Numerische Mathematik I

Wintersemester 2007/08

Dr. Serge Kräutle

geTEXt von A. Schleich und Ch. Basting

Vorlesungsskript

Department für Mathematik

AM1

Friedrich–Alexander–Universität Erlangen–Nürnberg

3 Nichtlineare Gleichungen

Problemstellung:

Definition 3.1 *Sei* $(X, \|\cdot\|)$ *ein normierter Vektorraum,* $U \subset X,\ \Phi : U \to X,\ b \in X$.
Das Problem

$$\text{“Finde } x \in U \text{ mit } \Phi(x) = 0\text{”} \tag{3.1}$$

heißt (allgemeines, nichtlineares) Nullstellenproblem*; eine Lösung heißt* Nullstelle *von* Φ.
Das Problem

$$\text{“Finde } x \in U \text{ mit } \Phi(x) = x\text{”} \tag{3.2}$$

heißt Fixpunktproblem*; eine Lösung heißt* Fixpunkt *von* Φ.
Ferner betrachten wir noch die allgemeine nichtlineare Gleichung (“Gleichungssystem” im Fall $X = \mathbb{R}^n$*)*

$$\text{“Finde } x \in U \text{ mit } \Phi(x) = b\text{”} \tag{3.3}$$

Die Probleme (3.1), (3.2), (3.3) sind “gleich schwer” bzw. “äquivalent” in dem Sinne, dass jedes Problem des einen Typus in ein Problem jedes anderen Typus umgewandelt werden kann:

(3.1) $\to$ (3.2):	Setze z.B.: $\tilde{\Phi}(x) := \Phi(x) + x$
	(auch $\tilde{\Phi}(x) := \alpha\Phi(x) + x,\ \alpha \in \mathbb{R} \setminus \{0\}$ ist möglich)
(3.2) $\to$ (3.3):	Setze z.B.: $\tilde{\Phi}(x) := \Phi(x) - x + b$
(3.3) $\to$ (3.1):	Setze z.B.: $\tilde{\Phi}(x) := \Phi(x) - b$

Daher können wir uns im folgenden zunächst auf Fixpunktprobleme konzentrieren. Im allgemeinen Fall (wenn also f echt nichtlinear ist) gibt es keinen Algorithmus, der nach endlich vielen Schritten (3.2) löst. Wir werden stattdessen ein *iteratives* Lösungsverfahren der Form

$$x^{(n+1)} := \Phi(x^{(n)})$$

herleiten. Wichtigstest Hilfsmittel: Der Fixpunktsatz vom Banach.

3.1 Fixpunktiterationen

Satz 3.2 *(Fixpunktsatz von Banach, 1922) Sei* $(X, \|\cdot\|)$ *ein Banach-Raum, sei* $U \subset X$ *eine abgeschlossene Teilmenge; Sei* $\Phi : U \to X$ *mit*

$$\Phi(U) \subseteq U \quad [\Phi \textit{ heißt } \text{selbstabbildend}]. \tag{3.4}$$

Es existiere ein $0 < k < 1$ *so dass*

$$\|\Phi(x) - \Phi(y)\| \le k\|x - y\|\ \forall x, y \in U \tag{3.5}$$

[“Kontraktionseigenschaft”].
Dann gilt:

1. *Es gibt* genau einen *Fixpunkt* $x^* \in U$ *von* Φ.

2. *Für* beliebigen *Startwert* $x^{(0)} \in U$ *konvergiert die Folge*

$$x^{(n+1)} := \Phi(x^{(n)}) \tag{3.6}$$

gegen x^*.

3. *Es gelten*

$$\|x^{(n)} - x^*\| \leq \frac{k}{1-k}\|x^{(n)} - x^{(n-1)}\| \quad \textit{a posteriori-Fehlerabschätzung} \tag{3.7}$$

$$\leq \frac{k^n}{1-k}\|x^{(1)} - x^{(0)}\| \quad \textit{a priori-Fehlerabschätzung} \tag{3.8}$$

Bemerkung 3.3 *:*

- *(3.6) heißt* Fixpunktiteration*, k heißt* Kontraktionskonstante*; der Wert von k ist entscheidend für die Konvergenzgeschwindigkeit.*
- *Oft wird der Satz mit $U = X$ angewendet. (3.4) ist dann trivialerweise erfüllt.*
- *Der Satz lässt sich auf vollständige metrische Räume übertragen.*
- *Über die Probleme aus Definition 3.1 im $\mathbb{R}^n$ hinaus wird der Satz auch z.B. angewendet, um die Existenz und Eindeutigkeit von Lösungen von* Anfangswertproblemen von gewöhnlichen Differenzialgleichungen $y'(t) = f(t, y(t))$, $y(t) = y$*, bei Lipschitzstetigem f, zu zeigen; dabei* $(X, \|\cdot\|) = (\mathcal{C}([a,b]), \|\cdot\|_\infty)$

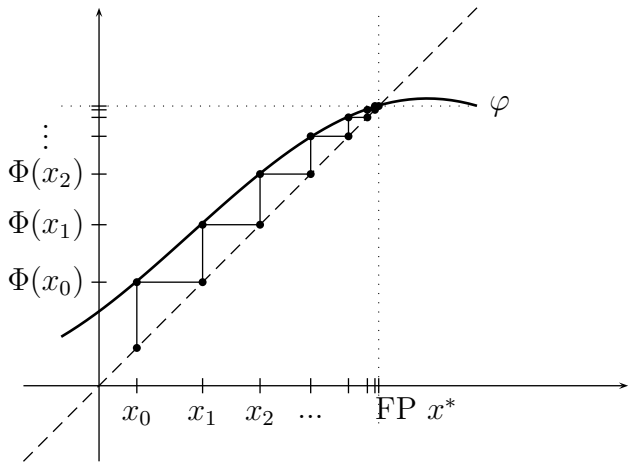

Abbildung 3: konvergente Fixpunktiteration, $X = \mathbb{R}$

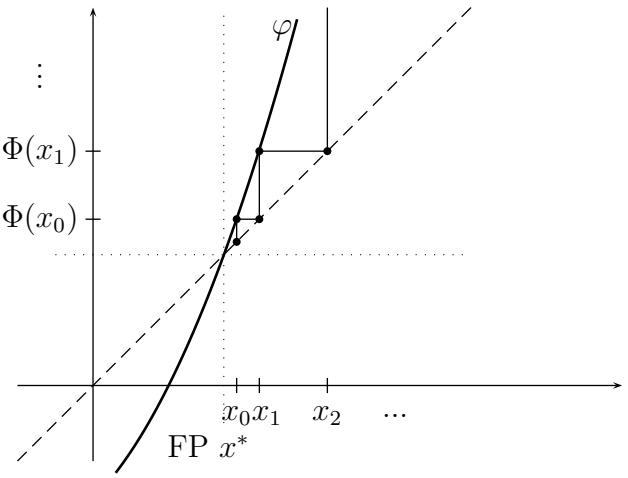

Abbildung 4: divergente Fixpunktiteration, $X = \mathbb{R}$

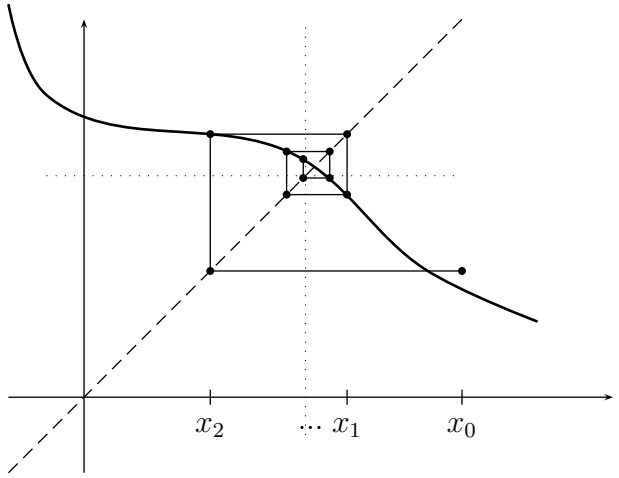

Abbildung 5: konvergente Fixpunktiteration, $X = \mathbb{R}$

Beweis : Die Folge $(x^{(n)})$ ist wegen (3.4) wohldefiniert. Wir zeigen: $(x^{(n)})_{n\in\mathbb{N}}$ ist CAUCHY-Folge:

$$
\begin{aligned}
\|x^{(n+1)} - x^{(n)}\| &\underset{\text{Def. Folge}}{=} \|\Phi(x^{(n)}) - \Phi(x^{(n-1)})\| \underset{\text{Kontraktion}}{\leq} k\|x^{(n)} - x^{(n-1)}\| \\
&\underset{\text{Def. Folge}}{\leq} k\|\Phi(x^{(n-1)}) - \Phi(x^{(n-2)})\| \underset{\text{Kontraktion}}{\leq} k^2\|x^{(n-1)} - x^{(n-2)}\| \\
&\ldots \\
&\leq k^n\|x^{(1)} - x^{(0)}\| \qquad (3.9) \\
\Rightarrow \|x^{(n+l)} - x^{(n)}\| &\underset{\text{Teleskopsumme}}{\leq} \|x^{(n+l)} - x^{(n+l-1)}\| + \ldots + \|x^{(n+1)} - x^{(n)}\| \\
&\underset{(3.9)}{\leq} (k^{n+l-1} + \ldots + k^n)\|x^{(1)} - x^{(0)}\| \\
&\leq k^n \left(\sum_{i=0}^{\infty} k^i\right) \|x^{(1)} - x^{(0)}\| \\
&= \underbrace{k^n \frac{1}{1-k}\|x^{(1)} - x^{(0)}\|}_{\leq \epsilon \text{ für } n \text{ hinreichend groß}} \Rightarrow \text{CAUCHY-Eigenschaft}
\end{aligned}
$$

Da $(X, \|\cdot\|)$ vollständig und $U \subset X$ abgeschlossen, ist $(x^{(n)})$ konvergent gegen ein $x^* \in U$. Wir zeigen: x^* ist Fixpunkt:

$$x^* = \lim_{n\to\infty} x^{(n)} \underset{\text{Def. Folge}}{=} \lim_{n\to\infty} \Phi(x^{(n-1)}) \underset{\Phi \text{ stetig}}{=} \Phi(\lim_{n\to\infty} x^{(n-1)}) = \Phi(x^*) \;\Rightarrow\; x^* \text{ist Fixpunkt}$$

Eindeutigkeit: Angenommen $x^*, x^{**} \in U$ seien Fixpunkte, dann:

$$\Rightarrow \quad \|x^* - x^{**}\| \underset{\text{Fixpunkt}}{=} \|\Phi(x^*) - \Phi(x^{**})\| \underset{\text{Kontraktion}}{\leq} k\|x^* - x^{**}\|$$

$$\underset{k<1}{\Rightarrow} \quad \|x^* - x^{**}\| = 0 \quad \Rightarrow \quad x^* = x^{**}$$

Zu den Abschätzungen (3.7) und (3.8):

$$\begin{aligned} \|x^{(n)} - x^*\| &\underset{\text{Def. Folge}}{=} \|\Phi(x^{(n-1)}) - \Phi(x^*)\| \underset{\text{Kontraktion}}{\leq} k\|x^{(n-1)} - x^*\| \qquad (3.10)\\ &\leq k \cdot (\|x^{(n-1)} - x^{(n)}\| + \|x^{(n)} - x^*\|) \\ \Rightarrow \|x^{(n)} - x^*\| &\leq \frac{k}{1-k}\|x^{(n-1)} - x^{(n)}\| \qquad \rightsquigarrow (3.7) \\ &\underset{(3.9)}{\leq} \frac{k^n}{1-k}\|x^{(1)} - x^{(0)}\| \qquad \rightsquigarrow (3.8) \end{aligned}$$

□

Lemma 3.4 *Eine stetig differenzierbare Abbildung $\Phi : U \to R$, $U \subset \mathbb{R}$ offenes Intervall, ist genau dann Lipschitz-stetig mit Lipschitz-Konstante $k > 0$, d.h.*

$$|\Phi(x) - \Phi(y)| \leq k|x - y| \quad \forall x, y \in U, \tag{3.11}$$

wenn

$$|\Phi'(x)| \leq k \quad \forall x \in U. \tag{3.12}$$

Beweis:

„⇐“ Es gelte (3.12). Der Mittelwertsatz liefert: $\Phi(x) = \Phi(y) + \Phi'(\zeta)(x - y)$, $\zeta \in U$.

$$\Rightarrow \quad |\Phi(x) - \Phi(y)| \leq |\Phi'(\zeta)| \cdot |x - y| \leq k \cdot |x - y|$$

„⇒“ Sei (3.12) falsch, es gibt also $x \in U$ mit $|\Phi'(x)| > k$. Dann gibt es (da f' stetig) eine zusammenhängende Umgebung $\tilde{U} \subseteq U$ von x, so dass $|\Phi'(\zeta)| > k \;\forall \zeta \in \tilde{U}$.

$$\Rightarrow \quad \text{für} \quad y \in \tilde{U},\ y \neq x \ |\Phi(x) - \Phi(y)| = |f'(\zeta)| \cdot |x - y| > k|x - y|$$

□

Für *stetig differenzierbares* $\Phi : U \to \mathbb{R}$ kann die Bedingung (3.5) also durch die einfacher überprüfbare Bedingung

$$k := \sup_{x\in U} |\Phi'(x)| < 1 \tag{3.13}$$

ersetzt werden.

