|x^1 - y^1\|$$

wegen $L < 1$

mit $L < 1$. Daraus folgt $x^* = y^*$ und die Eindeutigkeit des Fixpunktes.
Für die Fehlerabschätzungen haben wir (die Norm ist stetig)

$$\|x^* - x_n\| = \|\lim_{p\to\infty} x_{n+p} - x_n\| = \lim_{p\to\infty} \|x_{n+p} - x_n\| \leq \lim_{p\to\infty} \frac{1-L^p}{1-L} L^n \|x_1 - x_0\| = \frac{L^n}{1-L}\|x_1 - x_0\|.$$

Analog zu oben erhalten wir

$$\|x_{n+p} - x_n\| \leq \sum_{i=1}^{p} \|x_{n+i} - x_{n+i-1}\| \leq \sum_{i=1}^{p} L^i \|x_n - x_{n-1}\| = \|x_n - x_{n-1}\| L \sum_{i=0}^{p-1} L^i = \|x_n - x_{n-1}\| L \frac{1-L^p}{1-L}$$

und damit sofort

$$\|x^* - x_n\| = \lim_{p\to\infty} \|x_{n+p} - x_n\| \leq \lim_{p\to\infty} \|x_n - x_{n-1}\| L \frac{1-L^p}{1-L} = \frac{L}{1-L}\|x_n - x_{n-1}\|.$$

□

EINDEUTIGKEIT

Definition 3.2 (Lipschitz-Stetigkeit bzgl. y)
Die Funktion $f(x, y)$ mit $f : D \to \mathbb{R}^n$ mit $D \subset \mathbb{R} \times \mathbb{R}^n$ heißt Lipschitz-stetig bzgl. y, wenn eine Konstante $L \geq 0$ existiert, so dass für alle $(x, y), (x, \bar{y}) \in D$ gilt

$$\|f(x, y) - f(x, \bar{y})\| \leq L\|y - \bar{y}\|.$$

→ Funktionswertänderung ist proportional zu Argumentänderung (mit Faktor L)

Definition 3.3 (Lokale Lipschitz-Stetigkeit bzgl. y)
Die Funktion $f(x, y)$ mit $f : D \to \mathbb{R}^n$ mit $D \subset \mathbb{R} \times \mathbb{R}^n$ heißt lokal Lipschitz-stetig bzgl. y, wenn für alle $(x_0, y_0) \in D$ eine Umgebung $U = U(x_0, y_0)$ und eine Lipschitz-Konstante $L = L(x_0, y_0) \geq 0$ existieren, so dass f in $U \cap D$ Lipschitz-stetig bzgl. y ist.

Bemerkung 3.4 → Beweis für Lip.st. wenn f differenzierbar
Ist ∇f beschränkt, d.h. $\|\nabla f\| \leq L$, so folgt mit dem Mittelwertsatz, siehe Lemma 5.7,

$$\|f(x) - f(y)\| = \left\| \int_0^1 \nabla f(y + \tau(x - y))\, d\tau \cdot (x - y) \right\| \leq \int_0^1 \|\nabla f(y + \tau(x - y))\|\, d\tau \|x - y\| \leq L\|x - y\|.$$

Also ist f Lipschitz-stetig. Analog folgt, ist $\partial_y f$ stetig, so ist f lokal Lipschitz-stetig bzgl. y.

Satz 3.5
Sei $f : D = J \times G \subset \mathbb{R}^{n+1} \to \mathbb{R}^n$ stetig. Dann ist $y : J \subset \mathbb{R} \to \mathbb{R}^n$ genau dann eine Lösung des AWP $y' = f(x, y)$ mit $y(\xi) = \eta$ für $(\xi, \eta) \in D$, wenn y die Integralgleichung

$$y(x) = (Ty)(x) := \eta + \int_\xi^x f(t, y(t))\, dt \tag{3.1}$$

löst.

Beweis. Sei $y : J \to \mathbb{R}^n$ eine (differenzierbare) Lösung des AWP. Daraus folgt insbesondere, dass $y \in C^0(J)$ ist. Da f stetig ist, ist die Abbildung $x \mapsto f(x, y(x))$ als Verkettung stetiger Funktionen ebenfalls stetig. Weil $y' = f(x, y)$ gilt, ist somit $y \in C^1(J)$. Nach dem Hauptsatz der Differential- und Integralrechnung gilt

$$y(x) = y(\xi) + \int_\xi^x y'(t)\, dt = \eta + \int_\xi^x f(t, y(t))\, dt.$$

Für die Rückrichtung sei y eine Lösung von (3.1). Dann ist

$$y(\xi) = \eta + \int_\xi^\xi f(t, y(t))\, dt = \eta.$$

Ableiten von (3.1) liefert die DGL

$$y'(x) = \frac{d}{dx}\left(\eta + \int_\xi^x f(t, y(t))\,dt\right) = f(x, y(x)).$$

□

Lemma 3.6
Sei $D \subset \mathbb{R}^{n+1}, A \subset D$ kompakt (hier abgeschlossen und beschränkt) und $f : D \to \mathbb{R}^n$ stetig.

1. Ist ϕ eine Lösung der DGL $y' = f(x, y)$ in $[\xi, b)$ mit $\{(x, \phi(x)) : \xi \le x < b\} \subset A$, so lässt sich ϕ auf $[\xi, b]$ fortsetzten.
2. Ist ϕ eine Lösung der DGL $y' = f(x, y)$ in $[\xi, b]$ und ψ eine Lösung in $[b, \theta]$ sowie $\phi(b) = \psi(b)$, so ist

$$u(x) = \begin{cases} \phi(x), & x \in [\xi, b] \\ \psi(x), & x \in [b, \theta] \end{cases}$$

eine Lösung der DGL in $[\xi, \theta]$.

Beweis. 1. Weil f stetig ist, ist f auf A beschränkt, d.h. $\|f\| \le C$ für ein $C \in \mathbb{R}$. Damit ist $\|\phi'\| = \|f(x, \phi)\| \le C$ und somit existiert $\beta = \lim_{x \to b^-} \phi(x)$. Zudem ist $(b, \beta) \in A$, weil A kompakt ist. Für $\phi(b) := \beta$ ist $x \mapsto f(x, \phi(x))$ stetig in $[\xi, b]$. Weil ϕ die DGL löst gilt

$$\phi(x) = \phi(\xi) + \int_\xi^x f(t, \phi(t))\,dt$$

für $x \in [\xi, b)$ und damit ist

$$\phi(b) = \lim_{x \to b^-} \phi(x) = \lim_{x \to b^-} \phi(\xi) + \int_\xi^x f(t, \phi(t))\,dt = \phi(\xi) + \int_\xi^b f(t, \phi(t))\,dt.$$

Folglich ist ϕ eine Lösung von $\phi' = f(x, \phi)$ auf $[\xi, b]$.

2. Per Konstruktion ist u in b links- und rechtsseitig differenzierbar. Überdies gilt

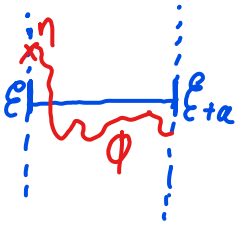
$$\begin{aligned}\lim_{x \to b^-} u'(x) &= \lim_{x \to b^-} \phi'(x) = \lim_{x \to b^-} f(x, \phi(x)) = f(b, \phi(b)) = f(b, \psi(b)) = \lim_{x \to b^+} f(x, \psi(x)) = \lim_{x \to b^+} \psi'(x) \\ &= \lim_{x \to b^+} u'(x).\end{aligned}$$

Somit ist u in b stetig differenzierbar mit $u'(b) = f(b, u(b))$ was zu zeigen war.

□

3.2 Existenz und Eindeutigkeit nach Picard-Lindelöf

Satz 3.7 (Picard-Lindelöf: Streifenversion)
Sei $S := J \times \mathbb{R}$ mit $J := [\xi, \xi + a]$. Die Funktion $f : S \to \mathbb{R}$ sei stetig und Lipschitz-stetig bzgl. y mit Lipschitz-Konstante L. Dann hat das AWP $y' = f(x, y)$ mit $y(\xi) = \eta$ genau eine Lösung. Diese Lösung existiert auf ganz J.



Beweis. Nach Satz 3.5 sind die Lösungen des AWP die Fixpunkte von $T : C(J) \to C(J)$ mit

$$Ty = \eta + \int_\xi^x f(t, y(t))\,dt.$$

$C(J)$ ist bzgl. der gewichteten Supremumsnorm $\|y\|_{\infty,\alpha} = \sup_{x \in J} |y(x)| e^{-\alpha \cdot x}$ abgeschlossen, weil jede bzgl. $\|\cdot\|_{\infty,\alpha}$ konvergente Funktionenfolge auch gleichmäßig konvergiert. Deshalb ist die Grenzfunktion solch einer

Funktionenfolge selber auch stetig. Um zu zeigen, dass T eine Kontraktion ist bemerken wir, dass für $y, z \in C(J)$ gilt

$$\begin{aligned}|(Ty)(x) - (Tz)(x)| &= \left| \int_\xi^x f(t, y(t)) - f(t, z(t))\, dt \right| \le \int_\xi^x |f(t, y(t)) - f(t, z(t))|\, dt \le \int_\xi^x L|y(t) - z(t)|\, dt \\ &= L \int_\xi^x |y(t) - z(t)| e^{-\alpha t} e^{\alpha t}\, dt \le L\|y - z\|_{\infty,\alpha} \int_\xi^x e^{\alpha \cdot t}\, dt = \frac{L}{\alpha}\|y - z\|_{\infty,\alpha}(e^{\alpha x} - e^{\alpha \xi}) \\ &\le \frac{L}{\alpha}\|y - z\|_{\infty,\alpha} e^{\alpha x}.\end{aligned}$$

Somit ist für alle $x \in J$ und alle y, $z \in C(J)$

$$|(Ty)(x) - (Tz)(x)| e^{-\alpha \cdot x} \le \frac{L}{\alpha}\|y - z\|_{\infty,\alpha}.$$

Folglich ist

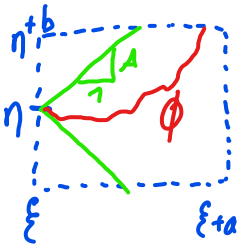
$$\|Ty - Tz\|_{\infty,\alpha} = \sup_{x \in J} |(Ty)(x) - (Tz)(x)| e^{-\alpha \cdot x} \le \frac{L}{\alpha}\|y - z\|_{\infty,\alpha}$$

und somit für $\alpha = 2L$ eine Kontraktion. Jetzt liefert der Banachsche Fixpunktsatz die behauptete Existenz und Eindeutigkeit. □

Korollar 3.8
Die Picard-Iteration

$$y_0(x) \equiv \eta \quad \text{und} \quad y_{n+1}(x) = \eta + \int_\xi^x f(t, y_n(t))\, dt$$

darf oben raus, aber nicht zu früh



konvergiert gegen die Lösung y des AWPs.

Satz 3.9 (Picard-Lindelöf: Lokale Version)
Sei $Q := \{(x, y) \in \mathbb{R}^2 : \xi \le x \le \xi + a,\ |y - \eta| \le b\}$ mit $a, b > 0$ und $(\xi, \eta) \in \mathbb{R}^2$. Weiterhin sei $f : Q \to \mathbb{R}$ stetig und Lipschitz-stetig bzgl. y. Dann existiert genau eine Lösung des AWP $y' = f(x, y)$ mit $y(\xi) = \eta$ im Intervall $J := [\xi, \xi + \alpha]$ für $\alpha := \min\{a, \frac{b}{A}\}$ mit $A := \max_{(x,y) \in Q} |f(x, y)|$.

Beweis. Sei $D := \{y \in C(J) : |y(x) - \eta| \le b\ \forall x \in J\}$. Wie im Beweis von Satz 3.7 ist D bzgl. $\|y\| = \sup_{x \in J} |y(x)| e^{-\alpha \cdot x}$ abgeschlossen, weil wir zusätzlich noch

$$|y(x) - \eta| = |\lim_{n \to \infty} y_n(x) - \eta| = \lim_{n \to \infty} |y_n(x) - \eta| \le \lim_{n \to \infty} b = b$$

für alle Folgen $\{y_n\} \subset D$ mit $y_n \to y$ haben. Außerdem ist $Ty \in D$ für $y \in D$, also eine Selbstabbildung, denn Ty ist offensichtlich stetig und

$$|(Ty)(x) - \eta| = \left| \int_\xi^x f(t, y(t))\, dt \right| \le \int_\xi^x |f(t, y(t))|\, dt \le A \int_\xi^x 1\, dt = A(x - \xi) \le A\alpha \le b$$

$\alpha \le \min(a, \frac{b}{A})$

für alle $x \in J$. Also $T : D \to D$. Analog zum Beweis von Satz 3.7 kann gezeigt werden, dass T eine Kontraktion ist. Der Banachsche Fixpunktsatz liefert dann die Behauptung. □

Satz 3.10
Sei $D \subset \mathbb{R}^2$ offen und $f : D \to \mathbb{R}$ stetig. Weiterhin sei f in D lokal Lipschitz-stetig bzgl. y. Dann ist das AWP $y' = f(x, y)$ mit $y(\xi) = \eta$ für $(\xi, \eta) \in D$ beliebig lokal eindeutig lösbar, d.h. in einer Umgebung der Anfangsdaten $(\xi, \eta) \in D$ existiert genau eine Lösung.

Beweis. Für $a, b > 0$ hinreichend klein existieren eindeutige Lösungen y_l für $f\big|_{Q_l}$ mit $Q_l := \{(x, y) \in D : \xi - a \le x \le \xi, |y - \eta| \le b\}$ und y_r für $f\big|_{Q_r}$ mit $Q_r := \{(x, y) \in D : \xi \le x \le \xi + a, |y - \eta| \le b\}$ nach Satz 3.9. Somit ist

$$y(x) = \begin{cases} y_l(x), & \xi - a \le x \le \xi \\ y_r(x), & \xi \le x \le \xi + a \end{cases}$$

die eindeutige Lösung nach Lemma 3.6. □

3.3 Existenz nach Peano

Für die bloße Existenz einer Lösung reicht bereits die Stetigkeit der rechten Seite, jedoch kann die Eindeutigkeit dabei verloren gehen.

Beispiel 3.11
TdV und die stationäre Lösung $y \equiv 0$ liefern uns für

weil nicht L.st. in y(0)

$$y' = \sqrt{y} \text{ mit } y(0) = 0$$

bereits zwei Lösungen. Darüber hinaus können wir zu einem beliebigen Zeitpunkt von der einen Lösung in die andere übergehen.

Definition 3.12 (Gleichgradige Stetigkeit)
Sei $M = \{f, g, \dots\} \subset C^0(J)$ eine Menge von Funktionen. Die Funktionenmenge heißt gleichgradig stetig, wenn für alle $\epsilon > 0$ ein $\delta = \delta(\epsilon) > 0$ existiert, so dass für alle $f \in M$ und alle $x, \bar{x} \in J$ mit $|x - \bar{x}| < \delta$ gilt

$$|f(x) - f(\bar{x})| < \epsilon. \tag{3.2}$$

Beispiel 3.13
Die Menge der Lipschitz-stetigen Funktionen mit der selben Lipschitz-Konstante L ist gleichgradig stetig, denn mit $\delta = \frac{\epsilon}{L}$ haben wir

$$|f(x) - f(\bar{x})| \leq L|x - \bar{x}| < L\delta < \epsilon.$$

Eine Punktmenge $A \subset J$ ist dicht in $J = [a, b]$, wenn jedes Teilintervall von J mindestens einen Punkt von A enthält.

Lemma 3.14
Ist die Funktionenfolge $\{f_n(x)\}_{n\in\mathbb{N}}$ in $J = [a, b]$ gleichgradig stetig und konvergiert sie für alle $x \in A$, wobei $A \subset J$ eine in J dichte Teilmenge ist, so konvergiert sie gleichmäßig für alle $x \in J$. Die Grenzfunktion $f(x) = \lim_{n\to\infty} f_n(x) \in C^0(J)$.

Beweis. Zu $\epsilon > 0$ sei $\delta = \delta(\epsilon)$ so gewählt, dass die Abschätzung (3.2) für alle Funktionen f_n gilt. Zerlege J in p abgeschlossene Teilintervalle $J_1, \dots, J_p$ der Länge $|J_i| < \delta$. Zu jedem J_i existiert wegen der Dichtheit von A zu J ein $x_i \in J_i \cap A$. Ferner gibt es nach vorausgesetzter Konvergenz ein $n_0 = n_0(\epsilon)$, so dass

$$|f_m(x_i) - f_n(x_i)| < \epsilon$$

für $m, n \geq n_0$ und $1 \leq i \leq p$. Nun sei $x \in J$ beliebig, o.B.d.A. sei $x \in J_i$. Wegen $|x - x_i| \leq |J_i| < \delta$ und der Abschätzung (3.2) folgt somit

$$\begin{aligned} |f_m(x) - f_n(x)| &= |f_m(x) - f_m(x_i) + f_m(x_i) - f_n(x_i) + f_n(x_i) - f_n(x)| \\ &\leq \underbrace{|f_m(x) - f_m(x_i)|}_{\text{gleichg. stetig}} + \underbrace{|f_m(x_i) - f_n(x_i)|}_{\text{Cauchyfolge}} + \underbrace{|f_n(x_i) - f_n(x)|}_{\text{gleichg. stetig}} \\ &< 3\epsilon \end{aligned}$$

für $m, n \geq n_0$. Damit ist f_n eine Cauchyfolge und somit konvergent. Ebenfalls ist gezeigt, dass $f_n(x)$ in J gleichmäßig konvergiert. Dies impliziert auch die Stetigkeit des Grenzwertes. □

Satz 3.15 (Ascoli-Arzelá)
Jede in $J = [a, b]$ gleichgradig stetige Folge von Funktionen $\{f_n(x)\}_{n\in\mathbb{N}}$ mit $|f_n(x)| \leq C$ für $x \in J$ enthält eine in J gleichmäßig konvergente Teilfolge.

Beweis. Sei $A = \{x_1, x_2, \dots\}$ eine abzählbare in J dichte Punktmenge, z.B. $A = \mathbb{Q} \cap J$. Die Zahlenfolge $\{f_n(x_1)\}_{n\in\mathbb{N}}$ ist beschränkt. Sie besitzt nach Bolzano-Weierstraß eine konvergente Teilfolge $\{f_{n_k}(x_1)\}_{k\in\mathbb{N}}$ welches die Funktionenteilfolge $\{f_{n_k}\}_{k\in\mathbb{N}}$ definiert. Um die Anzahl an Indizes zu beschränken schreiben wir $\{f_{1,n}\}_{n\in\mathbb{N}}$ für die erste Teilfolge $\{f_{n_k}\}_{k\in\mathbb{N}}$ von $\{f_n\}_{n\in\mathbb{N}}$. Während die Zahlenfolge $\{f_{1,n}(x_1)\}_{n\in\mathbb{N}}$ konvergiert, wird die Zahlenfolge $\{f_{1,n}(x_2)\}_{n\in\mathbb{N}}$ im Allgemeinen divergieren. Jedoch ist die (Teil-)Folge $\{f_{1,n}(x_2)\}_{n\in\mathbb{N}}$ beschränkt, und somit existiert eine konvergente Teilfolge $\{f_{2,n}(x_2)\}_{n\in\mathbb{N}}$ der (Teil-)Folge $\{f_{1,n}(x_2)\}_{n\in\mathbb{N}}$. Damit erhalten wir die zweite Funktionenteilfolge $\{f_{2,n}\}_{n\in\mathbb{N}}$ die für $x = x_1$ und $x = x_2$ konvergent ist. In dieser Weise führen wir nun fort und erhalten die (Teil)-Funktionenfolgen

$$
\begin{aligned}
\{f_{1,n}\}_{n\in\mathbb{N}} &= f_{1,1}, f_{1,2}, f_{1,3}, \dots \text{ konvergent für } x = x_1\\
\{f_{2,n}\}_{n\in\mathbb{N}} &= f_{2,1}, f_{2,2}, f_{2,3}, \dots \text{ konvergent für } x = x_1, x_2\\
\{f_{3,n}\}_{n\in\mathbb{N}} &= f_{3,1}, f_{3,2}, f_{3,3}, \dots \text{ konvergent für } x = x_1, x_2, x_3\\
\{f_{4,n}\}_{n\in\mathbb{N}} &= f_{4,1}, f_{4,2}, f_{4,3}, \dots \text{ konvergent für } x = x_1, x_2, x_3, x_4\\
&\vdots
\end{aligned}
$$

Die k-te Funktionenteilfolge $\{f_{k,n}\}_{n\in\mathbb{N}}$, also die k-te Zeile im obigen Schema, stellt eine Funktionenteilfolge von $\{f_{k-1,n}\}_{n\in\mathbb{N}}$, also der $(k-1)$-ten Zeile, dar. Sie konvergiert für $x = x_1, \dots, x_k$. Die Diagonalfolge $\{f_{n,n}\}_{n\in\mathbb{N}} = f_{11}, f_{22}, f_{33}, \dots$ konvergiert für alle $x \in A$. Sie ist nämlich jedenfalls von ihrem k-ten Glied an eine Funktionenteilfolge der k-ten Funktionenteilfolge $\{f_{k,n}\}_{n\in\mathbb{N}}$, und schon ganz $\{f_{k,n}\}_{n\in\mathbb{N}}$ konvergiert für $x = x_1, \dots, x_k$. Die gleichmäßige Konvergenz der Diagonalfolge ergibt sich aus Lemma 3.14. □

Satz 3.16 (Streifenversion von Peano)
Sei $f : S = J \times \mathbb{R} \to \mathbb{R}$ stetig und beschränkt mit $J = [\xi, \xi + a]$ für ein $a > 0$. Dann existiert mindestens eine Lösung y zu $y'(x) = f(x, y(x))$ in J und $y(\xi) = \eta$.

