Numerische Mathematik SS 2019

Dozent: Prof. Dr. Andreas Fischer

15. Mai 2019

Inhaltsverzeichnis

I Das gewöhnliche Iterationsverfahren **2**

1 Fixpunkte . . . 2
2 Der Fixpunktsatz von Banach . . . 3
3 Gewöhnliche Iterationsverfahren . . . 4
4 Das Newton-Verfahren als Fixpunktiteration . . . 7

II Iterative Verfahren für lineare Gleichungssysteme **9**

1 Fixpunktiteration . . . 9
1.1 Das Jacobi-Verfahren . . . 11
1.2 Das Gauss-Seidel-Verfahren . . . 12
1.3 SOR-Verfahren . . . 13
2 Krylov-Raum-basierte Verfahren . . . 15
2.1 Krylov-Räume . . . 15
2.2 Basisalgorithmen zur Lösung von $Ax = b$ . . . 16
2.3 Das CG-Verfahren . . . 17
2.4 Fehlerverhalten des CG-Verfahrens . . . 24
2.5 Vorkonditionierung . . . 27
2.6 Ausblick und Anmerkungen . . . 31

III Numerische Behandlung von Anfangswertaufgaben **33**

1 Aufgabe und Lösbarkeit . . . 33
2 Einschrittverfahren . . . 35
2.1 Grundlagen . . . 35
2.2 Lokaler Diskretisierungsfehler und Konsistenz . . . 36
2.3 Konvergenz von Einschrittverfahren . . . 37
2.4 Stabilität gegenüber Rundungsfehlern . . . 39
2.5 Runge-Kutta-Verfahren . . . 40
3 Mehrschrittverfahren . . . 43
3.1 Grundlagen . . . 43
3.2 Konsistenz- und Konvergenzordnung für lineare MSV . . . 44
4 A-Stabilität . . . 49
5 Einblick: Steife Probleme . . . 51
6 Ausblick . . . 53

Anhang **55**

Index **56**

Vorwort

Vorwort

Kapitel I

Das gewöhnliche Iterationsverfahren

1. Fixpunkte

Seien ein Vektorraum V, eine Menge $U \subseteq V$ und eine Abbildung $\Phi : U \to V$ gegeben. Dann heißt $x^* \in U$ Fixpunkt der Abbildung Φ, falls $\Phi(x^*) = x^*$ gilt. Die Aufgabe

$$\Phi(x) = x$$

(eigentlich die Aufgabe, diese Gleichung zu lösen) wird als Fixpunktaufgabe bezeichnet. Die Abbildung Φ heißt Fixpunktabbildung. Im Unterschied zur Fixpunktaufgabe heißt

$$F(x) = 0$$

Nullstellenaufgabe. Zu jeder Nullstellenaufgabe gibt es eine äquivalente Fixpunktaufgabe (z.B. $F(x) = 0 \Leftrightarrow \Phi(x) = x$ mit $\Phi(x) := F(x)+x$) und umgekehrt (z.B. $\Phi(x) = x \Leftrightarrow F(x) = 0$ mit $F(x) := \Phi(x)-x$).

2. Der Fixpunktsatz von Banach

Der folgende Satz gibt (unter gewissen Bedingungen) eine konstruktive Möglichkeit an, einen Fixpunkt näherungsweise zu ermitteln.

Satz 2.1 (Banach)
Seien $(V, \|\cdot\|)$ ein Banach-Raum, $U \subseteq V$ eine abgeschlossene Menge und $\Phi : U \to V$ eine Abbildung. Die Abbildung Φ sei selbstabbildend, d.h. es gilt

$$\Phi(U) \subseteq U.$$

Außerdem sei Φ kontraktiv, d.h. es gibt $\lambda \in [0, 1)$, so dass

$$\|\Phi(x) - \Phi(y)\| \leq \lambda \|x - y\|, \text{für alle } x, y \in U.$$

Dann besitzt Φ genau einen Fixpunkt $x^* \in U$. Weiterhin konvergiert die durch

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

erzeugte Folge $\{x^k\}$ für jeden Startwert $x^0 \in U$ gegen x^* und es gilt für alle $k \in \mathbb{N}$

$$\|x^{k+1} - x^*\| \leq \frac{\lambda}{1-\lambda} \|x^{k+1} - x^k\| \qquad \text{a posteriori Fehlerabschätzung,} \tag{2}$$

$$\|x^{k+1} - x^*\| \leq \frac{\lambda^{k+1}}{1-\lambda} \|x^1 - x^0\| \qquad \text{a priori Fehlerabschätzung,} \tag{3}$$

$$\|x^{k+1} - x^*\| \leq \frac{\lambda}{1-\lambda} \|x^k - x^*\| \qquad \text{Q-lineare Konvergenz mit Ordnung } \lambda. \tag{4}$$

Beweis. Verlesung zur Analysis. □

Die in Satz 2.1 vorkommende Zahl $\lambda \in [0, 1)$ wird Kontraktionskonstante genannt.

3. Gewöhnliche Iterationsverfahren

Durch Gleichung (1) erklärte Verfahren heißt gewöhnliches Interationsverfahren oder Fixpunktiteration. Kritisch ist dabei, ob die Voraussetzungen (Φ ist selbstabbildend und kontraktiv) erfüllt werden können. Dies wird in diesem Abschnitt im Fall $V = \mathbb{R}^n$ mit einer beliebigen aber festen Vektornorm $\|\cdot\|$ untersucht. Die zugeordnete Matrixnorm wurde mit $\|\cdot\|_*$ bezeichnet.

Lemma 3.1
Sei $S \subseteq \mathbb{R}^n$ offen und konvex und $\Phi : D \to \mathbb{R}^n$ stetig differenzierbar. Falls $L > 0$ existiert mit

$$\|\Phi'(x)\|_* \leq L \text{ für alle } x \in D, \tag{1}$$

dann ist Φ Lipschitz-stetig in D mit der Lipschitz-Konstante L, d.h. es gilt

$$\|\Phi(x) - \Phi(y)\| \leq L\|x - y\| \text{ für alle } x \in D. \tag{2}$$

Die Umkehrung dieser Aussage ist ebenfalls richtig.

Beweis. 1. Sei Gleichung (1) erfüllt. Mit Satz 5.1 aus der Vorlesung ENM folgt

$$\|\Phi(x) - \Phi(y)\|_* = \left\| \int_0^1 \Phi'(y + t(x - y))(x - y)\,\mathrm{d}t \right\| \leq \|x - y\| \sup_{t\in[0,1]} \left\|\Phi'(y + t(x - y))\right\|_* \tag{3}$$

für alle $x, y \in D$. Also liefert Gleichung (1) unter Beachtung der Konvexität von D die Behauptung.

2. Sei nun Gleichung (2) erfüllt. Angenommen es gibt $\hat{y} \in D$ mit

$$\left\|\Phi'(\hat{y})\right\|_* > L. \tag{4}$$

Unter Berücksichtigung der Definition der zugeordneten Matrixnorm $\|\cdot\|_*$ folgt, dass $d \in \mathbb{R}^n$ existiert mit $\|d\| = 1$ und $\|\Phi'(\hat{y}d)\| = \|\Phi(\hat{y})\|_*$. Wendet man nun ENM mit $x := \hat{y} + sd$ und $y := \hat{y}$ an, so folgt für alle $s > 0$ hinreichend klein

$$\|\Phi(\hat{y} + sd) - \Phi(\hat{y})\| \leq L\|sd\| = sL \tag{5}$$

und

$$\begin{aligned}
\|\Phi(\hat{y} + sd) - \Phi(\hat{y})\| &= \left\| \int_0^1 \Phi'(\hat{y} + tsd)(sd)\,\mathrm{d}t \right\| \\
&= \left\| \int_0^1 \Phi'(\hat{y} + tsd)(sd)\,\mathrm{d}t + \int_0^1 \Phi'(\hat{y})(sd)(sd)\,\mathrm{d}t - \int_0^1 \Phi'(\hat{y})(sd)(sd)\,\mathrm{d}t \right\| \\
&\geq s\left\|\Phi'(\hat{y}d)\right\| - s\|d\| \sup_{t\in[0,1]} \left\|\Phi'(\hat{y} + tsd) - \Phi'(\hat{y})\right\|_* \\
&= s(\left\|\Phi'(\hat{y})\right\|_* - \sup_{t\in[0,1]} \left\|\Phi'(\hat{y} + tsd) - \Phi'(\hat{y})\right\|_*) \\
&> sL,
\end{aligned}$$

wobei sich die letzte Ungleichung wegen Gleichung (4) und der Stetigkeit von Φ' ergibt. Offenbar hat man damit einen Widerspruch, so dass die Annahme falsch ist. □

■ **Beispiel 3.2**

Die Nullstellenaufgabe $\cos(x) - 2x = 0$ sei zu lösen. Eine mögliche Formulierung als Fixpunktaufgabe ist

$$\Phi(x) = x \text{ mit } \Phi(x) := -x + \cos(x)$$

Offenbar ist $\Phi : \mathbb{R} \to \mathbb{R}$ selbstabbildend. Weiter ergibt sich

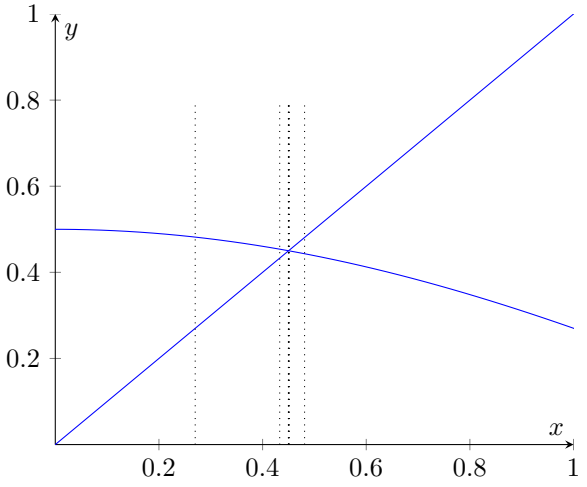
$$\Phi'(x) = -1 - \sin(x)$$

Für $x \in D := (0, 1)$ gilt daher $|\Phi'(x)| > 1$. Mit Lemma 3.1 folgt $|\Phi(x) - \Phi(y)| \geq |x - y|$ für mindestens ein Paar $(x, y) \in D \times D$. Somit ist Φ in D nicht kontrahierend. Definiert man Φ aber durch $\Phi(x) := 1/2 \cos(x)$, so ist die Fixpunktaufgabe $1/2 \cos(x) = x$ wiederum zur Nullstellenaufgabe äquivalent und es folgt

$$\Phi'(x) = \frac{1}{2} \sin(x).$$

Damit hat man $|\Phi'(x)| \leq 1/2$ für alle $x \in \mathbb{R}$. Also ist die zuletzt definierte Abbildung Φ kontrahierend auf $\mathbb{R}$ (und dort natürlich selbstabbildend), so dass die Voraussetzungen des Banachschen Fixpunktsatzes erfüllt sind. Die Fixpunktiteration mit $\Phi(x) = 1/2 \cos(x)$ und $x^0 := 1$ ergibt:

$$\begin{aligned}
x^1 &= 0.270\ldots \\
x^2 &= 0.481\ldots \\
x^3 &= 0.433\ldots \\
x^4 &= 0.4517\ldots \\
x^5 &= 0.4498\ldots \\
x^6 &= 0.45025\ldots \\
x^7 &= 0.450167\ldots \\
x^8 &= 0.450187\ldots
\end{aligned}$$



Nehmen wir an, die Voraussetzungen des Banachschen Fixpunktsatzes seien gegeben. Dann hängt die Konvergenzgeschwindigkeit der Fixpunktiteration offenbar von der Kontraktionskonstanten $\lambda \in [0,1)$ ab. Je kleiner λ ist, desto schneller ist ist die Konvergenzgeschwindigkeit. Unter Umständen kann die Umformulierung einer Fixpunktaufgabe mit Hilfe einer anderen Fixpunktabbildung helfen, die Konvergenzgeschwindigkeit zu verbessern (ggf. auf Kosten der Größe der Menge U, in der die Voraussetzungen des Banachschen Fixpunktsatzes erfüllt sind.) Ein Beispiel zu Konstruktion einer Fixpunktabbildung mit lokal beliebig kleiner Kontraktionskonstante gibt Abschnitt 1.4. In Abschnitt 2.1 wird gezeigt, wie Fixpunktabbildungen zu iterativen Lösung von linearen Gleichungssystemen eingesetzt werden können. Im Weiteren bezeichne $B(x^*, r) :=$ die abgeschlossene Kugel um x^* mit Radius r (bzgl. einer passenden Norm).

Satz 3.3 (Ostrowski)
Seien $D \subseteq \mathbb{R}^n$ offen und $\Phi : D \to \mathbb{R}^n$ stetig differenzierbar. Die Abbildung Φ besitze einen Fixpunkt $x^* \in D$ mit $\|\Phi'(x^*)\|_* < 1$. Dann existiert $r > 0$, so dass das gewöhnliche Iterationsverfahren für jeden Startpunkt $x^0 \in B(x^*, r)$ gegen x^* konvergiert.

Beweis. Da Φ stetig differenzierbar ist und $\|\Phi'(x^*)\|_* < 1$, gibt es $\lambda \in [0,1]$ und $r > 0$, sodass

$$\left\|\Phi'(x)\right\|_* \leq \lambda \quad \text{für alle } x \in B(x^*, r).$$

Nach Lemma 3.1 gilt daher

$$\|\Phi(x) - \Phi(y)\| \leq \lambda \|x - y\| \quad \text{für alle } x, y \in B(x^*, r). \tag{6}$$

Insbesondere folgt hieraus

$$\|\Phi(x) - \Phi(x^*)\| = \|\Phi(x) - x^*\| \leq \lambda \|x - x^*\| \quad \text{für alle } x \in B(x^*, r) \tag{7}$$

und damit $\Phi(x) \in B(x^*, r)$ für alle $x \in B(x^*, r)$. Also ist Φ bzgl. $B(x^*, r)$ selbstabbildend und kontraktiv. Daher liefert Satz 2.1 die gewünschte Aussage. □

4. Das Newton-Verfahren als Fixpunktiteration

Sei $D \subseteq \mathbb{R}^n$ offen und $F : D \to \mathbb{R}^n$ stetig differenzierbar. Die Nullstellenaufgabe

$$F(x) = 0$$

wird nun in eine äquivalente Fixpunktaufgabe überführt. Dazu nehmen wir an, dass x^* eine reguläre Nullstelle von F ist. Wegen der vorausgesetzten Stetigkeit von F' gibt es $r > 0$ hinreichend klein, so dass $F'(x)$ für $x \in B(x^*, r)$ regulär ist. Damit erhält man

$$F(x) = 0 \Leftrightarrow 0 = -F'(x)^{-1}F(x) \Leftrightarrow x = x - F'(x)^{-1}F(x).$$

für $x \in B(x^*, r)$. Definiert man $\Phi : B(x^*, r) \to \mathbb{R}^n$ durch

$$\Phi(x) := x - F^*(x)^{-1}F(x). \tag{1}$$

so kann das Newton-Verfahren als Fixpunktverfahren mit Φ als Fixpunktabbildung interpretiert werden. Ob Φ selbstabbildend und kontrahierend ist, müsste noch untersucht werden. Hier soll nur die Kontraktionseigenschaft in $B(x^*, r)$ für $r > 1$ hinreichend klein betrachtet werden. Die Eigenschaft der Selbstabbildung ergibt sich dann wie im Beweis zu Satz 3.3.

Lemma 4.1
Sei $D \subseteq \mathbb{R}^n$ offen und $F : D \to \mathbb{R}^n$ stetig differenzierbar. Weiter sei $x^* \in D$ eine reguläre Nullstelle von F. Dann ist Φ in x^* differenzierbar mit $\Phi'(x^*) = 0$.

Beweis. Wie zuvor gezeigt wurde, ist die durch Gleichung (1) definierte Abbildung Φ in $B(x^*, r) \subset D$ hinreichend kleines $r > 0$ wohldefiniert. Falls

$$\lim_{x \to x^*} \frac{\|\Phi(x) - \Phi(x^*) - G(x - x^*)\|}{\|x - x^*\|} \tag{2}$$

mit $G = 0 \in \mathbb{R}^{n \times n}$ gilt, folgt die Behauptung des Lemmas aus der Definition der Fréchet-Differenzierbarkeit. Unter Beachtung von $\Phi(x^*) = x^*$ ergibt sich

$$\Phi(x) - \Phi(x^*) = x - F'(x)^{-1}F(x) - x^* = -F'(x)^{-1}(F'(x))(x^* - x) + F(x))$$

und mit Satz 5.1 aus der Vorlesung ENM folgt weiter

$$\Phi(x) - \Phi(x^*) = F'(x)^{-1}\left(-F(x^*) + \int_0^1 (F'(x + t(x^* - x)) - F'(x))(x^* - x)\,\mathrm{d}t\right) \tag{3}$$

für alle $x \operatorname{int} B(x^*, r)$. Die Stetigkeit von F' auf der kompakten Menge $B(x^*, r)$ impliziert, dass F' dort auch gleichmäßig stetig ist. Also gibt es zu jedem $\varepsilon > 0$ ein $\delta(\varepsilon) > 0$, so dass auch

$$\|x + t(x^* - x) - x\| \leq \delta(\varepsilon) \quad \text{die Beziehung } \left\|F'(x + t(x^* - x)) - F'(x)\right\|_* \leq \varepsilon$$

für beliebige $x \in B(x^*, r)$ und $t \in [0, 1]$ folgt. Damit hat man

$$\lim_{x \to x^*} \max_{t \in [0,1]} \left\|F'(x + t(x^* - x)) - F'(x)\right\|_* = 0$$

und

$$\lim_{x \to x^*} \frac{\left\| \int_0^1 (F'(x + t(x^* - x)) - F'(x))(x^* - x) \, \mathrm{d}t \right\|_*}{\|x - x^*\|} = 0$$

Somit erhält man aus Gleichung (3) unter Beachtung von $F(x^*) = 0$ und der Regularität von $F'(x)$

$$\lim_{x \to x^*} \frac{\|\Phi(x) - \Phi(x^*)\|}{\|x - x^*\| O(x - x^*)} = 0,$$

d.h. Gleichung (2) ist für $G = 0$ erfüllt. □

▶ **Bemerkung 4.2**

Falls F in einer Umgebung von x^* sogar zweimal stetig differenzierbar und damit Φ dort stetig differenzierbar ist, zeigt Lemma 3.1, dass $\|\Phi'(x)\|_* \leq L$ für alle $x \in D \cap B(x^*, r(L))$ gilt. D.h. die Kontraktionskonstante der Fixpunktabbildung Φ in Gleichung (1) in einer Kugel $B(x^*, r)$ konvergiert gegen 0, wenn man den Radius r gegen 0 gehen lässt. Ferner gibt es Sätze, bei denen unter geeigneten Voraussetzungen eine bestimmte lokale Konvergenzgeschwindigkeit (Q-Ordnung) gezeigt wird (etwa die Q Ordnung 2, wenn insbesondere Φ' stetig ist und $\Phi'(x^*) = 0$ gilt).

