Prof. Dr. Irwin Yousept Universität Duisburg-Essen

Numerik gewöhnlicher Differentialgleichungen

Typeset und Layout: Roman Händler
Fassung vom 4. März 2017

Inhaltsverzeichnis

Kapitel 1

Nichtlineare Gleichungssysteme und Ausgleichsprobleme

1.1 Wiederholung zum Newton-Verfahren und Erweiterung

In Numerik I haben wir das Newton-Verfahren zur Lösung von

$$F(x) = 0, \quad F : \mathbb{R}^n \to \mathbb{R}^n$$

behandelt. Ausgehend von der aktuellen Iterierten $x^k \in \mathbb{R}^n$ löst man das lineare Gleichungssystem

$$F'(x^k)d^k = -F(x^k)$$

und dann definiert man

$$x^{k+1} = x^k + d^k \quad \Leftrightarrow \quad x^{k+1} = x^k - F'(x^k)^{-1}F(x^k).$$

Beachte, dass $F'(x^k) = (\partial_i F_j(x^k)) \in \mathbb{R}^{n\times n}$ die Jacobi-Matrix von F an der Stelle x^k bezeichnet.

Satz 1.1. *Es sei $D \subset \mathbb{R}^n$ offen und konvex, $F : \mathbb{R}^n \supset D \to \mathbb{R}^n$ stetig differenzierbar mit Jacobi-Matrix $F'(x)$, welche für alle $x \in D$ regulär (invertierbar) ist. Es existiere eine Konstante $\omega > 0$, so dass gilt*

$$\|F'(x)^{-1}(F'(x+sv) - F'(x))v\|_2 \leq s\omega \|v\|_2^2 \tag{1.1}$$

für alle $x \in D$, $v \in \mathbb{R}^n$ mit $x + v \in D$, $s \in [0,1]$. Es sei $F(\bar{x}) = 0$, $\bar{x} \in D$, und der Startwert $x^0 \in D$ so gewählt, dass

$$\rho := \|\bar{x} - x^0\|_2 < \frac{2}{\omega} \quad \textit{und} \quad B_\rho(\bar{x}) \subset D.$$

Dann erzeugt das Newton-Verfahren eine eindeutig definierte Folge $\{x^k\} \subset B_\rho(\bar{x})$ mit

$$\lim_{k\to\infty} x^k = \bar{x}.$$

Weiter gilt

$$\|x^{k+1} - \bar{x}\|_2 \leq \frac{\omega}{2}\|x^k - \bar{x}\|_2^2 \quad \textit{(quadratische Konvergenz).}$$

Bemerkung 1.2.

(i) Eine andere äquivalente Variante des Satzes haben wir bereits in Numerik I bewiesen.

(ii) Die Voraussetzung (1.1) sieht etwas kompliziert aus. Eingängiger ist die Folgende: $F'(\bar{x})$ sei regulär und F' sei lokal Lipschitz-stetig, das heißt es existieren $\Theta, L > 0$, so dass

$$\left\|F'(x) - F'(y)\right\|_2 \leq L \left\|x - y\right\|_2 \quad \forall x, y \in B_\Theta(\bar{x}).$$

Da $F'(\bar{x})$ regulär ist, gibt es ein $\varepsilon > 0$, so dass die Jacobi-Matrizen

$$F'(x) \in \mathbb{R}^{n\times n} \quad \forall x \in B_\varepsilon(\bar{x})$$

regulär sind und

$$\|F'(x)^{-1}\|_2 \leq 2\|F'(\bar{x})^{-1}\|_2 \quad \forall x \in B_\varepsilon(\bar{x}).$$

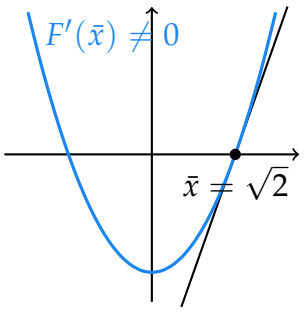
Es gilt also

$$\begin{aligned}
\|F'(x)^{-1}(F'(x+sv) - F'(x))v\|_2 &\leq \|F'(x)^{-1}\|_2 \left\|(F'(x+sv) - F'(x))v\right\|_2 \\
&\leq \|F'(x)^{-1}\|_2 \left\|F'(x+sv) - F'(x)\right\|_2 \|v\|_2 \\
&\leq 2\|F'(\bar{x})^{-1}\|_2 L \left\|sv\right\|_2 \|v\|_2 \\
&= s \underbrace{2\|F'(\bar{x})^{-1}\|_2 L}_{=:\omega} \|v\|_2^2
\end{aligned}$$

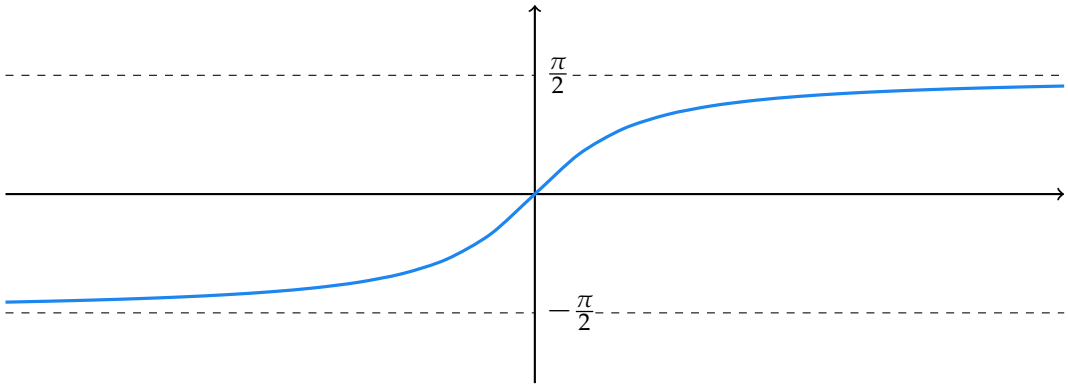
für alle $x \in B_{\tilde{\varepsilon}}(\bar{x})$ und $v \in \mathbb{R}^n$ mit $x + v \in B_{\tilde{\varepsilon}}(\bar{x})$, $s \in [0,1]$ mit $\tilde{\varepsilon} := \min\{\varepsilon, \Theta\}$. Wählen wir nun $D = B_{\tilde{\varepsilon}}(\bar{x})$, so sehen wir, dass (1.1) erfüllt ist.

Beispiele 1.3.

(i) Berechnung von $\sqrt{2}$. Definiere $F(x) := x^2 - 2$. Dann ist eine Nullstelle von $F(x)$ gegeben durch $\bar{x} = \sqrt{2}$. Wir erhalten die Ableitung $F'(x) = 2x$, welche global Lipschitz ist mit $F'(\bar{x}) = 2\sqrt{2} \neq 0$.



(ii) $F(x) = \arctan(x)$.



F hat genau eine Nullstelle bei $\bar{x} = 0$. Die Newtonvorschrift lautet

$$\begin{aligned} x^{k+1} = x^k + d^k &= x^k - F'(x^k)^{-1} F(x^k) \\ &= x^k - (1 + x^{k^2}) \arctan(x^k), \end{aligned}$$

denn

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

Gilt $\arctan(|x^0|) \geq \frac{2|x^0|}{1+x^{0^2}}$, so divergiert das Newton-Verfahren und bei „>“ gilt

$$\lim_{k \to \infty} |x^k| = \infty.$$

Bei $\arctan(|x^0|) = \frac{2|x^0|}{1+x^{0^2}}$ (o.B.d.A. $\arctan(x^0) = \frac{2x^0}{1+x^{0^2}}$) gilt

$$x^1 = x^0 + d^0 = x^0 - (1 + x^{0^2}) \frac{2x^0}{1 + x^{0^2}} = -x^0$$

und

$$x^2 = x^1 - (1 + x^{1^2}) \arctan(x^1) = -x^0 + (1 + x^{0^2}) \arctan(x^0) = x^0.$$

In diesem Fall gilt $x^k = (-1)^k x^0$.

Das obige Beispiel zeigt, dass das Newton-Verfahren im Allgemeinen nur lokal konvergiert. Bei der numerischen Umsetzung kann man nicht von vornherein wissen, ob man im Einzugsbereich des Verfahrens liegt. Deshalb benötigt man Konvergenztests und Strategien zur Erweiterung des Konvergenzbereiches.

1.1.1 Konvergenztests

Intuitiv nimmt man an, dass die aktuelle Iterierte umso näher an der Lösung $\bar{x}$ ist, je kleiner das Residuum $\|F(x^k)\|_2$ ist. Um $\|F(x^k)\|_2$ zu minimieren, wäre deshalb ein monotoner Abstieg

$$\|F(x^{k+1})\|_2 \leq \vartheta \|F(x^k)\|_2 \quad \text{„Standard-Monotonietest“}$$

mit einer positiven Zahl $0 < \vartheta < 1$ wünschenswert. Verletzt das Verfahren diese Bedingung, so bricht man ab. Ein anderer Monotonietest ist wie folgt:

$$\|F'(x^k)^{-1}F(x^{k+1})\|_2 \leq \vartheta \|F'(x^k)^{-1}F(x^k)\|_2 \quad \text{„Natürlicher Monotonietest.“}$$

Der natürliche Monotonietest lässt sich wie folgt motivieren:

$$x^{k+1} - x^k = d^k = -F'(x^k)^{-1}F(x^k) \quad \text{„Newton-Korrektur.“}$$

Die neue Newton-Korrektur lautet

$$d^{k+1} = -F'(x^{k+1})^{-1}F(x^{k+1}).$$

Wir verwenden aber die vereinfachte Newton-Korrektur

$$\bar{d}^{k+1} = -F'(x^k)^{-1}F(x^{k+1}).$$

Dann lautet der natürliche Monotonietest

$$\|\bar{d}^{k+1}\|_2 \leq \vartheta \|d^k\|_2.$$

Zur Durchführung des natürlichen Monotonietests müssen wir das lineare Gleichungssystem

$$F'(x^k)\bar{d}^{k+1} = -F(x^{k+1})$$

lösen. Das ist aber bei dem aktuellen Schritt einfach zu lösen, denn zur Lösung von

$$F'(x^k)d^k = -F(x^k)$$

hat man bereits eine LR-Zerlegung $F'(x^k) = LR$ gefunden. Also ist

$$F'(x^k)\bar{d}^{k+1} = -F(x^{k+1})$$

äquivalent zu

$$LR\bar{d}^{k+1} = -F(x^{k+1}).$$

Daraus ergibt sich

$$\begin{aligned} Ly &= -F(x^{k+1}) && \text{„Vorwärtssubstitution“} \\ R\bar{d}^{k+1} &= y && \text{„Rückwärtssubstitution“.} \end{aligned}$$

Als Abbruchkriterium verwendet man oft

1: **if** $\left\|\bar{d}^{k+1}\right\|_2 > \frac{1}{2}\left\|d^k\right\|_2$ **then**
2: STOP
3: **end if**

1.1.2 Erweiterung des Konvergenzbereichs

In der Praxis verwendet man als Standardmethode die Dämpfung des Newton-Verfahrens. An Stelle von

$$x^{k+1} = x^k + d^k \quad \text{(voller Newton-Schritt)}$$

verwendet man

$$x^{k+1} = x^k + \lambda_k d^k$$

mit Dämpfungsfaktor $\lambda_k \in (0, 1]$. Man geht vorsichtig vor, um ein „Überschießen“ wie beim arctan-Beispiel zu vermeiden.
Idee: Natürlicher Konvergenztest mit

$$\vartheta := (1 - \frac{\lambda_k}{2}).$$

Großer Schritt: $\lambda_k \approx 1$ und $\vartheta \approx \frac{1}{2}$ (enges Abbruchkriterium).
Vorsichtiger Schritt: $\lambda_k \ll 1$ und $\vartheta \approx 1$.
Gesucht ist nun ein geeignetes λ_k aus $(0, 1]$ mit

$$\|F'(x^k)^{-1} F(x^k + \lambda_k d^k)\|_2 \leq (1 - \frac{\lambda_k}{2}) \|d^k\|_2.$$

Daraus ergibt sich das folgende Verfahren:

Algorithmus 1.1 Gedämpftes Newton-Verfahren mit natürlichem Konvergenztest

---

(S0) Wähle einen Startwert $x^0 \in \mathbb{R}^n$ und setze $\lambda_0 = 1$, $k = 0$.
(S1) Abbruchkriterium:
1: **if** $\left\|F(x^k)\right\|_2 = 0$ **then**
2: STOP
3: **end if**
(S2) Bestimme $d^k \in \mathbb{R}^n$ als Lösung von

$$F'(x^k) d^k = -F(x^k).$$

(S3) Monotonietest:
4: **while** $\left\|F'(x^k)^{-1} F(x^k + \lambda_k d^k)\right\|_2 > (1 - \frac{\lambda_k}{2}) \left\|d^k\right\|_2$ **do**
5: $\lambda_k = 0.5 \lambda_k$
6: **end while**
(S4) Setze $x^{k+1} = x^k + \lambda_k d^k$ und

$$\lambda_{k+1} = \min\{1, 2\lambda_k\},$$

gehe zu (S1).

Bemerkung 1.4.

(i) Zur Bestimmung von d^k verwendet man zum Beispiel eine LR-Zerlegung für $F'(x^k)$. Die gespeicherte LR-Zerlegung wird dann für den Monotonietest in (S3) verwendet.

(ii) Hier kann man auch das vereinfachte Newton-Verfahren verwenden (Abschnitt 5.3, Numerik I).

(iii) Weglassen von kleinen unwichtigen Elementen in $F'(x^k)$ („sparsing“).

(iv) Fortsetzungsmethoden behandeln wir später.

1.2 Newton-Verfahren und Nullstellen von Polynomen

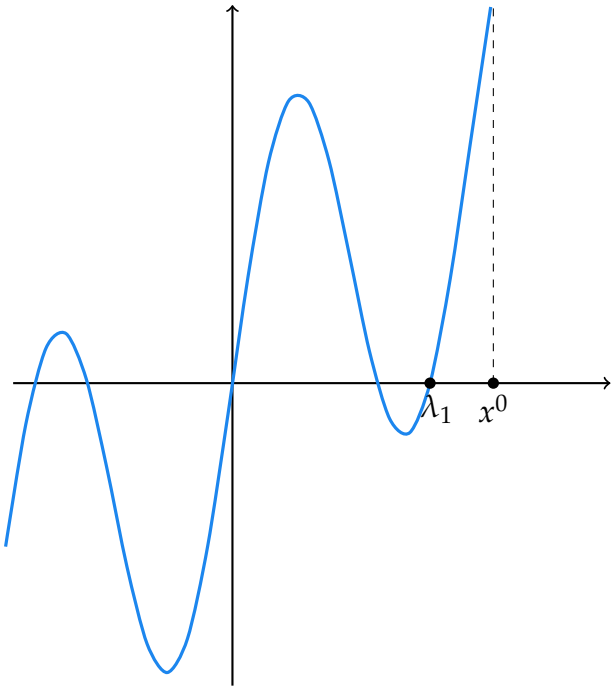
Die Nullstellenbestimmung von Polynomen ist überaus wichtig für die Berechnung von Eigenwerten. In diesem Abschnitt zeigen wir, wie man die größte reelle Nullstelle eines reellen Polynoms durch das Newton-Verfahren berechnen kann.

Satz 1.5 (Größte Nullstelle)**.** *Es sei $p \in \Pi_n$ ein reelles Polynom vom Grad $n \in \mathbb{N}$, welches eine reelle Nullstelle $\lambda_1 \in \mathbb{R}$ besitze, so dass gilt:*

$$\operatorname{Re} \zeta \leq \lambda_1$$

für alle anderen Nullstellen $\zeta \in \mathbb{C}$ von p. Dann konvergiert das Newton-Verfahren für jeden Startwert $x^0 > \lambda_1$ streng monoton fallend gegen λ_1.

Illustration 1.6.



Beweis. Wir zählen die Nullstellen von p wie folgt:

$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_l$	reelle Nullstellen,
$\zeta_1, \bar{\zeta}_1, \ldots, \zeta_m, \bar{\zeta}_m$	zueinander konjugierte komplexe Paare von Nullstellen $\notin \mathbb{R}$.

Daraus folgt

$$p(x) = \prod_{k=1}^{l}(x - \lambda_k) \prod_{k=1}^{m}(x - \zeta_k)(x - \bar{\zeta}_k). \tag{1.2}$$

Für die Ableitung $f'(x)$ eines Produkts

$$f(x) = \prod_{k=1}^{n}(x - \eta_k)$$

machen wir uns klar, dass

$$\begin{aligned} f'(x) = \sum_{k=1}^{n} \prod_{\substack{j=1 \\ j\neq k}}^{n}(x - \eta_j) &= \left(\sum_{k=1}^{n} \frac{1}{(x - \eta_k)} \right) \prod_{j=1}^{n}(x - \eta_j) \\ &= \left(\sum_{k=1}^{n} \frac{1}{(x - \eta_k)} \right) f(x) \end{aligned}$$

gilt. Nach diesem Prinzip erhalten wir für p':

$$p'(x) = \left(\sum_{k=1}^{l} \frac{1}{(x - \lambda_k)} + 2 \sum_{k=1}^{m} \frac{x - \operatorname{Re}\zeta_k}{(x - \zeta_k)(x - \bar{\zeta}_k)} \right) p(x), \tag{1.3}$$

denn $(x - \zeta_k)(x - \bar{\zeta}_k) = x^2 - 2x\operatorname{Re}(\zeta_k) + |\zeta_k|^2$, also

$$\frac{d}{dx} = 2(x - \operatorname{Re}\zeta_k) \quad (x \text{ ist reell}).$$

Weiter gilt

$$\begin{aligned} (x - \zeta_k)(x - \bar{\zeta}_k) &= x^2 - 2x\operatorname{Re}(\zeta_k) + |\zeta_k|^2 \\ &\geq x^2 - 2x\operatorname{Re}(\zeta_k) + (\operatorname{Re}\zeta_k)^2 \\ &= (x - \operatorname{Re}\zeta_k)^2 \\ &\geq 0. \end{aligned}$$

Somit liefert (1.2)-(1.3):

$$p(x) > 0 \quad \text{und} \quad p'(x) > 0 \quad \text{für } x > \lambda_1,$$

denn laut Voraussetzung gilt $\operatorname{Re}\zeta \leq \lambda_1$ für alle Nullstellen ζ. Hieraus ergibt sich

$$x - \frac{p(x)}{p'(x)} < x \quad \text{für} \quad x > \lambda_1.$$

Andererseits gilt auch für $x > \lambda_1$:

$$\left(\sum_{k=1}^{l} \frac{1}{x-\lambda_k} + 2 \sum_{k=1}^{m} \frac{x - \operatorname{Re} \zeta_k}{(x-\zeta_k)(x-\bar{\zeta}_k)} \right) > \frac{1}{x-\lambda_1},$$

also folgt mit (1.3)

$$\frac{p'(x)}{p(x)} = \left(\sum_{k=1}^{l} \frac{1}{x-\lambda_k} + 2 \sum_{k=1}^{m} \frac{x - \operatorname{Re} \zeta_k}{(x-\zeta_k)(x-\bar{\zeta}_k)} \right) \quad \Leftrightarrow \quad \frac{p(x)}{p'(x)} < x - \lambda_1$$

und somit

$$x - \frac{p(x)}{p'(x)} > x - (x - \lambda_1) = \lambda_1.$$

Insgesamt gilt

$$\lambda_1 < x - \frac{p(x)}{p'(x)} < x \quad \text{für} \quad x > \lambda_1. \tag{1.4}$$

Sei nun $x^0 > \lambda_1$ beliebig aber fest gewählt. Dann erzeugt das Newton-Verfahren

$$x^{k+1} = x^k - \frac{p(x^k)}{p'(x^k)}$$

eine beschränkte und streng monoton fallende Folge $\{x^k\}_{k \in \mathbb{N}}$. Somit ist diese Folge konvergent. Für den Grenzwert $\bar{x} \in \mathbb{R}$ gilt zum einen:

$$\bar{x} \geq \lambda_1 \quad \text{wegen (1.4)}.$$

Nun zeigen wir, dass $\bar{x}$ eine Nullstelle ist. Dazu gilt

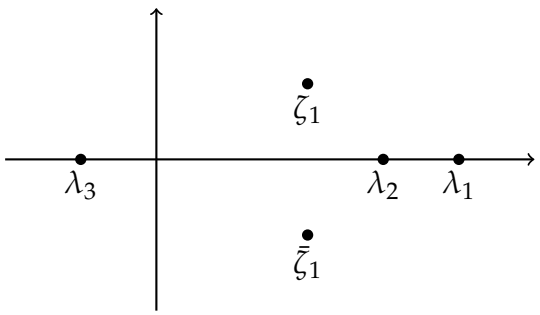
$$p'(x^k)(x^k - x^{k+1}) = p(x^k) \quad \text{(Newton-Iteration)}.$$

Daraus folgt mit der Stetigkeit von p und p':

$$p(\bar{x}) = \lim_{k \to \infty} p(x^k) = \lim_{k \to \infty} p'(x^k)(x^k - x^{k+1}) = p'(\bar{x}) \cdot 0 = 0.$$

Also ist $\bar{x}$ eine Nullstelle. Nach Voraussetzung gilt nun $\bar{x} \leq \lambda_1$ und insgesamt ergibt sich $\bar{x} = \lambda_1$. □

Die Anwendbarkeit des Satzes verdeutlicht die folgende Situation für ein Polynom vom Grad 5:



Die Nullstelle λ_1 berechnet sich wie im Satz, falls $x^0 > \lambda_1$. Für λ_2 verwendet man das deflationierte Polynom

$$\tilde{p}(x) := \frac{p(x)}{x - \lambda_1}.$$

Daraus lässt sich λ_2 berechnen. Für λ_3 kann man den Satz nicht anwenden.

Bemerkung 1.7. Die Nullstelle λ_1 ist a priori nicht bekannt. Daher ist es unklar, wie man x^0 sinnvoll wählen sollte. Dazu hilft die folgende Aussage:
Das reelle Polynom $p \in \Pi_n$

$$p(x) = a_0 + a_1 x + \ldots + a_n x^n \quad \text{mit} \quad a_n \neq 0$$

habe Nullstelle $\zeta \notin \mathbb{C}$. Dann gilt

$$|\zeta| \leq \max \left\{ 1, \sum_{k=0}^{n-1} \left| \frac{a_k}{a_n} \right| \right\}.$$

1.3 Nichtlineares Ausgleichsproblem

Bei nichtlinearen Ausgleichsproblemen wendet man in der Regel nicht das Newton-Verfahren an, sondern das einfache Gauß-Newton-Verfahren.

Definition 1.8. Es sei $F : \mathbb{R}^n \to \mathbb{R}^m$ eine Funktion, die hinreichend oft stetig differenzierbar ist mit $m \geq n$. Als *nichtlineares Ausgleichsproblem* bezeichnet man die folgende Aufgabe:

$$\min_{x \in \mathbb{R}^n} \|F(x)\|_2^2. \tag{P}$$

Bemerkung 1.9.

(i) Ist $\bar{x} \in \mathbb{R}^n$ eine Lösung von (P), so muss $F(\bar{x})$ nicht notwendigerweise verschwinden.

(ii) Ist $F : \mathbb{R}^n \supset D \to \mathbb{R}^m$ nur auf dem Definitionsbereich D definiert, so heißt die Aufgabe

$$\min_{x \in D} \|F(x)\|_2^2$$

nichtlineares restringiertes Ausgleichsproblem. In dieser Vorlesung betrachten wir nur den Fall $D = \mathbb{R}^n$.

Beispiel 1.10. Die gedämpfte Schwingung wird durch die Differentialgleichung

$$u''(t) + \frac{b}{m} u'(t) + \frac{D}{m} u(t) = 0$$

beschrieben. Hieraus erhält man eine Lösung der Gestalt

$$u(t) = u_0 e^{-\delta t} \sin(\omega t + \varphi_0),$$

wobei

$u(t)$	:	Auslenkung zum Zeitpunkt t,
u_0	:	Anfangswert,
ω	:	Frequenz,
δ	:	Abklingskonstante,
φ_0	:	Anfangsphase der Schwingung.

Liegen die Werte u_0, φ_0, ω, δ nicht vor, so lassen sich diese mittels einem nichtlinearen Ausgleichsproblem approximieren. Angenommen sind

$$(t_i, u_i), \quad i = 1, \ldots, m$$

m-verschiedene Messwerte der Auslenkung u_i zum Zeitpunkt t_i. Wir definieren die Fehlerfunktion $F : \mathbb{R}^4 \to \mathbb{R}^m$ durch

$$F(u_0, \delta, \omega, \varphi_0) := \begin{pmatrix} u_0 e^{\delta t_1} \sin(\omega t_1 + \varphi_0) - u_1 \\ \vdots \\ u_0 e^{\delta t_m} \sin(\omega t_m + \varphi_0) - u_m \end{pmatrix}.$$

Die Lösung des nichtlinearen Ausgleichproblems

$$\min_{(u_0,\delta,\omega,\varphi_0)\in\mathbb{R}^4} \|F((u_0, \delta, \omega, \varphi_0))\|_2^2$$

liefert eine Approximation für die tatsächlichen physikalischen Werte u_0, δ, ω, φ_0.

Idee zur Lösung von nichtlinearen Ausgleichsproblemen:
Es sei $F : \mathbb{R}^n \to \mathbb{R}^m$ zweimal stetig differenzierbar. Wir setzen

$$g : \mathbb{R}^n \to \mathbb{R}, \quad g(x) := \|F(x)\|_2^2 = (F(x), F(x))_{\mathbb{R}^m}.$$

Somit ist das nichtlineare Ausgleichsproblem (P) äquivalent zu

$$\min_{x\in\mathbb{R}^n} g(x).$$

Gesucht ist ein (lokales) Minimum von g. Die notwendige Optimalitätsbedingung lautet

$$\nabla g(x) = 0 \quad \Leftrightarrow \quad F'(x)^T F(x) = 0.$$

Idee: Anwendung des Newton-Verfahrens auf die Gleichung

$$F'(x)^T F(x) = 0.$$

Dazu setzen wir

$$G : \mathbb{R}^n \to \mathbb{R}^n, \quad G(x) := F'(x)^T F(x).$$

Die Newton-Vorschrift lautet

$$\begin{cases} G'(x^k)d^k = -G(x^k) \\ x^{k+1} = x^k + d^k \end{cases}$$

mit der Jacobi-Matrix $G'(x^k) \in \mathbb{R}^{n\times n}$ mit

$$G'(x^k) = F'(x^k)^T F'(x^k) + \sum_{i=1}^{m} F_i(x^k)F_i''(x^k),$$

wobei $F_i''(x^k) \in \mathbb{R}^{n\times n}$ die Hesse-Matrix der Funktion $F_i : \mathbb{R}^n \to \mathbb{R}$ bezeichnet.

1.3.1 Gauß-Newton-Verfahren

Beim Gauß-Newton-Verfahren vernachlässigen wir den zweiten Term in der Ableitung von G. An Stelle von

$$G'(x^k)d^k = -G(x^k)$$

lösen wir die einfachere Aufgabe

$$\begin{cases} F'(x^k)^T F'(x^k)d^k = -F'(x^k)^T F(x^k), \\ x^{k+1} = x^k + d^k. \end{cases}$$

Bemerkung 1.11. Die numerische Realisierung von $\sum_{i=1}^{m} F_i(x^k)F_i''(x^k)$ ist sehr aufwendig. Dieser Term ist oft klein und deshalb ist das Gauß-Newton-Verfahren sinnvoll.

Algorithmus 1.2 Gauß-Newton-Verfahren

(S0) Wähle einen Startwert $x^0 \in \mathbb{R}^n$ und setze $k = 0$.

(S1) Abbruchkriterium:

1: **if** $k > 0$ und $\left\|d^k\right\|_2 = 0$ **then**
2: STOP
3: **end if**

(S2) Bestimme $d^k \in \mathbb{R}^n$ als Lösung von

$$F'(x^k)^T F'(x^k)d^k = -F'(x^k)^T F(x^k).$$

(S3) Setze $x^{k+1} = x^k + d^k$ und $k = k + 1$ und gehe zu (S1).

Das folgende Lemma beantwortet die Frage, ob Algorithmus 1.2 immer durchführbar ist.

Lemma 1.12. *Es sei $F : \mathbb{R}^n \to \mathbb{R}^m$ differenzierbar. Dann ist das Gauß-Newton-Verfahren durchführbar. Mit anderen Worten besitzt*

$$F'(x^k)^T F'(x^k)d^k = -F'(x^k)^T F(x^k)$$

für jedes $k = 0, 1, 2, \ldots$ eine Lösung $d^k \in \mathbb{R}^n$.

Beweis. Sei $k \in \mathbb{N} \cup \{0\}$. Wir setzen

$$A := F'(x^k) \in \mathbb{R}^{m \times n}$$
$$b := -F(x^k) \in \mathbb{R}^m.$$

Betrachte nun das lineare Ausgleichsproblem

$$\min_{d^k \in \mathbb{R}^n} \|Ad^k - b\|_2.$$

Aus Numerik I wissen wir, dass die obige Aufgabe stets eine Lösung $d^k \in \mathbb{R}^n$ besitzt. Die notwendige und hinreichende Bedingung lautet

$$A^T A d^k = A^T b,$$

also in unserem Fall

$$F'(x^k)^T F'(x^k) d^k = -F'(x^k)^T F(x^k).$$

□

Folgerung 1.13. Das Gauß-Newton-Verfahren ist äquivalent zu einer Folge von linearen Ausgleichsproblemen

$$\begin{cases} \min\limits_{d^k \in \mathbb{R}^n} \left\|F(x^k) + F'(x^k) d^k\right\|_2^2 \\ x^{k+1} = x^k + d^k \end{cases}.$$

Bemerkung 1.14. Die Matrix $F'(x^k)^T F'(x^k) \in \mathbb{R}^{n \times n}$ ist im Allgemeinen nur symmetrisch und positiv semidefinit, so dass die Lösung $d^k \in \mathbb{R}^n$ von

$$F'(x^k)^T F'(x^k) d^k = -F'(x^k)^T F(x^k)$$

im Allgemeinen nicht eindeutig ist.

1.3.2 Levenberg-Marquardt-Verfahren

Wir haben gesehen, dass das Gauß-Newton-Verfahren die lineare Approximation

$$\|F(x)\|_2^2 \approx \|F(x^k) + F'(x^k)(x - x^k)\|_2^2$$

verwendet. Diese Approximation ist nur für x nahe bei x^k bzw. für kleine Zuwächse d^k gut. Daher sollte die Lösung $d^k \in \mathbb{R}^n$ von

$$\min_{d^k \in \mathbb{R}^n} \|F(x^k) + F'(x^k) d^k\|_2^2$$

klein bleiben, das heißt

$$\|d^k\|_2 \leq \varepsilon.$$

Dies erreicht man durch Addition von $\mu\|d^k\|_2^2$ mit $\mu > 0$ „klein“ im Gauß-Newton-Verfahren:

$$\min_{d^k \in \mathbb{R}^n} \|F(x^k) + F'(x^k)d^k\|_2^2 + \mu\|d^k\|_2^2.$$

Je größer μ wird, desto kleiner wird $\left\|d^k\right\|_2$. Man kann $\mu = 0$ wählen, falls $\left\|d^k\right\|_2 \leq \varepsilon$, ansonsten $\mu > 0$. Beim Levenberg-Marquardt-Verfahren löst man also

$$\underbrace{\left(F'(x^k)^T F'(x^k) + \mu I\right)}_{\text{s.p.d. für } \mu>0} d^k = -F'(x^k)^T F(x^k).$$

Beachte, dass die obige Gleichung genau eine Lösung $d^k \in \mathbb{R}^n$ besitzt, denn die Matrix

$$(F'(x^k)^T F'(x^k) + \mu I) \in \mathbb{R}^{n\times n}$$

ist symmetrisch und positiv definit.

Lemma 1.15. *Es sei $F : \mathbb{R}^n \to \mathbb{R}^m$ differenzierbar und $x^k \in \mathbb{R}^n$. Dann gilt:*

(i) Für jedes $\mu \in \mathbb{R}^+$ ist die Matrix $(F'(x^k)^T F'(x^k) + \mu I) \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit.

(ii) Die Abbildung

$$\Phi : (0,\infty) \to [0,\infty), \quad \Phi(\mu) = \|(F'(x^k)^T F'(x^k) + \mu I)^{-1} F'(x^k)^T F(x^k)\|_2$$

ist stetig und monoton fallend mit

$$\lim_{\mu\to\infty} \Phi(\mu) = 0.$$

(iii) Ist $F'(x^k)^T F'(x^k) \in \mathbb{R}^{n\times n}$ regulär, so gilt

$$\lim_{\mu\to 0} (F'(x^k)^T F'(x^k) + \mu I)^{-1} F'(x^k) F(x^k) = (F'(x^k)^T F'(x^k))^{-1} F'(x^k)^T F(x^k).$$

Beweis.

(i) Klar, da $F'(x^k)^T F'(x^k) \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semidefinit ist.

(ii) Die Matrix $F'(x^k)^T F'(x^k)$ ist symmetrisch und positiv semidefinit, und somit existiert eine Orthonormalbasis $\{v_j\}_{j=1}^n \subset \mathbb{R}^n$ aus Eigenvektoren v_j von $F'(x^k)^T F'(x^k)$ zu den Eigenwerten $\lambda_j \geq 0$. Folglich gilt

$$F'(x^k)^T F(x^k) = \sum_{j=1}^n c_j v_j$$

mit $c_j = v_j^T F'(x^k)^T F(x^k) \in \mathbb{R}$, da $(v_j, v_i)_{\mathbb{R}^n} = \delta_{ij}$. Also ist

$$(F'(x^k)^T F'(x^k) + \mu I)^{-1} F'(x^k)^T F(x^k) = \sum_{j=1}^n \frac{c_j}{\lambda_j + \mu} v_j,$$

da v_j Eigenvektor von $(F'(x^k)^T F'(x^k) + \mu I)^{-1}$ zum Eigenwert $\frac{1}{\lambda_j + \mu}$ ist. Daher ist

$$\begin{aligned}\left\| (F'(x^k)^T F'(x^k) + \mu I)^{-1} F'(x^k)^T F(x^k) \right\|_2^2 &= \left\| \sum_{j=1}^{n} \frac{c_j}{\lambda_j + \mu} v_j \right\|_2^2 \\ &= \left(\sum_{j=1}^{n} \frac{c_j}{\lambda_j + \mu} v_j, \sum_{k=1}^{n} \frac{c_k}{\lambda_k + \mu} v_k \right)_{\mathbb{R}^n} \\ &= \sum_{j=1}^{n} \left(\frac{c_j}{\lambda_j + \mu} \right)^2.\end{aligned}$$

Nun ist also

$$\Phi(\mu) = \sqrt{\sum_{j=1}^{n} \left(\frac{c_j}{\lambda_j + \mu} \right)^2}$$

und somit ist $\Phi : (0, \infty) \to [0, \infty)$ stetig und monoton fallend. Mit $\mu \to \infty$ ist

$$\lim_{\mu \to \infty} \Phi(\mu) = 0.$$

(iii) Übungsaufgabe.