Beweis. Es gilt eine Funktion $y \in C(J)$ zu finden, so dass

$$y(x) = \eta + \int_{\xi}^{x} f(t, y(t))\, dt$$

für alle $x \in J$ ist. Dazu konstruieren wir für jedes $\alpha > 0$ eine "Näherungslösung" $z_\alpha \in C(J)$, siehe Bild 3.1, gemäß

$$z_\alpha = \begin{cases} \eta, & x \le \xi \\ \eta + \int_{\xi}^{x} f(t, z_\alpha(t-\alpha))\, dt, & x \in J \end{cases}. \tag{3.3}$$

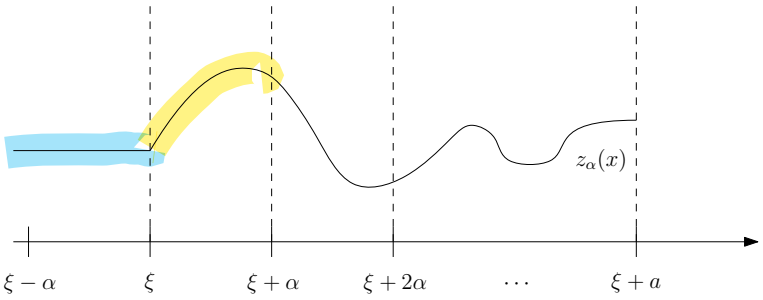


Abb. 3.1: Visualisierung von z_α.

Die Funktion z_α ist wohl definiert, denn durch den Shift $t - \alpha$ für $t \in [\xi, x]$ im Integranten wird z_α immer nur im Bereich $[\xi - \alpha, x - \alpha]$ ausgewertet, in welchem wir es bereits kennen. So ist im ersten Schritt $z_\alpha(x)$ für $x \in [\xi, \xi + \alpha]$ zu bestimmen. Wegen dem Shift ist für das Integral $z_\alpha(t - \alpha)$ mit $t \in [\xi, x] \subseteq [\xi, \xi + \alpha]$ bereits bekannt. $z_\alpha(t - \alpha)$ ist dort konstant η. Im zweiten Schritt wird $z_\alpha(x)$ für $x \in [\xi + \alpha, \xi + 2\alpha]$ mittels (3.3) bestimmt. Dafür wird wegen dem Shift für das Integral lediglich das bereits bekannte z_α eingeschränkt auf $[\xi - \alpha, \xi + \alpha]$ benötigt. Nach endlich vielen solcher Schritte haben wir ganz J abgedeckt und somit ist $z_\alpha(x)$ wohldefiniert.
Das $z_\alpha \in C(J)$ ist, ist trivial. Die Menge $M = \{z_\alpha\}_{\alpha>0}$ dieser in J stetigen Funktionen $z_\alpha(x)$ ist gleichgradig stetig, denn aus $|f| \le C$ folgt sofort $|z'_\alpha| \le C$. Nach Bemerkung 3.4 sind alle z_α Lipschitz-stetig mit der selben Lipschitz-Konstante C. Die Folge $z_1(x), z_{1/2}(x), z_{1/3}(x), \dots$ besitzt nach dem Satz 3.15 von Ascoli-Arzelá eine gleichmäßig konvergente Teilfolge $\{z_{\alpha_n}(x)\}_{n\in\mathbb{N}}$ mit $\alpha_n \to 0^+$. Ihren stetigen Limes bezeichnen wir mit $y(x) := \lim_{n\to\infty} z_{\alpha_n}$. Gemäß (3.3) ist

$$z_{\alpha_n}(x) = \eta + \int_\xi^x f(t, z_{\alpha_n}(t-\alpha_n))\,dt.$$

Nun folgt mit

$$|z_{\alpha_n}(t-\alpha_n) - y(t)| \le |z_{\alpha_n}(t-\alpha_n) - z_{\alpha_n}(t)| + |z_{\alpha_n}(t) - y(t)| \le C\alpha_n + |z_{\alpha_n}(t) - y(t)|,$$

dass auch $z_{\alpha_n}(t-\alpha_n)$ gleichmäßig in J gegen $y(t)$ konvergiert. Also konvergiert auch $f(t, z_{\alpha_n}(t-\alpha_n))$ gleichmäßig in J gegen $f(t, y(t))$. Mit der gleichmäßigen Konvergenz erhalten wir

$$y(x) := \lim_{n\to\infty} z_{\alpha_n}(x) = \eta + \lim_{n\to\infty} \int_\xi^x f(t, z_{\alpha_n}(t-\alpha_n))\,dt = \eta + \int_\xi^x \lim_{n\to\infty} f(t, z_{\alpha_n}(t-\alpha_n))\,dt = \eta + \int_\xi^x f(t, y(t))\,dt.$$

Mit Satz 3.5 folgt die Behauptung. □

Satz 3.17 (Rechteckversion von Peano)
Sei $Q = \{(x,y) \in \mathbb{R}^2 : \xi \le x \le \xi + a,\ |y-\eta| \le b\}$ mit $a, b > 0$ und $f : Q \to \mathbb{R}$ stetig. Weiterhin sei $A := \max_{(x,y)\in Q} |f(x,y)|$ sowie $\alpha := \min\{a, \frac{b}{A}\}$. Dann existiert mindestens eine Lösung zu dem AWP $y' = f(x,y)$ mit $y(\xi) = \eta$ im Intervall $J = [\xi, \xi + \alpha]$.

Beweis. Sei $\alpha > \tilde{\alpha}$ und $z_{\tilde{\alpha}}$ wie im Beweis von Satz 3.16. Dann ist $z_{\tilde{\alpha}}$ wohldefiniert, denn für $\xi \le x \le \xi + \tilde{\alpha}$ ist $t - \tilde{\alpha} \le \xi$ für $t \in [\xi, x]$ und $z_{\tilde{\alpha}}(t - \tilde{\alpha}) = \eta$. Somit ist $(t, z_{\tilde{\alpha}}(t-\tilde{\alpha})) \in Q$ und damit

$$z_{\tilde{\alpha}}(x) = \eta + \int_\xi^x f(t, \eta)\,dt$$

für $\xi \le x \le \xi + \tilde{\alpha}$ wohldefiniert. Überdies ist für diese x auch

$$|z_{\tilde{\alpha}}(x) - \eta| \le \int_\xi^x |f(t,\eta)|\,dt \le A(x-\xi) \le A\alpha \le b.$$

Also ist $(x, z_{\tilde{\alpha}}(x)) \in Q$ für $\xi \le x \le \xi + \tilde{\alpha}$. Analog folgen die restlichen x bis nach endlich vielen Schritten gezeigt wurde, dass $|z_{\tilde{\alpha}}(x) - \eta| \le b$ für alle $x \in J$. Analog zum Beweis von Satz 3.16 folgt die Behauptung. □

Korollar 3.18
Sei $D \subset \mathbb{R}^2$ ein Gebiet und $f : D \to \mathbb{R}$ stetig, dann existiert für jedes $(\xi, \eta) \in D$ mindestens eine (lokale) Lösung zum AWP $y' = f(x,y)$ mit $y(\xi) = \eta$.

3.4 Abhängigkeit der Lösung von den Daten

Lemma 3.19
Seien $a, b > 0$, $(\xi, \eta) \in \mathbb{R}^2$ und $Q = \{(x,y) \in \mathbb{R}^2 : |x-\xi| \le a, |y-\eta| \le b\}$. Weiterhin sei $f : Q \to \mathbb{R}$ stetig und Lipschitz-stetig bzgl. y mit Lipschitz-Konstante L. Darüber hinaus sein $A := \max_{(x,y)\in Q} |f(x,y)|$, $\alpha = \min\{a, \frac{b}{A}\}$ und $J = [\xi - \alpha, \xi + \alpha]$. Sei $\phi_0 \in C(J)$ mit $\max_{x\in J} |\phi_0(x) - \eta| = \mu \le b$ gegeben und

$$\phi_{n+1}(x) = \eta + \int_\xi^x f(t, \phi_n(t))\,dt$$

die Picard-Iterieration. Dann gilt die a priori Fehlerabschätzung

$$|y(x) - \phi_n(x)| \le \frac{A}{L} \sum_{i=n+1}^{\infty} \frac{[L(x-\xi)]^i}{i!} + \mu \sum_{i=n}^{\infty} \frac{[L(x-\xi)]^i}{i!} \le \left(\frac{A}{L} \frac{[L(x-\xi)]^{n+1}}{(n+1)!} + \mu \frac{[L(x-\xi)]^n}{n!} \right) e^{L(x-\xi)}$$

uniform für alle $x \in J$.

Restterm

Abschätzung

mit $e^{L(x-\xi)}$ Faktor

Beweis. Sei o.B.d.A. $x \geq \xi$. Dann gilt

$$|\phi_1(x) - \phi_0(x)| = \left|\eta - \phi_0(x) + \int_\xi^x f(t, \phi_0(t))\, dt\right| \leq |\eta - \phi_0(x)| + \int_\xi^x |f(t, \phi_0(t))|\, dt \leq \mu + \int_\xi^x A\, dt$$
$$= \mu + A(x - \xi).$$

Analog erhalten wir

$$|\phi_2(x) - \phi_1(x)| = \left|\int_\xi^x f(t, \phi_1(t)) - f(t, \phi_0(t))\, dt\right| \leq L \int_\xi^x |\phi_1(t) - \phi_0(t)|\, dt \leq L \int_\xi^x \mu + A(t - \xi)\, dt$$
$$= L\left(\mu(x - \xi) + A\frac{(x - \xi)^2}{2}\right)$$

und

$$|\phi_3(x) - \phi_2(x)| \leq L \int_\xi^x |\phi_2(t) - \phi_1(t)|\, dt \leq L^2 \int_\xi^x \mu(t - \xi) + A\frac{(t - \xi)^2}{2}\, dt$$
$$= L^2\left(\mu\frac{(x - \xi)^2}{2} + A\frac{(x - \xi)^3}{2 \cdot 3}\right).$$

Somit erhalten wir per Induktion

$$|\phi_n(x) - \phi_{n-1}(x)| \leq L \int_\xi^x |\phi_{n-1}(t) - \phi_{n-2}(t)|\, dt \leq L^{n-1} \int_\xi^x \mu\frac{(t - \xi)^{n-2}}{(n-2)!} + A\frac{(t - \xi)^{n-1}}{(n-1)!}\, dt$$
$$= L^{n-1}\left(\mu\frac{(x - \xi)^{n-1}}{(n-1)!} + A\frac{(x - \xi)^n}{n!}\right).$$

Schlussendlich gilt für $p \geq 1$ mit der Teleskopsumme und der Dreiecksungleichung

$$|\phi_{n+p}(x) - \phi_n(x)| \leq \sum_{i=1}^{p} |\phi_{n+i}(x) - \phi_{n+i-1}(x)|$$
$$\leq \sum_{i=1}^{p} L^{n+i-1}\left[\mu\frac{(x - \xi)^{n+i-1}}{(n+i-1)!} + A\frac{(x - \xi)^{n+i}}{(n+i)!}\right]$$
$$= \frac{A}{L}\sum_{i=1}^{p} \frac{[L(x - \xi)]^{n+i}}{(n+i)!} + \mu\sum_{i=1}^{p} \frac{[L(x - \xi)]^{n+i-1}}{(n+i-1)!}$$
$$= \frac{A}{L}\sum_{i=n+1}^{n+p} \frac{[L(x - \xi)]^{i}}{i!} + \mu\sum_{i=n}^{n+p-1} \frac{[L(x - \xi)]^{i}}{i!}.$$

Nach Picard-Lindelöf (genauer dem Banachschen Fixpunktsatz) gilt $y(x) = \lim_{n\to\infty} \phi_n(x)$ und somit

$$|y(x) - \phi_n(x)| = \lim_{p\to\infty} |\phi_{n+p}(x) - \phi_n(x)| \leq \frac{A}{L}\sum_{i=n+1}^{\infty} \frac{[L(x - \xi)]^{i}}{i!} + \mu\sum_{i=n}^{\infty} \frac{[L(x - \xi)]^{i}}{i!}.$$

Wegen $(i+1)(i+2)\cdots(i+n) \geq n!$ und

$$\sum_{i=n}^{\infty} \frac{x^i}{i!} = \sum_{i=0}^{\infty} \frac{x^{i+n}}{(i+n)!} = \sum_{i=0}^{\infty} \frac{x^i x^n}{i!(i+1)(i+2)\cdots(i+n)} \leq \frac{x^n}{n!}\sum_{i=0}^{\infty} \frac{x^i}{i!} = \frac{x^n}{n!}e^x$$

erhalten wir die übersichtlichere aber schwächere Aussage

$$|y(x) - \phi_n(x)| \leq \left(\frac{A}{L}\frac{[L(x - \xi)]^{n+1}}{(n+1)!} + \mu\frac{[L(x - \xi)]^{n}}{n!}\right)e^{L(x-\xi)}.$$

□

Satz 3.20 (Stetige Abhängigkeit von den Daten)
Es gelten die Voraussetzungen des Picard-Lindelöf-Satzes. Weiterhin sei $\tilde{J} \subseteq J$ abgeschlossen, $\tilde{f} : Q \to \mathbb{R}$ nur stetig, und $\tilde{y} : \tilde{J} \to \mathbb{R}$ eine Lösung von $y' = \tilde{f}(x, y)$ mit $y(\xi) = \tilde{\eta}$. Ist $|\tilde{\eta} - \eta| \leq \sigma \leq b$ und

$$|f(x, y) - \tilde{f}(x, y)| \leq \omega \quad \forall (x, y) \in Q,$$

so gilt uniform für alle $x \in \tilde{J}$

$$|y(x) - \tilde{y}(x)| \leq \sigma e^{L|x-\xi|} + \frac{\omega}{L}(e^{L|x-\xi|} - 1).$$

Beweis. Sei o.B.d.A. $x \geq \xi$. Sei ϕ_0 eine stetige, in Q bleibende Fortsetzung von $\tilde{y}$ auf ganz J, z.B.

$$\phi_0(x) = \begin{cases} \tilde{y}(x), & x \in \tilde{J} := [x_1, x_2] \\ \tilde{y}(x_1), & x \leq x_1 \\ \tilde{y}(x_2), & x \geq x_2 \end{cases}.$$

Für die Picard-Iteration

$$\phi_{n+1}(x) = \eta + \int_\xi^x f(t, \phi_n(t))\, dt$$

gilt $\lim_{n\to\infty} \phi_n(x) = y(x)$ für alle $x \in J$. Für $x \in \tilde{J} \subset J$ gilt somit

$$\phi_1(x) = \eta + \int_\xi^x f(t, \phi_0(t))\, dt = \eta + \int_\xi^x f(t, \tilde{y}(t))\, dt$$

nach Konstruktion von ϕ_0. Weil $\phi_0 = \tilde{y}$ auf $\tilde{J}$ ist gilt die Integralgleichung

$$\phi_0(x) = \tilde{y}(x) = \tilde{\eta} + \int_\xi^x \tilde{f}(t, \tilde{y}(t))\, dt.$$

Subtraktion dieser beiden Gleichungen liefert

$$|\phi_1(x) - \phi_0(x)| \leq |\eta - \tilde{\eta}| + \int_\xi^x |f(t, \tilde{y}(t)) - \tilde{f}(t, \tilde{y}(t))|\, dt \leq \sigma + \omega(x - \xi).$$

Analog zu Lemma 3.19 mit $n = 0$ folgt jetzt für alle $x \in \tilde{J}$

$$|y(x) - \phi_0(x)| \leq \frac{\omega}{L} \sum_{i=1}^{\infty} \frac{[L(x-\xi)]^i}{i!} + \sigma \sum_{i=0}^{\infty} \frac{[L(x-\xi)]^i}{i!} = \frac{\omega}{L}(e^{L(x-\xi)} - 1) + \sigma e^{L(x-\xi)}.$$

□

Die Lösung hängt nicht nur stetig, sondern häufig auch differenzierbar von den Daten ab.

Beispiel 3.21
Sei

$$\frac{\partial}{\partial x} y(x, \lambda) = f(x, y, \lambda) = \lambda y, \quad y(0, \lambda) = \eta.$$

Offensichtlich gilt

$$y(x, \lambda) = \eta e^{\lambda x}, \qquad \frac{\partial y}{\partial \eta} = e^{\lambda x}, \qquad \frac{\partial y}{\partial \lambda} = \eta x e^{\lambda x}.$$

Satz 3.22 (Differenzierbare Abhängigkeit)
Seien $a, b > 0$ und $Q = \{(x, y) \in \mathbb{R}^2 : |x - \xi| \leq a, |y - \eta| \leq b\}$. Weiterhin sei $\lambda \in I := (c, d)$ und $f : Q \times I \to \mathbb{R}$ stetig mit $\frac{\partial f}{\partial y} =: f_y$ und $\frac{\partial f}{\partial \lambda} =: f_\lambda$ ebenfalls stetig. Sei $A := \max_{(x,y,\lambda)\in Q\times I} |f(x, y, \lambda)|$, $\alpha := \min\{a, \frac{b}{A}\}$, $J := [\xi - \alpha, \xi + \alpha]$ und $y' = f(x, y, \lambda)$ mit $y(\xi, \lambda) = \eta$. Dann ist $y = y(x, \lambda)$ auf ganz J partiell nach λ differenzierbar.

Beweis. Sei $y = y(x, \lambda)$, $\tilde{y} = y(x, \tilde{\lambda})$ mit $y(\xi) = \eta = \tilde{y}(\xi)$ für λ, $\tilde{\lambda} \in I$. Dann gilt mit der Taylorreihenentwicklung von f

$$\begin{aligned}\frac{d}{dx}(\tilde{y} - y) &= f(x, \tilde{y}, \tilde{\lambda}) - f(x, y, \lambda)\\ &= f(x, \tilde{y}, \tilde{\lambda}) - f(x, y, \tilde{\lambda}) + f(x, y, \tilde{\lambda}) - f(x, y, \lambda)\\ &= f_y(x, y + \nu_1(\tilde{y} - y), \tilde{\lambda})(\tilde{y} - y) + f_\lambda(x, y, \lambda + \nu_2(\tilde{\lambda} - \lambda))(\tilde{\lambda} - \lambda)\end{aligned}$$

mit $0 < \nu_1, \nu_2 < 1$. Folglich haben wir für den Differenzenquotienten

$$\frac{d}{dx}\frac{\tilde{y} - y}{\tilde{\lambda} - \lambda} = f_y(x, y + \nu_1(\tilde{y} - y), \tilde{\lambda})\frac{\tilde{y} - y}{\tilde{\lambda} - \lambda} + f_\lambda(x, y, \lambda + \nu_2(\tilde{\lambda} - \lambda)) =: \tilde{g}(x)\frac{\tilde{y} - y}{\tilde{\lambda} - \lambda} + \tilde{h}(x).$$

Insbesondere sind $\tilde{g}(x)$ und $\tilde{h}(x)$ stetig in J, so dass

$$\tilde{z} := \frac{\tilde{y} - y}{\tilde{\lambda} - \lambda}$$

die eindeutige Lösung der linearen DGL 1. Ordnung

$$\frac{d\tilde{z}}{dx} = \tilde{g}(x)\tilde{z} + \tilde{h}(x), \quad \tilde{z}(\xi) = 0$$

ist. Darüber hinaus sind $g(x) := f_y(x, y, \lambda)$ und $h(x) := f_\lambda(x, y, \lambda)$ stetig in J nach Voraussetzung, so dass die lineare DGL 1. Ordnung

$$\frac{dz}{dx} = g(x)z + h(x), \quad z(\xi) = 0$$

genau eine Lösung z hat. Nach Satz 3.20 folgt $\tilde{y} \to y$ für $\tilde{\lambda} \to \lambda$, weil f in $Q \times I$ stetig ist. Damit erhalten wir, dass $\tilde{g} \to g$ und $\tilde{h} \to h$ für $\tilde{\lambda} \to \lambda$. Somit liefert eine erneutes Anwenden des Satzes 3.20, dass $\tilde{z} \to z$. Somit gilt

$$\frac{\partial}{\partial\lambda}y = \lim_{\tilde{\lambda}\to\lambda}\frac{\tilde{y} - y}{\tilde{\lambda} - \lambda} = \lim_{\tilde{\lambda}\to\lambda}\tilde{z} = z.$$

□

Numerik für DGLs (wenn Kochrezept nicht anwendbar)

In diesem Kapitel werden erste numerische Verfahren zum näherungsweise Lösen des AWP

$$y'(x) = f(x, y(x)) \text{ für } x \in I = [a, b] \quad \text{und} \quad y(a) \text{ gegeben} \tag{4.1}$$

untersucht. Dabei ist $f : G \subseteq I \times \mathbb{R}^n \to \mathbb{R}^n$ stetig und Lipschitz-stetig bzgl. y.
Fundamental für die klassische Analysis numerischer Verfahren für DGLs ist die Taylorreihenentwicklung. Ist $f : I \subset \mathbb{R} \to \mathbb{R}$, so heißt

$$T_n f(x; a) = \sum_{i=0}^{n} \frac{f^{(i)}(a)(x-a)^i}{i!} = f(a) + f'(a)(x-a) + \frac{f''(a)(x-a)^2}{2} + \frac{f'''(a)(x-a)^3}{6} + \cdots$$

für $f \in C^n(I)$ Taylorreihe von f mit Entwicklungspunkt $a \in I$. Die Größe

$$R_n f(x; a) = \frac{f^{(n+1)}(\xi)(x-a)^{n+1}}{(n+1)!} = f(x) - T_n f(x; a)$$

mit einem Zwischenpunkt $\xi = a + t(x-a)$ und $t \in [0, 1]$ heißt Lagrange-Restglied.
Für $f : \mathbb{R}^d \to \mathbb{R}$ setzen wir $F_{x,a} : \mathbb{R} \to \mathbb{R}$ mit $t \mapsto f(a + t(x-a))$. Insbesondere ist $F_{x,a}(0) = f(a)$ und $F_{x,a}(1) = f(x)$. Die Taylorreihenentwicklung von f mit Entwicklungspunkt a ist definiert über die Taylorreihe von $F_{x,a}(t)$ mit Entwicklungspunkt Null und Auswertungspunkt Eins. Sei $\alpha = (\alpha_1, \cdots, \alpha_d)$ mit $\alpha_i \geq 0$ ein Multiindex. Wir schreiben

$$|\alpha| = \sum_{i=1}^{d} \alpha_i, \quad \alpha! = \prod_{i=1}^{d} \alpha_i!, \quad x^\alpha = x_1^{\alpha_1} \cdots x_d^{\alpha_d}, \quad D^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}.$$

So ist

$$\begin{aligned} T_n f(x; a) = T_n F_{x,a}(1; 0) &= \sum_{i=1}^{n} \frac{F_{x,a}^{(i)}(0)}{i!} = \sum_{0 \leq |\alpha| \leq n} \frac{(x-a)^\alpha}{\alpha!} D^\alpha f(a) \\ &= f(a) + \sum_{i=1}^{d} (x_i - a_i) \partial_{x_i} f(a) + \sum_{i=1}^{d} \sum_{j=1}^{d} \frac{1}{2} (x_i - a_i)(x_j - a_j) \partial_{x_i} \partial_{x_j} f(a) + \cdots \\ &= f(a) + \nabla f(a)^\top (x-a) + \frac{1}{2} (x-a)^\top \nabla^2 f(a)(x-a) + \cdots \end{aligned}$$

die Taylorreihe, und

$$R_n f(x, a) = \sum_{|\alpha| = n+1} \frac{(x-a)^\alpha}{\alpha!} D^\alpha f(a + \vartheta(x-a))$$

mit $\vartheta \in [0, 1]$ das Lagrange-Restglied.
Ist die Funktion $f \in C^{n+1}$, so gilt für das Restglied, und somit für den Taylorreihenfehler

$$\|f(x) - T_n f(x; a)\| = \|R_n f(x, a)\| = \mathcal{O}(\|x - a\|^{n+1})$$

wobei

$$\mathcal{O}(f) := \{g : \mathbb{R}_+ \to \mathbb{R}_+ \mid \exists C, h_0 > 0 \; \forall h \le h_0 : g(h) \le Cf(h)\}.$$

Taylorreihenentwicklungen von vektorwertigen Funktionen erfolgt einfach durch komponentenweise Taylorreihenentwicklung.