Kapitel II

Iterative Verfahren für lineare Gleichungssysteme

Seien eine reguläre Matrix $A \in \mathbb{R}^{n\times n}$ und $b \in Rn$ gegeben. In diesem Kapitel werden iterative Verfahren zur Lösung des linearen Gleichungssystems

$$Ax = b \tag{1}$$

betrachtet.

1. Fixpunktiteration

Grundidee dieser Verfahren ist die geeignete Umformulierung des System $Ax = b$ als Fixpunktaufgabe und die Anwendung des gewöhnlichen Iterationsverfahrens. Die hier betrachtete (zu Gleichung (1) äquivalente) Fixpunktaufgabe lautet

$$x = x - B^{-1}(Ax - b),$$

wobei $B \in \mathbb{R}^{n\times n}$ eine noch zu wählende reguläre Matrix ist. Bei Wahl eines Startpunktes $x^0 \in \mathbb{R}^n$ ergibt sich das gewöhnliche Iterationsverfahren damit zu

$$x^{k+1} := x^k - B^{-1}(Ax^k - b) = (\mathbb{1} - B^{-1}A)x^k + B^{-1}b, \qquad k = 0, 1, 2, \ldots \tag{1}$$

Mit den Bezeichnung $M := \mathbb{1} - B^{-1}A$ und $c := B^{-1}b$ untersuchen wir deshalb die Iterationsvorschrift

$$x^{k+1} := Mx^k + c. \tag{2}$$

Die zugehörige Fixpunktabbildung $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ ist damit offenbar gegeben durch

$$\Phi(x) := Mx + c.$$

Satz 1.1
Es sei $B \in \mathbb{R}^{n\times n}$ regulär und mit $M := \mathbb{1} - B^{-1}A$ gelte

$$\lambda := \|M\|_* < 1 \tag{3}$$

wobei $\|\cdot\|_*$ die einer Vektornorm $\|\cdot\|$ zugeordnete Matrixnorm bezeichnet. Dann gilt:

(a) Die für eine beliebiges $x^0 \in \mathbb{R}^n$ durch Gleichung (2) erzeugte Folge $\{x^k\}$ konvergiert gegen die eindeutige Lösung $x*$ des linearen Gleichungssystems Gleichung (1).

(b) Die Abschätzungen Gleichung (2) - Gleichung (4) sind für alle $k \in \mathbb{N}$ erfüllt.

Beweis. Direkte Folgerung aus dem Banachschen Fixpunktsatz (Satz I.2.1) □

▶ **Bemerkung 1.2**
In Satz 1.1 (a) kann die Folgerung Gleichung (3) durch die Bedingung

$$\rho(M) < 1 \tag{4}$$

ersetzt werden. Da

$$\rho(C) \le \|C\|_* \quad \text{für alle } C \in \mathbb{R}^{n\times n}$$

für jede beliebige zugeordnete Matrixnorm $\|\cdot\|_*$ gilt (vgl. Übungsaufgabe), ist Gleichung (4) eine schwächere Forderung als Gleichung (3). Andererseits gibt es zu jedem Paar $(C, \varepsilon) \in \mathbb{R}^{n\times n} \times (0, \infty)$ eine zugeordnete Matrixnorm $\|\cdot\|_{(C,\varepsilon)}$, so dass

$$\|C\|_{(C,\varepsilon)} \le \rho(C) + \varepsilon.$$

Dabei ist $\rho(C)$ der Spektralradius der Matrix $C \in \mathbb{R}^{n\times n}$, d.h.

$$\rho(C) := \max_{i=1,\dots,n} |\lambda_i|,$$

wobei $\lambda_1, \dots, \lambda_n \in \mathbb{C}$ die Eigenwerte der Matrix $C \in \mathbb{R}^{n\times n}$ bezeichnen. Man kann weiter zeigen, dass Gleichung (4) auch notwendig dafür ist, dass die durch Gleichung (1) erzeugte Folge $\{x^k\}$ für jedes x^0 gegen $x*$ konvergiert.

Um eine Matrix B zu finden, so dass einerseits der Aufwand pro Iteration Gleichung (1) niedrig und andererseits die Bedingung Gleichung (3) bzw. Gleichung (4) erfüllt ist, betrachten wir die folgende Zerlegung

$$A = L + D + R$$

der Matrix A, wobei $D := \mathrm{diag}(a_{11}, \dots, a_{nn})$ die aus den Diagonalelementen von A bestehende Diagonalmatrix bezeichnet und L bzw. R eine untere bzw. obere Dreiecksmatrix ist mit

$$L = \begin{pmatrix} 0 & & & & \\ a_{21} & 0 & & & \\ a_{31} & a_{32} & 0 & & \\ \vdots & & \ddots & \ddots & \\ a_{n1} & \cdots & \cdots & a_{n,n-1} & 0 \end{pmatrix} \text{ bzw. } R = \begin{pmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1n} \\ & 0 & a_{23} & \cdots & a_{2n} \\ & & \ddots & \ddots & \vdots \\ & & & 0 & a_{n-1,n} \\ & & & a_{n,n-1} & 0 \end{pmatrix}.$$

1.1. Das Jacobi-Verfahren

Wir setzen hier voraus, dass D regulär ist und wählen

$$B := D \tag{5}$$

Damit ergibt sich die Iterationsvorschrift

$$x^{k+1} = x^k - D^{-1}(Ax^k - b) = -D^{-1}(L+R)x^k + D^{-1}b. \tag{6}$$

In Gleichung (2) ist entsprechend

$$M := M_J := -D^{-1}(L+R) \text{ und } c := c_J := D^{-1}b$$

zu wählen. Dieses Verfahren heißt Gesamtschrittverfahren oder Jacobi-Verfahren. Der Aufwand pro Schritt (Berechnung von x^{k+1} aus x^k) beträgt $\mathcal{O}(n^2)$ bei voll besetzter Matrix A und mindestens $\mathcal{O}(n)$, falls A schwach besetzt ist.

Satz 1.3
Die Matrix A sei streng diagonaldominant (vgl. Definition 3.1 der Vorlesung ENM). Dann ist die Matrix B aus Gleichung (5) regulär und es gilt

$$\|M_J\|_\infty \leq \lambda_{SD} := \max_{i=1,\ldots,n} \frac{1}{|a_{ii}|} \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| < 1.$$

Beweis. Die Regularität von B ergibt sich sofort aus der strengen Diagonaldominanz von A. Nutzt man die Definition der Zeilensummennorm $\|\cdot\|_\infty$ erhält man sofort

$$\|M_J\|_\infty = \left\|D^{-1}(L+R)\right\|_\infty = \max_{i=1,\ldots,n} \frac{1}{|a_{ii}|} \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| = \lambda_{SD}.$$

Die vorausgesetzte strenge Diagonaldominanz von A sichert $\lambda_{SD} < 1$. □

1.2. Das Gauss-Seidel-Verfahren

Wir setzen hier voraus, dass $L + D$ regulär ist und wählen

$$B := L + D \tag{7}$$

Damit ergibt sich die Iterationsvorschrift

$$x^{k+1} = x^k - (L + D)^{-1}(Ax^k - b) = -(L + D)^{-1}Rx^k + (L + D)^{-1}b.. \tag{8}$$

In Gleichung (2) ist entsprechend

$$M := MGS := -(L + D)^{-1}R \text{ und } c := c_{GS} := (L + D)^{-1}b$$

zu wählen. Dieses Verfahren heißt Einzelschrittverfahren oder Gauß-Seidel-Verfahren. Der Aufwand pro Schritt beträgt im ungünstigsten Fall $\mathcal{O}(n^2)$. Verbesserungen sind möglich, wenn eine Sparse-Struktur in A ausgenutzt werden kann.

Satz 1.4
Die Matrix A sei streng diagonaldominant (↗ Definition 3.1 der Vorlesung ENM). Dann ist die Matrix B aus Gleichung (7) regulär und es gilt

$$\|M_{GS}\|_\infty \leq \lambda_{SD} < 1.$$

Beweis. Die Regularität von B folgt sofort aus der strengen Diagonaldominanz von A. Weiter ergibt sich

$$\|M_{GS}\|_\infty = \left\|(L + D)^{-1}R\right\|_\infty = \sup_{\|y\|_\infty=1} \left\|(L + D)^{-1}Ry\right\|_\infty.$$

Um für einen festen Vektor y mit $\|y\|_\infty = 1$ eine Abschätzung für die rechte Seite zu erhalten, setzen wir $z := (L + D)^{-1}Ry$. Damit gilt

$$(D + L)z = Ry \tag{9}$$

und

$$z_1 = \frac{1}{a_{11}} \sum_{j=1}^{n} a_{1j}y_j.$$

Daraus folgt (da $\lambda_{SD} < 1$ wegen der strengen Diagonaldominanz von A)

$$|z_1| \leq \frac{1}{|a_{11}|} \sum_{j=2}^{n} |a_{1j}||y_j| \leq \sum_{j=2}^{n} |a_{1j}| \leq \lambda_{SD} < 1.$$

Nehmen wir nun an, dass

$$|z_1| \leq \text{ für } i = 1, \ldots, k - 1,$$

für ein $k \in \{2, \dots, n\}$ gilt. Dann folgt wegen Gleichung (9) und $\|y\|_\infty = 1$

$$|z_k| = \frac{1}{|a_{kk}|}\left|-\sum_{i=1}^{k-1} a_{ki} z_i + \sum_{i=k+1}^{n} a_{ki} y_i\right| \leq \frac{1}{|a_{kk}|}\left(\sum_{i=1}^{k-1}|a_{ki}| + \sum_{i=k+1}^{n}|a_{ki}|\right) \leq \lambda_{SD}.$$

Somit hat man induktiv $|z_k| \leq \lambda_{SD}$ für $k = 1, \dots, n$ und damit

$$\left\|(L+D)^{-1} R y\right\|_\infty = \|z\|_\infty \leq \lambda_{SD}$$

für beliebige y mit $\|y\|_\infty = 1$. □

1.3. SOR-Verfahren

Um dieses verfahren zu beschreiben, nehmen wir an, dass für ein $\omega \neq 0$ die Matrix

$$B := L + \frac{1}{\omega} D \tag{10}$$

regulär ist. Damit ergibt sich die Iterationsvorschrift

$$x^{k+1} := x^k - \left(L + \frac{1}{\omega} D\right)^{-1} (Ax^k - b) = M(\omega) x^k + c(\omega)$$

$$M(\omega) := \mathbb{1} - \left(L + \frac{1}{\omega} D^{-1} A\right) = \left(L + \frac{1}{\omega} D\right)^{-1} \tag{11}$$

und

$$c(\omega) := \left(L + \frac{1}{\omega} D\right)^{-1} b. \tag{12}$$

Für $\omega = 1$ erhält man offenbar als Spezialfall das Gauss-Seidel-Verfahren, so dass der folgende Satz auch dafür Anwendung finden kann. Man beachte dazu Bemerkung 1.2.

Satz 1.5
Die Matrix A sei symmetrisch und positiv definit. Dann ist die Matrix B aus Gleichung (10) regulär (für jedes $\omega \neq 0$). Falls $\omega \in (0, 2)$, dann gilt

$$\rho(M(\omega)) < 1$$

und umgekehrt.

Beweis. Da A positiv definit ist, gilt $e_i^T A e_i = a_{ii} > 0$ für $i = 1, \dots, n$. Also ist D positiv definit und damit B regulär für alle $\omega \neq 0$.
Sei $\lambda \in \mathbb{C}$ ein Eigenwert von $M(\omega)$ und $z \in \mathbb{C}^n$ ein zugehöriger Eigenvektor. Mit

$$A = A - M(\omega)^T A M(\omega) + M(\omega)^T A M(\omega)$$

sowie (unter Berücksichtigung der Definition von M und von $A = A^T$ und $R = L^T$)

$$\begin{aligned} A - M(\omega)^T A M(\omega) &= A - (\mathbb{1} - B^{-1}A)^T A(\mathbb{1} - B^{-1}A) \\ &= AB^T A + AB^{-1}A - AB^{T-1}AB^{-1}A \\ &= (B^{-1}A)^T(B + B^T - A)(B^{-1}A) \\ &= (B^{-1}A)^T\Big(L + \frac{1}{\omega}D + L^T + \frac{1}{\omega}D - L - D - L^T\Big))(B^{-1}A) \\ &= (B^{-1}A)^T\Big(\frac{2-\omega}{\omega}D\Big)(B^{-1}A) \end{aligned}$$

ergibt sich daher

$$z^H Az = (AB^{T-1}z)^H\Big(\frac{2-\omega}{\omega}D\Big)(B^{-1}Az) + z^H M(\omega)^T A M(\omega)z.$$

Da die Diagonalmatrix D positiv-definit ist, besitzt $\frac{2-\omega}{\omega}D$ dieselbe Eigenschaft für $\omega \in (0,2)$. Es folgt

$$(AB^{T-1}z)^H\Big(\frac{2-\omega}{\omega}D\Big)(B^{-1Az}) > 0$$

und damit

$$|\lambda| < 1. \tag{13}$$

Also gilt $\rho(M(\omega)) = \max_{i=1,\dots,n}|\lambda_i| < 1$, sofern $\omega \in (0,2)$. Die Umkehrung der Aussage ergibt sich aus dem Satz von Kahan (↗ Übungsaufgabe). □

Es ist nun naheliegend, dass man $\omega \in (0,2)$ so wählen möchte, dass $\rho(\omega)$ möglichst klein ist. Dies ist in bestimmten Fällen näherungsweise möglich, ansonsten beschränkt man sich auf geeignete Heuristiken zur Wahl von ω. Auf der Fixpunktiteration Gleichung (2) beruhende Verfahren werden häufig auch Splitting-Methoden genannt. Es gibt noch weitere solche Verfahren, auf die hier nicht eingegangen wird.

2. Krylov-Raum-basierte Verfahren

2.1. Krylov-Räume

Für $A \in \mathbb{R}^{n\times n}$, $r \in \mathbb{R}^n$ und $k \in \mathbb{N}$ ist der k-te Krylov-Raum gegeben durch $\mathcal{K}_0 = \{0\}$ und

$$\mathcal{K}_k(r, A) = \text{span}\{r, Ar, A^2r, ..., A^{k-1}r\} \quad \text{für } k > 0$$

Offenbar ist $\dim(\mathcal{K}_k(r, A)) \leq \min\{k, n\}$ für alle $k \in \mathbb{N}$.

Lemma 2.1
Es seien $A \in \mathbb{R}^{n\times n}$, $r \in \mathbb{R}^n\backslash\{0\}$ und $k \in \mathbb{N}$ gegeben. Dann sind folgende Aussagen äquivalent:

(a) $\dim(\mathcal{K}_{k+1}(r, A)) < k + 1$

(b) $\mathcal{K}_k(r, A) = \mathcal{K}_{k+1}(r, A)$

Beweis.

- (a) ⇒ (b): Nach Voraussetzung gibt es $l \in \{1, ..., k\}$ und $\alpha_0, ..., \alpha_l \in \mathbb{R}$, so dass

$$A^l r = \sum_{i=0}^{l-1} \alpha_i A^i r$$

Multiplikation mit A^{k-l} liefert

$$A^k r = \sum_{i=0}^{l-1} \alpha_i A^{k-l+i} r \in \mathcal{K}_k(r, A)$$

Also folgt $\mathcal{K}_k(r, A) = \mathcal{K}_{k+1}(r, A)$.

- (b) ⇒ (a): Offensichtlich □

Satz 2.2
Es seien $A \in \mathbb{R}^{n\times n}$ regulär, $x^0 \in \mathbb{R}^n$ mit $r^0 = b - Ax^0 \neq 0$ gegeben. Dann sind folgende Aussagen für $k \in \mathbb{N}$ äquivalent:

(a) $\mathcal{K}_k(r^0, A) = \mathcal{K}_{k+1}(r^0, A)$

(b) $x^* = A^{-1}b \in x^0 + \mathcal{K}_k(r^0, A)$

Beweis.

- (a) ⇒ (b): Wegen Lemma 2.1 gibt es $l \in \{0, ..., k\}$ und $\mu_l, ..., \mu_k \in \mathbb{R}$, so dass $\mu_l \neq 0$ und

$$\begin{aligned} 0 &= \sum_{i=l}^{k} \mu_i A^i r^0 \\ &= \mu_l A^l r^0 + \sum_{i=l+1}^{k} \mu_i A^i r^0 \end{aligned}$$

wobei der Summationsterm auf der rechten Seite entfällt, wenn $l = k$. Wegen der Regularität von A kann man die Gleichung mit $A^{-(l+1)}$ multiplizieren. Für $l = k$ liefert dies $0 = A^{-1}r^0 = A^{-1}b - x^0 \in \mathcal{K}_k(r^0, A)$.

Für $l < k$ folgt

$$\begin{aligned} A^{-1}b - x^0 &= A^{-1}r \\ &= -\frac{1}{\mu_l}\sum_{i=l+1}^{k} \mu_i A^{i-l-1} r^0 \\ &= -\frac{1}{\mu_l}\sum_{i=0}^{k-l-1} \mu_{i+l+1} A^i r^0 \in \mathcal{K}_k(r^0, A) \end{aligned}$$

Somit gilt Aussage (b).

- (b) $\Rightarrow$ (a): Nach Voraussetzung gilt $x^* \in x_0 + \mathcal{K}_k(r^0, A)$. Durch Multiplikation mit A folgt

$$\begin{aligned} b &\in Ax^0 + A\mathcal{K}_k(r^0, A) \\ &= Ax^0 + \text{span}\{Ar^0, A^2r^0, A^3r^0, ..., A^kr^0\} \end{aligned}$$

Also ist $r^0 = b - Ax^0$ eine Linearkombination der Vektoren $Ar^0, A^2r^0, A^3r^0, ..., A^kr^0$ und es gilt

$$\dim(\mathcal{K}_{k+1}(r^0, A)) < k + 1$$

Lemma 2.1 liefert damit die Gültigkeit von Aussage (a). □

▶ **Bemerkung 2.3**
Offenbar gibt es $k^* \in \{1, ..., n\}$, so dass Aussage (a) von Satz 2.2 für k^* zutrifft, aber für kein $k < k^*$ erfüllt ist. Satz 2.2 zeigt daher, dass die exakte Lösung x^* von $Ax = b$ in $x^0 + \mathcal{K}_{k^*}(r^0, A)$ liegt. Man kann also nun versuchen, eine Folge $\{x^k\}$ mit $x^k \in x^0 + \mathcal{K}_k(r^0, A)$ zu bestimmen, so dass x^k das Gleichungssystem $Ax = b$ (geeignet) näherungsweise löst. Dazu werden in nächsten Abschnitt zwei grundlegende Ansätze angegeben (Minimum-Residuum und Galerkin).

2.2. Basisalgorithmen zur Lösung von $Ax = b$

Für eine reguläre Matrix $B \in \mathbb{R}^{n\times n}$ ist durch

$$\|x\|_B = \|Bx\|_2 \quad \text{für } x|in\mathbb{R}^n$$

eine Vektornorm im $\mathbb{R}^n$ definiert. Unabhängig von der konkreten Wahl der regulären Matrix B sind das lineare Gleichungssystem $Ax = b$ und die Minimierungs aufgabe

$$\frac{1}{2}\|b - Ax\|_B^2 \to \min$$

offenbar äquivalent.