□

1.4 Konvergenzanalyse des Gauß-Newton-Verfahrens

In Numerik I haben wir das folgende Resultat gezeigt:

Lemma 1.16. *Es seien $A, B \in \mathbb{R}^{n \times n}$ mit $\|I - BA\|_2 < 1$. Dann sind A und B regulär und es gilt*

$$\|A^{-1}\|_2 \leq \frac{\|B\|_2}{1 - \|I - BA\|_2}.$$

Korollar 1.17. *Es sei $G : \mathbb{R}^n \to \mathbb{R}^{n \times n}$ stetig. Ferner sei $\bar{x} \in \mathbb{R}^n$, so dass $G(\bar{x}) \in \mathbb{R}^{n \times n}$ invertierbar ist. Dann existiert ein $\delta > 0$, so dass die Matrizen*

$$G(x) \in \mathbb{R}^{n \times n} \quad \forall x \in B_\delta(\bar{x})$$

invertierbar sind mit

$$\|G(x)^{-1}\|_2 \leq 2 \|G(\bar{x})^{-1}\|_2 \quad \forall x \in B_\delta(\bar{x}).$$

Beweis. Da die Abbildung $G : \mathbb{R}^n \to \mathbb{R}^{n \times n}$ stetig ist, existiert ein $\delta > 0$, so dass

$$\|G(x) - G(\bar{x})\|_2 \leq \frac{1}{2 \|G(\bar{x})^{-1}\|_2} \quad \forall x \in B_\delta(\bar{x}).$$

Folglich gilt

$$\begin{aligned}\|I - G(\bar{x})^{-1}G(x)\|_2 &\leq \|G(\bar{x})^{-1}(G(\bar{x}) - G(x))\|_2 \\ &\leq \|G(\bar{x})^{-1}\|_2 \|G(\bar{x}) - G(x)\|_2 \\ &\leq \frac{1}{2} \quad \forall x \in B_\delta(\bar{x}).\end{aligned}$$

Das obige Lemma mit

$$B := G(\bar{x})^{-1} \quad \text{und} \quad A := G(x)$$

liefert, dass $G(x)$ für alle $x \in B_\delta(\bar{x})$ invertierbar sind mit

$$\|G(x)^{-1}\|_2 \leq \frac{\left\|G(\bar{x})^{-1}\right\|_2}{1 - \|I - G(\bar{x})^{-1}G(x)\|_2} \leq 2\|G(\bar{x})^{-1}\|_2.$$

□

Die folgende Aussage haben wir auch bereits in Numerik I gezeigt.

Lemma 1.18. *Es sei $F : \mathbb{R}^n \to \mathbb{R}^m$ differenzierbar. Dann besitzt das Restglied*

$$r(x, y) := F(y) - F(x) - F'(x)(y - x) \quad \forall x, y \in \mathbb{R}^n$$

die folgende Integraldarstellung:

$$r(x, y) = \int_0^1 F'(x + t(y - x))(y - x) - F'(x)(y - x)\,\mathrm{d}t \quad \forall x, y \in \mathbb{R}^n.$$

Satz 1.19 (Lokale Konvergenz des Gauß-Newton-Verfahrens im Falle $F(\bar{x}) = 0$)**.** *Es sei $F : \mathbb{R}^n \to \mathbb{R}^m$ stetig differenzierbar und $\bar{x} \in \mathbb{R}^n$ eine Nullstelle von F. Ferner sei*

$$F'(\bar{x})^T F'(\bar{x}) \in \mathbb{R}^{n \times n}$$

invertierbar. Dann existiert ein $\varepsilon > 0$, so dass das Gauß-Newton-Verfahren für jeden Startwert $x^0 \in B_\varepsilon(\bar{x})$ superlinear *gegen $\bar{x}$ konvergiert.*

Beweis. Da $F'(\bar{x})^T F'(\bar{x}) \in \mathbb{R}^{n \times n}$ invertierbar ist und die Abbildung $x \mapsto F'(x)^T F'(x)$ stetig ist, existiert laut Korollar ein $\varepsilon_1 > 0$, so dass die Matrizen $F'(x)^T F'(x)$ für alle $x \in B_{\varepsilon_1}(\bar{x})$ invertierbar sind mit

$$\|(F'(x)^T F'(x))^{-1}\|_2 \leq 2\|(F'(\bar{x})^T F'(\bar{x}))^{-1}\|_2 \quad \forall x \in B_{\varepsilon_1}(\bar{x}). \tag{1.5}$$

Die Stetigkeit der Abbildung $x \mapsto F'(x)$ liefert die Existenz einer positiven Zahl $M > 0$, so dass gilt

$$\|F'(x)^T\|_2 \leq M \quad \forall x \in B_{\varepsilon_1}(\bar{x}). \tag{1.6}$$

Sei nun $\alpha \in (0, 1)$ beliebig aber fest. Da $F : \mathbb{R}^n \to \mathbb{R}^m$ stetig differenzierbar ist, existiert ein $\varepsilon_2 > 0$ mit

$$\int_0^1 \left\|F'(x + t(\bar{x} - x)) - F'(x)\right\|_2 \,\mathrm{d}t \leq \frac{\alpha}{2\left\|(F'(\bar{x})^T F'(\bar{x}))^{-1}\right\|_2 M} \quad \forall x \in B_{\varepsilon_2}(\bar{x}).$$

Aus dem vorherigen Lemma folgt nun

$$\begin{aligned}\|r(x,\bar{x})\|_2 &\leq \int_0^1 \left\|F'(x+t(\bar{x}-x)) - F'(x)\right\|_2 \,\mathrm{d}t\, \|\bar{x}-x\|_2 \\ &\leq \frac{\alpha}{2\,\|(F'(\bar{x})^T F'(\bar{x}))^{-1}\|_2\, M}\, \|\bar{x}-x\|_2 \quad \forall x \in B_{\varepsilon_2}(\bar{x}). \end{aligned} \tag{1.7}$$

Wir setzen $\varepsilon := \min\{\varepsilon_1, \varepsilon_2\}$. Sei $x^k \in B_\varepsilon(\bar{x})$ und $x^{k+1} \in \mathbb{R}^n$ die durch das Gauß-Newton-Verfahren erzeugte neue Iterierte. Dann gilt

$$\begin{aligned} x^{k+1} - \bar{x} &= (F'(x^k)^T F'(x^k))^{-1} (F'(x^k)^T F'(x^k)) (\underbrace{x^{k+1} - x^k}_{=d^k} + x^k - \bar{x}) \\ &= (F'(x^k)^T F'(x^k))^{-1} ((-F'(x^k)^T F(x^k)) \\ &\quad + (F'(x^k)^T F'(x^k))(x^k - \bar{x}) + F'(x^k) \underbrace{F(\bar{x})}_{=0}) \\ &= (F'(x^k)^T F'(x^k))^{-1} F'(x^k)^T (\underbrace{F(\bar{x}) - F(x^k) - F'(x^k)(\bar{x} - x^k)}_{=r(x^k,\bar{x})}). \end{aligned}$$

Daher ist mit (1.5)-(1.7)

$$\begin{aligned} \|x^{k+1} - \bar{x}\|_2 &\leq \|(F'(x^k)^T F'(x^k))^{-1}\|_2 \|F'(x^k)^T\|_2 \|r(x^k,\bar{x})\|_2 \|x^k - \bar{x}\|_2 \\ &\leq \alpha \|x^k - \bar{x}\|_2 \end{aligned}$$

mit $\alpha \in (0,1)$. Ist $x^0 \in B_\varepsilon(\bar{x})$, so folgt aus der obigen Ungleichung

$$x^k \in B_\varepsilon(\bar{x}) \quad \forall k \in \mathbb{N} \quad (\text{denn } \|x^k - \bar{x}\|_2 \leq \alpha^k \|x^0 - \bar{x}\|_2 < \varepsilon)$$

und

$$\|x^{k+1} - \bar{x}\|_2 \leq \alpha \|x^k - \bar{x}\|_2 \quad \forall k \in \mathbb{N}.$$

Somit konvergiert die Folge $\{x^k\}$ gegen $\bar{x}$ linear mit Konvergenzrate $\alpha \in (0,1)$, falls $x^0 \in B_\varepsilon(\bar{x})$. Die superlineare Konvergenz erfolgt wie beim Newton-Verfahren, das heißt

$$\lim_{k\to\infty} \frac{\left\|x^{k+1} - \bar{x}\right\|_2}{\left\|x^k - \bar{x}\right\|_2} = 0.$$

□

In der Praxis kann man leider nicht mehr erwarten, dass eine Lösung $\bar{x} \in \mathbb{R}^n$ von dem Ausgleichsproblem

$$\min_{x\in\mathbb{R}^n} \|F(x)\|_2^2$$

der Bedingung $F(\bar{x}) = 0$ genügt. Ist $\|F(\bar{x})\|_2$ hinreichend klein und ist die Abbildung

$$(F')^T : \mathbb{R}^n \to \mathbb{R}^{n\times m}$$

zusätzlich lokal Lipschitz-stetig in $\bar{x}$, so können wir zeigen, dass das Gauß-Newton-Verfahren lokal gegen $\bar{x}$ konvergiert.

Satz 1.20 (Lokale Konvergenz des Gauß-Newton-Verfahrens im Fall $F(\bar{x}) \neq 0$)**.** *Es sei $F : \mathbb{R}^n \to \mathbb{R}^m$ stetig differenzierbar und $\bar{x} \in \mathbb{R}^n$, so dass*

$$F'(\bar{x})^T F(\bar{x}) = 0 \quad \textit{und} \quad F'(\bar{x})^T F'(\bar{x}) \in \mathbb{R}^{n\times n}$$

invertierbar ist. Es existieren $\delta, L > 0$, so dass gilt

$$\|F'(x)^T - F'(\bar{x})^T\|_2 \leq L \|x - \bar{x}\|_2 \quad \forall x \in B_\delta(\bar{x}). \quad (\textit{lokale Lipschitzsannahme}) \tag{A1}$$

Es gebe eine weitere Konstante $\beta \in (0,1)$, so dass gilt

$$\|F(\bar{x})\|_2 \leq \frac{\beta}{2L \,\|(F'(\bar{x})^T F'(\bar{x}))^{-1}\|_2}. \tag{A2}$$

Dann existiert ein $\varepsilon > 0$, so dass das Gauß-Newton-Verfahren für jeden Startwert $x^0 \in B_\varepsilon(\bar{x})$ linear *gegen $\bar{x}$ konvergiert.*

Bemerkung 1.21. Anders als beim vorherigen Satz benötigen wir nun zusätzlich die lokale Lipschitzannahme (A1) und die Bedingung (A2). Beachte auch, dass der Satz nur die lokale lineare Konvergenz sichert.

Beweis. Es sei $\alpha \in (0,1)$ mit $\gamma := \alpha + \beta \in (0,1)$. Aus Stetigkeitsgründen existiert ein $\varepsilon > 0$, so dass gilt

$$\begin{cases} \left\|(F'(x)^T F'(x))^{-1}\right\|_2 \leq 2\left\|(F'(\bar{x})^T F'(\bar{x}))^{-1}\right\|_2 & \forall x \in B_\varepsilon(\bar{x}), \\ \left\|F'(x)^T\right\|_2 \leq M & \forall x \in B_\varepsilon(\bar{x}), \\ \|r(x,\bar{x})\|_2 \leq \frac{\alpha}{2\left\|(F'(\bar{x})^T F'(\bar{x}))^{-1}\right\|_2 M} \|x - \bar{x}\|_2 & \forall x \in B_\varepsilon(\bar{x}). \end{cases} \tag{1.8}$$

Wir setzen $\varepsilon := \min\{\varepsilon, \delta\}$. Sei $x^k \in B_\varepsilon(\bar{x})$ und $x^{k+1} \in \mathbb{R}^n$ die durch das Gauß-Newton-Verfahren erzeugte neue Iterierte. Dann gilt

$$\begin{aligned} x^{k+1} - \bar{x} &= (F'(x^k)^T F'(x^k))^{-1}(F'(x^k)^T F'(x^k))(\underbrace{x^{k+1} - x^k}_{=d^k} + x^k - \bar{x}) \\ &= (F'(x^k)^T F'(x^k))^{-1}(-F'(x^k)^T F(x^k) + F'(x^k)^T F'(x^k)(x^k - \bar{x}) \\ &\qquad + F'(\bar{x})^T F(\bar{x}) - F'(x^k)^T F(\bar{x}) + F'(x^k)^T F(\bar{x})) \\ &= (F'(x^k)^T F'(x^k))^{-1} F'(x^k)^T (F(\bar{x}) - F(x^k) + F'(x^k)(x^k - \bar{x})) \\ &\quad + (F'(x^k)^T F'(x^k))^{-1}(F'(\bar{x})^T - F'(x^k)^T) F(\bar{x}). \end{aligned}$$

Daraus ergibt sich

$$\begin{aligned} \|x^{k+1} - \bar{x}\|_2 &\leq \|(F'(x^k)^T F'(x^k))^{-1}\|_2 \|F'(x^k)^T\|_2 \|r(x^k,\bar{x})\|_2 \\ &\quad + \|(F'(x^k)^T F'(x^k))^{-1}\|_2 \|F'(\bar{x})^T - F'(x^k)^T\|_2 \|F(\bar{x})\|_2 \\ &\overset{(1.8)}{\leq} \alpha \|x^k - \bar{x}\|_2 + 2\|(F'(\bar{x})^T F'(\bar{x}))^{-1}\|_2 \\ &\quad \cdot L\|x^k - \bar{x}\|_2 \beta (2L\|(F'(\bar{x})^T F'(\bar{x}))^{-1}\|_2)^{-1} \\ &= (\alpha + \beta)\|x^k - \bar{x}\|_2. \end{aligned}$$

Ist $x^0 \in B_\varepsilon(\bar{x})$, so folgt aus der obigen Abschätzung

$$\begin{cases} x^k \in B_\varepsilon(\bar{x}) & \forall k \in \mathbb{N}, \\ \left\|x^{k+1} - \bar{x}\right\| \leq \gamma \left\|x^k - \bar{x}\right\| & \forall k \in \mathbb{N}. \end{cases}$$

Somit haben wir gezeigt, dass das Gauß-Newton-Verfahren gegen $\bar{x}$ linear mit Konvergenzrate $\gamma \in (0,1)$ konvergiert, wenn der Startwert klein genug ist. □

1.5 Parameterabhängige nichtlineare Gleichungssysteme

Wir untersuchen ein nichtlineares Problem der Gestalt

$$F(x, \lambda) = 0$$

mit einer Funktion $F : \mathbb{R}^n \times \mathbb{R}^p \supset D \to \mathbb{R}^n$, die von einem Parametervektor $\lambda \in \mathbb{R}^p$ abhängt. Beispielsweise kann folgende Situation vorliegen:
Eigentlich ist man interessiert an der Lösung von $G(x) = 0$, aber dieses System lässt sich schwer lösen. Man bettet dieses Problem ein in

$$F(x, \lambda) = 0$$

mit $\lambda \in [0,1]$ mit $F(x,0) = G(x)$ und $F(x,1) = 0$ leicht lösbar.

1.5.1 Beispiel: Innere-Punkte-Methode für lineare Optimierung

Es seien $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ und $c \in \mathbb{R}^n$ gegeben. Gesucht ist eine Lösung von

$$\begin{cases} \min & (c\,,x)_{\mathbb{R}^n} \\ \text{bei} & Ax = b, \\ & x_i \geq 0 \quad \forall i = 1, \ldots, n. \end{cases} \tag{PP}$$

Wir definieren die Lagrange-Funktion wie folgt:

$$\mathscr{L} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}, \quad \mathscr{L}(x, \mu) := (c\,,x)_{\mathbb{R}^n} + (b - Ax\,, \mu)_{\mathbb{R}^m}.$$

Hierbei heißt $\mu \in \mathbb{R}^n$ Lagrangescher Multiplikator.
Es gilt

$$\text{(PP)} \quad \Leftrightarrow \quad \min_{x \geq 0} (\max_{\mu \in \mathbb{R}^n} \mathscr{L}(x, \mu)),$$

denn

$$\max_{\mu \in \mathbb{R}^n} \mathscr{L}(x, \mu) = +\infty, \quad \text{falls } Ax \neq b.$$

Durch Vertauschung von min und max kommt man auf das sogenannte Dualproblem

$$\begin{aligned} &\max_{\mu\in\mathbb{R}^n}(\min_{x\geq 0}\mathscr{L}(x,\mu)) \\ &= \max_{\mu\in\mathbb{R}^n}(\min_{x\geq 0}(c\,,x)_{\mathbb{R}^n}+(b-Ax\,,\mu)_{\mathbb{R}^m}) \\ &= \max_{\mu\in\mathbb{R}^n}(\min_{x\geq 0}(c-A^T\mu\,,x)_{\mathbb{R}^n}+(b\,,\mu)_{\mathbb{R}^m}). \end{aligned} \tag{DP}$$

Somit ist (DP) äquivalent zu

$$\begin{cases} \max\limits_{\mu\in\mathbb{R}^n} & (b\,,\mu)_{\mathbb{R}^m} \\ \text{bei} & A^T\mu\leq c. \end{cases}$$

Durch Einführung eines Schlupfvektors $s\geq 0$ ergibt sich

$$\begin{cases} \max\limits_{\mu\in\mathbb{R}^n} & (b\,,\mu)_{\mathbb{R}^m} \\ \text{bei} & A^T\mu+s=c, \\ & s\geq 0. \end{cases} \tag{DP}$$

Aus der Optimierungsvorlesung ist bekannt, dass $x^*\in\mathbb{R}^n$ genau dann das Primalproblem (PP) löst, wenn zusammen mit einer Lösung des Dualproblems (DP), $(\mu^*,s^*)\in\mathbb{R}^m\times\mathbb{R}^n$, die sogenannten komplementären Schlupfbedingungen

$$\begin{cases} x^*\geq 0, \\ s^*\geq 0, \\ (x^*\,,s^*)_{\mathbb{R}^n}=0 \end{cases} \quad\Leftrightarrow\quad x_i^*s_i^*=0\quad\forall i=1,\dots,n$$

erfüllt sind. Insgesamt ergibt sich das folgende notwendige und hinreichende Optimalitätssystem

$$x^*\in\mathbb{R}^n \text{ löst (PP)} \quad\Leftrightarrow\quad \begin{cases} Ax^*=b, \\ A^T\mu^*+s^*=c, \\ x_i^*s_i^*=0\quad\forall i=1,\dots,n, \\ x^*\geq 0,\quad s^*\geq 0. \end{cases}$$

Dieses System bezeichnen wir mit (KKT), von Karush-Kuhn-Tucker.

Innere-Punkte-Methode

Wir führen $\lambda\in\mathbb{R}$ als Parameter ein, und weichen (KKT) etwas auf.

$$(IPM_\lambda)=\begin{cases} Ax^*=b, \\ A^T\mu^*+s^*=c, \\ x_i^*s_i^*=\lambda\quad\forall i=1,\dots,n, \\ x^*>0,\quad s^*>0. \end{cases}$$

Wir starten mit $\lambda \gg 0$ und lösen das System (IPM_λ). Dann verkleinern wir λ schrittweise, um den Wert $\lambda = 0$ zu approximieren. Das ist ein Beispiel für ein parameterabhängiges nichtlineares Gleichungssystem. Wir setzen also $z = (x, s, \mu)$ und

$$F : \mathbb{R}^{2n+m} \times \mathbb{R} \supset D \to \mathbb{R}^{2n+m}, \quad F(z, \lambda) := \begin{pmatrix} Ax - b \\ A^T\mu + s - c \\ XSe - \lambda e \end{pmatrix}$$

mit $X = \operatorname{diag}(x_1, \ldots, x_n)$, $S = \operatorname{diag}(s_1, \ldots, s_n)$, und $e = (1, \ldots, 1)^T$. Der Definitionsbereich ist

$$D = \mathbb{R}^n_+ \times \mathbb{R}^n_+ \times \mathbb{R}^m \times \mathbb{R}_+.$$

Es gilt also

$$(IPM_\lambda) \quad \Leftrightarrow \quad F(z, \lambda) = 0.$$

1.5.2 Fortsetzungsmethode

Im Folgenden sei $D \subset \mathbb{R}^n$ offen, $\lambda \in [a, b]$ mit $a \leq b$, und $F : D \times [a, b] \to \mathbb{R}^n$ sei stetig differenzierbar. Wir betrachten das parameterabhängige nichtlineare Gleichungssystem

$$F(x, \lambda) = 0.$$

Definition 1.22. Die Menge der Lösungen der Gleichung $F(x, \lambda) = 0$

$$S := \{(x, \lambda) \in \mathbb{R}^n \times [a, b] \mid F(x, \lambda) = 0\}$$

heißt **Lösungsstruktur**.

Beispiel 1.23. Es sei $D = \mathbb{R}$ und $[a, b] = [-1, 1]$. Ferner sei

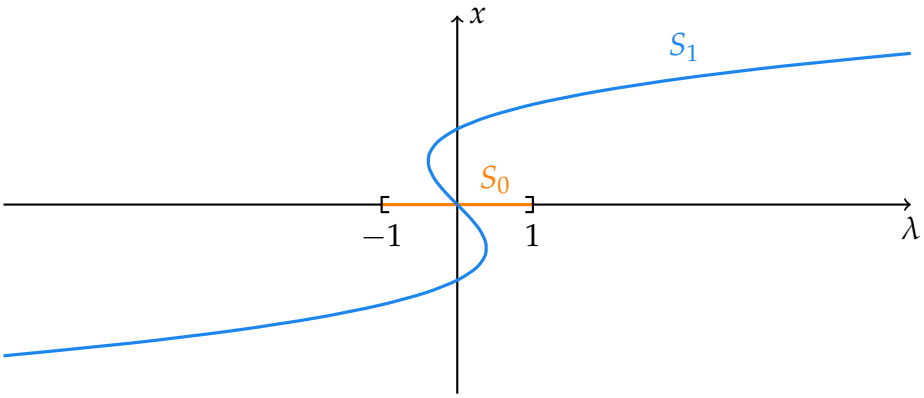
$$F(x, \lambda) = x(x^3 - x - \lambda).$$

Es gilt

$$F(x, \lambda) = 0 \quad \Leftrightarrow \quad x = 0 \text{ oder } \lambda = x^3 - x \in [-1, 1].$$

Somit ergeben sich zwei unterschiedliche Lösungsstrukturen:

$$S_0 := \{(x, \lambda) \mid x = 0 \text{ und } \lambda \in [-1, 1]\}$$
$$S_1 := \{(x, \lambda) \mid x \in \mathbb{R} \text{ und } \lambda = x^3 - x \in [-1, 1]\}.$$



Beachte, dass sich die Lösungskurven S_0 und S_1 im Nullpunkt schneiden. Der Nullpunkt ist somit ein Verzweigungspunkt. Im Nullpunkt gilt

$$\frac{\partial F}{\partial x}(0,0) = x^3 - x - \lambda + x(3x^2 - 1)\Big|_{(0,0)} = 0.$$

Daher ist der Satz über implizite Funktionen im Nullpunkt nicht anwendbar (sonst gäbe es dort keine Verzweigung).

Im Folgenden nehmen wir an, dass

$$F_x(x,\lambda) \in \mathbb{R}^{n\times n}$$

für alle $(x,\lambda) \in D \times [a,b]$ invertierbar ist. Somit kann man den Satz über implizite Funktionen anwenden. Unser paramterabhängiges nichtlineares Gleichungssystem lautet

$$F(x,\lambda) = 0.$$

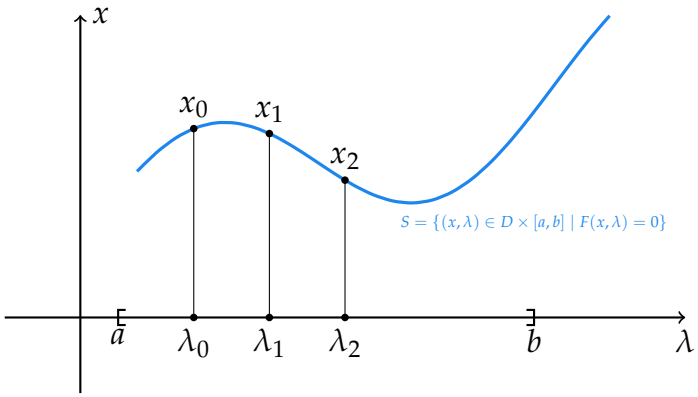
Dabei ist das nichtlineare Problem

$$F(x,a) = 0$$

einfach zu lösen und

$$F(x,b) = 0$$

ist die schwierigere Aufgabe. Es geht also darum, ausgehend von einem gewissen Wert für λ sukzessive weitere Punkte der Kurve $F(x,\lambda) = 0$ zu berechnen.



Idee:

(i) Lösung für $\lambda = \lambda_0$:

Wir nehmen an, dass für einen Startwert $\lambda = \lambda_0$ das Newton-Verfahren

$$\begin{cases} F_x(x^k,\lambda_0)d^k = -F(x^k,\lambda_0) \\ x^{k+1} = x^k + d^k \end{cases}$$

zu einer Lösung $\bar{x} = x_0$ der Aufgabe $F(x_0,\lambda_0) = 0$ konvergiert.

(ii) Fortsetzung $\lambda_1 > \lambda_0$:

Nun soll eine Lösung von

$$F(x, \lambda_1) = 0$$

mit $\lambda_1 > \lambda_0$ bestimmt werden. Es treten hier zwei Schwierigkeiten auf:

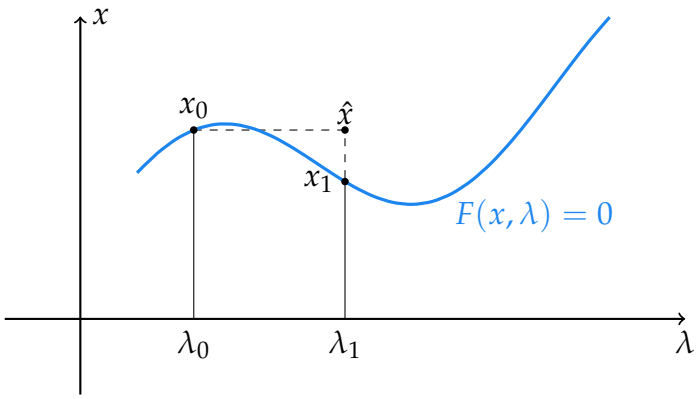
- Wie weit darf λ_1 von λ_0 entfernt sein, damit das Newton-Verfahren zur Bestimmung von x_1 konvergiert.
- Welchen Anfangswert x^0 sollen wir im Newton-Verfahren für

$$F(x, \lambda_1) = 0$$

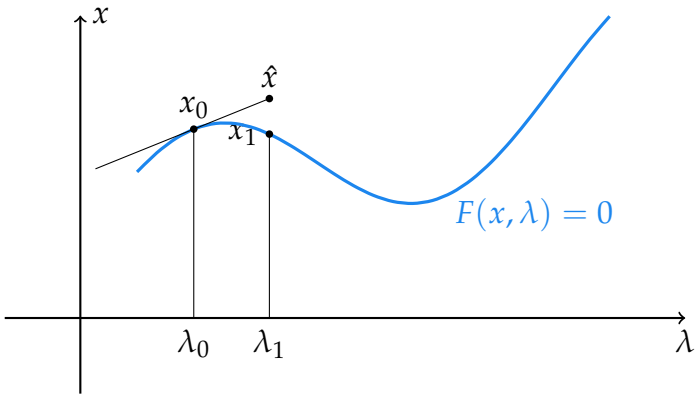
wählen?

Wir wollen nun zwei Ansätze untersuchen. Dazu nehmen wir an, dass λ_1 geschickt gewählt sei, so dass alles gut geht. Nun klären wir, wie man den Startwert für das Newton-Verfahren wählen kann.

1. Ansatz: Klassische Fortsetzungsmethode (Poincaré, 1892, Himmelsmechanik)
Man wählt $\hat{x} := x^0 = x_0$ als Anfangswert für das Newton-Verfahren zur Lösung von $F(x, \lambda_1) = 0$.



2. Ansatz: Methode der tangentialen Fortsetzung



Die Lösungskurve hat die Darstellung $(x(t), \lambda_0 + t)$ mit $x(0) = x_0$. Dann ergibt sich als Tangentialvektor im Punkt (x_0, λ_0), das heißt $t = 0$, der folgende Vektor:

$$\begin{pmatrix} x'(0) \\ 1 \end{pmatrix}.$$

Daher ergibt sich die Wahl des Startwertes $\hat{x}$ für das Newton-Verfahren wie folgt:

$$\begin{pmatrix} \hat{x} \\ \lambda_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ \lambda_0 \end{pmatrix} + \begin{pmatrix} x'(0) \\ 1 \end{pmatrix} (\lambda_1 - \lambda_0).$$

Also

$$\hat{x} = x_0 + x'(0)(\lambda_1 - \lambda_0) \quad \text{(tangentiale Fortsetzung)}.$$

Die Berechnung von $x'(0)$ erfolgt über implizites Differenzieren:

$$\begin{aligned} & F(x(t), \lambda_0 + t) = 0 \quad \forall t \geq 0 \\ \overset{|\frac{d}{dt},\ x(\cdot)\in C^1}{\Rightarrow} \quad & F_x(x(t), \lambda_0 + t)x'(t) + F_\lambda(x(t), \lambda_0 + t) = 0 \quad \forall t \geq 0. \end{aligned}$$

Im Nullpunkt $t = 0$ ergibt sich

$$F_x(\underbrace{x(0)}_{=x_0}, \lambda_0)x'(0) = -F_\lambda(\underbrace{x(0)}_{=x_0}, \lambda_0).$$

Das ist ein lineares Gleichungssystem, welches aufgrund unserer Annahme eindeutig lösbar ist.

Bemerkung 1.24. Man beachte, dass beide Ansätze aus zwei Teilschritten bestehen:

1. Wahl der Schrittweite
$$\tau := \lambda_1 - \lambda_0 > 0$$
und Wahl des Startwerts für das Newton-Verfahren

$\hat{x} := x_0$	„klassische Fortsetzungsmethode“
$\hat{x} := x_0 + \tau x'(0)$	„Methode der tangentialen Fortsetzung“.

2. Newton-Verfahren mit dem Startwert $\hat{x}$ zur Lösung von
$$F(x, \lambda_1) = 0.$$

Der erste Teilschritt heißt Prädikator-Schritt. Der zweite Teilschritt heißt Korrektor-Schritt (Newton-Verfahren bis zurück auf die Lösungskurve zum Punkt (x_1, λ_1)).
Der Gesamtprozess heißt Prädikator-Korrektor-Verfahren.
Hauptschwierigkeit: Wahl der Schrittweite τ

- τ zu klein: langsam zum Endziel $\lambda = b$.

- τ zu groß: keine Konvergenz im Newton-Verfahren.

Lemma 1.25. *Es sei $\varepsilon > 0$ und $x(\cdot) \in C^1([0,\varepsilon], \mathbb{R}^n)$. Dann gilt*

$$\|x(t) - x(0)\|_2 \leq \eta t \quad \forall t \in [0,\varepsilon]$$

mit $\eta := \max\limits_{0\leq t\leq\varepsilon} \|x'(t)\|_2$.

Beweis. Es sei $t \in [0,\varepsilon]$. Dann gilt

$$x(t) - x(0) = \int_0^t x'(\tau)\,\mathrm{d}\tau$$

und daher

$$\|x(t) - x(0)\|_2 \leq \int_0^t \underbrace{\|x'(\tau)\|_2}_{\leq \max\limits_{0\leq t\leq\varepsilon}\|x'(t)\|_2}\,\mathrm{d}\tau \leq \int_0^t \eta\,\mathrm{d}\tau = t\eta.$$

□

Lemma 1.26. *Es sei $\varepsilon > 0$ und $x(\cdot) \in C^2([0,\varepsilon], \mathbb{R}^n)$. Ferner sei*

$$\hat{x}(t) := x(0) + tx'(0) \quad \forall t \in [0,\varepsilon].$$

Dann gilt

$$\|x(t) - \hat{x}(t)\|_2 \leq \eta t^2 \quad \forall t \in [0,\varepsilon]$$

mit $\eta := \max\limits_{0\leq t\leq\varepsilon} \frac{1}{2}\|x''(t)\|_2$.

Beweis. Es sei $t \in [0,\varepsilon]$. Dann gilt

$$\begin{aligned}
x(t) - \hat{x}(t) &= x(t) - x(0) - tx'(0) \\
&= \int_0^t x'(\tau)\,\mathrm{d}\tau - tx'(0) \\
&= \int_0^1 tx'(\tau t)\,\mathrm{d}\tau - tx'(0) \\
&= \int_0^1 t(x'(\tau t) - x'(0))\,\mathrm{d}\tau \\
&= \int_0^1 t\int_0^{\tau t} x''(\sigma)\,\mathrm{d}\sigma\,\mathrm{d}\tau.
\end{aligned}$$

Daraus ergibt sich

$$\begin{aligned}
\|x(t) - \hat{x}(t)\|_2 &\leq \int_0^1 t\int_0^{\tau t} 2\eta\,\mathrm{d}\sigma\,\mathrm{d}\tau \\
&= \int_0^1 2\eta\tau t^2\,\mathrm{d}\tau \\
&= 2\eta t^2 \int_0^1 \tau\,\mathrm{d}\tau = \eta t^2.
\end{aligned}$$

□

Satz 1.27 (Wahl der Schrittweite $\tau > 0$ für die klassische Fortsetzung)**.** *Es sei $D \subset \mathbb{R}^n$ offen und konvex, $F : D \times [a,b] \to \mathbb{R}^n$ stetig differenzierbar mit auf ganz $D \times [a,b]$ invertierbarer Matrix $F_x(x,\lambda) \in \mathbb{R}^{n\times n}$. Es existiere ein $\omega > 0$, so dass*

$$\|F_x(x,\lambda)^{-1}(F_x(x+sv,\lambda) - F_x(x,\lambda))v\|_2 \leq s\omega \|v\|_2^2 \tag{1.9}$$

für alle $x \in D$, $\lambda \in [a,b]$, $v \in \mathbb{R}^n$ mit $x+v \in D$, und $s \in [0,1]$ gilt. Es sei $x(\cdot) \in C^1([0,\varepsilon],\mathbb{R}^n)$ mit $\varepsilon > 0$, so dass gilt

$$x(0) := x_0 \quad \textit{und} \quad F(x(t),\lambda_0 + t) = 0 \quad \forall t \in [0,\varepsilon].$$

Ist

$$0 < \tau < \tau_{max} := \min\left\{\varepsilon, \frac{2}{\omega \max\limits_{0\leq t\leq\varepsilon} \|x'(t)\|_2}\right\},$$

so konvergiert das Newton-Verfahren zum Startwert

$$\hat{x} := x_0 \quad \textit{(klassische Fortsetzungsmethode)}$$

gegen die Lösung $x(\tau)$ von

$$F(x(\tau),\lambda_0 + \tau) = 0.$$

Beweis. Wir verwenden den Konvergenzsatz des Newton-Verfahrens aus Abschnitt 1.1 auf die Aufgabe

$$F(x) := F(x,\lambda_0 + \tau) = 0.$$

Wir müssen also verifizieren, dass alle Voraussetzungen des genannten Satzes erfüllt sind. Die Lipschitz-Bedingung mit $\omega > 0$ ist bereits als Annahme (1.9) gefordert. Es bleibt also nur noch zu zeigen, dass

$$\|\bar{x} - x^0\|_2 < \frac{2}{\omega}$$

gilt. Bei uns sind

$$\bar{x} = x(\tau) \quad \text{und} \quad x^0 = \hat{x} = x_0 = x(0).$$

Also müssen wir die Ungleichung

$$\|x(\tau) - x(0)\|_2 < \frac{2}{\omega}$$

nachweisen. Aus dem (ersten) vorherigen Lemma wissen wir aber

$$\|x(\tau) - x(0)\|_2 \leq \eta\tau$$

mit $\eta := \max\limits_{0\leq t\leq\varepsilon} \|x'(t)\|_2$. Somit folgt die Konvergenz des Newton-Verfahrens gegen $x(\tau)$ aus

$$\|x(\tau) - x(0)\|_2 \leq \eta\tau < \eta\tau_{\max} \leq \eta\frac{2}{\omega\eta} = \frac{2}{\omega}.$$

□

Satz 1.28 (Wahl der Schrittweite $\tau > 0$ für die tangentiale Fortsetzung)**.** *Es sei $D \subset \mathbb{R}^n$ offen und konvex, $F : D \times [a,b] \to \mathbb{R}^n$ stetig differenzierbar mit auf ganz $D \times [a,b]$ invertierbarer Matrix $F_x(x,\lambda) \in \mathbb{R}^{n\times n}$. Es existiere ein $\omega > 0$, so dass*

$$\|F_x(x,\lambda)^{-1}(F_x(x+sv,\lambda) - F_x(x,\lambda))v\|_2 \leq s\omega \|v\|_2^2$$

für alle $x \in D$, $\lambda \in [a,b]$, $v \in \mathbb{R}^n$ mit $x+v \in D$, und $s \in [0,1]$ gilt. Es sei $x(\cdot) \in C^2([0,\varepsilon],\mathbb{R}^n)$ mit $\varepsilon > 0$, so dass gilt

$$x(0) := x_0 \quad \textit{und} \quad F(x(t),\lambda_0 + t) = 0 \quad \forall t \in [0,\varepsilon].$$

Ist

$$0 < \tau < \tau_{max} := \min\left\{\varepsilon, \sqrt{\frac{4}{\omega \max\limits_{0\leq t\leq\varepsilon} \|x''(t)\|_2}}\right\},$$

so konvergiert das Newton-Verfahren zum Startwert

$$\begin{cases} \hat{x} := x(0) + \tau x'(0) \\ F_x(x_0,\lambda_0)x'(0) = -F_\lambda(x_0,\lambda_0) \end{cases}$$

gegen die Lösung $x(\tau)$ von

$$F(x(\tau),\lambda_0+\tau) = 0.$$

Beweis. Analog zum vorheringen Satz müssen wir nur noch zeigen, dass

$$\|x(\tau) - \hat{x}\|_2 < \frac{2}{\omega}$$

gilt. Aus dem (zweiten) vorherigen Lemma folgt

$$\|x(\tau) - \hat{x}\|_2 \leq \eta\tau^2 < \eta\tau_{\max}^2 = \eta\frac{4}{\omega 2\eta} = \frac{2}{\omega}.$$

□

Mit den beiden Sätzen hätten wir das Problem analytisch gelöst. Aber wir haben dabei folgendes Problem:
Die Konstanten

$$\varepsilon > 0, \quad \omega > 0, \quad \max_{0\leq t\leq\varepsilon} \left\|x'(t)\right\|_2 \quad (\text{bzw. } \left\|x''(t)\right\|_2)$$

sind unbekannt. Im nächsten Abschnitt wollen wir eine numerische Strategie zur Schrittweitensteuerung betrachten, welche nur numerisch verfügbare Größen verwendet.