Beispiel 3.5 *Gesucht sind die Nullstellen von $f(x) = \cos x - 2x$ also $X = \mathbb{R}$, $U = \mathbb{R}$. Umwandlung in ein Fixpunktproblem:*

erster Versuch:

$$\begin{aligned} \cos x - 2x &\overset{!}{=} 0 \quad | \quad +x \\ \underbrace{\cos x - x}_{=:\Phi(x)} &= x \end{aligned}$$

Prüfe die Kontraktionseigenschaft. Wegen $\Phi \in \mathcal{C}^1(\mathbb{R}) \Rightarrow$ Lemma 3.4 ist anwendbar:

$$\Phi'(x) = -(\sin x + 1)$$

Erfüllt *nicht* (3.13)! Problem! Kann diese Problem durch Wahl von kleinerem $U \subsetneq \mathbb{R}$ behoben werden?
Wir erwarten Nullstelle von f im Bereich $[0, \frac{\pi}{2}]$, da $f(0) = 1 > 0,\ f(\frac{\pi}{2}) = -\pi < 0$. Aber selbst auf $[0, \frac{\pi}{2}] =: U$ ist $|\Phi'(x)| \geq 1$, da sin dort größer gleich 0 ist.

neuer Versuch:

$$f(x) = \cos x - 2x \overset{!}{=} 0 \qquad | \cdot \alpha,\ \alpha \neq 0,\ +x \quad \Leftrightarrow \quad \underbrace{\alpha(\cos x - 2x) + x}_{=:\Phi(x)} = x$$

Kontraktionseigenschaft :

$$\Phi'(x) = \overbrace{-\alpha \sin x}^{\in[-\alpha,\alpha]} - 2\alpha + 1 \in [1 - 3\alpha, 1 - \alpha] \overset{!}{\subseteq} (-1, 1)$$

Wähle zum Beispiel $\alpha := \frac{1}{2}$. Dann ist $\Phi'(x) \in [-\frac{1}{2}, \frac{1}{2}]$ und damit gilt für die Kontraktionskonstante

$$k = \sup_{x \in U} |\Phi'(x)| \leq \frac{1}{2}$$

($\alpha \in (0, \frac{2}{3})$, damit $k < 1$ gilt).
Der Fixpunktsatz von Banach liefert also: f hat genau eine Nullstelle x^*.
Die Folge $x^{(n+1)} := \Phi(x^{(n)}) = \frac{1}{2} \cos x^{(n)}$ konvergiert gegen x^* mit Kontraktionsrate $k = \frac{1}{2}$ ($\sim$ ca. 3 Iterationsschritte pro gewonnene Dezimalstelle), bei beliebigem Startwert $x^{(0)} \in U = \mathbb{R}$.

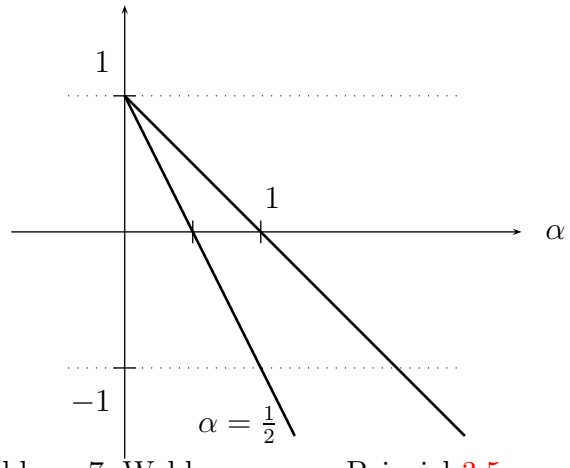

Abbildung 7: Wahl von α aus Beispiel 3.5, zweiter Versuch

3.2 Konvergenzordnung von Fixpunktverfahren und das Newton-Verfahren für skalare Nullstellenprobleme

Im Beispiel 3.5 haben wir das Nullstellenproblem $f(x) = 0$ umgewandelt in ein Fixpunktproblem $\Phi(x) = x$. Wir haben

$$\Phi(x) := x + \alpha f(x) \tag{3.14}$$

gewählt. Wie ist $\alpha \in \mathbb{R},\ \alpha \neq 0$, allgemein zu wählen? Für $\alpha := 1$ wäre die „Kontraktionsrate“

$$k := \sup_{x \in U = \mathbb{R}} |\Phi'(x)| > 1$$

$\Rightarrow$ Divergenz, für $\alpha := \frac{1}{2}$ ist $k = \frac{1}{2} < 1$. Welches α ist optimal? Oder, noch allgemeiner als (3.13): Wie soll Φ (in Abhängigkeit von f) gewählt werden? Bedingungen an Φ :

(a) **Konsistenz:** x^* soll genau dann Fixpunkt von Φ sein, wenn x^* Nullstelle von f ist.

(b) **schnelle Konvergenz:** $k := \sup\limits_{x \in U} |\Phi'(x)|$, wobei U eine Umgebung von x^* ist, sollte möglichst klein sein, bzw.

(b^*) **asymptotisch** für $x^{(k)} \approx x^*$, sollte $k^* := |\Phi'(x^*)|$, die sogenannte asymptotische Kontraktionskonstante, möglichst klein sein.

Wenn wir Φ von der Form (3.14) wählen, so ist (a) erfüllt; und für (b) bzw. (b^*) wäre $\alpha := \frac{1}{f'(x^*)}$ günstig (falls $f'(x^*) \neq 0$), da dann

$$\Phi(x) = x - \frac{1}{f'(x^*)} f(x) \quad \Rightarrow \quad \Phi'(x^*) = 1 - \frac{f'(x^*)}{f'(x^*)} = 0$$

(optimale asymptotische Kontraktionskonstante).

[Wir können erwarten, dass $\Phi'(x)$ auch in einer Umgebung U von x^* *klein* ist und damit die Kontraktionskonstante k klein ist.]

Aber: x^* ist ja a priori unbekannt, daher ist die Wahl $\alpha := -\frac{1}{f'(x^*)}$ unpraktikabel. Eine praktisch durchführbare Wahl, die um so besser ist, je näher $x^{(n)}$ bei x^* liegt, ist es, bei der Berechnung von $x^{(n+1)}$ aus $x^{(n)}$,

$$\alpha = \alpha(n) := -\frac{1}{f'(x^{(n)})}$$

zu wählen, also:

$$\begin{aligned} \Phi(x) &:= x - \frac{f(x)}{f'(x)} \\ x^{(n+1)} &:= \Phi(x^{(n)}) = x^{(n)} - \frac{f(x^{(n)})}{f'(x^{(n)})} \end{aligned} \tag{3.15}$$

Dann ist

$$\Phi'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2},$$

und damit $\Phi'(x^*) = 0$ (falls $f'(x^*) \neq 0$).

Diese Iteration zur Bestimmung von Nullstellen von f heißt NEWTON-*Iteration*/ NEWTON-*Verfahren*. Nach obigen Überlegungen erwarten wir für das NEWTON-Verfahren besonders gute Konvergenzeigenschaften. Dies werden wir im folgenden genauer untersuchen (Def. 3.6 - Satz 3.10).

Zuvor noch eine *geometrische* Interpretation des NEWTON-Verfahrens (für skalare Probleme):
Sei $x^{(n)}$ eine Nährung für die Nullstelle x^* von f. Anstatt die Nullstelle von f direkt zu berechnen, wird f approximiert durch eine Funktion $\tilde{f}$, deren Nullstelle $\tilde{x}^*$ man *leicht* berechnen kann. $\tilde{x}^*$ wird als neue, bessere Approximation $x^{(n+1)}$ an die Nullstelle x^* von f verwendet. Als $\tilde{f}$ nehmen wir die *Tangente* (=Linearisierung von f) an f in der Stelle $x^{(n)}$ (setzt $f \in \mathcal{C}^1$ voraus, außerdem $f'(x^{(n)}) \neq 0$). Steigungsdreieck: $f'(x^{(n)}) = \frac{f(x^{(n)})-0}{x^{(n)}-x^{(n+1)}} \Leftrightarrow \quad x^{(n)} - x^{(n+1)} = \frac{f(x^{(n)})}{f'(x^{(n)})} \quad \Leftrightarrow$ (3.15)

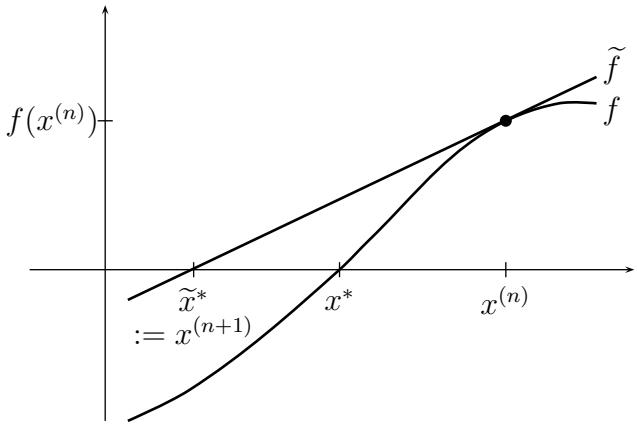

Abbildung 8: geometrische Interpretation des NEWTON-Verfahrens

Um das Konvergenzverhalten des NEWTON-Verfahrens beschreiben zu können, sind folgende Definitionen nützlich:

Definition 3.6 *Sei $(X, \|\cdot\|)$ ein normierter Vektorraum, $U \subset X$ offen. Eine Folge $x^{(n+1)} := \Phi(x^{(n)})$, $\Phi : U \to U$ heißt* lokal konvergent *gegen $x^* \in U$, falls es eine Umgebung $V \subseteq U$ von x^* gibt, so dass für* beliebigen *Startwert $x^{(0)} \in V$ die Folge gegen x^* konvergiert. Die Folge heißt* global konvergent*, falls für* beliebigen *Startwert $x^{(0)} \in U$ die Folge gegen x^* konvergiert.*

Wir wissen (Satz 3.2): Unter den Vorraussetzungen des Fixpunktsatzes von Banach ist die Fixpunktiteration *global konvergent*.

Definition 3.7 *Sei $(X, \|\cdot\|)$ ein normierter Vektorraum, $U \subseteq X$ offen, $x^* \in U$, $x^{(0)} \in U$, $x^{(n+1)} := \Phi(x^{(n)}), p \in \mathbb{R}, p \geq 1$. Falls es ein $c > 0$ gibt mit*

$$\|x^{(n+1)} - x^*\| \leq c\|x^{(n)} - x^*\|^p \quad \forall n \in \mathbb{N} \tag{3.16}$$

und, nur im Fall $p = 1$, zusätzlich $c < 1$ gilt, so heißt die Folge $(x^{(n)})$ (lokal) konvergent von der Ordnung p.

Im Fall $p = 1$ sprechen wir von (lokal) ***linearer*** *Konvergenz.*
Im Fall $p = 2$ sprechen wir von (lokal) ***quadratischer*** *Konvergenz.*
Im Fall $p = 3$ sprechen wir von (lokal) ***kubischer*** *Konvergenz.*

Bemerkung: In Büchern wird das Wort "lokal" in obiger Definition meist weggelassen. Das ist aber irreführend, denn die Bedingung (3.16) impliziert im Allgemeinen nicht für beliebige Startwerte $x^{(0)} \in U$ Konvergenz der Folge $(x^{(n)})$ (=„globale Konvergenz"), sondern nur für $x^{(0)}$ hinreichend nahe bei x^* (= „lokale Konvergenz") (z.B. für $\|x^{(0)} - x^*\| < 1$, wenn $c \leq 1$) siehe dazu auch das folgende Lemma 3.8.

Wir wissen (siehe (3.10) im Beweis des Fixpunktsatzes von Banach): Unter den Vorraussetzungen des Fixpunktsatzes ist die Fixpunktiteration global konvergent *von der Ordnung mindestens* 1.

Lemma 3.8 *Sei $(X, \|\cdot\|)$ normierter Vektorraum, $U \subset X$ offen, $\Phi : U \to U$ stetig, $x^{(0)} \in U$ gegeben, $x^{(n+1)} := \Phi(x^{(n)})$, $x^* \in U$, $p \geq 1$, $c > 0$, sowie $c < 1$ falls $p = 1$. Es gelte*

$$\|\Phi(x) - x^*\| \leq c\|x - x^*\|^p \quad \forall x \in U \tag{3.17}$$

Dann konvergiert die Iteration lokal *(d.h. es gibt eine Umgebung $\tilde{U} \subset U$ von x^* und für alle $x^{(0)} \in \tilde{U}$ konvergiert sie) gegen x^* mit Konvergenzordnung mindestens p, und x^* ist Fixpunkt von Φ.*

Beweis : Wähle $r > 0$ so dass die Kugel $B_r(x^*) \subseteq U$ und

$$c \cdot r^{p-1} =: k < 1 \tag{3.18}$$

Es ist $x^{(n)} \in U$ $\forall n \in \mathbb{N}$, und für $x^{(n)} \in B_r(x^*)$ gilt

$$d^{(n+1)} := \|x^{(n+1)} - x^*\| = \|\Phi(x^{(n)}) - x^*\| \underset{(3.17)}{\leq} c\|x^{(n)} - x^*\|^p \underset{x^{(n)} \in B_r(x^*)}{\leq} cr^p \underset{(3.18)}{<} r \tag{3.19}$$

d.h. auch $x^{(n+1)} \in B_r(x^*)$. Falls also der Anfangswert $x^{(0)} \in B_r(x^*) =: \tilde{U}$ gewählt wird, ist die gesamte Folge $(x^{(n)})$ in $\tilde{U}$. Weiter gilt

$$d^{(n+1)} \underset{(3.19)}{\leq} c \cdot \underbrace{\|x^{(n)} - x^*\|^{p-1}}_{\leq r \text{ da } x^{(n)} \in \tilde{U}} \underbrace{\|x^{(n)} - x^*\|}_{= d^{(n)}} \leq cr^{p-1} d^{(n)} \underset{(3.18)}{=} kd^{(n)}, \; k < 1$$

$\Rightarrow x^{(n)} \to x^*$, $x^* = \lim\limits_n x^{(n)} = \lim\limits_n \Phi(x^{(n-1)}) = \Phi(\lim\limits_n x^{(n-1)}) = \Phi(x^*)$
Die Konvergenzordnung ist (mindestens) p nach (3.19). □

Wir wissen: Fixpunktiterationen konvergieren unter den Voraussetzungen von Satz 3.2 *mindestens linear*. Aus Lemma 3.8 folgt der folgende Satz 3.9, der uns sagt, unter welcher Voraussetzungen an Φ wir sogar Konvergenz p-ter Ordnung der Folge $x^{(n+1)} = \Phi(x^{(n)})$ haben; vgl. die Konstruktionsversuche von Φ am Anfang von Kapitel 3.2, wo wir versucht haben, möglichst gute Konvergenz zu erreichen und so die Wahl $\Phi(x) := x - \frac{f(x)}{f'(x)}$ zur Lösung von $f(x) = 0$ motiviert haben.