■ **Algorithmus 2.4 (Minimum-Residuum Basisalgorithmus)**
Input: $x^0 \in \mathbb{R}$, $A, B \in \mathbb{R}^{n\times n}$ regulär, $b \in \mathbb{R}^n$

```
1 while r^k ≠ 0 do
2  k = k + 1;
3  compute x^k ∈ ℝ^n als Lösung von
4    f_B(x) = ½‖b − Ax‖²_B → min   bei x ∈ x^0 + K_k(r^0, A)
```

```
5   r^k = b - Ax^k
6  enddo
```

Output: $x^* = x^k$, $k^* = k$

Satz 2.5
Algorithmus 2.4 ist wohldefiniert und bricht (bei exakter Arithmetik) nach höchstens n (das heißt $k^* \leq n$) it der exakten Lösung x^* des Gleichungssystems $Ax = b$ ab.

Beweis. Die Funktion $f_B : \mathbb{R}^n \to \mathbb{R}$ (Zielfunktion der Optimierungsaufgabe) ist gleichmäßig konvex, da A regulär ist. Die Menge $x^0 + \mathcal{K}_k(r^0, A)$ (zulässiger Bereich der Optimierungsaufgabe) ist außerdem nichtleer, abgeschlossen und konvex. Nach einem bekannten Resultat der kontinuierlichen Optimierung besitzt die Optimierungsaufgabe eine eindeutige Lösung x^k. Somit ist Algorithmus 2.4 wohldefiniert. Wegen $\mathcal{K}_k(r^0, A) = \mathcal{K}_{k+1}(r^0, A)$ spätestens für $k = n$ muss Algorithmus 2.4 auf Grund von Satz 2.2 nach höchstens n Schritten mit der exakten Lösung x^* abbrechen, vgl. auch Bemerkung 2.3 □

Algorithmus 2.4 ist somit ein direktes Verfahren. Jedoch besteht die praktische Bedeutung von Realisierungen dieses Verfahrens darin, dass man unter bestimmten Bedingungen bereits für $k \ll k^*$ eine brauchbare Näherung x^k für x^* erhalten kann. Dabei ist es von entscheidender Bedeutung, dass jeder Schritt des Algorithmus möglichst wenig Aufwand erfordert (etwa $\mathcal{O}(n)$ für gewisse schwach besetzte Matrizen). Analoges trifft auch für folgende Klasse von Algorithmen zu.

■ **Algorithmus 2.6 (Galerkin Basisalgorithmus)**
Input: $x^0 \in \mathbb{R}$, $A, B \in \mathbb{R}^{n\times n}$ regulär, $b \in \mathbb{R}^n$

```
1  r^0 = b - Ax^0;
2  k = 0;
3  while r^k ≠ 0 do
4   k = k + 1;
5   determine Untervektorraum L_k von R^n mit dim(L_k) = k;
6   compute x^k ∈ x^0 + K_k(r^0, A) mit b - Ax^k ⊥ L_k;
7   r^k = b - Ax^k;
8  enddo
```

Output: $x^* = x^k$, $k^* = k$

Unter der Voraussetzung der Durchführbarkeit bricht Algorithmus 2.6 nach höchstens n Schritten mit der Lösung x^* von $Ax = b$ ab, da $\mathcal{L}_n = \mathbb{R}^n$ und $b - Ax^n \perp \mathcal{L}_n$ somit $b - Ax^n = 0$ impliziert.

2.3. Das CG-Verfahren

Es seien $A \in \mathbb{R}^{n\times n}$ und $b \in \mathbb{R}^n$ gegeben. Die Funktion $f : \mathbb{R}^n \to \mathbb{R}$ sei gegeben durch

$$f(x) = \frac{1}{2}x^T Ax - b^T x \quad \text{für } x \in \mathbb{R}^n$$

Falls die Matrix $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit ist (und das wird beim CG-Verfahren immer vorausgesetzt werden), gibt es eine orthogonale Matrix $Q \in \mathbb{R}^{n\times n}$ und eine Diagonalmatrix

$\Lambda = \text{diag}(\lambda_1, ..., \lambda_n)$ mit den n positiven Eigenwerten $\lambda_1, ..., \lambda_n$ von A, so dass

$$A = Q\Lambda Q^T$$

Mit $\Lambda^{1/2} = \text{diag}(\sqrt{\lambda_1}, ..., \sqrt{\lambda_n})$ und $A^{1/2} = Q\Lambda^{1/2}Q^T$ gilt $A^{1/2}A^{1/2} = A$. Die Matrix $A^{1/2}$ wird als Wurzel von A bezeichnet und ist symmetrisch und positiv definit. Es gibt keine weitere symmetrisch positiv definite Matrix, die Wurzel von A ist.

Lemma 2.7
Seien $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, $b \in \mathbb{R}^n$, $G \subseteq \mathbb{R}^n$ eine abgeschlossene nichtleere Menge und $x^* = A^{-1}b$. Dann sind folgende Aussagen äquivalent:

(a) y minimiert f auf G

(b) y minimiert $x \mapsto \|x - x^*\|_{A^{1/2}}$ auf G

(c) y minimiert $x \mapsto \|b - Ax\|_{A^{-1/2}}$ auf G

Beweis.

- (a) $\Rightarrow$ (b): Für jedes $x \in \mathbb{R}^n$ gilt

$$\begin{aligned}
\|x - x^*\|^2_{A^{1/2}} &= \left\|A^{1/2}(x - x^*)\right\|^2 \\
&= (x - x^*)^T A(x - x^*) \\
&= x^T Ax - 2x^T Ax^* + (x^*)^T Ax^* \\
&= x^T Ax - 2x^T b + b^T x^* \\
&= 2f(x) + b^T x^*
\end{aligned}$$

  Da $b^T x^*$ nicht von x abhängt, ist eine Minimalstelle y von $f(x)$ auf G auch eine Minimalstelle von $x \mapsto \|x - x^*\|_{A^{1/2}}$ auf G und umgekehrt.

- (b) $\Rightarrow$ (c): Für jedes $x \in \mathbb{R}^{n\times n}$ gilt

$$\begin{aligned}
\|x - x^*\|^2_{A^{1/2}} &= (x - x^*)^T A(x - x^*) \\
&= (x - x^*)^T AA^{-1}A(x - x^*) \\
&= (A(x - x^*))^T A^{-1}A(x - x^*) \\
&= (Ax - b)^T A^{-1}(Ax - b) \\
&= \left\|A^{-1/2}(b - Ax)\right\|^2 \\
&= \|b - Ax\|^2_{A^{-1/2}}
\end{aligned}$$

  Damit folgt die Behauptung unmittelbar. □

Das CG-Verfahren ergibt sich nn formal aus Algorithmus 2.4 (Minimum-Residuum Basisalgorithmus), indem dort $B = A^{-1/2}$ gesetzt wird. Lemma 2.7 zeigt uns dann, dass das Teilproblem, also

$$\frac{1}{2}\|b - Ax\|^2_{A^{-1/2}} \to \min \quad \text{bei } x \in x^0 + \mathcal{K}_k(r^0, A) \tag{1}$$

zur Aufgabe

$$\frac{1}{2}x^T Ax - b^T x \to \min \quad \text{bei } x \in x^0 + \mathcal{K}_k(r^0, A) \tag{2}$$

äquivalent ist. Es stellt sich nun die Frage, wie diese Teilprobleme effizient zu lösen sind. Dazu gibt nachfolgendes Lemma einen Hinweis. Zu seiner Formulierung benötigen wir noch

Definition 2.8 (*A*-konjugiert, *A*-orthogonal)
Seien $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit und $k \in \mathbb{N}$ mit $k \leq n$. Dann nennt man Vektoren $d^0, ..., d^{k-1} \in \mathbb{R}^n \backslash \{0\}$ mit der Eigenschaft

$$(d^i)^T A d^j = 0 \quad \forall \text{ Paare } (i,j) \text{ mit } i,j \in \{0,...,k-1\} \text{ und } i \neq j$$

A-konjugiert oder A-orthogonal.

Die lineare Hülle von Vektoren $d^0, ..., d^{k-1} \in \mathbb{R}^k$ soll durch

$$\mathcal{D}_k = \text{span}\{d^0, ..., d^{k-1}\}$$

bezeichnet werden, wobei durch den Kontext klar ist, welche Vektoren $d^0, ..., d^{k-1}$ gemeint sind.

Lemma 2.9
Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Weiter seien $x^0 \in \mathbb{R}^n$ und A-konjugierte Vektoren $d^0, ..., d^{n-1} \in \mathbb{R}^n \backslash \{0\}$ gegeben. Dann besitzt die Optimierungsaufgabe

$$\frac{1}{2} x^T A x - b^T x \to \min \quad \text{bei } x \in x^0 + \mathcal{D}_k \tag{3}$$

eine eindeutig bestimmte Lösung x^k für $k = 1, ..., n$. Dabei gilt für $k = 0, ..., n-1$

$$x^{k+1} = x^k + t_k d^k \quad \text{mit} \quad t_k = \frac{(r^k)^T d^k}{(d^k)^T A d^k} \quad \text{und} \quad r^k = b - Ax^k \tag{4}$$

und

$$(r^{k+1})^T d^j = 0 \quad \text{für } j = 0, ..., k \tag{5}$$

Beweis. Zunächst sei die Funktion $\varphi : \mathbb{R}^k \to \mathbb{R}$ für ein $k \in \{1, ..., n\}$ wie folgt definiert:

$$\varphi(\alpha) = f\left(x^0 + \sum_{i=0}^{k-1} \alpha_i d^i\right) \quad \forall \alpha \in \mathbb{R}^k$$

Offenbar ist dann jede Lösung α^* der freien Optimierungsaufgabe

$$\varphi(\alpha) \to \min \tag{6}$$

vermöge

$$x^k = x^0 + \sum_{i=0}^{k-1} \alpha_i^* d^i \tag{7}$$

eine Lösung von Gleichung (3) und entsprechend umgekehrt. Wir untersuchen daher nun Gleichung (6). Für

$\varphi(\alpha)$ erhält man

$$\varphi(\alpha) = \frac{1}{2}\left(x^0 + \sum_{i=0}^{k-1}\alpha_i d^i\right)^T A\left(x^0 + \sum_{i=0}^{k-1}\alpha_i d^i\right) - b^T\left(x^0 + \sum_{i=0}^{k-1}\alpha_i d^i\right)$$

Da $d^0, ..., d^{k-1}$ A-konjugiert sind, folgt

$$\begin{aligned}\varphi(\alpha) &= \frac{1}{2}\sum_{i=0}^{k-1}\alpha_i^2 (d^i)^T A d^i + (Ax^0 - b)^T\left(x^0 + \sum_{i=0}^{k-1}\alpha_i d^i\right) - \frac{1}{2}(x^0)^T A x^0 \\ &= \frac{1}{2}\sum_{i=0}^{k-1}\alpha_i^2 (d^i)^T A d^i - \sum_{i=0}^{k-1}\alpha_i (r^0)^T d^i + f(x^0)\end{aligned}$$

Jede Lösung α^* von Gleichung (6) muss der notwendigen Optimalitätsbedingung $\nabla\varphi(\alpha) = 0$ genügen. Wegen der Konvexität von φ (beachte $\nabla^2\varphi = A$ ist positiv definit) ist diese Bedingung auch hinreichend. Aus $\nabla\varphi(\alpha^*) = 0$ folgt

$$0 = \frac{\partial\varphi}{\partial\alpha_i}(\alpha^*) = \alpha_i^*(d^i)^T A d^i - (r^0)^T d^i \quad \text{für } i = 0, ..., k-1$$

und damit (beachte positive Definitheit von A) nach Umstellen

$$\alpha_i^* = \frac{(r^0)^T d^i}{(d^i)^T A d^i} \quad \text{für } i = 0, ..., k-1 \tag{8}$$

Also besitzt Gleichung (6) genau eine Lösung α^*. Entsprechend Gleichung (3) eine eindeutige Lösung, die sich aus Gleichung (7) ergibt. Man sieht nun sofort, dass der Wert α_i^* nicht vom gewählten k abhängig. Mit Gleichung (7) folgt daraus

$$x^{k+1} - x^k = \alpha_k^* d^k \quad \text{für } k = 0, ..., n-1 \tag{9}$$

Damit erhält man für $i \in \{1, ..., k-1\}$

$$\begin{aligned}(r^0)^T d^i &= \left(r^i + \sum_{j=1}^{i}(r^{j-1} - r^j)\right)^T d^i \\ &= \left(r^i + \sum_{j=1}^{i}(-Ax^{j-1} + Ax^j)\right)^T d^i \\ &= (r^i)^T d^i + \sum_{j=1}^{i}(x^j - x^{j-1})^T A d^i \\ &= (r^i)^T d^i + \sum_{j=1}^{i}\alpha_{j-1}^*(d^{j-1})^T A d^i\end{aligned}$$

und weiter unter Ausnutzung der A-Konjugiertheit der Vektoren $d^0, ..., d^{n-1}$

$$(r^0)^T d^i = (r^i)^T d^i$$

Setzt man dies in die Darstellung von α_i^* ein ,so folgt wegen Gleichung (9) und Gleichung (8) Teilbehauptung Gleichung (4).

Die letzte Teilbehauptung Gleichung (5) ergibt sich für $j = k$ aus

$$\begin{aligned}(r^{k+1})^T d^k &= (b - Ax^{k+1})^T d^k \\ &= (b - Ax^k - t_k A d^k)^T d^k \\ &= (r^k)^T d^k - t_k (d^k)^T A d^k \\ &= 0\end{aligned} \tag{10}$$

wobei die letzte Gleichung mit Hilfe der Darstellung von t_k in Gleichung (4) klar wird. Für $j \in \{0, ..., k-1\}$ zeigt man Gleichung (5) schließlich unter Benutzung von Gleichung (10) und der A-Konjugiertheit der $d^0, ..., d^{n-1}$

$$\begin{aligned}(r^{k+1})^T d^j &= (r^{j+1})^T d^j + \sum_{i=j+1}^{k} (r^{i+1} - r^i)^T d^j \\ &= \sum_{i=j+1}^{k} (b - Ax^{i+1} - b + Ax^i)^T d^j \\ &= \sum_{i=j+1}^{k} (-t_i A d^i)^T d^j \\ &= 0\end{aligned}$$

□

Um nun die Teilprobleme Gleichung (2) mit Hilfe von Lemma 2.9 lösen zu können, bietet es an, die Vektoren $d^0, d^1, ...$ möglichst so zu erzeugen, dass sie einerseits A-konjugiert sind und andererseits $\mathcal{K}_k(r^0, A) = \mathcal{D}_k$ für $k = 1, 2, ...$ gilt. Zuerst setzt man dazu

$$d^0 = r^0$$

so dass $\mathcal{K}_1(r^0, A) = \text{span}\{r^0\} = \mathcal{D}_1$ folgt. Zur Abkürzung sei noch

$$\mathcal{R}_k = \text{span}\{r^0, ..., r^{k-1}\}$$

definiert. Wir nehmen nun an, dass für ein $k \in \{1, ..., n-1\}$ bereits

$$\mathcal{R}_k = \mathcal{K}_k(r^0, A) = \mathcal{D}_k \tag{11}$$

mit

$$A\text{-konjugierten Vektoren } d^0, ..., d^{k-1} \in \mathbb{R}^n \backslash \{0\} \tag{12}$$

erfüllt ist. Falls $r^k \neq 0$ soll nun gezeigt werden, dass Gleichung (11) und Gleichung (12) mit einem geeignet zu bestimmenden d^k auch für $k+1$ anstelle von k gelten.

Offenbar zeiht Gleichung (4) in Lemma 2.9

$$\begin{aligned}r^k &= b - Ax^k \\ &= b - A(x^{k-1} + t_{k-1} d^{k-1}) \\ &= r^{-1} - t_{k-1} A d^{k-1}\end{aligned}$$

nach sich. Da

$$r^{k-1} \in \mathcal{K}_k(r^0, A) \subseteq \mathcal{K}_{k+1}(r^0, A) \text{ und}$$
$$Ad^{k-1} \in A\mathcal{K}_k(r^0, A) \subseteq \mathcal{K}_{k+1}(r^0, A)$$

folgt

$$r^k \in \mathcal{K}_{k+1}(r^0, A)$$

Wegen $r^k \in \mathcal{R}_{k+1} \backslash \{0\}$ hat man weiter

$$\mathcal{K}_{k+1}(r^0, A) = \mathcal{R}_{k+1}$$

Um auch $\mathcal{K}_{k+1}(r^0, A) = \mathcal{D}_{k+1}$ zu erreichen, muss man also

$$\mathcal{D}_{k+1} = \text{span}\{\mathcal{D}_k, r^k\}$$

setzen. Zur Bestimmung eines Vektors $d^k \neq 0$ derart, dass $\mathcal{D}_{k+1} = \text{span}\{\mathcal{D}_k, d^k\}$ und $d^0, ..., d^k$ A-konjugiert sind, kann (analog zum Orthogonalisierungsverfahren nach GRAM-SCHMIDT) der Ansatz

$$d^k = r^k + \sum_{i=0}^{k-1} \beta_i d^i \tag{13}$$

verwendet werden. Wegen Gleichung (5) in Lemma 2.9 und der Voraussetzung $r^k \neq 0$ hat man zunächst $d^k \neq 0$. Die gewünschte A-Konjugiertheit liefert nun die Bedingungen

$$\begin{aligned} 0 &= (d^k)^T A d^j \\ &= (r^k)^T A d^j + \sum_{i=0}^{k-1} \beta_i (d^i)^T A d^j \quad \text{für } j = 0, ..., k-1 \end{aligned}$$

Auf Grund der entsprechend Gleichung (11) vorausgesetzten A-Konjugiertheit von $d^0, ..., d^{k-1}$ ist dies äquivalent zu

$$0 = (r^k)^T A d^j + \beta_j (d^j)^T A d^j \quad \text{für } j = 0, ..., k-1$$

Wegen der positiven Definitheit von A hat man

$$\beta_j = -\frac{(r^k)^T A d^j}{(d^j)^T A d^j} \quad \text{für } j = 0, ..., k-1 \tag{14}$$

Für $j \in \{0, ..., k-1\}$ ergibt sich für den Zähler des Bruches in Gleichung (14)

$$\begin{aligned} (r^k)^T A d^j &= \frac{1}{t_j} (r^k)^T A (x^{j+1} - x^j) \\ &= \frac{1}{t_j} (r^k)^T (r^j - r^{j+1}) \end{aligned}$$

Da $r^j \in \mathcal{R}_k = \mathcal{D}_k = \text{span}\{d^0, ..., d^{k-1}\}$ nach Gleichung (11) für $j \in \{0, ..., k-1\}$ gilt, liefert Gleichung (5)

in Lemma 2.9, dass

$$(r^k)Ad^j = \begin{cases} 0 & \text{für } j = 0, ..., k-2 \\ \frac{-1}{t_{k-1}}(r^k)^T r^k & \text{für } j = k-1 \end{cases}$$

Somit und wegen Gleichung (4) folgt aus Gleichung (14)

$$\beta_0 = \cdots = \beta_{k-2} = 0 \quad \text{und} \quad \beta_{k-1} = \frac{\|r^k\|_2^2}{\|r^{k-1}\|_2^2} \tag{15}$$

sowie mit Gleichung (13)

$$d^k = r^k + \beta_{k-1} d^{k-1} \tag{16}$$

Da Gleichung (11) und Gleichung (12) trivialerweise auch für $k = 0$ richtig waren, ist gezeigt, dass Gleichung (11) und Gleichung (12) mit Gleichung (15) und Gleichung (16) für $k = 1, 2, ...$ gelten, sofern $r^{k-1} \neq 0$.