1.5.3 Natürlicher Monotonietest zur Schrittweitenbestimmung bei Fortsetzungsmethoden

Ausgehend von einer Lösung $(x_0, \lambda_0) \in D \times [a, b]$ mit

$$F(x_0, \lambda_0) = 0$$

suchen wir eine Schrittweite $\tau > 0$, so dass das Newton-Verfahren zur Lösung von

$$F(x, \lambda_0 + \tau) = 0$$

konvergiert. Unsere Idee besteht darin, den natürlichen Monotonietest anzuwenden:

$$\|\bar{d}^{k+1}\|_2 \leq \vartheta \|d^k\|_2 \quad (0 < \vartheta < 1, \text{ z.B. } \vartheta = \frac{1}{2}),$$

wobei $d^k \in \mathbb{R}^n$ bzw. $\bar{d}^{k+1} \in \mathbb{R}^n$ die Newton-Korrektur bzw. vereinfachte Newton-Korrektur bezeichnen. Also:

$$\begin{cases} F_x(x^k, \lambda_0 + \tau) d^k = -F(x^k, \lambda_0 + \tau), \\ F_x(x^k, \lambda_0 + \tau) \bar{d}^{k+1} = -F(x^{k+1}, \lambda_0 + \tau). \end{cases}$$

Gilt irgendwann

$$\|\bar{d}^{k+1}\|_2 > \vartheta \|d^k\|_2$$

für ein $k \in \mathbb{N}$, so war der Schritt des Prädiktors zu weit, das heißt τ ist zu groß. In diesem Fall verkleinern wir die Schrittweite, etwa wie

$$\tau = \beta \tau$$

mit einem Verkleinerungsfaktor $\beta \in (0, 1)$, z.B. $\beta = \frac{1}{2}$, und starten den Newton-Prozess neu.
Aber $\tau > 0$ kann auch zu klein sein. Man könnte das annehmen, wenn

$$\|\bar{d}^1\| \leq \frac{\vartheta}{2} \|d^0\|_2, \quad \text{(nur bei Start des Newton-Verfahrens)}$$

so sollte man τ vergrößern, etwa

$$\tau = \frac{1}{\beta} \tau$$

und neu starten. Weitere Algorithmen findet man im Buch von Deufelhard und Hohmann.

1.5.4 Homotopiemethode

Gegeben seien $F : \mathbb{R}^n \to \mathbb{R}^n$ und $G : \mathbb{R}^n \to \mathbb{R}^n$. Wir betrachten die nichtlineare Aufgabe

$$\begin{cases} F(x) = 0, \\ G(x) = 0. \end{cases}$$

Dabei ist die Aufgabe $F(x) = 0$ schwer zu lösen, während $G(x) = 0$ leicht ist. Die Idee ist, sich in einer Folge zu lösender Probleme an $F(x) = 0$ heranzutasten. Mit anderen Worten konstruiert man ein parameterabhängiges nichtlineares Problem

$$H(x, \lambda) = 0,$$

wobei $H : \mathbb{R}^n \times [0, 1] \to \mathbb{R}^n$, mit der Eigenschaft

$$H(x, 0) = G(x) \quad \text{und} \quad H(x, 1) = F(x).$$

Dann löst man $H(x, \lambda) = 0$ mit Hilfe der Fortsetzungsmethode, angefangen mit $\lambda_0 = 0$, also die einfache Aufgabe

$$H(x, 0) = G(x) = 0.$$

Definition 1.29. Ein solches H heißt *Einbettung* des Problems $F(x) = 0$ oder auch *Homotopie*.

Beispiel 1.30.

$$H(x, \lambda) := (1 - \lambda)G(x) + \lambda F(x).$$

Beispiel 1.31. Es sei

$$F : \mathbb{R}^{10} \to \mathbb{R}^{10}, \quad F(x) = x - \phi(x)$$

mit

$$\phi(x) = \begin{pmatrix} \phi_1(x) \\ \vdots \\ \phi_{10}(x) \end{pmatrix} \quad \text{mit} \quad \phi_i(x) = \exp\left(\cos\left(i \sum_{j=1}^{10} x_j\right)\right).$$

Mögliche Einbettungen sind dann gegeben durch

1. $H(x, \lambda) = x - \lambda\phi(x)$.
2. $H(x, \lambda) = \begin{pmatrix} H_1(x, \lambda) \\ \vdots \\ H_{10}(x, \lambda) \end{pmatrix}$ mit

$$H_i(x, \lambda) = x_i - \exp\left(\lambda \cos\left(i \sum_{j=1}^{10} x_j\right)\right).$$

Kapitel 2

Eigenwertprobleme

2.1 Vorbetrachtungen

Im Folgenden sei $A \in \mathbb{C}^{n\times n}$ bzw. $A \in \mathbb{R}^{n\times n}$ eine Matrix, von welcher wir alle Eigenwerte und Eigenvektoren bestimmen wollen. Gesucht sind also die Zahlen $\lambda \in \mathbb{C}$ und Vektoren $x \in \mathbb{C}^n \setminus \{0\}$, so dass gilt

$$Ax = \lambda x.$$

Definition 2.1. Die *Menge aller Eigenwerte* von A bezeichnen wir mit

$$\Lambda(A) := \{\lambda \in \mathbb{C} \mid \exists\, x \in \mathbb{C}^n \setminus \{0\} : Ax = \lambda x\} \quad (\textit{Spektrum} \text{ von } A).$$

Bemerkung 2.2. Aus der linearen Algebra gilt

$$\lambda \in \Lambda(A) \quad \Leftrightarrow \quad P_A(\lambda) := \det(A - \lambda I) = 0,$$

wobei P_A das charakteristische Polynom von A bezeichnet. Dieses Polynom hat n (ggf. mehrfache) Nullstellen $\lambda_1, \ldots, \lambda_n$. Folglich ist

$$\Lambda(A) = \{\lambda_1, \ldots, \lambda_n\}.$$

Definition 2.3. Ist $\lambda \in \Lambda(A)$ eine k-fache Nullstelle von P_A, so heißt k die *algebraische Vielfachheit* von λ. Mit

$$E_A(\lambda) := \{x \in \mathbb{C}^n : (A - \lambda I)x = 0\}$$

bezeichnen wir den zu λ gehörenden Eigenraum. Die Dimension des Eigenraums $E_A(\lambda)$ ist die *geometrische Vielfachheit* von λ.

Notation 2.4.

$$\sigma(A) = \text{algebraische Vielfachheit},$$
$$\rho(A) = \text{geometrische Vielfachheit}.$$

Aus der linearen Algebra gilt

$$\rho(A) \leq \sigma(A),$$

der Beweis erfolgt beispielsweise durch die Jordansche Normalform.
Im Prinzip scheint die Bestimmung der Eigenwerte numerisch einfach zu sein, denn man braucht nur die Nullstellen von P_A zu bestimmen. Es gibt jedoch wichtige Argumente dagegen!

(i) Aufstellen von $P_A(\lambda)$ erfordert die Berechnung von $\det(A - \lambda I)$, welches für großes n sehr aufwendig ist.

(ii) Das Berechnen aller komplexen Nullstellen von P_A ist schwierig.

(iii) Kleine Störungen des Polynoms P_A führt zu großen Fehlern in den Eigenwerten.

Beispiel 2.5 (Wilkinson)**.** Wir betrachten das Polynom

$$p(\lambda) = (\lambda - 1)(\lambda - 2) \cdot \ldots \cdot (\lambda - 20).$$

Die Nullstellen sind also $\lambda_k = k$ für $k = 1, \ldots, 20$. Ausmultipliziert erhalten wir

$$p(\lambda) = a_0 + a_1\lambda + \ldots + a_n\lambda^n$$

mit Größenordnungen der Koeffizienten $1, \ldots, 10^{18}(20!)$. Nun stören wir a_{19} um $\varepsilon = 2^{-23} \approx 10^{-7}$ und betrachten das gestörte Polynom

$$p^*(\lambda) = a_0 + a_1\lambda + \ldots + a_{18}\lambda^{18} + (a_{19} - \varepsilon)\lambda^{19} + a_{20}\lambda^{20} = p(\lambda) - \varepsilon\lambda^{19}.$$

Die (exakten) Nullstellen von p^* sind:

1.000	2.000	3.000	4.000	4.999
6.000	7.000	8.007	8.917	$10.095 \pm 0.643i$
$11.793 \pm 1.652i$	$13.992 \pm 2.518i$	$16.730 \pm 2.812i$	$19.502 \pm 1.940i$	20.846

2.2 Nullstellen von gestörten reellen Polynomen

Im Folgenden sei

$$p(\lambda) := \lambda^n + p_{n-1}\lambda^{n-1} + \ldots + p_1\lambda + p_0$$

mit $p_0, \ldots, p_{n-1} \in \mathbb{R}$. Ferner sei

$$p^*(\lambda) := p_n^*\lambda^n + p_{n-1}^*\lambda^{n-1} + \ldots + p_1^*\lambda + p_0^*$$

ein gestörtes Polynom mit Koeffizienten

$$\begin{cases} p_j^* := p_j + \varepsilon q_j, & j = 0, \ldots, n-1 \\ p_n^* := 1 + \varepsilon q_n \end{cases}$$

mit $q_j \in \mathbb{R}$ für alle $j = 0, \ldots, n$ und $\varepsilon \in \mathbb{R}$ „klein“. Folglich ist

$$p^*(\lambda) = p(\lambda) + \varepsilon q(\lambda)$$

mit

$$q(\lambda) = \sum_{j=1}^{n} \lambda^j q_j.$$

Wir nehmen an, dass p eine einfache reelle Nullstelle $\lambda_0 \in \mathbb{R}$ habe:

$$p(\lambda_0) = 0 \quad \text{und} \quad p'(\lambda_0) \neq 0.$$

Wir wollen nun die Nullstelle von dem gestörten Polynom p^* um λ_0 untersuchen. Das heißt, wir betrachten die Aufgabe

$$p^*(\lambda) = p(\lambda) + \varepsilon q(\lambda) = 0$$

in Abhängigkeit von ε und λ. Dazu setzen wir

$$F : \mathbb{R} \times \mathbb{R} \to \mathbb{R}, \quad F(\lambda, \varepsilon) := p(\lambda) + \varepsilon q(\lambda).$$

Laut Voraussetzungen gilt

$$F(\lambda_0, 0) = p(\lambda_0) + 0q(\lambda_0) = p(\lambda_0) = 0.$$

Ferner gilt

$$\frac{\partial F}{\partial \lambda}(\lambda_0, 0) = p'(\lambda_0) + \varepsilon q'(\lambda_0)\big|_{\varepsilon=0} = p'(\lambda_0) + 0q'(\lambda_0) = p'(\lambda_0) \neq 0.$$

Somit liefert der Satz über implizite Funktionen die Existenz offener Umgebungen $U \subset \mathbb{R}$ um 0 und $V \subset \mathbb{R}$ um λ_0 mit den folgenden Eigenschaften:

(i) Zu jedem $\varepsilon \in U$ gibt es genau ein $\lambda \in V$, so dass gilt

$$F(\lambda, \varepsilon) = 0.$$

(ii) Die nach (i) eindeutig bestimmte Funktion $\lambda : U \to V$ mit

$$F(\lambda(\varepsilon), \varepsilon) = 0$$

ist stetig differenzierbar. Für jedes $\varepsilon \in U$ ist also $\lambda(\varepsilon) \in V$ eine Nullstelle von p^*.

(iii) Für jedes $\varepsilon \in U$ gilt

$$\frac{\partial F}{\partial \lambda}(\lambda(\varepsilon), \varepsilon) \neq 0.$$

Aus (i)-(ii) ergibt sich

$$g(\varepsilon) := F(\lambda(\varepsilon), \varepsilon) = 0 \quad \forall \varepsilon \in U.$$

Hieraus folgt

$$0 = g'(\varepsilon) = \partial_\varepsilon F(\lambda(\varepsilon), \varepsilon) = p'(\lambda(\varepsilon))\lambda'(\varepsilon) + \varepsilon q'(\lambda(\varepsilon))\lambda'(\varepsilon) + q(\lambda(\varepsilon)) \quad \forall \varepsilon \in U,$$

was äquivalent ist zu

$$\left[p'(\lambda(\varepsilon)) + \varepsilon q'(\lambda(\varepsilon))\right] \lambda'(\varepsilon) = -q(\lambda(\varepsilon)) \quad \forall \varepsilon \in U.$$

Folglich gilt

$$\lambda'(\varepsilon) = \frac{-q(\lambda(\varepsilon))}{p'(\lambda(\varepsilon)) + \varepsilon q'(\lambda(\varepsilon))} \quad \forall \varepsilon \in U,$$

denn laut (iii) ist $\frac{\partial F}{\partial \lambda}(\lambda(\varepsilon), \varepsilon) \neq 0$ für alle $\varepsilon \in U$. Insbesondere gilt für $\varepsilon = 0$

$$\lambda'(0) = \frac{-q(\lambda_0)}{p'(\lambda_0)}.$$

Somit ist

$$\lambda(\varepsilon) \approx \lambda_0 - \varepsilon \frac{q(\lambda_0)}{p'(\lambda_0)},$$

falls $\varepsilon \approx 0$, also sehr klein. Der Faktor

$$\left| \frac{q(\lambda_0)}{p'(\lambda_0)} \right|$$

ist also wichtig. Ist dieser groß im Vergleich mit $\varepsilon > 0$, so erhält man einen großen Fehler für die Nullstelle.

Beispiel 2.6. Betrachte das Polynom

$$p(\lambda) = (\lambda - 1)(\lambda - 2) \cdot \ldots \cdot (\lambda - 12).$$

Wir stören dieses Polynom nun wie folgt:

$$p^*(\lambda) = p(\lambda) + \varepsilon p_j \lambda^j = p(\lambda) + \varepsilon q(\lambda)$$

mit $q(\lambda) = p_j \lambda^j$ und $0 \leq j \leq 12$. Die Nullstellen von p sind

$$\lambda_k = k, \quad k = 1, \ldots, 12.$$

Wir wollen nun den Verstärkungsfaktor

$$\left| \frac{q(\lambda_0)}{p'(\lambda_0)} \right|$$

berechnen. Dazu ist

$$p'(\lambda_k) = p'(k) = (-1)^{12-k}(k-1)!(12-k)! \quad \text{für } k = 1, \ldots, 12.$$

Daraus folgt

$$c_{kj} := \left| \frac{q(\lambda_k)}{p'(\lambda_k)} \right| = \frac{|p_j| k^j}{(k-1)!(12-k)!} \quad \text{für } k = 1, \ldots, 12, \quad j = 0, \ldots, 12.$$

Die Werte c_{kj} können jedoch sehr groß sein:

j \ k	1	9	10
0	1.20 E01	1.98 E03	6.06 E02
3	3.54 E01	4.26 E06	1.95 E06
7	1.74 E−01	1.37 E08	9.54 E07
11	1.35 E−06	1.01 E07	1.07 E07

Im Vergleich mit anderen Werten liegt das Maximum bei $k = 9$ und $j = 7$. Also ist die Nullstelle $\lambda_9 = 9$ bezüglich Störungen von $p_7 = -6926634$ am sensibelsten. Betrachte eine kleine Störung

$$p_7^* = -6926634.001.$$

Dann ist

$$\lambda_9^* \approx \lambda_9 - \varepsilon \frac{q(\lambda_9)}{p'(\lambda_9)} = 8.98$$

bei dem relativen Fehler der Störung

$$\frac{0.001}{|p_7|} = 1.444 \cdot 10^{-10}.$$

2.3 Die Schur-Zerlegung

In diesem Abschnitt wollen wir wichtige theoretische Grundlagen für numerische Berechnung von Eigenwerten bereitstellen.

Definition 2.7. Eine Menge $S \subset \mathbb{C}^n$ heißt *invarianter Unterraum* von $A \in \mathbb{C}^{n \times n}$, falls gilt

$$Ax \in S \quad \forall x \in S \quad \Leftrightarrow \quad AS \subset S.$$

In diesem Fall heißt S auch A-invarianter Unterraum.

Lemma 2.8. *Es sei $A \in \mathbb{C}^{n \times n}$ und $\lambda \in \mathbb{C}$ ein Eigenwert von A. Dann ist der zu λ gehörende Eigenraum*

$$E_A(\lambda) \subset \mathbb{C}^n$$

A-invariant.

Beweis. Sei $x \in E_A(\lambda)$. Dann ist

$$(A - \lambda I)x = 0 \quad \Leftrightarrow \quad 0 = A0 = A(A - \lambda I)x = (A^2 - \lambda A)x = (A - \lambda I)Ax$$

und daraus folgt die Behauptung. □

Lemma 2.9. *Es sei $A \in \mathbb{C}^{n\times n}$ und $X \in \mathbb{C}^{n\times k}$. Existiert eine Matrix $B \in \mathbb{C}^{k\times k}$ mit*

$$AX = XB,$$

so ist das Bild

$$\text{Bild}(X) = \text{Span}\{X_1, \ldots, X_k\}$$

A-invariant. Beachte, dass X_i, $i = 1, \ldots, k$, den i-ten Spaltenvektor von X bezeichnet.

Beweis. Sei $x \in \text{Bild}(X)$. Zu zeigen: $Ax \in \text{Bild}(X)$. Es gilt

$$x \in \text{Bild}(X) \quad \Leftrightarrow \quad x = y_1X_1 + \ldots + y_kX_k = Xy$$

mit $y = (y_1, \ldots, y_k)^T \in \mathbb{C}^k$. Folglich ist

$$Ax = AXy = XBy$$

und daraus folgt

$$Ax = z_1X_1 + \ldots + z_kX_k$$

und schließlich

$$Ax \in \text{Bild}(x).$$

□

Vertauschen wir die Rolle von B und X, so erhalten wir das folgende Resultat.

Lemma 2.10. *Es seien $A \in \mathbb{C}^{n\times n}$ und $B \in \mathbb{C}^{k\times k}$. Existiert eine Matrix $X \in \mathbb{C}^{n\times k}$ mit* $\text{Rang}(X) = k$ *und*

$$AX = XB,$$

so gilt

$$\Lambda(B) \subset \Lambda(A).$$

Beweis. Es sei $\lambda \in \Lambda(B)$. Dann gibt es einen Vektor $y \in \mathbb{C}^k \setminus \{0\}$ mit

$$By = \lambda y.$$

Unsere Annahme liefert

$$AXy = XBy = \lambda Xy.$$

Da $\text{Rang}(X) = k$ und $y \neq 0$ ist, so ist

$$Xy \in \mathbb{C}^n \setminus \{0\}$$

und daraus folgt

$$\lambda \in \Lambda(A).$$

□

Korollar 2.11. *Es seien $A, B \in \mathbb{C}^{n\times n}$. Existiert eine reguläre Matrix $X \in \mathbb{C}^{n\times n}$ mit*

$$AX = XB,$$

so gilt

$$\Lambda(A) = \Lambda(B).$$

Beweis. Die Inklusion

$$\Lambda(B) \subset \Lambda(A)$$

ist bereits erfüllt. Die Voraussetzung liefert auch

$$X^{-1}A = X^{-1}AXX^{-1} = X^{-1}XBX^{-1} = BX^{-1},$$

also mit anderen Worten

$$BX^{-1} = X^{-1}A.$$

Somit liefert das obige Lemma mit $B = A$ die Aussage

$$\Lambda(A) \subset \Lambda(B).$$

□

Bemerkung 2.12. Die Voraussetzung $AX = XB$ mit einer regulären Matrix X ist äquivalent zu

$$B = X^{-1}AX.$$

Definition 2.13. Zwei Matrizen $A, B \in \mathbb{C}^{n\times n}$ heißen *ähnlich*, falls eine reguläre Matrix $X \in \mathbb{C}^{n\times n}$ existiert mit

$$B = X^{-1}AX.$$

Die Transformation

$$T : A \to T^{-1}AT$$

heißt *Ähnlichkeitstransformation*.

Folgerung 2.14. Sind $A, B \in \mathbb{C}^{n\times n}$ ähnlich, so besitzen A und B die gleichen Eigenwerte.

Lemma 2.15. *Sei $A \in \mathbb{C}^{n\times n}$. Es existiere $B \in \mathbb{C}^{k\times k}$, $k \leq n$, und $X \in \mathbb{C}^{n\times k}$ mit* $\mathrm{Rang}(X) = k$ *und*

$$AX = XB.$$

Dann existiert eine unitäre Matrix $Q \in \mathbb{C}^{n\times n}$ (das heißt $Q^*Q = I$), so dass gilt

$$Q^*AQ = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}$$

und

$$\Lambda(T_{11}) = \Lambda(A) \cap \Lambda(B) = \Lambda(B).$$

Beweis. Aus der linearen Algebra wissen wir, dass X eine volle QR-Zerlegung besitzt, also

$$X = Q \begin{bmatrix} R \\ 0 \end{bmatrix}$$

mit einer unitären Matrix $Q \in \mathbb{C}^{n\times n}$ und einer oberen Dreiecksmatrix $R \in \mathbb{C}^{k\times k}$. Hieraus folgt

$$AX = XB \quad \Rightarrow \quad AQ \begin{bmatrix} R \\ 0 \end{bmatrix} = Q \begin{bmatrix} R \\ 0 \end{bmatrix} B \quad \Rightarrow \quad Q^* AQ \begin{bmatrix} R \\ 0 \end{bmatrix} = \begin{bmatrix} RB \\ 0 \end{bmatrix}. \tag{2.1}$$

Wir schreiben nun

$$Q^* AQ = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}. \tag{2.2}$$

Wir zeigen zuerst, dass $T_{21} = 0$ ist. Wegen $\text{Rang}(X) = k$ muss

$$R = \begin{pmatrix} r_{11} & & * \\ & \ddots & \\ & & r_{kk} \end{pmatrix}$$

regulär sein, denn

$$X = Q \begin{bmatrix} R \\ 0 \end{bmatrix}.$$

Aus (2.1)-(2.2) gilt

$$\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} R \\ 0 \end{bmatrix} = \begin{bmatrix} RB \\ 0 \end{bmatrix} \quad \Leftrightarrow \quad \begin{matrix} T_{11}R + 0 = RB \\ T_{21}R + 0 = 0 \end{matrix}. \tag{2.3}$$

Aufgrund der Regularität von R ist nun $T_{21} = 0$. Es bleibt nur noch zu zeigen:

$$\Lambda(T_{11}) = \Lambda(A) \cap \Lambda(B).$$

Aus (2.3) wissen wir

$$R^{-1} T_{11} R = B,$$

denn $R \in \mathbb{C}^{k\times k}$ ist regulär. Mit anderen Worten sind $T_{11}, B \in \mathbb{C}^{k\times k}$ ähnlich. Somit gilt

$$\Lambda(T_{11}) = \Lambda(B).$$

Außerdem ist

$$\Lambda(B) \subset \Lambda(A)$$

nach dem vorherigen Lemma. Insgesamt gilt

$$\Lambda(T_{11}) = \Lambda(A) \cap \Lambda(B) = \Lambda(B).$$

□

Korollar 2.16. *Sei $A \in \mathbb{C}^{n\times n}$. Es existieren $B \in \mathbb{C}^{k\times k}$, $k \leq n$, $X \in \mathbb{C}^{n\times k}$ mit* $\mathrm{Rang}(X) = k$ *und*

$$AX = XB.$$

Dann existiert eine unitäre Matrix $Q \in \mathbb{C}^{n\times n}$ mit

$$Q^* AQ = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}$$

und

$$\Lambda(A) = \Lambda(T_{11}) \cup \Lambda(T_{22}).$$

Beweis. Aus dem vorherigen Lemma wissen wir bereits, dass es eine unitäre Matrix $Q \in \mathbb{C}^{n\times n}$ gibt, so dass gilt

$$Q^* AQ = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} \quad \text{und} \quad \Lambda(T_{11}) = \Lambda(A) \cap \Lambda(B).$$

Aus der Darstellung sind $A \in \mathbb{C}^{n\times n}$ und

$$T = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} \in \mathbb{C}^{n\times n}$$

ähnlich und somit sind die Eigenwerte von A und T gleich, das heißt

$$\Lambda(A) = \Lambda(T).$$

Also müssen wir nur noch nachweisen, dass

$$\Lambda(T) = \Lambda(T_{11}) \cup \Lambda(T_{22})$$

gilt. Es sei $\lambda \in \Lambda(T)$. Dann gilt

$$0 = P_T(\lambda) \quad \Leftrightarrow \quad 0 = \det(T - \lambda I)$$

und

$$\begin{aligned} & \det\left(\begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} - \lambda I\right) && = 0 \\ \Leftrightarrow \quad & \det\left(\begin{bmatrix} T_{11} - \lambda I & T_{12} \\ 0 & T_{22} - \lambda I \end{bmatrix}\right) && = 0 \\ \Leftrightarrow \quad & \det(T_{11} - \lambda I)\det(T_{22} - \lambda I) && = 0 \\ \Leftrightarrow \quad & \lambda \in \Lambda(T_{11}) \quad \text{oder} \quad \lambda \in \Lambda(T_{22}). \end{aligned}$$

□

Satz 2.17 (Schur). *Zu jeder Matrix $A \in \mathbb{C}^{n\times n}$ existiert eine unitäre Matrix $Q \in \mathbb{C}^{n\times n}$, so dass gilt*

$$Q^* A Q = \begin{pmatrix} t_{11} & & * \\ & \ddots & \\ 0 & & t_{nn} \end{pmatrix} \in \mathbb{C}^{n\times n}$$

und

$$\Lambda(A) = \{t_{11}, \ldots, t_{nn}\}.$$

Beweis durch Induktion.

Induktionsanfang $n = 1$: $Q = (1)$.

Induktionsannahme Die Aussage gelte für $n = m \in \mathbb{N}$. Wir zeigen nun, dass die Aussage für $n = m + 1$ gilt.

Sei $A \in \mathbb{C}^{(m+1)\times(m+1)}$ und sei $\lambda \in \Lambda(A)$ beliebig aber fest. Dann existiert ein Eigenvektor $x \in \mathbb{C}^{m+1} \setminus \{0\}$ mit

$$Ax = \lambda x.$$

Das obige Lemma mit $B = (\lambda) \in \mathbb{C}^{1\times 1}$ (das heißt $k = 1$) und $X := x \in \mathbb{C}^{(m+1)\times 1}$ ($\mathrm{Rang}(X) = 1$) liefert die Existenz einer unitären Matrix $U \in \mathbb{C}^{(m+1)\times(m+1)}$, so dass gilt

$$U^* A U = \begin{pmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix}$$

mit

$$\Lambda(T_{11}) = \Lambda(A) \cap \Lambda(B) = \Lambda(B) = \{\lambda\}.$$

Somit ist $T_{11} = (\lambda)$. Außerdem gilt

$$\Lambda(A) = \{\lambda\} \cup \Lambda(T_{22}). \tag{2.4}$$

Da $T_{22} \in \mathbb{C}^{m\times m}$ ist, existiert laut Induktionsannahme eine unitäre Matrix $\tilde{U} \in \mathbb{C}^{m\times m}$ mit

$$\tilde{U}^* T_{22} \tilde{U} = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_m \end{pmatrix}$$

und

$$\Lambda(T_{22}) = \{\lambda_1, \ldots, \lambda_m\}. \tag{2.5}$$

Wir setzen nun

$$Q := U \begin{pmatrix} 1 & 0 \\ 0 & \tilde{U} \end{pmatrix} \in \mathbb{C}^{(m+1)\times(m+1)}.$$

Folglich gilt

$$\begin{aligned} Q^*AQ = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{U}^* \end{pmatrix} U^*AU \begin{pmatrix} 1 & 0 \\ 0 & \tilde{U} \end{pmatrix} &= \begin{pmatrix} 1 & 0 \\ 0 & \tilde{U}^* \end{pmatrix} \begin{pmatrix} \lambda & T_{12} \\ 0 & T_{22} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \tilde{U} \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & \tilde{U}^* \end{pmatrix} \begin{pmatrix} \lambda & T_{12}\tilde{U} \\ 0 & T_{22}\tilde{U} \end{pmatrix} \\ &= \begin{pmatrix} \lambda & T_{12}\tilde{U} \\ 0 & \tilde{U}^*T_{22}\tilde{U} \end{pmatrix} \\ &= \begin{pmatrix} \lambda & & * \\ & \lambda_1 & \\ & & \ddots \\ 0 & & \lambda_m \end{pmatrix} \end{aligned}$$

und aufgrund von (2.4)-(2.5) gilt

$$\Lambda(A) = \{\lambda, \lambda_1, \ldots, \lambda_m\}$$

und daraus folgt die Behauptung. □

Definition 2.18. Es sei $A \in \mathbb{C}^{n\times n}$. Dann heißt die Zerlegung

$$Q^*AQ = \begin{pmatrix} t_{11} & & * \\ & \ddots & \\ 0 & & t_{nn} \end{pmatrix} \in \mathbb{C}^{n\times n}$$

mit einer unitären Matrix $Q \in \mathbb{C}^{n\times n}$ und der Eigenschaft, dass

$$\Lambda(A) = \{t_{11}, \ldots, t_{nn}\}$$

gilt, *Schur-Zerlegung* von A. Die Spaltenvektoren von Q heißen *Schur-Vektoren*.

Bemerkung 2.19. Die Schur-Zerlegung ist nicht eindeutig.

Korollar 2.20. *Es sei $A \in \mathbb{C}^{n\times n}$ und $q_1, \ldots, q_n \in \mathbb{C}^n$ Schur-Vektoren einer Schur-Zerlegung*

$$Q^*AQ = \begin{pmatrix} t_{11} & & * \\ & \ddots & \\ 0 & & t_{nn} \end{pmatrix} \in \mathbb{C}^{n\times n}.$$

Dann ist

$$S_k = \operatorname{Span}\{q_1, \ldots, q_k\} \subset \mathbb{C}^n$$

für jedes $k = 1, \ldots, n$ A-invariant.

Bemerkung 2.21. Der vorherige Satz sichert die Existenz einer Schur-Zerlegung für jede Matrix. Aus der Schur-Zerlegung kann man sofort die Eigenwerte von A ablesen. Leider kann man ohne weitere Annahme die dazugehörigen Eigenvektoren nicht bestimmen.

Definition 2.22. Eine Matrix $A \in \mathbb{C}^{n\times n}$ heißt *normal*, falls gilt

$$A^*A = AA^*.$$

Beispiel 2.23. Jede unitäre Matrix ist normal, symmetrische reelle Matrizen sowie hermitsche Matrizen auch.

Lemma 2.24. *Es sei $A \in \mathbb{C}^{n\times n}$ eine quadratische Matrix und $Q \in \mathbb{C}^{n\times n}$ eine unitäre Matrix. Dann ist A normal genau dann, wenn Q^*AQ normal ist.*

Beweis.

„⇒“: Es sei $A \in \mathbb{C}^{n\times n}$ normal. Dann gilt

$$\begin{aligned}(Q^*AQ)^*(Q^*AQ) = Q^*A^*QQ^*AQ = Q^*A^*AQ = Q^*AA^*Q = Q^*AQQ^*A^*Q \\ = (Q^*AQ)(Q^*AQ)^*.\end{aligned}$$

„⇐“: Analog.

□

Satz 2.25. *Eine Matrix $A \in \mathbb{C}^{n\times n}$ ist genau dann normal, wenn es eine unitäre Matrix $Q \in \mathbb{C}^{n\times n}$ gibt, so dass gilt*

$$Q^*AQ = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

und

$$\Lambda(A) = \{\lambda_1, \ldots, \lambda_n\}.$$

Insbesondere ist λ_i für jedes $i = 1, \ldots, n$ ein Eigenwert von A zum Eigenvektor Qe_i.

Beweis.