Satz 3.9 *(Konvergenzordnung von Fixpunktverfahren)*
Sei $X = \mathbb{R}$, $U \subset \mathbb{R}$ offen, $\Phi : U \to U$ p-mal stetig differenzierbar. Sei $x^* \in U$ Fixpunkt von Φ, sei $p \in \mathbb{N}$. Ist

$$\Phi'(x^*) = \ldots = \Phi^{(p-1)}(x^*) = 0, \; \Phi^{(p)}(x^*) \neq 0 \quad \textit{im Fall } p > 1$$

bzw.

$$\Phi'(x^*) \neq 0, \; |\Phi'(x^*)| < 1 \quad \textit{im Fall } p = 1$$

dann konvergiert die durch Φ definierte Fixpunktiteration $x^{(n+1)} = \Phi(x^{(n)})$ lokal *gegen x^** genau mit Ordnung *p, und x^* ist Fixpunkt von Φ.*

Beweis : TAYLOR-Entwicklung von Φ um x^*:

$$\Phi(x) = \underbrace{\Phi(x^*)}_{=x^*} + \frac{\Phi^{(p)}(\zeta)}{p!}(x-x^*)^p \quad \forall x \in U \text{ mit } \zeta = \zeta(x) \in U \quad | - x^* \tag{3.20}$$

$\Rightarrow \quad |\Phi(x) - x^*| \leq c|x-x^*|^p$ mit $c := \max\limits_{\zeta \in \tilde{U}} \frac{|\Phi^{(p)}(\zeta)|}{p!}$, $\tilde{U} \subset U$ kompakte Umgebung von x^*

$\Rightarrow$ Lemma 3.8 liefert lokale Konvergenz der Ordnung mindestens p. Angenommen die Ordnung sei höher als p. Dann müsste (3.17) für ein $\tilde{p} > p$ statt p gelten:

$$|\Phi(x) - x^*| \leq c|x-x^*|^{\tilde{p}} \quad \forall x \in U \tag{3.21}$$

Dies kann nicht sein, da

$$\Phi^{(p)}(x^*) \underset{\text{nach Vor.}}{\neq} 0, \text{ also } \Phi^{(p)}(\zeta) \neq 0$$

in einer Umgebung $\tilde{U}$ von x^* in (3.20), also

$$|\Phi(x) - x^*| \geq \min_{\zeta \in \tilde{U}} |\Phi^{(p)}(\zeta)| \; |x-x^*|^p \; \forall x \in \tilde{U},$$

was ein Widerspruch zu (3.21) ist. □

Satz 3.9 im Fall $p = 2$ besagt: Wir bekommen *quadratische* Konvergenz (als Verbesserung der linearen Konvergenz, die allgemein in der Situation des Fixpunktsatzes von Banach gilt) genau dann wenn

$$\Phi'(x^*) = 0 \quad , \quad \Phi''(x^*) \neq 0.$$

Dies wird in der Tat vom NEWTON-Verfahren

$$\Phi(x) = x - \frac{f(x)}{f'(x)}, \; x^{(n+1)} = \Phi(x^{(n)}), \tag{3.22}$$

wenn $f'(x^*) = 0$, erfüllt:

Satz 3.10 *Sei $f \in \mathcal{C}^3(\mathbb{R})$ und x^* eine* einfache *Nullstelle von f (also $f'(x^*) \neq 0$). Dann ist das* NEWTON*-Verfahren lokal konvergent, mindestens von Ordnung 2.*

Beweis : Wegen der Stetigkeit existiert eine offene Umgebung V von x^*, auf der $f'(x) \neq 0$ ist, d.h Φ aus (3.22) ist wohldefiniert und stetig auf V, es ist

$$\Phi(x^*) = x^* - \frac{\overbrace{f(x^*)}^{=0}}{f'(x^*)} = x^*$$

(*Konsistenz* des Nullstellenproblems $f(x) = 0$ mit dem Fixpunktproblem $\Phi(x) = x$), es ist

$$\Phi'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2}, \text{ also } \Phi'(x^*) = 0,$$

und $\Phi''(x)$ existiert und ist stetig.
Nach Satz 3.9 folgt die Behauptung. □

Zur Vorgehensweise bei *mehr*facher Nullstelle (siehe Übung): Setze $\Phi(x) := x - \gamma\frac{f(x)}{f'(x)}$ wobei $\gamma = p$ die Vielfachheit der Nullstelle ist. Hier für $p = 2$ (= doppelte Nullstellen), also $f(x^*) = f'(x^*) = 0, f''(x^*) \neq 0$ hergeleitet:

$$\begin{aligned}\Phi(x) &:= x - \gamma\frac{f(x)}{f'(x)}, \quad x \neq x^* \\ \lim_{x\to x^*}\Phi(x) &= x - \gamma\lim_{x\to x^*}\frac{f'(x)}{f''(x)} = x^*\end{aligned}$$

$\Rightarrow$ Mit der Setzung $\Phi(x^*) := x^*$ wird die Definitionslücke von Φ stetig geschlossen.

$$\Phi'(x) = 1 - \gamma\frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = 1 - \gamma + \gamma\frac{f(x)f''(x)}{f'(x)^2}, \quad x \neq x^*$$

Es folgt mit L'Hopital:

$$\begin{aligned}\Phi'(x^*) &= \lim_{x\to x^*}\frac{f(x) - f(x^*)}{x - x^*} = \lim_{x\to x^*}\frac{x - \gamma\frac{f(x)}{f'(x)} - x^*}{x - x^*} = 1 - \gamma\lim_{x\to x^*}\frac{f(x)}{f'(x)(x - x^*)} \\ &= 1 - \frac{\gamma}{\lim\limits_{x\to x^*}\frac{f'(x)(x-x^*)}{f(x)}} = 1 - \frac{\gamma}{\lim\limits_{x\to x^*}\frac{f''(x)(x-x^*)+f'(x)}{f'(x)}} \\ &= 1 - \frac{\gamma}{1 + f''(x^*)\lim\limits_{x\to x^*}\frac{x-x^*}{f'(x)}} = 1 - \frac{\gamma}{1 + f''(x^*)\lim\limits_{x\to x^*}\frac{1}{f''(x)}} \\ &= 1 - \frac{\gamma}{2}\end{aligned}$$

sowie:

$$\begin{aligned}\lim_{x\to x^*}\Phi'(x) &= 1 - \gamma + \gamma f''(x^*)\lim_{x\to x^*}\frac{f(x)}{f'^2(x)} = 1 - \gamma + \gamma f''(x^*)\cdot\lim\frac{f'(x)}{2f'(x)f''(x)} \\ &= 1 - \gamma + \gamma f''(x^*)\cdot\frac{1}{2f''(x^*)} = 1 - \frac{\gamma}{2}.\end{aligned}$$

Also: Φ ist stetig differenzierbar an der Stelle $x = x^*$, und die Forderung

$$\Phi'(x^*) = 1 - \frac{\gamma}{2} \stackrel{!}{=} 0$$

erfordert $\gamma := 2$.

Fazit soweit: Das Newton-Verfahren zur Nullstellenbestimmung konvergiert nur dann sicher gegen eine Nullstelle x^*, wenn der Startwert hinreichend nahe bei x^* liegt. Der Konvergenzbereich kann sehr klein sein, Insbesondere wenn Extrem- und Wendepunkte in der Nähe von x^* (zwischen x^* und $x^{(0)}$) liegen: Um sich eine bessere Startlösung $x^{(0)}$ fürs Newton-Verfahren zu verschaffen, kann man zunächst einige Schritte eines anderen Verfahrens (z.B. Bisektion, s. Kapitel 3.4) vorab durchführen, oder gegebenenfalls ein "gedämpftes Newton-Verfahren" durchführen (s. Kapitel 3.3).

Wenn das Newton-Verfahren konvergiert, dann extrem schnell. (**lokal quadratische** Konvergenz)

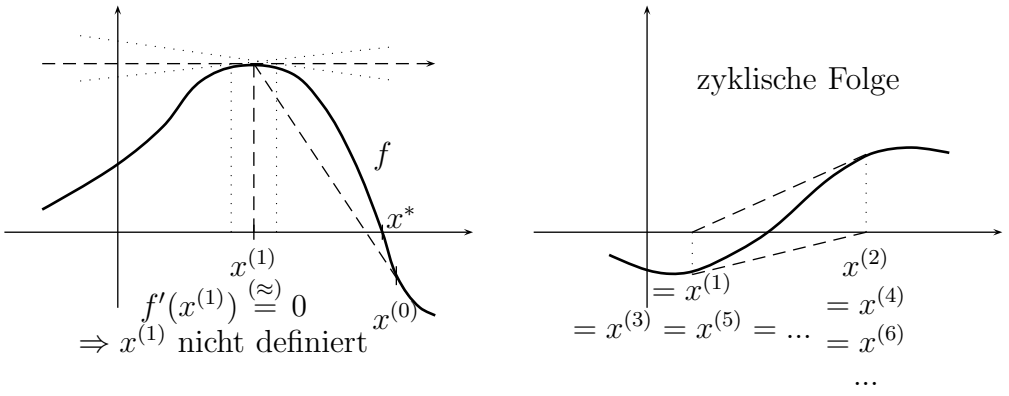

Abbildung 9: NEWTON-Verfahren bei Extrem- und Wendepunkten

3.3 Das Bisektions- und das Sekantenverfahren

Als Alternative zum NEWTON-Verfahren für skalare Nullstellenprobleme gibt es das *Bisektionsverfahren* (= *Intervallhalbierungsverfahren*) sowie das *Sekantenfahren*.

Bisektionsverfahren : Sei $f : \mathbb{R} \to \mathbb{R}$ stetig, sei $x^{(0)} < y^{(0)}$ bekannt mit

$$f(x^{(0)}) \cdot f(y^{(0)}) \leq 0$$

$\Rightarrow f$ besitzt eine Nullstelle im Intervall $I^{(0)} := [x^{(0)}, y^{(0)}]$. Man zerteilt das Intervall in

$$[x^{(0)}, m^{(0)}] \text{ und } [m^{(0)}, y^{(0)}],\ m^{(0)} := \frac{x^{(0)} + y^{(0)}}{2}\ .$$

In mindestens einem der beiden Intervalle muss eine Nullstelle liegen, man testet also: Falls

$$f(x^{(0)}) \cdot f(m^{(0)}) < 0,$$

dann

$$x^{(1)} := x^{(0)},\ y^{(1)} := m^{(0)},$$

andernfalls

$$x^{(1)} := m^{(0)}\ y^{(1)} := y^{(0)},$$

und setzt

$$I^{(1)} := [x^{(1)}, y^{(1)}]$$

dies wird iteriert. $\Rightarrow$ Jedes der Intervalle $I^{(n)}$ enthält eine Nullstelle, es ist

$$|I^{(n)}| = |y^{(n)} - x^{(n)}| = \left(\frac{1}{2}\right)^n \cdot |I^{(0)}|$$

d.h. wir wissen

$$\begin{aligned} |x^* - x^{(n)}| &\leq \left(\frac{1}{2}\right)^n |I^0|, \\ |x^* - y^{(n)}| &\leq \left(\frac{1}{2}\right)^n |I^0|. \end{aligned}$$

Deswegen ist das Verfahren (nur) *linear* konvergent. Es wird *Bisektions-* oder *Intervallhalbierungsverfahren* genannt.

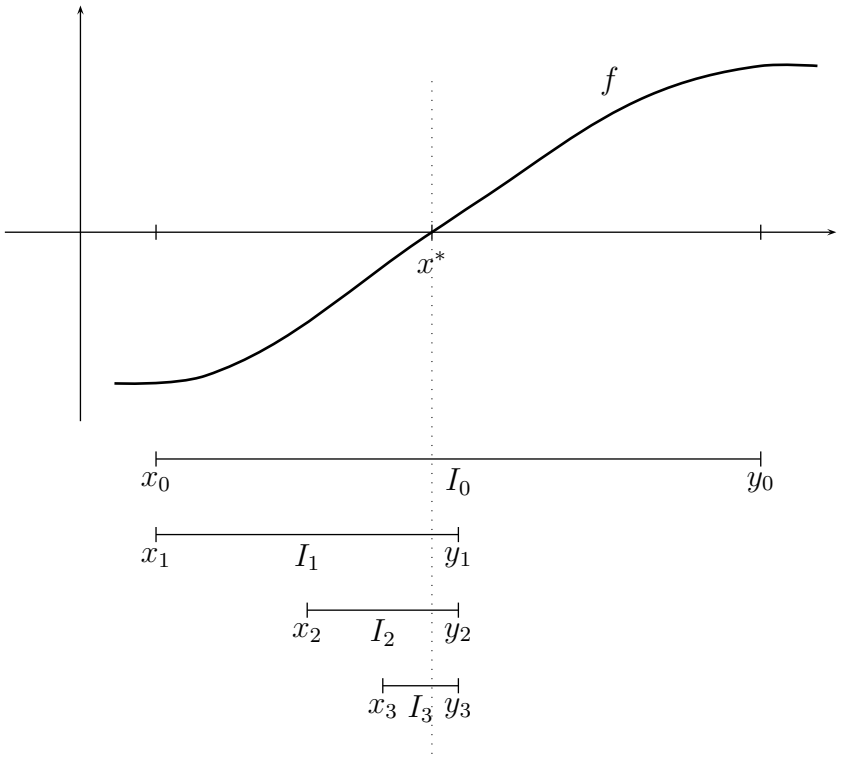

Abbildung 10: Bisektionsverfahren

Vorteile:

- Konvergiert *immer* (*globale* Konvergenz), sobald $x^{(0)}, y^{(0)}$ mit $f(x^{(0)}) \cdot f(y^{(0)}) < 0$ bekannt.
- Braucht kein f'; Stetigkeit von f reicht.
- Liefert *Einschließung* der Lösung.