Multipliziert man d^k in Gleichung (16) von links mit $(r^k)^T$ und berücksichtigt man Gleichung (5), so ergibt sich noch folgenden Vereinfachung für t_k aus Gleichung (4):

$$\begin{aligned} t_k &= \frac{(r^k)^T d^k}{(d^k)^T A d^k} \\ &= \frac{\|r^k\|_2^2 + \beta_{k-1}(r^k)^T d^{k-1}}{(d^k)^T A d^k} \\ &= \frac{\|r^k\|_2^2}{(d^k)^T A d^k} \end{aligned}$$

Mit Gleichung (15), Gleichung (16), Abschnitt 2.3 und unter Beachtung der Äquivalenz der Teilprobleme Gleichung (1) und Gleichung (2) ergibt sich mit $B = A^{-1/2}$ folgende Realisierung des Algorithmus Algorithmus 2.4

■ **Algorithmus 2.10 (CG-Verfahren)**

Input: $x^0 \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit, $b \in \mathbb{R}^n$

```
1  d^0 = r^0 = b - Ax^0;
2  k = 0;
3  while r^k ≠ 0 do
4   t_k = ||r^k||_2^2 / ((d^k)^T A d^k);
5   x^{k+1} = x^k + t_k d^k;
6   r^{k+1} = b - Ax^{k+1} = r^k - t_k A d^k;
7   β_k = ||r^{k+1}||_2^2 / ||r^k||_2^2;
8   d^{k+1} = r^{k+1} - β_k d^k;
9   k = k + 1;
10 enddo
```

Output: $x^* = x^k$

Der Rechenaufwand pro Schritt des CG-Verfahrens ist durch den Aufwand zur Berechnung des Matrix-mal-Vektor-Produkts Ad^k (sowie einige weitere Operationen wie Skalarprodukte) gegeben. Falls die Matrix A geeignet schwach besetzt ist, kommt man daher mit $\mathcal{O}(n)$ Operationen pro Schritt des CG-Verfahrens aus. Von besonderer Bedeutung zur Erreichung dieses geringen Aufwandes ist die sehr einfache Bestimmung der A-konjugierten Basisvektoren $d^0, d^1, ...$ der KRYLOV-Räume $\mathcal{K}_1(r^0, A), \mathcal{K}_2(r^0, A), ...$ Obwohl wegen Satz 2.5 bekannt ist, dass das CG-Verfahren (bei exakter Arithmetik) nach höchstens n Schritten mit einer Lösung von $Ax = b$ abbricht, gilt das Interesse hier der Frage, wie sich der Fehler der Näherung x^k gegenüber einer Lösung x^* abschätzen lässt. Dies ist besonders für $k \ll n$ von praktischem Interesse.

2.4. Fehlerverhalten des CG-Verfahrens

Definition 2.11 (n-tes Tschebyschow-Polynom)
Für $n \in \mathbb{N}$ wird die durch

$$T_n(x) = \begin{cases} \frac{1}{2}\left[\left(x + i\sqrt{1-x^2}\right)^n + \left(x - i\sqrt{1-x^2}\right)^n\right] & \text{falls } |x| \leq 1 \\ \frac{1}{2}\left[\left(x + i\sqrt{x^2-1}\right)^n + \left(x - i\sqrt{x^2-1}\right)^n\right] & \text{falls } |x| > 1 \end{cases}$$

definierte Funktion $T_n : \mathbb{R} \to \mathbb{R}$ als n-tes TSCHEBYSCHOW-Polynom bezeichnet.

Für die Sinnhaftigkeit der Definition und weitere Darstellungen der TSCHEBYSCHOW-Polynome sei auf eine entsprechende Übungsaufgabe verwiesen, insbesondere sei noch bemerkt, dass

$$T_n(x) = \cos(n \cdot \arccos(x)) \quad \forall x \in [-1, 1] \tag{17}$$

für jedes $n \in \mathbb{N}$ erfüllt ist.

Lemma 2.12
Für jedes $n \in \mathbb{N}$ gilt

$$T_n\left(\frac{\kappa+1}{\kappa-1}\right) \geq \frac{1}{2}\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^n \quad \forall \kappa > 1$$

Beweis. Sei $x = \frac{\kappa+1}{\kappa-1}$. Dann gilt $x > 1$ wegen $\kappa > 1$. Mit Definition 2.11 folgt

$$2T_n(x) \geq \left(x + \sqrt{x^2-1}\right)^n$$

Die Behauptung ergibt sich nun, da

$$\begin{aligned} x + \sqrt{x^2-1} &= \frac{\kappa + 1 + \sqrt{(\kappa+1)^2 - (\kappa-1)^2}}{\kappa - 1} \\ &= \frac{\kappa + 2\sqrt{\kappa} + 1}{\kappa - 1} \\ &= \frac{(\sqrt{\kappa}+1)^2}{(\sqrt{\kappa}-1)(\sqrt{\kappa}+1)} \\ &= \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1} \end{aligned}$$

□

Es sei daran erinnert, dass die Konditionszahl einer regulären Matrix A in der Spektralnorm durch

$$\begin{aligned}\text{cond}_2(A) &= \|A\|_2 \left\|A^{-1}\right\|_2 \\ &= \sqrt{\lambda_{max}(A^T A)}\sqrt{\lambda_{max}((A^T)^{-1}A^{-1})} \\ &= \sqrt{\frac{\lambda_{max}(A^T A)}{\lambda_{min})(A^T A)}}\end{aligned}$$

gegeben ist, wobei λ_{max} bzw. λ_{min} den größten bzw. kleinsten Eigenwert der jeweiligen Matrix bezeichnen. Falls A sogar symmetrisch und positiv definit ist, folgt

$$\text{cond}_2(A) = \frac{\lambda_{max}(A)}{\lambda_{min}(A)}$$

Satz 2.13
Es seien $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit und $b \in \mathbb{R}^n$. Weiter seien $x^* = A^{-1}b$, $\kappa = \text{cond}_2(A)$ und $\{x^k\}$ eine durch Algorithmus 2.10 (CG-Verfahren) erzeugte Folge. Dann gilt

$$\left\|x^k - x^*\right\|_{A^{1/2}} \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k \left\|x^0 - x^*\right\|_{A^{1/2}}$$

für $k = 1, 2, ...$

Beweis. Es bezeichne $e^k = x^k - x^*$. Wegen $r^0 = b - Ax^0 = Ax^* - Ax^0 = -Ae^0$ folgt

$$A^j r^0 = -A^{j+1}e^0$$

für alle $j \in \mathbb{N}$ und damit

$$\begin{aligned}\mathcal{K}_k(r^0, A) &= \text{span}\{r^0, ..., A^{k-1}r^0\} \\ &= \text{span}\{Ae^0, ..., A^k e^0\}\end{aligned} \tag{18}$$

Berücksichtigt man, dass x^k Lösung von Gleichung (2) ist, so liefert die Äquivalenz (a) $\Leftrightarrow$ (b) in Lemma 2.7

$$\begin{aligned}\left\|e^k\right\|_{A^{1/2}} &= \left\|x^k - x^*\right\|_{A^{1/2}} \\ &= \min_{x\in x^0+\mathcal{K}_k(r^0,A)} \|x - x^*\|_{A^{1/2}}\end{aligned}$$

Mit Gleichung (18) erhält man daraus weiter

$$\begin{aligned}\left\|e^k\right\|_{A^{1/2}} &= \min_{z\in\mathcal{K}_k(r^0,A)} \left\|x^0 - z - x^*\right\|_{A^{1/2}} \\ &= \min_{\alpha_1,...,\alpha_k} \left\|e^0 + \sum_{i=1}^{k} \alpha_i A^i e^0\right\|_{A^{1/2}} \\ &= \min_{p\in\Pi_k^0} \left\|p(A)e^0\right\|_{A^{1/2}}\end{aligned} \tag{19}$$

wobei Π_k^0 die Menge aller Polynome p mit dem Höchstgrad k und $p(0) = 1$ bezeichnet. Da A symmetrisch ist,

gibt es eine Orthogonalbasis des $\mathbb{R}^n$ aus Eigenvektoren $v^1, ..., v^n$ von A. Also existieren $\gamma_1, ..., \gamma_n \in \mathbb{R}$, so dass

$$e^0 = \sum_{i=1}^{n} \gamma_i v^i \tag{20}$$

Bezeichnet man mit λ_i den zu v^i gehörenden Eigenwert, so ergibt sich

$$\begin{aligned} \left\|e^0\right\|_{A^{1/2}}^2 &= (e^0)^T A e^0 \\ &= \left(\sum_{i=1}^{n} \gamma_i v^i\right)^T \left(\sum_{i=1}^{n} \gamma_i A v^i\right) \\ &= \left(\sum_{i=1}^{n} \gamma_i v^i\right)^T \left(\sum_{i=1}^{n} \gamma_i \lambda_i v^i\right) \end{aligned}$$

und schließlich (mit der Orthogonalität der Eigenvektoren)

$$\left\|e^0\right\|_{A^{1/2}}^2 = \sum_{i=1}^{n} \gamma_i^2 \lambda_i \tag{21}$$

Man überlegt sich nun, dass für jedes $p \in \Pi_k^0$

$$\begin{aligned} p(A)v^i &= \mathbb{1}v^i + \sum_{j=1}^{k} \alpha_j A^j v^i \\ &= v^i + \sum_{j=1}^{k} \alpha_j (\lambda_i)^j v^i \\ &= \left(1 + \sum_{j=1}^{k} \alpha_j (\lambda_i)^j\right) v^i \\ &= p(\lambda_i) v^i \end{aligned} \tag{22}$$

das heißt die Eigenwerte von $p(A)$ sind durch $p(\lambda_1), ..., p(\lambda_n)$ gegeben. Mit Gleichung (20) folgt

$$\begin{aligned} \left\|p(A)e^0\right\|_{A^{1/2}}^2 &= \left(p(A)e^0\right)^T A p(A) e^0 \\ &= \left(\sum_{i=1}^{n} \gamma_i p(A) v^i\right)^T \left(\sum_{i=1}^{n} \gamma_i A p(A) v^i\right) \end{aligned}$$

und weiter mit Gleichung (22)

$$\begin{aligned} \left\|p(A)e^0\right\|_{A^{1/2}}^2 &= \left(\sum_{i=1}^{n} \gamma_i p(\lambda_i) v^i\right)^T \left(\sum_{i=1}^{n} \gamma_i \lambda_i p(\lambda_i) v^i\right) \\ &= \sum_{i=1}^{n} \gamma_i^2 p(\lambda_i)^2 \lambda_i \end{aligned}$$

Aus Gleichung (19) ergibt sich damit

$$\begin{aligned} \left\|e^k\right\|_{A^{1/2}} &= \min_{p \in \Pi_k^0} \left\|p(A)e^0\right\|_{A^{1/2}} \\ &= \min_{p \in \Pi_k^0} \left(\sum_{i=1}^{n} \gamma_i^2 p(\lambda_i)^2 \lambda_i\right)^{1/2} \end{aligned}$$

Wegen Gleichung (21) und mit $\lambda_{min} = \lambda_{min}(A)$ sowie $\lambda_{max} = \lambda_{max}(A)$ hat man

$$\begin{aligned} \left\|e^k\right\|_{A^{1/2}} &\leq \min_{p\in\Pi_k^0} \max_{\lambda\in[\lambda_{min},\lambda_{max}]} |p(\lambda)| \left(\sum_{i=1}^{n} \gamma_i^2 \lambda_i\right)^{1/2} \\ &= \left\|e^0\right\|_{A^{1/2}} \left(\min_{p\in\Pi_k^0} \max_{\lambda\in[\lambda_{min},\lambda_{max}]} |p(\lambda)|\right) \end{aligned} \tag{23}$$

Um den Term $(\dots)$ auf der rechten Seite von Gleichung (23) weiter abzuschätzen, sei zunächst bemerkt, dass im Fall $\lambda_{min} = \lambda_{max}$ dieser Term gleich 0 ist (man wähle $p(\lambda) = (\lambda_{max} - \lambda)\lambda_{max}^{-1}$). Also können wir $\lambda_{max} > \lambda_{min}$ annehmen und wählen dann $q \in \Pi_k^0$ speziell durch

$$q(\lambda) = \frac{T_k(l(\lambda))}{T_k(l(0))} \quad \text{mit} \quad l(\lambda) = \frac{2\lambda - (\lambda_{max} + \lambda_{min})}{\lambda_{min} - \lambda_{max}}$$

wobei T_k das k-te TSCHEBYSCHOW-Polynom ist. Die Funktion $l : \mathbb{R} \to \mathbb{R}$ ist offenbar monoton fallend mit $l(\lambda_{min}) = 1$ und $l(\lambda_{max}) = -1$. Wegen Gleichung (17) gilt daher

$$\max_{\lambda\in[\lambda_{min},\lambda_{max}]} |T_k(l(\lambda))| \leq 1$$

Mit Lemma 2.12 folgt noch

$$\begin{aligned} T_k(l(0)) &= T_k\left(\frac{\lambda_{max} + \lambda_{min}}{\lambda_{max} - \lambda_{min}}\right) \\ &= T_k\left(\frac{\frac{\lambda_{max}}{\lambda_{min}} + 1}{\frac{\lambda_{max}}{\lambda_{min}} - 1}\right) \\ &= T_k\left(\frac{\kappa + 1}{\kappa - 1}\right) \\ &\geq \frac{1}{2}\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^k \end{aligned}$$

Somit hat man

$$\begin{aligned} \min_{p\in\Pi_k^0} \left(\max_{\lambda\in[\lambda_{min},\lambda_{max}]} |p(\lambda)|\right) &\leq \max_{\lambda\in[\lambda_{min},\lambda_{max}]} |q(\lambda)| \\ &\leq 2\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^k \end{aligned}$$

und wegen Gleichung (23) die Behauptung. □

Satz 2.14
Es seien $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit und $b \in \mathbb{R}^n$. Wenn A genau m paarweise verschiedene Eigenwerte besitzt, so bricht Algorithmus 2.10 (bei exakter Rechnung) nach höchstens m Schritten mit der Lösung x^* des Gleichungssystems $Ax = b$ ab.

Beweis. Übungsaufgabe □

2.5. Vorkonditionierung

Satz 2.13 zeigt, dass das CG-Verfahren den Fehler $\left\|x^k - x^*\right\|_{A^{1/2}}$ insbesondere dann schnell reduziert, wenn die Kondition der symmetrischen und positiv definiten Matrix A klein ist, das heißt, wenn der größte und der kleinste Eigenwert von A eng beieinander liegen. Falls dem nicht so ist, dass wäre es

sinnvoll, das CG-Verfahren nicht auf $Ax = b$ sondern auf ein (in einem bestimmten Sinn) äquivalentes Gleichungssystem anzuwenden, dessen Systemmatrix ebenfalls symmetrisch und positiv definit ist, aber eine kleinere Kondition als A besitzt. Um ein solches Gleichungssystem zu erhalten, sei zunächst $C \in \mathbb{R}^{n\times n}$ eine beliebige reguläre Matrix. Dann gilt

$$Ax = b \quad \Leftrightarrow \quad (C^{-1}A(C^T)^{-1})C^T x = C^{-1}b$$

Mit der symmetrischen und positiv definiten Matrix $\tilde{A} = C^{-1}A(C^T)^{-1}$ und $\tilde{b} = C^{-1}b$ ergibt sich das Gleichungssystem

$$\tilde{A}\tilde{x} = \tilde{b}$$

wobei man aus dessen Lösung $\tilde{x}^* = \tilde{A}^{-1}\tilde{b}$ die Lösung $x^* = A^{-1}b$ vermittels $x^* = (C^T)^{-1}\tilde{x}^*$ erhalten kann.

■ **Algorithmus 2.15 (CG-Verfahren für $\tilde{A}\tilde{x} = \tilde{b}$)**
Input: $\tilde{x}^0 \in \mathbb{R}^n$, $\tilde{A} \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, $\tilde{b} \in \mathbb{R}^n$

1 $\tilde{d}^0 = \tilde{r}^0 = \tilde{b} - \tilde{A}\tilde{x}^0$;
2 `k = 0;`
3 `while` $\tilde{r}^k \neq 0$ `do`
4 $\tilde{t}_k = \frac{\|\tilde{r}^k\|_2^2}{(\tilde{d}^k)^T\tilde{A}\tilde{d}^k}$;
5 $\tilde{x}^{k+1} = \tilde{x}^k + \tilde{t}_k\tilde{d}^k$;
6 $\tilde{r}^{k+1} = \tilde{b} - \tilde{A}\tilde{x}^{k+1} = \tilde{r}^k - \tilde{t}_k\tilde{A}\tilde{d}^k$;
7 $\tilde{\beta}_k = \frac{\|\tilde{r}^{k+1}\|_2^2}{\|\tilde{r}^k\|_2^2}$;
8 $\tilde{d}^{k+1} = \tilde{r}^{k+1} - \tilde{\beta}_k\tilde{d}^k$;
9 `k = k + 1;`
10 `enddo`

Output: $x^* = x^k$

Setzt man für $k = 1, 2, ...$

$$x^k = (C^T)^{-1}\tilde{x}^k \quad r^k = C\tilde{r}^k \quad d^k = (C^T)^{-1}\tilde{d}^k \tag{24}$$

so folgt

$$\begin{aligned}\tilde{t}_k &= \frac{\|\tilde{r}^k\|_2^2}{(\tilde{d}^k)^T\tilde{A}\tilde{d}^k} \\ &= \frac{(r^k)^T(C^T)^{-1}C^{-1}r^k}{(C^Td^k)^T(C^{-1}A(C^T)^{-1})C^Td^k} \\ &= \frac{(r^k)^T(C^T)^{-1}C^{-1}r^k}{(d^k)^TAd^k}\end{aligned}$$

sowie

$$\tilde{x}^{k+1} = \tilde{x}^k + \tilde{t}_k \tilde{d}^k = C^T x^k + \tilde{t}_k C^T d^k$$

Wegen Gleichung (24) ergibt sich außerdem

$$x^{k+1} = x^k + \tilde{t}_k d^k \tag{25}$$

Für $\tilde{r}^{k+1}$ erhält man mit Gleichung (24)

$$\begin{aligned}\tilde{r}^{k+1} &= \tilde{r}^k - \tilde{t}_k \tilde{A}\tilde{d}^k \\ &= C^{-1}r^k - \tilde{t}_k C^{-1}A(C^T)^{-1}C^T d^k \\ &= C^{-1}(r^k - \tilde{t}_k A d^k)\end{aligned}$$