Beispiel 4.1
Sei $y : \mathbb{R} \to \mathbb{R}$ die Lösung von $y'(x) = f(x, y(x))$, sowie y und f hinreichend glatt. So ist

$$y''(x) = \frac{d}{dx} f(x, y(x)) = f_x(x, y(x)) + f_y(x, y(x))y'(x) = f_x(x, y(x)) + f_y(x, y(x))f(x, y(x))$$

mit f_x und f_y die partiellen Ableitungen von f. Für die dritte Ableitung gilt

$$\begin{aligned} y'''(x) &= \frac{d}{dx}\left(f_x(x, y(x)) + f_y(x, y(x))f(x, y(x))\right) \\ &= f_{xx}(x, y(x)) + f_{yx}(x, y(x))y'(x) + \left(f_{xy}(x, y(x)) + f_{yy}(x, y(x))y'(x)\right) f(x, y(x)) + f_{yy}(x, y(x))y''(x) \\ &= f_{xx}(x, y(x)) + 2f_{xy}(x, y(x))f(x, y(x)) + f_{yy}(x, y(x))f^2(x, y(x)) f_y(x, y(x)) f_x(x, y(x)) \\ &\quad + f_y^2(x, y(x))f(x, y(x)). \end{aligned}$$

Damit erhalten wir die Taylorreihenentwicklung für y mit Entwicklungspunkt x ausgewertet in $x + h$

$$\begin{aligned} y(x + h) &= y(x) + hy'(x) + \frac{h^2}{2} y''(x) + \mathcal{O}(h^3) \\ &= y(x) + hf(x, y(x)) + \frac{h^2}{2}\left(f_x(x, y(x)) + f_y(x, y(x))f(x, y(x))\right) + \mathcal{O}(h^3). \end{aligned}$$

4.1 Explizite Einschrittverfahren

Idee: x-Achse in kleine Stücke teilen, und Fkt.-Wert durch Steigung im letzten Abschnitt abschätzen

4.1.1 Explizites Euler-Verfahren

Das explizite Euler-Verfahren, auch Polygonzugverfahren genannt, zum Lösen von (4.1) ist

$$y_0 = y(x_0), \qquad y_{i+1} = y_i + h_i f(x_i, y_i), \quad 0 \le i \le n - 1 \tag{4.2}$$

mit $x_0 = a$ und $x_{i+1} = x_i + h_i$. Hierbei gilt für die lokale Schrittweite $h_i > 0$, dass $\sum_{i=0}^{n-1} h_i = |I|$, also $x_n = b$. Die Größe $y_i \in \mathbb{R}^n$ approximiert dabei den exakten Wert $y(x_i)$. Eine stückweise lineare Verbindung der $n+1$ Punkte (x_i, y_i) liefert einen Polygonzug als Approximation von y.
Es gibt mindestens drei Interpretationsmöglichkeiten des expliziten Euler-Verfahrens.

1. **Numerisches Differenzieren:** Ignorieren des Grenzwertprozessen in der Ableitung liefert den Vorwärtsdifferenzenquotienten

$$y'(x_i) := \lim_{\delta \to 0} \frac{y(x_i + \delta) - y(x_i)}{\delta} \approx \frac{y(x_i + h_i) - y(x_i)}{h_i} = \frac{y(x_{i+1}) - y(x_i)}{h_i}$$

für eine Schrittweite $h_i > 0$. Wird die DGL (4.1) nur in den Knoten x_i betrachtet und die Ableitung durch den Vorwärtsdifferenzenquotienten ersetzt, so erhalten wir das explizite Euler-Verfahren

$$f(x_i, y(x_i)) = y'(x_i) \approx \frac{y(x_{i+1}) - y(x_i)}{h_i} \quad \Rightarrow \quad f(x_i, y_i) = \frac{y_{i+1} - y_i}{h_i}. \quad \Rightarrow y_{i+1} = y_i + h_i f(x_i, y_i)$$

2. **Numerisches Integrieren:** Approximieren der Flächen $\int_{x_i}^{x_{i+1}} f(t, y(t))\, dt$ durch ein Rechteck in der zu (4.1) äquivalenten Integralgleichung liefert

$$y(x_{i+1}) = y(x_i) + \int_{x_i}^{x_{i+1}} f(t, y(t))\, dt \approx y(x_i) + \int_{x_i}^{x_{i+1}} f(x_i, y(x_i))\, dt = y(x_i) + \underbrace{(x_{i+1} - x_i)}_{h_i} f(x_i, y(x_i)).$$

und damit das explizite Euler-Verfahren

$$y_{i+1} = y_i + h_i f(x_i, y_i).$$

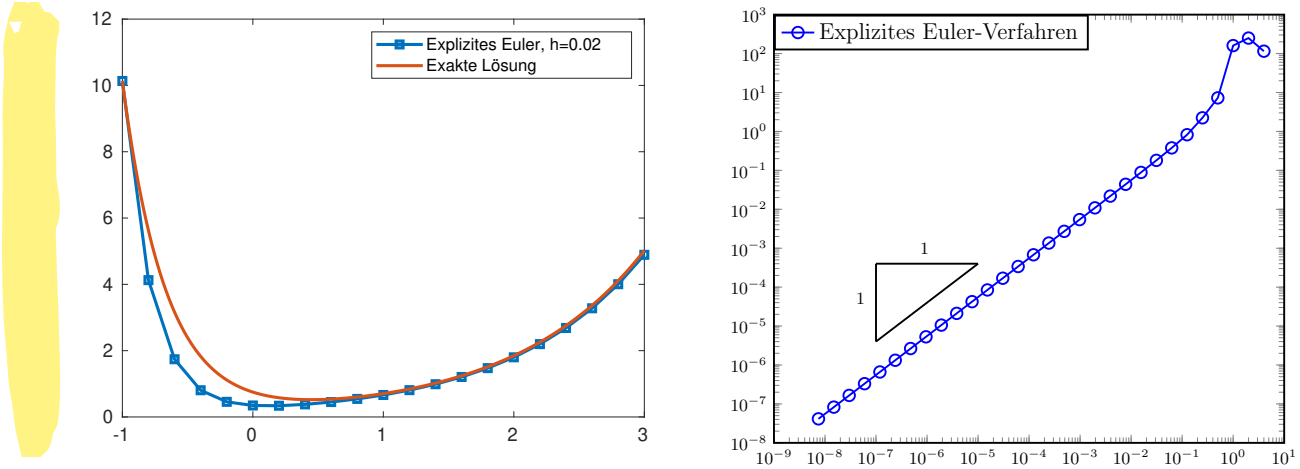
3. **Taylorreihenentwicklung:** Taylorreihenentwicklung der exakten Lösung mit Vernachlässigung des Restterms liefert

$$y(x_i + h_i) = y(x_i) + h_i y'(x_i) + \frac{h_i^2}{2} y''(\xi_i) = y(x_i) + h_i f(x_i, y(x_i)) + \frac{h_i^2}{2} y''(\xi_i) \approx y(x_i) + h_i f(x_i, y(x_i))$$

für ein Zwischenpunkt $\xi_i \in [x_i, x_i + h_i]$. Damit folgt erneut das explizite Euler-Verfahren

$$y_{i+1} = y_i + h_i f(x_i, y_i).$$

Beispiel 4.2
Sei $y' = -3y + e^x$ mit $y(-1) = 0.25e^{-1} + 0.5e^3$ und exakter Lösung $y(x) = 0.25e^x + 0.5e^{-3x}$. Sei $I = [-1, 3]$ und $h_i \equiv h$, also $x_i = -1 + ih$ mit $i = 0, 1, \cdots, n$ wobei $n = |I|/h = 4h^{-1}$. Es gilt $y_0 = y(-1)$ und $y_{i+1} = y_i + hf(x_i, y_i) = (1 - 3h)y_i + he^{x_i}$.



(a) Lösungen (b) Fehlerplot

Abb. 4.1: Euler-Lösung zu $h = 0.2$ und exakte Lösung (links). $\max_i |y(x_i) - y_i|$ (y-Achse) gegen Schrittweite h (x-Achse) doppelt-logarithmisch abgetragen (rechts).

Sei

$$\|\epsilon(h)\| = C_0 + C_1 h^\alpha + C_2 h^{2\alpha} + \dots$$

eine Potenzreihenentwicklung des Fehlers $\epsilon(h)$. Weil $\|\epsilon(h)\| \to 0$ für $h \to 0^+$ bei einem konvergenten Verfahren sind $C_0 = 0$ und $\alpha > 0$. Somit sind für hinreichend kleines h die Terme $|C_2 h^{2\alpha} + \dots| \ll |C_1 h^\alpha|$ von höherer Ordnung. Also gilt

$$\|\epsilon(h)\| \approx C_1 h^\alpha.$$

Sind die Fehler $\|\epsilon_i\| := \|\epsilon(h_i)\|$ und $\|\epsilon_{i+1}\|$ zu den zwei Schrittweiten h_i und h_{i+1} bekannt, so lässt sich der Exponent α, die Konvergenzrate, experimentell ermitteln. Division der beiden Gleichungen

$$\|\epsilon_i\| \approx C_1 h_i^\alpha \quad \text{und} \quad \|\epsilon_{i+1}\| \approx C_1 h_{i+1}^\alpha$$

liefert

$$\frac{\|\epsilon_{i+1}\|}{\|\epsilon_i\|} \approx \left(\frac{h_{i+1}}{h_i}\right)^\alpha \quad \Leftrightarrow \quad \alpha \approx \frac{\log \frac{\|\epsilon_{i+1}\|}{\|\epsilon_i\|}}{\log \frac{h_{i+1}}{h_i}}.$$

Lemma 4.1. *Für alle $x \in \mathbb{R}$ gilt $1 + x \le e^x$, und für alle $x \ge -1$ gilt $0 \le (1 + x)^m \le e^{mx}$ mit $m \in \mathbb{N}_0$ beliebig.*

Beweis. Die Taylorreihenentwicklung liefert $e^x = 1 + x + 0.5x^2 e^{\xi x} \ge 1 + x$ mit $\xi \in [0, 1]$. Für $x \ge -1$ ist $1 + x \ge 0$. Die Monotonie von x^m liefert sofort die Behauptung. □

Satz 4.3
Sei $y \in C^2(I)$ die Lösung von (4.1) und y_i die Euler-Approximation zu $y(x_i)$ mit den äquidistanten Knoten $a = x_0 < x_1 < \cdots < x_n = b$, d.h. $h_i \equiv h$. Dann gilt

$$\max_{0\le i\le n} \|y(x_i) - y_i\| \le e^{L|I|}\|y(x_0) - y_0\| + \frac{e^{L|I|} - 1}{2L}\|y''\|_{\infty,I} \cdot h.$$

Wenn $\|y(x_0) - y_0\| = \mathcal{O}(h)$, dann ist $\max_{0\le i\le n} \|y(x_i) - y_i\| = \mathcal{O}(h)$.

Beweis. Sei $\epsilon_i = y(x_i) - y_i$ der Fehler in x_i. Taylorreihenentwicklung für y liefert

$$y(x_{i+1}) = y(x_i) + h_i f(x_i, y(x_i)) + \frac{h_i^2}{2} y''(\xi_i)$$

mit $\xi_i \in [x_i, x_{i+1}] =: I_i$, wohingegen das explizite Euler-Verfahren bereits

$$y_{i+1} = y_i + h_i f(x_i, y_i)$$

ist. Subtraktion dieser beiden Gleichungen ergibt

$$\epsilon_{i+1} = \epsilon_i + h_i \left[f(x_i, y(x_i)) - f(x_i, y_i)\right] + \frac{h_i^2}{2} y''(\xi_i).$$

Folglich ist mit Dreiecksungleichung und der Lipschitz-Stetigkeit von f (es sei L_i die Lipschitz-Konstante von f bzgl. y für $x \in I_i$)

$$\begin{aligned}
\|\epsilon_{i+1}\| &\le \|\epsilon_i\| + h_i\|f(x_i, y(x_i)) - f(x_i, y_i)\| + \frac{h_i^2}{2}\|y''(\xi_i)\| \\
&\le \|\epsilon_i\| + h_i L_i \|\epsilon_i\| + \frac{h_i^2}{2}\|y''(\xi_i)\| \\
&\le (1 + h_i L_i)\|\epsilon_i\| + \frac{h_i^2}{2}\|y''\|_{\infty,I_i}.
\end{aligned}$$

Rekursion liefert

$$\|\epsilon_i\| \le \|\epsilon_0\| \prod_{j=0}^{i-1}(1 + h_j L_j) + \sum_{j=0}^{i-1} \frac{h_j^2}{2}\|y''\|_{\infty,I_j} \prod_{k=j+1}^{i-1}(1 + h_k L_k). \tag{4.3}$$

Speziell für $h_i \equiv h$ und $L = \max_j L_j$ die globale Lipschitz-Konstante folgt zusammen mit der geometrischen Reihe, dass

$$\begin{aligned}
\|\epsilon_i\| &\le (1 + hL)^i\|\epsilon_0\| + \left[1 + (1 + hL) + \cdots + (1 + hL)^{i-1}\right] \frac{h^2}{2}\|y''\|_{\infty,I} \\
&= (1 + hL)^i\|\epsilon_0\| + \left[\frac{(1 + hL)^i - 1}{hL}\right] \frac{h^2}{2}\|y''\|_{\infty,I} \\
&\le e^{ihL}\|\epsilon_0\| + \left[\frac{e^{ihL} - 1}{L}\right] \frac{h}{2}\|y''\|_{\infty,I}.
\end{aligned}$$

Mit $nh = |I|$ gilt

$$\max_{0\le i\le n} \|\epsilon_i\| \le e^{L|I|}\|\epsilon_0\| + \frac{e^{L|I|} - 1}{L} \frac{\|y''\|_{\infty,I}}{2} h.$$

Die zweite Aussage folgt sofort aus der Ersten. □

Obiger Satz besagt, dass sich die obere Schranke des Fehlers bei Halbierung der Schrittweite h ebenfalls halbiert. D.h. das explizite Euler-Verfahren konvergiert "nur" linear.
Die rechte Seite von (4.3) lässt sich auch unter beibehalten der lokalen Schrittweiteninformation abschätzen. Die Fehlerkonstante wird bei dieser groben Abschätzung jedoch größer als im obigen Satz. Es gilt

$$\begin{aligned}\|\epsilon_i\| &\le \|\epsilon_0\| \prod_{j=0}^{i-1}(1+h_jL_j) + \sum_{j=0}^{i-1}\frac{h_j^2}{2}\|y''\|_{\infty,I_j}\prod_{k=j+1}^{i-1}(1+h_kL_k)\\ &\le \|\epsilon_0\| \prod_{j=0}^{i-1} e^{h_jL_j} + \sum_{j=0}^{i-1} h_j \prod_{k=j+1}^{i-1} e^{h_kL_k} \max_{0\le j\le i-1}\frac{h_j}{2}\|y''\|_{\infty,I_j}\\ &\le \|\epsilon_0\| e^{\sum_{j=0}^{i-1}h_jL_j} + \sum_{j=0}^{i-1} h_j e^{\sum_{k=j+1}^{i-1}h_kL_k} \max_{0\le j\le i-1}\frac{h_j}{2}\|y''\|_{\infty,I_j}\\ &\le e^{L|I|}\|\epsilon_0\| + |I|e^{L|I|}\max_j \frac{h_j}{2}\|y''\|_{\infty,I_j}.\end{aligned}$$

Wir sehen, dass es geschickter ist dort kleine Schrittweiten h_j zu nehmen wo die zweite Ableitung von y betragsmäßig groß ist, und große Schrittweiten wo y fast linear ist, als überall die gleiche uniforme Schrittweite h.

Logisch

4.1.2 Allgemeines explizites Runge-Kutta-Verfahren 2. Ordnung

Sei

$$y_{i+1} = y_i + ak_1 + bk_2 \quad \text{mit} \quad k_1 = h_i f(x_i, y_i),\ k_2 = h_i f(x_i+\alpha h_i, y_i+\beta k_1) \tag{4.4}$$

für gegebene $a, b, \alpha, \beta \in \mathbb{R}$. Damit (4.4) von zweiter Ordnung ist, muss die Taylorreihenentwicklung der exakten Lösung y mit der Taylorreihenentwicklung von (4.4) bis zur Ordnung zwei übereinstimmen. Taylorreihenentwicklung von y liefert mit $y' = f(x,y)$

$$\begin{aligned}y(x_{i+1}) &= y(x_i) + h_i y'(x_i) + \frac{h_i^2}{2}y''(x_i) + \frac{h_i^3}{6}y'''(x_i) + \mathcal{O}(h_i^4)\\ &= y(x_i) + h_i f(x_i, y(x_i)) + \frac{h_i^2}{2}(f_x + f_y f) + \frac{h_i^3}{6}\left(f_{xx} + 2f_{xy}f + f_{yy}f^2 + f_x f_y + f_y^2 f\right) + \mathcal{O}(h_i^4)\end{aligned} \tag{4.5}$$

wie „Interpretation Taylor" oben

wobei $f_x = \partial_x f(x_i, y(x_i))$ und f_y, f_{xx}, f_{xy} und f_{yy} analog definiert sind. Ebenso liefert die Taylorreihenentwicklung

$$f(x_i+\alpha h_i, y_i+\beta k_1) = f(x_i,y_i) + \alpha h_i f_x + \beta k_1 f_y + \frac{\alpha^2 h_i^2}{2}f_{xx} + \alpha h_i \beta k_1 f_{xy} + \frac{\beta^2 k_1^2}{2}f_{yy} + \mathcal{O}(h_i^3).$$

$k_1 = h_i \cdot f$

Einsetzen dieser Gleichung in (4.4) liefert

(für k_2)

$$y_{i+1} = y_i + (a+b)h_i f + bh_i^2(\alpha f_x + \beta f_y f) + bh_i^3\left(\frac{\alpha^2}{2}f_{xx} + \alpha\beta f f_{xy} + \frac{\beta^2}{2}f^2 f_{yy}\right) + \mathcal{O}(h_i^4) \tag{4.6}$$

wobei f, f_x, f_y, f_{xx}, f_{xy} und f_{yy} in (x_i, y_i) auszuwerten sind. Vergleichen der Koeffizienten zu gleichen Potenzen von h_i mit (4.5) liefert

$$a+b=1, \quad b\alpha = b\beta = \frac{1}{2}.$$

Wegen dem Term $6^{-1}h_i^3 f_x f_y$ in (4.5) können die Koeffizienten zu h_i^3 von (4.5) und (4.6) im Allgemeinen nicht übereinstimmen. Wir haben drei Gleichungen bei vier Unbekannte. Es gibt somit unendlich viele Runge-Kutta-Verfahren 2. Ordnung. Typische Koeffizienten sind $a = b = 1/2$ und $\alpha = \beta = 1$.

4.1.3 Allgemeine explizite Runge-Kutta-Verfahren

Die Idee des expliziten Runge-Kutta-Verfahrens 2. Ordnung lässt sich leicht für höhere Ordnungen verallgemeinern. Der Nachweis, dass es sich dabei tatschächlich um ein Verfahren höherer Ordnung handelt erfolgt analog zum vorherigen Abschnitt.
Seien $a_{jl} \in \mathbb{R}$ mit $1 \le l \le s-1$, $l+1 \le j \le s$, und $b_j \in \mathbb{R}$ $(1 \le j \le s)$ und $c_j \in [0,1]$ $(2 \le j \le s)$ gegeben. Das s-stufige explizite Runge-Kutta-Verfahren ist

$$y_{i+1} = y_i + h_i \sum_{j=1}^{s} b_j k_j \quad \text{mit} \quad k_j = f(x_i + c_j h_i, y_i + h_i \sum_{l=1}^{j-1} a_{jl} k_l).$$

Neben den Runge-Kutta-Verfahren 2. Ordnung wir häufig auch das klassisches Runge-Kutta-Verfahren 4. Ordnung

$$y_{i+1} = y_i + \frac{k_0 + 2k_1 + 2k_2 + k_3}{6}$$

mit

$$k_0 = h_i f(x_i, y_i), \quad k_1 = h_i f(x_i + \frac{h_i}{2}, y_i + \frac{k_0}{2}), \quad k_2 = h_i f(x_i + \frac{h_i}{2}, y_i + \frac{k_1}{2}), \quad k_3 = h_i f(x_i + h_i, y_i + k_2)$$

eingesetzt.