Nachteile:

- Nur linear konvergent.
- Nicht auf $X = \mathbb{R}^n$ zu verallgemeinern (vgl. Kapitel 3.4: NEWTON-Verfahren im $\mathbb{R}^n$)

Das Sekantenverfahren : Man ersetzt im NEWTON-Verfahren die Ableitung $f'(x^{(n)})$ durch einen Differenzenquotienten $\frac{f(x^{(n)})-f(x^{(n-1)})}{x^{(n)}-x^{(n-1)}}$, also

$$\begin{aligned} x^{(n+1)} &= x^{(n)} - \frac{f(x^{(n)})}{\frac{f(x^{(n)})-f(x^{(n-1)})}{x^{(n)}-x^{(n-1)}}} = \frac{x^{(n)}(f(x^n)) - f(x^{(n-1)})}{f(x^{(n)}) - f(x^{(n-1)})} - \frac{f(x^{(n)})(x^{(n)} - x^{(n-1)})}{f(x^{(n)} - f(x^{(n-1)}))} = \\ &= \frac{x^{(n-1)} f(x^{(n)}) - x^{(n)} f(x^{(n-1)})}{f(x^{(n)}) - f(x^{(n-1)})} \end{aligned} \tag{3.23}$$

Es hängt $x^{(n+1)}$ also nicht nur von $x^{(n)}$, sondern auch von $x^{(n-1)}$ ab es handelt sich um ein *Zweischrittverfahren*. Ein solches hat allgemein die Form:

$$x^{(n+1)} = \Phi(x^{(n-1)}, x^{(n)}), \ \Phi : \mathbb{R}^2 \to \mathbb{R} \text{ (bzw. } \Phi : X^2 \to X \text{ im allgemeinen Fall)}$$

Man benötigt also 2 Startwerte $x^{(0)}, x^{(1)}$. Beim Sekantenverfahren ist offensichtlich $f(x^{(0)}) \neq f(x^{(1)})$ erforderlich. Rein formal kann man jedes Zweischrittverfahren in $\mathbb{R}$ (bzw X) als ein *Ein*schrittverfahren in $\mathbb{R}^2$ (bzw. X^2) auffassen, indem man jeweils Paare von aufeinanderfolgenden Iterationen zusammenfasst und den Übergang

$$\tilde{x}^{(n)} = (x^{(n-1)}, x^{(n)}) \to (x^{(n)}, x^{(n+1)}) = \tilde{x}^{(n+1)}$$

durch ein

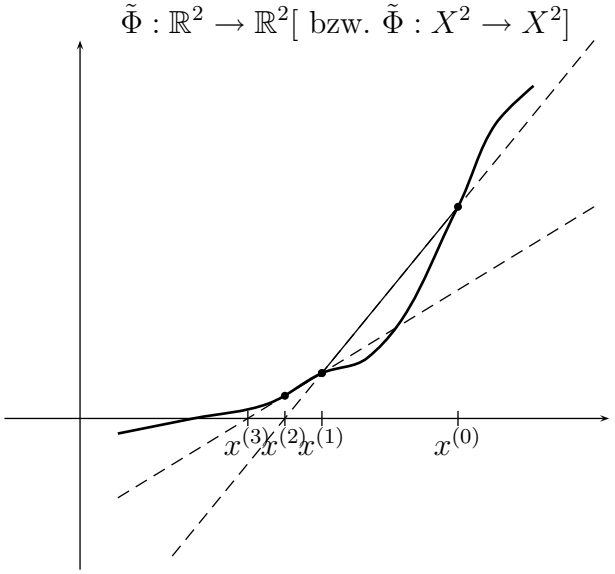

Abbildung 11: geometrische Interpretation des Sekantenverfahrens

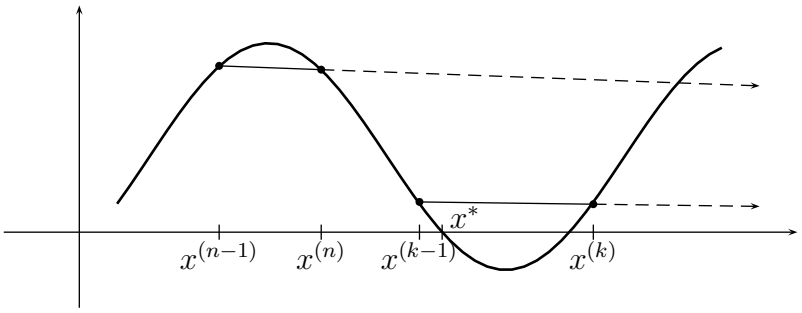

Abbildung 12: Probleme bei lokalen Extrema

Lemma 3.11 *Sei $f \in \mathcal{C}^2(U)$, $U \subset \mathbb{R}$ offen, $x^* \in U$ einfache Nullstelle, $x^{(0)} \neq x^{(1)}$. Dann ist das Sekantenverfahren lokal konvergent gegen x^* mit Konvergenzordnung*

$$p = \frac{1}{2}(1 + \sqrt{5}) \approx 1.618 \qquad (\text{„goldener Schnitt“, Lösung von } p^2 = p + 1) \, .$$

Beweisskizze : (nur für Konvergenz*ordnung* und Wohldefiniertheit)

Da $f'(x^*) \neq 0$, ist f in einer Umgebung $\tilde{U} \subseteq U$ von x^* streng monoton, also injektiv. Wir gehen davon aus, dass wir ein $n_0 \in \mathbb{N}$ haben mit $x^{(n)} \in \tilde{U}$ $\forall n \geq n_0$ (ohne Beweis). Aus

$$x^{(n)} \neq x^{(n-1)}$$

folgt

$$x^{(n+1)} = x^{(n)} - \frac{\overbrace{(x^{(n)} - x^{(n-1)})}^{\neq 0} f(x^{(n)})}{f(x^{(n)}) - f(x^{(n-1)})} \neq x^{(n)}$$

(es sei denn, $f(x^{(n)}) = 0$, d.h. $x^{(n)} = x^*$, in welchem Falle man abbrechen kann). Per Induktion folgt:

$$x^{(n+1)} \neq x^{(n)} \ \forall n \geq n_0$$

(bis ggf. zum Abbruch $x^{(n)} = x^*$).
Sei $e^{(n)} := x^{(n)} - x^*$ der Fehler. Subtraktion von x^* von (3.23) ergibt:

$$\begin{aligned} e^{(n+1)} &= x^{(n+1)} - x^* = \frac{x^{(n-1)} f(x^{(n)}) - x^{(n)} f(x^{(n-1)})}{f(x^{(n)}) - f(x^{(n-1)})} - x^* \\ &= \frac{e^{(n-1)} f(x^{(n)}) - e^{(n)} f(x^{(n-1)})}{f(x^{(n)}) - f(x^{(n-1)})}. \end{aligned} \tag{3.24}$$

Unter der Annahme $|e^{(n)}|,\ |e^{(n-1)}| \ll 1$ bekommen wir näherungsweise mittels Taylor-Entwicklung:

$$\begin{aligned} \text{Nenner} &\approx f(x^*) + f'(x^*)(x^{(n)} - x^*) - [f(x^*) + f'(x^*)(x^{(n-1)} - x^*)] \\ &= f'(x^*)(e^{(n)} - e^{(n-1)}) \\ \text{Zähler} &\approx e^{(n-1)} \cdot \left[\underbrace{f(x^*)}_{=0} + f'(x^*) e^{(n)} + f''(x^*) \frac{e^{(n)^2}}{2} \right] \\ &\quad - e^{(n)} \left[\underbrace{f(x^*)}_{=0} + f'(x^*) e^{(n-1)} + f''(x^*) \frac{e^{(n-1)^2}}{2} \right] \\ &= \frac{1}{2} e^{(n-1)} e^{(n)} (e^{(n)} - e^{(n-1)}) f''(x^*) \qquad [\text{Annahme: } f''(x^*) \neq 0] \\ \Rightarrow \quad e^{(n+1)} &\approx \frac{f''(x^*)}{2f'(x^*)} e^{(n-1)} e^{(n)} \end{aligned}$$

Insbesondere gilt

$$e^{(n+1)} \leq \underbrace{\left(\left| \frac{f''(x^*)}{2f'(x^*)} \right| + \epsilon \right)}_{=:c} e^{(n-1)} e^{(n)} \ .$$

Sei $E_{n_0} \geq |e^{(n_0)}|,\ E_{n_0+1} \geq |e^{(n_0+1)}|,\ E_{n+1} \underset{(*)}{:=} c E_n E_{n-1}$ $(\Rightarrow |e^{(n)}| \leq E_n \ \forall n \geq n_0$, d.h. es reicht

E_n abzuschätzen)[2]. Der Ansatz $E_n \stackrel{!}{=} k \cdot E_{n-1}^p$ wird eingesetzt in die Rekursionsgleichung $(*)$:

$$\left.\begin{array}{r} \text{links } E_{n+1} = k \cdot E_n^p = k \cdot (kE_{n-1}^p)^p \\ \text{rechts } cE_nE_{n-1} = ckE_{n-1}^pE_{n-1} \end{array}\right\} \stackrel{!}{=} \quad \Rightarrow k^p E_{n-1}^{p^2} = cE_{n-1}^{p+1} \; \forall n \geq n_0.$$

Dies wird gelöst durch $k = \sqrt[p]{c}$, $p^2 = p + 1 \Rightarrow \; p = \frac{1}{2}(1 \pm \sqrt{5})$. Die negative Wurzel kann man ausschließen, indem man die "Anfangswerte" E_{n_0+1} und E_{n_0} so wählt, dass

$$E_{n_0+1} = \sqrt[\frac{1}{2}(1+\sqrt{5})]{c} E_{n_0}^{\frac{1}{2}(1+\sqrt{5})}.$$

□

Effizienzanalyse :

Durch Zwischenspeicherung der Werte $f(x^{(n)})$ kann man das Sekantenverfahren so formulieren, dass in jedem Iterationsschritt (außer dem ersten) nur *eine* Funktionsauswertung nötig ist. Beim NEWTON-Verfahren dagegen sind Auswertung von f und f' nötig.
Unter der Annahme, dass daher die Durchführung eines NEWTON-Schritts in etwa so lange dauert wie die Durchführung von zwei Schritten des Sekantenverfahrens ($x^{(n)} \to x^{(n+2)}$), ergibt sich: Fehlerordung eines Doppel-Schritts des Sekantenverfahrens: $p^2 = p + 1 \approx 2.61 > 2$!
In diesem Sinne kann das Sekantenverfahren durchaus effizienter sein als das NEWTON-Verfahren!

Bewertung :

+ sehr schnell;
+ braucht nur f, nicht f';
+ realistische a priori-Abschätzung verfügbar (Stummel & Hainer: Praktische Mathematik);
- keine globale Konvergenz, Probleme bei Extrema;
- Probleme bei mehrfachen Nullstellen möglich (siehe Übung);
- schwierig auf mehrdimensionale Probleme ($X = \mathbb{R}^n$) zu verallgemeinern.

Weitere Verfahren:

- **Regula Falsi:** s. Stummel & Hainer: Praktische Mathematik, Knabner-Skript;
  Kann als "Mischung" aus Sekanten und Bisektionsverfahren betrachtet werden. Startet mit $x^{(0)}, x^{(1)}$ so dass $f(x^{(0)}) \cdot f(x^{(1)}) < 0$, $x^{(2)}$ sei Nullstelle der Sekante durch $x^{(0)}, x^{(1)}$; liefert *Einschließung* der Lösung;
  aber Vorsicht: im Allgemeinen $|I^{(n)}| = |x^{(n+1)} - x^{(n)}| \not\to 0$, sondern nur $dist(x^*, \partial I^{(n)}) \to 0$ für $n \to \infty$, im Allgemeinen von erster Ordnung; nur historisch interessant;

[2]Rekursionsgleichungen der Form $x_n = b + \alpha_1 x_{n-1} + \ldots + \alpha_k x_{n-k}$ ($b, \alpha_1, \ldots, \alpha_{n-k}$ gegeben) werden auch Differenzengleichungen genannt; für diese gibt es eine Lösungstheorie, die die Struktur der Lösungsmenge charakterisiert. Gleichung $(*)$ hat, wenn logarithmiert, diese Struktur bzgl. der Unbekannten $\ln E_n$. Wir verwenden einen *Ansatz*, um Lösungen auch ohne theoretische Betrachtung zu bestimmen.

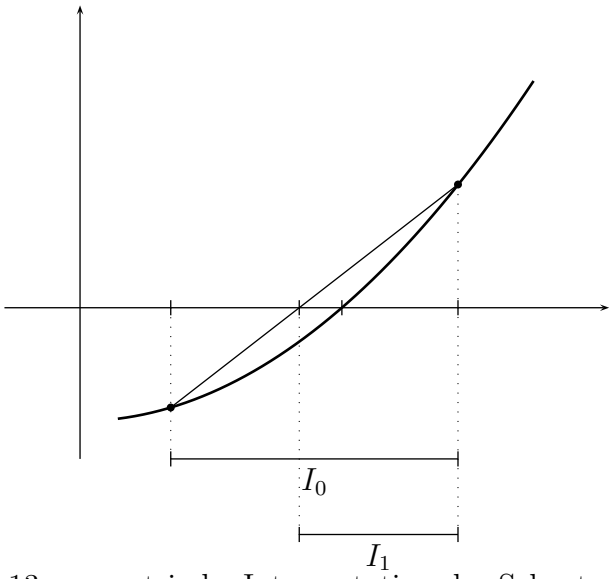

Abbildung 13: geometrische Interpretation des Sekantenverfahrens

- **Hybridverfahren**, wie z.B.: Starte mit einschließendem Intervall $x^* \in I^{(0)} = [a^{(0)}, b^{(0)}]$, probiere NEWTON-Schritt; falls das Ergebnis $x^{(1)} \in I^{(0)}$, so akzeptiere; bilde je nach Vorzeichen von $f(x^{(1)})$ das Intervall $I^{(1)} = [a^{(0)}, x^{(1)}]$ bzw. $I^{(1)} = [x^{(1)}, b^{(0)}]$; falls dagegen $x^{(1)} \notin I^{(0)}$, so verwerfe das Ergebnis des NEWTON-Schritts und führe Intervallhalbierungsschritt stattdessen durch.