Wegen Gleichung (24) gilt außerdem

$$r^{k+1} = C\tilde{r}^{k+1} = r^k - \tilde{t}_k A d^k \tag{26}$$

Mit Gleichung (24) erhält man

$$\tilde{\beta}_k = \frac{\left\|\tilde{r}^{k+1}\right\|_2^2}{\|\tilde{r}^k\|_2^2} = \frac{(r^{k+1})^T(C^T)^{-1}C^{-1}r^{k+1}}{(r^k)^T(C^T)^{-1}C^{-1}r^k}$$

sowie

$$\begin{aligned}\tilde{d}^{k+1} &= \tilde{r}^{k+1} + \tilde{\beta}_k \tilde{d}^k \\ &= C^{-1}r^{k+1} + \tilde{\beta}_k C^T d^k \\ &= C^T((C^T)^{-1}C^{-1}r^{k+1} + \tilde{\beta}_k d^k)\end{aligned}$$

Wegen Gleichung (24) liefert das Letztere

$$d^{k+1} = (C^T)^{-1}C^{-1}r^{k+1} + \tilde{\beta}_k d^k \tag{27}$$

Zur weiteren Vereinfachung der Darstellung in Algorithmus 2.15 ersetzen wir wegen Gleichung (25), Gleichung (26) und Gleichung (27) nun die Tilde-Größen $\tilde{x}^k$, $\tilde{r}^k$ und $\tilde{d}^k$ durch die in Gleichung (24) definierten Größen ohne Tilde. Des weiteren wird die symmetrische positiv definite Matrix

$$P = CC^T$$

benutzt, um C aus allen Rechnungen zu eliminieren. Insbesondere wird dazu ein Vektor z^k als Lösung des linearen Gleichungssystems

$$CC^T z = r^k \quad \text{bzw. äquivalent von} \quad Pz = r^k$$

definiert. Man beachte weiter, dass nach Gleichung (24) und Algorithmus 2.15

$$\begin{aligned} d^0 &= (C^T)^{-1}\tilde{d}^0 \\ &= (C^T)^{-1}\tilde{r}^0 \\ &= (C^T)^{-1}C^{-1}r^0 \\ &= P^{-1}r^0 \end{aligned}$$

Schließlich ersetzen wir noch die Bezeichnungen $\tilde{t}_k$ bzw. $\tilde{\beta}_k$ durch t_k bzw. β_k. Damit hat man folgenden Algorithmus:

■ **Algorithmus 2.16 (Vorkonditioniertes CG-Verfahren)**
Input: $x^0 \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, $b \in \mathbb{R}^n$

```
1  r^0 = b - Ax^0;
2  compute z^0 als Lösung von Pz = r^0
3  d^0 = z^0
4  k = 0;
5  while r^k ≠ 0 do
6   t_k = (r^k)^T / ((d^k)^T A d^k);
7   x^{k+1} = x^k + t_k d^k;
8   r^{k+1} = r^k - t_k A d^k;
9   compute z^{k+1} als Lösung von Pz = r^{k+1}
10  β_k = (r^{k+1})^T z^{k+1} / (r^k)^T z^k;
11  d^{k+1} = z^{k+1} + β_k d^k;
12  k = k + 1;
13 enddo
```

Output: $x^* = x^k$

Die Matrix $P = CC^T$ wird als Vorkonditionierer bezeichnet und sollte so gewählt werden, dass einerseits $\text{cond}_2(C^{-1}A(C^T)^{-1})$ möglichst klein und andererseits der Aufwand zur Lösung des linearen Gleichungssystems mit der Systemmatrix P nicht wesentlich größer ist als der Aufwand in einem Schritt des CG-Verfahrens ohne Vorkonditionierung. Um das erste Ziel zu erreichen, wäre zwar $C = A^{1/2}$ optimal, da $\text{cond}_2(C^{-1}A(C^T)^{-1}) = \text{cond}_2(\mathbb{1}) = 1$. Jedoch haben die pro Schritt von Algorithmus 2.16 zu lösenden Gleichungssysteme dann die Systemmatrix $P = CC^T = A$ des Ausgangssystems $Ax = b$.

In der Literatur findet man sowohl einige generelle Ideen, um zur Matrix A einen Vorkonditionierer P zu erhalten als auch spezielle Techniken, die Strukturen bzw. Eigenschaften von A zu berücksichtigen.

Es ist nicht erforderlich, die Matrix P explizit anzugeben, vielmehr genügt es, wenn man die Lösungen der linearen Systeme $Pz = r^k$ (mit wenig Aufwand) bestimmen kann.

Eine Möglichkeit der Vorkonditionierung besteht in der Nutzung einer sogenannten unvollständigen Cholesky-Faktorisierung. Dabei wird eine untere Dreiecksmatrix $\hat{L} \in \mathbb{R}^{n\times n}$ bestimmt, so dass

$$\hat{L}\hat{L}^T \approx A$$

und $\hat{L}$ nur verhältnismäßig wenige Nichtnull-Elemente aufweist. Üblicherweise entsteht bei der klassischen Cholesky-Faktorisierung $A = LL^T$ ein sogenannter *Fill-in*, das heißt L hat deutlich mehr Nichtnull-Elemente als das untere Dreieck von A. Diesen Fill-in versucht man mit einer unvollständigen Cholesky-Faktorisierung zu verringern, vgl. auch entsprechende MATLAB-Routine.

2.6. Ausblick und Anmerkungen

Es gibt zahlreiche weitere sogenannte KRYLOV-Raum basierte Verfahren. Auf einige dieser Methoden wird hier kurz hingewiesen.

- **CGNR-Verfahren für reguläre Matrizen:** Da A regulär ist, muss das System $Ax = b$ äquivalent zu $A^TAx = A^Tb$ sein. Offenbar ist A^TA eine symmetrische und positiv definite Matrix. Damit kann das CG-Verfahren prinzipiell auf $A^TAx = A^Tb$ angewendet werden. Dieses Vorgehen wird als CGNR-Verfahren bezeichnet, dabei gilt jedoch $\text{cond}_2(A^TA) = \text{cond}_2(A)^2$.

- **GMRES-Verfahren für reguläre Matrizen:** Das GMRES-Verfahren von SAAD und SCHULTZ kann dann als Spezialfall des Minimum-Residuum Basisalgorithmus (Algorithmus 2.4) aufgefasst werden, wenn dort $B = \mathbb{1}$ gesetzt wird. Die dabei notwendige Erzeugung und Abspeicherung einer Orthonormalbasis der KRYLOV-Räume $\mathcal{K}_k(r^0, A)$ ist zu aufwendig, so dass man zu einem GMRES(m)-Verfahren übergeht, dass nach m Schritten und einem Restart mit der letzten erhaltenen Iterierten als Startvektor durchführt und dieses Verfahren ggf. mehrfach wiederholt.

- **MINRES-Verfahren für symmetrische reguläre Matrizen:** Theoretisch ist dieses Verfahren von PAIGE und SAUNDERS identisch mit dem GMRES-Verfahren, jedoch gestattet die zusätzliche Symmetrie von A wesentliche Vereinfachungen bei der Lösung der Teilprobleme.

- **BiCG-Verfahren für reguläre Matrizen:** Dieses Verfahren von LANCZOS kann als Realisierung des GALERKIN Basisalgorithmus (Algorithmus 2.6) angesehen werden, wobei die pro Schritt zu lösenden Teilprobleme dann durch

  $$1 \quad \texttt{compute} \;\; x^k \in x^0 + \mathcal{K}_k(r^0, A) \;\; \texttt{mit} \;\; b - Ax^k \perp \mathcal{K}_k(\bar{r}^0, A^T)$$

  gegeben sind mit einem $\bar{r}^0$ derart, dass $(\bar{r}^0)^T r^0 \neq 0$.

- **CG-Verfahren als Galerkin-Verfahren:** Beim CG-Verfahren (Algorithmus 2.4 mit $B = A^{-1/2}$ bzw. Algorithmus 2.16) ergibt sich aus Lemma 2.9

  $$(r^k)^T d^j = 0 \quad \text{für } j = 0, ..., k-1$$

  Außerdem hatten wir induktiv gezeigt (vgl. Gleichung (11)), dass $\mathcal{R}_k = \mathcal{K}_k(r^0, A) = \mathcal{D}_k$ gilt sofern $r^{k-1} \neq 0$. Also folgt

  $$b - Ax^k = r^k \perp \mathcal{K}_k(r^0, A)$$

  und weiter (vgl. die Teilprobleme in Algorithmus 2.4)

  $$x^k \in x^0 + \mathcal{K}_k(r^0, A) \quad \text{mit} \quad b - Ax^k \perp \mathcal{K}_k(r^0, A)$$

  Damit kann man das CG-Verfahren auch als ein GALERKIN-Verfahren betrachten.

- **Abbruchkriterien bei iterativen Lösern:** Anstelle der theoretischen Abbruchbedingung $b -$

$Ax^k = 0$ muss ein praktikableres Kriterium verwendet werden. Dazu bietet sich die Bedingung

$$\frac{\left\|b - Ax^k\right\|}{\|b - Ax^0\|} \leq \varepsilon$$

an das relative Residuum an, wobei $\varepsilon > 0$ vorzugeben ist. Zur Vermeidung einer unendlichen Anzahl von Iterationsschritten sind noch weitere Bedingungen (etwa an die maximale Iterationszahl oder die Differenz $x^k - x^{k-1}$) erforderlich.

Kapitel III

Numerische Behandlung von Anfangswertaufgaben

1. Aufgabe und Lösbarkeit

Es seien $a, b \in \mathbb{R}$ mit $a < b$, eine stetige Funktion $f\colon [a,b] \times \mathbb{R}^m \to \mathbb{R}^m$ und $y^0 \in \mathbb{R}^m$ gegeben. Unter Anfangswertaufgabe (AWA) 1. Ordnung versteht man das Problem, eine stetige Funktion $y\colon [a,b] \to \mathbb{R}^m$ zu ermitteln, so dass y auf (a,b) stetig differenzierbar ist und

$$y'(x) = f(x, y(x)) \quad \text{mit} \quad y(a) = y^0$$

für alle $x \in [a,b]$ gilt. Eine solche Funktion wollen wir Lösung der AWA nennen. Kürzer schreibt man für die AWA auch

$$y' = f(x,y) \quad \text{mit} \quad y(a) = y^0 \tag{1}$$

Die Existenz und Eindeutigkeit einer Lösung einer AWA hängen von den Eingangsinformationen a, b, f und y^0 ab. Es gilt folgender Satz zur (globalen) Existenz und Eindeutigkeit einer Lösung auf $[a,b]$:

Satz 1.1 (Picard-Lindelöf: eine globale Version)
Es sein $f\colon [a,b] \times \mathbb{R}^m \to \mathbb{R}^m$ stetig und es existiere $L > 0$, so dass

$$\|f(x,y) - f(x,z)\| \leq L\|y - z\| \quad \forall (x,y),(x,z) \in [a,b] \times \mathbb{R}^m \tag{2}$$

Dann besitzt Gleichung (1) für jedes $y^0 \in \mathbb{R}^m$ eine eindeutige Lösung.

Die Bedingung Gleichung (2) ist eine globale Lipschitz-Bedingung an f bezüglich der zweiten Veränderlichen. Es ist leicht, AWA anzugeben, in denen diese Bedingung nicht erfüllt ist und keine Lösung in ganz $[a,b]$ existiert, zum Beispiel

$$y' = y^2 \quad \text{mit} \quad y(0) = 1$$

Dafür erhält man für beliebige $x, y, z \in \mathbb{R}$

$$|f(x,y) - f(x,z)| = \left|y^2 - z^2\right| = |y+z||y-z|$$

das heißt die Bedingung Gleichung (2) kann in diesem Beispiel (global) nicht gelten. Die Lösung der AWA lautet $y(x) = {}^{-1}/_{x-1}$ für $x \in [0,1)$. Für Intervalle $[0,b]$ mit $b \geq 1$ existiert keine Lösung. Eine Abschwächung der Lipschitz-Bedingung Gleichung (2) gestattet folgender

Satz 1.2 (Picard-Lindelöf: eine lokale Version)
Es sei f: $[a,b] \times \mathbb{R}^m \to \mathbb{R}^m$ stetig und zu jeder kompakten Menge $\mathcal{Y} \subset \mathbb{R}^m$ existiere $L_Y > 0$, so dass

$$\|f(x,y) - f(x,z)\| \leq L_y \|y - z\| \quad \forall (x,y),(x,z) \in [a,b] \times \mathcal{Y}$$

Dann gibt es für jedes $y^0 \in \mathbb{R}^m$ ein Teilintervall $\mathcal{I} \subseteq [a,b]$ mit $a \in \mathcal{I}$, so dass die AWA Gleichung (1) auf $\mathcal{I}$ eine eindeutige Lösung besitzt.

Seien g: $[a,b] \times \mathbb{R}^n \to \mathbb{R}$ stetig und $\eta \in \mathbb{R}^n$. Jede explizite Differentialgleichung n-ter Ordnung

$$y^{(n)} = g(x, y, y', y'', ..., y^{(n-1)})$$

mit den Anfangsbedingungen

$$y(a) = \eta_1, \quad y'(a) = \eta_2, \quad y''(a) = \eta_3, \quad \dots \quad y^{(n-1)}(a) = \eta_n$$

kann mittels Substitution

$$y_1 = y, \quad y_2 = y', \quad y_3 = y'', \quad \dots \quad y_n = y^{(n-1)}$$

in eine AWA 1. Ordnung überführt werden:

$$\begin{pmatrix} y_1' \\ \vdots \\ y_n' \end{pmatrix} = \begin{pmatrix} y_2 \\ \vdots \\ y_n \\ g(x, y_1, ..., y_n) \end{pmatrix} \quad \text{mit} \quad \begin{pmatrix} y_1(a) \\ \vdots \\ y_n(a) \end{pmatrix} = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix}$$

2. Einschrittverfahren

2.1. Grundlagen

Anstelle der gesuchten Lösungsfunktion $y\colon [a,b] \to \mathbb{R}^m$ einer AWA ist man an möglichst guten Näherungen $y^k \in \mathbb{R}^m$ $(k = 0,1,2,...,N)$ für die Funktionswerte $y(x_k) \in \mathbb{R}^m$ der Funktion y an Gitterpunkten $x_k \in [a,b]$ interessiert. Auf Grundlage der Paare (x_k, y^k) $(k = 0,1,...,N-1)$ ist es auch möglich, eine Näherungsfunktion y zu erzeugen (etwa durch Interpolation).

Einschrittverfahren bilden eine Klasse von Verfahren, die Näherungen y^k zu erzeugen. Das Gitter $\{x_0,...,x_N\}$ is so gewählt, dass

$$x_0 = a < x_1 < x_2 < \cdots < x_{N-1} < x_N = b$$

Außerdem setzen wir

$$h_k = x_{k+1} - x_k \quad \text{für} k = 0,...,N-1$$

und bezeichnen h_k als Schrittweite. Falls $h = h_0 = \cdots = h_{N-1}$, so heißen die Gitterpunkte bzw. das Gitter gleichabständig oder äquidistant.

Ein Verfahren zur Erzeugung einer Folge $y^0,...,y^N$ heißt Einschrittverfahren für das AWA Gleichung (1), wenn

$$y^{k+1} = y_k + h_k \Phi(x_k, y_k, y^{k+1}, h_k) \quad \text{für} k = 0,...,N-1 \tag{1}$$

Dabei bezeichnet $\Phi(x,y,z,h)$ den Funktionswert einer Verfahrensfunktion

$$\Phi : [a,b] \times \mathbb{R}^m \times \mathbb{R}^m \times (0, b-a] \to \mathbb{R}^m$$

die das jeweilige Einschrittverfahren definiert. Man beachte, dass y^0 bereits durch die Anfangsbedingung in Gleichung (1) gegeben ist. Ein Einschrittverfahren heißt implizit, falls Φ tatsächlich von z abhängt. Dann ist zur Bestimmung von y^{k+1} aus Gleichung (1) die Lösung eines im Allgemeinen nichtlinearen Gleichungssystems erforderlich. Falls Φ nicht von z abhängt, heißt das Einschrittverfahren explizit. Das explizite Euler-Verfahren (auch Polygonzugverfahren genannt) ist gegeben durch

$$\Phi(x,y,z,h) = f(x,y) \tag{2}$$

das heißt

$$y^{k+1} = y^k + h_k f(x_k, y^k)$$

Für das implizite Euler-Verfahren gilt die Vorschrift

$$y^{k+1} = y^k + h_k f(x_k + h_k, y^{k+1})$$

Um die Güte der Näherungen y^k zu beurteilen, untersuchen wir zunächst den lokalen Diskretisierungsfehler eines Einschrittverfahrens.

2.2. Lokaler Diskretisierungsfehler und Konsistenz

Definition 2.1 (lokaler Diskretisierungsfehler)

Seien $y\colon [a,b] \to \mathbb{R}^m$ Lösung des Differentialgleichung $y' = f(x,y)$ und Φ die Verfahrensfunktion eines Einschrittverfahrens. Für $x \in [a,b)$ und $h > 0$ mit $x + h \leq b$ heißt

$$\Delta(x,h) = y(x+h) - \Big(y(x) + h\Phi\big(x, y(x), y(x+h), h\big)\Big) \tag{3}$$

lokaler Diskretisierungsfehler und

$$\frac{\Delta(x,h)}{h} = \frac{y(x+h) - y(x)}{h} - \Phi(x, y(x), y(x+h), h) \tag{4}$$

relativer lokaler Diskretisierungsfehler des Einschrittverfahrens.

Der lokale Diskretisierungsfehler gibt also die Abweichung zwischen exakter Lösung $y(x+h)$ an der Stelle $x+h$ und der Näherung an dieser Stelle an, wobei angenommen wird, dass die Näherung unter Verwendung der exakten Lösung $y(x)$ (und ggf. $y(x+h)$) berechnet wird. Die Bezeichnung relativer Diskretisierungsfehler ist bezüglich der Schrittweite h zu verstehen.

Definition 2.2 (konsistent, Konsistenzordnung)

Ein Einschrittverfahren heißt konsistent zur Differentialgleichung $y' = f(x,y)$, wenn

$$\lim_{h \downarrow 0} \left\| \frac{\Delta(x,h)}{h} \right\| = 0 \quad \forall x \in [a,b)$$

für jede Lösung $y\colon [a,b] \to \mathbb{R}^m$ der Differentialgleichung gilt. Gibt es außerdem $p \geq 1$, $M > 0$, $\tilde{h} > 0$, so dass

$$\left\| \frac{\Delta(x,h)}{h} \right\| \leq Mh^p \quad \forall (x,h) \in [a,b) \times (0,\tilde{h}) \text{ mit } x + h \leq b$$

für jede Lösung $y\colon [a,b] \to \mathbb{R}^m$ der Differentialgleichung gilt, so hat das Einschrittverfahren (für diese Differentialgleichung) die Konsistenzordnung p.

Satz 2.3

Sei $f\colon [a,b] \times \mathbb{R}^m \to \mathbb{R}^m$ stetig differenzierbar. Dann hat das explizite Euler-Verfahren die Konsistenzordnung 1.