„⇐“: Es gebe eine unitäre Matrix $Q \in \mathbb{C}^{n\times n}$ mit

$$Q^*AQ = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}.$$

Dann gilt

$$A = Q \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} Q^*.$$

Daraus folgt

$$\begin{aligned} A^*A &= Q \begin{pmatrix} \bar{\lambda}_1 & & 0 \\ & \ddots & \\ 0 & & \bar{\lambda}_n \end{pmatrix} Q^*Q \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} Q^* \\ &= Q \begin{pmatrix} |\lambda_1|^2 & & 0 \\ & \ddots & \\ 0 & & |\lambda_n|^2 \end{pmatrix} Q^* \\ &= \cdots = AA^*. \end{aligned}$$

„⇒“: Es sei A normal. Wir haben bereits gezeigt, dass jede Matrix eine Schur-Zerlegung besitzt. Es existiere also eine unitäre Matrix $Q \in \mathbb{C}^{n\times n}$, so dass gilt

$$Q^*AQ = \begin{pmatrix} t_{11} & & * \\ & \ddots & \\ 0 & & t_{nn} \end{pmatrix} =: T$$

und

$$\Lambda(A) = \{t_{11}, \ldots, t_{nn}\}.$$

Wir zeigen nun, dass dann $t_{ij} = 0$ für alle $j > i$ gilt. Da $A \in \mathbb{C}^{n\times n}$ normal ist und $Q \in \mathbb{C}^{n\times n}$ unitär ist, so ist

$$T = Q^*AQ$$

ebenso normal, das heißt

$$T^*T = TT^*.$$

Wir vergleichen nun die Diagonalelemente von T^*T und TT^*.

$$\begin{aligned} T^*T &= \begin{pmatrix} \bar{t}_{11} & & & & 0 \\ \bar{t}_{12} & \bar{t}_{22} & & & \\ \bar{t}_{13} & \bar{t}_{23} & \bar{t}_{33} & & \\ \vdots & \vdots & \vdots & \ddots & \\ \bar{t}_{1n} & \bar{t}_{2n} & \bar{t}_{3n} & \cdots & \bar{t}_{nn} \end{pmatrix} \begin{pmatrix} t_{11} & t_{12} & t_{13} & \cdots & t_{1n} \\ & t_{22} & t_{23} & \cdots & t_{2n} \\ & & t_{33} & \cdots & t_{3n} \\ & & & \ddots & \vdots \\ 0 & & & & t_{nn} \end{pmatrix} \\ &= \begin{pmatrix} |t_{11}|^2 & & & & * \\ & |t_{12}|^2 + |t_{22}|^2 & & & \\ & & \sum_{j=1}^{3} |t_{j3}|^2 & & \\ & & & \ddots & \\ * & & & & \sum_{j=1}^{n} |t_{jn}|^2 \end{pmatrix} \end{aligned}$$

und

$$TT^* = \begin{pmatrix} t_{11} & t_{12} & t_{13} & \cdots & t_{1n} \\ & t_{22} & t_{23} & \cdots & t_{2n} \\ & & t_{33} & \cdots & t_{3n} \\ & & & \ddots & \vdots \\ 0 & & & & t_{nn} \end{pmatrix} \begin{pmatrix} \bar{t}_{11} & & & & 0 \\ \bar{t}_{12} & \bar{t}_{22} & & & \\ \bar{t}_{13} & \bar{t}_{23} & \bar{t}_{33} & & \\ \vdots & \vdots & \vdots & \ddots & \\ \bar{t}_{1n} & \bar{t}_{2n} & \bar{t}_{3n} & \cdots & \bar{t}_{nn} \end{pmatrix}$$
$$= \begin{pmatrix} \sum_{j=1}^{n} |t_{1j}|^2 & & & & * \\ & \sum_{j=2}^{n} |t_{2j}|^2 & & & \\ & & \sum_{j=3}^{n} |t_{2j}|^2 & & \\ & & & \ddots & \\ * & & & & |t_{nn}|^2 \end{pmatrix}.$$

Mit $T^*T = TT^*$ folgt

$$|t_{11}|^2 = \sum_{j=1}^{n} |t_{1j}|^2 \quad \Rightarrow \quad 0 = \sum_{j=2}^{n} |t_{1j}|^2 \quad \Rightarrow \quad t_{1j} = 0 \quad \forall j > 1,$$
$$|t_{12}|^2 + |t_{22}|^2 = \sum_{j=2}^{n} |t_{2j}|^2 \quad \Rightarrow \quad 0 = \sum_{j=3}^{n} |t_{2j}|^2 \quad \Rightarrow \quad t_{2j} = 0 \quad \forall j > 2,$$
$$|t_{13}|^2 + |t_{23}|^2 + |t_{33}|^2 = \sum_{j=3}^{n} |t_{3j}|^2 \quad \Rightarrow \quad 0 = \sum_{j=4}^{n} |t_{3j}|^2 \quad \Rightarrow \quad t_{3j} = 0 \quad \forall j > 3,$$

und so weiter. Folglich gilt

$$t_{ij} = 0 \quad \forall j > i$$

und daraus ergibt sich die Behauptung.

□

Bemerkung 2.26. Ist $A \in \mathbb{R}^{n \times n}$ symmetrisch, so wissen wir bereits, dass A reelle Eigenwerte besitzt. In diesem Fall existiert eine orthogonale Matrix $Q \in \mathbb{R}^{n \times n}$, so dass gilt

$$Q^T A Q = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

und

$$\Lambda(A) = \{\lambda_1, \ldots, \lambda_n\}.$$

Ist A nicht symmetrisch, dann ist diese Aussage im Allgemeinen nicht richtig.

Es sei $\lambda \in \mathbb{C}$ ein Eigenwert von $A \in \mathbb{R}^{n \times n}$ mit Eigenvektor $x \in \mathbb{C}^n \setminus \{0\}$. Dann gilt

$$Ax = \lambda x \quad \Leftrightarrow \quad \bar{A}\bar{x} = \bar{\lambda}\bar{x} \quad \Leftrightarrow \quad A\bar{x} = \bar{\lambda}\bar{x},$$

da $A \in \mathbb{R}^{n\times n}$. Deshalb gilt

$$\lambda \in \Lambda(A) \text{ mit Eigenvektor } x \quad \Leftrightarrow \quad \bar{\lambda} \in \Lambda(A) \text{ mit Eigenvektor } \bar{x}.$$

Wir zerlegen

$$\begin{aligned} \lambda &= \mu + i\beta, \quad \mu, \beta \in \mathbb{R}, \\ \bar{x} &= u + iv, \quad u, v \in \mathbb{R}^n. \end{aligned}$$

Folglich haben wir

$$A(u+iv) = (\mu + i\beta)(u+iv)$$

und daher

$$\begin{cases} \mathrm{Re}(A(u+iv)) = \mu u - \beta v, \\ \mathrm{Im}(A(u+iv)) = \beta u + \mu v, \end{cases}$$

sowie

$$\begin{cases} Au = \mu u - \beta v, \\ Av = \beta u + \mu v. \end{cases}$$

Schließlich erhalten wir

$$A \begin{pmatrix} u & v \end{pmatrix} = \begin{pmatrix} \mu u - \beta v & \beta u + \mu v \end{pmatrix} = \begin{pmatrix} u & v \end{pmatrix} \begin{pmatrix} \mu & \beta \\ -\beta & \mu \end{pmatrix}. \tag{2.6}$$

Korollar 2.27. *Es sei $A \in \mathbb{R}^{n\times n}$ und $\lambda = \mu + i\beta \in \Lambda(A)$. Ist*

$$x = u + iv, \quad u, v \in \mathbb{R}^n$$

ein Eigenvektor von A zum Eigenwert λ, so ist

$$\mathrm{Span}\{u, v\} \subset \mathbb{R}^n$$

A-invariant.

Beweis. Es sei $z \in \mathrm{Span}\{u, v\}$. Das heißt

$$\exists \alpha_1, \alpha_2 \in \mathbb{R} : \quad z = \alpha_1 u + \alpha_2 v = \begin{pmatrix} u & v \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}.$$

Folglich gilt

$$\begin{aligned} Az = A \begin{pmatrix} u & v \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} &\overset{(2.6)}{=} \begin{pmatrix} u & v \end{pmatrix} \begin{pmatrix} \mu & \beta \\ -\beta & \mu \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \\ &= \begin{pmatrix} u & v \end{pmatrix} \begin{pmatrix} \mu\alpha_1 + \beta\alpha_2 \\ -\beta\alpha_1 + \mu\alpha_2 \end{pmatrix} \\ &= (\mu\alpha_1 + \beta\alpha_2)u + (-\beta\alpha_1 + \mu\alpha_2)v \in \mathrm{Span}\{u, v\}. \end{aligned}$$

□

Satz 2.28 (Reelle Schur-Form). *Es sei $A \in \mathbb{R}^{n\times n}$. Dann existiert eine orthogonale Matrix $Q \in \mathbb{R}^{n\times n}$, so dass gilt*

$$Q^T A Q = \begin{pmatrix} R_{11} & & * \\ & \ddots & \\ 0 & & R_{mm} \end{pmatrix} \in \mathbb{R}^{n\times n}$$

mit Blockmatrizen $R_{jj} \in \mathbb{R}^{k_j\times k_j}$, $k_j \in \{1,2\}$, und $\sum_{j=1}^m k_j = n$. Dabei gilt für $k_j = 1$, dass $R_{jj} \in \Lambda(A)$ und für $k_j = 2$, dass $\Lambda(R_{jj}) = \{\lambda, \bar{\lambda}\} \subset \Lambda(A)$.

2.4 Störungssätze

Wir beginnen mit dem bekannten Satz über die Jordan-Normalform.

Satz 2.29 (Jordansche-Normalform). *Es sei $A \in \mathbb{C}^{n\times n}$. Dann existiert eine reguläre Matrix $X \in \mathbb{C}^{n\times n}$, so dass gilt*

$$X^{-1} A X = \begin{pmatrix} J(\lambda_1) & & 0 \\ & \ddots & \\ 0 & & J(\lambda_k) \end{pmatrix}$$

mit den paarweise verschiedenen Eigenwerten $\lambda_1, \ldots, \lambda_k$. Die Jordan-Blöcke $J(\lambda_l)$, $l = 1, \ldots, k$ haben die Form

$$J(\lambda_l) = \begin{pmatrix} J_{l,1} & & 0 \\ & \ddots & \\ 0 & & J_{l,\rho(\lambda_l)} \end{pmatrix}, \quad J_{l,r} = \begin{pmatrix} \lambda_l & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_l \end{pmatrix} \in \mathbb{C}^{m_{l,r}\times m_{l,r}}.$$

Es gilt $\rho(\lambda_l) = \sigma(\lambda_l)$ genau dann, wenn $m_{l,r} = 1$ für alle $r = 1, \ldots, \rho(\lambda_l)$ gilt.

Definition 2.30. Es sei $A \in \mathbb{C}^{n\times n}$ und $\lambda \in \Lambda(A)$.

(i) Ist $\rho(\lambda) < \sigma(\lambda)$, so heißt λ *defektiver Eigenwert*.

(ii) Besitzt A einen defektiven Eigenwert, so heißt A *defektiv*.

(iii) Gilt $\rho(\lambda) = \sigma(\lambda)$ für alle $\lambda \in \Lambda(A)$, so heißt A *diagonalisierbar*.

Lemma 2.31. *Die Inklusion*

$$\{A \in \mathbb{C}^{n\times n} \mid A \text{ ist diagonalisierbar}\} \subset \mathbb{C}^{n\times n}$$

ist dicht. Mit anderen Worten:

$$\forall A \in \mathbb{C}^{n\times n} \quad \forall \delta > 0 \quad \exists\, A_\delta \in \mathbb{C}^{n\times n} \text{ diagonalisierbar} : \quad \|A_\delta - A\|_2 \le \delta.$$

Notation:

$$\overline{\{A \in \mathbb{C}^{n\times n} \mid A \text{ ist diagonalisierbar}\}}^{\|\cdot\|_2} = \mathbb{C}^{n\times n}.$$

Beweis. Es sei $A \in \mathbb{C}^{n\times n}$. Ist $J_{l,r}$ ein gegebener Jordan-Block zum mehrfachen Eigenwert λ_l

$$J_{l,r} = \begin{pmatrix} \lambda_l & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_l \end{pmatrix},$$

so kann man durch Addition beliebig kleiner $\varepsilon_i \neq 0$, $i = 1, \ldots, m_{l,r}$, diesen Block ändern zu

$$\tilde{J}_{l,r} = \begin{pmatrix} \lambda_l - \varepsilon_1 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_l - \varepsilon_{m_{l,r}} \end{pmatrix},$$

welcher paarweise verschiedene Eigenwerte hat, wenn die ε_i unterschiedlich gewählt sind. So erzeugt man insgesamt durch beliebig kleine Änderung eine diagonalisierbare Matrix $\tilde{A} \in \mathbb{C}^{n\times n}$. □

Folgerung 2.32. Jede Matrix ist numerisch diagonalisierbar.

Beispiel 2.33. Die Matrix

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

ist defektiv (nicht diagonalisierbar), denn

$$\rho(A) = \dim E_A(1) = 1 < 2 = \sigma(1),$$

denn 1 ist zweifache Nullstelle von

$$\det(A - \lambda I) = 0 \quad \Leftrightarrow \quad (1-\lambda)^2 = 0.$$

Die klein gestörte Matrix

$$\tilde{A} = \begin{pmatrix} 1 & 1 \\ \varepsilon & 1 \end{pmatrix} \quad \text{mit } \varepsilon > 0 \text{ „klein“}$$

ist diagonalisierbar, denn $\Lambda(\tilde{A}) = \{1 + \sqrt{\varepsilon}, 1 - \sqrt{\varepsilon}\}$.

Es sei $X \in \mathbb{C}^{n\times n}$ regulär. Wir betrachten die Transformation

$$T : A \to X^{-1}AX.$$

Sei $\tilde{A}$ eine gestörte Matrix, also

$$\tilde{A} = A + E.$$

Dann gilt

$$T(\tilde{A}) = T(A) + T(E),$$
$$\|T(\tilde{A}) - T(A)\|_2 = \|T(E)\|_2 \leq \|X^{-1}\|_2 \, \|E\|_2 \, \|X\|_2 = \operatorname{cond}_2(X) \, \|E\|_2 .$$

Korollar 2.34. *Ist $X \in \mathbb{C}^{n\times n}$ unitär, so gilt*

$$\|T(\tilde{A}) - T(A)\|_2 \leq \|E\|_2 = \|A - \tilde{A}\|_2$$

für alle Matrizen $A, \tilde{A} \in \mathbb{C}^{n\times n}$.

Beweis. Für die Spektralkondition gilt $\operatorname{cond}_2(X) = \|X^{-1}\|_2 \|X\|_2 = 1$, denn

$$\|X\|_2 = \sqrt{\lambda_{\max}(X^*X)} = 1.$$

□

Satz 2.35 (Bauer-Fike)**.** *Es sei $A \in \mathbb{C}^{n\times n}$ diagonalisierbar, das heißt es existiere eine reguläre Matrix $X \in \mathbb{C}^{n\times n}$, so dass gilt*

$$X^{-1}AX = D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}, \quad \Lambda(A) = \{\lambda_1, \ldots, \lambda_n\}.$$

Ferner sei $E \in \mathbb{C}^{n\times n}$ eine Störung und $\mu \in \Lambda(A + E)$. Dann gilt

$$\min_{\lambda\in\Lambda(A)} |\mu - \lambda| \leq \operatorname{cond}_2(X) \|E\|_2 .$$

Beweis. Ist $\mu \in \Lambda(A + E) \cap \Lambda(A)$, dann ist die Aussage trivial. Nun sei $\mu \in \Lambda(A + E)$, aber $\mu \notin \Lambda(A)$. Es gilt

$$X^{-1}(A + E - \mu I)X = X^{-1}AX + X^{-1}EX - \mu I = D - \mu I + X^{-1}EX.$$

Da $\mu \notin \Lambda(A) = \{\lambda_1, \ldots, \lambda_n\}$ gilt, so ist

$$\det(D - \mu I) = \det \begin{pmatrix} \lambda_1 - \mu & & 0 \\ & \ddots & \\ 0 & & \lambda_n - \mu \end{pmatrix} \neq 0.$$

Daher ist $D - \mu I$ regulär. Folglich ist

$$X^{-1}(A + E - \mu I)X = (D - \mu I)(I + (D - \mu I)^{-1}X^{-1}EX).$$

Berechnung der Determinante auf beiden Seiten liefert

$$\begin{aligned} \underbrace{\det(D - \mu I)}_{\neq 0} \det(I + (D - \mu I)^{-1}X^{-1}EX) &= \det(X^{-1}(A + E - \mu I)X) \\ &= \det(X^{-1}) \underbrace{\det(A + E - \mu I)}_{=0} \det(X), \end{aligned}$$

also

$$\det(I + (D - \mu I)^{-1} X^{-1} E X) = 0$$

und somit ist $I + (D - \mu I)^{-1} X^{-1} E X$ nicht regulär. Dann existiert ein $z \in \mathbb{C}^n \setminus \{0\}$ mit

$$(I + (D - \mu I)^{-1} X^{-1} E X) z = 0.$$

Es folgt

$$\|z\|_2 = \|(D - \mu I)^{-1} (X^{-1} E X) z\|_2$$

und daher

$$\|z\|_2 \leq \|(D - \mu I)^{-1}\|_2 \|X^{-1}\|_2 \, \|E\|_2 \, \|X\|_2 \, \|z\|_2 .$$

Mit $\|z\|_2 \neq 0$ folgt

$$1 \leq \|(D - \mu I)^{-1}\|_2 \operatorname{cond}_2(X) \, \|E\|_2 ,$$

also

$$\|(D - \mu I)^{-1}\|_2^{-1} \leq \operatorname{cond}_2(X) \, \|E\|_2 .$$

Laut Definition ist aber

$$\begin{aligned} \left\| (D - \mu I)^{-1} \right\|_2 = \sqrt{\lambda_{\max}((D - \mu I)^{-*} (D - \mu I)^{-1})} &= \sqrt{\lambda_{\max} \begin{pmatrix} \frac{1}{|\lambda_1 - \mu|^2} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{|\lambda_n - \mu|^2} \end{pmatrix}} \\ &= \sqrt{\min_{1 \leq i \leq n} |\lambda_i - \mu|^{-2}} \\ &= \frac{1}{\min_{1 \leq i \leq n} |\lambda_i - \mu|}, \end{aligned}$$

denn

$$(D - \mu I)^{-1} = \begin{pmatrix} \frac{1}{\lambda_1 - \mu} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\lambda_n - \mu} \end{pmatrix}.$$

Insgesamt gilt

$$\min_{1 \leq i \leq n} |\lambda_i - \mu| \leq \operatorname{cond}_2(X) \, \|X\|_2 .$$

□

Korollar 2.36. *Ist zusätzlich zu den Voraussetzungen des vorherigen Satzes die Matrix A normal, dann gilt*

$$\min_{\lambda \in \Lambda(A)} |\lambda - \mu| \leq \|E\|_2 .$$

Der Satz von Bauer-Fike liefert ein wichtiges Resultat der Rückwärtsanalyse:
Es sei $A \in \mathbb{C}^{n\times n}$ diagonalisierbar und zu lösen sei

$$Ax = \lambda x.$$

Das numerische Verfahren liefert ein fehlerbehaftetes $\tilde{x}$ und $\tilde{\lambda}$ mit o.B.d.A. $\|\tilde{x}\|_2 = 1$ und

$$A\tilde{x} = \tilde{\lambda}\tilde{x} + r$$

mit einem kleinen Fehler $r \neq 0$. Wäre $r = 0$, so wäre $\tilde{\lambda} \in \Lambda(A)$ mit exaktem Eigenvektor $\tilde{x}$. Mit dem Satz von Bauer-Fike kann man zeigen, dass

$$\min_{\lambda\in\Lambda(A)} |\lambda - \tilde{\lambda}| \leq \operatorname{cond}_2(X) \|r\|_2 .$$

Beispiel 2.37. Es sei

$$A = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & \varepsilon & 1 \end{pmatrix}.$$

Es gilt

$$0 = \det(A - \lambda I) \quad \Leftrightarrow \quad (2-\lambda)((1-\lambda)^2 - \varepsilon) = 0.$$

Somit ist

$$\Lambda(A) = \{2, 1+\sqrt{\varepsilon}, 1-\sqrt{\varepsilon}\}.$$

Als Eigenvektoren erhält man ohne Normierung

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \\ \sqrt{\varepsilon} \end{pmatrix}, \quad \begin{pmatrix} 0 \\ -1 \\ \sqrt{\varepsilon} \end{pmatrix}.$$

Die Matrix A ist nach Definition diagonalisierbar. Für

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & \sqrt{\varepsilon} & \sqrt{\varepsilon} \end{pmatrix}, \quad X^{-1} = \frac{1}{2\sqrt{\varepsilon}} \begin{pmatrix} 2\sqrt{\varepsilon} & 0 & 0 \\ 0 & \sqrt{\varepsilon} & 1 \\ 0 & -\sqrt{\varepsilon} & 1 \end{pmatrix}$$

gilt

$$X^{-1}AX = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1+\sqrt{\varepsilon} & 0 \\ 0 & 0 & 1-\sqrt{\varepsilon} \end{pmatrix}.$$

Daraus ergibt sich mit dem Satz von Bauer-Fike und $\mu \in \Lambda(A+E)$

$$\min_{\lambda\in\Lambda(A)} |\lambda - \mu| \leq \operatorname{cond}_2(X) \|E\|_2 \to \infty \quad \text{für } \varepsilon \searrow 0.$$

Beachte, dass die Berechnung des Eigenwerts 2 für obiges Beispiel eigentlich unproblematisch ist. Aber der Satz von Bauer-Fike liefert nur eine allgemeine Aussage für alle

Eigenwerte, so dass numerische Probleme wie oben für alle Eigenwerte angedeutet werden.
Daher wäre es sinnvoll, die Sensitivität eines Eigenwertes nur über seinen zugeordneten Eigenvektor zu untersuchen.

Satz 2.38 (Störung einfacher Eigenwerte)**.** *Es sei $A \in \mathbb{C}^{n\times n}$ und $\lambda_0 \in \Lambda(A)$ mit $\sigma(\lambda_0) = 1$, das heißt λ_0 sei einfach. Ferner seien $x_0, y_0 \in \mathbb{C}^n \setminus \{0\}$ rechte bzw. linke normierte Eigenvektoren zum Eigenwert λ_0:*

$$Ax_0 = \lambda_0 x_0 \quad \textit{sowie} \quad y_0^* A = \lambda_0 y_0^*, \quad \|x_0\|_2 = \|y_0\|_2 = 1.$$

Es sei $E \in \mathbb{C}^{n\times n}$ eine Störung. Dann existiert eine stetig differenzierbare Abbildung

$$\lambda : (-\varepsilon, \varepsilon) \to B_\delta(\lambda_0) \subset \mathbb{C}$$

mit $\varepsilon, \delta \in \mathbb{R}^+$, so dass gilt

(i) $\lambda(0) = \lambda_0$,

(ii) $\Lambda(A + tE) \cap B_\delta(\lambda_0) = \{\lambda(t)\} \quad \forall t \in (-\varepsilon, \varepsilon)$,

(iii) $\sigma(\lambda(t)) = 1 \quad \forall t \in (-\varepsilon, \varepsilon)$,

(iv) $\lambda'(0) = \frac{(y_0, Ex_0)_{\mathbb{C}^n}}{(y_0, x_0)_{\mathbb{C}^n}}$.

Beweis. Laut Voraussetzung ist $\lambda_0 \in \Lambda(A)$ ein einfacher Eigenwert und daher ist λ_0 eine einfache Nullstelle des charakteristischen Polynoms

$$P_A(\lambda) := \det(A - \lambda I).$$

Mit anderen Worten:

$$P_A(\lambda_0) = 0 \quad \text{und} \quad P_A'(\lambda_0) \neq 0.$$

Wir definieren die Funktion

$$F : \mathbb{C} \times \mathbb{R} \to \mathbb{C}, \quad F(\lambda, t) := \det(A + tE - \lambda I) = P_{A+tE}(\lambda).$$

Diese Funktion ist unendlich oft differenzierbar mit

$$\begin{aligned} F(\lambda_0, 0) &= P_A(\lambda_0) = 0, \\ F_\lambda(\lambda_0, 0) &= P_A'(\lambda_0) \neq 0. \end{aligned}$$

Somit liefert der Satz über implizite Funktionen die Existenz von $\varepsilon, \delta \in \mathbb{R}^+$, so dass gilt:

(i) Zu jedem $t \in (-\varepsilon, \varepsilon)$ gibt es genau ein $\lambda \in B_\delta(\lambda_0) \subset \mathbb{C}$, so dass gilt

$$F(\lambda, t) = 0.$$

(ii) Die nach (i) eindeutig bestimmte Abbildung

$$\lambda : (-\varepsilon, \varepsilon) \to B_\delta(\lambda) \quad \text{mit} \quad F(\lambda(t), t) = 0 \quad \forall t \in (-\varepsilon, \varepsilon)$$

ist stetig differenzierbar.

Da $F_\lambda(\lambda_0, 0) = P'_A(\lambda_0) \neq 0$ gilt und die Abbildung $t \to F_\lambda(\lambda(t), t)$ stetig ist, können wir ε verkleinern, so dass gilt

$$F_\lambda(\lambda(t), t) \neq 0 \quad \forall t \in (-\varepsilon, \varepsilon) \quad \Leftrightarrow \quad P'_{A+tE}(\lambda(t)) \neq 0 \quad \forall t \in (-\varepsilon, \varepsilon).$$

Daher ist

$$\sigma(\lambda(t)) = 1 \quad \forall t \in (-\varepsilon, \varepsilon).$$

Insgesamt erfüllt die Funktion $\lambda : (-\varepsilon, \varepsilon) \to B_\delta(\lambda_0)$ die Eigenschaften:

(i) $\lambda(0) = \lambda_0$,

(ii) $\Lambda(A + tE) \cap B_\delta(\lambda_0) = \{\lambda(t)\} \quad \forall t \in (-\varepsilon, \varepsilon)$,

(iii) $\sigma(\lambda(t)) = 1 \quad \forall t \in (-\varepsilon, \varepsilon)$.

Es bleibt nur noch zu zeigen, dass

$$\lambda'(0) = \frac{(y_0, Ex_0)_{\mathbb{C}^n}}{(y_0, x_0)_{\mathbb{C}^n}}$$

gilt. Beachte, dass $(y_0, x_0)_{\mathbb{C}^n} \neq 0$ ist, da $\sigma(\lambda_0) = 1$ ist (siehe Hausaufgabe). Da $\sigma(\lambda(t)) = 1$ für alle $t \in (-\varepsilon, \varepsilon)$ gilt, so gibt es zu jedem $t \in (-\varepsilon, \varepsilon)$ genau einen normierten (rechten) Eigenvektor

$$x = x(t) \in \mathbb{C}^n \setminus \{0\} \quad \text{mit } \|x\|_2 = 1$$

von $A + tE$ zum Eigenwert $\lambda(t)$. Laut Definition ist

$$\begin{aligned} \frac{(y_0, Ax(t) - Ax_0)_{\mathbb{C}^n}}{t} = \frac{(A^* y_0, x(t) - x_0)_{\mathbb{C}^n}}{t} &= \frac{(\bar{\lambda}_0 y_0, x(t) - x_0)_{\mathbb{C}^n}}{t} \\ &= \frac{\lambda_0 (y_0, x(t) - x_0)_{\mathbb{C}^n}}{t}. \end{aligned} \tag{2.7}$$

Andererseits gilt

$$\begin{aligned} \frac{(y_0, Ax(t) - Ax_0)_{\mathbb{C}^n}}{t} &= \frac{(y_0, (A + tE)x(t) - Ax_0)_{\mathbb{C}^n}}{t} - \frac{(y_0, tEx(t))_{\mathbb{C}^n}}{t} \\ &= \frac{(y_0, \lambda(t)x(t) - \lambda_0 x_0)_{\mathbb{C}^n}}{t} - (y_0, Ex(t))_{\mathbb{C}^n} \\ &= \frac{\lambda(t) - \lambda_0}{t}(y_0, x_0)_{\mathbb{C}^n} + \frac{\lambda(t)}{t}(y_0, x(t) - x_0)_{\mathbb{C}^n} - (y_0, Ex(t))_{\mathbb{C}^n}. \end{aligned} \tag{2.8}$$

Somit liefern (2.7)-(2.8)

$$(y_0, Ex(t))_{\mathbb{C}^n} = \frac{\lambda(t)-\lambda_0}{t}(y_0, x_0)_{\mathbb{C}^n} + \frac{\lambda(t)-\lambda(0)}{t}(y_0, x(t)-x_0)_{\mathbb{C}^n}.$$

Mit $t \to 0$ erhalten wir

$$(y_0, Ex(0))_{\mathbb{C}^n} = \lambda'(0)(y_0, x_0)_{\mathbb{C}^n} + \lambda'(0)(y_0, 0)_{\mathbb{C}^n}$$

und mit $(y_0, x_0)_{\mathbb{C}^n} \neq 0$ folgt insgesamt

$$\lambda'(0) = \frac{(y_0, Ex_0)_{\mathbb{C}^n}}{(y_0, x_0)_{\mathbb{C}^n}}.$$

□

Beachte, dass man hier auch die Stetigkeit der Abbildung

$$x : (-\varepsilon, \varepsilon) \to \mathbb{C}^n \setminus \{0\}$$

im Punkt 0 angewendet hat. In der Tat ist diese Abbildung stetig (siehe Hausaufgabe).

Korollar 2.39. *Es sei $A \in \mathbb{C}^{n\times n}$ und $\lambda_0 \in \Lambda(A)$ einfach. Ferner seien $x_0, y_0 \in \mathbb{C}^n \setminus \{0\}$ rechte bzw. linke normierte Eigenvektoren von A zum Eigenwert λ_0. Ferner sei $E \in \mathbb{C}^{n\times n}$ eine Störung und*

$$\lambda : (-\varepsilon, \varepsilon) \to B_\delta(\lambda_0)$$

die im vorherigen Satz wohldefinierte Abbildung der einfachen Eigenwerte mit

$$\lambda(0) = \lambda_0, \quad \Lambda(A+tE) \cap B_\delta = \{\lambda(t)\} \quad \forall t \in (-\varepsilon, \varepsilon) \quad \text{und} \quad \sigma(\lambda(t)) = 1 \quad \forall t \in (-\varepsilon, \varepsilon).$$

Dann gilt

$$|\lambda(t) - \lambda(0)| \leq \frac{|t|}{|(y_0, x_0)_{\mathbb{C}^n}|} \|E\|_2 + \mathcal{O}(|t|).$$

Beweis. Die Abbildung $\lambda : (-\varepsilon, \varepsilon) \to B_\delta(\lambda_0)$ ist differenzierbar und somit ist

$$\lambda(t) = \lambda(0) + \lambda'(0)t + \mathcal{O}(|t|)$$

und nach dem vorherigen Satz gilt

$$\lambda(t) - \lambda(0) = \frac{t(y_0, Ex_0)_{\mathbb{C}^n}}{(y_0, x_0)_{\mathbb{C}^n}} + \mathcal{O}(|t|).$$

Insgesamt ergibt sich also

$$|\lambda(t) - \lambda(0)| \leq \frac{|t| \, \|y_0\|_2 \, \|E\|_2 \, \|x_0\|_2}{|(y_0, x_0)_{\mathbb{C}^n}|} + \mathcal{O}(|t|)$$

und somit die Behauptung. □

Definition 2.40. Sei $A \in \mathbb{C}^{n\times n}$ eine Matrix und $\lambda_0 \in \Lambda(A)$ einfach. Ferner seien $x_0, y_0 \in \mathbb{C}^n \setminus \{0\}$ rechte bzw. linke normierte Eigenvektoren von A zum Eigenwert λ_0. Dann heißt

$$\operatorname{cond}(\lambda_0) := \frac{1}{|(y_0, x_0)_{\mathbb{C}^n}|}$$

Kondition des einfachen Eigenwerts λ_0.

Bemerkung 2.41.

(i) Machen sie sich klar, dass $\operatorname{cond}(\lambda_0) \in [1, \infty)$ gilt.

(ii) Ideal wäre $\operatorname{cond}(\lambda_0) = 1$. Dies gilt z.B. wenn A normal ist.

2.5 Rayleigh-Ritz-Quotient und Courant-Fischer-Variationsprinzip

Im Folgenden untersuchen wir Eigenwertprobleme mit hermiteschen Matrizen.

Definition 2.42. Eine Matrix A heißt *hermitesch*, falls gilt

$$A = A^* \quad \Leftrightarrow \quad a_{ij} = \bar{a}_{ji} \quad \forall i, j = 1, \ldots, n.$$

Lemma 2.43. *Jeder Eigenwert einer hermiteschen Matrix ist immer reell.*

Beweis. Sei $A \in \mathbb{C}^{n\times n}$ hermitesch und $\lambda \in \Lambda(A)$ mit dem Eigenvektor $x \in \mathbb{C}^n \setminus \{0\}$. Dann gilt

$$\bar{\lambda}(x, x)_{\mathbb{C}^n} = (\lambda x, x)_{\mathbb{C}^n} = (Ax, x)_{\mathbb{C}^n} = (x, A^* x)_{\mathbb{C}^n} = (x, Ax)_{\mathbb{C}^n} = (x, \lambda x)_{\mathbb{C}^n} = \lambda (x, x)_{\mathbb{C}^n}.$$

Da $\|x\|_2 \neq 0$ ist, folgt $\bar{\lambda} = \lambda$. □

Korollar 2.44. *Eine Matrix $A \in \mathbb{C}^{n\times n}$ ist genau dann hermitesch, wenn es eine unitäre Matrix $Q \in \mathbb{C}^{n\times n}$ gibt, so dass gilt*

$$Q^* A Q = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^{n\times n}.$$

Definition 2.45. Es sei $A \in \mathbb{C}^{n\times n}$. Dann bezeichnen wir mit

$$R : \mathbb{C}^n \setminus \{0\} \to \mathbb{C}, \quad R(x) := \frac{(x, Ax)_{\mathbb{C}^n}}{\|x\|_2^2}$$

den zu A gehörigen *Rayleigh-Ritz-Quotienten*.

Es hat sich herausgestellt, dass der Rayleigh-Ritz-Quotient ein wichtiges analytisches Werkzeug zur Untersuchung von Eigenwertproblemen ist. Zunächst wollen wir einige elementare Eigenschaften des Rayleigh-Ritz-Quotienten beweisen. Dazu definieren wir das Bild von R wie folgt:

$$\operatorname{Bild}(R) := \{R(x) \mid x \in \mathbb{C}^n \setminus \{0\}\}.$$

Lemma 2.46. *Es sei $A \in \mathbb{C}^{n\times n}$. Dann gilt*

$$\Lambda(A) \subset \text{Bild}(R).$$

Beweis. Sei $\lambda \in \Lambda(A)$ und $x \in \mathbb{C}^n \setminus \{0\}$ ein dazugehöriger Eigenvektor. Dann gilt

$$\lambda = \lambda \frac{(x,x)_{\mathbb{C}^n}}{\|x\|_2^2} = \frac{(x,\lambda x)_{\mathbb{C}^n}}{\|x\|_2^2} = \frac{(x,Ax)_{\mathbb{C}^n}}{\|x\|_2^2} = R(x) \in \text{Bild}(R).$$

Daher ist $\Lambda(A) \subset \text{Bild}(R)$. □

Lemma 2.47. *Es sei $A \in \mathbb{C}^{n\times n}$ normal, das heißt*

$$A^*A = AA^*.$$

Dann gilt

$$\text{Bild}(R) = \left\{ \sum_{j=1}^n a_j\lambda_j \mid a_j \in \mathbb{R},\ a_j \geq 0 \quad \forall j = 1,\dots,n,\ \sum_{j=1}^n a_j = 1 \right\},$$

wobei $\Lambda(A) = \{\lambda_1,\dots,\lambda_n\}$ ist. Mit anderen Worten ist das Bild von R nichts anderes als die konvexe Hülle von $\Lambda(A) = \{\lambda_1,\dots,\lambda_n\}$.