3.4 Das Newton-Verfahren im $\mathbb{R}^m$

Für $f : \mathbb{R} \to \mathbb{R}$ hatten wir das Newton-Verfahren durch *Linearisierung* gewonnen: Die TAYLOR-Entwicklung von f um $x^{(n)}$ (wenn wir uns an die geometrische Konstruktion auf S. 46 erinnern):

$$f(x) = \underbrace{f(x^{(n)}) + f'(x^{(n)}) \cdot (x - x^{(n)})}_{\text{lineare Approximation an } f \text{ (=Tangente!)}} + O(|x - x^{(n)}|^2)$$

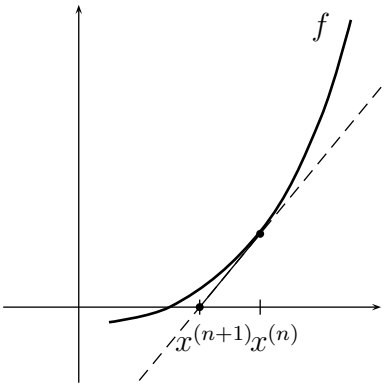

Abbildung 14: Linearisierung von f

$x^{(n+1)}$ wurde durch *Nullsetzen* der *Linearisierung*

$$f(x^{(n)}) + f'(x^{(n)}) \cdot (x^{(n+1)} - x^{(n)}) \stackrel{!}{=} 0$$

$$\Leftrightarrow x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})}{f'(x^{(n)})}$$

gewonnen. Naheliegend im Fall $f : \mathbb{R}^m \to \mathbb{R}^m$ ist: Wir verwenden die TAYLOR-Entwicklung

$$f(x) = \underbrace{f(x^{(n)}) + (Jf)(x^{(n)})(x - x^{(n)})}_{\textit{lineare} \text{ Approximation an } f} + O(\|x - x^{(n)}\|^2),$$

und $x^{(n+1)}$ sei wieder die Nullstelle der Linearisierung:

$$f(x^{(n)}) + Jf(x^{(n)})(x^{(n+1)} - x^{(n)}) \stackrel{!}{=} 0 \quad \Leftrightarrow \quad \boxed{x^{(n+1)} = x^{(n)} - Jf(x^{(n)})^{-1} f(x^{(n)})} \tag{3.25}$$

NEWTON-Verfahren im$\mathbb{R}^m$

Erforderlich: Die Matrix $Jf(x^{(n)})$ muss invertierbar sein. Nach (3.25) scheint es, das $(Jf(x^{(n)}))^{-1}$ explizit berechnet werden muss. Folgende Umformulierung zeigt, dass es stattdessen reicht, ein lineares Gleichungssystem zu lösen:

Setze $\Delta x^{(n)} := x^{(n+1)} - x^{(n)}$. Dann ist (3.25) $\Leftrightarrow$ $Jf(x^{(n)})\Delta x^{(n)} = -f(x^{(n)})$. Der NEWTON-Schritt (3.25) wurde so äquivalent umgeschrieben zu:

$$\begin{array}{ll} (1) & \text{Löse das Lineare Gleichungssystem } Jf(x^{(n)})\Delta x^{(n)} = -f(x^{(n)}) \\ (2) & \text{Setze } x^{(n+1)} := x^{(n)} + \Delta x^{(n)} \end{array} \tag{3.26}$$

$\Delta x^{(n)}$ bzw. Teilschritt (2) heißt auch *Newton-Update*.

Das nichtlineare Problem $f(x) = 0$ wurde ersetzt durch eine *Folge* von *linearen* Problemen.

Aufwandsvergleich: Bei Verwendung von LR erfordert (3.26)

$$\underbrace{\frac{1}{3}m^3}_{LR\text{-Zerlegung}} + \underbrace{\frac{m^2}{2} + \frac{m^2}{2}}_{\text{Vor- und Rückwärtseinsetzen}} \doteq \frac{1}{3}m^3 \text{ Additionen und Multiplikationen}$$

Bei (3.25) mit LR brauchen wir ebenfalls eine LR-Zerlegung für $A := Jf(x^{(n)})$, allerdings brauchen wir zum Aufstellen von A^{-1} m-maliges Vor-/Rückwärtseinsetzen:

$$LRu_j = e_j; \{e_1, \ldots, e_n\} = \text{ Standard-Basis}$$

Die u_j bilden die Spalten von A^{-1}, insgesamt $\frac{1}{3}m^3 + m \cdot m^2 = \frac{4}{3}m^3$ Operationen!

(3.25) ist also viermal so teuer wie (3.26)! Man kann auch im vektoriellen Fall $X = \mathbb{R}^m$ zeigen, dass das NEWTON-Verfahren (3.25) bzw. (3.26) *lokal quadratisch* konvergent ist; man kann sogar konkrete Angaben über die *Mindest*größe des Konvergenzbereichs machen (die aber in der Praxis mühsam zu überprüfen sind):

Satz 3.12 *(Konvergenz Newton-Verfahren für $X = \mathbb{R}^m$)*
Sei $U \subset \mathbb{R}^m$ offen und konvex, sei $f : U \to \mathbb{R}^m$ differenzierbar, sei $x^{(0)} \in U$. Es gebe Zahlen $\alpha, \beta, \gamma, h, r > 0$ mit

$$h := \frac{\alpha\beta\gamma}{2} < 1 \quad , \quad r := \frac{\alpha}{1-h} \quad , \quad \overline{B_r}(x^{(0)}) \subseteq U.$$

Ferner sei vorausgesetzt:

(a) $\|Jf(x) - Jf(y)\| \leq \gamma\|x - y\|$ $\forall x, y \in \tilde{U} := B_{r+\epsilon}(x^{(0)}) \subset U$ *für ein* $\epsilon > 0$, *("Jf ist Lipschitz-stetig".)*[3]

(b) Für alle $x \in \overline{B_r}(x^{(0)})$ *existiert* $(Jf(x))^{-1}$ *und* $\|Jf(x)^{-1}\| \leq \beta$

(c) $\|Jf(x^{(0)})^{-1} f(x^{(0)})\| \leq \alpha$

Dann gilt:

(1) Die Newton-Iteration (3.25)/(3.26) ist wohldefiniert und $x^{(n)} \in B_r(x^{(0)})$ $\forall n \in \mathbb{N}$

[3]Dies ist eine Abschwächung der Voraussetzung "$f \in \mathcal{C}^3$" aus Satz 3.10.

(2) $(x^{(n)})$ *konvergiert gegen eine Nullstelle* x^* *von* f*.*

(3) $\|x^{(n+1)} - x^*\| \leq \frac{\beta\gamma}{2}\|x^{(n)} - x^*\|^2$ *(d.h. quadratische Konvergenz)*
sowie $\|x^{(n)} - x^*\| \leq \alpha\frac{h^{2^n-1}}{1-h^{2n}}$ $\forall n \in \mathbb{N}$ *(a priori-Abschätzung)*

Beweis: s. Knabner-Skript S. 107-109

Bemerkung:

- Die Aussagen gelten natürlich erst recht im skalaren Fall $X = \mathbb{R}$.
- Durch die Wahl von $x^{(0)}$ hinreichend nahe bei einer Nullstelle x^* (also $f(x^{(0)})$ hinreichend nahe bei 0) kann $\alpha > 0$ beliebig klein gemacht werden (vgl. (c)), so dass die Bedingungen des Satzes erfüllt werden ($\rightarrow$ *lokale* Konvergenz).
- Stellt man die Frage nach der *Größe* des Konvergenzbereichs *nicht*, so lässt sich Satz 3.12 vereinfachen zu:

  f differenzierbar auf einer Umgebung U von x^* mit $f(x^*) = 0$ und $Jf(x^*)$ invertierbar, Jf L-stetig auf U. Dann ist das Newton-Verfahren *lokal quadratisch* konvergent.

  Beweis: (sowie weitere Infos:)
  P. Deuflhard, Newton methods for nonlinear problems, Springer, 2004, 430 Seiten.

In der Praxis wird man, um das Verfahren auf Konvergenz/Divergenz zu prüfen, weniger die Voraussetzung von Satz 3.12 überprüfen, sondern eher einen sogenannten *Monotonietest* durchführen:

Das Nullstellenproblem $f(x) = 0$ ist offensichtlich äquivalent zum Lösen von:

$$\text{minimiere } g(x) := \|f(x)\|^2 \quad \text{für } x \in \mathbb{R}^n \tag{3.27}$$

Man könnte also einen "Abstieg" $g(x^{(n+1)}) \overset{!}{\leq} g(x^{(n)})$ in diesem Funktional erwarten und daher

$$\|f(x^{(n+1)})\| \leq \Theta\|f(x^{(n)})\| \quad \text{für ein } \Theta \in (0,1) \tag{3.28}$$

testen. Falls (3.28) nicht erfüllt ist, wird für die neue Newton-Iterierte $x^{(n+1)} = x^{(n)} + \Delta x^{(n)}$, so kann man den Newton-Schritt "dämpfen" d.h. man setzt

$$x^{(n+1)} := x^{(n)} + t\Delta x^{(n)} \quad \text{mit } t \in (0,1], \tag{3.29}$$

wobei ein t so zu finden ist, dass (3.28) erfüllt ist oder besser

$$\|f(x^{(n)} + t\Delta x^{(n)})\| \leq (1 - \Theta t)\|f(x^{(n)})\|$$

erfüllt ist („Armijo-Regel").
In der Praxis probiert man nacheinander $t = 1$, $t = \gamma$, $t = \gamma^2$, ... für ein festes $\gamma \in (0,1)$, z.B.: $\gamma = 0,5$, aus, bis (3.28) erfüllt wird:

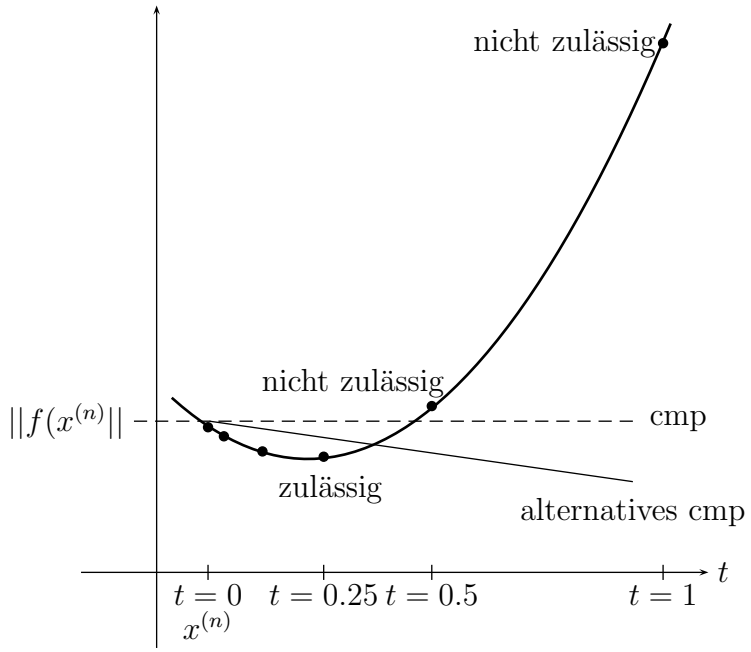

Abbildung 15: "Newton-Armijo-Regel"

gegeben: Startwert x, Parameter γ, $\tilde{\Theta} \in (0, 1]$, $l_{\max} \in \mathbb{N}$;

```
1:  wiederhole {
2:      t := 1
3:      loese Jf(x) Δ x = -f(x)
4:      l := 0
5:      solange || f(x + t Δ x) || > (1 - t Θ̃) || f(x) ||
        und l ≤ l_max
6:          t := t · γ; l := l + 1
7:      falls l > l_max: Abbruch
8:      x := x + t Δ x
9:  } bis Abbruchkriterium erfuellt
```

Dieser Algorithmus (gedämpftes Newton-Verfahren nach Armijo) hat im allgemeinen einen größeren Konvergenzbereich als das reine Newton-Verfahren.

Literatur: Kelley: Iterative methods for linear and nonlinear problems

Die Suche nach der geeigneten "Schrittweite" t nennt man auch *Line Search*, da die neue Iteration entlang der durch (3.29) beschriebenen *Geraden* gesucht wird.

Zur Übung: Der Line Search sollte, zumindest für $\Theta \approx 1$, bzw. $\tilde{\Theta} = 0$ sofern $Jf(x)$ immer invertierbar und $x \mapsto Jf(x)$ stetig differenzierbar ist, irgendwann erfolgreich ein t finden, denn die Richtung

$$r := \Delta x^{(n)} = -(Jf(x^{(n)})^{-1} f(x^{(n)})$$

stellt eine *Abstiegs*richtung für das Funktional $g(x) = \|f(x)\|^2$ dar, sofern $\|\cdot\| = \|\cdot\|_2$:

$$\begin{aligned}\frac{\partial g(x^{(n)})}{\partial r} = \langle \nabla g(x^{(n)}), \Delta x^{(n)} \rangle &= \langle 2Jf(x^{(n)})^t f(x), -Jf(x^{(n)})^{-1} f(x^{(n)}) \rangle \\ &= -2\langle f(x^{(n)}), \underbrace{Jf(x^{(n)}) Jf(x^{(n)})^{-1}}_{=Id} f(x^{(n)}) \rangle \\ &= -2\|f(x^{(n)})\|^2 \leq 0\end{aligned}$$

(sogar < 0 solange $f(x^{(n)}) \neq 0$).

Um bei Nichtexistenz von $(Jf(x^{(n)}))^{-1}$ einen Abbruch des Verfahrens zu vermeiden, kann man in diesem Fall die Suchrichtung $\Delta x^{(n)} = (Jf(x^{(n)}))^{-1} f(x^{(n)})$ durch die Richtung des stärksten Abstiegs des Funktionals g, also $-\nabla g(x^{(n)}) = -2Jf(x^{(n)})^t f(x^{(n)})$, multipliziert mit einer geeigneten Schrittweite t, ersetzen (s. Hanke-Bourgeois).

Vereinfachtes Newton-Verfahren :

Um Rechenzeiten zu sparen, kann man alle (oder zumindest jeweils 2 oder 3) Newton-Schritte mit ein und derselben Jacobi-Matrix $Jf(x^{(0)})$ durchführen;

$$x^{(n+1)} = x^{(n)} - (Jf(x^{(0)})^{-1} f(x^{(n)}). \tag{3.30}$$

Diese LR-Zerlegung der Jacobi-Matrix ist dann nur einmal durchzuführen. Diesen Rechenzeitvorteil erkauft man sich allerdings mit dem Verlust der quadratischen Konvergenz; die Konvergenz ist nur noch linear (wenn auch mit kleiner Kontraktionskonstanten k, falls $x^{(0)}$ nahe bei x^* ist). Beachte:
(3.30) ist von der Form (3.14), d.h.

$$\Phi(x) = x + \alpha f(x), \text{ mit } \alpha = -Df(x^{(0)})^{-1} \text{ bzw. } \alpha = -\frac{1}{f'(x^{(0)})} \text{ im } 1-D\text{-Fall.}$$

Eine weitere Vereinfachungsmöglichkeit ist es, beim Assemblieren (=Aufstellen) der linearen Gleichungssysteme die Ableitung $\frac{\partial f_i}{\partial x_j}$ durch diskrete Differenzenquotienten (vgl. Kapitel 1.5) zu ersetzen, um das Assemblieren zu beschleunigen, da die Ableitungen oft kompliziertere Funktionen sind als die f_i selbst.