Beispiel 4.4
Sei $y' = -3y + e^x$ mit $y(-1) = 0.25e^{-1} + 0.5e^3$ und exakter Lösung $y(x) = 0.25e^x + 0.5e^{-3x}$. Sei $I = [-1, 3]$ und $h_i \equiv h$, also $x_i = -1 + ih$ mit $i = 0, 1, \cdots, n$ wobei $n = |I|/h = 4h^{-1}$. Im Bild 4.2 ist der Fehler $\max_i |y(x_i) - y_i|$ gegen die Schritteweite h im doppelt-logarithmischen Maßstab für das explizite Euler-Verfahren und jeweils ein explizites Runge-Kutta-Verfahren 2. und 4. Ordnung abgetragen. Die Konvergenzraten im asymptotischen Bereich, aber bevor die Rundungsfehler dominieren, sind neben den Steigungsdreicken eingetragen.

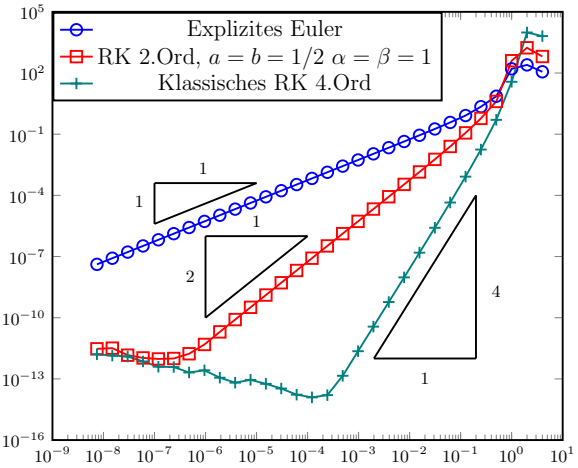


Abb. 4.2: $\max_i |y(x_i) - y_i|$ gegen Schrittweite h doppelt-logarithmisch abgetragen.

(wie schnell konvergiert)

Es ist leicht einzusehen, dass die erreichbare Konvergenzrate (Ordnung) α eines s-stufigen Runge-Kutta-Verfahrens maximal s ist. Genauer gelten die sogenannten Butcherschranken

s	1	2	3	4	5	6	7	8	9	$s \geq 9$
α	1	2	3	4	4	5	6	6	7	$\alpha \leq s - 2$

wobei die Schranke $\alpha \leq s - 2$ für $s \geq 9$ nicht scharf ist. So wird zur Zeit noch $s = 17$ für $\alpha = 10$ benötigt.

4.2 Konvergenz von allgemeinen expliziten Einschrittverfahren

Definition 4.5 (Methodenfunktion)
Das Gitter Δ_h bestehe aus den Knoten $a = x_0 < x_1 < \cdots < x_n = b$ mit $h_i = x_{i+1} - x_i$. Die Funktion $\Phi(\cdot, \cdot, \cdot) : G \times \mathbb{R}_{>0} \to \mathbb{R}^n$ in dem Einschrittverfahren

$$y_0 = y(x_0), \qquad y_{i+1} = y_i + h_i \Phi(x_i, y_i, h_i) \tag{4.7}$$

heißt Methodenfunktion bzw. Steigungsschätzer.

↳ (wir haben bis jetzt immer genaue Steigung verwendet)

Beispiel 4.6
Für das explizite Euler-Verfahren ist $\Phi(x, y, h) = f(x, y)$.

Definition 4.7 (Konvergenz)
Sei $y_h : \Delta_h \to \mathbb{R}^n$ eine Näherungslösung für das Anfangswertproblem (4.1). Der Ausdruck

$$\epsilon_h(x) := y(x) - y_h(x)$$

definiert für $x \in \Delta_h$ heißt globaler Fehler. y_h ist konvergent wenn

$$\|\epsilon_h\|_h := \max_{x_i \in \Delta_h} \|y(x_i) - y_h(x_i)\| \to 0 \quad \text{für } h := \max_i h_i \to 0.$$

(≙ größer glob. Fehler → 0 für h → 0)

Das Verfahren hat die Konvergenzordnung $p > 0$, wenn $\|\epsilon_h\|_h = \mathcal{O}(h^p)$.

Dabei ist zu beachten, dass die Anzahl an Schritten n, bzw. Knoten x_i, wegen $\sum_{i=0}^{n-1} h_i = |I|$ von der Schrittweite abhängt.

Definition 4.8 (Konsistenz)
Sei y die Lösung des Anfangswertproblems (4.1) und x_i, $x_{i+1} = x_i + h_i \in I$. Dann heißt

$$d(x_i, y(x_i), h_i) := \frac{y(x_i + h_i) - y(x_i) - h_i \Phi(x_i, y(x_i), h_i)}{h_i}$$

der lokale Diskretisierungsfehler in $(x_i, y(x_i))$. Gilt uniform für alle $x_i \in \Delta_h$

$$\|d(x_i, y(x_i), h_i)\| \to 0 \quad \text{für } h := \max_i h_i \to 0,$$

so ist das Verfahren konsistent. Das Verfahren hat die Konsistenzordnung $p > 0$, wenn $\max_{x_i \in \Delta_h} \|d(x_i, y(x_i), h_i)\| = \mathcal{O}(h^p)$.

Beispiel 4.9
Für das explizite Euler-Verfahren gilt mit $y' = f(x, y) = \Phi(x, y, h)$ und Taylorreihenentwicklung

$$d(x_i, y(x_i), h_i) = \frac{y(x_i + h_i) - y(x_i) - h_i \Phi(x_i, y_i, h_i)}{h_i} = \frac{y(x_i + h_i) - y(x_i)}{h_i} - y'(x_i)$$

wir kennen y ja nicht genau, darum Taylor

$$= \frac{y(x_i) + h_i y'(x_i) + 0.5 h_i^2 y''(\xi_i) - y(x_i)}{h_i} - y'(x_i) = \frac{h_i}{2} y''(\xi_i)$$

(Zwischenwertsatz Lagrange = statt >)

für einen Zwischenpunkt $\xi_i \in [x_i, x_{i+1}]$. Folglich ist $\|d(x_i, y(x_i), h_i)\| \le C h_i \le C h$ für alle $y \in C^2(I)$ und die Konsistenzordnung ist somit Eins.

$C = \frac{y''(\xi_i)}{2}$ ↳ h^1

↳ = 2. Abl. stetig, wichtig für Zwischenwertsatz

Lemma 4.10 (diskretes Gronwall-Lemma)
Seien $a_i, b_i, w_i \ge 0$ und $b_i \le b_{i+1}$ für $i = 0, 1, 2, \ldots$. Dann gilt

1. Ist $a_n < 1$ sowie $w_n \le b_n + \sum_{r=0}^{n} a_r w_r$, so ist
   $$w_n \le b_n e^{\sum_{r=0}^{n} \sigma_r a_r}$$
   mit $\sigma_r = (1 - a_r)^{-1} > 0$ für $n = 0, 1, 2, \ldots$.
2. Ist $w_n \le b_n + \sum_{r=0}^{n-1} a_r w_r$, so ist
   $$w_n \le b_n e^{\sum_{r=0}^{n-1} a_r}$$
   für $n = 0, 1, 2, \ldots$.

Beweis. "Teil 1.: Der Beweis erfolgt per Induktion. Für

$$s_n := b_n + \sum_{r=0}^{n} a_r w_r \ge w_n$$

gilt

$$s_n - s_{n-1} = a_n w_n + b_n - b_{n-1} \le a_n s_n + b_n - b_{n-1} \quad \Rightarrow \quad (1 - a_n) s_n \le s_{n-1} + b_n - b_{n-1}.$$

Insbesondere gilt für $n = 0$

$$s_0 = a_0 w_0 + b_0 \leq a_0 s_0 + b_0$$

und somit

$$s_0 \leq (1-a_0)^{-1} b_0 = \sigma_0 b_0 = (1+\sigma_0 a_0) b_0 \leq b_0 e^{\sigma_0 a_0}$$

nach Lemma 4.1. Gelte $s_{n-1} \leq b_{n-1} \exp(\sum_{r=0}^{n-1} \sigma_r a_r)$. Dann gilt

$$\begin{aligned}
s_n &\leq (1-a_n)^{-1}(s_{n-1} + b_n - b_{n-1}) \\
&= (1+\sigma_n a_n)(s_{n-1} + b_n - b_{n-1}) \\
&\leq e^{\sigma_n a_n} \left(b_{n-1} e^{\sum_{r=0}^{n-1} \sigma_r a_r} + b_n - b_{n-1} \right) \\
&= e^{\sigma_n a_n} \left(b_{n-1} \left(e^{\sum_{r=0}^{n-1} \sigma_r a_r} - 1 \right) + b_n \right) \\
&\leq e^{\sigma_n a_n} \left(b_n \left(e^{\sum_{r=0}^{n-1} \sigma_r a_r} - 1 \right) + b_n \right) \\
&= b_n e^{\sum_{r=0}^{n} \sigma_r a_r}.
\end{aligned}$$

Folglich ist

$$w_n \leq s_n \leq b_n e^{\sum_{r=0}^{n} \sigma_r a_r}.$$

"Teil 2.: Folgt analog. □

Lemma 4.11
Seien $a_i, b_i, c_i \geq 0$ und $a_i \leq b_{i-1} + (1+c_{i-1}) a_{i-1}$. Dann gilt

$$a_i \leq \left(a_0 + \sum_{r=0}^{i-1} b_r \right) e^{\sum_{r=0}^{i-1} c_r}.$$

Beweis. Aus $a_i \leq a_{i-1} + c_{i-1} a_{i-1} + b_{i-1}$ folgt rekursiv

$$a_i \leq \cdots \leq a_0 + \sum_{r=0}^{i-1} c_r a_r + \sum_{r=0}^{i-1} b_r = \left(a_0 + \sum_{r=0}^{i-1} b_r \right) + \sum_{r=0}^{i-1} c_r a_r.$$

Das diskrete Gronwall-Lemma 4.10 Teil 2. liefert jetzt die Behauptung. □

Satz 4.12
Gilt für das Einschrittverfahren (4.7)

1. die Methodenfunktion Φ ist Lipschitz-stetig bzgl. y mit Lipschitz-Konstante L_Φ
2. und das Verfahren ist konsistent,

so ist das Einschrittverfahren konvergent und die Konvergenzordnung stimmt mit der Konsistenzordnung überein.

! Klausurrelevant für Konvergenzanalyse

Beweis. Für den Fehler in $x_i \in \Delta_h$ gilt mit der Lipschitz-Stetigkeit von Φ, dass

$$\begin{aligned}
\|\epsilon_{i+1}\| &= \|y(x_{i+1}) - y_{i+1}\| \\
&= \|y(x_{i+1}) - y(x_i) - h_i \Phi(x_i, y(x_i), h_i) + y(x_i) - y_i + h_i \Phi(x_i, y(x_i), h_i) - h_i \Phi(x_i, y_i, h_i)\| \\
&\leq h_i \|d(x_i, y(x_i), h_i)\| + \|\epsilon_i\| + h_i L_\Phi \|\epsilon_i\| \\
&= h_i \|d(x_i, y(x_i), h_i)\| + (1 + h_i L_\Phi) \|\epsilon_i\|.
\end{aligned}$$

Sei $d_i := d(x_i, y(x_i), h_i)$, so liefert das Lemma 4.11

$$\|\epsilon_i\| \leq e^{\sum_{r=0}^{i-1} h_r L_\Phi} \left(\|\epsilon_0\| + \sum_{r=0}^{i-1} h_r \|d_r\| \right) \leq e^{L_\Phi |I|} \left(\|\epsilon_0\| + |I| \max_{0 \leq r \leq n-1} \|d_r\| \right).$$

Daraus folgt mit $y_0 = y(x_0)$

$$\|\epsilon_h\|_h = \max_{0 \leq i \leq n} \|\epsilon_i\| \leq e^{L_\Phi |I|} \left(\|y(x_0) - y_0\| + |I| \max_{0 \leq r \leq n-1} \|d_r\| \right) \leq |I| e^{L_\Phi |I|} \max_{0 \leq r \leq n-1} \|d_r\|.$$

Diese Abschätzung impliziert Konvergenz bei Konsistenz, und eine Konvergenzordnung von mindestens der Konsistenzordnung. Da die Konsistenzordnung trivialerweise größer gleich der Konvergenzordnung ist folgt die Gleichheit der beiden Ordnungen. □

4.3 Stabilität von Einschrittverfahren

Wird das Einschrittverfahren (4.7) mit Hilfe eines Rechners umgesetzt, so treten bei der Durchführung unvermeidbare Rundungsfehler auf. Ein stabiles Verfahren dämpft diese Fehler in den darauffolgenden Iterationsschritten, wohingegen ein Instabiles diese immer weiter verstärkt. Letzteres kann zu unbrauchbaren numerischen Lösungen führen.
Rundungsfehler sind kleinste Störungen, so dass lineare Terme quadratische und andere Terme höherer Ordnung dominieren. Wir linearisieren deshalb die rechte Seite f in der DGL bzgl. y. Dies liefert

$$y' = f(x,y) = f(x,y_0) + \frac{df}{dy}(x,y_0)(y-y_0) + \ldots \approx \underbrace{\frac{df}{dy}(x,y_0)}_{=:\lambda(x)}\, y + \underbrace{f(x,y_0) - \frac{df}{dy}(x,y_0)y_0}_{=:a(x)},$$

also eine linear DGL

$$y' = \lambda y + a. \tag{4.8}$$

Beispiel 4.13
Sei

$$y' = -100y + 100, \quad y(0.05) = e^{-5} + 1 \approx 1.00673.$$

VdK liefert die exakte Lösung $y = e^{-100x} + 1$ und das explizite Euler-Verfahren zur Approximation dieser reduziert sich auf

$$y_{i+1} = y_i + hf(x_i, y_i) = (1 - 100h)y_i + 100h, \qquad y_0 = 1.00673.$$

Weil $y \in C^\infty$ reduziert sich der Fehler linear mit der Schrittweite h. Für große Schrittweiten beobachten wir aber folgendes instabiles Verhalten.

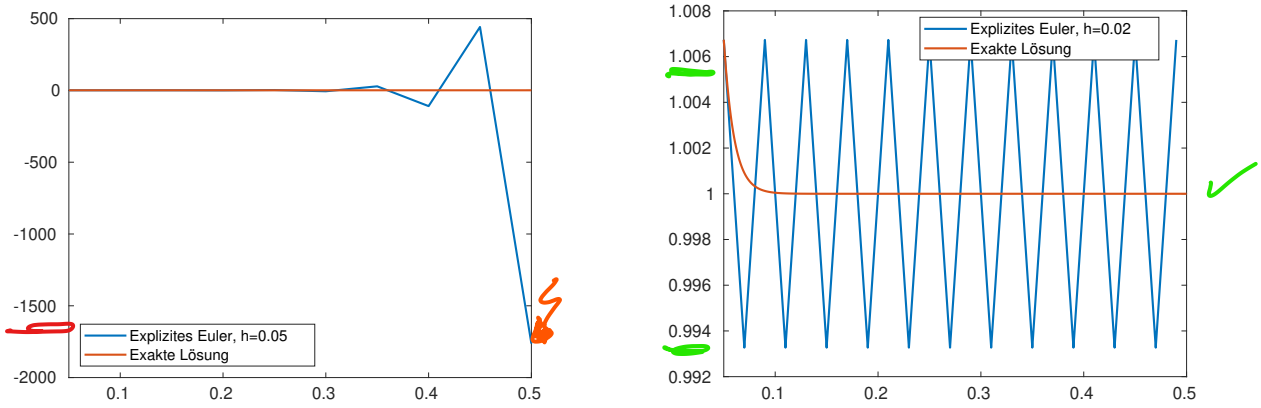


Abb. 4.3: Lösung des expliziten Eulerverfahrens zu $h = 0.05$ (links, blau) und $h = 0.02$ (rechts, blau). Exakte Lösung in rot.

$h = 0.05$			$h = 0.02$		
x_i	y_i	$y(x_i) - y_i$	x_i	y_i	$y(x_i) - y_i$
0.05	$1.0067 \cdot 10^{-0}$	$7.9470 \cdot 10^{-6}$	0.05	$1.0067 \cdot 10^{-0}$	$7.9470 \cdot 10^{-6}$
0.10	$9.7308 \cdot 10^{-1}$	$2.6965 \cdot 10^{-2}$	0.07	$9.9327 \cdot 10^{-1}$	$7.6419 \cdot 10^{-3}$
0.15	$1.1077 \cdot 10^{-0}$	$-1.0768 \cdot 10^{-1}$	0.09	$1.0067 \cdot 10^{-0}$	$-6.6066 \cdot 10^{-3}$
0.20	$5.6928 \cdot 10^{-1}$	$4.3072 \cdot 10^{-1}$	0.11	$9.9327 \cdot 10^{-1}$	$6.7467 \cdot 10^{-3}$
0.25	$2.7229 \cdot 10^{-0}$	$-1.7229 \cdot 10^{-0}$	0.13	$1.0067 \cdot 10^{-0}$	$-6.7277 \cdot 10^{-3}$
0.30	$-5.8915 \cdot 10^{-0}$	$6.8915 \cdot 10^{-0}$	0.15	$9.9327 \cdot 10^{-1}$	$6.7303 \cdot 10^{-3}$
0.35	$2.8566 \cdot 10^{+1}$	$-2.7566 \cdot 10^{+1}$	0.17	$1.0067 \cdot 10^{-0}$	$-6.7300 \cdot 10^{-3}$
0.40	$-1.0926 \cdot 10^{+2}$	$1.1026 \cdot 10^{+2}$	0.19	$9.9327 \cdot 10^{-1}$	$6.7300 \cdot 10^{-3}$
0.45	$4.4206 \cdot 10^{+2}$	$-4.4106 \cdot 10^{+2}$	0.21	$1.0067 \cdot 10^{-0}$	$-6.7300 \cdot 10^{-3}$
0.50	$-1.7632 \cdot 10^{+3}$	$1.7642 \cdot 10^{+3}$	0.23	$9.9327 \cdot 10^{-1}$	$6.7300 \cdot 10^{-3}$
			0.25	$1.0067 \cdot 10^{-0}$	$-6.7300 \cdot 10^{-3}$
			0.27	$9.9327 \cdot 10^{-1}$	$6.7300 \cdot 10^{-3}$
			0.29	$1.0067 \cdot 10^{-0}$	$-6.7300 \cdot 10^{-3}$
			0.31	$9.9327 \cdot 10^{-1}$	$6.7300 \cdot 10^{-3}$
			0.33	$1.0067 \cdot 10^{-0}$	$-6.7300 \cdot 10^{-3}$
			0.35	$9.9327 \cdot 10^{-1}$	$6.7300 \cdot 10^{-3}$
			0.37	$1.0067 \cdot 10^{-0}$	$-6.7300 \cdot 10^{-3}$
			0.39	$9.9327 \cdot 10^{-1}$	$6.7300 \cdot 10^{-3}$
			0.41	$1.0067 \cdot 10^{-0}$	$-6.7300 \cdot 10^{-3}$
			0.43	$9.9327 \cdot 10^{-1}$	$6.7300 \cdot 10^{-3}$
			0.45	$1.0067 \cdot 10^{-0}$	$-6.7300 \cdot 10^{-3}$
			0.47	$9.9327 \cdot 10^{-1}$	$6.7300 \cdot 10^{-3}$
			0.49	$1.0067 \cdot 10^{-0}$	$-6.7300 \cdot 10^{-3}$

Für große $h > 0.02$ werden die Rundungsfehler verstärkt und die numerische Lösung ist unbrauchbar. Die Schrittweite $h = 0.02$ ist der Schwellenwert und erst für $h < 0.02$ werden die Rundungsfehler gedämpft. Das Verfahren ist somit nur bedingt (h hinreichend klein) stabil.

Der Störterm $a(x)$ in (4.8) entscheidet nicht, ob ein numerisches Verfahren stabil ist oder nicht. Ebenso ob λ konstant ist oder nicht. Deshalb reicht es aus die Stabilität anhand des Modellproblems

$$y' = \lambda y, \quad y(0) = 1, \quad \text{für } \lambda \in \mathbb{C} \tag{4.9}$$

zu analysieren. Dies hat die exakte Lösung $y(t) = e^{\lambda t}$, welche exponentiell schnell gegen Null geht wenn $\Re(\lambda) < 0$.

Beispiel 4.14 (Stabilitätsgebiet des expliziten Euler-Verfahrens)
Für das expliziten Euler-Verfahren angewendet auf das Modellproblem (4.9) gilt $y_{i+1} = y_i + h\lambda y_i = (1+h\lambda)y_i$. Sei y_i die Eulerlösung zu den Anfangsdaten y_0 und $\tilde{y}_i$ die Eulerlösung zu den Anfangsdaten $\tilde{y}_0 = y_0 - \epsilon_0$ mit $|\epsilon_0| \ll 1$. So erfüllt die Differenz $\epsilon_i = y_i - \tilde{y}_i$ die Gleichung

$$\epsilon_{i+1} = y_{i+1} - \tilde{y}_{i+1} = (1 + h\lambda)y_i - (1 + h\lambda)\tilde{y}_i = (1 + h\lambda)\epsilon_i.$$

Somit gilt

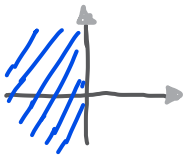
$$\epsilon_i = (1 + h\lambda)\epsilon_{i-1} = \ldots = (1 + h\lambda)^i \epsilon_0.$$

Ist $|1 + h\lambda| < 1$, so wird der Anfangsfehler ϵ_0 im Laufe der Iterationen gedämpft. Im vorherigen Beispiel ist $\lambda = -100$, woraus sich ergibt, dass $h < 0.02$ sein muss damit solche Fehler gedämpft werden.