Beweis. Da $A \in \mathbb{C}^{n\times n}$ normal ist, existiert eine unitäre Matrix $Q \in \mathbb{C}^{n\times n}$ mit

$$Q^*AQ = D = \text{diag}(\lambda_1,\dots,\lambda_n) \in \mathbb{C}^{n\times n}.$$

Hieraus ergibt sich

$$\begin{aligned}
\text{Bild}(R) = \{R(x) \mid x \neq 0\} &= \left\{ \frac{(x,Ax)_{\mathbb{C}^n}}{(x,x)_{\mathbb{C}^n}} \mid x \neq 0 \right\} \\
&= \left\{ \frac{x^*Ax}{x^*x} \mid x \neq 0 \right\} \\
&= \left\{ \frac{x^*QDQ^*x}{x^*QQ^*x} \mid x \neq 0 \right\} \\
&= \left\{ \frac{y^*Dy}{y^*y} \mid y \neq 0 \right\} \\
&= \left\{ \frac{\sum_{j=1}^n \lambda_j|y_j|^2}{\sum_{j=1}^n |y_j|^2} \mid y \neq 0 \right\} \\
&= \left\{ \sum_{j=1}^n a_j\lambda_j \mid a_j \in \mathbb{R},\ a_j \geq 0,\ \sum_{j=1}^n a_j = 1 \right\}.
\end{aligned}$$

Daraus folgt

$$\text{Bild}(R) = \text{ konvexe Hülle von } \Lambda(A).$$

□

Satz 2.48. *Es sei $A \in \mathbb{C}^{n\times n}$ hermitesch, das heißt $A^* = A$. Dann gilt*

$$\text{Bild}(R) = [\lambda_{\min}, \lambda_{\max}] \subset \mathbb{R}$$

(das Bild ist ein abgeschlossenes Intervall in $\mathbb{R}$ (Kompaktum)) und

$$\lambda_{\min} = \min_{x\in\mathbb{C}^n\setminus\{0\}} R(x), \quad \lambda_{\max} = \max_{x\in\mathbb{C}^n\setminus\{0\}} R(x),$$

wobei $\lambda_{\min}$ und $\lambda_{\max}$ den kleinsten bzw. den größten Eigenwert von A bezeichnen.

Beweis. Da die Matrix $A \in \mathbb{C}^{n\times n}$ hermitesch ist, besitzt diese nur reelle Eigenwerte

$$\lambda_{\min} := \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n =: \lambda_{\max}.$$

Außerdem wissen wir, dass jede hermitesche Matrix normal ist. Somit liefert das vorherige Lemma

$$\text{Bild}(R) = \left\{ \sum_{j=1}^{n} a_j\lambda_j \mid a_j \in \mathbb{R},\ a_j \geq 0,\ \sum_{j=1}^{n} a_j = 1 \right\} = [\lambda_{\min}, \lambda_{\max}].$$

Sind $x_{\min}, x_{\max} \in \mathbb{C}^n \setminus \{0\}$ Eigenvektoren von A zu Eigenwerten $\lambda_{\min}$ und $\lambda_{\max}$, so gilt

$$R(x_{\min}) = \frac{(x_{\min}, Ax_{\min})_{\mathbb{C}^n}}{\|x_{\min}\|_2^2} = \frac{(x_{\min}, \lambda_{\min}x_{\min})_{\mathbb{C}^n}}{(x_{\min}, x_{\min})_{\mathbb{C}^n}} = \lambda_{\min}$$
$$R(x_{\max}) = \frac{(x_{\max}, Ax_{\max})_{\mathbb{C}^n}}{\|x_{\max}\|_2^2} = \frac{(x_{\max}, \lambda_{\max}x_{\max})_{\mathbb{C}^n}}{(x_{\max}, x_{\max})_{\mathbb{C}^n}} = \lambda_{\max}$$

und somit folgt insgesamt die Aussage. □

Satz 2.49 (Variationsprinzip von Rayleigh-Ritz)**.** *Es sei $A \in \mathbb{C}^{n\times n}$ hermitesch mit den reellen Eigenwerten*

$$\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$$

und zugehörigen Eigenvektoren $y_1, \ldots, y_n \in \mathbb{C}^n \setminus \{0\}$ mit $(y_i, y_j)_{\mathbb{C}^n} = \delta_{ij}$. Dann gilt für jedes $j = 1, \ldots, n$:

$$\lambda_j = \begin{cases} \min & R(x) \\ \textit{bei} & (x, y_k)_{\mathbb{C}^n} = 0 \quad \forall k = 1, \ldots, j-1 \\ & x \in \mathbb{C}^n \setminus \{0\} \end{cases}$$
$$= \begin{cases} \max & R(x) \\ \textit{bei} & (x, y_k)_{\mathbb{C}^n} = 0 \quad \forall k = j+1, \ldots, n\,. \\ & x \in \mathbb{C}^n \setminus \{0\} \end{cases}$$

Beweis. Sei $j \in \{1, \dots, n\}$ und $x \in \mathbb{C}^n \setminus \{0\}$ mit $(x, y_k)_{\mathbb{C}^n} = 0$ für alle $k = 1, \dots, j-1$. Der Vektor x hat die Darstellung

$$x = \sum_{k=1}^{n} a_k y_k.$$

Daraus folgt

$$\begin{aligned} R(x) = \frac{(x, Ax)_{\mathbb{C}^n}}{(x,x)_{\mathbb{C}^n}} = \frac{\sum_{k=1}^n (x, a_k A y_k)_{\mathbb{C}^n}}{\sum_{k=1}^n (x, a_k y_k)_{\mathbb{C}^n}} &= \frac{\sum_{k=1}^n \lambda_k a_k (x, y_k)_{\mathbb{C}^n}}{\sum_{k=1}^n a_k (x, y_k)_{\mathbb{C}^n}} \\ &= \frac{\sum_{k=j}^n \lambda_k a_k (x, y_k)_{\mathbb{C}^n}}{\sum_{k=j}^n a_k (x, y_k)_{\mathbb{C}^n}} \\ &\geq \lambda_j \frac{\sum_{k=j}^n |a_k|^2}{\sum_{k=j}^n |a_k|^2} = \lambda_j. \end{aligned}$$

Also ergibt sich

$$R(x) \geq \lambda_j.$$

Für $x = y_j$ ergibt sich

$$R(y_j) = \frac{(y_j, Ay_j)_{\mathbb{C}^n}}{\|y_j\|_2^2} = \lambda_j.$$

Also ist

$$\lambda_j = \begin{cases} \min & R(x) \\ \text{bei} & (x, y_k)_{\mathbb{C}^n} = 0 \quad \forall k = 1, \dots, j-1 \\ & x \in \mathbb{C}^n \setminus \{0\} \end{cases}.$$

□

Fazit 2.50. Ist die Matrix A hermitesch, so lassen sich alle reellen Eigenwerte von A als Minimierungs- bzw. Maximierungsproblem mit dem Kostenfunktional R darstellen:

$$\lambda_1 = \min_{x \in \mathbb{C}^n \setminus \{0\}} R(x), \quad \lambda_n = \max_{x \in \mathbb{C}^n \setminus \{0\}} R(x)$$

sowie

$$\begin{aligned} \lambda_j &= \begin{cases} \min & R(x) \\ \text{bei} & (x, y_k)_{\mathbb{C}^n} = 0 \quad \forall k = 1, \dots, j-1 \\ & x \in \mathbb{C}^n \setminus \{0\} \end{cases} \\ &= \begin{cases} \max & R(x) \\ \text{bei} & (x, y_k)_{\mathbb{C}^n} = 0 \quad \forall k = j+1, \dots, n \\ & x \in \mathbb{C}^n \setminus \{0\} \end{cases}. \end{aligned}$$

Satz 2.51 (Variationsprinzip von Courant und Fischer)**.** *Es sei $A \in \mathbb{C}^{n\times n}$ hermitesch mit Eigenwerten*

$$\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n.$$

Dann gilt für jedes $j = 1, \ldots, n$:

$$\lambda_j = \min_{\substack{V\subset\mathbb{C}^n \\ \dim(V)=j}} \max_{x\in V\setminus\{0\}} R(x) = \max_{\substack{V\subset\mathbb{C}^n \\ \dim(V)=n-j+1}} \min_{x\in V\setminus\{0\}} R(x).$$

Beweis. Da $A \in \mathbb{C}^{n\times n}$ hermitesch ist, existiert eine Orthonormalbasis $\{y_k\}_{k=1}^n \subset \mathbb{C}^n \setminus \{0\}$ aus Eigenvektoren von A zu Eigenwerten λ_k. Sei nun $j \in \{1, \ldots, n\}$ beliebig aber fest und $V \subset \mathbb{C}^n$ ein Unterraum mit $\dim(V) = j$. Es gilt

$$V \cap \text{Span}\{y_j, \ldots, y_n\} \neq \{0\},$$

denn wäre $V \cap \text{Span}\{y_j, \ldots, y_n\} = \{0\}$, so wäre

$$\dim(V + \text{Span}\{y_j, \ldots, y_n\}) = \dim(V) + \dim(\text{Span}\{y_j, \ldots, y_n\}) = j + n - j + 1 = n + 1.$$

Sei nun $x \in V \cap \text{Span}\{y_j, \ldots, y_n\}$ mit $x \neq 0$. Dann gilt

$$x = \sum_{k=j}^{n} a_k y_k.$$

Folglich gilt

$$R(x) = \frac{(x, Ax)_{\mathbb{C}^n}}{(x,x)_{\mathbb{C}^n}} = \frac{(\sum_{k=j}^n a_k y_k, \sum_{k=j}^n a_k A y_k)_{\mathbb{C}^n}}{(\sum_{k=j}^n a_k y_k, \sum_{k=j}^n a_k y_k)_{\mathbb{C}^n}} = \frac{\sum_{k=j}^n \lambda_k |a_k|^2}{\sum_{k=j}^n |a_k|^2} \geq \lambda_j.$$

Hieraus ergibt sich

$$\max_{x\in V\setminus\{0\}} R(x) \geq \lambda_j.$$

Da diese Ungleichung für alle Unterräume V mit $\dim(V) = j$ gilt, so muss gelten

$$\min_{\substack{V\subset\mathbb{C}^n \\ \dim(V)=j}} \max_{x\in V\setminus\{0\}} R(x) \geq \lambda_j.$$

Um die andere Ungleichung zu zeigen, setzen wir

$$\hat{V} = \text{Span}\{y_1, \ldots, y_j\}$$

mit $\dim(\hat{V}) = j$. Dann gilt

$$\min_{\substack{V\subset\mathbb{C}^n \\ \dim(V)=j}} \max_{x\in V\setminus\{0\}} R(x) \leq \max_{x\in \hat{V}\setminus\{0\}} R(x) \leq \lambda_j,$$

denn $x \in \hat{V} \setminus \{0\}$. Daher hat x die Darstellung

$$x = \sum_{j=1}^{j} a_k y_k.$$

Also ist

$$R(x) = \frac{\sum_{k=1}^{j} \lambda_k |a_k|^2}{\sum_{k=1}^{j} |a_k|^2} \leq \lambda_j.$$

Insgesamt gilt

$$\lambda_j \leq \min_{\substack{V \subset \mathbb{C}^n \\ \dim(V)=j}} \max_{x \in V \setminus \{0\}} R(x) \leq \lambda_j.$$

Der zweite Teil des Satzes ist Hausaufgabe. □

Eine Folgerung des Variationsprinzips von Courant und Fischer ist der folgende Störungssatz.

Satz 2.52. *Es seien $A, E \in \mathbb{C}^{n\times n}$ hermitesch. Für $B \in \{A, E, A+E\}$ bezeichnen wir mit*

$$\lambda_1(B) \leq \lambda_2(B) \leq \ldots \leq \lambda_n(B)$$

die monoton steigenden Eigenwerte von B. Dann gilt

$$\lambda_j(A) + \lambda_1(E) \leq \lambda_j(A+E) \leq \lambda_j(A) + \lambda_n(E) \quad \forall j = 1, \ldots, n.$$

Beweis. Es sei $j \in \{1, \ldots, n\}$. Das Variationsprinzip von Courant und Fischer liefert

$$\begin{aligned}
\lambda_j(A+E) &= \min_{\substack{V \subset \mathbb{C}^n \\ \dim(V)=j}} \max_{x \in V \setminus \{0\}} R(x) \\
&= \min_{\substack{V \subset \mathbb{C}^n \\ \dim(V)=j}} \max_{x \in V \setminus \{0\}} \frac{(x, (A+E)x)_{\mathbb{C}^n}}{(x,x)_{\mathbb{C}^n}} \\
&= \min_{\substack{V \subset \mathbb{C}^n \\ \dim(V)=j}} \max_{x \in V \setminus \{0\}} \left(\frac{(x, Ax)_{\mathbb{C}^n}}{(x,x)_{\mathbb{C}^n}} + \frac{(x, Ex)_{\mathbb{C}^n}}{(x,x)_{\mathbb{C}^n}} \right).
\end{aligned}$$

Da $E \in \mathbb{C}^{n\times n}$ hermitesch ist, so ist

$$\text{Bild}(R_E) = \left\{ \frac{(x, Ex)_{\mathbb{C}^n}}{(x,x)_{\mathbb{C}^n}} \mid x \in \mathbb{C}^n \setminus \{0\} \right\} = [\lambda_1(E), \lambda_n(E)].$$

Daher ist

$$\begin{aligned}
\lambda_j(A+E) &= \min_{\substack{V \subset \mathbb{C}^n \\ \dim(V)=j}} \max_{x \in V \setminus \{0\}} \frac{(x, (A+E)x)_{\mathbb{C}^n}}{(x,x)_{\mathbb{C}^n}} \\
&\leq \min_{\substack{V \subset \mathbb{C}^n \\ \dim(V)=j}} \max_{x \in V \setminus \{0\}} \frac{(x, Ax)_{\mathbb{C}^n}}{(x,x)_{\mathbb{C}^n}} + \lambda_n(E) \\
&= \lambda_j(A) + \lambda_n(E).
\end{aligned}$$

Analog gilt

$$\begin{aligned}\lambda_j(A+E) &= \max_{\substack{V\subset\mathbb{C}^n\\ \dim(V)=n-j+1}} \min_{x\in V\setminus\{0\}} \frac{(x,(A+E)x)_{\mathbb{C}^n}}{(x,x)_{\mathbb{C}^n}} \\ &\geq \max_{\substack{V\subset\mathbb{C}^n\\ \dim(V)=n-j+1}} \min_{x\in V\setminus\{0\}} \frac{(x,Ax)_{\mathbb{C}^n}}{(x,x)_{\mathbb{C}^n}} + \lambda_1(E) \\ &= \lambda_j(A) + \lambda_1(E).\end{aligned}$$

Insgesamt gilt also

$$\lambda_j(A) + \lambda_1(E) \leq \lambda_j(A+E) \leq \lambda_j(A) + \lambda_n(E) \quad \forall j = 1,\dots,n.$$

□

Korollar 2.53. *Es seien $A, E \in \mathbb{C}^{n\times n}$ hermitesch. Dann gilt*

$$|\lambda_k(A+E) - \lambda_k(A)| \leq \|E\|_2 \quad \forall k = 1,\dots,n.$$

2.6 Numerische Behandlung von Eigenwerten

2.6.1 Vektoriterationen

Die einfachste Idee geht auf Richard von Mises zurück:
Ausgehend von einem Startvektor $x^0 \in \mathbb{R}^n \setminus \{0\}$ iterieren wir

$$x^{k+1} = Ax^k, \quad k = 0,1,\dots,$$

wobei $A \in \mathbb{R}^{n\times n}$ symmetrisch ist. In der Literatur heißt dieses Verfahren „Power-method“.

Satz 2.54. *Es sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und $\lambda_n \in \Lambda(A)$ ein einfacher Eigenwert mit*

$$0 \neq |\lambda_n| > |\lambda_{n-1}| \geq \dots \geq |\lambda_1|$$

für alle Eigenwerte von A. Mit anderen Worten ist λ_n der betragsmäßig größte Eigenwert von A. Ist der Startvektor $x^0 \in \mathbb{R}^n \setminus \{0\}$ nicht orthogonal zum Eigenraum von λ_n, das heißt

$$(x^0, y) \neq 0 \quad \forall y \in E_A(\lambda_n) \setminus \{0\},$$

dann konvergiert

$$y^k := \frac{x^k}{\|x^k\|_2} \operatorname{sign}(\lambda_n)^k, \quad \operatorname{sign}(\lambda_n) = \frac{\lambda_n}{|\lambda_n|},$$

bei der Power-method $x^{k+1} = Ax^k$ gegen einen normierten Eigenvektor zum Eigenwert λ_n, das heißt

$$\lim_{k\to\infty} y^k = y^*, \quad Ay^* = \lambda_n y^*, \quad \|y^*\|_2 = 1.$$

Beweis. Da $A \in \mathbb{R}^{n\times n}$ symmetrisch ist, existiert eine Orthonormalbasis $\{y_j\}_{j=1}^n \subset \mathbb{R}^n \setminus \{0\}$ aus Eigenvektoren von A zu Eigenwerten λ_j, das heißt

$$Ay_j = \lambda_j y_j \quad \forall j = 1, \ldots, n.$$

Sei $x^0 \in \mathbb{R}^n \setminus \{0\}$ beliebig aber fest mit

$$(x^0, y_n)_{\mathbb{C}^n} \neq 0.$$

Wir schreiben nun x^0 als Linearkombination von $\{y_j\}_{j=1}^n$:

$$x^0 = \sum_{j=1}^n a_j y_j.$$

Da $(x^0, y_n)_{\mathbb{C}^n} \neq 0$ ist, so muss

$$0 \neq (x^0, y_n)_{\mathbb{C}^n} = \left(\sum_{j=1}^n a_j y_j, y_n \right)_{\mathbb{C}^n} = a_n$$

gelten, also ist $a_n \neq 0$. Andererseits ist

$$x^k = Ax^{k-1} = \ldots = A^k x^0 = A^k \sum_{j=1}^n a_j y_j = \sum_{j=1}^n a_j \underbrace{A^k y_j}_{=\lambda_j^k y_j}$$

und daher

$$x^k = \sum_{j=1}^n a_j \lambda_j^k y_j.$$

Da aber $a_n \neq 0$ und $\lambda_n \neq 0$ ist, so erhalten wir

$$x^k = \sum_{j=1}^{n-1} a_j \lambda_j^k y_j + a_n \lambda_n^k y_n = a_n \lambda_n^k \left(\sum_{j=1}^{n-1} \frac{a_j}{a_n} \left(\frac{\lambda_j}{\lambda_n} \right)^k y_j + y_n \right).$$

Der Term $\left| \frac{\lambda_j}{\lambda_n} \right|$ erfüllt

$$\left| \frac{\lambda_j}{\lambda_n} \right| < 1 \quad \forall j = 1, \ldots, n-1,$$

denn $|\lambda_n| > |\lambda_{n-1}| \geq \ldots \geq |\lambda_1|$. Hieraus folgt

$$\left(\sum_{j=1}^{n-1} \frac{a_j}{a_n} \left(\frac{\lambda_j}{\lambda_n} \right)^k y_j + y_n \right) \to y_n \quad \text{für } k \to \infty$$

und insgesamt

$$y^k = \frac{x^k}{\|x^k\|_2} \operatorname{sign}(\lambda_n)^k \to \pm y_n.$$

□

Bemerkung 2.55. Der Vektor x^k muss nicht konvergieren, denn:
Ist $\lambda_n > 1$, so ist

$$x_k \to \infty.$$

Nachteile:

(i) Die Methode berechnet nur den betragsmäßig größten Eigenwert von A und den zugehörigen Eigenvektor.

(ii) Die Konvergenzgeschwindigkeit hängt im Wesentlichen von der Größe $\left|\frac{\lambda_{n-1}}{\lambda_n}\right|$ ab. Liegen $|\lambda_n|$ und $|\lambda_{n-1}|$ dicht beieinander, dann konvergiert das Verfahren langsam.

Ein besseres Verfahren ist die inverse Vektoriteration[1]. Seine Idee ist wie folgt:
Es sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und $\hat{\lambda} \in \Lambda(A)$ und $\bar{\lambda} \notin \Lambda(A)$ mit

$$|\hat{\lambda} - \bar{\lambda}| < |\lambda - \bar{\lambda}| \quad \forall \lambda \in \Lambda(A) \setminus \{\hat{\lambda}\}.$$

Mit anderen Worten ist

$$|\hat{\lambda} - \bar{\lambda}|^{-1} > |\lambda - \bar{\lambda}|^{-1} \quad \forall \lambda \in \Lambda(A) \setminus \{\hat{\lambda}\}.$$

Also ist $|\hat{\lambda} - \bar{\lambda}|^{-1}$ der betragsmäßig größte Eigenwert der Matrix $(A - \bar{\lambda}I)^{-1}$. Beachte, dass $\bar{\lambda}$ als eine Approximation oder Störung des Eigenwerts $\hat{\lambda} \in \Lambda(A)$ interpretiert werden kann. Da $\bar{\lambda} \notin \Lambda(A)$ gilt, so ist

$$(A - \bar{\lambda}I)$$

regulär, denn $\det(A - \bar{\lambda}I) \neq 0$. Das Verfahren der inversen Vektoriteration lautet

$$(A - \bar{\lambda}I)x^{k+1} = x^k, \quad k = 0, 1, 2, \ldots$$

Satz 2.56 (Konvergenzaussage für inverse Vektoriterationen)**.** *Es sei $A \in \mathbb{R}^{n\times n}$ symmetrisch. Ferner seien $\hat{\lambda} \in \Lambda(A)$, $\bar{\lambda} \notin \Lambda(A)$ mit*

$$|\hat{\lambda} - \bar{\lambda}| < |\lambda - \bar{\lambda}| \quad \forall \lambda \in \Lambda(A) \setminus \{\hat{\lambda}\},$$

und $(\hat{\lambda} - \bar{\lambda})^{-1}$ sei ein einfacher Eigenwert von $(A - \bar{\lambda}I)^{-1}$. Genügt der Startvektor $x^0 \in \mathbb{R}^n \setminus \{0\}$ der Bedingung

$$(x^0, y)_{\mathbb{C}^n} \neq 0 \quad \forall y \in E_{(A-\bar{\lambda}I)^{-1}}((\hat{\lambda} - \bar{\lambda})^{-1}) \setminus \{0\},$$

so konvergiert

$$y^k := \frac{x^k}{\|x^k\|_2} \operatorname{sign}((\hat{\lambda} - \bar{\lambda})^{-1})^k$$

bei der inversen Vektoriteration $(A - \bar{\lambda}I)x^{k+1} = x^k$ gegen einen normierten Eigenvektor von A zum Eigenwert $\hat{\lambda}$, das heißt

$$\lim_{k\to\infty} y^k = y^*, \quad Ay^* = \hat{\lambda}y^*, \quad \|y^*\|_2 = 1.$$

[1] Helmut Wielandt, 1945

Beweis. Das Verfahren der inversen Vektoriteration ist nichts anderes als

$$x^{k+1} = (A - \bar{\lambda} I)^{-1} x^k,$$

also die Power-method angewendet auf die Matrix $(A - \bar{\lambda} I)^{-1}$. Alle Voraussetzungen des vorherigen Satzes sind erfüllt und somit konvergiert $y^k = \frac{x^k}{\|x^k\|_2} \operatorname{sign}(\lambda_n)^k$ gegen einen normierten Eigenvektor von $(A - \bar{\lambda} I)^{-1}$, das heißt

$$\lim_{k \to \infty} y^k = y^*, \quad (A - \bar{\lambda})^{-1} y^* = (\hat{\lambda} - \bar{\lambda})^{-1} y^*, \quad \|y^*\|_2 = 1.$$

Hieraus folgt

$$y^* = (A - \bar{\lambda} I)(A - \bar{\lambda} I)^{-1} y^* = (A - \bar{\lambda} I)(\hat{\lambda} - \bar{\lambda})^{-1} y^*,$$

was äquivalent ist zu

$$(\hat{\lambda} - \bar{\lambda}) y^* = (A - \bar{\lambda} I) y^*,$$

also schließlich

$$\hat{\lambda} y^* - \bar{\lambda} y^* = A y^* - \bar{\lambda} y^*$$

und somit die Behauptung. □

Vorteile der Vektoriteration:

(i) Theoretisch ist es möglich, alle Eigenwerte von A mithilfe der inversen Vektoriteration zu bestimmen.

(ii) Ist $\bar{\lambda} \approx \hat{\lambda}$, dann ist $|\hat{\lambda} - \bar{\lambda}|$ sehr klein im Vergleich mit

$$|\hat{\lambda} - \lambda| \quad \forall \lambda \in \Lambda(A) \setminus \{\hat{\lambda}\}$$

und somit konvergiert das Verfahren recht schnell gegen $\hat{\lambda}$, denn in diesem Fall gilt

$$\max_{\lambda \in \Lambda(A) \setminus \{\hat{\lambda}\}} \left| \frac{\hat{\lambda} - \bar{\lambda}}{\lambda - \bar{\lambda}} \right| \ll 1.$$

Nachteil des Verfahrens ist:

(i) Man muss die Inverse $(A - \bar{\lambda} I)^{-1}$ bestimmen.

Beispiel 2.57. Wir betrachten die Matrix

$$A = \begin{pmatrix} -1 & 3 \\ -2 & 4 \end{pmatrix}.$$

Die Eigenwerte sind $\lambda_1 = 1$ und $\lambda_2 = 2$. Sei $\bar{\lambda} = 1 - \varepsilon$ eine Approximation von λ_1. Somit ist

$$(A - \bar{\lambda} I) = \begin{pmatrix} -2 + \varepsilon & 3 \\ -2 & 3 + \varepsilon \end{pmatrix}.$$

Diese Matrix ist fast singulär, denn

$$(A - \bar{\lambda} I)^{-1} = \frac{1}{\varepsilon(\varepsilon+1)} \begin{pmatrix} 3+\varepsilon & -3 \\ 2 & -2+\varepsilon \end{pmatrix}.$$

Bei $\varepsilon \to 0$ ist $(A - \bar{\lambda} I)^{-1}$ singulär (nicht regulär). Durch Normierung kürzt sich der Faktor $\frac{1}{\varepsilon(\varepsilon+1)}$ heraus. Folglich ist die Berechnung einer normierten Lösung von

$$(A - \bar{\lambda} I)x = b$$

gut konditioniert.

Bemerkung 2.58. Ist $\bar{\lambda} \approx \hat{\lambda}$ gewählt, so ist $(A - \bar{\lambda} I)$ fast singulär. Wir rechnen aber nicht den Eigenvektor $\hat{x}$ aus, sondern nur $\frac{\hat{x}}{\|\hat{x}\|_2}$, also nur dessen Richtung (vgl. Deufelhard/Hohmann, Bsp. 2.33).

2.6.2 Das QR-Verfahren zur Eigenwertbestimmung

In Numerik I haben wir die folgende Aussage gezeigt.

Satz 2.59 (Existenz und Eindeutigkeit der QR-Zerlegung)**.** *Es sei $A \in \mathbb{R}^{n\times n}$ regulär. Dann besitzt A eine QR-Zerlegung*

$$A = QR$$

mit einer orthogonalen Matrix $Q \in \mathbb{R}^{n\times n}$ und einer oberen Dreiecksmatrix $R \in \mathbb{R}^{n\times n}$. Ist

$$A = \tilde{Q}\tilde{R}$$

eine andere QR-Zerlegung, so existiert eine Diagonalmatrix

$$D = \begin{pmatrix} \pm 1 & & 0 \\ & \ddots & \\ 0 & & \pm 1 \end{pmatrix},$$

so dass

$$Q = \tilde{Q}D \quad \textit{und} \quad \tilde{R} = DR.$$

Bemerkung 2.60. Die QR-Zerlegung lässt sich zum Beispiel durch Householder-Spiegelungen bestimmen.

Das QR-Verfahren zur Eigenwertbestimmung sieht wie folgt aus.

Algorithmus 2.1 QR-Verfahren

(S0) Setze $A_0 := A$ und $k = 0$.
(S1) Bestimme eine QR-Zerlegung von A_k

$$A_k = Q_k R_k.$$

(S2) Setze $A_{k+1} = R_k Q_k$, $k = k+1$, und gehe zu (S1).

Lemma 2.61. *Es sei $A \in \mathbb{R}^{n\times n}$ regulär. Dann ist der QR-Algorithmus durchführbar, und die erzeugten Matrizen A_k erfüllen die folgenden Bedingungen:*

(i) Alle A_k sind ähnlich zu A.

(ii) Ist A symmetrisch, so ist auch A_k symmetrisch für alle $k \in \mathbb{N}$.

Beweis. Da $A \in \mathbb{R}^{n\times n}$ regulär ist, ist der Algorithmus durchführbar.

(i) Sei $k = 0, 1, \ldots$. Dann gilt

$$\begin{aligned} A_k &= Q_k R_k, \\ A_{k+1} &= R_k Q_k. \end{aligned}$$

Damit ist

$$Q_k A_{k+1} Q_k^T = Q_k R_k Q_k Q_k^T = Q_k R_k = A_k,$$

also sind A_{k+1} und A_k ähnlich für alle $k \in \mathbb{N}$. Insgesamt sind also alle A_k ähnlich zu A.

(ii) Mit (i) folgt

$$Q_k A_{k+1} Q_k^T = A_k$$

und daher

$$A_{k+1} = Q_k^T A_k Q_k.$$

Ist A_k symmetrisch, so ist

$$A_{k+1}^T = (Q_k^T A_k Q_k)^T = Q_k^T A_k^T Q_k = Q_k^T A_k Q_k = A_{k+1}.$$

Mit anderen Worten ist A_{k+1} ebenso symmetrisch.

Insgesamt ergibt sich somit die Behauptung. □

Satz 2.62 (Konvergenz des QR-Verfahrens)**.** *Sei $A \in \mathbb{R}^{n\times n}$ regulär und diagonalisierbar mit reellen Eigenwerten, so dass gilt*

$$|\lambda_1| > |\lambda_2| > |\lambda_3| > \ldots > |\lambda_n| > 0.$$

Die Inverse der Matrix

$$T = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix} \in \mathbb{R}^{n\times n},$$

mit orthonormalen Eigenvektoren y_j zu λ_j, besitze ohne Pivotisierung die LR-Zerlegung

$$T^{-1} = LR, \quad L = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ * & & 1 \end{pmatrix} \in \mathbb{R}^{n\times n}, \quad R = \begin{pmatrix} * & & * \\ & \ddots & \\ & & * \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

Dann gilt für das QR-Verfahren:

$$A_k = S_k U S_k + \mathcal{O}(q^k) \quad \textit{für } k \to \infty$$

mit

$$q = \max_{j\in\{1,\dots,n-1\}} \left|\frac{\lambda_{j+1}}{\lambda_j}\right| \in (0,1),$$
$$S_k = \operatorname{diag}\{S_{k_1},\dots,S_{k_n}\}, \quad S_{k_j} \in \{\pm 1\} \quad \forall j = 1,\dots,n$$

und

$$U = \begin{bmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

Bemerkung 2.63. Wir schreiben $f_k \in \mathcal{O}(q^k)$, bzw. $f_k = \mathcal{O}(q^k)$ für $k \to \infty$ genau dann, wenn

$$\exists\, C > 0 \quad \forall k \in \mathbb{N}: \quad \|f_k\|_2 \le C q^k.$$

Da $\lim\limits_{k\to\infty} q^k = 0$, so folgt auch $\lim\limits_{k\to\infty} f_k = 0$.