Beweis. Mit Gleichung (2) folgt

$$\Delta(x,h) = y(x+h) - y(x) - hf(x,y(x))$$

Da y die Differentialgleichung $y' = f(x,y)$ löst und f stetig differenzierbar ist, muss y zweimal stetig differenzierbar sein. Aus der Taylor-Formel erhält man für $i \in \{1, ..., m\}$

$$\begin{aligned} \Delta(x,h)_i &= y_i'(x)h + \frac{1}{2}y_i''(\xi_i(x,h))h^2 - hf_i(x,y(x)) \\ &= \frac{1}{2}y_i''(\xi_i(x,h))h^2 \end{aligned}$$

für ein $\xi_i(x,h) \in (x, x+h)$. Die Stetigkeit von y'' auf $[a,b]$ und Division durch h liefert die Behauptung mit $M = \frac{1}{2}\max_{1\leq i\leq m}\max_{\xi\in[a,b]} \|y_i''(\xi)\|$ und $\tilde{h} = b - a$. □

2.3. Konvergenz von Einschrittverfahren

Zum Gitter $G = \{x_0, ..., x_N\} \subset [a, b]$ mit $x_0 = a$ und $x_N = b$ seien $y^0, ..., y^N \in \mathbb{R}^m$ durch ein Einschrittverfahren erzeugt. Weiter bezeichne y: $[a, b] \to \mathbb{R}^m$ die eindeutige Lösung der AWA Gleichung (1). Dann seien

$$e(x_k) = y(x_k) - y^k$$
$$e(G) = \max_{x \in G} \|e(x)\|$$

sowie

$$h_{max}(G) = \max_{k=0,...,N-1} h_k$$

definiert.

Definition 2.4 (konvergent)
Die AWA Gleichung (1) besitzt die eindeutige Lösung y: $[a, b] \to \mathbb{R}^m$. Ein Einschrittverfahren für diese AWA heißt dann konvergent, falls

$$\lim_{l \to \infty} e(G_l) = 0$$

für alle Gitterfolgen $\{G_t\}$ gilt, für die $\lim_{l \to \infty} h_{max}(G_l) = 0$. Gibt es außerdem $p \geq 1$, $C > 0$, $\tilde{h} > 0$, so dass

$$e(G) \leq C \cdot h_{max}(G)^p$$

für jedes Gitter mit $h_{max}(G) \leq \tilde{h}$, so hat das Einschrittverfahren für die gegebene AWA die Konvergenzordnung p.

Lemma 2.5 (diskretes Grönwallsches Lemma)
Falls die Zahlenfolgen $\{\alpha_k\}$, $\{\beta_k\}$, $\{v_k\} \subset [0, \infty)$ den Bedingungen

$$v_0 = 0 \quad \text{und} \quad v_{k+1} = (1 + \alpha_k)v_k + \beta_k \quad \forall k = 0, ..., N - 1$$

genügen, dann folgt

$$v_{k+1} \leq \sum_{i=0}^{k} \beta_i \cdot \exp\left(\sum_{j=i+1}^{k} \alpha_j \right) \quad \text{für } k = 0, ..., N - 1$$

gilt zusätzlich $\alpha_k = \alpha > 0$ und $\beta_k = \beta > 0$ für jedes $k = 0, ..., N - 1$, dann folgt

$$v_k \leq \frac{\beta}{\alpha}(\exp(k\alpha) - 1) \quad \text{für } k = 0, ..., N - 1$$

Beweis. Zum Beispiel durch vollständige Induktion (vgl. Übungsaufgabe). □

In der Literatur findet man für vorstehende und ähnliche Aussagen die Bezeichnung *diskretes* GRÖNWALL*sches Lemma*.

Satz 2.6
Die AWA Gleichung (1) besitze die eindeutige Lösung y: $[a,b] \to \mathbb{R}^m$. Ein Einschrittverfahren mit der Verfahrensfunktion Φ habe für die Differentialgleichung $y' = f(x,y)$ die Konsistenzordnung p. Es gebe ferner $L_\Phi > 0$ und $H > 0$, so dass die Lipschitz-Bedingung

$$\|\Phi(x,y,z,h) - \Phi(x,\tilde{y},\tilde{z},h)\| \leq L_\Phi(\|y-\tilde{y}\| + \|z-\tilde{z}\|) \tag{5}$$

für alle $(x,y,z,h),(x,\tilde{y},\tilde{z},h) \in [a,b]\times\mathbb{R}^m\times\mathbb{R}^m\times(0,H]$ gilt. Dann besitzt das Einschrittverfahren die Konvergenzordnung p.

Beweis. Entsprechend Gleichung (1) und Gleichung (3) gilt

$$y^{k+1} = y^k + h_k\Phi(x_k, y^k, y^{k+1}, h_k)$$

und

$$y(x_{k+1}) = y(x_k) + h_k\Phi(x_k, y(x_k), y(x_k+h_k), h_k) + \Delta(x_k,h_k)$$

also folgt

$$\begin{aligned} e(x_{k+1}) &= y(x_{k+1}) - y^{k+1} \\ &= y(x_k) - y^k + h_k\Big(\Phi((x_k,y(x_k),y(x_k+h_k),h_k) - \Phi(x_k,y^k,y^{k+1},h_k)\Big) + \Delta(x_k,h_k) \end{aligned}$$

und weiter mit Gleichung (5) für $0 < h_k \leq \tilde{h} = \min\{H, \tilde{h}, \frac{1}{2L_\Phi}\}$

$$\|e(x_{k+1})\| \leq \|e(x_k)\| + \|\Delta(x_k,h_k)\| + h_kL_\Phi(\|e(x_k)\| + \|e(x_{k+1})\|)$$

Durch Umstellen und Beachtung der Konsistenzordnung ergibt sich

$$\|e(x_{k+1})\| \leq \frac{1+h_kL_\Phi}{1-h_kL_\Phi}\|e(x_k)\| + \frac{M}{1-h_kL_\Phi}h_k^{p+1} \tag{6}$$

Mit $\alpha_k = 4h_kL_\Phi$ hat man (wegen $2h_kL\Phi \leq 1$)

$$\frac{1+h_kL_\Phi}{1-h_kL_\Phi} = 1 + \frac{2h_kL_\Phi}{1-h_kL_\Phi} \leq 1+\alpha_k$$

Setzt man weiter $v_k = \|e(x_k)\|$ und $\beta_k = 2Mh_k^{p+1}$, so erhält man aus Gleichung (6)

$$v_{k+1} \leq (1+\alpha_k)v_k + \beta_k \quad k = 0,...,N-1$$

Nach Lemma 2.5 folgt daraus (beachte $v_0 = \|e(x_0)\| = \left\|y(x_0) - y^0\right\| = 0$)

$$v_{k+1} \leq \left(\sum_{i=0}^{k}\beta_i\right)\exp\left(\sum_{i=0}^{k}\alpha_i\right) \quad \text{für } k = 0,...,N-1$$

und damit

$$\begin{aligned} \|e(x_{k+1})\| = v_{k+1} &\leq 2M\left(\sum_{i=0}^{k}h_i^{p+1}\right)\exp\left(4L_\Phi\sum_{i=0}^{k}h_i\right) \\ &\leq 2Mh_{max}(G)^p(x_{k+1}-x_0)\exp(4L_\Phi(x_{k+1}-x_0)) \end{aligned}$$

für $k = 0, ..., N-1$ sowie

$$e(G) \leq 2M(b-a)\exp(4L_\Phi(b-a))h_{max}(G)^p$$

Also besitzt das Einschrittverfahren die Konvergenzordnung p. □

2.4. Stabilität gegenüber Rundungsfehlern

Wir betrachten das Einschrittverfahren Gleichung (1) für ein gleichabständiges Gitter ($h_k = h$) bei exakter Rechnung, das heißt

$$y^{k+1} = y^k + h\Phi(x_k, y^k, y^{k+1}, h) \quad \text{für } k = 0, ..., N-1 \tag{7}$$

Weiter beschreibe

$$\tilde{y}^0 = y^0 \quad \text{und} \quad \tilde{y}^{k+1} = \tilde{y}^k + h\Phi(x_k, \tilde{y}^k, \tilde{y}^{k+1}, h) + \varepsilon_k \quad \text{für } k = 0, ..., N-1 \tag{8}$$

ein gestörtes Verfahren, das heißt $\tilde{y}^1, ..., \tilde{y}^N$ sind die tatsächlich im Computer berechneten Größen.

Satz 2.7
Sei $y^o \in \mathbb{R}^m$ gegeben und $y^1, ..., y^N$ sowie $\tilde{y}^1, ..., \tilde{y}^N$ entsprechend Gleichung (7) und Gleichung (8) berechnet, wobei $\|\varepsilon_k\| < \varepsilon$ für alle $k = 0, ..., N-1$ mit einem $\varepsilon > 0$ gelte. Außerdem sei für gewisse $L_\Phi, H > 0$ die Lipschitz-Bedingung Gleichung (5) für alle $(x, y, z, h), (x, \tilde{y}, \tilde{z}, h) \in [a, b] \times \mathbb{R}^m \times \mathbb{R}^m \times (0, H]$ erfüllt. Dann gibt es $\tilde{h} > 0$, so dass

$$\left\|y^k - \tilde{y}^k\right\| \leq \frac{\varepsilon}{2hL_\Phi}(\exp(4L_\Phi(x_k - a)) - 1) \quad \text{für } k = 0, ..., N$$

falls $0 > h < \tilde{h}$.

Beweis. Für $z^k = y^k - \tilde{y}^k$ folgt

$$\begin{aligned} z^{k+1} &= y^{k+1} - \tilde{y}^{k+1} \\ &= y^k - \tilde{y}^k + h\big(\Phi(x_k, y^k, y^{k+1}, h) - \Phi(x_k, \tilde{y}^k, \tilde{y}^{k+1}, h)\big) - \varepsilon_k \end{aligned}$$

und weiter

$$\left\|z^{k+1}\right\| \leq \left\|z^k\right\| + hL_\Phi(\left\|z^k\right\| + \left\|z^{k+1}\right\|) + \varepsilon$$

Mit $v_k = \left\|z^k\right\|$, $\alpha = 4hL_\Phi$ und $\beta = 2\varepsilon$ hat man für $0 < h \leq \tilde{h} = \min H, \frac{1}{2L_\Phi}$ die Differenzenungleichung

$$v_{k+1} \leq (1+\alpha)v_k + \beta$$

für $k = 0, ..., N-1$. Lemma 2.5 liefert

$$\begin{aligned} \left\|y^k - \tilde{y}^k\right\| = \left\|z^k\right\| &= v_k \\ &\leq \frac{\varepsilon}{2hL_\Phi}(\exp(k4hL_\Phi) - 1) \\ &= \frac{\varepsilon}{2hL_\Phi}(\exp(4L_\Phi(x_k - a)) - 1) \end{aligned}$$

□

Selbst wenn die Abschätzung in Satz 2.7 nicht scharf ist, muss man damit rechnen, dass der Rundungs-

fehler wie $1/h$ wächst. Der Gesamtfehler eines Einschrittverfahrens ans einer Stelle x_k setzt sich aus dem Verfahrensfehler $\|y(x_k) - y^k\|$ und dem Rundungsfehler $\|y^k - \tilde{y}^k\|$ zusammen. Für ein Verfahren der Konvergenzordnung p ergibt sich also (bei äquidistantem Gitter) für den Gesamtfehler

$$\begin{aligned}\|y(x_k) - \tilde{y}^k\| &\leq \|y(x_k) - y^k\| + \|y^k - \tilde{y}^k\| \\ &\leq Ch^p + \tilde{C}\frac{\varepsilon}{h}\end{aligned}$$

Minimiert man die rechte Seite der Abschätzung in Abhängigkeit von h, so folgt, dass man h nicht kleiner als $\sim \sqrt[p+1]{\varepsilon}$ wählen sollte. Setzt man speziell $h = \sqrt[p+1]{\varepsilon}$, dann folgt

$$Ch^p + \tilde{C}\frac{\varepsilon}{h} = C\exp\left(\frac{p}{p+1}\right) + \tilde{C}\exp\left(\frac{p}{p+1}\right)$$

Durch Erhöhung der Konvergenzordnung p kann man also versuchen, mit einer größeren Schrittweite einen kleineren Gesamtfehler zu erreichen. Ein weiterer Grund für das Interesse an Verfahren mit höherer Konvergenzordnung liegt in der Möglichkeit, die Gesamtzahl der erforderlichen Funktionswertbestimmungen der Funktion f zu verringern.

2.5. Runge-Kutta-Verfahren

Die Klasse der Runge-Kutta-Verfahren (RKV) ist eine Möglichkeit, Einschrittverfahren mit höheren Konsistenz- bzw. Konvergenzordnungen zu konstruieren. Betrachten wir folgende Idee, eine Näherung y^{k+1} für $y(x_{k+1})$ aus einer Näherung y^k für $y(x_k)$ zu erzeugen.

Wegen $y' = f(x, y)$ liefert der Hauptsatz der Differential- und Integralrechnung

$$y(x_{k+1}) = y(x_k) + \int_{x_k}^{x_{k+1}} f(x, y(x))\,\mathrm{d}x \tag{9}$$

Approximiert man das Integral durch eine gewichtet Summe von Funktionswerten (vgl. Newton-Cotes Formeln), so ergibt sich die folgende Verfahrensidee

$$y^{k+1} = y^k + h_k \sum_{i=1}^{s} c_i f(s_i, y(s_i)) \tag{10}$$

wobei $c_1, ..., c_s$ die Gewichte und $s_1, ..., s_s$ Stützstellen bezeichnen. Zur Darstellung der Stützstellen sei

$$s_i = x_k + \alpha_i h_k \quad \text{für } i = 1, ..., s$$

mit $\alpha_1 = 0$ und den Parametern $\alpha_2, ..., \alpha_s$. Da $y(s_i)$ unbekannt ist, ersetzt man $f(s_i, y(s_i))$ zunächst durch einen Parameter k^i, wobei $k^1 = f(x_k, y^k) \approx f(x_k, y(x_k))$ gesetzt wird. Um $y(s_i)$ und damit $f(s_i, y(s_i))$ zu approximieren, verwendet man (bei expliziten RKV) den Ansatz

$$y(s_i) \approx y^k + h_k \sum_{j=1}^{i-1} \beta_{ij} k^j$$

mit Parametern β_{ij}. Bei sogenannten impliziten RKV läuft die Summation von $j = 1$ bis $j = s$ (und mindestens ein β_{ij} mit $j \geq 1$ ist ungleich 0). Für die Parameter α_i, k^i, β_{ij} ergibt sich (im expliziten

Fall) somit das folgende Gleichungssystem

$$\begin{aligned} k^1 &= f(x_k, y^k) \\ k^2 &= f(x_k + \alpha_2 h_k, y^k + h_k \beta_{21} k^1) \\ k^3 &= f(x_k + \alpha_3 h_k, y_k + h_k(\beta_{31} k^1 + \beta_{32} k^2)) \\ &\vdots \\ k^s &= f(x_k + \alpha_s h_k, y^k + h_k(\beta_{s1} k^1 + \cdots + \beta_{s,s-1} k^{s-1})) \end{aligned} \tag{11}$$

Ersetzt man in Gleichung (10) die unbekannten Vektoren $f(s_i, y(s_i))$ durch die Näherungen k^i, so hat man das s-stufige RUNGE-KUTTA-Verfahren

$$y^{k+1} = y^k + h_k \sum_{i=1}^{s} c_i k^i \tag{12}$$

mit den Parametern $c_1, ..., c_s$. Die Verfahrensfunktion eines expliziten RKV ist damit gegeben durch

$$\Phi(x, y, h) = \sum_{i=1}^{s} c_i f\left(x + \alpha_i h_i y + h \sum_{j=1}^{i-1} \beta_{ij} k^j(x, y, h)\right)$$

wobei $k^i = k^i(x, y, h)$ entsprechend Gleichung (11) verwendet wird (die bei expliziten Verfahren nicht vorhandene Abhängigkeit der Funktion Φ von z wurde weggelassen). Zum Beispiel ist das explizite EULER-Verfahren $y^{k+1} = y^k + h_k f(x_k, y^k)$ ein einstufiges RKV mit $c_1 = 1$.

Satz 2.8
Sei f: $[a, b] \times \mathbb{R}^m \to \mathbb{R}^m$ stetig differenzierbar. Ein explizites RKV Gleichung (12) mit

$$\sum_{i=1}^{s} c_i = 1 \tag{13}$$

hat dann (mindestens) die Konsistenzordnung 1.

Beweis. Sei $x \in \mathbb{R}$ fest gegeben. Weiter sei $\eta = y(x)$. Da f stetig differenzierbar ist, gibt es $L_f > 0$, so dass

$$\|f(x, \eta) - f(x + \delta x, \eta + \delta\eta)\| \leq L_f(|\delta x| + \|\delta\eta\|)$$

für alle $(\delta x, \delta\eta) \in \mathbb{R} \times \mathbb{R}^m$ mit $|\delta x| + \|\delta\eta\| \leq 1$. Induktiv folgt damit, dass $\tilde{h} > 0$ existiert, so dass

$$\left\|k^i(x, \eta, h) - f(x, \eta)\right\| = \mathcal{O}(h) \quad \forall h \in [0, \tilde{h}]$$

für alle $i = 1, ..., s$. Also gilt wegen Gleichung (13)

$$\|\Phi(x, \eta, h) - f(x, \eta)\| = \mathcal{O}(h) \quad \forall h \in [0, \tilde{h}]$$

Daraus erhält man (da f in $[a, b]$ stetig differenzierbar und somit y zweimal stetig differenzierbar ist, vgl. Beweis

zu Satz 2.3)

$$\left\| \frac{\Delta(x,h)}{h} \right\| = \left\| \frac{y(x+h)-y(x)}{h} - f(x,y(x)) + f(x,y(x)) - \Phi(x,y(x),h) \right\|$$
$$\leq \mathcal{O}(h)$$

□

Satz 2.9
Sei $f\colon [a,b] \times \mathbb{R}^m \to \mathbb{R}^m$ zweimal stetig differenzierbar. Ein explizites RKV Gleichung (12) mit Gleichung (13),

$$\sum_{j=1}^{i-1} \beta_{ij} = \alpha_i \quad \text{für } i = 2, ..., s \tag{14}$$

und

$$\sum_{i=2}^{s} c_i \alpha_i = \tfrac{1}{2} \tag{15}$$

hat dann (mindestens) die Konvergenzordnung 2.

Beweis. Übungsaufgabe □

Verwendet man zur Approximation des bestimmten Integrals in Gleichung (9) die Trapezregel, das heißt

$$\int_{x_k}^{x_{k+1}} f(x,y(x))\,\mathrm{d}x \approx \frac{1}{2}\Big(f(x_k, y(x_k)) + f(x_k + h_k, y(x_k + h_k)) \Big)$$

und ersetzt man $f(x_k, y(x_k))$ und $f(x_k + h_k, y(x_k + h_k))$ im RKV durch $k^1 = f(x_k, y^k)$ bzw. $k^2 = f(x_k + h_k, y^k + h_k k^1)$, dann ergibt sich ein 2-stufiges RKV mit

$$c_1 = c_2 = \frac{1}{2}, \quad \alpha_2 = 1, \quad \beta_{21} = 1$$

das heißt die Bedingungen Gleichung (13), Gleichung (14) und Gleichung (15) sind erfüllt. Also besitzt dieses explizite RKV die Konsistenzordnung 2. es ist als Verfahren von Heun bekannt.