Definition 4.15 (Absolute Stabilität)
Lässt sich ein Einschrittverfahren angewendet auf (4.9) schreiben als $y_i = F(\lambda h)y_{i-1}$ für eine Funktion $F : \mathbb{C} \to \mathbb{C}$, so heißt

$$B := \{\mu \in \mathbb{C} : |F(\mu)| \leq 1\}$$

das Gebiet absoluter Stabilität. Ist $B \cap \{\mathbb{R}_{\leq 0} \times \mathbb{R}\} = \{\mathbb{R}_{\leq 0} \times \mathbb{R}\}$, so heißt das Verfahren absolut stabil.



Lemma 4.16
Das Stabilitätsgebiet des expliziten Euler-Verfahrens ist $B = \{\mu \in \mathbb{C} : |1 + \mu| \leq 1\}$, also eine Kugel mit Mittelpunkt $(-1, 0)^\top$ und Radius 1.
Das Stabilitätsgebiet jedes expliziten Runge-Kutta-Verfahrens 2. Ordnung ist $B = \{\mu \in \mathbb{C} : |1+\mu+\mu^2/2| \leq 1\}$.

Beweis. Das explizite Euler-Verfahren für (4.9) ist $y_{i+1} = y_i + h\lambda y_i = (1+h\lambda)y_i$. Daraus folgt $F(\mu) = 1+\mu$. Die Definition 4.15 liefert die Behauptung.

Für das Runge-Kutta-Vefahren 2. Ordnung gilt

$$y_{i+1} = y_i + ahf(x_i, y_i) + bhf(x_i + \alpha h, y_i + \beta h f(x_i, y_i))$$

mit $a + b = 1$ und $b\alpha = b\beta = 1/2$. Somit erhalten wir für (4.9)

$$y_{i+1} = y_i + ah\lambda y_i + bh\lambda(y_i + \beta h\lambda y_i) = (1 + (a+b)h\lambda + b\beta h^2\lambda^2)y_i = \left(1 + h\lambda + \frac{h^2\lambda^2}{2}\right) y_i$$

und $F(\mu) = 1 + \mu + \mu^2/2$. □

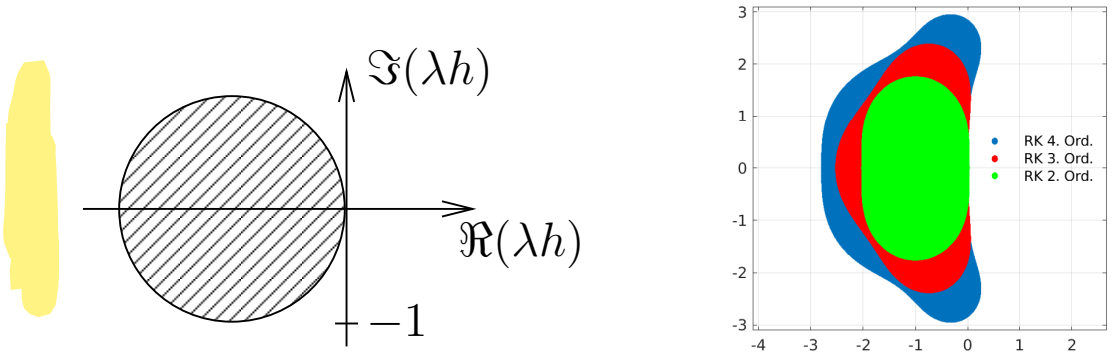


Abb. 4.4: Stabilitätsgebiet des expliziten Euler-Verfahrens (links) und der expliziten Runge-Kutta-Verfahren 2.-4. Ordnung (rechts).

Explizite Verfahren basierend auf Taylorreihenentwicklung sind niemals absolut stabil, weil $F(\mu) = 1 + a_1\mu + a_2\mu^2 + \cdots$. Das einfachste absolut stabile Verfahren ist das implizite Euler-Verfahren. Dazu wird das Rechteck zur Approximation des Flächeninhalts $\int_{x_i}^{x_{i+1}} f(t, y(t))\, dt$ nicht über den Punkt $(x_i, y(x_i))$ sondern über $(x_{i+1}, y(x_{i+1}))$ definiert. Dies liefert

$$y(x_{i+1}) = y(x_i) + \int_{x_i}^{x_{i+1}} f(t, y(t))\, dt \approx y(x_i) + \int_{x_i}^{x_{i+1}} f(x_{i+1}, y(x_{i+1}))\, dt = y(x_i) + h_i f(x_{i+1}, y(x_{i+1})).$$

Somit ist das implizite Euler-Verfahren gegeben durch

$$y_0 = y(x_0), \qquad y_{i+1} = y_i + h_i f(x_{i+1}, y_{i+1}), \quad 0 \le i \le n-1.$$

(absolut stabil)

Es kann gezeigt werden, dass der Fehler des implizite Euler-Verfahrens sich ebenfalls linear in h verhält.

Beispiel 4.17
Sei wieder

$$y' = -100y + 100, \quad y(0.05) = e^{-5} + 1 \approx 1.00673.$$

VdK liefert die exakte Lösung $y = e^{-100x} + 1$ und das implizite Euler-Verfahren zur Approximation dieser reduziert sich auf

$$y_{i+1} = y_i + hf(x_{i+1}, y_{i+1}) = y_i - 100hy_{i+1} + 100h \quad \Leftrightarrow \quad y_{i+1} = \frac{y_i + 100h}{1 + 100h}.$$

Weil $y \in C^\infty$ reduziert sich der Fehler linear in der Schrittweite h. Bereits für große Schrittweiten beobachten wir folgendes stabiles Verhalten.

$h = 0.05$			$h = 0.02$		
x_i	y_i	$y(x_i) - y_i$	x_i	y_i	$y(x_i) - y_i$
0.05	1.006730	$7.9470 \cdot 10^{-6}$	0.05	1.006730	$7.9470 \cdot 10^{-6}$
0.10	1.001121	$-1.0763 \cdot 10^{-3}$	0.07	1.002243	$-1.3315 \cdot 10^{-3}$
0.15	1.000186	$-1.8664 \cdot 10^{-4}$	0.09	1.000747	$-6.2437 \cdot 10^{-4}$
0.20	1.000031	$-3.1155 \cdot 10^{-5}$	0.11	1.000249	$-2.3256 \cdot 10^{-4}$
0.25	1.000005	$-5.1929 \cdot 10^{-6}$	0.13	1.000083	$-8.0826 \cdot 10^{-5}$
0.30	1.000000	$-8.6548 \cdot 10^{-7}$	0.15	1.000027	$-2.7390 \cdot 10^{-5}$
0.35	1.000000	$-1.4425 \cdot 10^{-7}$	0.17	1.000009	$-9.1904 \cdot 10^{-6}$
0.40	1.000000	$-2.4041 \cdot 10^{-8}$	0.19	1.000003	$-3.0717 \cdot 10^{-6}$
0.45	1.000000	$-4.0069 \cdot 10^{-9}$	0.21	1.000001	$-1.0250 \cdot 10^{-6}$
0.50	1.000000	$-6.6781 \cdot 10^{-10}$	0.23	1.000000	$-3.4182 \cdot 10^{-7}$
			0.25	1.000000	$-1.1396 \cdot 10^{-7}$
			0.27	1.000000	$-3.7989 \cdot 10^{-8}$
			0.29	1.000000	$-1.2663 \cdot 10^{-8}$
			0.31	1.000000	$-4.2212 \cdot 10^{-9}$
			0.33	1.000000	$-1.4071 \cdot 10^{-9}$
			0.35	1.000000	$-4.6902 \cdot 10^{-10}$
			0.37	1.000000	$-1.5634 \cdot 10^{-10}$
			0.39	1.000000	$-5.2114 \cdot 10^{-11}$
			0.41	1.000000	$-1.7371 \cdot 10^{-11}$
			0.43	1.000000	$-5.7903 \cdot 10^{-12}$
			0.45	1.000000	$-1.9300 \cdot 10^{-12}$
			0.47	1.000000	$-6.4326 \cdot 10^{-13}$
			0.49	1.000000	$-2.1427 \cdot 10^{-13}$

Lemma 4.18
Das Stabilitätsgebiet des impliziten Euler-Verfahrens ist $B = \{\mu \in \mathbb{C} : |1 - \mu| \geq 1\}$, also der ganze $\mathbb{C}$ ohne die Kugel mit Mittelpunkt $(1, 0)^\top$ und Radius 1. Insbesondere ist das Verfahren absolut stabil.

Beweis. Das implizite Euler-Verfahren für (4.9) ist $y_{i+1} = y_i + h\lambda y_{i+1}$, also $y_{i+1} = (1 - h\lambda)^{-1} y_i$. Daraus folgt $F(\mu) = (1 - \mu)^{-1}$. Die Definition 4.15 liefert die Behauptung. □

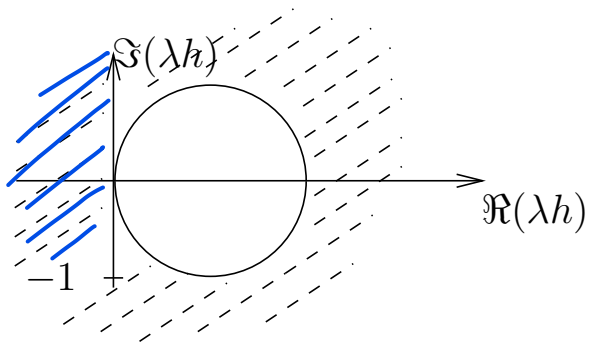


Abb. 4.5: Stabilitätsgebiet des impliziten Euler-Verfahrens.

Wie das allgemeine explizite Einschrittverfahren lassen sich implizite Einschrittverfahren allgemein schreiben als

$$y_0 = y(x_0), \quad y_{i+1} = y_i + h_i \Psi(x_i, y_i, y_{i+1}, h_i) \tag{4.10}$$

mit einer Methodenfunktion Ψ. Weil y_{i+1} auch in Ψ vorkommt, ist das Verfahren implizit.
Der Vorteil impliziter Verfahren liegt darin, dass sie absolut stabil sind, also große Schrittweiten h zulassen. Jedoch muss in jedem Zeitschritt bzw. Iterationsschritt eine potenziell mehr dimensionale nicht-lineare Gleichung gelöst werden. Genauer ein Fixpunktproblem. Dieses Fixpunktproblem kann mit der Banachschen Fixpunktiteration gelöst werden, wenn die rechte Seite eine Kontraktion ist.

Satz 4.19
Sei Ψ in y_{i+1} (der dritten Komponente) Lipschitz-stetig mit Lipschitz-Konstante L. Ist h_i hinreichend klein, wobei $h_i < L^{-1}$ hinreichend ist, so ist (4.10) eindeutig lösbar.

Beweis. Sei $\psi(\eta) = y_i + h_i\Psi(x_i, y_i, \eta, h_i)$. y_{i+1} ist eine Lösung von (4.10) genau dann, wenn es die Fixpunktgleichung

$$y_{i+1} = \psi(y_{i+1})$$

erfüllt. Offensichtlich ist ψ Lipschitz-stetig mit Lipschitz-Konstante h_iL, denn

$$\begin{aligned}\|\psi(\eta) - \psi(\tilde{\eta})\| &= \|y_i + h_i\Psi(x_i, y_i, \eta, h_i) - y_i - h_i\Psi(x_i, y_i, \tilde{\eta}, h_i)\| \\ &= h_i\|\Psi(x_i, y_i, \eta, h_i) - \Psi(x_i, y_i, \tilde{\eta}, h_i)\| \\ &\leq h_iL\|\eta - \tilde{\eta}\|.\end{aligned}$$

Ist $h_i < L^{-1}$, so ist ψ eine Kontraktion und der Banachsche Fixpunktsatz liefert die Behauptung. □

Der Beweis der Konvergenz impliziter Einschrittverfahren ist denkbar einfach wenn man die Konvergenz expliziter Einschrittverfahren bereits bewiesen hat. Denn zu jedem impliziten Einschrittverfahren gibt es ein äquivalentes explizites Einschrittverfahren.

Lemma 4.20
Sei das impliziten Einschrittverfahren (4.10) gegeben. Die Methodenfunktion Ψ sei Lipschitz-stetig bezüglich y_i und y_{i+1}, und sei $0 < h \leq h_0$ für ein hinreichend kleines h_0. Dann existiert ein $\Phi(x, y, h)$ mit $\Phi(x_i, y_i, h_i) = \Psi(x_i, y_i, y_{i+1}, h_i)$, das Lipschitz-stetig bezüglich y_i ist.

Beweis. Nach Satz 4.19 ist (4.10) eindeutig lösbar. Nach dem Satz über implizite Funktionen existiert damit eine Umkehrfunktion $\psi(x, y_i, h_i)$ mit $y_{i+1} = \psi(x_i, y_i, h_i)$, wobei ψ Lipschitz-stetig bezüglich y_i ist. Setze

$$\Phi(x, y, h) = \Psi(x, y, \psi(x, y, h), h).$$

Für die Lipschitz-Konstante gilt dann $L_{\Phi \text{ bzgl. } y} \leq L_{\Psi \text{ bzgl. } y} + L_{\Psi \text{ bzgl. } y_{i+1}}L_{\psi \text{ bzgl. } y}$. □

Laut Satz 4.19 lässt sich das nicht-lineare Gleichungssystem (4.10) mit der Banach'schen Fixpunktiteration lösen falls $h < L^{-1}$. Für große $h \geq L^{-1}$ wird (4.10) als Nullstellenproblem,

$$\text{finde } y_{i+1} \in \mathbb{R}^n : \quad y_{i+1} - y_i - h_i\Psi(x_i, y_i, y_{i+1}, h_i) = 0$$

formuliert und numerisch gelöst. In beiden Fällen wird y_i (konstante Extrapolation) als erste Näherung an y_{i+1} genommen.

5

Berechnung von Nullstellen

5.1 Bisektionsverfahren

Für ein gegebenes $f \in C^0(I)$ mit $I = [a, b]$ ist eine Nullstelle $x^* \in I$ zu finden, d.h. $f(x^*) = 0$ ist zu lösen.

Satz 5.1 (Zwischenwertsatz)
Sei $f \in C^0([a, b])$. Dann existiert zu jedem $\lambda \in [0, 1]$ ein $x^* \in [a, b]$ mit $f(x^*) = \lambda f(a) + (1 - \lambda)f(b)$.

Korollar 5.2 (Nullstellensatz)
Sei $f \in C^0([a, b])$ mit $f(a)f(b) < 0$. Dann existiert ein $x^* \in [a, b]$ mit $f(x^*) = 0$.

Das Bisektionsverfahren basiert auf der trivialen Feststellung, dass für die Näherung $x = m := (a + b)/2$ an x^* gilt $|x - x^*| \leq (b - a)/2 = |I|/2$. Geht die Länge des Intervalls I welches x^* beinhaltet gegen Null, so muss der Mittelpunkt dieses Intervalls gegen die gesuchte Nullstelle x^* gehen. Dies führt zu folgendem Algorithmus, graphisch dargestellt in Abbildung 5.1.

Algorithmus 5.1 Bisektionsverfahren für $f(x^*) = 0$

```
1: Wähle a_0 < b_0 ∈ ℝ mit f(a_0)f(b_0) < 0 und tol > 0
2: for k = 0, 1, ... do
3:     Setze m_k = (a_k + b_k)/2
4:     if |f(m_k)| ≤ tol then
5:         Stop
6:     end if
7:     if f(a_k)f(m_k) < 0 then
8:         Setze a_{k+1} = a_k und b_{k+1} = m_k
9:     else
10:        Setze a_{k+1} = m_k und b_{k+1} = b_k
11:    end if
12: end for
```

Beispiel 5.3
Sei $f(x) = x^2 - 2$, $I = [0, 400]$. Also ist $m_0 = 200$ die Startnäherung an $x^* = \sqrt{2}$. Das Abbruchkriterium ist $|f(m_k)| \leq 10^{-15}$. Für den Fehlerplot siehe Bild 5.2.

Satz 5.4
Sei $f \in C^0(I_0)$ mit $I_0 = [a_0, b_0]$ und $f(a_0)f(b_0) < 0$. Wenn die Intervallfolge $I_k = [a_k, b_k]$ nach Algorithmus 5.1 konstruiert wurde, dann gilt

1. $a_k \leq a_{k+1} \leq b_{k+1} \leq b_k$
2. $|I_k| = 2^{-k}|I_0|$
3. $|(a_k + b_k)/2 - x^*| \leq 2^{-(k+1)}|I_0|$

Beweis. Teil 1. folgt sofort aus der Konstruktion. Ebenfalls aus der Konstruktion folgt unmittelbar

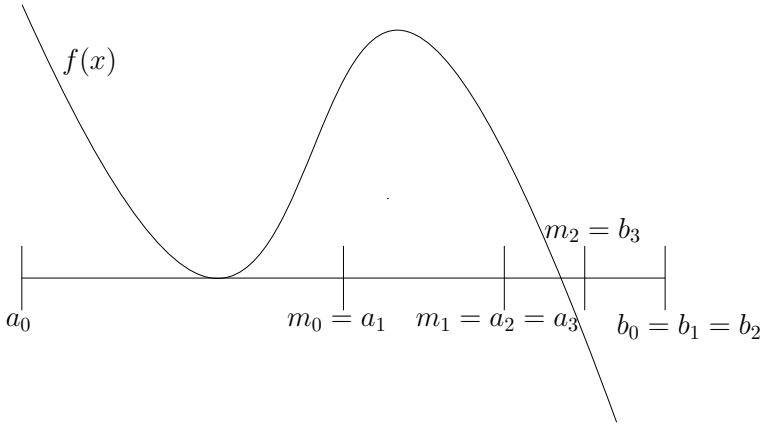


Abb. 5.1: Visualisierung des Bisektionsverfahrens.

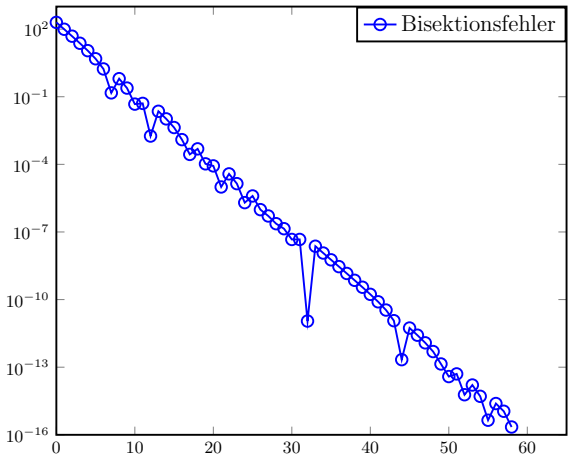


Abb. 5.2: Bisektionsfehler $|m_k - \sqrt{2}|$ gegen Anzahl an Iterationsschritten abgetragen, $m_0 = 200$.

$$|I_k| = \frac{1}{2}|I_{k-1}| = \left(\frac{1}{2}\right)^2 |I_{k-2}| = \cdots = \left(\frac{1}{2}\right)^k |I_0|.$$

Nach Konstruktion gilt $f(a_k)f(b_k) < 0$ und somit nach dem Nullstellensatz $x^* \in I_k$. Folglich ist $|(a_k+b_k)/2 - x^*| \leq 2^{-1}|I_k|$. Die Behauptung folgt mit Teil 2. □

Das Bisektionsverfahren ist numerisch stabil, konvergiert aber nur langsam. So ist nach 10 Iterationsschritten die Fehlerschranke nur um dem Faktor $2^{-10} = 1024^{-1} \approx 10^{-3}$ reduziert.
Es gibt weitere Nullstellenverfahren die deutlich schneller konvergieren und sich leicht für Nullstellenprobleme in höheren Dimensionen verallgemeinern lassen. Besonders prominent ist das Newton-Verfahren.

5.2 Newton-Verfahren

Ist die Funktion f glatter (mind. $f \in C^1$), so kann das Newton-Verfahren eine Nullstelle unter Umständen deutlich schneller finden als das Bisektionsverfahren. Die Grundidee basiert auf einer Linearisierung von f um eine aktuelle Näherung x^k an x^*. So gilt mit der Taylorreihenentwicklung in linearer Näherung

$$f(x) \approx f(x^k) + \nabla f(x^k)(x - x^k).$$

Für $x = x^*$ ist also

$$f(x^*) = 0 \approx f(x^k) + \nabla f(x^k)(x^* - x^k) \quad \Leftrightarrow \quad x^* \approx x^k - \nabla f(x^k)^{-1} f(x^k).$$

Die Größe

$$x^{k+1} = x^k - \nabla f(x^k)^{-1} f(x^k)$$

verwenden wir als neue Näherung an x^* und linearisieren f nun um x^{k+1}. Dies liefert den Algorithmus 5.2. Dabei wird anstatt das Inverse der Jakobimatrix $\nabla f(x^k)$ zu berechnen nur das dazugehörige LGS (Newton-Gleichung)

$$\nabla f(x^k) d^k = -f(x^k) \tag{5.1}$$

mit $x^{k+1} = x^k + d^k$ gelöst. Der Vektor $d^k \in \mathbb{R}^n$ heißt Inkrement bzw. Suchrichtung. Um den Konvergenzradius bzw. die Konvergenzeigenschaften im prä-asymptomatischen Bereich zu verbessern, wird häufig d_k mit einem $t_k \in (0, 1]$ skaliert.