Beweis. Für die Eigenvektormatrix T nehmen wir eine QR-Zerlegung vor:

$$T = Q\hat{R} \tag{2.9}$$

mit einer Orthogonalmatrix $Q \in \mathbb{R}^{n\times n}$ und einer oberen Dreiecksmatrix $\hat{R} \in \mathbb{R}^{n\times n}$. Wir zeigen:

$$A_k = S_k(\hat{R}D\hat{R}^{-1})S_k + \mathcal{O}(q^k) \quad \text{für } k \to \infty,$$

wobei S_k Vorzeichenmatrizen sind und $D = \operatorname{diag}(\lambda_1,\dots,\lambda_n)$. Dann folgt die Aussage mit $U = \hat{R}D\hat{R}^{-1}$.
Schritt 1: Laut Voraussetzung gilt

$$T^{-1} = LR \quad \text{mit} \quad L = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ * & & 1 \end{pmatrix} \quad \text{und} \quad R = \begin{pmatrix} * & & * \\ & \ddots & \\ 0 & & * \end{pmatrix}. \tag{2.10}$$

Wir setzen

$$L_k := D^k L D^{-k} = \begin{pmatrix} \lambda_1^k & & 0 \\ & \ddots & \\ 0 & & \lambda_n^k \end{pmatrix} \begin{pmatrix} 1 & & & \\ l_{21} & \ddots & & \\ \vdots & & \ddots & \\ l_{n1} & \dots & l_{n,n-1} & 1 \end{pmatrix} \begin{pmatrix} \lambda_1^{-k} & & 0 \\ & \ddots & \\ 0 & & \lambda_n^{-k} \end{pmatrix}. \tag{2.11}$$

Daraus ergibt sich

$$(L_k)_{ij} = \left(\frac{\lambda_i}{\lambda_j}\right)^k l_{ij} = \begin{cases} 1 & \text{für } i = j \\ 0 & \text{für } i < j \\ \left(\frac{\lambda_i}{\lambda_j}\right)^k l_{ij} & \text{für } i > j \end{cases}.$$

Es gilt für $i > j$

$$\left|\frac{\lambda_i}{\lambda_j}\right| \leq \max_{l \in \{1,\ldots,n-1\}} \left|\frac{\lambda_{l+1}}{\lambda_l}\right| = q \in (0,1),$$

denn $|\lambda_1| > |\lambda_2| > |\lambda_3| > \ldots > |\lambda_n| > 0$. Hieraus folgt

$$L_k = D^k L D^{-k} = I + \mathcal{O}(q^k) \quad \text{für } k \to \infty. \tag{2.12}$$

Für jedes $k \in \mathbb{N}$ ist L_k regulär, und somit ist

$$\hat{R} L_k \in \mathbb{R}^{n \times n}$$

ebenso regulär. Demzufolge hat $\hat{R} L_k$ eine QR-Zerlegung

$$\hat{R} L_k = \hat{Q}_k \hat{R}_k \tag{2.13}$$

mit einer Orthogonalmatrix $\hat{Q}_k \in \mathbb{R}^{n \times n}$ und einer oberen Dreiecksmatrix $\hat{R}_k \in \mathbb{R}^{n \times n}$. Folglich gilt

$$\begin{aligned} \hat{Q}_k \hat{R}_k &= \hat{R} L_k \\ &\overset{(2.12)}{=} \hat{R}(I + \mathcal{O}(q^k)) \quad \text{für } k \to \infty \\ &= I\hat{R} + \mathcal{O}(q^k) \quad \text{für } k \to \infty. \end{aligned}$$

Insbesondere gilt

$$\lim_{k \to \infty} \hat{Q}_k \hat{R}_k = I\hat{R}.$$

Da die QR-Zerlegung stetig von der gewählten Matrix abhängt, liefert die obige Aussage

$$\begin{aligned} \hat{Q}_k &= I + \mathcal{O}(q^k) \quad \text{für } k \to \infty, \\ \hat{R}_k &= \hat{R} + \mathcal{O}(q^k) \quad \text{für } k \to \infty, \end{aligned}$$

das heißt insbesondere

$$\begin{aligned} \lim_{k \to \infty} \hat{Q}_k &= I, \\ \lim_{k \to \infty} \hat{R}_k &= \hat{R}. \end{aligned}$$

2. Schritt: Das QR-Verfahren lautet

$$\begin{cases} A_1 := A \\ A_k := Q_k R_k \\ A_{k+1} := R_k Q_k \end{cases}.$$

Außerdem ist

$$TDT^{-1} = A \quad \Leftrightarrow \quad D = T^{-1}AT.$$

Folglich gilt

$$\begin{aligned} A^2 &= AA = TDT^{-1}TDT^{-1} = TD^2T^{-1} \\ &\vdots \\ A^k &= TD^kT^{-1} \end{aligned}$$

und daher

$$A^k \overset{(2.9)-(2.10)}{=} Q\hat{R}D^kLR = Q\hat{R}D^kLD^{-k}D^kR \overset{(2.11)}{=} Q\hat{R}L_kD^kR \overset{(2.13)}{=} Q\hat{Q}_k\hat{R}_kD^kR.$$

Insgesamt ist

$$A^k = \underbrace{Q\hat{Q}_k}_{\substack{\text{orthogonale} \\ \text{Matrix}}} \underbrace{\hat{R}_kD^kR}_{\substack{\text{obere} \\ \text{Dreiecksmatrix}}} \qquad \forall k \in \mathbb{N}.$$

Wir haben also eine QR-Zerlegung für A^k gefunden. Wir berechnen nun eine andere QR-Zerlegung für A^k anhand des QR-Verfahrens. Dazu setzen wir

$$\begin{aligned} Q_{1\cdots k} &:= Q_1Q_2 \cdot \ldots \cdot Q_k \in \mathbb{R}^{n\times n} \quad \text{orthogonal,} \\ R_{k\cdots 1} &:= R_kR_{k-1} \cdot \ldots \cdot R_1 \in \mathbb{R}^{n\times n} \quad \text{obere Dreiecksmatrix.} \end{aligned}$$

Nun gilt

$$A_{k+1} = R_kQ_k = Q_k^TQ_kR_kQ_k = Q_k^TA_kQ_k \quad \forall k \in \mathbb{N}.$$

Also ist

$$\begin{aligned} A_{k+1} &= Q_k^TQ_{k-1}^T \cdot \ldots \cdot Q_1^TAQ_1 \cdot \ldots \cdot Q_k \\ &= (Q_1 \cdot \ldots \cdot Q_k)^TA(Q_1 \cdot \ldots \cdot Q_k) \\ &= Q_{1\cdots k}^TAQ_{1\cdots k} \end{aligned}$$

und somit

$$Q_{1\cdots k}A_{k+1} = AQ_{1\cdots k}.$$

Das heißt

$$Q_{1\cdots k-1}A_k = AQ_{1\cdots k-1} \quad \forall k \in \mathbb{N}.$$

Somit ist

$$\begin{aligned} Q_{1\cdots k}R_{k\cdots 1} &= Q_1 \cdot \ldots \cdot Q_k R_k \cdot \ldots \cdot R_1 \\ &= Q_{1\cdots k-1}A_k R_{k-1\cdots 1} \\ &= AQ_{1\cdots k-1}R_{k-1\cdots 1} \quad \forall k \in \mathbb{N}. \end{aligned}$$

Wir erhalten

$$\begin{aligned} Q_{1\cdots 2}R_{2\cdots 1} &= AQ_1R_1 = AA_1 = AA = A^2 \\ Q_{1\cdots 3}R_{3\cdots 1} &= AQ_{1\cdots 2}R_{2\cdots 1} = A^3 \\ &\vdots \\ Q_{1\cdots k}R_{k\cdots 1} &= A^k \quad \forall k \in \mathbb{N}. \end{aligned}$$

Das heißt wir haben eine zweite QR-Zerlegung für A^k gefunden:

$$A^k = Q_{1\cdots k}R_{k\cdots 1} \quad \forall k \in \mathbb{N}.$$

Insgesamt gilt

$$\begin{aligned} A^k &= Q\hat{Q}_k\hat{R}_k D^k R \quad \forall k \in \mathbb{N}, \\ A^k &= Q_{1\cdots k}R_{k\cdots 1} \quad \forall k \in \mathbb{N}. \end{aligned}$$

Somit liefert der Satz über die Eindeutigkeit der QR-Zerlegung eine Vorzeichenmatrix

$$S_{k+1} = \operatorname{diag}(\pm 1, \ldots \pm 1), \quad k \in \mathbb{N},$$

so dass gilt

$$\begin{aligned} Q_{1\cdots k} &= Q\hat{Q}_k S_{k+1} \quad \forall k \in \mathbb{N}, \\ R_{k\cdots 1} &= S_{k+1}\hat{R}_k D^k R \quad \forall k \in \mathbb{N}. \end{aligned}$$

Diese Identitäten liefern zum Einen

$$Q_{1\cdots k-1}^T = S_k \hat{Q}_{k-1}^T Q^T \tag{2.14}$$

und zum Anderen

$$Q_1 \cdot \ldots \cdot Q_k = Q_{1\cdots k} = Q\hat{Q}_k S_{k+1}$$

und somit

$$\begin{aligned} Q_k &= Q_{k-1}^T \cdot \ldots \cdot Q_1^T Q\hat{Q}_k S_{k+1} \\ &= (Q_1 \cdot \ldots \cdot Q_{k-1})^T Q\hat{Q}_k S_{k+1} \\ &= Q_{1\cdots k-1}^T Q\hat{Q}_k S_{k+1}. \end{aligned} \tag{2.15}$$

Aus (2.14)-(2.15) ergibt sich

$$\begin{aligned} Q_k &= Q_{1\cdots k-1}^T Q \hat{Q}_k S_{k+1} \\ &= S_k \hat{Q}_{k-1}^T Q^T Q \hat{Q}_k S_{k+1} \\ &= S_k \hat{Q}_{k-1}^T \hat{Q}_k S_{k+1}. \end{aligned}$$

Analog gilt

$$\begin{aligned} R_k &= R_{k\cdots 1} R_{k-1\cdots 1}^{-1} \\ &= S_{k+1} \hat{R}_k D^k R R^{-1} (D^{-1})^{k-1} \hat{R}_{k-1}^{-1} S_k \\ &= S_{k+1} \hat{R}_k D \hat{R}_{k-1}^{-1} S_k. \end{aligned}$$

Insgesamt gilt

$$A_k = Q_k R_k = S_k \hat{Q}_{k-1}^T \hat{Q}_k S_{k+1} S_{k+1} \hat{R}_k D \hat{R}_{k-1}^{-1} S_k.$$

Aus Schritt 1 erhalten wir

$$A_k = S_k \hat{R} D \hat{R}^{-1} S_k + \mathcal{O}(q^k) \quad \text{für } k \to \infty.$$

□

Bemerkung 2.64.

(i) Die Voraussetzung an T ist a priori nicht nachprüfbar, weil man die Eigenvektoren nicht kennt. Man wendet das Verfahren an und hofft, dass alles gut geht.

(ii) Der Satz gilt nicht mehr für komplexe Eigenwerte.

(iii) Entscheidend für die Geschwindigkeit des Verfahrens ist

$$q = \max_{j \in \{1,\ldots,n-1\}} \left| \frac{\lambda_{j+1}}{\lambda_j} \right| \in (0,1).$$

Liegen die Eigenwerte $|\lambda_{j+1}|$ und $|\lambda_j|$ dicht beieinander, so ist

$$q \approx 1$$

und wir erhalten eine langsame Konvergenz.

Korollar 2.65. *Unter den Voraussetzungen des vorherigen Satzes gilt*

$$\lim_{k\to\infty} (A_k)_{jj} = \lambda_j \quad \forall j = 1, \ldots, n.$$

Beweis. Im vorherigen Satz haben wir gezeigt:

$$A_k = S_k U S_k + \mathcal{O}(q^k) \quad \text{für } k \to \infty$$

mit $S_k = \operatorname{diag}\{\pm 1, \dots, \pm 1\}$, $q \in (0,1)$ und

$$U = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}.$$

Hieraus ergibt sich

$$S_k U S_k = \begin{pmatrix} S_{k1} & & 0 \\ & \ddots & \\ 0 & & S_{kn} \end{pmatrix} \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \begin{pmatrix} S_{k1} & & 0 \\ & \ddots & \\ 0 & & S_{kn} \end{pmatrix} = \begin{pmatrix} \lambda_1 S_{k1}^2 & & * \\ & \ddots & \\ 0 & & \lambda_n S_{kn}^2 \end{pmatrix}$$
$$= \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}.$$

Mit $q^k \to 0$ für $k \to \infty$ ergibt sich insgesamt

$$\lim_{k\to\infty} (A_k)_{jj} = \lambda_j \quad \forall j = 1, \dots, n.$$

□

Kapitel 3

Das GMRES-Verfahren

In diesem Abschnitt untersuchen wir das GMRES-Verfahren (**G**eneralized **M**inimal **Res**idual Method) zur Lösung von linearen Gleichungssystemen der Form

$$Ax = b$$

mit einer Matrix $A \in \mathbb{R}^{n\times n}$ und $b \in \mathbb{R}^n \setminus \{0\}$. Dieses Problem lässt sich zum Beispiel durch das CG-Verfahren lösen, wie wir bereits in Numerik I gelernt haben. Als Voraussetzung für die Konvergenz benötigt das CG-Verfahren symmetrische und positiv definite Matrizen. Im Gegensatz dazu benötigt das GMRES-Verfahren diese Annahmen nicht.

3.1 Arnoldi-Prozess

Die folgende Definition kennen wir aus Numerik I.

Definition 3.1 (Krylov-Raum)**.** Es seien $A \in \mathbb{R}^{n\times n}$ und $b \in \mathbb{R}^n$. Dann heißt die Menge

$$\mathcal{K}_m(A, b) = \text{Span}\{b, Ab, A^2b, \ldots, A^{m-1}b\}$$

Krylov-Raum.

Es ist klar, dass

$$\mathcal{K}_m \subset \mathcal{K}_{m+1}$$

für alle $m \in \mathbb{N}$ gilt. Als Nächstes wiederholen wir das Gram-Schmidt-Verfahren. Sei $\{a_j\}_{j=1}^m \subset \mathbb{R}^n \setminus \{0\}$ linear unabhängig. Wir setzen

$$\hat{q}_j = a_j - \sum_{i=1}^{j-1} (q_i, a_j) q_i \quad \forall j = 1, \ldots, m,$$
$$q_j = \frac{\hat{q}_j}{\|\hat{q}_j\|_2}.$$

Insbesondere gilt für $j = 1$:

$$\hat{q}_1 = a_1, \quad q_1 = \frac{a_1}{\|a_1\|_2}.$$

Dann folgt

$$(q_i, q_j) = \delta_{ij} \quad \forall i, j \in \{1, \dots, m\}.$$

Definition 3.2. Es seien $A \in \mathbb{R}^{n\times n}$ und $b \in \mathbb{R}^n \setminus \{0\}$. Der Arnoldi-Prozess ist definiert wie folgt:
(S0) $q_1 := \frac{b}{\|b\|_2}$ und setze $m = 1$.
(S1) $\hat{q}_{m+1} := Aq_m - \sum_{k=1}^{m} (Aq_m, q_k) q_k$
(S2) Abbruchkriterium:
1: **if** $\hat{q}_{m+1} = 0$ **then**
2: STOP
3: **end if**
(S3) Setze $q_{m+1} = \frac{\hat{q}_{m+1}}{\|\hat{q}_{m+1}\|_2}$ (Normierung), $m = m + 1$, und gehe zu (S1).

Beim Arnoldi-Prozess wird also geprüft, ob mit nochmaliger Anwendung der Matrix A nach $b, Ab, \dots$ noch eine neue Dimension ins Spiel kommt. Im Folgenden bezeichnen wir mit $m^* \in \{1, \dots, n\}$ den Abbruchindex, das heißt den kleinsten Index mit

$$\hat{q}_{m^*+1} = 0 \quad \Leftrightarrow \quad Aq_{m^*} \in \text{Span}\{q_1, \dots, q_{m^*}\}.$$

Lemma 3.3. *Die vom Arnoldi-Prozess erzeugten Vektoren $q_1, \dots, q_{m^*}$ sind paarweise orthogonal, und es gilt*

$$\text{Span}\{q_1, \dots, q_m\} = \text{Span}\{q_1, \dots, q_{m-1}, Aq_{m-1}\} = \mathcal{K}_m(A, b) \quad \forall m \in \{1, \dots, m^*\}.$$

Ist A regulär, so gilt

$$x^* = A^{-1}b \in \mathcal{K}_{m^*}(A, b).$$

Beweis. Es sei $m \in \{1, \dots, m^*\}$. Nach Definition sind $q_1, \dots, q_m$ paarweise orthogonal und

$$\text{Span}\{q_1, \dots, q_m\} = \text{Span}\{q_1, \dots, q_{m-1}, Aq_{m-1}\}. \tag{3.1}$$

Mit Induktion zeigen wir nun die andere Identität

$$\text{Span}\{q_1, \dots, q_m\} = \mathcal{K}_m(A, b). \tag{3.2}$$

Induktionsanfang $m = 1$: $\text{Span}\{q_1\} = \text{Span}\{b\} = \mathcal{K}_1(A, b)$.

Induktionsannahme: Es gelte

$$\text{Span}\{q_1, \dots, q_m\} = \mathcal{K}_m(A, b)$$

für ein $m \in \{1, \dots, m^* - 1\}$. Wir zeigen nun, dass die Aussage für $m + 1$ gilt.

Dazu betrachten wir

$$Aq_m = A\left(\sum_{i=1}^{m} \lambda_i A^{i-1} b\right) = \sum_{i=1}^{m} \lambda_i A^i b \in \mathcal{K}_{m+1}(A,b).$$

Nun gilt

$$\text{Span}\{q_1,\ldots,q_{m+1}\} \overset{(3.1)}{=} \text{Span}\{q_1,\ldots,q_m,Aq_m\} \subset \mathcal{K}_{m+1}(A,b)$$

und die Gleichheit folgt unmittelbar aus Dimensionsgründen:

$$m+1 = \dim(\text{Span}\{q_1,\ldots,q_{m+1}\}) \leq \dim(\mathcal{K}_{m+1}(A,b)) = m+1.$$

Es sei nun A regulär. Nach Definition ist m^* der Abbruchindex, und somit haben wir

$$Aq_{m^*} \in \text{Span}\{q_1,\ldots,q_{m^*}\} = \mathcal{K}_{m^*}(A,b).$$

Außerdem gilt für $j \in \{1,\ldots,m^*-1\}$:

$$Aq_j \in \text{Span}\{q_1,\ldots,q_j,Aq_j\} = \text{Span}\{q_1,\ldots,q_{j+1}\} = \mathcal{K}_{j+1}(A,b) \subset \mathcal{K}_{m^*}(A,b).$$

Daraus folgt

$$Aq_j \in \mathcal{K}_{m^*}(A,b) \quad \forall j \in \{1,\ldots,m^*\}$$

und somit

$$A(\mathcal{K}_{m^*}(A,b)) \overset{(3.2)}{=} A(\text{Span}\{q_1,\ldots,q_{m^*}\}) \subset \mathcal{K}_{m^*}(A,b).$$

Das heißt, die Funktion

$$F : \mathcal{K}_{m^*}(A,b) \to K_{m^*}(A,b), \quad F(x) = Ax,$$

ist wohldefiniert und bijektiv, da A regulär ist. Daher gilt

$$x^* = A^{-1}b \in K_{m^*}(A,b),$$

da $b \in K_{m^*}(A,b)$. □

3.1.1 Matrixversion des Arnoldi-Prozesses

Wir führen nun die Matrixversion des Arnoldi-Prozesses wie folgt ein.

Definition 3.4. Es seien $A \in \mathbb{R}^{n\times n}$ und $b \in \mathbb{R}^n \setminus \{0\}$. Ferner seien $q_1,\ldots,q_{m^*}$ die vom Arnoldi-Prozess erzeugten Vektoren mit dem Abbruchindex $m^* \in \{1,\ldots,n\}$. Wir definieren

$$\begin{aligned} h_{km} &:= (Aq_m, q_k), \quad k = 1,\ldots,m, \\ h_{m+1,m} &:= \|\hat{q}_{m+1}\|_2, \quad m = 1,\ldots,m^*-1. \end{aligned}$$

Nach Definition des Arnoldi-Prozesses haben wir für jedes $m \in \{1, \ldots, m^* - 1\}$:

$$h_{m+1,m} q_{m+1} = \|\hat{q}_{m+1}\|_2 \, q_{m+1} = \hat{q}_{m+1} = Aq_m - \sum_{k=1}^{m} \underbrace{(Aq_m, q_k)}_{=h_{km}} q_k.$$

Folglich gilt

$$Aq_m = \sum_{k=1}^{m+1} h_{km} q_k.$$

Also sind

$$\begin{aligned} Aq_1 &= h_{11}q_1 + h_{21}q_2 \\ Aq_2 &= h_{12}q_1 + h_{22}q_2 + h_{32}q_3 \\ &\vdots \\ Aq_m &= h_{1m}q_1 + \ldots + h_{m+1,m}q_{m+1} \end{aligned}$$

und wir erhalten

$$A \underbrace{\begin{bmatrix} q_1 & \cdots & q_m \end{bmatrix}}_{\in \mathbb{R}^{n \times m}} = \underbrace{\begin{bmatrix} q_1 & \cdots & q_{m+1} \end{bmatrix}}_{\in \mathbb{R}^{n \times (m+1)}} \underbrace{\begin{bmatrix} h_{11} & \cdots & \cdots & h_{1m} \\ h_{21} & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & \ddots & h_{mm} \\ 0 & & & h_{m+1,m} \end{bmatrix}}_{\in \mathbb{R}^{(m+1) \times m}}.$$

Bei $m = m^*$ ist laut Definition $\hat{q}_{m^*+1} = 0$ und somit haben wir

$$A \underbrace{\begin{bmatrix} q_1 & \cdots & q_{m^*} \end{bmatrix}}_{\in \mathbb{R}^{n \times m^*}} = \underbrace{\begin{bmatrix} q_1 & \cdots & q_{m^*} \end{bmatrix}}_{\in \mathbb{R}^{n \times m^*}} \underbrace{\begin{bmatrix} h_{11} & \cdots & \cdots & h_{1m^*} \\ h_{21} & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ 0 & & h_{m^*,m^*-1} & h_{m^*m^*} \end{bmatrix}}_{\in \mathbb{R}^{m^* \times m^*}}.$$

Somit haben wir den folgenden Satz bewiesen:

Satz 3.5. *Es seien $A \in \mathbb{R}^{n \times n}$ und $b \in \mathbb{R}^n \setminus \{0\}$. Ferner seien $q_1, \ldots, q_{m^*}$ und h_{ij} die vom Arnoldi-Prozess erzeugten Größen. Dann gilt*

$$\begin{aligned} AQ_m &= Q_{m+1} H_m \quad \forall m = 1, \ldots, m^* - 1, \\ AQ_{m^*} &= Q_{m^*} H_{m^*} \end{aligned}$$

mit

$$Q_m := \begin{bmatrix} q_1 & \cdots & q_m \end{bmatrix} \in \mathbb{R}^{n\times m},$$

$$H_m := \begin{bmatrix} h_{11} & \cdots & \cdots & h_{1m} \\ h_{21} & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & \ddots & h_{mm} \\ 0 & & & h_{m+1,m} \end{bmatrix} \in \mathbb{R}^{(m+1)\times m},$$

$$H_{m^*} := \begin{bmatrix} h_{11} & \cdots & \cdots & h_{1m^*} \\ h_{21} & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ 0 & & h_{m^*,m^*-1} & h_{m^*m^*} \end{bmatrix} \in \mathbb{R}^{m^*\times m^*}.$$

Definition 3.6. Eine Matrix $A \in \mathbb{R}^{n\times n}$ der Gestalt

$$A = \begin{pmatrix} * & \cdots & \cdots & * \\ * & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ 0 & & * & * \end{pmatrix}$$

heißt *obere Hessebergmatrix*.

Korollar 3.7. *Gilt $m^* = n$, das heißt der Arnoldi-Prozess bricht nicht vorzeitig ab, so gilt*

$$Q_n^T A Q_n = H_n.$$

Mit anderen Worten ist A in diesem Fall ähnlich zur Hessebergmatrix H_n.

3.2 Definition des GMRES-Verfahrens

Grundidee: Zur Lösung des Gleichungssystems

$$Ax = b$$

werden schrittweise die Minimierungsaufgaben

$$\min_{x\in\mathcal{K}_m(A,b)} \|Ax - b\|_2, \quad m = 1,\ldots,m^*$$

gelöst. Man sucht also $x_m \in \mathcal{K}_m(A,b)$, so dass

$$\|Ax_m - b\|_2 = \min_{x\in\mathcal{K}_m(A,b)} \|Ax - b\|_2$$

gilt. Mit anderen Worten minimiert x_m das Residuum

$$r = Ax - b \quad \text{auf } \mathcal{K}_m(A,b) = \text{Span}\{b, Ab, \ldots, A^{m-1}b\}.$$

Daher kommt der Name

Generalized Minimal Residual Method

als Verallgemeinerung von MinRes-Verfahren (diese funktionieren für symmetrische Matrizen).

Lemma 3.8. *Es seien $A \in \mathbb{R}^{n\times n}$ und $b \in \mathbb{R}^n \setminus \{0\}$. Mit den Bezeichnungen aus dem Arnoldi-Prozess gelten für die vom GMRES-Verfahren erzeugten Vektoren $x_m \in \mathcal{K}_m(A,b)$, $m = 1, \ldots, m^*$, die Darstellungen*

$$x_m = Q_m z_m, \quad m = 1, \ldots, m^*,$$

und $x_m \in \mathcal{K}_m(A,b)$ löst genau dann

$$\min_{x\in\mathcal{K}_m(A,b)} \|Ax - b\|_2,$$

wenn $z_m \in \mathbb{R}^m$ die Aufgabe

$$\min_{z\in\mathbb{R}^m} \|H_m z - c_m\|_2$$

löst mit

$$c_m = \begin{pmatrix} \|b\|_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{\min\{m^*, m+1\}}.$$

Beweis.

(i) $x_m \in \mathcal{K}_m(A,b)$ heißt $x_m \in \text{Span}\{q_1, \ldots, q_m\}$, das heißt

$$x_m = \sum_{i=1}^{m} q_i z_i =: Q_m z_m.$$

(ii) Sei $m \in \{1, \ldots, m^* - 1\}$. Dann gilt

$$\begin{aligned} \|Ax_m - b\|_2 = \|AQ_m z_m - b\|_2 &= \|Q_{m+1}H_m z_m - b\|_2 \\ &= \left\|Q_{m+1}H_m z_m - Q_{m+1}Q_{m+1}^T b\right\|_2 \\ &= \left\|Q_{m+1}(H_m z_m - Q_{m+1}^T b)\right\|_2 \\ &= \|Q_{m+1}(H_m z_m - c_m)\|_2 \\ &= \|H_m z_m - c_m\|_2, \end{aligned}$$

denn als Matrix mit orthonormalen Spaltenvektoren hat Q_{m+1} die Isometrieeigenschaft $\|Q_{m+1}y\|_2 = \|y\|_2$.
Nebenrechnungen:

$$Q_{m+1}^T b = \begin{bmatrix} q_1^T \\ \vdots \\ q_{m+1}^T \end{bmatrix} b = \begin{bmatrix} q_1^T \\ \vdots \\ q_{m+1}^T \end{bmatrix} q_1 \|b\|_2 = \begin{bmatrix} \|b\|_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = c_m,$$

weil $(q_i, q_k) = \delta_{ik}$. Hieraus folgt auch

$$Q_{m+1}Q_{m+1}^T b = \begin{bmatrix} q_1 & \cdots & q_{m+1} \end{bmatrix} \begin{bmatrix} \|b\|_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = q_1 \|b\|_2 = b.$$

Für $m = m^*$ bekommen wir analog

$$\|Ax_{m^*} - b\|_2 = \|Q_{m^*}(H_{m^*}z_{m^*} - c_{m^*})\|_2 = \|H_{m^*}z_{m^*} - c_{m^*}\|_2 .$$

Insgesamt gilt

$$\|Ax_m - b\|_2 = \|H_m z_m - c_m\|_2 \quad \forall m \in \{1, \ldots, m^*\}.$$

□

3.3 Vorgehensweise zur Lösung der Minimierungsprobleme beim GMRES-Verfahren

Es sei $m \in \{1, \ldots, m^* - 1\}$. Wir wissen bereits:

$$\min_{x \in \mathcal{K}_m(A,b)} \|Ax - b\|_2 \quad \Leftrightarrow \quad \min_{z \in \mathbb{R}^m} \|H_m z - c_m\|_2$$

mit

$$H_m = \begin{bmatrix} h_{11} & \cdots & \cdots & h_{1m} \\ h_{21} & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & \ddots & h_{mm} \\ 0 & & & h_{m+1,m} \end{bmatrix} \in \mathbb{R}^{(m+1)\times m}, \quad c_m = \begin{pmatrix} \|b\|_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{m+1}.$$

Wir bestimmen die volle QR-Zerlegung der Matrix H_m

$$H_m = T_m \begin{bmatrix} R_m \\ 0^T \end{bmatrix}$$

mit einer Orthogonalmatrix $T_m \in \mathbb{R}^{(m+1)\times(m+1)}$ und einer oberen Dreiecksmatrix $R_m \in \mathbb{R}^{m\times m}$, sowie $0 \in \mathbb{R}^m$. Mit der QR-Zerlegung lässt sich das obige Minimierungsproblem einfach lösen:

(i) $$\|H_m z - c_m\|_2 = \left\| T_m \begin{bmatrix} R_m \\ 0^T \end{bmatrix} z - c_m \right\|_2 = \left\| T_m \begin{bmatrix} R_m z \\ 0^T \end{bmatrix} - T_m T_m^T c_m \right\|_2$$
$$= \left\| \begin{bmatrix} R_m z \\ 0^T \end{bmatrix} - T_m^T c_m \right\|_2 .$$

(ii) Partitionierung:
$$T_m^T c_m = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}.$$

(iii) Löse
$$R_m z = b_1 \quad \text{(Rückwärtssubstitution)}.$$
Aus der daraus resultierenden Lösung z_m erhalten wir
$$x_m = Q_m z_m.$$

Für $m = m^*$ macht man analog wie oben

$$H_{m^*} = \begin{bmatrix} h_{11} & \cdots & \cdots & h_{1m^*} \\ h_{21} & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ 0 & & h_{m^*,m^*-1} & h_{m^*m^*} \end{bmatrix} = T_{m^*} R_{m^*}$$

mit einer Orthogonalmatrix $T_{m^*} \in \mathbb{R}^{m^*\times m^*}$ und einer oberen Dreiecksmatrix $R_{m^*} \in \mathbb{R}^{m^*\times m^*}$.

3.3.1 Detaillierte Beschreibung der QR-Zerlegungen

Die obigen QR-Zerlegungen der Matrizen H_m lassen sich sehr effizient erledigen, indem jeweils die vorangegangene Zerlegung verwendet wird.

(i) Initialzerlegung $m = 1$ (Annahme $m^* \neq 1$):
Für $m = 1$, $m^* \neq 1$, gilt
$$H_1 = \begin{bmatrix} h_{11} \\ h_{21} \end{bmatrix}$$
mit
$$h_{11} = (Aq_1, q_1) = (Ab, b)\frac{1}{\|b\|_2^2}$$
$$h_{21} = \|\hat{q}_2\|_2 \neq 0 \quad \text{(da } m^* \neq 1 \text{ ist)}.$$

Wir suchen eine Orthogonalmatrix $T_1 \in \mathbb{R}^{2\times 2}$ mit

$$T_1 = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}, \quad H_1 = T_1 \begin{bmatrix} r_{11} \\ 0 \end{bmatrix} \quad \text{bzw.} \quad T_1^T H_1 = \begin{bmatrix} r_{11} \\ 0 \end{bmatrix}$$

mit $r_{11} \neq 0$. Hier können wir die Orthogonalmatrix $T_1 \in \mathbb{R}^{2\times 2}$ mittels Householder-Spiegelungen bestimmen.

(ii) $m-1 \to m \in \{1, \ldots, m^* - 1\}$:

Es seien nun T_{m-1}, R_{m-1} bereits bestimmt, das heißt

$$H_{m-1} = T_{m-1} \begin{bmatrix} R_{m-1} \\ 0^T \end{bmatrix}$$

mit einer Orthogonalmatrix $T_{m-1} \in \mathbb{R}^{m\times m}$ und einer oberen Dreiecksmatrix $R_{m-1} \in \mathbb{R}^{(m-1)\times(m-1)}$. Wir machen den Ansatz

$$\tilde{T}_m = \left[\begin{array}{c|c} T_{m-1} & 0 \\ \hline 0^T & 1 \end{array}\right] \in \mathbb{R}^{(m+1)\times(m+1)}.$$

Es ist klar, dass $\tilde{T}_m \in \mathbb{R}^{(m+1)\times(m+1)}$ orthogonal ist. Laut Definition wissen wir auch

$$\mathbb{R}^{(m+1)\times m} \ni H_m = \left[\begin{array}{ccc|c} h_{11} & \cdots & \cdots & h_{1m} \\ h_{21} & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ 0 & & \ddots & h_{mm} \\ \hline 0^T & & & h_{m+1,m} \end{array}\right] = \left[\begin{array}{c|c} & h_{1m} \\ H_{m-1} & \vdots \\ & h_{mm} \\ \hline 0^T & h_{m+1,m} \end{array}\right].$$

Daraus folgt

$$\tilde{T}_m^T H_m = \left[\begin{array}{c|c} T_{m-1}^T & 0 \\ \hline 0^T & 1 \end{array}\right] \left[\begin{array}{c|c} & h_{1m} \\ H_{m-1} & \vdots \\ & h_{mm} \\ \hline 0^T & h_{m+1,m} \end{array}\right] = \left[\begin{array}{c|c} T_{m-1}^T H_{m-1} & T_{m-1}^T \begin{pmatrix} h_{1m} \\ \vdots \\ h_{mm} \end{pmatrix} \\ \hline 0^T & h_{m+1,m} \end{array}\right]$$

$$= \left[\begin{array}{c|c} R_{m-1} & r_{1m} \\ & \vdots \\ \hline 0^T & r_{mm} \\ \hline 0^T & r_{m+1,m} \end{array}\right]$$

mit

$$\begin{pmatrix} r_{1m} \\ \vdots \\ r_{mm} \end{pmatrix} = T_{m-1}^T \begin{pmatrix} h_{1m} \\ \vdots \\ h_{mm} \end{pmatrix} \quad \text{und} \quad r_{m+1,m} = h_{m+1,m}.$$

Leider ist $r_{m+1,m} = h_{m+1,m} = \|\hat{q}_{m+1}\|_2 \neq 0$, da $m < m^*$ ist. Deshalb erfüllt der Ansatz mit $\tilde{T}_m$ noch nicht die gewünschte Eigenschaft. Aus diesem Grund suchen wir eine 2×2 Orthogonalmatrix

$$W_m = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix},$$

so dass gilt

$$W_m \begin{pmatrix} r_{mm} \\ r_{m+1,m} \end{pmatrix} = \begin{pmatrix} * \\ 0 \end{pmatrix}.$$

Insgesamt haben wir

$$\left[\begin{array}{c|c|c} I_{m-1} & 0 & 0 \\ \hline 0^T & & \\ 0^T & \multicolumn{2}{c}{W_m} \end{array}\right] \left[\begin{array}{c|c} T_{m-1}^T & 0 \\ \hline 0^T & 1 \end{array}\right] H_m = \left[\begin{array}{c|c|c} I_{m-1} & 0 & 0 \\ \hline 0^T & & \\ 0^T & \multicolumn{2}{c}{W_m} \end{array}\right] \left[\begin{array}{c|c} R_{m-1} & r_{1m} \\ & \vdots \\ \hline 0^T & r_{mm} \\ \hline 0^T & r_{m+1,m} \end{array}\right]$$

$$= \begin{bmatrix} R_m \\ 0^T \end{bmatrix},$$

mit einer oberen Dreiecksmatrix $R_m \in \mathbb{R}^{m \times m}$, denn es ist

$$W_m \begin{pmatrix} r_{mm} \\ r_{m+1,m} \end{pmatrix} = \begin{pmatrix} * \\ 0 \end{pmatrix}.$$

Zusammengefasst erhalten wir

$$H_m = T_m \begin{bmatrix} R_m \\ 0^T \end{bmatrix}, \quad T_m^T = \begin{bmatrix} I_{m-1} & 0 & 0 \\ 0^T & & \\ 0^T & \multicolumn{2}{c}{W_m} \end{bmatrix} \tilde{T}_m^T.$$

(iii) Der Fall $m = m^*$:

Hier ist $H_{m^*} \in \mathbb{R}^{m^* \times m^*}$ also eine quadratische Matrix. Es gilt dann

$$H_{m^*} = \left[\begin{array}{c|c} H_{m^*-1} & \begin{matrix} h_{1m^*} \\ \vdots \\ h_{m^*m^*} \end{matrix} \end{array}\right]$$

und daraus folgt

$$T_{m^*-1}^T H_{m^*} = \left[\begin{array}{c|c} T_{m^*-1}^T H_{m^*-1} & \begin{matrix} r_{1m^*} \\ \vdots \\ r_{m^*m^*} \end{matrix} \end{array}\right] = \left[\begin{array}{c|c} R_{m^*-1} & \begin{matrix} r_{1m^*} \\ \vdots \\ r_{m^*m^*} \end{matrix} \\ \hline 0^T & \end{array}\right],$$

mit

$$\begin{pmatrix} r_{1m^*} \\ \vdots \\ r_{m^*m^*} \end{pmatrix} = T_{m^*-1}^T \begin{pmatrix} h_{1m^*} \\ \vdots \\ h_{m^*m^*} \end{pmatrix}.$$

Insgesamt erhalten wir

$$T_{m^*} = T_{m^*-1}.$$

Kapitel 4

Numerische Lösung gewöhnlicher DGL

4.1 Einführung zu gewöhnlichen Differentialgleichungen

Definition 4.1. Es sei $f : [a,b] \times \mathbb{R}^n \to \mathbb{R}^n$ und $y_0 \in \mathbb{R}^n$. Ein Anfangswertproblem für ein System von n gewöhnlichen Differentialgleichungen erster Ordnung ist von der Form:

$$y'(t) = f(t, y(t)) \quad \forall t \in (a,b), \tag{DGL}$$
$$y(a) = y_0. \tag{AB}$$

Gesucht ist also eine differenzierbare Funktion $y : [a,b] \to \mathbb{R}^n$, die das Anfangswertproblem (DGL)-(AB) erfüllt. Die Funktion heißt Lösung von (DGL)-(AB).

Bemerkung 4.2.

(i) Die Aufgabe (DGL) heißt gewöhnliche Differentialgleichung, weil die gesuchte Funktion y nur von einer Variablen $t \in [a,b]$ abhängt.

(ii) Die Schreibweise

$$y'(t) = f(t, y(t))$$

ist äquivalent zu

$$\begin{aligned} y_1'(t) &= f_1(t, y_1(t), y_2(t), \ldots, y_n(t)) \\ y_2'(t) &= f_2(t, y_1(t), y_2(t), \ldots, y_n(t)) \\ &\vdots \\ y_n'(t) &= f_n(t, y_1(t), y_2(t), \ldots, y_n(t)). \end{aligned}$$

(iii) Viele Anwendungsprobleme lassen sich als Anfangswertproblem (DGL)-(AB) auffassen, wie beispielsweise die Newtonsche-Bewegungsgleichung, die Berechnung der Flugbahn, das Wachstum einer Population, und vieles mehr.

Satz 4.3 (Existenz und Eindeutigkeit). *Es sei $f : [a,b] \times \mathbb{R}^n \to \mathbb{R}^n$ stetig und erfülle die Lipschitz-Bedingung:*
Es existiere eine Konstante $L > 0$, so dass gilt

$$\|f(t,v) - f(t,w)\|_2 \leq L \|v - w\|_2 \quad \forall t \in [a,b] \text{ und } v, w \in \mathbb{R}^n.$$

Dann hat das Anfangswertproblem (DGL)-(AB) *mit $y_0 \in \mathbb{R}^n$ genau eine stetig differenzierbare Lösung $y \in C^1([a,b], \mathbb{R}^n)$. Für Lösungen $y, \tilde{y} : [a,b] \to \mathbb{R}^n$ von*

$$\begin{cases} y'(t) = f(t, y(t)) & \text{in } (a,b) \\ y(a) = y_0 \end{cases}, \qquad \begin{cases} \tilde{y}'(t) = f(t, \tilde{y}(t)) & \text{in } (a,b) \\ \tilde{y}(a) = \tilde{y}_0 \end{cases}$$

gilt die Abschätzung

$$\|y(t) - \tilde{y}(t)\|_2 \leq e^{L(t-a)} \|y_0 - \tilde{y}_0\| \quad \forall t \in [a,b].$$

Bemerkung 4.4.