4 Iterative Verfahren für lineare Gleichungssysteme. Teil I: Fixpunktverfahren

4.1 Einführung

Wir betrachten das lineare Gleichungssystem $Ax = b,\ A \in \mathbb{C}^{n\times n}$, invertierbar, $b \in \mathbb{C}^n$. Wir wollen dieses Problem auf *Fixpunktgestalt* bringen und per Fixpunktverfahren iterativ lösen. Die Konvergenz wird mit dem Fixpunktsatz von Banach untersucht.

Mögliche elementare Ansätze:

1. (In Anlehnung an den nichtlinearen Fall Kapitel 3.2)

$$\begin{aligned} Ax - b &= 0 && | \cdot(-\omega),\ \omega \neq 0 \\ (-\omega)(Ax-b) &= 0 && | + x \\ \underbrace{(Id - \omega A)x + \omega b}_{=:\Phi(x)} &= x && \text{Fixpunktgleichung} \end{aligned}$$

   Fixpunktiteration: $x^{(k+1)} := (Id - \omega A)x^{(k)} + \omega b$.

2. Statt mit $\omega \in \mathbb{R}$ kann man in 1. auch mit einer nichtsingulären Matrix $-B \in \mathbb{C}^{n\times n}$ multiplizieren. Man erhält die Fixpunktgleichung

$$\underbrace{(Id - BA)x + Bb}_{=:\Phi(x)} = x$$

   und die Fixpunktiteration $x^{(k+1)} := (Id - BA)x^{(k)} + Bb$.

3. Man zerlegt A in einen "wesentlichen Anteil" W und einen "Rest" S:

$$A = W + S. \tag{4.1}$$

   W soll dabei leicht invertierbar sein (z.B. bei diagonaldominantem A: $W := \mathrm{diag}(a_{ii})$ und $S := A - W$):

$$Ax = b \Leftrightarrow Wx = -Sx + b \Leftrightarrow x = \underbrace{-W^{-1}Sx + W^{-1}b}_{=:\Phi(x)} \tag{4.2}$$

In allen Fällen hat Φ die Form

$$\Phi(x) = Mx + \tilde{b} \tag{4.3}$$

wobei $M, \tilde{b}$ derart, dass

$$\Phi(x) = x \Leftrightarrow Ax = b$$

("*Konsistenz*" des linearen Gleichungssystems und des Fixpunktproblems) (mit $M = Id - \omega A$ bzw. $M = Id - BA$ bzw. $M = -W^{-1}S$). Wann liefert der Fixpunktsatz von Banach Konvergenz der Fixpunktiteration

$$x^{(k+1)} = \Phi(x^{(k)})\ ? \tag{4.4}$$

Wir wollen eine Matrixnorm $\|\cdot\|$ als *Operatornorm* bezeichnen, falls es eine zugehörige Vektornorm $\|\cdot\|$ gibt mit

$$\|A\| = \sup_{0\neq x\in\mathbb{C}^n} \frac{\|Ax\|}{\|x\|} \ \forall A \in \mathbb{C}^{n\times n}.$$

Lemma 4.1 *Sei $x^{(0)} \in \mathbb{C}^n$, sei $\|\cdot\|$ eine beliebige Norm des $\mathbb{C}^n$ sowie eine verträgliche Matrixnorm (z.B. die zugeordnete Matrixnorm/Operatornorm). Falls $\|M\| < 1$, so ist die Iteration (4.4)/(4.3) konvergent (sogar in* jeder *beliebigen Norm des $\mathbb{C}^n$).*

Beweis: $\Phi : \mathbb{C}^n \to \mathbb{C}^n$ ist trivialerweise selbstabbildend.

$$\|\Phi(x) - \Phi(y)\| = \|M(x-y)\| \underset{\text{verträglich}}{\leq} \underbrace{\|M\|}_{<1} \cdot \|x-y\| \quad \forall x, y \in \mathbb{C}^n$$

$$\overset{\text{Fixpunktsatz}}{\underset{\text{von Banach}}{\Rightarrow}} \text{ Behauptung}$$

□

Es kann vorkommen, dass für *eine* Matrix M *eine* Matrixnorm $\|M\|_a > 1$, und eine *andere* Matrixnorm $\|M\|_b < 1$ erfüllt (in diesem Fall hätten wir nach Lemma 4.1 Konvergenz!). Welche Norm ist zu prüfen? Kann man (für festes M)

$$\inf\{\|M\| \mid \|\cdot\| \text{ ist zu einer Vektornorm des } \mathbb{C}^n \text{ verträgliche Matrixnorm}\} \tag{4.5}$$

charakerisieren? (Prüfe dann, ob dies < 1 ist!)
Es reicht in (4.5) nur die *Operator*normen zu betrachten, denn für eine *beliebige* verträgliche Matrixnorm $\|\cdot\|$, d.h. $\|Ax\| \leq \|A\| \cdot \|x\| \ \forall A, x$, also $\|A\| \geq \frac{\|Ax\|}{\|x\|} \ \forall A, \ x \neq 0$, gibt es immer eine Operatornorm $|||\cdot|||$ mit $|||A||| \leq \|A\|$, nämlich $|||A||| := \sup\limits_{x\neq 0} \frac{\|Ax\|}{\|x\|} \leq \|A\|$. Also gilt:

$$(4.5) = \inf\{\|M\| \mid \|\cdot\| \text{ ist eine Operatornorm}\}$$

Lemma 4.2 *Sei $\varrho(M) = \max\{|\lambda| \mid \lambda \in \mathbb{C}$ ist Eigenwert von $M\}$ der Spektralradius von $M \in \mathbb{C}^{n\times n}$. Dann gilt*

(a) Für jede *Operatornorm $\|M\| = \sup\limits_{0\neq x\in\mathbb{C}^n} \frac{\|Ax\|}{\|x\|}$ gilt:*

$$\varrho(M) \leq \|M\|$$

(d.h. $\varrho(M)$ist eine untere Schranke für die Menge (4.5)).

(b) Für jedes $\epsilon > 0$ und jede Matrix $M \in \mathbb{C}^{n\times n}$ gibt es eine Operatornorm

$$\|M\|_M = \sup_{0\neq x\in\mathbb{C}^n} \frac{\|Mx\|_M}{\|x\|_M}, \tag{4.6}$$

so dass

$$\|M\|_M \leq \varrho(M) + \epsilon.$$

Also: $\varrho(M)$ ist der gesuchte Wert (4.5)!

Bemerkung: Wir wissen: Für *symmetrisches* M können wir in (b) $\|\cdot\|_M := \|\cdot\|_2$ wählen (für beliebiges $\epsilon > 0$); es gilt $\|M\|_2 = \varrho(M)$.

Beweis von Lemma 4.2:

(a) Mit $Mx = \lambda x, x \neq 0$, folgt $\|M\| \underset{\text{Operatornorm}}{\geq} \frac{\|Mx\|}{\|x\|} = |\lambda| \Rightarrow \|M\| \geq \varrho(M)$.

(b) siehe Stoer & Bulirsch, Satz 6.8.2 oder Schwarz, 3. Auflage, Satz 11.4.

□

Lemma 4.2 besagt: $\varrho(M) < 1$ ist notwendig und hinreichend für die Existenz einer Operatornorm $\|\cdot\|_M$, so dass $\|M\|_M < 1$ ist. Damit folgt:

Lemma 4.3 *Sei $M \in \mathbb{C}^{n\times n}$, $\tilde{b} \in \mathbb{C}^n$. Die Fixpunktiteration $x^{(k+1)} = \Phi(x^{(n)})$, mit $\Phi(x) = Mx + \tilde{b}$, ist genau dann für beliebige Startwerte $x^{(0)}$ konvergent gegen den eindeutig bestimmten Fixpunkt x^* von Φ (d.h. $Mx^* + \tilde{b} = x^*$), wenn $\varrho(M) < 1$.*

Beweis:

"$\varrho(M) < 1 \Rightarrow$ Konvergenz": Aus $\varrho(M) < 1$ folgt mit Lemma 4.2 (b) die Existenz einer Vektornorm $\|\cdot\|_M$, so dass die zugehörige Matrixnorm $\|M\|_M < 1$ erfüllt. Die Behauptung folgt mit Lemma 4.1 (d.h. Fixpunktsatz von BANACH in $(\mathbb{C}^n, \|\cdot\|_M)$)

"$\varrho(M) \geq 1 \Rightarrow$ keine Konvergenz für beliebige $x^{(0)}$" : Sei $\varrho(M) \geq 1$. Angenommen $x^{(k)} \to x^*$ (x^* Fixpunkt von Φ), für beliebigen Startpunkt $x^{(0)} \in \mathbb{C}^n$. Sei x Eigenvektor von M zu einem Eigenwert $|\lambda| \geq 1$: Setze $x^{(0)} := x + x^*$

$$\begin{aligned}
\Rightarrow x^{(1)} &= Mx^{(0)} + \tilde{b} = Mx + \underbrace{Mx^* + \tilde{b}}_{=x^*} = \lambda x + x^* \\
x^{(2)} &= Mx^{(1)} + \tilde{b} = \lambda Mx + \underbrace{Mx^* + \tilde{b}}_{=x^*} = \lambda^2 x + x^* \\
&\vdots \\
\|x^{(k)} - x^*\| &= \|\lambda^k x\| = \underbrace{|\lambda|^k}_{\geq 1} \|x\| \geq \underbrace{\|x\|}_{\neq 0} \qquad \forall k \in \mathbb{N} \\
&\Rightarrow \quad x^{(k)} \nrightarrow x^*
\end{aligned}$$

□

Für die Verfahrensklassen (1) - (3) ist also sicherzustellen:

$$\varrho(Id - \omega A) < 1 \quad \text{bzw.} \quad \varrho(Id - BA) < 1 \quad \text{bzw.} \quad \varrho(W^{-1}S) < 1.$$

Im Folgenden konzentrieren wir uns auf Verfahren der Kategorie (3). Zu Kategorie (1) siehe Übung. Kategorie (2) und (3) lassen sich ineinander umwandeln (siehe ebenfalls Übung) durch:

$$B := (Id + W^{-1}S)A^{-1} \quad \text{bzw.} \quad W := B^{-1}, \quad S := A - B^{-1}.$$

4.2 Jacobi- und Gauß-Seidel-Verfahren

Man zerlegt $A = L + D + R$, wobei L und R eine strikte (linke bzw. rechte) Dreiecksmatrix und $D = \mathrm{diag}(a_{ii})$ eine Diagonalmatrix ist. Wir setzen voraus, dass $a_{ii} \neq 0\ \forall i$, d.h. D^{-1} existiert. In der Notation von Verfahrensklasse (3) in Kapitel 4.1 setzt man $W := D$ und $S = L + R$, d.h.:

$$\begin{aligned} Ax = b \quad &\Leftrightarrow \quad Dx &&= -(L+R)x + b \\ &\Leftrightarrow \quad x &&= \underbrace{-D^{-1}(L+R)x + D^{-1}b}_{\Phi(x)} \end{aligned}$$

Die zugehörige Fixpunktiteration

$$\begin{aligned} x^{(k+1)} = \Phi(x^{(k)}) &= M_{\mathrm{J}} x^{(k)} + \tilde{b} \\ \text{mit} \quad M_{\mathrm{J}} &= -D^{-1}(L+R), \quad \tilde{b} = D^{-1}b \end{aligned} \tag{4.7}$$

heißt Jacobi-*Verfahren* (Carl Jacobi, 1804-51, Berlin) oder *Gesamtschrittverfahren*. Komponentenweise geschrieben:

$$x_i^{(k+1)} := \frac{1}{a_{ii}}(-\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij} x_j^{(k)} + b_i), \quad i = 1, \ldots, n.$$

Das Update des i-ten Vektoreintrags verwendet also die Komponenten $x_1^{(k)}, \ldots, x_{i-1}^{(k)}, x_{i+1}^{(k)}, \ldots, x_n^{(k)}$ der alten Iterierten $x^{(k)}$. Naheliegend ist die Vermutung, dass man schnellere Konvergenz bekommt, wenn man stattdessen die bereits berechneten Werte $x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}$ (sowie weiterhin $x_{i+1}^{(k)}, \ldots, x_n^{(k)}$) benutzt:

$$x_i^{(k+1)} := \frac{1}{a_{ii}}(-\sum_{j<i} a_{ij} x_j^{(k+1)} - \sum_{j>i} a_{ij} x_j^{(k)} + b_i), \ i = 1, \ldots, n. \tag{4.8}$$

(4.8) heißt Gauss-Seidel-*Verfahren* oder *Einzelschrittverfahren* (Seidel: 1821-96, München). Im Computer kann (4.8) einfach mittels Überschreiben implementiert werden:

$$x_i := \frac{1}{a_{ii}}(-\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij} x_j + b_i) \tag{4.9}$$

Um die Iterationsmatrix M_{GS} des Gauß-Seidel-Verfahrens (4.8) zu identifizieren, schreiben wir es als

$$x^{(k+1)} = D^{-1}(-Lx^{(k+1)} - Rx^{(k)} + b) \qquad |D\cdot \tag{4.10}$$

$$\Leftrightarrow \quad (D+L)x^{(k+1)} = -Rx^{(k)} + b$$

$$\Leftrightarrow \quad x^{(k+1)} = -(D+L)^{-1}Rx^{(k)} + (D+L)^{-1}b, \tag{4.10a}$$

$$\text{also} \quad M_{GS} = -(D+L)^{-1}R, \quad \tilde{b} = (D+L)^{-1}b.$$

Die Invertierung von $D + L$ in (4.10) muss nur *scheinbar* erfolgen, siehe Darstellungen (4.8), (4.9). Ein Schritt des Jacobi- und des Gauss-Seidel-Verfahrens sind gleich aufwändig.

Satz 4.4 *Sei A (zeilenweise)* diagonaldominant*, d.h.*

$$\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| < |a_{ii}| \ \forall i = 1, \ldots, n \ .$$

Dann gilt:

(a) $\|M_J\|_\infty < 1$.