Algorithmus 5.2 Lokales Newton-Verfahren für $f(x^*) = 0$

1: Wähle $x^0 \in \mathbb{R}^n$ und $tol > 0$
2: **for** $k = 0, 1, \ldots$ **do**
3: **if** $\|f(x^k)\| \leq tol$ **then**
4: Stop
5: **end if**
6: Löse das LGS $\nabla f(x^k) d^k = -f(x^k)$ nach $d^k \in \mathbb{R}^n$
7: Setze $x^{k+1} = x^k + d^k$
8: **end for**

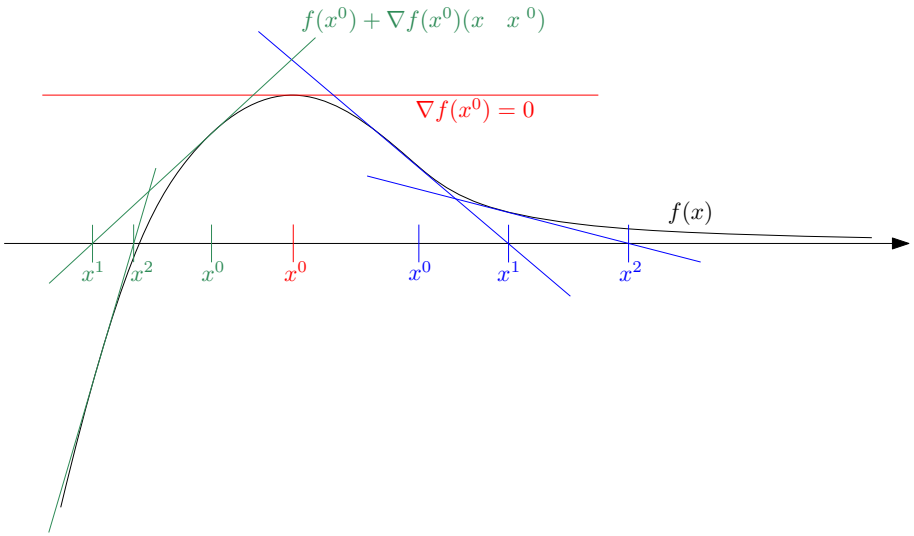


Abb. 5.3: Visualisierung des Newton-Verfahrens.

Beispiel 5.5
Sei $f(x) = x^2 - 2$ und $x^0 = 200$ die Startnäherung an $x^* = \sqrt{2}$. Das Abbruchkriterium ist $|f(x^k)| \leq 10^{-15}$. Für den Fehlerplot siehe Bild 5.4.

5.2.1 Technische Hilfsresultate

Um einen allgemeinen Konvergenzbeweis zu führen benötigen wir mehrere Hilfsaussagen.

Definition 5.6
Sei $\{x^k\} \subset \mathbb{R}^n$ eine gegen x^* konvergente Folge.

1. $x^k \to x^*$ linear, falls ein $C \in [0, 1)$ und ein $k_0 \in \mathbb{N}$ existieren mit

$$\|x^{k+1} - x^*\| \leq C \|x^k - x^*\| \quad \forall\, k \geq k_0 \quad \Leftrightarrow \quad \lim_{k \to \infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = C.$$

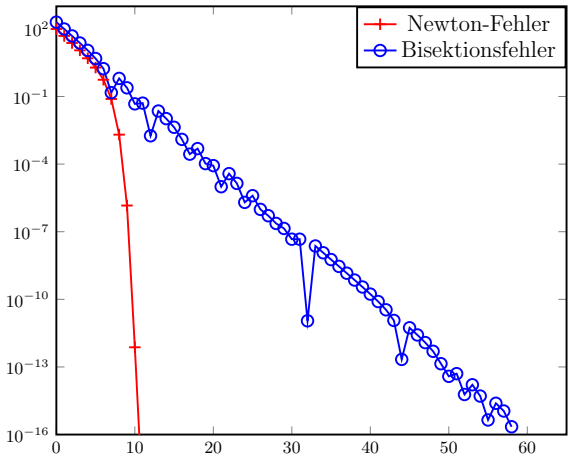


Abb. 5.4: Newton-Fehler $|x^k - \sqrt{2}|$ und Bisektionsfehler $|m_k - \sqrt{2}|$ gegen Anzahl an Iterationsschritten abgetragen, $x^0 = m_0 = 200$.

2. $x^k \to x^*$ superlinear, falls eine Nullfolge $\varepsilon_k \to 0^+$ und ein $k_0 \in \mathbb{N}$ existieren mit

$$\|x^{k+1} - x^*\| \le \varepsilon_k \|x^k - x^*\| \quad \forall\, k \ge k_0 \quad \Leftrightarrow \quad \lim_{k\to\infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = 0.$$

3. $x^k \to x^*$ quadratisch, falls ein $C \ge 0$ und ein $k_0 \in \mathbb{N}$ existieren mit

$$\|x^{k+1} - x^*\| \le C\|x^k - x^*\|^2 \quad \forall\, k \ge k_0 \quad \Leftrightarrow \quad \lim_{k\to\infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|^2} < \infty.$$

Für eine s-mal stetig differenzierbare Funktion $f : A \to B$ schreiben wir kurz $f \in C^s(A; B)$.

Lemma 5.7 (Mittelwertsatz in Integralform)
Sei $f \in C^1(\mathbb{R}^n; \mathbb{R}^m)$, dann gilt

$$f(x) = f(y) + \int_0^1 \nabla f(y + \tau(x - y))(x - y)\, d\tau.$$

Im Folgenden brauchen wir die durch die Vektornorm $\|\cdot\|$ induzierte Matrixnorm

$$\|M\| := \sup_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|Mx\|}{\|x\|} = \sup_{x\in\mathbb{R}^n,\ \|x\|=1} \|Mx\|$$

für jede Matrix $M \in \mathbb{R}^{m\times n}$. Die Norm im Zähler kann eine andere sein als im Nenner, insbesondere wenn $n \ne m$.

Bemerkung 5.8
Für induzierte Matrixnormen gilt die Verträglichkeitsbedingung

$$\|Mx\| = \left\| \|x\| M \frac{x}{\|x\|} \right\| = \left\| M \frac{x}{\|x\|} \right\| \|x\| \le \sup_{z\in\mathbb{R}^n,\ \|z\|=1} \|Mz\| \|x\| = \|M\| \|x\|.$$

Lemma 5.9
Sei $f \in C^1(\mathbb{R}^n; \mathbb{R}^n)$ und $x^k \to x^*$ eine konvergente Folge.

1. Es gilt

$$\lim_{x^k \to x^*} \frac{\|f(x^k) - f(x^*) - \nabla f(x^k)(x^k - x^*)\|}{\|x^k - x^*\|} = 0.$$

2. Ist ∇f um x^* lokal Lipschitz-stetig, so existiert ein $k_0 \in \mathbb{N}$ und eine Konstante $C > 0$, so dass

$$\|f(x^k) - f(x^*) - \nabla f(x^k)(x^k - x^*)\| \le C\|x^k - x^*\|^2 \quad \forall\, k \ge k_0.$$

Beweis. **Teil 1:** Additionen von Null, Dreiecksungleichung und die Verträglichkeitsbedingung der Matrixnorm liefert

$$\begin{aligned}\|f(x^k) - f(x^*) - \nabla f(x^k)(x^k - x^*)\| &= \|f(x^k) - f(x^*) - \nabla f(x^*)(x^k - x^*) + [\nabla f(x^*) - \nabla f(x^k)](x^k - x^*)\| \\ &\leq \|f(x^k) - f(x^*) - \nabla f(x^*)(x^k - x^*)\| + \|\nabla f(x^*) - \nabla f(x^k)\|\|x^k - x^*\|.\end{aligned}$$

Aus der Definition der Ableitung und $\nabla f \in C^0(\mathbb{R}^n; \mathbb{R}^{n\times n})$ folgt

$$\begin{aligned}0 &\leq \lim_{x^k \to x^*} \frac{\|f(x^k) - f(x^*) - \nabla f(x^k)(x^k - x^*)\|}{\|x^k - x^*\|} \\ &\leq \lim_{x^k \to x^*} \frac{\|f(x^k) - f(x^*) - \nabla f(x^*)(x^k - x^*)\|}{\|x^k - x^*\|} + \lim_{x^k \to x^*} \frac{\|\nabla f(x^*) - \nabla f(x^k)\|\|x^k - x^*\|}{\|x^k - x^*\|} \\ &= 0.\end{aligned}$$

Teil 2: Der Mittelwertsatz Lemma 5.7 liefert

$$\begin{aligned}f(x^k) - f(x^*) - \nabla f(x^k)(x^k - x^*) &= \int_0^1 \nabla f(x^* + \tau(x^k - x^*))(x^k - x^*)\, d\tau - \nabla f(x^k)(x^k - x^*) \\ &= \int_0^1 \left[\nabla f(x^* + \tau(x^k - x^*)) - \nabla f(x^k)\right](x^k - x^*)\, d\tau.\end{aligned}$$

Folglich erhalten wir mit der lokalen Lipschitz-Stetigkeit von ∇f sobald k hinreichend groß ist, dass

$$\begin{aligned}\|f(x^k) - f(x^*) - \nabla f(x^k)(x^k - x^*)\| &\leq \int_0^1 \|\nabla f(x^* + \tau(x^k - x^*)) - \nabla f(x^k)\|\|x^k - x^*\|\, d\tau \\ &\leq L\|x^k - x^*\| \int_0^1 \|x^* - x^k + \tau(x^k - x^*)\|\, d\tau \\ &= L\|x^k - x^*\|^2 \int_0^1 |\tau - 1|\, d\tau \\ &= \frac{L}{2}\|x^k - x^*\|^2,\end{aligned}$$

wobei L die Lipschitz-Konstante von ∇f ist. □

Lemma 5.10
Sei $M \in \mathbb{R}^{n\times n}$ mit $\|M\| < 1$. Dann ist $I - M$ regulär mit

$$\|(I - M)^{-1}\| \leq \frac{1}{1 - \|M\|}.$$

Beweis. Für $x \in \mathbb{R}^n$ beliebig gilt mit der Dreiecksungleichung ($\|x\| = \|x - Mx + Mx\| \leq \|x - Mx\| + \|Mx\|$) und der Verträglichkeitsbedingung für Matrixnormen

$$\|(I - M)x\| = \|x - Mx\| \geq \|x\| - \|Mx\| \geq (1 - \|M\|)\|x\|. \tag{5.2}$$

Weil $\|M\| < 1$ ist $1 - \|M\| > 0$ und folglich ist $(I - M)x \neq 0$ für $x \neq 0$. Dies bedeutet, dass $I - M$ regulär ist. Es hat nur den trivialen Kern $\ker(I - M) = \{0\}$. Für $x = (I - M)^{-1}y$ mit $y \in \mathbb{R}^n$ beliebig folgt aus (5.2), dass

$$\|y\| \geq (1 - \|M\|)\|(I - M)^{-1}y\|.$$

Mit der Definition der Matrixnorm erhalten wir nun die Behauptung

$$\|(I - M)^{-1}\| := \sup_{y \in \mathbb{R}^n \setminus \{0\}} \frac{\|(I - M)^{-1}y\|}{\|y\|} \leq \frac{1}{1 - \|M\|}.$$

□

Lemma 5.11 (Banach-Lemma)
Seien A, $B \in \mathbb{R}^{n\times n}$ mit $\|I - BA\| < 1$. Dann sind A und B regulär mit

$$\|B^{-1}\| \leq \frac{\|A\|}{1 - \|I - BA\|} \quad \text{und} \quad \|A^{-1}\| \leq \frac{\|B\|}{1 - \|I - BA\|}.$$

Beweis. Sei $M = I - BA$. Nach Lemma 5.10 ist $I - M = BA$ regulär. Wegen $0 \neq \det(BA) = \det(B)\det(A)$ sind $\det(B) \neq 0$ und $\det(A) \neq 0$. Also sind A und B regulär. Somit folgt aus $I - M = BA$, dass $B^{-1} = A(I - M)^{-1}$, und mit Lemma 5.10, dass

$$\|B^{-1}\| \leq \|A\|\|(I - M)^{-1}\| \leq \frac{\|A\|}{1 - \|I - BA\|}.$$

Die zweite Abschätzung folgt analog mit $A^{-1} = (I - M)^{-1}B$. □

Lemma 5.12
Sei $f \in C^1(\mathbb{R}^n;\mathbb{R}^n)$ und $x^* \in \mathbb{R}^n$ mit $\nabla f(x^*)$ regulär. Dann existiert ein $\varepsilon > 0$ und eine Konstante $C > 0$, so dass für alle $x \in U_\varepsilon(x^*)$ $\nabla f(x)$ regulär ist mit $\|\nabla f(x)^{-1}\| \leq C$.

Beweis. Da ∇f stetig in x^* ist, existiert ein $\varepsilon > 0$ mit

$$\|\nabla f(x^*) - \nabla f(x)\| \leq \frac{1}{2\|\nabla f(x^*)^{-1}\|} \quad \forall x \in U_\varepsilon(x^*).$$

Mit der Submuliplikativität der Matrixnorm erhalten wir somit

$$\|I - \nabla f(x^*)^{-1}\nabla f(x)\| = \|\nabla f(x^*)^{-1}[\nabla f(x^*) - \nabla f(x)]\| \leq \|\nabla f(x^*)^{-1}\|\|\nabla f(x^*) - \nabla f(x)\| \leq \frac{1}{2}.$$

Das Banach-Lemma 5.11 liefert, dass $\nabla f(x)$ regulär ist mit

$$\|\nabla f(x)^{-1}\| \leq \frac{\|\nabla f(x^*)^{-1}\|}{1 - \|I - \nabla f(x^*)^{-1}\nabla f(x)\|} \leq 2\|\nabla f(x^*)^{-1}\|.$$

□

Lemma 5.13
Sei $f \in C^1(\mathbb{R}^n;\mathbb{R}^n)$, $x^k \to x^*$ eine konvergente Folge mit $f(x^*) = 0$ und $\nabla f(x^*)$ regulär. Dann existiert ein $k_0 \in \mathbb{N}$ und ein $\beta > 0$, so dass

$$\|f(x^k)\| \geq \beta\|x^k - x^*\| \quad \forall\, k \geq k_0.$$

Beweis. Da $f \in C^1(\mathbb{R}^n;\mathbb{R}^n)$ gilt nach der Definition der Ableitung

$$\lim_{x^k \to x^*} \frac{\|f(x^k) - f(x^*) - \nabla f(x^*)(x^k - x^*)\|}{\|x^k - x^*\|} = 0.$$

Deshalb existiert für jedes $\varepsilon > 0$ ein $k_0 \in \mathbb{N}$, so dass

$$\|f(x^k) - f(x^*) - \nabla f(x^*)(x^k - x^*)\| \leq \varepsilon\|x^k - x^*\| \quad \forall\, k \geq k_0.$$

Für $\varepsilon < \|\nabla f(x^*)^{-1}\|^{-1}$ gilt mit der Dreiecksungleichung ($\|x\| \geq \|x + y\| - \|y\|$), $f(x^*) = 0$ und $\|x\| = \|\nabla f(x^*)^{-1}\nabla f(x^*)x\| \leq \|\nabla f(x^*)^{-1}\|\|\nabla f(x^*)x\|$, dass

$$\begin{aligned}\|f(x^k)\| &\geq \|\nabla f(x^*)(x^k - x^*)\| - \|f(x^k) - f(x^*) - \nabla f(x^*)(x^k - x^*)\| \\ &\geq \|\nabla f(x^*)^{-1}\|^{-1}\|x^k - x^*\| - \varepsilon\|x^k - x^*\| \\ &= \beta\|x^k - x^*\|\end{aligned}$$

für $\beta = \|\nabla f(x^*)^{-1}\|^{-1} - \varepsilon > 0$. □

Dieses Lemma liefert überdies die Rechtfertigung für ein Abbruchkriterium im Newton-Algorithmus der Form $\|f(x^k)\| \leq tol$.

Lemma 5.14
Sei $f \in C^1(\mathbb{R}^n;\mathbb{R}^n)$ und $x^k \to x^*$ eine konvergente Folge. Dann gilt für $k \to \infty$

$$\int_0^1 \|\nabla f(x^k + \tau(x^{k+1} - x^k)) - \nabla f(x^*)\|\, d\tau \to 0.$$

Beweis. Aus $x^k \to x^*$ folgt $x^k + \tau(x^{k+1} - x^k) \to x^*$ gleichmäßig für alle $\tau \in [0,1]$. Wegen $\nabla f \in C^0(\mathbb{R}^n;\mathbb{R}^{n\times n})$ gilt für alle $\varepsilon > 0$ existiert ein $k_0 \in \mathbb{N}$, so dass

$$\|\nabla f(x^k + \tau(x^{k+1} - x^k)) - \nabla f(x^*)\| \le \varepsilon \quad \forall\, k \ge k_0\ \forall\, \tau \in [0,1].$$

Folglich gilt

$$\int_0^1 \|\nabla f(x^k + \tau(x^{k+1} - x^k)) - \nabla f(x^*)\|\, d\tau \le \int_0^1 \varepsilon\, d\tau = \varepsilon \quad \forall\, k \ge k_0.$$

□

Korollar 5.15
Sei $f \in C^1(\mathbb{R}^n;\mathbb{R}^n)$ und $x^k \to x^*$ eine konvergente Folge. Dann gilt für $k \to \infty$

$$\int_0^1 \|\nabla f(x^* + \tau(x^* - x^k)) - \nabla f(x^*)\|\, d\tau \to 0.$$

Lemma 5.16
Sei $x^k \to x^*$ eine superlinear konvergierende Folge mit $x^k \neq x^*$. Dann gilt

$$\lim_{k\to\infty} \frac{\|x^{k+1} - x^k\|}{\|x^k - x^*\|} = 1.$$

Beweis. Mit der inversen Dreiecksungleichung ($\|x \pm y\| \ge |\,\|x\| - \|y\|\,|$) und der superlinearen Konvergenz folgt

$$0 \le \lim_{k\to\infty} \left| \frac{\|x^{k+1} - x^k\|}{\|x^k - x^*\|} - 1 \right| = \lim_{k\to\infty} \left| \frac{\|x^{k+1} - x^k\| - \|x^k - x^*\|}{\|x^k - x^*\|} \right| \le \lim_{k\to\infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = 0.$$

□

Dieses Lemma liefert überdies die Rechtfertigung für ein Abbruchkriterium im Newton-Algorithmus der Form $\|x^{k+1} - x^k\| \le tol$.

Satz 5.17
Sei $f \in C^1(\mathbb{R}^n;\mathbb{R}^n)$, $x^k \to x^*$ eine konvergente Folge mit $x^k \neq x^*$ und $\nabla f(x^*)$ regulär. Dann sind äquivalent

1. $x^k \to x^*$ superlinear und $f(x^*) = 0$.
2. $$\lim_{k\to\infty} \frac{\|f(x^k) + \nabla f(x^k)(x^{k+1} - x^k)\|}{\|x^{k+1} - x^k\|} = 0.$$
3. $$\lim_{k\to\infty} \frac{\|f(x^k) + \nabla f(x^*)(x^{k+1} - x^k)\|}{\|x^{k+1} - x^k\|} = 0.$$

Beweis. "3. ⇒ 1.: Addition von Nullen und anwenden des Mittelwertsatzes Lemma 5.7 liefert

$$\begin{aligned} f(x^{k+1}) &= f(x^{k+1}) - f(x^k) - \nabla f(x^*)(x^{k+1} - x^k) + f(x^k) + \nabla f(x^*)(x^{k+1} - x^k) \\ &= \int_0^1 \left[\nabla f(x^k + \tau(x^{k+1} - x^k)) - \nabla f(x^*)\right](x^{k+1} - x^k)\, d\tau + f(x^k) + \nabla f(x^*)(x^{k+1} - x^k). \end{aligned} \tag{5.3}$$

Folglich ist

$$\|f(x^{k+1})\| \le \int_0^1 \|\nabla f(x^k + \tau(x^{k+1} - x^k)) - \nabla f(x^*)\|\, d\tau \|x^{k+1} - x^k\| + \|f(x^k) + \nabla f(x^*)(x^{k+1} - x^k)\|.$$

Nach Lemma 5.14 und Voraussetzung aus Teil 3. folgt die Existenz einer Nullfolge $\varepsilon_k \to 0^+$ mit

$$\|f(x^{k+1})\| \leq \varepsilon_k \|x^{k+1} - x^k\|.$$

Somit gilt mit $x^k \to x^*$, dass $f(x^{k+1}) \to 0$ für $k \to \infty$ was $f(x^*) = 0$ impliziert. Nach Lemma 5.13 existiert ein $\beta > 0$ und eine $k_0 \in \mathbb{N}$ mit

$$\|f(x^{k+1})\| \geq \beta \|x^{k+1} - x^*\| \quad \forall\, k \geq k_0.$$

Somit gilt für $k \geq k_0$

$$\beta \|x^{k+1} - x^*\| \leq \|f(x^{k+1})\| \leq \varepsilon_k \|x^{k+1} - x^k\| \leq \varepsilon_k \left(\|x^{k+1} - x^*\| + \|x^k - x^*\| \right).$$

Folglich liegt mit

$$\frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} \leq \frac{\varepsilon_k}{\beta - \varepsilon_k} \xrightarrow{k \to \infty} 0$$

für k hinreichend groß, so dass $\beta - \varepsilon_k > 0$, superlineare Konvergenz vor.
"1. $\Rightarrow$ 3.: Weil $f \in C^1(\mathbb{R}^n; \mathbb{R}^n)$ lokale Lipschitz-Stetigkeit impliziert, folgt aus $x^k \to x^*$ die Existenz einer Konstante $L > 0$ und eines Index $k_0 \in \mathbb{N}$ mit

$$\|f(x^k) - f(x^*)\| \leq L \|x^k - x^*\| \quad \forall\, k \geq k_0.$$

Wegen $f(x^*) = 0$ gilt damit

$$\|f(x^{k+1})\| = \|f(x^{k+1}) - f(x^*)\| \leq L \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} \frac{\|x^k - x^*\|}{\|x^{k+1} - x^k\|} \|x^{k+1} - x^k\|.$$