(i) Dieser Satz geht zurück auf Picard-Lindelöf. Der Beweis ist zum Beispiel im Buch von Heuser „Gewöhnliche Differentialgleichungen“ zu finden.

(ii) Die Abschätzung

$$\|y(t) - \tilde{y}(t)\|_2 \leq e^{L(t-a)} \|y_0 - \tilde{y}_0\| \quad \forall t \in [a,b]$$

bedeutet, dass die Lösung y stetig von dem Anfangswert abhängt.

Lemma 4.5. *Es sei $y : [a,b] \to \mathbb{R}^n$ eine stetig differenzierbare Lösung von* (DGL)-(AB). *Dann erfüllt y die Integralgleichung*

$$y(t) = y_0 + \int_a^t f(s, y(s)) \, \mathrm{d}s$$

für alle $t \in [a,b]$.

Beweis. Laut Voraussetzung erfüllt $y : [a,b] \to \mathbb{R}^n$ die folgende Gleichung:

$$y'(t) = f(t, y(t)) =: g(t) \quad \forall t \in [a,b].$$

Demzufolge ist $y : [a,b] \to \mathbb{R}^n$ die Stammfunktion von $g : [a,b] \to \mathbb{R}^n$. Die Funktion g ist außerdem laut Annahme stetig. Somit liefert der Hauptsatz der Differential- und Integralrechnung:

$$\int_a^t g(s) \, \mathrm{d}s = y(t) - y(a) \quad \forall t \in [a,b].$$

Wir erhalten also insgsamt

$$y(t) = y(a) + \int_a^t g(s) \, \mathrm{d}s = y_0 + \int_a^t f(s, y(s)) \, \mathrm{d}s \quad \forall t \in [a,b].$$

□

4.2 Bekannte numerische Lösungsverfahren

Um das Anfangswertproblem (DGL)-(AB) numerisch zu lösen, müssen wir zunächst das Zeitintervall $[a, b]$ diskretisieren. Am einfachsten verwenden wir die äquidistante Zerlegung

$$a := t_0 < t_1 < \ldots < t_N =: b,$$
$$t_j = a + jh \quad \text{mit } h = \frac{b-a}{N} \quad \text{(Schrittweite)}.$$

Gesucht sind nun Approximationen

$$u_j \approx y(t_j) \quad \forall j = 0, \ldots, N.$$

4.2.1 Das explizite Euler-Verfahren (Polygonzug-Verfahren)

Ist $y : [a, b] \to \mathbb{R}^n$ eine Lösung von (DGL)-(AB), so gilt

$$y(t_j) = y(t_{j-1} + h) = y(t_{j-1}) + y'(t_{j-1})h + \mathcal{O}(h) = y(t_{j-1}) + f(t_{j-1}, y(t_{j-1}))h + \mathcal{O}(h).$$

Daraus ergibt sich das Verfahren

$$\begin{cases} u_j = u_{j-1} + f(t_{j-1}, u_{j-1})h, & j = 1, \ldots, N, \\ u_0 = y_0. \end{cases}$$

Beachte, dass es sich beim expliziten Euler-Verfahren um die Vorwärtsdifferenz handelt:

$$y'(t_{j-1}) \approx \frac{y(t_j) - y(t_{j-1})}{h} \approx \frac{u_j - u_{j-1}}{h}.$$

4.2.2 Das implizite Euler-Verfahren

Beim impliziten Euler-Verfahren erhalten wir

$$y(t_{j-1}) = y(t_j - h) = y(t_j) + y'(t_j)(-h) + \mathcal{O}(h) = y(t_j) - hf(t_j, y(t_j)) + \mathcal{O}(h).$$

Das ist äquivalent zu

$$y(t_j) = y(t_{j-1}) + hf(t_j, y(t_j)) - \mathcal{O}(h).$$

Daraus ergibt sich

$$\begin{cases} u_j = u_{j-1} + hf(t_j, u_j), & j = 1, \ldots, N, \\ u_0 = y_0. \end{cases}$$

Es handelt sich also um eine Rückwärtsdifferenz:

$$y'(t_j) \approx \frac{y(t_j) - y(t_{j-1})}{h} \approx \frac{u_j - u_{j-1}}{h}.$$

Im Gegensatz zum expliziten Euler-Verfahren ist das implizite Euler-Verfahren schwieriger zu lösen, denn der gesuchte Vektor u_j taucht sowohl links als auch rechts im Argument der Funktion f auf. Dieses Verfahren ist im Allgemeinen besser (stabiler) als die explizite Version.

4.2.3 Trapezmethode

Wir verwenden die Integralgleichung

$$y(t) = y_0 + \int_a^t f(s, y(s))\,\mathrm{d}s \quad \forall t \in [a, b].$$

Diese Integralgleichung liefert die folgende Darstellung

$$\begin{aligned} y(t_{j+1}) = y_0 + \int_a^{t_{j+1}} f(s, y(s))\,\mathrm{d}s &= y_0 + \int_a^t f(s, y(s))\,\mathrm{d}s + \int_t^{t_{j+1}} f(s, y(s))\,\mathrm{d}s \\ &= y(t_j) + \int_t^{t_{j+1}} f(s, y(s))\,\mathrm{d}s \quad \forall j = 0, \ldots, N-1. \end{aligned}$$

Wird das Integral mit der Trapezregel approximiert, so ergibt sich

$$u_{j+1} = u_j + \frac{h}{2}[f(t_j, u_j) + f(t_{j+1}, u_{j+1})] \quad \forall j = 0, \ldots, N-1.$$

4.2.4 Einfache Beispiele

Beispiel 4.6. Betrachte das Anfangswertproblem

$$\begin{cases} y'(t) = -2ty^2(t), & t \in (0, 1), \\ y(0) = 1. \end{cases}$$

Es ist leicht zu sehen, dass

$$y : [0, 1] \to \mathbb{R}, \quad y(t) = \frac{1}{t^2 + 1}$$

die exakte Lösung ist ($y \in C^1([0, 1], \mathbb{R})$). Wir verwenden die Schrittweite $h = 0.1$.
Explizites Euler-Verfahren:

$$\begin{aligned} &u_0 = y_0 = 1, \\ &u_1 = u_0 + hf(t_0, u_0) = 1 + 0.1 \cdot 0 = 1, \\ &u_2 = u_1 + hf(t_1, u_1) = 1 + 0.1 \cdot (-2 \cdot 0.1 \cdot 1^2) = 0.98, \\ &\quad \vdots \end{aligned}$$

Implizites Euler-Verfahren:

$$\begin{aligned} &u_0 = y_0 = 1, \\ &u_1 = u_0 + hf(t_1, u_1) = 1 + 0.1 \cdot (-2 \cdot 0.1 \cdot u_1^2) = 1 - 0.02u_1^2, \\ &\quad \Rightarrow \quad u_1^2 + 50u_1 = 500, \\ &\quad \Rightarrow \quad u_1 = 0.98076211 \quad \text{(die andere Nullstelle kommt nicht in Betracht)} \\ &u_2 = u_1 + hf(t_2, u_2) \\ &\quad \Rightarrow \quad u_2 = 0.94503822, \\ &\quad \vdots \end{aligned}$$

Die exakte Lösung lautet:

$$
\begin{aligned}
&y(t_0) = y_0 = 1,\\
&y(t_1) = (0.1^2 + 1)^{-1} = 0.99009901,\\
&y(t_2) = (0.2^2 + 1)^{-1} = 0.96153846.
\end{aligned}
$$

Beispiel 4.7. Betrachte das Anfangswertproblem

$$
\begin{cases} y'(t) = \lambda y(t) \quad \forall t \in (0,1), \\ y(0) = 1 \end{cases}
$$

mit $\lambda \in \mathbb{R}$. Die exakte Lösung ist $y(t) = e^{\lambda t}$. Wir wählen die Schrittweite $h = 0.1$.
Explizites Euler-Verfahren:

$$
\begin{aligned}
&u_0 = 1,\\
&u_1 = u_0 + hf(t_0, u_0) = u_0 + h\lambda u_0 = 1 + \lambda h,\\
&u_2 = u_1 + hf(t_1, u_1) = 1 + \lambda h + h\lambda(1 + \lambda h) = (1 + \lambda h)^2,\\
&\quad \vdots\\
&u_j = (1 + \lambda h)^j.
\end{aligned}
$$

Implizites Euler-Verfahren:

$$
\begin{aligned}
&u_0 = 1,\\
&u_1 = u_0 + hf(t_1, u_1) = 1 + \lambda h u_1 \quad \Rightarrow \quad u_1 = \frac{1}{(1 - \lambda h)},\\
&u_2 = u_1 + hf(t_2, u_2) = (1 - \lambda h)^{-1} + \lambda h u_2 \quad \Rightarrow \quad u_2 = \frac{1}{(1 - \lambda h)^2},\\
&\quad \vdots\\
&u_j = \frac{1}{(1 - \lambda h)^j}.
\end{aligned}
$$

	$\lambda = 1, \quad h = 0.1$			$\lambda = -21, \quad h = 0.1$		
t	Explizit	Implizit	Exakt	Explizit	Implizit	Exakt
0	1.0	1.0	1.0	1.0	1.0	1.0
0.1	0.9	0.909	0.9048	-1.1	0.3225	0.1225
0.2	0.81	0.826	0.8187	1.21	0.104	0.015
$\vdots$						
1	0.349	0.385	0.3679	2.6	$1.2 \cdot 10^{-5}$	$7.6 \cdot 10^{-10}$

4.3 Theorie der expliziten Einschrittverfahren

Im Folgenden sei $y \in C^1([a,b],\mathbb{R}^n)$ die eindeutige Lösung des Anfangswertproblems (DGL)-(AB):

$$\begin{cases} y'(t) = f(t,y(t)) & \forall t \in [a,b], \\ y(0) = y_0. \end{cases}$$

Wir betrachten eine allgemeine Zerlegung (Diskretisierung) des Zeitintervalls $[a,b]$ wie folgt:

$$a = t_0 < t_1 < t_2 < \ldots < t_N = b$$

mit der Schrittweite

$$\begin{aligned} h_j &= t_{j+1} - t_j, \quad j = 0,\ldots,N-1, \\ h_{\max} &= \max_{0 \le j \le N-1} h_j. \end{aligned}$$

Definition 4.8. Ein explizites Einschrittverfahren zur Approximation des Anfangswertproblems (DGL)-(AB) ist von der Gestalt

$$\begin{cases} u_{j+1} = u_j + h_j \varphi(t_j, u_j, h_j), \quad j = 0,\ldots,N-1, \\ u_0 = y_0, \end{cases} \tag{EV}$$

mit einer Verfahrensfunktion

$$\varphi : [a,b] \times \mathbb{R}^n \times \mathbb{R}_0^+ \to \mathbb{R}^n \quad (\mathbb{R}_0^+ := \{x \in \mathbb{R} \mid x \ge 0\}).$$

Beispiel 4.9. Das explizite Euler-Verfahren

$$\begin{cases} u_{j+1} = u_j + h_j f(t_j, u_j), \quad j = 0,\ldots,N-1, \\ u_0 = y_0. \end{cases}$$

ist ein explizites Einschrittverfahren mit der Verfahrensfunktion

$$\varphi(t,u,h) = f(t,u) \quad \forall t \in [a,b], u \in \mathbb{R}^n, h \in \mathbb{R}_0^+.$$

Bemerkung 4.10. Die Approximation u_{j+1} hängt nur von u_j ab. Deshalb heißt diese Methode explizites Einschrittverfahren.

Bemerkung 4.11. Das implizite Euler-Verfahren lässt sich nicht als ein explizites Einschrittverfahren auffassen.

Im nächsten Abschnitt untersuchen wir implizite Einschrittverfahren gemeinsam mit Mehrschrittverfahren.

Definition 4.12 (Verfahrensfehler)**.** Es sei $\varphi : [a,b] \times \mathbb{R}^n \times \mathbb{R}_0^+ \to \mathbb{R}^n$ eine Verfahrensfunktion für das explizite Einschrittverfahren (EV). Dann heißt

$$\eta(t,h) = \underbrace{y(t) + h\varphi(t,y(t),h)}_{\text{Verfahrensvorschrift}} - y(t+h) \quad \text{für } t \in [a,b] \text{ und } h \in [0, b-t]$$

lokaler Fehler im Punkt $(t+h, y(t+h))$ bezüglich der Schrittweite h.

Definition 4.13 (Konsistenz)**.** Das explizite Einschrittverfahren (EV) zur Lösung von (DGL)-(AB) besitzt die *Konsistenzordnung* $p \geq 1$, falls für den lokalen Fehler die Ungleichung

$$\|\eta(t,h)\|_2 \leq \hat{C} h^{p+1} \quad \forall t \in [a,b],\ h \in [0, b-t]$$

mit einer von t und h unabhängigen Konstanten $\hat{C} > 0$ erfüllt ist.

Die Konsistenzordnung ist entscheidend für die Konvergenz von (EV). Zusammen mit der Lipschitz-Annahme für die Verfahrensfunktion φ liefert die Konsistenzordnung die Konvergenz mit p als Konvergenzgeschwindigkeit.
Lipschitz-Annahme: Es existiere eine Konstante $L > 0$, so dass gilt

$$\|\varphi(t,u,h) - \varphi(t,v,h)\|_2 \leq L \|u - v\|_2 \quad \forall u, v \in \mathbb{R}^n,\ t \in [a,b],\ h \in [0, b-t].$$

Beachte, dass L unabhängig von u, t und h ist.

Satz 4.14 (Konvergenz des expliziten Einschrittverfahrens)**.** *Das explizite Einschrittverfahren besitze die Konsistenzordnung $p \geq 1$, und die Verfahrensfunktion $\varphi : [a,b] \times \mathbb{R}^n \times \mathbb{R}_0^+ \to \mathbb{R}^n$ erfülle die Lipschitz-Annahme. Dann konvergiert* (EV) *gegen die Lösung $y \in C^1([a,b], \mathbb{R}^n)$ des Anfangswertproblems* (DGL)-(AB) *mit der Konvergenzrate p. Genauer gilt*

$$\max_{0 \leq j \leq N} \|u_j - y(t_j)\|_2 \leq C h_{\max}^p$$

mit der Konstanten

$$C := \frac{\hat{C}}{L}(e^{L(b-a)} - 1).$$

Lemma 4.15. *Für reelle Zahlen $L > 0$, $\xi \geq 0$, $h_j > 0$ und $\kappa \geq 0$ sei*

$$\xi_{j+1} \leq (1 + Lh_j)\xi_j + h_j\kappa \quad \forall j = 0, \ldots, N-1$$

erfüllt. Dann gilt

$$\xi_j \leq \frac{e^{Lx_j} - 1}{L}\kappa + e^{Lx_j}\xi_0 \quad \text{mit } x_j = \sum_{k=0}^{j-1} h_k,\ x_0 := 0,\ x_1 := h_0$$

für alle $j = 0, \ldots, N$.

Beweis durch Induktion.

Induktionsanfang $j = 0$: Es ist

$$x_0 = 0,$$

und daraus folgt

$$\xi_0 = \underbrace{\frac{e^{Lx_0} - 1}{L}}_{=0} \underbrace{\kappa}_{\geq 0} + \underbrace{e^{Lx_0}}_{=1} \xi_0 = \xi_0.$$

Induktionsannahme: Die Aussage gelte für alle $j = 0, \ldots, m$ mit $m \in \{0, \ldots, N-1\}$.

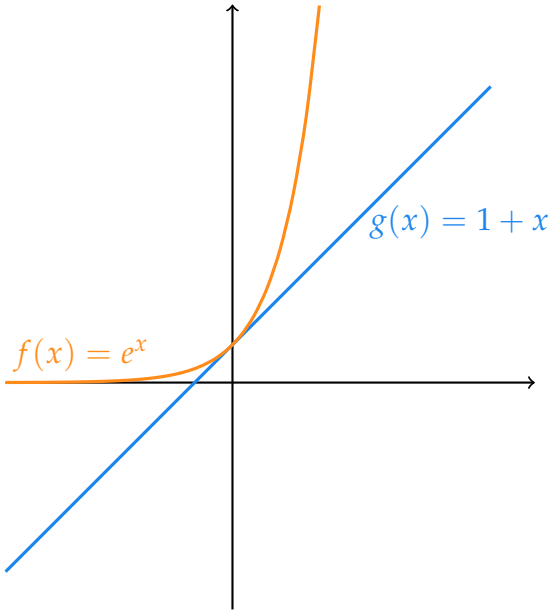
Dann gilt

$$\begin{aligned}
\xi_{m+1} &\leq (1 + Lh_m)\xi_m + h_m\kappa \\
&\leq (1 + Lh_m)\left[\frac{e^{Lx_m} - 1}{L}\kappa + e^{Lx_m}\xi_0\right] + h_m\kappa \\
&= \frac{(1 + Lh_m)e^{Lx_m} - (1 + Lh_m)}{L}\kappa + (1 + Lh_m)e^{Lx_m}\xi_0 + h_m\kappa \\
&= \frac{(1 + Lh_m)e^{Lx_m} - 1}{L}\kappa + (1 + Lh_m)e^{Lx_m}\xi_0 \\
&\leq \frac{e^{Lh_m}e^{Lx_m} - 1}{L}\kappa + e^{Lh_m}e^{Lx_m}\xi_0 \\
&= \frac{e^{L(x_m + h_m)} - 1}{L}\kappa + e^{L(x_m + h_m)}\xi_0 \\
&= \frac{e^{x_{m+1}} - 1}{L}\kappa + e^{x_{m+1}}\xi_0.
\end{aligned}$$

Dabei haben wir bei der letzten Abschätzung verwendet, dass

$$1 + x \leq e^x$$

für alle $x \in \mathbb{R}$ gilt.



□

Beweis des Satzes. Wir setzen

$$e_j := u_j - y(t_j), \quad j = 0, \dots, N-1.$$

Laut Definition ist

$$\begin{cases} u_{j+1} = u_j + h_j \varphi(t_j, u_j, h_j), & j = 0, \dots, N-1, \\ u_0 = y_0, \end{cases}$$

und

$$\eta(t_j, h_j) = y(t_j) + h_j \varphi(t_j, y(t_j), h_j) - y(\underbrace{t_j + h_j}_{t_{j+1}}), \quad j = 0, \dots, N-1.$$

Somit gilt für jedes $j = 0, \dots, N-1$:

$$\begin{aligned} e_{j+1} &= u_{j+1} - y(t_{j+1}) \\ &= u_j + h_j \varphi(t_j, u_j, h_j) - y(t_{j+1}) \\ &= u_j - y(t_j) + h_j(\varphi(t_j, u_j, h_j) - \varphi(t_j, y(t_j), h_j)) + \eta(t_j, h_j). \end{aligned}$$

Mit der Dreiecksungleichung ergibt sich nun

$$\left\|e_{j+1}\right\|_2 \leq \left\|u_j - y(t_j)\right\|_2 + h_j \left\|\varphi(t_j, u_j, h_j) - \varphi(t_j, y(t_j), h_j)\right\|_2 + \left\|\eta(t_j, h_j)\right\|_2.$$

Die Konsistenzordnung und die Lipschitz-Annahme liefern

$$\begin{aligned} \left\|e_{j+1}\right\|_2 &\leq \left\|e_j\right\|_2 + h_j L \left\|u_j - y(t_j)\right\|_2 + \hat{C} h_j^{p+1} \\ &= (1 + L h_j) \left\|e_j\right\|_2 + \hat{C} h_j^{p+1} \quad \forall j = 0, \dots, N-1. \end{aligned}$$

Mit $h_j \leq h_{\max}$ folgt

$$\underbrace{\left\|e_{j+1}\right\|_2}_{=:\xi_{j+1}} \leq (1 + L h_j) \underbrace{\left\|e_j\right\|_2}_{=:\xi_j} + \underbrace{\hat{C} h_{\max}^p}_{=:\kappa} h_j \quad \forall j = 0, \dots, N-1$$

und das vorherige Lemma liefert

$$\left\|e_j\right\|_2 \leq \left(\frac{e^{L x_j} - 1}{L}\right) \hat{C} h_{\max}^p + e^{L x_j} \xi_0.$$

Insgesamt folgt

$$\left\|e_j\right\|_2 \leq \underbrace{\left(\frac{e^{L(b-a)} - 1}{L}\right) \hat{C}}_{=:C} h_{\max}^p \quad \forall j = 0, \dots, N-1,$$

denn $e_0 = u_0 - y(0) = u_0 - y_0 = 0$ und

$$\begin{aligned} x_j = \sum_{k=0}^{j-1} h_k = h_{j-1} + h_{j-2} + \ldots + h_0 &= (t_j - t_{j-1}) + (t_{j-1} - t_{j-2}) + \ldots + (t_1 - t_0) \\ &= t_j - t_0 \\ &\leq b - a. \end{aligned}$$

□

4.3.1 Explizites Einschrittverfahren der Konsistenzordnung p = 1

Unter milden Voraussetzungen hat das explizite Euler-Verfahren die Konsistenzordnung $p = 1$. Zur Erinnerung ist das explizite Euler-Verfahren wie folgt definiert:

$$\begin{cases} u_{j+1} = u_j + h_j f(t_j, u_j), \quad j = 1, \ldots, N-1, \\ u_0 = y_0. \end{cases}$$

Satz 4.16. *Erfüllt die Lösung des Anfangswertproblems* (DGL)-(AB) *die Regularität*

$$y \in C^2([a,b], \mathbb{R}^n),$$

so besitzt das explizite Euler-Verfahren die Konsistenzordnung $p = 1$, das heißt es existiert ein $\hat{C} > 0$ mit

$$\|\eta(t,h)\|_2 \leq \hat{C} h^2$$

für alle $t \in [a,b]$ und $h \in [0, b-t]$.

Beweis. Da $y \in C^2([a,b], \mathbb{R}^n)$ ist, so liefert der Satz von Taylor für jedes $t \in [a,b]$ und $h \in [0, b-t]$:

$$y_i(t+h) = y_i(t) + y_i'(t)h + y_i''(\xi_i)\frac{h^2}{2} \quad \forall i = 1, \ldots, n$$

mit Zwischenstellen $\xi_i \in [a,b]$ für alle $i = 1, \ldots, n$. Dies ist äquivalent zu

$$y(t+h) = y(t) + y'(t)h + (y_i''(\xi_i))_{i=1}^n \frac{h^2}{2}. \tag{4.1}$$

Somit ergibt sich für jeden lokalen Verfahrensfehler:

$$\begin{aligned} \eta(t,h) &= y(t) + h\varphi(t, y(t), h) - y(t+h) \\ &\overset{\text{Euler}}{=} y(t) + hf(t, y(t)) - y(t+h) \\ &\overset{\text{(DGL)}}{=} y(t) + hy'(t) - y(t+h) \\ &\overset{\text{(4.1)}}{=} y(t) + hy'(t) - y(t) - y'(t)h - (y''(\xi))_{i=1}^n \frac{h^2}{2} \\ &= -(y''(\xi_i))_{i=1}^n \frac{h^2}{2}. \end{aligned}$$

Insgesamt folgt nun

$$\|\eta(t,h)\|_2 \leq \hat{C}h^2 \quad \forall t \in [a,b],\ h \in [0, b-t]$$

mit

$$\hat{C} = \frac{1}{2} \max_{a \leq t \leq b} \left\|y''(t)\right\|_\infty .$$

□

Korollar 4.17. *Es existiere eine Konstante $L > 0$, so dass gilt*

$$\|f(t,u) - f(t,v)\|_2 \leq L \|u - v\|_2 \quad \forall t \in [a,b],\ u, v \in \mathbb{R}^n.$$

Ferner erfülle die Lösung des Anfangswertproblems (DGL)-(AB) *die Regularität*

$$y \in C^2([a,b], \mathbb{R}^n).$$

Dann konvergiert das explizite Euler-Verfahren mit der Konvergenzrate $p = 1$. Genauer gilt

$$\max_{0 \leq j \leq N} \left\|u_j - y(t_j)\right\|_2 \leq C h_{\max}$$

mit

$$C = \frac{\hat{C}}{L}(e^{L(b-a)} - 1) = \frac{1}{2L} \max_{a \leq t \leq b} \left\|y''(t)\right\|_\infty (e^{L(b-a)} - 1).$$

Beweis. Gemäß dem Satz zur Konvergenz des expliziten Einschrittverfahrens implizieren die Lipschitz-Annahme und die Konsistenzordnung $p = 1$ die Konvergenz mit Rate p. □

4.3.2 Explizite Einschrittverfahren höherer Konsistenzordnung

Unter Differenzierbarkeitsannahme lassen sich explizite Einschrittverfahren mit einer höheren Konsistenzordnung $p > 1$ konstruieren.

(i) **Das modifizierte Euler-Verfahren:**
Hier verwenden wir die Verfahrensfunktion

$$\varphi(t,u,h) := f(t + \frac{h}{2}, u + \frac{h}{2} f(t,u)) \quad \forall t \in [a,b],\ u \in \mathbb{R}^n,\ h \in [0, b-t].$$

Die Verfahrensvorschrift lautet

$$\begin{cases} u_0 = y_0, \\ u_{j+\frac{1}{2}} = u_j + \frac{h_j}{2} f(t_j, u_j), \quad \forall j = 0, \ldots, N-1, \\ u_{j+1} = u_j + h_j f(t_j + \frac{h_j}{2}, u_{j+\frac{1}{2}}), \quad \forall j = 0, \ldots, N-1. \end{cases}$$

Unter gewisser Differenzierbarkeitsannahme hat dieses Verfahren die Konsistenzordnung $p = 2$.

(ii) **Das klassische Runge-Kutta-Verfahren:** Die Verfahrensfunktion lautet

$$\varphi(t,u,h) = \frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4) \quad \forall t \in [a,b],\ u \in \mathbb{R}^n,\ h \in [0, b-t]$$

mit

$$\begin{aligned} K_1 &= f(t,u), \\ K_2 &= f(t + \frac{h}{2}, u + \frac{h}{2}K_1), \\ K_3 &= f(t + \frac{h}{2}, u + \frac{h}{2}K_2), \\ K_4 &= f(t + \frac{h}{2}, u + hK_3). \end{aligned}$$

Unter gewisser Differenzierbarkeitsannahme hat das klassische Runge-Kutta-Verfahren die Konsistenzordnung $p = 4$.

4.4 Mehrschrittverfahren

4.4.1 Definition des Mehrschrittverfahrens

Wir betrachten zur Vereinfachung eine äquidistante Zerlegung des Intervalls $[a,b]$:

$$\begin{cases} a = t_0 < t_1 < \ldots < t_N = b, \\ t_j = a + jh, \quad j = 0, \ldots, N, \\ h = \frac{b-a}{N}. \end{cases}$$

Definition 4.18. Ein m-Schrittverfahren, $m \in \mathbb{N}$, zur numerischen Lösung des Anfangswertproblems

$$\begin{cases} y'(t) = f(t, y(t)), \quad t \in [a,b], \\ y(0) = y_0 \end{cases}$$

ist von der Form:

$$\sum_{k=0}^{m} a_k u_{j+k} = h\varphi(t_j, u_j, u_{j+1}, \ldots, u_{j+m}, h) \quad \forall j = 0, \ldots, N-m \tag{MV}$$

mit Koeffizienten $a_k \in \mathbb{R}$, $a_m \neq 0$, und einer Verfahrensfunktion

$$\varphi : [a,b] \times (\mathbb{R}^n)^{m+1} \times \mathbb{R}_0^+ \to \mathbb{R}^n.$$

Ein m-Schrittverfahren wird im allgemeinen als Mehrschrittverfahren bezeichnet.

Bemerkung 4.19. Das m-Schrittverfahren benötigt Startwerte $u_0, \ldots, u_{m-1} \in \mathbb{R}^n$. Üblicherweise setzt man $u_0 = y_0$ und bestimmt $u_{m-1}, \ldots, u_1$ durch ein anderes Verfahren (z. B. Einschrittverfahren).

Beispiel 4.20. Das implizite Euler-Verfahren

$$\begin{cases} u_{j+1} = u_j + hf(t_{j+1}, u_{j+1}), & j = 0, \dots, N-1, \\ u_0 = y_0 \end{cases}$$

ist ein 1-Schrittverfahren (Mehrschrittverfahren). Die Verfahrensfunktion lautet

$$\varphi(t_j, u_j, u_{j+1}, h) = f(t_j + h, u_{j+1})$$

und die Koeffizienten sind

$$a_0 = -1, \quad a_m = 1.$$

Das implizite Euler-Verfahren ist nichts anderes als

$$\sum_{k=0}^{1} a_k u_{j+k} = h\varphi(t_j, u_j, u_{j+1}, h) \quad \forall j = 0, \dots, N-1.$$

Beispiel 4.21. Die Mittelpunktregel

$$u_{j+2} = u_j + 2hf(t_{j+1}, u_{j+1}) \quad \forall j = 0, \dots, N-2$$

ist ein 2-Schrittverfahren (Mehrschrittverfahren) mit der Verfahrensfunktion

$$\varphi(t_j, u_j, u_{j+1}, u_{j+2}, h) = 2f(t_j + h, u_{j+1})$$

und Koeffizienten

$$a_0 = -1, \quad a_1 = 0, \quad a_2 = 1.$$

Definition 4.22 (lokaler Verfahrensfehler)**.** Für ein Mehrschrittverfahren (MV) zur Lösung des Anfangswertproblems (DGL)-(AB) bezeichnet

$$\eta(t,h) = \sum_{k=0}^{m} a_k y(t+kh) - h\varphi(t, y(t), y(t+h), \dots, y(t+mh), h)$$

für $t \in [a,b]$, $h \in [0, \frac{b-t}{m}]$ den *lokalen Verfahrensfehler* im Punkt $(t, y(t))$ der Schrittweite h.

Definition 4.23 (Konsistenzordnung)**.** Ein Mehrschrittverfahren (MV) besitzt die Konsistenzordnung $p \geq 1$, falls es Konstanten $\hat{C} > 0$ und $\bar{h} > 0$ gibt, so dass gilt

$$\|\eta(t,h)\|_2 \leq \hat{C}h^{p+1}, \quad t \in [a,b],\ h \in \left[0, \min\left\{\bar{h}, \frac{b-t}{m}\right\}\right].$$

Lipschitz-Annahme: Es existiere eine Konstante $L > 0$, so dass gilt

$$\left\|\varphi(t, u^0, u^1, \dots, u^m, h) - \varphi(t, v^0, v^1, \dots, v^m, h)\right\|_2 \leq L\left(\sum_{k=0}^{m} \left\|u^k - v^k\right\|_2\right)$$

für alle $t \in [a,b]$, $u^0, \dots, u^m, v^0, \dots, v^m \in \mathbb{R}^n$, $h \in [0, b-t]$.
Bei expliziten Einschrittverfahren sind die Konsistenzordnung $p \geq 1$ und die Lipschitz-Annahme hinreichend für die Konvergenz mit Rate p. Beim Mehrschrittverfahren braucht man eine weitere Voraussetzung, und zwar die sogenannte Nullstabilität.

Definition 4.24 (Nullstabilität)**.** Das m-Schrittverfahren (MV) heißt *nullstabil*, falls das Polynom

$$p(x) := a_0 + a_1 x + \ldots + a_m x^m$$

der Bedingung

$$\begin{cases} p(x) = 0 \quad \Rightarrow \quad |x| \leq 1, \\ p(x) = 0 \text{ und } |x| = 1 \quad \Rightarrow \quad x \text{ ist einfache Nullstelle von } p \end{cases}$$

genügt.

4.4.2 Konvergenzaussage

Satz 4.25. *Ein m-Schrittverfahren (MV) für das Anfangswertproblem (DGL)-(AB) sei nullstabil und die Verfahrensfunktion erfülle die Lipschitz-Annahme. Dann existieren zwei Konstanten $K > 0$ und $\bar{h} > 0$, so dass gilt*

$$\max_{0 \leq j \leq N} \left\| u_j - y(t_j) \right\|_2 \leq K \left[\underbrace{\max_{0 \leq k \leq m-1} \|u_k - y(t_k)\|_2}_{\text{Fehler der Startwerte}} + \frac{1}{h} \max_{t \in [a, b-mh]} \|\eta(t,h)\|_2 \right] \quad \forall h \in (0, \bar{h}).$$

Beweis. Wir zeigen die Aussage o.B.d.A. für $a_m = 1$ (sonst teilen wir (MV) durch $a_m \neq 0$) und $n = 1$. Wir setzen

$$\begin{aligned} e_j &:= u_j - y(t_j), \quad j = 0, \ldots, N, \\ \eta_j &:= \eta(t_j, h), \quad j = 0, \ldots, N - m. \end{aligned}$$

Laut Definition ist

$$\sum_{k=0}^{m} a_k u_{j+k} = h\varphi(t_j, u_j, \ldots, u_{j+m}, h), \quad j = 0, \ldots, N - m,$$

$$\sum_{k=0}^{m} a_k y(t_j + kh) = h\varphi(t_j, y(t_j), \ldots, y(t_j + mh, h) + \eta(t_j, h), \quad j = 0, \ldots, N - m.$$

Folglich gilt

$$\begin{aligned} \sum_{k=0}^{m} a_k e_{j+k} &= h(\varphi(t_j, u_j, \ldots, u_{j+m}, h) - \varphi(t_j, y(t_j), \ldots, y(t_j + mh), h)) - \eta(t_j, h) \\ &=: \delta_j - \eta_j, \quad j = 0, \ldots, N - m. \end{aligned}$$

Diese Gleichung lässt sich wie folgt darstellen: Sei $j \in \{0, \ldots, N - m\}$ fest, so gilt

$$\underbrace{\begin{pmatrix} e_{j+1} \\ \vdots \\ e_{j+m} \end{pmatrix}}_{=:E_{j+1}} = \underbrace{\begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ -a_0 & & \cdots & -a_{m-2} & -a_{m-1} \end{pmatrix}}_{=:A} \underbrace{\begin{pmatrix} e_j \\ \vdots \\ e_{j+m-1} \end{pmatrix}}_{=:E_j} + \underbrace{\begin{pmatrix} 0 \\ \vdots \\ 0 \\ \delta_j - \eta_j \end{pmatrix}}_{=:F_j} \tag{4.2}$$

mit der Matrix $A \in \mathbb{R}^{m\times m}$ und den Vektoren $E_j, E_{j+1}, F_j \in \mathbb{R}^m$. Mittels vollständiger Induktion zeigen wir:

$$E_j = A^j E_0 + \sum_{k=0}^{j-1} A^{j-1-k} F_k, \quad j = 0, \ldots, N-m+1. \tag{4.3}$$

Induktionsanfang $j = 0$: $E_0 = IE_0 + 0 = E_0$.

Induktionsannahme: Die Aussage gelte für $j \in \{0, \ldots, N-m\}$.