(b) $\|M_{GS}\|_\infty \leq \|M_J\|_\infty \ \ (< 1)$.

d.h. das Jacobi- *und das* Gauss-Seidel-*Verfahren sind konvergent.* [4]

Bemerkung: Ist A *spaltenweise* diagonaldominant, so gelten in Satz 4.4 die entsprechenden Aussagen für $\|\cdot\|_1$ statt $\|\cdot\|_\infty$.
Beweis:

(a) $\|M_J\|_\infty = \|D^{-1}(L+R)\|_\infty = \max\limits_i \sum\limits_{\substack{j=1\\ j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| < 1.$

(b) Wir zeigen:

$$\|\underbrace{\overbrace{(D+L)^{-1}R}^{-M_{GS}}x}_{=y}\|_\infty \leq \|\overbrace{D^{-1}(L+R)}^{-M_J}\|_\infty \ \forall x \in \mathbb{C}^n \ \text{ mit } \|x\|_\infty = 1. \tag{4.11}$$

daraus folgt durch sup-Bildung die Behauptung.

Setze $y := \overset{x}{(D+L)^{-1}}Rx$, also $(D+L)y = Rx, \ Dy = -Ly + Rx$; komponentenweise heißt das:

$$a_{ii}y_i = -\sum_{j<i} a_{ij}y_i + \sum_{j>i} a_{ij}x_j \ \ \forall i = 1, \ldots, n. \tag{4.12}$$

Wir zeigen per Induktion nach i, dass

$$|y_i| \leq \|D^{-1}(L+R)\|_\infty, \ \ \forall i = 1, \ldots, n \tag{4.13}$$

(woraus dann $\|y\|_\infty \leq \|D^{-1}(L+R)\|_\infty$, also (4.11), und somit die Behauptung, folgt)

$\underline{i=1}$: $|y_1| \underset{(4.12)}{=} \left|\frac{1}{a_{11}}\sum\limits_{j=2}^{n} a_{1j}x_j\right| \leq \sum\limits_{j=2}^{n} \left|\frac{a_{1j}}{a_{11}}\right| \cdot \underbrace{|x_j|}_{\leq 1} \leq \sum\limits_{j=2}^{n} \left|\frac{a_{1j}}{a_{11}}\right|$ = erste Zeilensumme von $D^{-1}(L+R)$,
also $\leq \|D^{-1}(L+R)\|_\infty$.

[4] Aussage (b) "passt" zu der zuvor geäußerten Intuition, dass das Gauß-Seidel-Verfahren schneller sein sollte als das Jacobi-Verfahren, ist aber *kein* strenger Beweis einer solchen Behauptung. Vergleiche auch S.76, L.A. II.

$\underline{\{1,\dots,i-1\} \to i}$: $\|y_i\| \underset{(4.12)}{\leq} \sum_{j<i} \left|\frac{a_{ij}}{a_{ii}}\right| \cdot \underbrace{|x_j|}_{<1} + \sum_{j>i} \left|\frac{a_{ij}}{a_{ii}}\right| \cdot \underbrace{|x_j|}_{\leq 1}$.

Nach Induktionsvorraussetzung ist

$$\begin{aligned} |y_j| &\underset{\text{I.V. für } j<i}{\leq} \|D^{-1}(L+R)\|_\infty \underset{\text{(a)}}{<} 1 \text{ für } j<i \\ \Rightarrow |y_i| &\leq \sum_{j\neq i} \left|\frac{a_{ij}}{a_{ii}}\right| = i\text{-te Zeilensumme von } D^{-1}(L+R) \\ &\leq \|D^{-1}(L+R)\|_\infty \end{aligned}$$

Also: (4.13) gilt für alle $i = 1,\dots,n$.
$\Rightarrow \|(D+L)^{-1}R\|_\infty = \max_{\|x\|_\infty=1} \| \underbrace{(D+L)^{-1}Rx}_{=y} \|_\infty \underset{(4.13)}{\leq} \|D^{-1}(L+R)\|_\infty$

□

Die Forderung der Diagonaldominanz in Satz 4.4 ist in der Praxis oft zu restriktiv. Siehe zum Beispiel die Diskretisierung der POISSON-Gleichung

$$-\Delta u = f \quad (\text{bzw. } -u'' = f \text{ in 1-D})$$

[vgl. L.A.I Blatt 7 sowie L.A. II Seite 75 f., L.A. II Blatt 8]. Dort hat man Matrizen A, die nur

$$|a_{ii}| \geq \sum_{\substack{j=1 \\ j\neq n}} |a_{ij}| \quad \forall i = 1,\dots,n \tag{4.14}$$

und

$$\exists\, i \in \{1,\dots,n\} \quad |a_{ii}| > \sum_{\substack{j=1 \\ j\neq n}}^{n} |a_{ij}| \tag{4.15}$$

erfüllen.

Eine Matrix $A \in \mathbb{C}^{n\times n}$, die (4.14) und (4.15) erfüllt, heißt *schwach diagonaldominant*. Man kann zeigen, dass Satz 4.4 auch für schwach diagonaldominante Matrizen gültig ist, sofern zusätzlich vorausgesetzt wird, dass A *irreduzibel* ist. Eine Matrix $A \in \mathbb{C}^{n\times n}$ heißt *irreduzibel*, falls es für jede disjunkte Zerlegung der Indexmenge

$$\{1,\dots,n\} = I_1 \cup I_2,\ I_1, I_2 \neq \emptyset,\ I_1 \cap I_2 = \emptyset,$$

stets Indizes $i \in I_1,\ j \in I_2$ gibt mit $a_{ij} \neq 0$. Äquivalent dazu ist:
Es gibt keine Permutationsmatrix P, so dass

$$P^{-1}AP = \begin{pmatrix} B_1 & 0 \\ B_2 & B_3 \end{pmatrix},$$

wobei B_1 und B_3 quadratische Blöcke sind. Es gibt einen grafentheoretischen Algorithmus zum Testen der Irreduzibilität (s. Schwarz & Köckler, Seite 497 f.). Ist A *reduzibel*, so kann das Lösen von $Ax = b$ offensichtlich in zwei (oder mehr) getrennte lineare Gleichungssysteme, deren Matrix jeweils irreduzibel ist, gesplittet werden, d.h. die Forderung der Irreduzibilität ist keine große Einschränkung. Zum Beweis der Verallgemeinerung von Satz 4.4 auf schwach diagonaldominante irreduzible Matrizen siehe z.B. Schwarz & Köckler.

4.3 Relaxation für Gauß-Seidel- und Jacobi-Verfahren

Iterative Verfahren lohnen sich im Allgemeinen dann, wenn nicht allzuviele Iterationsschritte nötig sind[5]. Daher ist es aus Effizienzgründen erforderlich, dass die Kontraktionsrate $\varrho(T)$ hinreichend klein ist. Für lineare Gleichungssysteme, die aus der Diskretisierung von PDEs entstehen, hat man oft

$$\varrho(M_{\mathrm{GS}}) \to 1, \varrho(M_{\mathrm{J}}) \to 1$$

für Diskretisierungsparameter $h \to 0$ (d.h. $n \to \infty$), d.h. asymptotisch immer schlechter werdende Konvergenzraten (s. Übung). Eine Möglichkeit, die Konvergenzrate zu verbessern, bietet oft die *Relaxation*. Beim JACOBI-Verfahren

$$\begin{aligned} x^{(k+1)} &:= -D^{-1}(L+R)x^{(k)} + D^{-1}b \\ &= x^{(k)} \underbrace{-D^{-1}(L+D+R)x^{(k)} + D^{-1}b}_{=:\Delta x^{(k)}} \end{aligned}$$

ersetzt man einfach das Update $\Delta x^{(k)} = x^{(k+1)} - x^{(k)}$ durch $\omega \cdot \Delta x^{(k)}$, $\omega \in \mathbb{R}$, d.h.

$$\begin{aligned} x^{(k+1)} &= x^{(k)} + \omega \Delta x^{(k)} \\ &= (1-\omega)x^{(k)} - \omega D^{-1}(L+R)x^{(k)} + \omega D^{-1}b. \end{aligned} \tag{4.16}$$

(4.16) heißt *relaxiertes* JACOBI-*Verfahren*, $\omega \in \mathbb{R}$ heißt *Relaxationsparameter*. Für $\omega > 1$ spricht man von *Überrelaxation*, für $\omega < 1$ von *Unterrelaxation*. Nichtsdestotrotz hat sich für (4.16), $\omega \in \mathbb{R}$, der Begriff *JOR-Verfahren* (Jacobi overrelaxation) eingebürgert. Das JOR-Verfahren hat die Iterationsmatrix

$$\begin{aligned} M_{\mathrm{JOR}}(\omega) &= (1-\omega)Id - \omega D^{-1}(L+R) \\ &= (1-\omega)Id + \omega M_{\mathrm{J}}, \\ M_{\mathrm{JOR}}(1) &= M_{\mathrm{J}}. \end{aligned} \tag{4.17}$$

Beim GAUSS-SEIDEL-Verfahren geht man ähnlich vor. Die Darstellungen (4.8)/(4.10) (*nicht* Darstellung (4.10a)!) werden verwendet, um $\Delta x^{(k)}$ zu erklären:

$$\begin{aligned} x^{(k+1)} &= D^{-1}(-Lx^{(k+1)} - Rx^{(k)} + b) \\ &= x^{(k)} \underbrace{-D^{-1}Lx^{(k+1)} - D^{-1}(D+R)x^{(k)} + D^{-1}b}_{=:\Delta x^{(k)}} \end{aligned}$$

Dies wird modifiziert zu

$$\begin{aligned} x^{(k+1)} &= x^{(k)} + \omega \Delta x^{(k)} \\ &= (1-\omega)x^{(k)} - \omega D^{-1}(Lx^{(k+1)} + Rx^{(k)}) + \omega D^{-1}b. \end{aligned} \tag{4.18}$$

Für dieses Verfahren hat sich der Name *SOR-Verfahren* (successive overrelaxation) eingebürgert. Die Darstellung (4.18) des SOR-Verfahrens ist ideal für die Implementierung. Zur Ermittlung der

[5] z.B. wenn die Anzahl der Schritte $\approx n^{\alpha}$ mit $\alpha < 1$ für $n \times n$ Matrizen, die aus der Diskretisierung von partiellen Differentialgleichungen hervorgehen.

Iterationsmatrix formen wir um:

$$\begin{aligned}(4.18) \quad &\Leftrightarrow \quad (Id + \omega D^{-1}L)x^{(k+1)} = (1-\omega)x^{(k)} - \omega D^{-1}Rx^{(k)} + \omega D^{-1}b \quad |D\cdot \\ &\Leftrightarrow \quad (D+\omega L)x^{(k+1)} = (1-\omega)D^{(k)} - \omega R x^{(k)} + \omega b \\ &\Leftrightarrow \quad x^{(k+1)} = (D+\omega L)^{-1}[(1-\omega)D - \omega R]x^{(k)} + \omega(D+\omega L)^{-1}b,\end{aligned}$$

die Iterationsmatrix lautet also:

$$\boxed{M_{\mathrm{SOR}}(\omega) = (D+\omega L)^{-1}[(1-\omega)D - \omega R]}$$

Lemma 4.5 *Falls das* JACOBI-*Verfahren konvergent ist, so ist auch das JOR-Verfahren konvergent für* $0 < \omega \leq 1$.

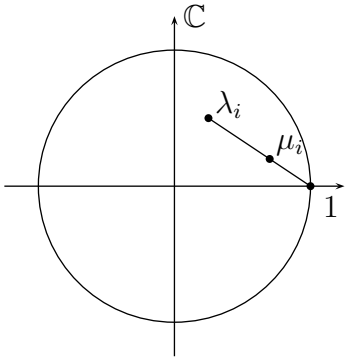

Abbildung 16: Konvexkombination

Beweis:

Nach Voraussetzung erfüllen alle Eigenwerte $\lambda_i \in \mathbb{C}$ von M_J : $|\lambda_i| < 1$. Nach (4.17) lauten die Eigenwerte von $M_{\mathrm{JOR}}(\omega)$: $\mu_i = (1-\omega)\cdot 1 + \omega\lambda_i$.
Das heißt $\mu_i \in \mathbb{C}$ ist Konvexkombination von λ_i und 1 (wobei der Koeffizient von 1 nicht null ist). Daraus folgt $|\mu_i| < 1$.

□

Lemma 4.6 *Es sei* $A \in \mathbb{R}^{n\times n}$ *s.p.d., und das* JACOBI-*Verfahren sei konvergent. Dann ist das JOR-Verfahren konvergent für alle* $\omega \in \mathbb{R}$ *mit*

$$0 < \omega < \frac{2}{1-\mu_{\min}} \left(\underset{(a)}{\leq} 2 \right),$$

wobei $\mu_{\min}$ *der kleinste Eigenwert von* M_J *ist; dieser ist insbesondere kleiner gleich 0.*

Beweis : A symmetrisch $\Rightarrow$ $L+R$ symmetrisch und hat nur reelle Eigenwerte, A positiv definit $\Rightarrow a_{ii} > 0 \Rightarrow D^{\frac{1}{2}} := \mathrm{diag}(\sqrt{a_{ii}})$ existiert.

$$M_J = -D^{-1}(L+R)$$

ist *ähnlich* zur Matrix (d.h. hat die gleichen Eigenwerte wie)

$$S := -D^{\frac{1}{2}} M_{\mathrm{J}} D^{-\frac{1}{2}} = -D^{-\frac{1}{2}}(L+R)D^{-\frac{1}{2}},$$

welche *symmetrisch* ist und daher nur reelle Eigenwerte hat. S hat Spur $= 0$, die Spur ist die Summe aller Eigenwerte $\Rightarrow$ S (also auch M_{J}) hat kleinsten Eigenwert $\mu_{\min} \leq 0$. Wegen der vorausgesetzten Konvergenz des Jacobi-Verfahrens hat M_{J} Eigenwerte $-1 < \mu_i < 1$. $M_{\mathrm{JOR}}(\omega)$ hat (nach (4.17)) Eigenwerte $\nu_j = 1 - \omega + \omega\mu_j$. Notwendige und hinreichende Bedingung für die Konvergenz des JOR-Verfahrens ist also

$$\begin{aligned} -1 < \quad & 1-\omega+\omega\mu_j \quad < 1 \qquad && |-1 \\ -2 < \quad & -\omega+\omega\mu_j \quad < 0 \qquad && |\cdot(-1) \\ 2 > \quad & \omega(1-\mu_j) \quad > 0 \qquad && |\cdot \frac{1}{1-\mu_j} \end{aligned}$$

Da nach Voraussetzung das Jacobi-Verfahren konvergent ist, ist $1-\mu_j > 0 \ \ \forall j$, also

$$\underbrace{\frac{2}{1-\mu_j}}_{\text{wird minimal für} \mu_j = \mu_{\min}} > \omega > 0 \quad \forall \text{ Eigenwerte } \mu_j \text{ von } M_{\text{J}}$$

Dies ist äquivalent zu

$$\frac{2}{1-\mu_{\min}} > \omega > 0.$$

□

Lemma 4.7 *(optimale Wahl des Relaxationsparameters für JOR)*
Sei A symmetrisch (M_J ist dann ähnlich zu einer symmetrischen Matrix, s. Beweis Lemma 4.6) und die (dann reellen) Eigenwerte von M_J erfüllen

$$\mu_1 \leq \mu_2 \leq \ldots \leq \mu_n < 1$$

(ist z.B. erfüllt, wenn das Jacobi-Verfahren konvergent ist; z.B. wenn A diagonal dominant ist). Dann ist die optimale Wahl von ω, d.h. ω_{opt} so, dass

$$\varrho(M_{JOR}(\omega_{opt})) = \min_{\omega \in \mathbb{R}} \varrho(M_{JOR}(\omega)),$$

gegeben durch

$$\omega_{opt} = \frac{2}{2-\mu_n-\mu_1}.$$

Es ist dann $\varrho(M_{JOR}(\omega_{opt})) = \frac{\mu_n-\mu_1}{2-\mu_n-\mu_1}$.