Verwendet man zur Quadratur des Integrals in Gleichung (9) die Simpson-Regel, das erhält man ein 4-stufiges RKV mit folgenden Parametern (im sogenannten Butcher-Schema)

0					0				
α_2	β_{21}				$1/2$	$1/2$			
α_3	β_{31}	β_{32}			$1/2$	0	$1/2$		
α_4	β_{41}	β_{42}	β_{43}		1	0	0	1	
	c_1	c_2	c_3	c_4		$1/6$	$1/3$	$1/3$	$1/6$

Dieses Verfahren hat die Konsistenzordnung 4 (sofern f hinreichend glatt).

3. Mehrschrittverfahren

3.1. Grundlagen

Bei Mehrschrittverfahren (MSV) wird eine Näherung y^{k+l} für $y(x_{k+l})$ in bestimmter Weise aus l vorhergehenden Näherungen $y^k, y^{k+1}, ..., y^{k+l-1}$ bestimmt. Um dies genau zu beschreiben, seien zusätzlich zu y^0 (aus AWA) die Startwerte $y^1, ..., y^{l-1} \in \mathbb{R}^m$ gegeben. Im Folgenden wollen wir von einem äquidistanten Gitter $G_h = \{x_0, ..., x_N\}$ mit Schrittweite $h = \frac{b-a}{N}$ ausgehen. Ein lineares Mehrschrittverfahren mit l Schritten erzeugt dann für $k = 0, ..., N-l$ die Iterierte y^{k+l} aus $y^k, y^{k+1}, ..., y^{k+l-1}$ entsprechend

$$\sum_{\nu=0}^{l} \alpha_\nu y^{k+\nu} = h \sum_{\nu=0}^{l} \beta_\nu f(x_{k+\nu}, y^{k+\nu}) \tag{1}$$

wobei α_ν, β_ν ($\nu = 0, ..., l$) reelle Parameter sind mit $\alpha_l \neq 0$ und $|\alpha_0| + |\beta_0| \neq 0$. Falls $\beta_l = 0$, dann spricht man von einem expliziten (sonst impliziten) linearen MSV. Die MSV Gleichung (1) heißen linear, da die rechte Seite von Gleichung (1) linear von den Funktionswerten $f(x_{k+\nu}, y^{k+\nu})$ abhängt. Einem linearen MSV ordnet man sein erstes und zweites charakteristisches Polynom $\rho : \mathbb{C} \to \mathbb{C}$ und $\sigma : \mathbb{C} \to \mathbb{C}$ zu durch

$$\rho(z) = \sum_{\nu=0}^{l} \alpha_\nu z^\nu \quad \text{und} \quad \sigma(z) = \sum_{\nu=0}^{l} \beta_\nu z^\nu \quad \forall z \in \mathbb{C} \tag{2}$$

Das lineare MSV nach Adams-Bashford (1883) geht von

$$y(x_{k+l}) - y(x_{k+l-1}) = \int_{x_{k+l-1}}^{x_{k+l}} f(x, y(x))\,\mathrm{d}x \tag{3}$$

aus und approximiert den Integranden $f(x, y(x))$ durch ein Interpolationspolynom, nämlich

$$\sum_{\nu=0}^{l-1} L_\nu(x) f(x_{k+\nu}, y(x_{k+\nu})) \tag{4}$$

Dabei bezeichnen $L_\nu : \mathbb{R} \to \mathbb{R}$ für $\nu = 0, ..., l-1$ die Langrange-Polynome mit

$$L\nu(x) = \prod_{\substack{i=k \\ i \neq k+\nu}}^{k+l-1} \frac{x - x_i}{x_{k+\nu} - x_i} \quad \text{für } x \in \mathbb{R}$$

Definiert man β_ν durch

$$\int_{x_{k+l-1}}^{x_{k+l}} L_\nu(x)\,\mathrm{d}x = h\beta_\nu \tag{5}$$

so liefert die Approximation von Gleichung (3) die Näherungsformel

$$\begin{aligned} y^{k+l} - y^{k+l-1} &= \sum_{\nu=0}^{l-1} \left(\int_{x_{k+l-1}}^{x_{k+l}} L_\nu(x)\,\mathrm{d}x \right) f(x_{k+\nu}, y^{k+\nu}) \\ &= h \sum_{\nu=0}^{l-1} \beta_\nu f(x_{k+\nu}, y^{k+\nu}) \end{aligned}$$

also ein explizites l-schrittiges lineares MSV mit $\alpha_l = 1$, $\alpha_{l-1} = -1$ und den durch Gleichung (5) definierten $\beta_0, ..., \beta_{l-1}$ sowie $\beta_l = 0$.

Beim linearen MSV nach ADAMS-MOULTON (1926) wird die Summation in Gleichung (4) von $\nu = 0$ bis $\nu = l$ erstreckt und dann analog vorgegangen. Dies ergibt das implizite lineare MSV

$$y^{k+l} - y^{k+l-1} = h \sum_{\nu=0}^{l} \beta_\nu f(x_{k+\nu}, y^{k+\nu}) \tag{6}$$

Es erfolgt die (ggf. näherungsweise) Lösung eines im Allgemeinen nichtlinearen Gleichungssystems für y^{k+l} und kann mit Hilfe des Prädiktor-Korrektor-Prinzips erfolgen. Dabei ermittelt man mit Hilfe eines expliziten linearen MSV (Prädiktor) eine erste Näherung ζ^0 für y^{k+l} und verbessert diese dann mit einem (näherungsweisen) Schritt eines impliziten linearen MSV (Korrektor). Zum Beispiel bestimme man ζ^0 mit ADAMS-BASHFORD, das heißt

$$\zeta^0 = y^{k+l} + h \sum_{\nu=0}^{l-1} \beta_\nu f(x_{k+\nu}, y^{k+\nu})$$

Danach wird eine Näherungslösung von Gleichung (6) (ADAMS-MOULTON) etwa mittels Fixpunktiteration ermittelt

$$\zeta^j = y^{k+l-1} + h\beta_l^C f(x_{k+l}, \zeta^{j-1}) + h \sum_{\nu=0}^{l-1} \beta_\nu^C f(x_{k+\nu}, y^{k+\nu})$$

die für ein vorgegebenes $j \geq 1$ abgebrochen wird. Die Bezeichnung β_ν^C dient der Unterscheidung von den im Prädiktor verwendeten Parametern β_ν. Für $j = 1$ ergibt sich ein nichtlineares MSV (ADAMS-BASHFORD-MOULTON-Verfahren). Für $j \to \infty$ kann unter bestimmten Voraussetzungen für hinreichend kleine $h > 0$ die Konvergenz der Folge $\{\zeta^j\}$ gegen den eindeutigen Fixpunkt y^{k+l} gezeigt werden.

Eine Klasse von impliziten linearen MSV (sogenannte Backward Differentiation Formulas bzw. BDF-Verfahren) erhält man aus der Idee $y'(x_{k+l}) = f(x_{k+l}, y(x_{k+l}))$ durch $\frac{1}{h}\sum_{\nu=0}^{l} \alpha_\nu y(x_{k+\nu})$ (verallgemeinerte Sekantensteigung) zu approximieren. Man hat dann ein lineares MSV der Form

$$\sum_{\nu=0}^{l} \alpha_\nu y^{k+\nu} = h f(x_{k+l}, y^{k+l})$$

3.2. Konsistenz- und Konvergenzordnung für lineare MSV

Die Konsistenzordnung eines linearen MSV wird in Analogie zur entsprechenden Definition bei den ESV eingeführt. Verallgemeinerungen für beliebige MSV werden hier nicht betrachtet. Der lokale Diskretisierungsfehler eines MSV ergibt sich zu

$$\begin{aligned}\Delta(x,h) &= y(x+lh) - \frac{1}{\alpha_l}\left(-\sum_{\nu=0}^{l-1}\alpha_\nu y(x+\nu h) + h\sum_{\nu=0}^{l}\beta_\nu f(x+\nu h, y(x+\nu h))\right) \\ &= \frac{1}{\alpha_l}\left(\sum_{\nu=0}^{l}\alpha_\nu y(x+\nu h) - h\sum_{\nu=0}^{l}\beta_\nu f(x+\nu h, y(x+\nu h))\right)\end{aligned} \tag{7}$$

Wenn es also $p \geq 1$, $M > 0$ und $\tilde{h} > 0$ gibt, so dass

$$\left\| \frac{\Delta(x,h)}{h} \right\| \leq Mh^p \quad \forall (x,h) \in [a,b) \times (0,\tilde{h}] \text{ mit } x+h \leq b$$

für jede Lösung $y\colon [a,b] \to \mathbb{R}^m$ der Differentialgleichung $y' = f(x,y)$ gilt, dann sagt man, dass das MSV die Konsistenzordnung $p \geq 1$ besitzt. Unter der Voraussetzung, dass f und damit die Lösungen y der Differentialgleichung hinreichend glatt sind, gelten die Entwicklungen

$$y(x+\nu h) = \sum_{q=0}^{p} \frac{\nu^q}{q!} y^{(q)}(x) h^q + R_p(x,h)$$

und (mit $\frac{\mathrm{d}y(x+\nu h)}{\mathrm{d}h} = y'(x+\nu h)\nu$)

$$f(x+\nu h, y(x+\nu h)) = y'(x+\nu h) = \sum_{q=1}^{p} \frac{\nu^{q-1}}{(q-1)!} y^{(q)}(x) h^{q-1} + r_p(x,h)$$

wobei für die Restglieder $\|R_p(x,h)\| \leq M_R h^{p+1}$ und $\|r_p(x,h)\| \leq M_r h^p$ bei festem x mit gewissen Konstanten $M_R, M_r > 0$ gilt. Aus Gleichung (7) hat man daher

$$\alpha_l \Delta(x,h) = c_0 y(x) + \sum_{q=1}^{p} c_q y(x)^{(q)} h^q + Q(x,h) \quad \text{mit} \quad \|Q(x,h)\| \leq M_Q h^{p+1}$$

wobei $M_Q > 0$ sowie

$$c_0 = \sum_{\nu=0}^{l} \alpha_\nu \quad \text{und} \quad c_q ? \sum_{\nu=0}^{l} \left(\frac{\nu^q \alpha_\nu}{q!} - \frac{\nu^{q-1}\beta_\nu}{(q-1)!} \right) \quad \text{für } q = 1, \ldots, p \tag{8}$$

(mit $0^0 = 1$). Falls $c_0 = \cdots = c_p = 0$, folgt damit

$$\left\| \frac{\Delta(x,h)}{h} \right\| \leq \alpha_l M_Q h^p$$

Also gilt

Satz 3.1
Die Funktion $f\colon [a,b] \times \mathbb{R}^m \to \mathbb{R}^m$ sei p mal stetig differenzierbar. Dann hat das MSV Gleichung (1) (mindestens) die Konsistenzordnung p, wenn $c_0 = \cdots = c_p = 0$.

Unter Beachtung von Gleichung (2) und Gleichung (8) gilt $c_0 = c_1 = 0$ (und damit entsprechend Satz Satz 3.1 Konsistenzordnung $p \geq 1$), genau dann wenn

$$\rho(1) = 0 \quad \text{und} \quad \rho'(1) - \sigma(1) = 0$$

Ein l-schrittiges explizites lineares MSV Gleichung (1) hat die $2l$ freien Parameter $\alpha_0, \ldots, \alpha_{l-1}$ und $\beta_0, \ldots, \beta_{l-1}$ (o.B.d.A. kann $\alpha_l = 1$ gewählt werden, β_l kommt nicht vor, da MSV explizit sein sollte). Durch geeignete Wahl dieser Parameter könnte man $c_0 = 0, \ldots, c_{2l-1} = 0$ erreichen und damit die Konsistenzordnung $p = 2l - 1$. Wie wir aber sehen werden, sind solche Verfahren im Allgemeinen nicht

konvergent.

■ **Beispiel 3.2**

Sei $l = 2$ und $m = 1$. Dann lautet Gleichung (1)

$$y_{k+2} + \alpha_1 y_{k+1} + \alpha_0 y_k = h\big(\beta_1 f(x_{k+1}, y_{k+1}) + \beta_0 f(x_k, y_k)\big)$$

Um $c_0 = c_1 = c_2 = c_3 = 0$ und damit Konsistenzordnung 3 zu erreichen, muss man also $\alpha_0, \alpha_1, \beta_0, \beta_1$ so wählen, dass (mit $\alpha_2 = 1$ und $\beta_2 = 0$)

$$\begin{aligned}
\alpha_0 + \quad \alpha_1 \qquad\qquad + 1 = 0 \quad & c_0 \\
\alpha_1 - \beta_0 - \quad \beta_1 + 2 = 0 \quad & c_1 \\
\tfrac{1}{2}\alpha_1 \qquad - \quad \beta_1 + 2 = 0 \quad & c_2 \\
\tfrac{1}{6}\alpha_1 \qquad - \tfrac{1}{2}\beta_1 + \tfrac{4}{3} = 0 \quad & c_3
\end{aligned}$$

gilt. Die Lösung dieses Systems ist gegeben durch $\alpha_0 = -5$, $\alpha_1 = 4$, $\beta_0 = 2$ und $\beta_1 = 4$. Also besitzt das MSV

$$y_{k+2} + 4y_{k+1} - -5y_k = h\big(2f(x_k, y_k) + 4f(x_{k+1}, y_{k+1})\big) \tag{9}$$

die Konsistenzordnung 3. Für das Testproblem

$$y' = -y \quad \text{mit} \quad y(0) = 1$$

lautet die Lösung $y(x) = \exp(-x)$. Das Verfahren Gleichung (9) geht wegen $f(x, y) = -y$ über die homogene lineare Differentialgleichung mit konstanten Koeffizienten

$$y_{k+2} + (4 + 4h)y_{k+1} + (-5 + 2h)y_k = 0 \tag{10}$$

Mit dem Ansatz $y_k = z^k$ für ein $z \in \mathbb{R}\backslash\{0\}$ erhält man aus Gleichung (10) (nach Division durch z^k)

$$z^2 + (4 + 4h)z + (-5 + 2h) = \rho(z) - h\sigma(z) = 0 \tag{11}$$

Die Lösungen dieser quadratischen Gleichungen lauten

$$\begin{aligned}
z_{1/2} = z_{1/2}(h) &= -2(1 + h) \pm \sqrt{4(1 + h)^2 - 2h + 5} \\
&= -2(1 + h) \pm \sqrt{1 + \frac{2h}{3} + \frac{4h^2}{9}}
\end{aligned}$$

Für $h \to 0$ hat man

$$z_1(h) = 1 - h + \mathcal{O}(h^2) \quad \text{und} \quad z_2(h) = -5 + \mathcal{O}(h)$$

Die allgemeine Lösung der Differentialgleichung Gleichung (10) ist gegeben durch

$$y_k = c_1 z_1^k + c_2 z_2^k$$

Gibt man sich die Startwerte $y_0 = y(0) = 1$ und $y_1 = y(h) = \exp(-h)$ als exakte Funktionswerte vor, bestimmen sich die Konstanten $c_1 = c_1(h)$ und $c_2 = c_2(h)$ aus $y_0 = c_1 + c_2$ und $y_1 = c_1 z_1 + c_2 z_2$. Bei genauerer Betrachtung der Abhängigkeit von z_1, z_2 von h ergibt sich dafür $c_1(h) = 1 + \mathcal{O}(h^2)$ und $c_2(h) = -\frac{h^4}{216} + \mathcal{O}(h^5)$. Für festes $x > 0$ und $x = kh$ folgt für $k \to \infty$ und $h \to 0$

$$\left|c_2(h) z_2(h)^k\right| = \left|\mathcal{O}\left(\left(\frac{x}{k}\right)^4\right)\right| \cdot \left|-5 + \mathcal{O}(h)\right|^k \to \infty$$

sowie

$$\begin{aligned} c_1(h) z_1(h) &= (1 + \mathcal{O}(h^2)) \cdot (1 - h + \mathcal{O}(h^2))^k \\ &= \left(1 - \frac{x}{k}\right)^k + \mathcal{O}(h^2) \\ &\to \exp(-x) \end{aligned}$$

Da die sogenannte parasitäre Lösungskomponente $c_2 z_2^k$ die andere sinnvolle Lösungskomponente der Differentialgleichung beliebig übersteigt, kann man für das MSV Gleichung (9) keine Konvergenz erwarten. Um dieses Verhalten bei einem linearen MSV zu verhindern, darf zumindest der Betrag jeder Lösung (Wurzel) der Polynomgleichung $\rho(z) = 0$ den Wert 1 nicht übersteigen, vergleiche Gleichung (11) für $h \to 0$.

Definition 3.3 (D-stabil, nullstabil)
Das lineare MSV Gleichung (1) heißt D-stabil (oder nullstabil), falls es die Wurzelbedingung erfüllt, das heißt wenn der Betrag jeder Nullstelle seines ersten charakteristischen Polynoms ρ durch 1 beschränkt ist und der Betrag jeder mehrfachen Nullstelle von ρ kleiner als 1 ist.

Die Bezeichnung D-stabil ist zu Ehren von Dahlquist (1925-2005) für seine Arbeiten zur Stabilität von linearen MSV gewählt worden.

Zur formalen Definition der Konvergenzordnung eines linearen MSV nehmen wir (wie bei ESV) an, dass die AWA Gleichung (1) die eindeutige Lösung $y\colon [a, b] \to \mathbb{R}^m$ besitzt. Weiter nehmen wir an, dass zu jeder Schrittweite h Startvektoren $y_0^h, ..., y_h^{l-1}$ gegeben sind, aus denen das MSV die Näherungen $y_h^l, ..., y_h^N$ erzeugt. Falls für jede Schrittweite $h = \frac{b-a}{N}$ die Startvektoren die Bedingung

$$\|y_h^\nu - y(x_o + \nu h)\| \leq C_1 h^p \quad \forall \nu = 0, ..., l-1 \tag{12}$$

genügen (mit einem von h unabhängigen $C_1 > 0$), dann heißt ein linearen Mehrschrittverfahren konvergent mit der Ordnung $p \geq 1$, wenn es $C_2 > 0$ und $\tilde{h} > 0$ gibt, so dass

$$\left\|y_h^k - y(x_0 + kh)\right\| \leq C_2 h^p$$

für alle $k = l, ..., N$ und alle Schrittweiten $h \in (0, \tilde{h}]$.

Die beiden folgenden Sätze geben wir ohne Beweis an.

Satz 3.4
Die AWA Gleichung (1) besitze die eindeutige Lösung y: $[a,b] \to \mathbb{R}^m$. Das lineare MSV Gleichung (1) sei D-stabil und habe die Konsistenzordnung p. Es gebe $L_f > 0$, so dass die Lipschitz-Bedingung

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

für alle $(x,y),(x,\bar{y}) \in [a,b] \times \mathbb{R}^m$ gilt. Weiter gelte die Bedingung Gleichung (12) an die Startvektoren. Dann ist das MSV konvergent mit der Ordnung p.

Satz 3.5 (Erste Dahlquist-Barriere)
Ein l-schrittiges lineares MSV Gleichung (1) sei D-stabil. Dann gilt für seine Konsistenzordnung

$$p \leq \begin{cases} l+1 & \text{falls } l \text{ ungerade} \\ l+2 & \text{falls } l \text{ gerade} \\ l & \text{falls } \frac{\beta_l}{\alpha_l} \leq 0 \end{cases}$$

Für den Einfluss von Rundungsfehlern lassen sich für lineare MSV zu Abschnitt 2.4 vergleichbare Überlegungen durchführen.