Lemma 5.16 und die Definition der superlinearen Konvergenz implizieren jetzt die Existenz einer Nullfolge $\varepsilon_k \to 0^+$ mit

$$\|f(x^{k+1})\| \leq \varepsilon_k \|x^{k+1} - x^k\|.$$

Aus der Identität (5.3) folgt

$$\begin{aligned} \|f(x^k) + \nabla f(x^*)(x^{k+1} - x^k)\| &\leq \|f(x^{k+1})\| + \int_0^1 \|\nabla f(x^k + \tau(x^{k+1} - x^k)) - \nabla f(x^*)\|\, d\tau \|x^{k+1} - x^k\| \\ &\leq \left(\varepsilon_k + \int_0^1 \|\nabla f(x^k + \tau(x^{k+1} - x^k)) - \nabla f(x^*)\|\, d\tau \right) \|x^{k+1} - x^k\|. \end{aligned}$$

Jetzt folgt die Behauptung aus Teil 3. mit Lemma 5.14.
"2. $\Leftrightarrow$ 3.: Hausübung. □

Satz 5.18
Sei $f \in C^1(\mathbb{R}^n; \mathbb{R}^n)$, $x^k \to x^*$ eine konvergente Folge mit $x^k \neq x^*$, $\nabla f(x^*)$ regulär und ∇f lokal um x^* Lipschitz-stetig. Dann sind äquivalent:

1. $x^k \to x^*$ quadratisch und $f(x^*) = 0$.
2. Es existiert ein $k_0 \in \mathbb{N}$ und eine Konstante $C > 0$ mit

$$\|f(x^k) + \nabla f(x^k)(x^{k+1} - x^k)\| \leq C \|x^{k+1} - x^k\|^2 \quad \forall\, k \geq k_0.$$

3. Es existiert ein $k_0 \in \mathbb{N}$ und eine Konstante $C > 0$ mit

$$\|f(x^k) + \nabla f(x^*)(x^{k+1} - x^k)\| \leq C \|x^{k+1} - x^k\|^2 \quad \forall\, k \geq k_0.$$

Beweis. "2. $\Rightarrow$ 1.: Die Aussage 2. impliziert

$$\lim_{k \to \infty} \frac{\|f(x^k) + \nabla f(x^k)(x^{k+1} - x^k)\|}{\|x^{k+1} - x^k\|} = 0$$

und somit folgt nach Satz 5.17 bereits $f(x^*) = 0$ wie auch $x^k \to x^*$ superlinear. Durch Addition von Nullen erhalten wir die Identität

$$\nabla f(x^k)(x^{k+1}-x^*) = \left[f(x^k)+\nabla f(x^k)(x^{k+1}-x^k)\right] - \left[f(x^k)-f(x^*)-\nabla f(x^k)(x^k-x^*)\right]. \tag{5.4}$$

Lemma 5.12 liefert jetzt die Existenz einer Konstante $C>0$ und eines Index $k_0\in\mathbb{N}$, so dass für alle $k\geq k_0$ $\nabla f(x^k)$ regulär ist mit $\|\nabla f(x^k)^{-1}\|\leq C$. Multiplikation von (5.4) mit $\nabla f(x^k)^{-1}$, Normabschätzungen und Division mit $\|x^k-x^*\|^2$ liefert

$$\frac{\|x^{k+1}-x^*\|}{\|x^k-x^*\|^2} \leq C\left(\frac{\|f(x^k)+\nabla f(x^k)(x^{k+1}-x^k)\|}{\|x^{k+1}-x^k\|^2}\frac{\|x^{k+1}-x^k\|^2}{\|x^k-x^*\|^2} + \frac{\|f(x^k)-f(x^*)-\nabla f(x^k)(x^k-x^*)\|}{\|x^k-x^*\|^2}\right).$$

Mit Lemma 5.16, Lemma 5.9 und Voraussetzung aus Teil 2. folgt jetzt die quadratische Konvergenz.
"2. $\Leftarrow$ 1.: Aus der allgemein gültigen Gleichung (5.4) folgt

$$\|f(x^k)+\nabla f(x^k)(x^{k+1}-x^k)\| \leq \|f(x^k)-f(x^*)-\nabla f(x^k)(x^k-x^*)\| + \|\nabla f(x^k)\|\|x^{k+1}-x^*\|.$$

Mit $0<\|\nabla f(x^k)\|<\infty$ für k hinreichend groß nach Lemma 5.16, liefert die quadratische Konvergenz und Lemma 5.9 jetzt

$$\|f(x^k)+\nabla f(x^k)(x^{k+1}-x^k)\| \leq C\|x^k-x^*\|^2$$

mit einer Konstante $C>0$. Lemma 5.16 liefert damit die Behauptung aus Teil 2.
"2. $\Leftrightarrow$ 3.: Hausübung. □

5.2.2 Lokales Newton-Verfahren

Satz 5.19 (Konvergenz des lokalen Newton-Verfahrens)
Sei $f\in C^1(\mathbb{R}^n;\mathbb{R}^n)$, $x^*\in\mathbb{R}^n$ mit $f(x^*)=0$ und $\nabla f(x^*)$ regulär. Dann existiert ein $\varepsilon>0$, so dass für jedes $x^0\in U_\varepsilon(x^*)$ gilt:

1. Das lokale Newton-Verfahren 5.2 ist wohldefiniert und die erzeugte Folge $\{x^k\}$ konvergiert gegen x^*.
2. Die Konvergenzgeschwindigkeit ist superlinear.
3. Ist ∇f lokal um x^* Lipschitz-stetig, so ist die Konvergenzgeschwindigkeit quadratisch.

Beweis. Nach Lemma 5.12 existiert ein $\varepsilon_1>0$ und eine Konstante $C>0$ mit

$$\|\nabla f(x)^{-1}\|\leq C \quad \forall\, x\in U_{\varepsilon_1}(x^*).$$

Insbesondere ist $\nabla f(x)$ regulär in dieser Umgebung. Nach Lemma 5.9 existiert ein weiteres $\varepsilon_2>0$ mit

$$\|f(x)-f(x^*)-\nabla f(x)(x-x^*)\| \leq \frac{1}{2C}\|x-x^*\| \quad \forall\, x\in U_{\varepsilon_2}(x^*)$$

mit derselben Konstante C von oben. Sei $\varepsilon=\min\{\varepsilon_1,\varepsilon_2\}$ und $x^0\in U_\varepsilon(x^*)$. Dann ist x^1 wohldefiniert, weil $\nabla f(x^0)$ invertierbar ist und damit die Newton-Gleichung (5.1) (eindeutig) lösbar ist. Weiterhin gilt

$$\|x^1-x^*\| = \|x^0-x^*-\nabla f(x^0)^{-1}f(x^0)\| \leq \|\nabla f(x^0)^{-1}\|\|f(x^0)-f(x^*)-\nabla f(x^0)(x^0-x^*)\| \leq \frac{C}{2C}\|x^0-x^*\|.$$

Damit folgt, dass $x^1\in U_\varepsilon(x^*)$ und somit gilt per Rekursion, dass x^k immer wohldefiniert ist und

$$\|x^k-x^*\| \leq \frac{1}{2}\|x^{k-1}-x^*\| \leq \cdots \leq \left(\frac{1}{2}\right)^k\|x^0-x^*\|, \quad k=0,1,\ldots.$$

Insbesondere konvergiert $x^k\to x^*$ linear. Wegen der Newton-Gleichung (5.1) ist

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

und die behaupteten Konvergenzgeschwindigkeiten folgen direkt aus den Sätzen 5.17 und 5.18. □

Beispiel 5.20
Die Voraussetzung $\nabla f(x^*)\neq 0$ ist für die superlineare bzw. quadratische Konvergenz wichtig, aber nicht für das Verfahren selbst. So konvergiert das Newton-Verfahren für $f(x)=x^2$ immer noch linear. Für den Fehlerplot mit $x_0=200$ siehe Bild 5.5.

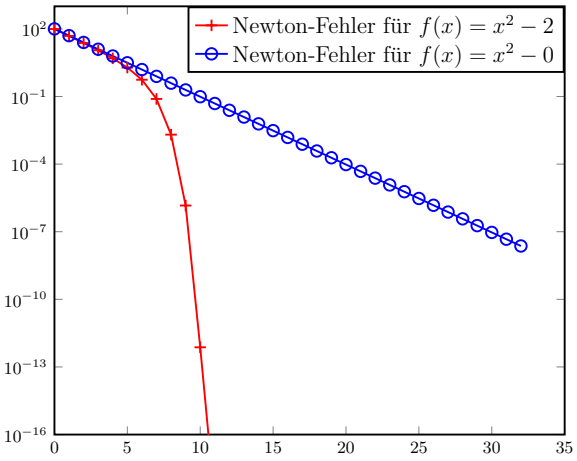


Abb. 5.5: Newton-Fehler gegen Anzahl an Iterationsschritten abgetragen, $x^0 = 200$.

5.2.3 Inexaktes Newton-Verfahren

Das Lösen der Newton-Gleichung (5.1), also das Finden eines $d^k \in \mathbb{R}^n$, so dass

$$\nabla f(x^k) d^k = -f(x^k)$$

kann in hohen Dimensionen sehr aufwändig sein. Da die neue Lösung x^{k+1} sowieso in einem weiteren Newton-Schritt verbessert werden muss, kommt es gar nicht darauf an, dass die Suchrichtung/Inkrement d^k exakt ist, sondern sie muss nur grob in die richtige Richtung zeigen. Wir wollen die Newton-Gleichung nur noch näherungsweise mit einem iterativen Verfahren (siehe Vorlesung wissenschaftliches Rechnen) lösen. Sei $\eta_k \geq 0$ eine gegebene Toleranz. Die Newton-Gleichung wird jetzt abgeändert zu dem unterbestimmten Problem: Finde ein $d^k \in \mathbb{R}^n$, so dass

$$\|\nabla f(x^k) d^k + f(x^k)\| \leq \eta_k \|f(x^k)\|. \tag{5.5}$$

Algorithmus 5.3 Inexaktes lokales Newton-Verfahren für $f(x^*) = 0$

1: Wähle $x^0 \in \mathbb{R}^n$ und $tol > 0$
2: **for** $k = 0, 1, \ldots$ **do**
3: **if** $\|f(x^k)\| \leq tol$ **then**
4: Stop
5: **end if**
6: Wähle $\eta_k \geq 0$
7: Finde ein $d^k \in \mathbb{R}^n$: $\|\nabla f(x^k) d^k + f(x^k)\| \leq \eta_k \|f(x^k)\|$ (durch iterativen Löser für die Newton-Gleichung (5.1))
8: Setze $x^{k+1} = x^k + d^k$
9: **end for**

Satz 5.21 (Konvergenz inexaktes lokales Newton-Verfahren)
Sei $f \in C^1(\mathbb{R}^n; \mathbb{R}^n)$, $x^* \in \mathbb{R}^n$ mit $f(x^*) = 0$ und $\nabla f(x^*)$ regulär. Dann existiert ein $\varepsilon > 0$, so dass für jedes $x^0 \in U_\varepsilon(x^*)$ gilt:

1. Ist $\eta_k \leq \bar{\eta}$ für ein hinreichend kleines $\bar{\eta} \in (0, 1)$, so ist das inexakte lokale Newton-Verfahren 5.3 wohldefiniert und die erzeugte Folge $\{x^k\}$ konvergiert linear gegen x^*.
2. Falls $\eta_k \to 0$, so ist die Konvergenzgeschwindigkeit superlinear.
3. Ist $\eta_k \leq C\|f(x^k)\|$ für eine beliebige Konstante $C > 0$, und ist ∇f lokal um x^* Lipschitz-stetig, so ist die Konvergenzgeschwindigkeit quadratisch.

Beweis. Weil $f \in C^1(\mathbb{R}^n; \mathbb{R}^n)$ ist f lokal Lipschitz-stetig, d.h. es existiert ein $\varepsilon_1 > 0$ und ein $L > 0$, so dass

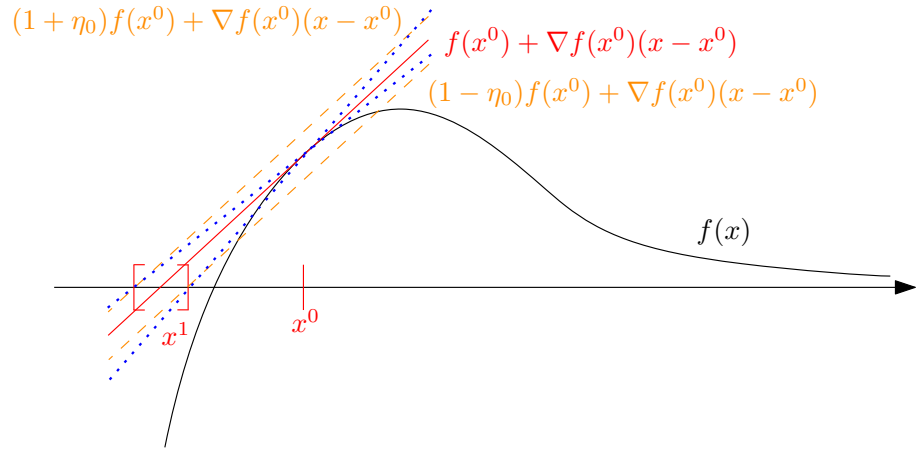


Abb. 5.6: Visualisierung des inexakten Newton-Verfahrens.

$$\|f(x)\| = \|f(x) - f(x^*)\| \leq L\|x - x^*\| \quad \forall\, x \in U_{\varepsilon_1}(x^*).$$

Nach Lemma 5.12 existiert ein $\varepsilon_2 > 0$ und eine Konstante $C > 0$ mit

$$\|\nabla f(x)^{-1}\| \leq C \quad \forall\, x \in U_{\varepsilon_2}(x^*).$$

Insbesondere ist $\nabla f(x)$ regulär in dieser Umgebung. Nach Lemma 5.9 existiert ein drittes $\varepsilon_3 > 0$ mit

$$\|f(x) - f(x^*) - \nabla f(x)(x - x^*)\| \leq \frac{1}{4C}\|x - x^*\| \quad \forall x \in U_{\varepsilon_3}(x^*)$$

mit derselben Konstante C von oben. Sei $\varepsilon = \min\{\varepsilon_1, \varepsilon_2, \varepsilon_3\}$ und $\bar{\eta} = (4CL)^{-1}$. Für $x^0 \in U_\varepsilon(x^*)$ ist $\nabla f(x^0)$ regulär und damit hat (5.5) mindestens die Lösung $d^0 = -\nabla f(x^0)^{-1} f(x^0)$. Folglich ist x^1 wohldefiniert. Weiterhin gilt

$$\begin{aligned}
\|x^1 - x^*\| &= \|x^0 - x^* + d^0\| \\
&= \|x^0 - x^* - \nabla f(x^0)^{-1} f(x^0) + \nabla f(x^0)^{-1}\left(\nabla f(x^0) d^0 + f(x^0)\right)\| \\
&\leq \|\nabla f(x^0)^{-1}\|\left(\|f(x^0) - f(x^*) - \nabla f(x^0)(x^0 - x^*)\| + \|\nabla f(x^0) d^0 + f(x^0)\|\right) \\
&\leq C\left(\frac{1}{4C}\|x^0 - x^*\| + \eta_0 \|f(x^0)\|\right) \\
&\leq C\left(\frac{1}{4C} + \bar{\eta} L\right)\|x^0 - x^*\| \\
&= \frac{1}{2}\|x^0 - x^*\|.
\end{aligned}$$

Damit folgt, dass $x^1 \in U_\varepsilon(x^*)$ und somit gilt rekursiv, dass x^k immer wohldefiniert ist und

$$\|x^k - x^*\| \leq \frac{1}{2}\|x^{k-1} - x^*\| \leq \cdots \leq \left(\frac{1}{2}\right)^k \|x^0 - x^*\|, \quad k = 0, 1, \ldots$$

Insbesondere konvergiert $x^k \to x^*$ linear.
Um die garantierte Konvergenzgeschwindigkeit zu superlinear zu verbessern erhalten wir analog zu oben

$$\begin{aligned}
\|x^{k+1} - x^*\| &\leq \|\nabla f(x^k)^{-1}\|\left(\|f(x^k) - f(x^*) - \nabla f(x^k)(x^k - x^*)\| + \|\nabla f(x^k) d^k + f(x^k)\|\right) \\
&\leq C\left(\|f(x^k) - f(x^*) - \nabla f(x^k)(x^k - x^*)\| + \eta_k \|f(x^k)\|\right) \\
&\leq C\left(\frac{\|f(x^k) - f(x^*) - \nabla f(x^k)(x^k - x^*)\|}{\|x^k - x^*\|} + \eta_k L\right)\|x^k - x^*\|.
\end{aligned}$$

Mit Teil 1. des Lemmas 5.9 und $\eta_k \to 0$ geht der Faktor vor $\|x^k - x^*\|$ gegen Null. Es liegt also superlineare Konvergenz vor.
Wieder analog zu oben erhalten wir

$$\begin{aligned}\|x^{k+1}-x^*\| &\le C\left(\|f(x^k)-f(x^*)-\nabla f(x^k)(x^k-x^*)\|+\eta_k\|f(x^k)\|\right)\\ &\le C\left(\|f(x^k)-f(x^*)-\nabla f(x^k)(x^k-x^*)\|+\tilde{C}\|f(x^k)\|^2\right)\\ &\le C\left(\frac{\|f(x^k)-f(x^*)-\nabla f(x^k)(x^k-x^*)\|}{\|x^k-x^*\|^2}+\tilde{C}L^2\right)\|x^k-x^*\|^2.\end{aligned}$$

Lemmas 5.9 Teil 2. liefert jetzt die quadratische Konvergenz. □

Bemerkung 5.22
Im vorherigen Satz wurde für die lineare Konvergenz benötigt, dass $\bar{\eta} \in (0,1)$ hinreichend klein ist, ohne dass $\bar{\eta}$ praktisch berechnet werden könnte. Wiederholt man die Analysis mit der gewichteten, aber i.A. nicht berechenbare, Norm $\|\cdot\|_* = \|\nabla f(x^*)\cdot\|$ (die Regularität von $\nabla f(x^*)$ impliziert, dass $\|\cdot\|_*$ wieder eine Norm ist), so lässt sich lineare Konvergenz in dieser Norm für alle $\bar{\eta} \in (0,1)$ zeigen. Weil alle Normen im $\mathbb{R}^n$ äquivalent sind, konvergiert somit das inexakte Newton-Verfahren in jeder Norm für $\bar{\eta} \in (0,1)$, aber nicht mehr notwendigerweise linear.

Beispiel 5.23
Sei $f(x) = x^2 - 2$ und $x^0 = 200$ die Startnäherung an $x^* = \sqrt{2}$. Das Abbruchkriterium ist $|f(x^k)| \le 10^{-15}$. Die Newton-Gleichung (5.1) wird mit der Splitting-Methode, $d^{k,i+1} = d^{k,i}/2 - f(x^k)/(4x^k)$ für $i = 0,1,2,\ldots$ und $d^{k,0} = 0$, gelöst. Gleichung (5.5) definiert das Abbruchkriterium für diese Fixpunktiteration (Für die Konvergenz vergleiche VL "wissenschaftliches Rechen"). Es wird $\eta_k \equiv 0.5$ (lineare Konvergenz), $\eta_k = 0.5(k+1)^{-3} \to 0^+$ (superlineare Konvergenz) und $\eta_k = 10^{-5}|f(x^k)|$ (quadratische Konvergenz) gewählt. Das Bild 5.7 zeigt die Verhältnisse $|x^{k+1}-x^*|/|x^k-x^*|$ und $|x^{k+1}-x^*|/|x^k-x^*|^2$.

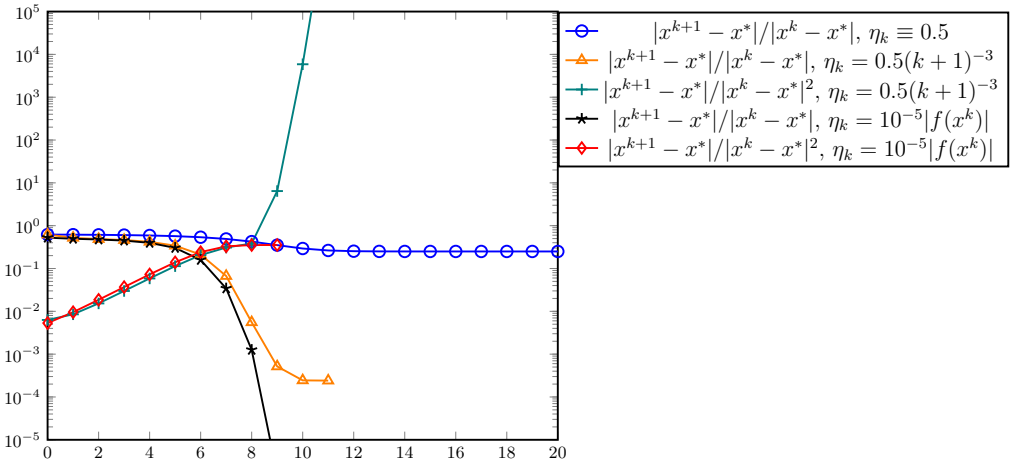


Abb. 5.7: Newton-Fehlerquotienten gegen Anzahl an Iterationsschritten abgetragen.

5.3 Weitere Nullstellenverfahren

Im Folgenden werden Modifikationen des Newton-Verfahrens und das mit dem Newton-Verfahren eng verwandte Sekanten-Verfahren betrachtet. Von zentraler Bedeutung ist die präzise Bestimmung der Konvergenzordnung des Fixpunkt-Verfahrens.