Bemerkung: Leider kennt man im Allgemeinen die Eigenwerte von M_{J} nicht (noch nicht einmal die von A), d.h. man bestimmt ein "gutes" ω durch Ausprobieren.

Beweis : Es ist $\mu_1 \leq 0, \mu_n \geq 0$ (wie im Beweis von Lemma 4.7). Es ist

$$M_{\text{JOR}}(\omega) = (1-\omega)Id + \omega M_{\text{J}},$$

d.h. M_{JOR} hat die Eigenwerte $\nu_i = (1-\omega) + \omega\mu_i$. Offensichtlich (siehe grafische Darstellung, Abbildung 17) kommen nur $\omega > 0$ in Frage. Für $\omega > 0$ hängen die ν_i monoton von den μ_i ab:

$$\begin{gathered} \nu_1 \leq \ldots \leq \nu_n < 1 \\ \Rightarrow \varrho(M_{\text{JOR}}(\omega)) = \max\{|\nu_1|, \ldots, |\nu_n|\} = \max\{|\nu_1|, |\nu_n|\}. \end{gathered}$$

Dieser Wert wird *minimal*, wenn

$$\begin{aligned} & -\nu_1 = \nu_n \\ \text{d.h.} \quad & -1+\omega-\omega\mu_1 = 1-\omega+\omega\mu_n \\ \Leftrightarrow \quad & 2-2\omega+\omega(\mu_1+\mu_n) = 0 \\ \Leftrightarrow \quad & 2 = \omega(2-\mu_1-\mu_n) \\ \Leftrightarrow \quad & \omega = \frac{2}{2-\mu_1-\mu_n}. \end{aligned}$$

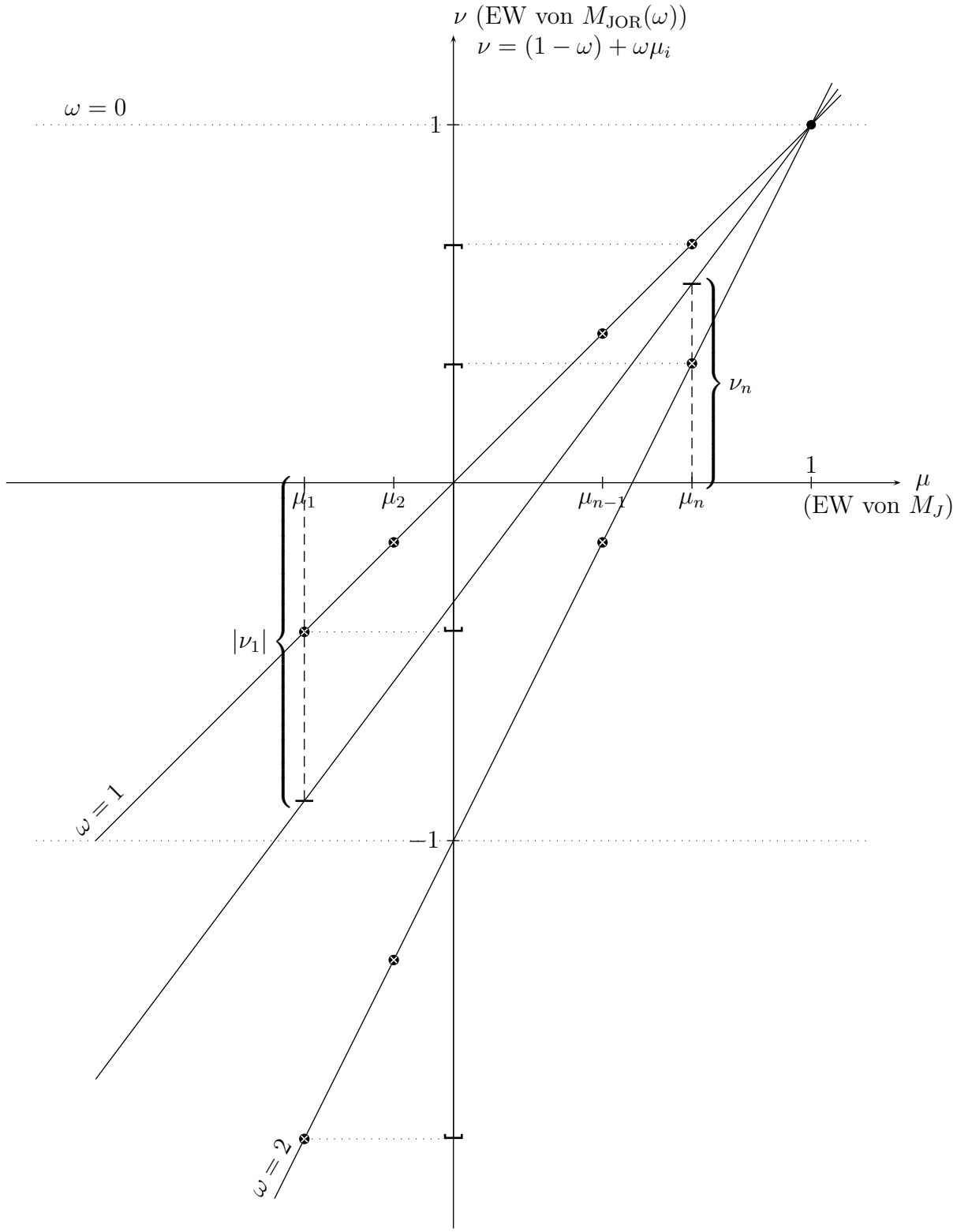

Abbildung 17: Eigenwerte von M_J und M_{JOR} für verschiedene ω

Für dieses ω ist

$$\varrho(M_{\text{JOR}}(\omega)) = -\nu_1 = +\nu_n = 1 + \omega \cdot (\mu_n - 1) = 1 + \frac{2}{2 - \mu_1 - \mu_n} \cdot (\mu_n - 1) = \frac{\mu_n - \mu_1}{2 - \mu_1 - \mu_n}.$$

□

Bemerkung:
Falls $\mu_1 = -\mu_n$, was bei der Diskretisierung der LAPLACE-Gleichung zum Beispiel der Fall ist, so ist $\omega_{\text{opt}} = 1$, d.h. die Relaxation bringt keine Verbesserung.

Auch für das *SOR*-Verfahren lassen sich Konvergenzsätze sowie eine Formel für die optimale Wahl von ω angeben. Die Analyse des SOR-Verfahrens ist allerdings schwieriger als die des JOR-Verfahrens (Lemma 4.5 - 4.7), da bei SOR die Iterationsmatrix

$$M_{\text{SOR}}(\omega) = (D + \omega L)^{-1}[(1 - \omega)D - \omega R]$$

nichtlinear von ω abhängt und außerdem auch bei symmetrisch positiv definitem A im Allgemeinen nicht ähnlich zu einer spd-Matrix ist.
Hier einige Resultate:

Lemma 4.8

a) Es gilt $\varrho(M_{SOR}(\omega)) \geq |1 - \omega|$,
d.h. für $\omega \leq 0$ und für $\omega \geq 2$ ist das SOR-Verfahren divergent.

b) Für diagonaldominantes A (auch für schwach diagonaldominantes und irreduzibles A) ist das SOR-Verfahren für $0 < \omega \leq 1$ konvergent.

c) Für A s.p.d. ist das SOR-Verfahren für $0 < \omega < 2$ konvergent.

Beweis:

a)

$$detM_{\text{SOR}}(\omega) \underset{\text{Det. Prod. Satz}}{=} \frac{det[(1 - \omega)D - \omega R]}{det(D + \omega L)} \underset{det \text{ von } \Delta-\text{Matrizen}}{=} \frac{(1 - \omega)^n detD}{detD} = (1 - \omega)^n$$

Da Determinante = Produkt der Eigenwerte $\in \mathbb{C}$:

$$\prod_{i=1}^{n} \lambda_i = (1 - \omega)^n$$

Ferner

$$\varrho(M_{\text{SOR}}(\omega))^n = \max_{i=1\ldots n} |\lambda_i|^n \geq \prod_{i=1}^{n} |\lambda_i| = |1 - \omega|^n$$

$\Rightarrow$ Behauptung

b) und c) : siehe Schwarz & Köckler Satz 11.13, 11.15

□

Definition 4.9 *Eine Matrix $A \in \mathbb{R}^{n\times n}$ heißt* T - Matrix*, falls sie die blockweise Tridiagonalgestalt*

$$A = \begin{pmatrix} D_1 & R_1 & & & & \\ U_1 & \ddots & \ddots & & 0 & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & \ddots & \ddots & R_{s-1} \\ & & & & U_{s-1} & D_s \end{pmatrix}$$

hat, wobei die D_i quadratische *Diagonalmatrizen sind (die R_i, U_i sind im Allgemeinen rechteckig).*

Bemerkung: Matrizen, die bei der Diskretisierung von partiellen Differentialgleichungen mittels sogenannter finiter Differenzen entstehen, sind, bei geeigneter Nummerierung der Gleichungen/Unbekannten, oft T-Matrizen[6]. T-Matrizen haben die Eigenschaft, dass $det(\alpha L + \beta D + \frac{1}{\alpha}R)$ unabhängig von α ist. Diese Eigenschaft kann genutzt werden, um die Eigenwerte der komplizierten Matrix $M_{\mathrm{SOR}}(\omega)$ mit denen der einfach aufgebauten Matrix M_{J} in Beziehung setzen. So lässt sich beweisen (s. Schwarz & Köckler, Kapitel 11.2.3):

Lemma 4.10 *(Varga '62, Young '71)*
Ist A eine T-Matrix mit $a_{ii} \neq 0$, und besitzt $M_J = -D^{-1}(L+R)$ nur reelle Eigenwerte μ_j, und gilt $\varrho(M_J) < 1$, dann ist der optimale Relaxationsparameter ω_{opt} des SOR-Verfahrens gegeben durch

$$\omega_{opt} = \frac{2}{1+\sqrt{1-\varrho(M_J)^2}} \in [1,2)$$

und der zugehörige optimale Spektralradius ist

$$\varrho(M_{SOR}(\omega_{opt})) = \omega_{opt} - 1 \in [0,1).$$

$\varrho(M_{\mathrm{J}})$	ω_{opt}	$\varrho(M_{\mathrm{SOR}}(\omega_{\mathrm{opt}}))$
0.9	1.393	0.393
0.99	1.753	0.753
0.999	1.914	0.914

Bemerkung: Die Abbildung $\omega \to \varrho(M_{\mathrm{SOR}}(\omega))$ hat die Form

d.h. es ist besser ein etwas zu großes als ein etwas zu kleines ω zu wählen. In der Praxis: ω durch Ausprobieren bestimmen.

Allgemeine Bemerkung zu Fixpunktverfahen:

[6]Die Gitterpunkte werden schachbrettartig schwarz oder weiß "eingefärbt", so dass jeder Gitterpunkt von vier Gitterpunkten der jeweils anderen Farbe umgeben ist; dann startet man mit den "weißen" Gleichungen/Unbekannten und endet mit den "schwarzen". Es ergibt sich die obige Struktur mit $s = 2$.

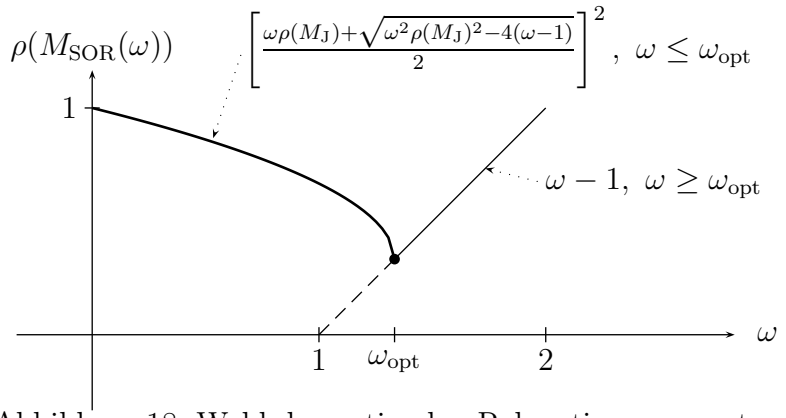

Abbildung 18: Wahl des optimalen Relaxationsparameters.

- Ein prinzipieller Vorteil von Fixpunktverfahren zum Lösen von linearen Gleichungssystemen gegenüber direkten Verfahren ist, dass sich Rundungsfehler nicht akkumulieren können!
- In der Praxis kennt man die Eigenwerte von M_{J} nicht (nicht einmal die von A)
  $\rightarrow$ gutes ω für SOR und JOR durch *Ausprobieren* finden!