Zum Beispiel hat man bei den Adams-Bashford-Verfahren für die Schrittzahl $l = 2$ ein explizites lineares MSV, nämlich

$$y^{k+2} - y^{k+1} = h(\beta_0 f(x_k, y^k) + \beta_1 f(x_{k+1}, y^{k+1}))$$

Bezogen aus die allgemeine Form Gleichung (1) linearer MSV gilt hier $\alpha_2 = 1$, $\alpha_1 = -1$, $\alpha_0 = 0$ und $\beta_2 = 0$. Also ist für dieses spezielle MSV $\frac{\beta_2}{\alpha_2} = 0$ und nach Satz 3.5 daher bei gewünschter Konvergenz (und damit D-Stabilität) maximal die Konsistenzordnung $p = 2$ erreichbar. Um diese zu sichern, müssen $c_0 = 0$, $c_1 = 0$ und $c_2 = 0$ nach Satz 3.1 gelten. Wegen $c_0 = \rho(1) = \alpha_0 + \alpha_1 + \alpha_2 = 0 - 1 + 1 = 0$ sind noch $c_1 = \rho'(1) - \sigma(1) = 2\alpha_2 + \alpha_1 - (\beta_0 + \beta_1 + \beta_2) = 1 - \beta_0 - \beta_1 = 0$ und $c_2 = \frac{\alpha_1}{2} - \beta_1 + \frac{4\alpha_2}{2} - 2\beta_2 = -1/2 - \beta_1 + 2 = 0$ zu erfüllen. Dies liefert $\beta_1 = 3/2$ und $\beta_0 = -1/2$. Also hat das Verfahren

$$y^{k+2} - y^{k+1} = \frac{h}{2}\Big(- f(x_k, y^k) + 3f(x_{k+1}, y^{k+1})\Big)$$

nach Satz 3.1 die Konvergenzordnung 2. Das charakteristische Polynom ρ des linearen MSV ist offenbar gegeben durch $\rho(z) = z^2 - z$. Seine Nullstellen sind $z_1 = 0$ und $z_2 = 1$. Folglich ist das Verfahren auch D-stabil und damit nach Satz 3.4 konvergent mit der Ordnung 2.

4. A-Stabilität

Wir betrachten die Test-AWA

$$y' = \lambda y \quad \text{mit} \quad y(0) = 1 \tag{1}$$

wobei $\lambda \in \mathbb{C}$ ein Parameter ist. Die eindeutige Lösung dieser Aufgabe ist gegeben durch $y(x) = \exp(\lambda x)$ und es gilt insbesondere

$$\begin{aligned} \mathfrak{Re}(\lambda) < 0 \quad &\Rightarrow \quad |y(x)| \to 0 \quad \text{für } x \to \infty \\ \mathfrak{Re}(\lambda) = 0 \quad &\Rightarrow \quad |y(x)| = 1 \quad \text{für alle } x \in [0, \infty) \end{aligned}$$

Definition 4.1 (A-Stabilität)
Ein Verfahren erzeuge zu einem beliebigen Paar $(h, \lambda) \in (0, \infty) \times \mathbb{C}$ eine Folge $\{y_k\}$. Dann heißt das Verfahren A-stabil, wenn

$$|y_{k+1}| \leq |y_k| \quad \forall k \in \mathbb{N}$$

für jedes $(h, \lambda) \in (0, \infty) \times \mathbb{C}$ mit $\mathfrak{Re}(\lambda) \leq 0$.

Bei ESV gilt $y_{k+1} = y_k + h\Phi(x_k, y_k, y_{k+1}, h)$. Wir nehmen an, dass für $f(x, y) = \lambda y$ eine Darstellung des ESV in der Form

$$y_{k+1} = g(h\lambda) y_k$$

mit einer Funktion g: $\mathbb{C} \to \mathbb{C}$ existiert. Die Funktion g heißt dann auch Stabilitätsfunktion. Falls der Stabilitätsbereich (Bereich der absoluten Stabilität)

$$\mathcal{S} = \{z \in \mathbb{C} \mid |g(z)| \leq 1\}$$

die Halbebene $\mathbb{C}_- = \{z \in \mathbb{C} \mid \mathfrak{Re}(z) \leq 0\}$ enthält, dann ist das ESV A-stabil (und umgekehrt), denn es gilt $|y_{k+1}| = |g(h\lambda)||y_k| \leq |y_k|$ für $k \in \mathbb{N}$ und beliebige $(h, \lambda) \in (0, \infty) \times \mathbb{C}$ mit $h\lambda \in \mathbb{C}_-$. Für die Trapenzregel (ein implizites ESV)

$$y_{k+1} = y_k + \frac{h}{2}\Big(f(x_k, y_k) + f(x_{k+1}, y_{k+1})\Big)$$

erhält aus der Test-AWA Gleichung (1) $y_{k+1} = y_k + \frac{h}{2}(\lambda y_k + \lambda y_{k+1})$ und somit

$$\left(1 - \frac{h}{2}\lambda\right) y_{k+1} = y_k \left(1 + \frac{h}{2}\lambda\right)$$

das heißt die Stabilitätsfunktion der Trapezregel ist gegeben durch

$$g(z) = \frac{1 + \frac{z}{2}}{1 - \frac{z}{2}} = \frac{2 + z}{2 - z}$$

Falls $\mathfrak{Re}(z) \leq 0$, so folgt

$$|g(z)|^2 = \frac{(2+\mathfrak{Re}(z))^2 + (\mathfrak{Im}(z))^2}{(2-\mathfrak{Re}(z))^2 + (\mathfrak{Im}(z))^2} \leq 1$$

also die A-Stabilität der Trapezregel.

Für das explizite EULER-Verfahren ergibt sich (wegen Gleichung (1))

$$y_{k+1} = y_k + h\lambda y_k = (1+h\lambda)y_k \quad \text{und} \quad g(z) = 1+z$$

Damit gilt

$$|g(z)|^2 = |1+z|^2 = (1+\mathfrak{Re}(z))^2 + (\mathfrak{Im}(z))^2 \leq 1$$

genau dann, wenn $z = h\lambda$ im Einheitskreis um $(-1,0) \in \mathbb{C}$ liegt. Da der Stabilitätsbereich beim expliziten EULER-Verfahren nicht alle $z \in \mathbb{C}$ mit $\mathfrak{Re}(z) \leq 0$ enthält, ist dieses Verfahren nicht A-stabil. Das explizite EULER-Verfahren hat Konsistenzordnung 1 (vgl. Satz 2.3) und ist mit dieser Ordnung auch konvergent (vgl. Satz 2.6). Die fehlende A-Stabilität hat zur Folge, dass zur erfolgreichen numerischen Lösung der Test-AWA Gleichung (1) für $\lambda < 0$ zumindest $-2 \leq h\lambda$ gelten muss. Dies erfordert $h \sim \frac{1}{|\lambda|}$, also gegebenenfalls sehr kleine Schrittweiten. Dies ist neben einem hohen Aufwand auch die Gefahr des Überwiegens von Rundungsfehlern verbunden, vgl. Abschnitt 2.4. Verfahren, die A-stabil sind, bzw. einen hinreichend großen Bereich absoluter Stabilität besitzen, haben außerdem Vorteile bei sogenannten steifen AWA, vgl. Abschnitt 5.

Bei RKV kann man den Stabilitätsbereich untersuchen, indem man sich die Stabilitätsfunktion beschafft. Zum Beispiel betrachten wir das 2-stufige explizite RKV

$$\begin{aligned} y_{k+1} &= y_k + hc_1k_1 + hc_2k_2 \\ &= y_k + hc_1 f(x_k, y_k) + hc_2 f(x_k + \alpha_2 h, y_k + h\beta_{21} f(x_k, y_k)) \end{aligned}$$

Mit der Test-AWA Gleichung (1) folgt

$$\begin{aligned} y_{k+1} &= y_k + h\lambda c_1 y_k + h\lambda c_2 (y_k + h\lambda\beta_{21} y_k) \\ &= y_k (1 + h\lambda c_1 + h\lambda c_2 + (h\lambda)^2 c_2 \beta_{21}) \end{aligned}$$

Beim Verfahren von HEUN (vgl. Abschnitt 2.5) mit $c_1 = c_2 = 1/2$ und $\beta_{21} = 1$ ergibt sich

$$y_{k+1} = y_k \left(1 + h\lambda + \frac{1}{2}(h\lambda)^2\right) \quad \text{und also} \quad g(z) = 1 + z + \frac{1}{2}z^2$$

Man sieht schnell, dass dieses Verfahren nicht A-stabil ist (man wähle $z = (a, 0)$ mit $a < -2$).

▶ **Bemerkung 4.2**

Es gibt kein explizites lineares MSV und kein explizites RKV, dass A-stabil ist und die A-stabilen impliziten MSV haben höchstens Konsistenzordnung 2 (zweite DAHLQUIST-Barriere).

5. Einblick: Steife Probleme

Für $A \in \mathbb{R}^{m\times m}$ werde die AWA

$$y' = Ay \quad \text{mit} \quad y(a) = y^0 \tag{1}$$

für $x \in [a,b]$ betrachtet. Wir setzen in diesem Abschnitt voraus, dass A eine diagonalisierbare Matrix ist, das heißt es gibt eine reguläre Matrix $S \in \mathbb{C}^{m\times m}$ und eine Diagonalmatrix $D \in \mathbb{C}^{m\times m}$ mit $A = SDS^{-1}$. Dann ist die allgemeine Lösung y: $[a,b] \to \mathbb{R}^m$ von $y' = Ay$ gegeben durch

$$y(x) = \sum_{i=1}^{m} c_i \exp(\lambda_i(x-a))v^i$$

wobei $\lambda_1, ..., \lambda_m \in \mathbb{C}$ die Eigenwerte von A und $v^1, ..., v^m \in \mathbb{C}^m$ ein zugehöriges System linear unabhängiger Eigenvektoren bezeichnet (A diagonalisierbar!). Die Koeffizienten $c_1, ..., c_m$ ergeben sich damit eindeutig aus der Anfangsbedingung $y(a) = c_1v^1 + \cdots + c_mv^m = y^0$.

Falls $\mathfrak{Re}(\lambda_i) < 0$ für $i = 1, ..., m$ wird die Zahl

$$\frac{\max_{1\le i\le m}|\mathfrak{Re}(\lambda_i)|}{\min_{1\le i\le m}|\mathfrak{Re}(\lambda_i)|}$$

als Steifigkeitsquotient von A bezeichnet. Ist dieser Quotient groß, dann dient dies als Indikator für ein Phänomen, das bei der Anwendung bestimmter numerischer Verfahren aus Gleichung (1) auftreten kann und als Steifheit (stiffness) der AWA Gleichung (1) bezeichnet wird. Ein solches Phänomen wird im folgenden Beispiel beschrieben und führt bei bestimmten Lösungsverfahren (hier explizites Euler-Verfahren) zum Erfordernis sehr kleiner Schrittweiten.

■ **Beispiel 5.1**
Für $a = 0$ und

$$A = \begin{pmatrix} -80.6 & 119.4 \\ 79.6 & -120.4 \end{pmatrix}$$

ergibt sich als allgemeine Lösung von $y' = Ay$

$$y(x) = c_1\exp(-x)v^1 + c_2\exp(-200x)v^2 \quad \text{mit} \quad v^1 = \begin{pmatrix} 3 \\ 2 \end{pmatrix} \text{ und } v^2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

Für $y^0 = (2,3)^T$ hat man als exakte Lösung von Gleichung (1) $y(x) = c_1\exp(-x)v^1 + c_2\exp(-200x)v^2$. Das explizite Euler-Verfahren liefert

$$y^{k+1} = y^k + hAy^k = (\mathbb{1} + hA)y^k$$

Da A diagonalisierbar ist, gilt $A = SDS^{-1}$ mit $S = (v^1, v^2)$ und $D = \text{diag}(-1, -200)$ und

$$S^{-1} = \frac{1}{5}\begin{pmatrix} 1 & 1 \\ -2 & 3 \end{pmatrix}$$

Damit folgt

$$\begin{aligned} S^{-1}y^{k+1} &= S^{-1}y^k + hS^{-1}ASS^{-1}y^k \\ &= S^{-1}y^k + hDS^{-1}y^k \\ &= (\mathbb{1} + hD)S^{-1}y^k \end{aligned}$$

Setzt man $z^k = S^{-1}y^k$ ergibt sich weiter

$$z^{k+1} = (\mathbb{1} + hD)u^k$$

für $k = 0, ...$ Wegen $z^0 = S^{-1}y^0 = S^{-1}(v^1 + v^2) = (1, 0)^T + (0, 1)^T = (1, 1)^T$ erhält man

$$z^k = (\mathbb{1} + hD)^k \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Für $k \to \infty$ folgt $x_k \to \infty$ und $y(x_k) \to 0$. Um die Konvergenz der Folge $\{z^k\}$ und damit der Folge $\{y^k\}$ gegen 0 zu sichern, müssen

$$|1 + \lambda_1 h| = |1 - h| < 1 \quad \text{und} \quad |1 + \lambda_2 h| = |1 - 200h| < 1$$

erfüllt sein. Dies impliziert $h < 1/100$. Der für die exakte Lösung eigentlich unwesentliche (das heißt sehr schnell abklingende) Anteil $\exp(-200x)v^2$ verursacht beim expliziten Euler-Verfahren sehr kleine Schrittweiten.

Ähnliche Phänomene können bei der Anwendung anderer Verfahren, die nicht A-stabil sind, bzw. deren Bereich der absoluten Stabilität ungeeignet ist, auftreten. Auch bei allgemeineren AWA als Gleichung (1) treten Phänomene der Steifheit auf und erfordern angepasste Verfahren.

6. Ausblick

Die Theorie zur numerischen Behandlung von AWA ist natürlich wesentlich umfangreicher als hier dargestellt werden konnte. Das bedeutet einerseits, dass von den behandelten Themen nur wichtige Grundlagen des vorhandenen Wissens präsentiert wurden. Beispielsweise gibt es eine Reihe von speziellen Verfahren, die nicht oder nur beispielhaft beschrieben und analysiert wurden. Andererseits konnten verschiedene weitere Themen in der Vorlesung gar nicht angesprochen werden. Dazu zählen insbesondere Fragen der Fehlerabschätzung und Schrittweitensteuerung sowie angepasste Stabilitätsbegriffe. Zur Vertiefung stehen neben der Literatur auch weitere Lehrveranstaltungen zur Verfügung.

Anhang

Literaturverzeichnis

[1] Bierbaum, F., Preuss, W., and Wenisch, G. Lehr- und übungsbuch numerische Mathematik. Fachbuchverlag Leipzig im Carl Hanser Verlag, 2001.

[2] Bollhöfer, M., and Mehrmann, V. Numerische Mathematik, 1 ed. Vieweg Teubner Verlag, 2004.

[3] Bärwolff, G. Numerik für Ingenieure, Physiker und Informatiker. Springer Spektrum, 2016.

[4] Freund, R. W., and Hoppe, R. H. Stoer, Bulirsch: Numerische Mathematik 1. Springer, 2007.

[5] Hanke-Bourgeois, M. Grundlagen der numerischen Mathematik und des wissenschaftlichen Rechnens. Vieweg Teubner, 2009.

[6] Hestenes, M. R., and Stiefel, E. Methods of conjugate gradients for solving linear systems. NBS, 1952.

[7] Kanzow, C. Numerik linearer Gleichungssysteme: direkte und iterative Verfahren. Springer, 2005.

[8] Knorrenschild, M. Numerische Mathematik. Fachbuchverlag Leipzig im Carl Hanser Verlag, 2017.

[9] Lanczos, C. Solution of systems of linear equations by minimized iterations. Journal of Research of the National Bureau of Standards 49, 1 (1952), 33–53.

[10] Lin, C.-J., and Moré, J. J. Incomplete cholesky factorizations with limited memory. SIAM Journal on Scientific Computing 21, 1 (1999), 24–45.

[11] Meister, A. Numerik linearer Gleichungssysteme: eine Einführung in moderne Verfahren; mit MATLAB-Implementierungen von C. Vömel. Springer Spektrum, 2015.

[12] Paige, C. C., and Saunders, M. A. Solution of sparse indefinite systems of linear equations. SIAM Journal on Numerical Analysis 12, 4 (1975), 617–629.

[13] Quarteroni, A., Sacco, R., and Saleri, F. Numerische Mathematik 1. Springer, 2002.

[14] Roos, H.-G., and Schwetlick, H. Numerische Mathematik. Teubner, 1999.

[15] Saad, Y., and Schultz, M. Gmres: A generalized minimal residual algorithm for solving non-symmetric linear systems. SIAM Journal on Scientific and Statistical Computing 7, 3 (1986), 856–869.

[16] Schaback, R., and Wendland, H. Numerische Mathematik. Springer-Verlag Berlin Heidelberg, 2005.

[17] Schwarz, H. R. Numerische Mathematik. Teubner, 1997.

[18] Steinbach, O. Losungsverfahren fur lineare Gleichungssysteme: Algorithmen und Anwendungen. Friedrick Vieweg und Son, 2005.

[19] Stoer, J., and Bulirsch, R. Numerische Mathematik 2. Springer, 2005.

[20] Strehmel, K., Weiner, R., and Podhaisky, H. Numerik gewöhnlicher Differentialgleichungen. Vieweg und Teubner, 2012.

Index

A-konjugiert, 19
A-orthogonal, 19
Butcher-Schema, 42
Euler-Verfahren, 35
    explizite Euler-Verfahren, 50
Runge-Kutta-Verfahren, 41
Tschebyschow-Polynom, 24

A-stabil, 49
Anfangswertaufgabe, 33
    Lösung, 33

Backward Differentiation Formulas, 44
BDF-Verfahren, 44

Einschrittverfahren, 35
    explizit, 35
    implizit, 35
    konvergent, 37
    Konvergenzordnung, 37
Einzelschrittverfahren, 12

Fixpunkt, 2
Fixpunktabbildung, 2
Fixpunktaufgabe, 2
Fixpunktiteration, 4

Gauß-Seidel-Verfahren, 12
Gesamtschrittverfahren, 11
gewöhnliches Interationsverfahren, 4
Gitter, 35
    äquidistant, 35
    gleichabständig, 35
Gitterpunkten, 35

Jacobi-Verfahren, 11

konsistent, 36
Konsistenzordnung, 36
Kontraktionskonstante, 3
kontraktiv, 3
Krylov-Raum, 15

lineares Mehrschrittverfahren, 43
    charakteristisches Polynom, 43
    expliziten, 43
    impliziten, 43
lokaler Diskretisierungsfehler, 36

Mehrschrittverfahren
    D-stabil, 47
    Konsistenzordnung, 45
    konvergent mit der Ordnung, 47
    nullstabil, 47

Nullstellenaufgabe, 2

Polygonzugverfahren, 35
Prädiktor-Korrektor-Prinzips, 44

Schrittweite, 35
selbstabbildend, 3
Spektralradius, 10
Splitting-Methoden, 14
Stabilitätsbereich, 49
Stabilitätsfunktion, 49
Steifheit, 51
Steifigkeitsquotient, 51

Trapenzregel, 49

Verfahren von Heun, 42
Verfahrensfunktion, 35
Vorkonditionierer, 30

Wurzelbedingung, 47