Mit (4.2) folgt

$$\begin{aligned} E_{j+1} = AE_j + F_j &= A(A^j E_0 + \sum_{k=0}^{j-1} A^{j-1-k} F_k) + F_j \\ &= A^{j+1} E_0 + \sum_{k=0}^{j-1} A^{j-k} F_k + F_j \\ &= A^{j+1} E_0 + \sum_{k=0}^{j} A^{j-k} F_k \end{aligned}$$

und daraus folgt die Aussage (4.3).
Mit der Nullstabilität zeigen wir später, dass eine Konstante $C > 0$ existiert, so dass gilt

$$\left\| A^k \right\|_\infty \leq C \quad \forall k \in \mathbb{N}.$$

Die Beschränktheit der Folge $\{\left\| A^k \right\|_\infty\}_{k=1}^\infty \subset \mathbb{R}_0^+$ zusammen mit (4.3) liefert

$$\begin{aligned} \|E_j\|_\infty &\leq \left\| A^j \right\|_\infty \|E_0\|_\infty + \sum_{k=0}^{j-1} \left\| A^{j-1-k} \right\|_\infty \|F_k\|_\infty \\ &\leq C \left(\|E_0\|_\infty + \sum_{k=0}^{j-1} \|F_k\|_\infty \right), \quad j = 0, \ldots, N-m+1. \end{aligned} \tag{4.4}$$

Laut Definition ist

$$\begin{aligned} \|F_k\|_\infty = |\delta_k - \eta_k| &\leq hL \sum_{l=0}^{m} |u_{k+l} - y(t_{k+l})| + |\eta_k| \\ &\leq hL \sum_{l=0}^{m} |e_{k+l}| + \max_{0\leq k\leq N-m} |\eta_k|. \end{aligned}$$

Andererseits ist

$$\begin{aligned} \sum_{l=0}^{m} |e_{k+l}| = &\underbrace{|e_k|}_{\leq \|E_k\|_\infty} + \underbrace{|e_{k+1}|}_{\leq \|E_k\|_\infty} + \ldots + \underbrace{|e_{k+m-1}|}_{\leq \|E_k\|_\infty} + \underbrace{|e_{k+m}|}_{\leq \|E_{k+1}\|_\infty} \\ &\leq m \|E_k\|_\infty + \|E_{k+1}\|_\infty . \end{aligned}$$

Insgesamt gilt

$$\|F_k\|_\infty \leq hL(m\,\|E_k\|_\infty + \|E_{k+1}\|_\infty) + \max_{0\leq k\leq N-m} |\eta_k| \quad \forall k = 0,\dots,N-m.$$

Somit haben wir

$$\begin{aligned}\sum_{k=0}^{j-1} \|F_k\|_\infty &\leq h\underbrace{L(m+1)}_{=:c_1}\left(\sum_{k=0}^{j-1}\|E_k\|_\infty\right) + hL\left\|E_j\right\|_\infty + j\max_{0\leq k\leq N-m}|\eta_k| \\ &\leq hc_1\left(\sum_{k=0}^{j-1}\|E_k\|_\infty\right) + hL\left\|E_j\right\|_\infty + N\max_{0\leq k\leq N-m}|\eta_k| \quad \forall j = 0,\dots,N-m+1\end{aligned}$$

und mit (4.4) folgt

$$\left\|E_j\right\|_\infty \leq C\left(\|E_0\|_\infty + N\max_{0\leq k\leq N-m}|\eta_k| + hc_1\sum_{k=0}^{j-1}\|E_k\|_\infty\right) + ChL\left\|E_j\right\|_\infty.$$

Ist $h < \dfrac{1}{CL}$, so folgt

$$\underbrace{(1-ChL)}_{>0}\left\|E_j\right\|_\infty \leq C\left(\|E_0\|_\infty + N\max_{0\leq k\leq N-m}|\eta_k| + hc_1\sum_{k=0}^{j-1}\|E_k\|_\infty\right)$$

und daraus ergibt sich

$$\left\|E_j\right\|_\infty \leq \frac{C}{1-ChL}\left(\|E_0\|_\infty + N\max_{0\leq k\leq N-m}|\eta_k| + hc_1\sum_{k=0}^{j-1}\|E_k\|_\infty\right).$$

Wir verwenden nun die diskrete Version des Lemmas von Gronwall:
Für reelle Zahlen $v_0,\dots,v_r \in \mathbb{R}$, positive reelle Zahlen $h_0,\dots,h_{r-1} \in \mathbb{R}^+$, und $\alpha,\beta \geq 0$ seien $|v_0| \leq \alpha$, $|v_j| \leq \alpha + \beta\sum_{k=0}^{j-1} h_k|v_k|$ für alle $j = 1,\dots,r$. Dann ist

$$|v_j| \leq \alpha\exp\left(\beta\sum_{k=0}^{j-1}h_k\right) \quad \forall j = 0,\dots,r.$$

Das Lemma von Gronwall liefert dann

$$\left\|E_j\right\|_\infty \leq \frac{C}{1-ChL}\left(\|E_0\|_\infty + N\max_{0\leq k\leq N-m}|\eta_k|\right)\exp\left(\frac{Cc_1}{1-ChL}hj\right)$$

für alle $j = 0,\dots,N-m+1$ und daraus folgt

$$\begin{aligned}\left\|E_j\right\|_\infty &\leq \frac{C}{1-ChL}\exp\left(\frac{Cc_1}{1-ChL}hj\right)\left(\max_{0\leq k\leq m-1}|u_k - y(t_k)| + \frac{b-a}{h}\max_{0\leq k\leq N-m}|\eta_k|\right) \\ &\leq \frac{C}{1-ChL}\exp\left(\frac{Cc_1}{1-ChL}\frac{b-a}{N}N\right)\cdot \\ &\qquad \cdot\max\{1,b-a\}\left(\max_{0\leq k\leq m-1}|u_k - y(t_k)| + \frac{1}{h}\max_{t\in[a,b-mh]}|\eta(t,h)|\right),\end{aligned}$$

denn

$$t_{N-m} = a + (N-m)h = a + Nh - mh = a + b - a - mh = b - mh.$$

Diese Aussage gilt für jedes $h < \dfrac{1}{CL}$. Wählen wir nun ein festes $\bar{h} < \dfrac{1}{CL}$, so folgt

$$\left\|E_j\right\|_\infty \leq K \left(\max_{0 \leq k \leq m-1} |u_k - y(t_k)| + \max_{t \in [a,\, b-mh]} |\eta(t,h)| \right)$$

für alle $h \in (0, \bar{h})$ und für alle $j = 0, \ldots, N-m+1$ mit

$$K := \frac{C}{1 - C\bar{h}L} \exp\left(\frac{Cc_1(b-a)}{1 - C\bar{h}L} \right) \max\{1, b-a\}.$$

□

4.4.3 Lemma von Gronwall

Im Beweis des Konvergenzsatzes haben wir das Lemma von Gronwall verwendet. Das Lemma von Gronwall ist in der Tat ein überaus wichtiges Werkzeug zur Untersuchung von partiellen Differentialgleichungen, insbesondere von Evolutionsgleichungen.

Lemma von Gronwall in der Integraldarstellung. *Es seien $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}^+$. Ferner sei $f : [0,T] \to \mathbb{R}$ eine Riemann-integrierbare Funktion mit*

$$f(t) \leq \alpha + \beta \int_0^t f(s)\,\mathrm{d}s \quad \forall t \in [0,T].$$

Dann gilt

$$f(t) \leq \alpha \exp(\beta t) \quad \forall t \in [0,T].$$

Beweis. Wir definieren

$$M := \sup_{t \in [0,T]} f(t) < \infty,$$

da $f : [0,T] \to \mathbb{R}$ laut Annahme Riemann-integrierbar ist, also ist f insbesondere beschränkt. Mittels vollständiger Induktion zeigen wir:

$$f(t) \leq \alpha \left(\sum_{j=0}^{n} \frac{(\beta t)^j}{j!} \right) + M \frac{(\beta t)^{n+1}}{(n+1)!} \quad \forall t \in [0,T]\ \forall n \in \mathbb{N} \cup \{0\}. \tag{4.5}$$

Induktionsanfang $n = 0$: Es gilt laut Voraussetzung

$$\begin{aligned} f(t) \leq \alpha + \beta \int_0^t f(s)\,\mathrm{d}s &\leq \alpha + \beta \int_0^t M\,\mathrm{d}s \\ &= \alpha + \beta M \int_0^t \mathrm{d}s \\ &= \alpha + M\beta t \quad \forall t \in [0,T]. \end{aligned}$$

Induktionsannahme: Die Aussage (4.5) gelte für $n-1$, mit $n \in \mathbb{N}$.

Dann gilt für $t \in [0,T]$:

$$\begin{aligned}
f(t) &\leq \alpha + \beta \int_0^t f(s)\,\mathrm{d}s \\
&\leq \alpha + \beta \left(\alpha \sum_{j=0}^{n-1} \frac{\beta^j}{j!} \int_0^t s^j \,\mathrm{d}s + M \frac{\beta^n}{n!} \int_0^t s^n \,\mathrm{d}s \right) \\
&= \alpha + \beta \left(\alpha \sum_{j=0}^{n-1} \frac{\beta^j}{j!} \frac{1}{j+1} t^{j+1} + M \frac{\beta^n}{n!} \frac{1}{n+1} t^{n+1} \right) \\
&= \alpha + \alpha \left(\sum_{j=0}^{n-1} \frac{(\beta t)^{j+1}}{(j+1)!} \right) + M \frac{(\beta t)^{n+1}}{(n+1)!} \\
&= \alpha + \alpha \left(\sum_{j=1}^{n} \frac{(\beta t)^{j}}{j!} \right) + M \frac{(\beta t)^{n+1}}{(n+1)!} \\
&= \alpha \left(\sum_{j=0}^{n} \frac{(\beta t)^{j}}{j!} \right) + M \frac{(\beta t)^{n+1}}{(n+1)!}
\end{aligned}$$

und daraus folgt die Behauptung (4.5). Der Übergang zum Grenzwert $n \to \infty$ in (4.5) liefert

$$f(t) \leq \lim_{n\to\infty} \left(\alpha \left(\sum_{j=0}^{n} \frac{(\beta t)^j}{j!} \right) + M \frac{(\beta t)^{n+1}}{(n+1)!} \right) = \alpha \exp(\beta t).$$

□

Lemma 4.26 (Diskrete Version des Lemmas von Gronwall)**.** *Für reelle Zahlen $v_0, \ldots, v_r \in \mathbb{R}$, positive reelle Zahlen $h_0, \ldots, h_{r-1} \in \mathbb{R}^+$, und $\alpha, \beta \geq 0$ seien*

$$|v_0| \leq \alpha, \quad |v_j| \leq \alpha + \beta \sum_{k=0}^{j-1} h_k |v_k| \quad \forall j = 0, \ldots, r.$$

Dann gilt

$$|v_j| \leq \alpha \exp\left(\beta \sum_{k=0}^{j-1} h_k \right) \quad \forall j = 0, \ldots, r.$$

Beweis. Für eine Menge $M \subset [0,T]$ definieren wir die charakteristische Funktion

$$\chi_M(t) = \begin{cases} 1, & \text{falls } t \in M, \\ 0, & \text{falls } t \notin M. \end{cases}$$

Wir setzen

$$\begin{aligned} x_0 &:= 0, \\ x_j &:= \sum_{k=0}^{j-1} h_k, \quad j = 0, \dots, r, \\ T &:= x_r. \end{aligned}$$

Wir definieren die folgende Treppenfunktion

$$f : [0, T] \to \mathbb{R}, \quad f(t) = \sum_{j=0}^{r-1} \chi_{[x_j, x_{j+1})}(t)|v_j| + \chi_{\{x_r\}}(t)|v_r|.$$

Für jedes $k \in \{0, \dots, r-1\}$ gilt

$$\int_{x_k}^{x_{k+1}} f(s)\,\mathrm{d}s = \int_{x_k}^{x_{k+1}} 1|v_k|\,\mathrm{d}s = |v_k|(x_{k+1} - x_k) = |v_k|h_k. \tag{4.6}$$

Außerdem gibt es zu jedem $t \in [0, T]$ ein $j \in \{0, \dots, r-1\}$ mit

$$t \in [x_j, x_{j+1})$$

oder $j = r$ mit $t = x_r = T$. Demzufolge gilt für jedes $t \in [0, T]$:

$$\begin{aligned} f(t) &= |v_j| \quad \text{für ein } j \in \{0, \dots, r\} \\ &\leq \alpha + \beta \sum_{k=0}^{j-1} \int_{x_k}^{x_{k+1}} f(s)\,\mathrm{d}s \\ &= \alpha + \beta \left(\int_{x_0}^{x_1} f(s)\,\mathrm{d}s + \int_{x_1}^{x_2} f(s)\,\mathrm{d}s + \ldots + \int_{x_{j-1}}^{x_j} f(s)\,\mathrm{d}s \right) \\ &= \alpha + \beta \int_0^{x_j} f(s)\,\mathrm{d}s \\ &\leq \alpha + \beta \int_0^t f(s)\,\mathrm{d}s. \end{aligned}$$

Insgesamt gilt also

$$f(t) \leq \alpha + \beta \int_0^t f(s)\,\mathrm{d}s \quad \forall t \in [0, T].$$

Mit dem Lemma von Gronwall folgt nun

$$f(t) \leq \alpha \exp(\beta t) \quad \forall t \in [0, T]$$

und mit $t = x_j$ folgt schließlich

$$|v_j| = f(x_j) \leq \alpha \exp(\beta x_j) = \alpha \exp\left(\beta \sum_{k=0}^{j-1} h_k \right) \quad \forall j = 0, \dots, r.$$

□

4.4.4 Beschränktheit der Folge $\{\|A^k\|_\infty\}_{k=0}^\infty$

Im Konvergenzsatz haben wir behauptet, dass für die Matrix

$$A = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ -a_0 & & \cdots & -a_{m-2} & -a_{m-1} \end{pmatrix} \in \mathbb{R}^{m\times m}$$

die Folge $\{\|A^k\|_\infty\}_{k=0}^\infty \subset \mathbb{R}_0^+$ beschränkt ist. Wir wollen diese Aussage nun beweisen. Dazu ist die Nullstabilität des Mehrschrittverfahrens (MV) hinreichend.

Lemma 4.27. *Das m-Schrittverfahren* (MV) *(o.B.d.A. $a_m = 1$) sei nullstabil, das heißt das Polynom*

$$p(x) = a_0 + a_1 x + \ldots + \underbrace{a_m}_{=1} x^m$$

erfülle die Bedingung

$$\begin{cases} p(x) = 0 \quad \Rightarrow \quad |x| \le 1, \\ p(x) = 0 \text{ und } |x| = 1 \quad \Rightarrow \quad x \text{ ist einfache Nullstelle von } p. \end{cases}$$

Dann gilt für alle Eigenwerte der Matrix A

$$\begin{cases} |\lambda_k| \le 1 \quad \forall k = 1, \ldots, m, \\ |\lambda_k| = 1 \quad \Rightarrow \quad \sigma(\lambda_k) = 1. \end{cases}$$

Beweis. Es genügt zu zeigen:

$$\det(\lambda I - A) = P_A(\lambda) = p(\lambda).$$

Laut Definition ist

$$
\begin{aligned}
&\det(\lambda I - A) \\
&= \det \begin{pmatrix} \lambda & -1 & & & \\ & \lambda & -1 & & \\ & & \ddots & \ddots & \\ & & & \lambda & -1 \\ a_0 & & \cdots & a_{m-2} & \lambda + a_{m-1} \end{pmatrix} \\
&= a_0(-1)^{m+1} \det \begin{pmatrix} -1 & & & \\ \lambda & -1 & & \\ & \ddots & \ddots & \\ & & \lambda & -1 \end{pmatrix} + a_1(-1)^{m+2} \det \begin{pmatrix} \lambda & & & \\ & -1 & & \\ & \lambda & \ddots & \\ & & \ddots & \ddots \\ & & & \lambda & -1 \end{pmatrix} \\
&+ a_2(-1)^{m+3} \det \begin{pmatrix} \lambda & -1 & & & & \\ & \lambda & & & & \\ & & -1 & & & \\ & & \lambda & \ddots & & \\ & & & \ddots & \ddots & \\ & & & & \lambda & -1 \end{pmatrix} + \ldots + \\
&+ (\lambda + a_{m-1})(-1)^{m+m} \det \begin{pmatrix} \lambda & -1 & & \\ & \ddots & \ddots & \\ & & \ddots & -1 \\ & & & \lambda \end{pmatrix} \\
&= a_0(-1)^{m+1}(-1)^{m-1} + a_1(-1)^{m+2}\lambda(-1)^{m-2} + a_2(-1)^{m+3}\lambda^2(-1)^{m-3} + \ldots + \\
&\qquad + (\lambda + a_{m-1})(-1)^{m+m}\lambda^{m-1} \\
&= (-1)^{2m}(a_0 + a_1\lambda + a_2\lambda^2 + \ldots + a_{m-1}\lambda^{m-1} + \lambda^m) = p(\lambda).
\end{aligned}
$$

□

Lemma 4.28. *Das m-Schrittverfahren mit o.B.d.A. $a_m = 1$ sei nullstabil. Dann gibt es eine Konstante $C > 0$, so dass gilt*

$$\left\| A^k \right\|_\infty \leq C \quad \forall k = 0, 1, 2, \ldots.$$

Dann ist die Folge $\{\left\| A^k \right\|_\infty\}_{k=0}^\infty$ insbesondere beschränkt.

Beweis. Wir verwenden die Jordansche Normalform: Es existiere eine reguläre Matrix

$X \in \mathbb{C}^{m\times m}$ mit

$$X^{-1}AX = J = \begin{pmatrix} J(\lambda_1) & & 0 \\ & \ddots & \\ 0 & & J(\lambda_k) \end{pmatrix}$$

mit paarweise verschiedenen Eigenwerten $\lambda_1, \dots, \lambda_k$ sowie den Jordan-Blöcken

$$J(\lambda_l) = \begin{pmatrix} \lambda_l & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_l \end{pmatrix} \in \mathbb{C}^{m_l \times m_l}.$$

Im vorherigen Lemma haben wir gezeigt:

$$\begin{cases} |\lambda_k| \leq 1 \quad \forall k = 1, \dots, m, \\ |\lambda_k| = 1 \quad \Rightarrow \quad \sigma(\lambda_k) = 1 \quad \Rightarrow \quad J(\lambda_l) = (\lambda_l) \in \mathbb{C}^{1\times 1}. \end{cases}$$

Wir wählen $\varepsilon > 0$ hinreichend klein, so dass für jedes $l \in \{1, \dots, k\}$ mit

$$|\lambda_l| < 1$$

gilt, dass

$$|\lambda_l| + \varepsilon \leq 1$$

ist. Wir definieren

$$\hat{J} := D^{-1}JD \quad \text{mit } D = \operatorname{diag}(1, \varepsilon^1, \varepsilon^2, \dots, \varepsilon^{m-1}) \in \mathbb{R}^{m\times m}.$$

Somit haben wir

$$\hat{J} = \begin{pmatrix} \hat{J}(\lambda_1) & & 0 \\ & \ddots & \\ 0 & & \hat{J}(\lambda_k) \end{pmatrix},$$

mit

$$\hat{J}(\lambda_l) = \begin{pmatrix} \lambda_l & \varepsilon & 0 & \\ & \ddots & \ddots & \\ & & \ddots & \varepsilon \\ & & & \lambda_l \end{pmatrix}, \quad \text{falls } |\lambda_l| < 1,$$

$$\hat{J}(\lambda_l) = (\lambda_l), \quad \text{falls } |\lambda_l| = 1.$$

Demzufolge gilt

$$\|\hat{J}\|_\infty = \max_{1\leq l\leq k} \|\hat{J}(\lambda_k)\|_\infty \leq 1,$$

denn

$$\max_{1\le l\le k}\|\hat{J}(\lambda_l)\|_\infty = \|\hat{J}(\lambda_j)\|_\infty = \begin{cases} |\lambda_j|+\varepsilon, & \text{falls } |\lambda_l|<1, \\ |\lambda_j|, & \text{falls } |\lambda_l|=1, \end{cases}$$

wobei j der Index sei, in dem das Maximum angenommen wird. Folglich ist

$$\|\hat{J}^n\|_\infty \le \underbrace{\|\hat{J}\|_\infty \cdot \|\hat{J}\|_\infty \cdot \ldots \cdot \|\hat{J}\|_\infty}_{n\text{-mal}} = \|\hat{J}\|_\infty^n \le 1 \quad \forall n \in \mathbb{N}.$$

Andererseits ist

$$\begin{aligned} & X^{-1}AX = J = D\hat{J}D^{-1} \\ \Rightarrow \quad & A = (XD)\hat{J}(XD)^{-1} \\ \Rightarrow \quad & AA = (XD)\hat{J}(XD)^{-1}(XD)\hat{J}(XD)^{-1} = (XD)\hat{J}^2(XD)^{-1} \\ & \quad \vdots \\ \Rightarrow \quad & A^n = (XD)\hat{J}^n(XD)^{-1} \quad \forall n \in \mathbb{N}. \end{aligned}$$

Demzufolge ist

$$\begin{aligned} \|A^n\|_\infty = \|(XD)\hat{J}^n(XD)^{-1}\|_\infty &\le \|XD\|_\infty\|\hat{J}^n\|_\infty\|(XD)^{-1}\|_\infty \\ &\le \|XD\|_\infty\|(XD)^{-1}\|_\infty =: C \quad \forall n \in \mathbb{N}. \end{aligned}$$

□

Somit ist die Konvergenzaussage von Abschnitt 4.4.2 vollständig bewiesen. Als Folgerung erhalten wir das folgende Korollar:

Korollar 4.29. *Das m-Schrittverfahren (MV) sei nullstabil und besitze die Konsistenzordnung $p \ge 1$. Ferner erfülle die Verfahrensfunktion φ die Lipschitz-Annahme. Existiert eine Konstante $C_0 > 0$, so dass für die Startwerte $u_0, \ldots, u_{m-1}$ gilt*

$$\max_{0\le j\le m-1}\|u_j - y(t_j)\|_2 \le C_0 h^p \quad \forall h > 0,$$

dann existieren Konstanten $\hat{K} > 0$ und $\bar{h} > 0$, so dass gilt

$$\max_{0\le j\le N}\|u_j - y(t_j)\|_2 \le \hat{K}h^p \quad \forall h \in (0, \bar{h}).$$

Beweis. Die Nullstabilität und die Lipschitz-Annahme liefern zwei Konstanten $K > 0$ und $\bar{h} > 0$, so dass gilt

$$\max_{0\le j\le N}\left\|u_j - y(t_j)\right\|_2 \le K\left[\max_{0\le k\le m-1}\|u_k - y(t_k)\|_2 + \frac{1}{h}\max_{t\in[a,b-mh]}\|\eta(t,h)\|_2\right] \quad \forall h \in (0,\bar{h}).$$

Die Annahme für die Startwerte und die Konsistenzordnung $p \ge 1$ liefern

$$\max_{0\le j\le N}\left\|u_j - y(t_j)\right\|_2 \le K(C_0h^p + \frac{1}{h}\hat{C}h^{p+1}) = \hat{K}h^p$$

mit

$$\hat{K} := K(C_0 + \hat{C}).$$

□

4.4.5 Adams-Verfahren

Wir betrachten in diesem Abschnitt das bekannte m-Schrittverfahren von Adams. Eine Grundlage hierzu ist die Integraldarstellung der Lösung $y \in C^1([a,b],\mathbb{R}^n)$ des Anfangswertproblems (DGL)-(AB):

$$y(t) = y_0 + \int_a^t f(s,y(s))\,\mathrm{d}s \quad \forall t \in [a,b].$$

Hieraus ergibt sich

$$\begin{aligned} y(t_{j+m}) - y(t_{j+m-1}) &= y_0 + \int_a^{t_{j+m}} f(s,y(s))\,\mathrm{d}s - y_0 - \int_a^{t_{j+m-1}} f(s,y(s))\,\mathrm{d}s \\ &= \int_{t_{j+m-1}}^{t_{j+m}} f(s,y(s))\,\mathrm{d}s. \end{aligned}$$

Durch Ersetzen des Integranden durch geeignete Polynome p erhalten wir das Verfahren von Adams.

4.4.5.1 Adams-Bashfort-Verfahren

Wir erklären die Methode für $m = 4$. Wir betrachten die äquidistante Zerlegung

$$\begin{cases} a = t_0 < t_1 < \ldots < t_N = b, \\ t_j = a + jh, \quad j = 0,\ldots,N, \\ h = \frac{b-a}{N}. \end{cases}$$

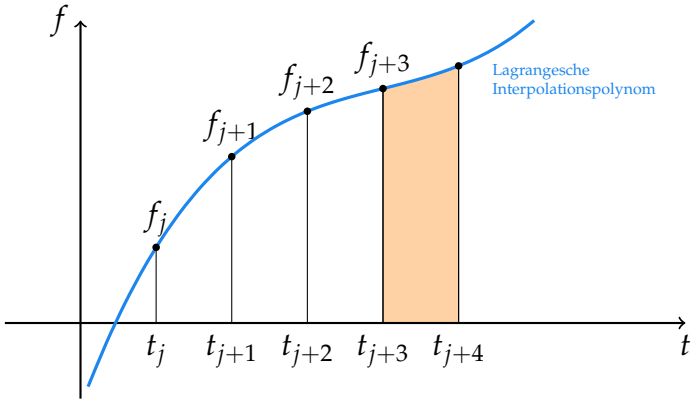
Sei $j \in \{0,\ldots,N-1\}$. Wir nehmen an, dass die Werte

$$\begin{aligned} f_j &:= f(t_j,u_j), & f_{j+1} &:= f(t_{j+1},u_{j+1}), \\ f_{j+2} &:= f(t_{j+2},u_{j+2}), & f_{j+3} &:= f(t_{j+3},u_{j+3}) \end{aligned}$$

berechnet worden sind. Mit diesen vier Stützstellen soll das Integral

$$\int_{t_{j+3}}^{t_{j+4}} f(s,y(s))\,\mathrm{d}s$$

mit dem Lagrangeschen Interpolationspolynom approximiert werden.



Wir approximieren

$$\int_{t_{j+3}}^{t_{j+4}} f(s, y(s))\,\mathrm{d}s \approx \int_{t_{j+3}}^{t_{j+4}} p(t)\,\mathrm{d}t$$

mit

$$p(t) = \sum_{k=0}^{3} f_{j+k} L_{j+k}(t).$$

Somit lautet das Verfahren von Adams-Bashfort für $m = 4$ wie folgt:

$$u_{j+4} - u_{j+3} = \sum_{k=0}^{3} \int_{t_{j+3}}^{t_{j+4}} f_{j+k} L_{j+k}(t)\,\mathrm{d}t.$$

Die Integrale lassen sich explizit berechnen, zum Beispiel bei $k = 0$:

$$\begin{aligned} I_0 &= \int_{t_{j+3}}^{t_{j+4}} \frac{(t - t_{j+3})(t - t_{j+2})(t - t_{j+1})}{(t_j - t_{j+3})(t_j - t_{j+2})(t_j - t_{j+1})}\,\mathrm{d}t \\ &= \int_{t_{j+3}}^{t_{j+4}} \frac{(t - t_{j+3})(t - t_{j+2})(t - t_{j+1})}{(-3h)(-2h)(-h)}\,\mathrm{d}t. \end{aligned}$$

Wir verwenden die Substitution

$$t = t_{j+3} + sh, \quad \mathrm{d}t = h\mathrm{d}s, \quad s \in [0, 1],$$

und daraus erhalten wir

$$\begin{aligned} I_0 = h \int_0^1 \frac{(sh) \cdot (s+1)h \cdot (s+2)h}{(-3h)(-2h)(-h)}\,\mathrm{d}s &= -h \int_0^1 \frac{s(s+1)(s+2)}{6}\,\mathrm{d}s \\ &\quad - \frac{h}{6} \int_0^1 s^3 + 3s^2 + 2s\,\mathrm{d}s \\ &= -\frac{h}{6}\left(\frac{1}{4} + 1 + 1\right) \\ &= -\frac{9}{24}h. \end{aligned}$$

Analog berechnen sich I_1, I_2, I_3. Insgesamt erhält man das „4-Schrittverfahren von Adams-Bashfort“:

$$u_{j+4} = u_{j+3} + \frac{h}{24}(55f_{j+3} - 59f_{j+2} + 37f_{j+1} - 9f_j) \quad \forall j = 0, \ldots, N-4.$$

Analog leitet man das allgemeine m-Schrittverfahren von Adams-Bashfort her:

$$\begin{aligned} m = 1: &\quad u_{j+1} = u_j + hf_j, \quad j = 0, \ldots, N-1, \\ m = 2: &\quad u_{j+2} = u_{j+1} + \frac{h}{2}(3f_{j+1} - f_j), \quad j = 0, \ldots, N-2, \\ m = 3: &\quad u_{j+3} = u_{j+2} + \frac{h}{12}(23f_{j+2} - 16f_{j+1} + 5f_j), \quad j = 0, \ldots, N-3, \end{aligned}$$

und so weiter.

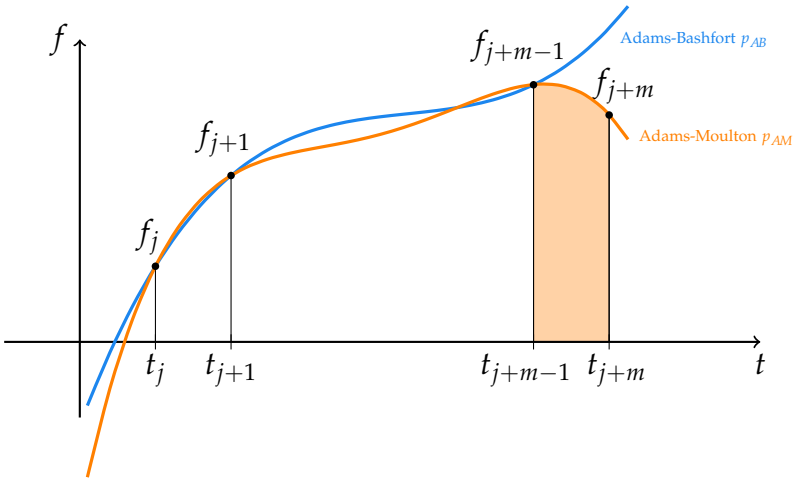
Bemerkung 4.30. Für $m = 1$ ist das Adams-Bashfort-Verfahren nichts anderes als das explizite Euler-Verfahren.

4.4.5.2 Adams-Moulton-Verfahren

Bei dem Adams-Bashfort-Verfahren verwendet man den unbekannten Wert u_{j+m} nicht. Somit sind Verfahren dieser Klasse explizit. Nun soll bei der Konstruktion des Interpolationspolynoms p des Adams-Moulton-Verfahrens der Wert

$$f_{j+m} := f(t_{j+m}, u_{j+m})$$

mit verwendet werden. Somit hat man bei m-Schrittverfahren von Adams-Moulton $m+1$ Stützstellen (statt m bei Adams-Bashfort).



Mit dem Interpolationspolynom

$$p_{AM}(t_{j+k}) = f_{j+k} \quad \forall k = 0, \ldots, m$$

setzen wir

$$u_{j+m} - u_{j+m-1} = \int_{t_{j+m-1}}^{t_{j+m}} p_{AM}(t)\,\mathrm{d}t \quad \forall j = 0, \ldots, N-m.$$

Dieses Verfahren bezeichnen wir als m-Schrittverfahren von Adams-Moulton.

Beispiele 4.31.

$m = 1$: $u_{j+1} = u_j + \frac{h}{2}(f_{j+1} + f_j) \quad \forall j = 0, \ldots, N-1.$

$m = 2$: $u_{j+2} = u_{j+1} + \frac{h}{12}(5f_{j+2} + 8f_{j+1} - f_j) \quad \forall j = 0, \ldots, N-2.$

$m = 3$: $u_{j+3} = u_{j+2} + \frac{h}{24}(9f_{j+3} + 19f_{j+2} - 5f_{j+1} - f_j) \quad \forall j = 0, \ldots, N-3.$

$m = 4$: $u_{j+4} = u_{j+3} + \frac{h}{720}(251f_{j+4} + 646f_{j+3} - 264f_{j+2} + 106f_{j+1} - 19f_j) \quad \forall j = 0, \ldots, N-4.$

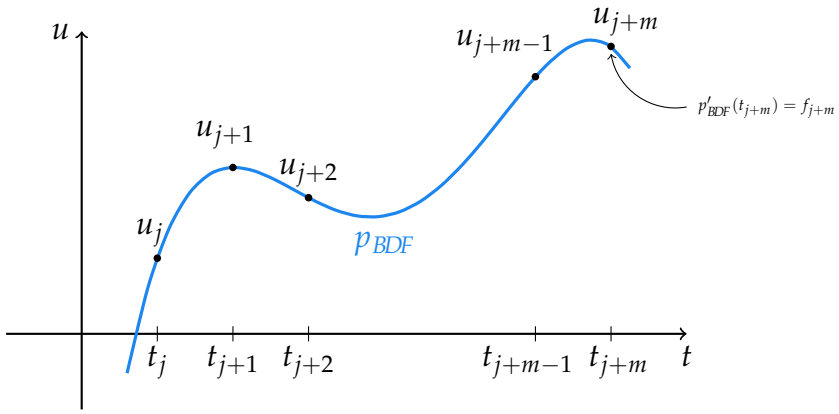
4.4.6 BDF-Verfahren

Wir präsentieren das bekannte BDF-Verfahren zur numerischen Lösung des Anfangswertproblems (DGL)-(AB). Es handelt sich um ein implizites Verfahren der Art:

$$u_{j+m} - u_{j+m-1} = \int_{t_{j+m-1}}^{t_{j+m}} p_{BDF}(t)\,\mathrm{d}t \quad \forall j = 0, \ldots, N-m,$$

mit dem Interpolationspolynom

$$\begin{cases} p_{BDF}(t_{j+k}) = u_{j+k} & \forall k = 0, \ldots, m, \\ p'_{BDF}(t_{j+m}) = f(t_{j+m}, u_{j+m}). \end{cases}$$



Das m-schrittige BDF-Verfahren lässt sich eindeutig wie folgt darstellen:

$$\sum_{k=0}^{m} a_k u_{j+k} = h f(t_{j+m}, u_{j+m}).$$

Beispiele 4.32.

$m = 1$: $u_{j+1} - u_j = h f(t_{j+1}, u_{j+1}) \quad \forall j = 0, \ldots, N-1.$

$m = 2$: $\frac{1}{3}(3u_{j+2} - 4u_{j+1} + u_j) = h f(t_{j+2}, u_{j+2}) \quad \forall j = 0, \ldots, N-2.$

$m = 3$: $\frac{1}{6}(11u_{j+3} - 18u_{j+2} + 9u_{j+1} - 2u_j) = h f(t_{j+3}, u_{j+3}) \quad \forall j = 0, \ldots, N-3.$

$m = 4$: $\frac{1}{12}(25u_{j+4} - 48u_{j+3} + 36u_{j+2} - 16u_{j+1} + 3u_j) = h f(t_{j+4}, u_{j+4}) \quad \forall j = 0, \ldots, N-4.$

Bemerkung 4.33. BDF steht für „Backward Differentiation Formula“.

Fazit 4.34.
Adams-Bashfort:

$$u_{j+m} - u_{j+m-1} = \int_{t_{j+m-1}}^{t_{j+m}} p_{AB}(t)\,\mathrm{d}t$$

mit

$$p_{AB}(t_{j+k}) = f_{j+k} \quad \forall k = 0, \ldots, m-1.$$

Adams-Moulton:

$$u_{j+m} - u_{j+m-1} = \int_{t_{j+m-1}}^{t_{j+m}} p_{AM}(t)\,\mathrm{d}t$$

mit

$$p_{AM}(t_{j+k}) = f_{j+k} \quad \forall k = 0, \ldots, m.$$

BDF-Verfahren:

$$u_{j+m} - u_{j+m-1} = \int_{t_{j+m-1}}^{t_{j+m}} p_{BDF}(t)\,\mathrm{d}t$$

mit

$$\begin{cases} p_{BDF}(t_{j+k}) = u_{j+k} & \forall k = 0, \ldots, m, \\ p'_{BDF}(t_{j+m}) = f(t_{j+m}, u_{j+m}). \end{cases}$$

Bei $m = 1$ im Adams-Bashfort-Verfahren erhält man das explizite Euler-Verfahren, $m = 1$ im BDF-Verfahren liefert das implizite Euler-Verfahren.

Stichwortverzeichnis