Satz 5.24
Die Funktion $\phi : D(\phi) \subset \mathbb{R} \to \mathbb{R}$ sei $(p+1)$-mal stetig differenzierbar und habe einen Fixpunkt $x^* \in D(\phi)$. Ferner sei $p \ge 2$ und

$$\phi'(x^*) = \ldots = \phi^{(p-1)}(x^*) = 0 \quad \text{und} \quad \phi^{(p)}(x^*) \neq 0.$$

Dann konvergiert die Fixpunktiteration $x_{k+1} = \phi(x_k)$ lokal gegen x^* mit der Ordnung p.

Beweis. Mit $\phi'(x^*) = 0$ und $\phi \in C^{p+1}$ gilt in einer abgeschlossenen Umgebung $U_\epsilon(x^*) \subset D(\phi)$ von x^* mit $\epsilon > 0$, dass $|\phi'(x)| \leq q < 1$ für alle $x \in U_\epsilon(x^*)$. Zusätzlich liefert der Mittelwertsatz, dass

$$|\phi(y) - \phi(x)| \leq \int_0^1 |\phi'(x + t(y-x))||x-y| \, dt \leq q|x-y| \quad \forall x, y \in U_\epsilon(x^*).$$

Speziell für $y = x^*$ gilt somit

$$|\phi(x) - x^*| = |\phi(x) - \phi(x^*)| \leq q|x - x^*| \leq q\epsilon < \epsilon \quad \forall x \in U_\epsilon(x^*).$$

Damit ist $\phi : U_\epsilon(x^*) \to U_\epsilon(x^*)$ eine Kontraktion und der Banachsche Fixpunktsatz liefert die (lineare) Konvergenz.
Das Taylorpolynom von ϕ entwickelt um x^* liefert

$$\begin{aligned} x_{k+1} = \phi(x_k) &= \phi(x^*) + \sum_{i=1}^{p} \frac{\phi^{(i)}(x^*)}{i!}(x_k - x^*)^i + \mathcal{O}(|x_k - x^*|^{p+1}) \\ &= x^* + \frac{\phi^{(p)}(x^*)}{p!}(x_k - x^*)^p + \mathcal{O}(|x_k - x^*|^{p+1}) \\ &= x^* + \frac{\phi^{(p)}(x^*)}{p!}(x_k - x^*)^p + \xi_k(x_k - x^*)^p \end{aligned}$$

mit $\xi_k = \xi_k(x_k) = \mathcal{O}(|x_k - x^*|)$. Da $\phi^{(p)}(x^*) \neq 0$ existiert ein $\rho > 0$ mit $U_\rho(x^*) \subseteq U_\epsilon(x^*)$ und

$$|\xi_k| \leq \frac{1}{2}\frac{|\phi^{(p)}(x^*)|}{p!} \quad \forall x_k \in U_\rho(x^*).$$

Damit folgt, dass

$$\frac{1}{2}\frac{|\phi^{(p)}(x^*)|}{p!}|x_k - x^*|^p \leq |x_{k+1} - x^*| \leq \frac{3}{2}\frac{|\phi^{(p)}(x^*)|}{p!}|x_k - x^*|^p \quad \forall x_k \in U_\rho(x^*)$$

und die Konvergenzordnung ist somit genau p. □

Bemerkung 5.25
Unter den Voraussetzungen von Satz 5.24 aber mit $\phi^{(p)}(x^*) = 0$ zulässig ist die Konvergenzordnung mindestens p.

5.3.1 Erneute Betrachtung des Newton-Verfahrens

Im folgenden sei zunächst $f : \mathbb{R} \to \mathbb{R}$ hinreichend glatt mit einer Nullstelle x^*. Das Problem $f(x) = 0$ wird in das Fixpunktproblem

$$x = x + g(x)f(x) =: \phi(x)$$

umgeschrieben. Hier ist g eine hinreichend glatte Funktion mit $g(x) \neq 0$ für alle $x \in U_\epsilon(x^*)$. Damit Satz 5.24 angewendet werden kann, muss gelten

$$\phi'(x^*) = 1 + g'(x^*)f(x^*) + g(x^*)f'(x^*) = 1 + g(x^*)f'(x^*) = 0,$$

also $g(x^*) = \frac{-1}{f'(x^*)}$. Die naheliegende Wahl $g(x) = \frac{-1}{f'(x)}$ liefert das ein dimensionale Newton-Verfahren

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

Mit Satz 5.24 kann jetzt ein leichter aber dafür schwächerer Alternativbeweis zur Konvergenz des Newton-Verfahrens geführt werden.

Satz 5.26
Sei $f \in C^3(a,b)$ mit $x^* \in (a,b)$, $f(x^*) = 0$ und $f'(x^*) \neq 0$. Dann konvergiert das Newton-Verfahren lokal (mindestens) quadratisch gegen x^*.

Beweis. Sei $\phi(x) = x - \frac{f(x)}{f'(x)}$. Dann ist

$$\phi'(x) = 1 - \frac{f'(x)f'(x) - f(x)f''(x)}{[f'(x)]^2} = \frac{f(x)f''(x)}{[f'(x)]^2}$$
$$\phi''(x) = \frac{[f'(x)f''(x) + f(x)f^{(3)}(x)][f'(x)]^2 - f(x)f''(x)\cdot 2f'(x)f''(x)}{[f'(x)]^4}.$$

Einsetzten von x^* liefert

$$\phi'(x^*) = 0 \quad \text{und} \quad \phi''(x^*) = \frac{f''(x^*)}{f'(x^*)}.$$

Mit Satz 5.24 und der darauf folgenden Bemerkung folgt die Behauptung. □

Als nächstes soll die Bedingung $f'(x^*) \neq 0$ fallen gelassen werden, aber dennoch soll das Newton-Verfahren, also

$$\phi(x) = x - \frac{f(x)}{f'(x)},$$

für $U_\epsilon(x^*) \ni x \neq x^*$ weiterhin wohl definiert sein.

Beispiel 5.27
Sei $f(x) = x^\alpha$ mit $\alpha > 1$. Dann konvergieren die Newton-Iterierten

$$x_{k+1} = x_k - \frac{x_k^\alpha}{\alpha x_k^{\alpha-1}} = \left(1 - \frac{1}{\alpha}\right) x_k = \ldots = \left(1 - \frac{1}{\alpha}\right)^{-(k+1)} x_0$$

linear gegen die Nullstelle $x^* = 0$.

Satz 5.28
Sei $p \in \mathbb{N} \setminus \{1\}$, $f \in C^{p+1}(a,b)$ und x^* eine p-fache Nullstelle von f, d.h.

$$f(x^*) = f'(x^*) = \ldots = f^{(p-1)}(x^*) = 0 \quad \text{und} \quad f^{(p)}(x^*) \neq 0.$$

Dann konvergiert das Newton-Verfahren lokal linear mit Fehlerkonstante $C = 1 - p^{-1}$.

Beweis. Die Taylorentwicklung liefert

$$f(x) = f(x^*) + \sum_{i=1}^{p} \frac{f^{(i)}(x^*)}{i!}(x - x^*)^i + \mathcal{O}(|x - x^*|^{p+1})$$
$$= \frac{f^{(p)}(x^*)}{p!}(x - x^*)^p + \mathcal{O}(|x - x^*|^{p+1})$$

und analog

$$f'(x) = \frac{f^{(p)}(x^*)}{(p-1)!}(x - x^*)^{p-1} + \mathcal{O}(|x - x^*|^p).$$

Damit folgt

$$\phi(x) = x - \frac{1}{p}(x - x^*) + \mathcal{O}(|x - x^*|^2)$$

und für das Newton-Verfahren $x_{k+1} = \phi(x_k)$ gilt somit

$$|x_{k+1} - x^*| = |\phi(x_k) - x^*| = \left(1 - \frac{1}{p}\right)|x_k - x^*| + \mathcal{O}(|x_k - x^*|^2).$$

Damit ist die Konvergenz lokal linear mit Fehlerreduktionsfaktor $C = 1 - p^{-1} \to 1$ für $p \to \infty$. □

Bemerkung 5.29
Wird statt $\phi(x) = x - \frac{f(x)}{f'(x)}$ jetzt $\phi(x) = x - p\frac{f(x)}{f'(x)}$ gewählt, also das Newton-Update skaliert, so erhalten wir analog zum obigen Satz, dass

$$\phi(x) = x - \frac{p}{p}(x - x^*) + \mathcal{O}(p|x - x^*|^2) = x^* + \mathcal{O}(|x - x^*|^2)$$

und die Konvergenzordnung ist wieder quadratisch. Jedoch sind beide Verfahren, mit und ohne Skalierung, numerisch instabil, weil $f(x)$ und $f'(x)$ für $x \to x^*$ klein werden und zu Auslöschung führen können. Damit sind beide Varianten sehr Rundungsfehler empfindlich.

Das Aufwendige beim (n-dimensionalen) Newton-Verfahren ist die Berechnung von $\nabla f(x^{(k)})$ sowie von $\nabla f(x^{(k)})^{-1} f(x^{(k)})$. Beim vereinfachten Newton-Verfahren ist die Rechenvorschrift

$$x^{(k+1)} = x^{(k)} - \nabla f(\bar{x})^{-1} f(x^{(k)}) \tag{5.6}$$

mit einer festen Jakobimatrix $\nabla f(\bar{x})$.

Satz 5.30
Sei $f \in C^1(\mathbb{R}^n; \mathbb{R}^n)$, $x^* \in \mathbb{R}^n$ mit $f(x^*) = 0$ und $\bar{x} \in \mathbb{R}^n$. Gilt $\|I - \nabla f(\bar{x})^{-1} \nabla f(x^*)\| < 1$ in einer Matrixnorm, welche zu einer Vektornorm verträglich ist, so konvergiert das vereinfachte Newton-Verfahren (5.6) lokal linear.

Beweis. Offensichtlich gilt mit $f(x^*) = 0$ und $\phi(x) := x - \nabla f(\bar{x})^{-1} f(x)$, dass $\phi(x^*) = x^*$ und $\nabla\phi(x) = I - \nabla f(\bar{x})^{-1} \nabla f(x)$. Mit den Voraussetzungen folgt jetzt, dass $\|\nabla\phi(x)\| \leq q < 1$ in einer Umgebung von x^* ist. Die Behauptung folgt somit mit dem Banachschen Fixpunktsatz. □

Gerne wird $\bar{x}$ alle K Schritte durch die aktuelle Iterierte $x^{(jK)}$ aktualisiert, denn damit ist häufig $\|I - \nabla f(\bar{x})^{-1} \nabla f(x^*)\| \ll 1$ und es kann eine „immer schneller werdende“ lineare Konvergenz beobachtet werden.

5.3.2 Das Sekanten-Verfahren

Die Idee beim Sekanten-Verfahren ist die Ableitung im Newton-Verfahren durch einen Differenzenquotienten zu ersetzen und bereits erfolge Funktionsauswertungen wiederzuverwenden. Also

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

Ausgehend von den zwei Startwerten x_0 und x_1 erhalten wir die nächste Iterierte des Sekanten-Verfahrens durch

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} f(x_k) = \frac{x_{k-1} f(x_k) - x_k f(x_{k-1})}{f(x_k) - f(x_{k-1})}. \tag{5.7}$$

Tatsächlich wird jetzt die Sekante anstelle der Tangente zur linearen Approximation von f verwendend, siehe auch Abbildung 5.8.

Lemma 5.31 (Verallgemeinerter Mittelwertsatz)
Für $a, b \in \mathbb{R}$ mit $a < b$ seien die Funktionen $f, g \in C^0[a, b]$ differenzierbar auf (a, b). Gilt $g'(x) \neq 0$ für alle $x \in (a, b)$, dann existiert eine Stelle $\zeta \in (a, b)$ mit

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(\zeta)}{g'(\zeta)}.$$

Satz 5.32 (Sekanten-Verfahren)
Sei $f \in C^2(a, b)$ mit $x^* \in (a, b)$, $f(x^*) = 0$, $f'(x^*) \neq 0$ und $f''(x^*) \neq 0$. Dann konvergiert das Sekanten-Verfahren lokal gegen x^* mit der Ordnung

$$p = \frac{1}{2}\left(1 + \sqrt{5}\right) \approx 1.61803.$$

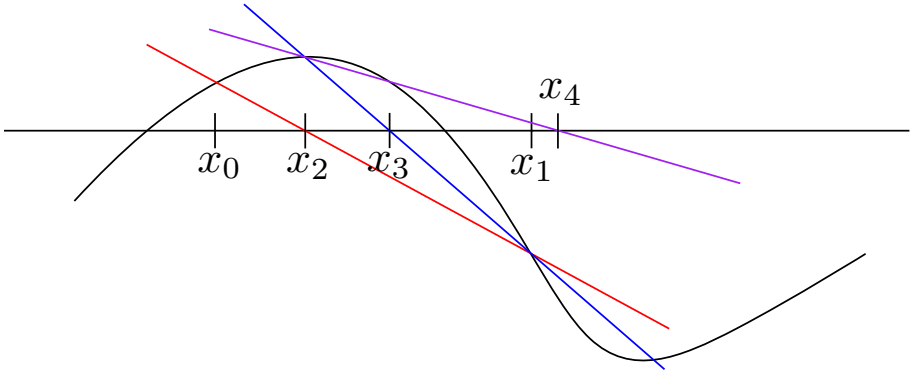


Abb. 5.8: Visualisierung des Sekanten-Verfahrens.

Beweis. Wegen $f \in C^2(a,b)$ und $f'(x^*) \neq 0$ ist f lokal um $x^* \in (a,b)$ injektiv und damit ist das Sekanten-Verfahren wohldefiniert. Der Konvergenzbeweis erfolgt in drei Schritten:
Schritt 1: Sei $e_k := x^* - x_k$. Mit (5.7) folgt

$$e_{k+1} = e_k - \frac{e_k - e_{k-1}}{f(x_k) - f(x_{k-1})} f(x_k) = \frac{e_{k-1} f(x_k) - e_k f(x_{k-1})}{f(x_k) - f(x_{k-1})},$$

also

$$\frac{e_{k+1}}{e_k e_{k-1}} = \frac{1}{f(x_k) - f(x_{k-1})} \left(\frac{f(x_{k-1})}{x_{k-1} - x^*} - \frac{f(x_k)}{x_k - x^*} \right) = \frac{g(x_k) - g(x_{k-1})}{f(x_k) - f(x_{k-1})}$$

mit

$$g(x) := \frac{-f(x)}{x - x^*} \quad \text{und} \quad g'(x) = \frac{-f'(x)(x - x^*) + f(x)}{(x - x^*)^2}.$$

Die Funktion g kann durch $g(x^*) := -f'(x^*)$ und $g'(x^*) := -0.5 f''(x^*)$ stetig differenzierbar in $x = x^*$ fortgesetzt werden. Aus dem verallgemeinerten Mittelwertsatz Lemma 5.31 folgt die Existenz eines ζ_k zwischen x_{k-1} und x_k, so dass mit der Taylor-Entwicklung folgt, dass

$$\frac{e_{k+1}}{e_k e_{k-1}} = \frac{g'(\zeta_k)}{f'(\zeta_k)} = \frac{1}{f'(\zeta_k)} \frac{f(\zeta_k) + f'(\zeta_k)(x^* - \zeta_k)}{(\zeta_k - x^*)^2} = -\frac{1}{2} \frac{f''(\eta_k)}{f'(\zeta_k)} \tag{5.8}$$

mit einem η_k zwischen x^* und ζ_k. In einer Umgebung von x^* ist nach Voraussetzung die rechte Seite von (5.8) beschränkt und wir erhalten

$$|e_{k+1}| \leq (C \cdot |e_{k-1}|)\, |e_k|$$

für eine Konstante $C > 0$. Sind x_0 und x_1 hinreichend nah bei x^*, d.h. $|e_0| < C^{-1}$ und $|e_1| < C^{-1}$, so konvergiert der Fehler monoton und mindestens linear gegen Null.

Schritt 2: Sei

$$\epsilon_k := \frac{|e_k|}{|e_{k-1}|^p}$$

mit $p = \frac{1}{2}\left(1 + \sqrt{5}\right)$ und $\gamma_k := \log \epsilon_k$. Wegen

$$\frac{1}{p} = \frac{2}{1 + \sqrt{5}} = \frac{1}{2}\left(\sqrt{5} - 1\right) = p - 1 \tag{5.9}$$

folgt mit (5.8), dass

$$\epsilon_{k+1} = \frac{|e_{k+1}|}{|e_k|^p} = \frac{|e_{k+1}|}{|e_k|} |e_k|^{1-p} = \frac{|e_{k+1}|}{|e_k|} |e_k|^{-(p-1)} = \alpha_k |e_{k-1}| |e_k|^{-1/p} = \alpha_k \epsilon_k^{-1/p}$$

mit

$$\alpha_k := \frac{|f''(\eta_k)|}{2|f'(\zeta_k)|}.$$

Das heißt

$$\gamma_{k+1} = \log \alpha_k - \frac{\gamma_k}{p}. \tag{5.10}$$

Schritt 3: Jetzt liefert die Rekursionsformel (5.10), dass

$$\gamma_{k+1} = \log \alpha_k - \frac{\gamma_k}{p} = \log \alpha_k - \frac{\log \alpha_{k-1}}{p} + \frac{\gamma_{k-1}}{p^2} = \ldots = \sum_{j=1}^{k} \left(\frac{-1}{p}\right)^{k-j} \log \alpha_j + \left(\frac{-1}{p}\right)^k \gamma_1.$$

Nach Voraussetzung existiert eine Umgebung um x^* mit $|f''|$ und $|f'|$ nach unten durch ein $\delta > 0$ und nach oben beschränkt. Mit der monotonen Konvergenz aus dem ersten Schritt folgt jetzt die Existenz eines $a > 0$, so dass $|\log \alpha_j| \leq a < \infty$ für alle $j \in \mathbb{N}$. Wegen $1/p < 1$ folgt somit

$$|\gamma_{k+1}| < |\gamma_1| + a \sum_{j=1}^{\infty} p^{-j} =: \tilde{C} < \infty$$

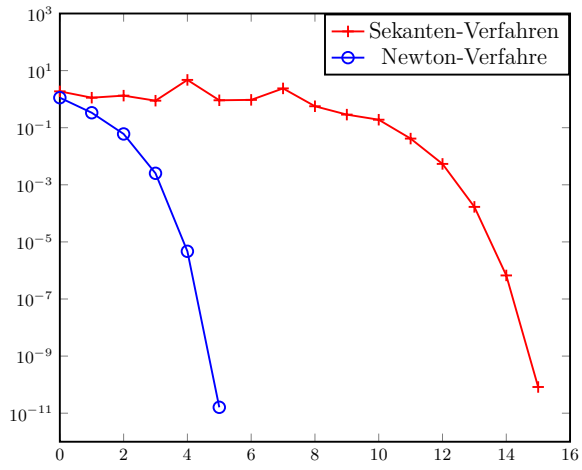
für alle $k \in \mathbb{N}_0$. Somit ist $\epsilon_k = \exp(\gamma_k) \in (\exp(-\tilde{C}), \exp(\tilde{C}))$ und wir erhalten mit $\epsilon_k := \frac{|e_k|}{|e_{k-1}|^p}$, dass

$$e^{-\tilde{C}} |e_{k-1}|^p \leq |e_k| \leq e^{\tilde{C}} |e_{k-1}|^p \qquad \forall k \in \mathbb{N}.$$

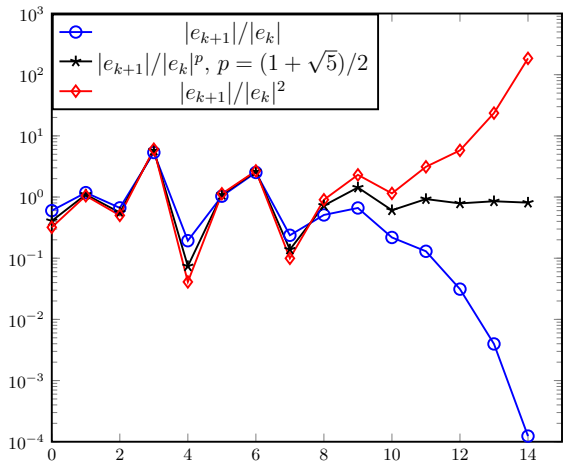
Das Sekanten-Verfahren konvergiert also genau mit Ordnung p. □

Das Sekanten-Verfahren konvergiert langsamer als das Newton-Verfahren, braucht also mehr Iterationsschritte um die selbe Fehlertoleranz zu unterschreiten, jedoch ist auch jeder Sekanten-Schritt billiger als ein Newton-Schritt. Damit ist das Sekanten-Verfahren oft konkurrenzfähig zum Newton-Verfahren, jedoch ist es wegen der Gefahr der Auslöschung numerisch weniger stabil.

Beispiel 5.33
Sei $f(x) = \cos(x)\cosh(x)+1$ mit $f'(x) = \cos(x)\sinh(x)-\cosh(x)\sin(x)$, $f(x^*) = 0$ für $x^* \approx 1.875104068711961$. Weiters sei $x_0 = 3$ und (beim Sekanten-Verfahren) $x_1 = 0$.



(a) Sekanten- bzw. Newton-Fehler gegen Anzahl an Iterationsschritten abgetragen



(b) Fehlerquotienten des Sekanten-Verfahrens gegen Anzahl an Iterationsschritten abgetragen

Abb. 5.9: Verhalten des Sekanten-Verfahrens.

6

Verwendete und weiterführende Literatur

[1] W. Walter: Gewöhnliche Differentialgleichungen. Eine Einführung, 2000, Springer
[2] H. Heuser: Gewöhnliche Differentialgleichungen. Einführung in Lehre und Gebrauch, 2004, Teubner
[3] H. Heuser: Lehrbuch der Analysis. Teil 2, 2002, Vieweg + Teubner
[4] H.-R. Schwarz und N. Köckler: Numerische Mathematik, 2004, Vieweg + Teubner
[5] C. Geiger und C. Kanzow: Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben, 1999, Springer