Seminar: Numerik gewöhnlicher Differentinalgleichungen
Diagonal implizite Runge-Kutta Verfahren

Manuel Hofmann

14.12.2010

1 Einleitung

Ziel dieser Arbeit ist es den Begriff der S-Stabilität einzuführen und im 1. Teil hinreichende und notwendige Bedingungen herzuleiten. Im 2. Teil werden zwei S-stabile Verfahren hergeleitet, jeweils von Ordnung $p = 2$ und $p = 3$. Weiter wird gezeigt, dass gewisse (S-stabile) Verfahren mit vorgegebener Ordung nicht existieren.
Den Begriff der A- bzw. L-Stabilität muss man deshalb erweitern, da bei der Anwendung solcher auf grosse nichtlineare steife Systeme folgende Probleme auftreten können:

- einige A-stabile Verfahren liefern sehr unstabile Lösungen
- die Genauigkeit scheint unabhängig von der Ordnung zu sein.

Diagonal implizite Runge-Kutta Verfahren haben z.B. gegenüber vollimpliziten Verfahren den Vorteil, dass man wesentlich weniger Gleichungen pro Zeitschritt lösen muss. Verwendet man das Newton-Iterationsverfahren, so löst man in jedem Schritt ein lineares Gleichungssystem von der Form $I - ha_{ii}\frac{\partial f}{\partial x}$. Ist die Matrix A eine untere Dreiecksmatrix und sind die Diagonalelemte alle gleich, so kann man z.B. die LU-Zerlegung speichern und das Gleichungssystem damit sukzessive lösen.

2 Definitionen

Definition 1 (Steife Anfangswertaufgabe)**.** *Eine Anfangswertaufgabe* $u : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$

$$u'(t) = f(t, u(t)) \quad ; \quad u(t_0) = u_0 \tag{1}$$

nennt man steif, wenn für die Eigenwerte λ_i *der Jacobi-Matrix* $f_x(t, u(t))$ *mit* $Re(\lambda_i) < 0$ *gilt:*

$$\frac{\max\limits_i |Re(\lambda_i)|}{\min\limits_j |Re(\lambda_j)|} >> 1 \tag{2}$$

Definition 2 (Allgemeines Runge-Kutta Verfahren)**.** *Das allgemeine Runge-Kutta Verfahren der Stufe* s *ist definiert als:*

$$y_{n+1} = y_n + h\sum_{i=1}^{s} b_i k_i \tag{3}$$

mit

$$k_i = f(t_n + hc_i, y_n + h\sum_{j=1}^{s} a_{ij}k_i) \quad i = 1, \ldots, s. \tag{4}$$

Andere Schreibweise als Butcher-Tableau: $\begin{array}{c|c} A & c \\ \hline b^T & \end{array}$

$s \in \mathbb{N}$ *ist die Stufenanzahl. Man nennt das Verfahren ein*

- *DIRK Verfahren, wenn* $a_{ii} \neq 0$ *für* $i = 1, \ldots, s$ *und* $a_{ij} = 0$ *für* $j \geq i + 1$
  ("diagonally implicit Runge-Kutta")
- *SDIRK Verfahren, wenn zusätzlich* $a_{ii} = \alpha$ *für* $i = 1, \ldots, s$ *gilt.*
  ("singly diagonally implicit Runge-Kutta")

Definition 3 (A-stabil / L-stabil)**.** *Wendet man ein Runge-Kutta Verfahren auf das Modellproblem* $u' = \lambda u$ *an, so erhält man* $y_{n+1} = R(h\lambda)y_n$ *mit* $R(h\lambda) = 1 + h\lambda b^T(I - h\lambda A)^{-1}e$ *wobei* $e = (1, \ldots, 1)^T \in \mathbb{R}^s$
Ein Runge-Kutta Verfahren ist ***A-stabil****, wenn* $|R(h\lambda)| < 1$ *für alle* $h\lambda \in \mathbb{C}$ *mit* $Re(h\lambda) < 0$*.*
Gilt zusätzlich $\lim_{h\lambda \to \infty} R(h\lambda) = 0$*, so nennt man es* ***L-stabil****.*

Man kann das Modellproblem wie folgt erweitern:

$$u'(t) = g'(t) + \lambda(u(t) - g(t)) \quad mit \quad Re(\lambda) < 0 \tag{5}$$

wobei $g(t)$ und $g'(t)$ beschränkte skalare Funktionen auf $[0, T]$ sind.

Definition 4 (S-stabil)**.** *Man nennt ein Runge-Kutta Verfahren* ***S-stabil****, wenn es für jedes* $\lambda_0 < 0$ *ein* $h_0 > 0$ *gibt, so dass für die numerische Lösung* y_n *gilt:*

$$\left| \frac{y_{n+1} - g(t_{n+1})}{y_n - g(t_n)} \right| < 1 \tag{6}$$

für $y_n \neq g(t_n)$ *und für alle* $0 < h < h_0$ *und alle* $\lambda \in \mathbb{C}$ *mit* $Re(\lambda) \leq \lambda_0$*. Gilt zusätzlich*

$$\frac{y_{n+1} - g(t_{n+1})}{y_n - g(t_n)} \longrightarrow 0 \quad für \quad Re(\lambda) \to -\infty \tag{7}$$

für alle h *mit* $[t_n, t_n + h] \subset [0, T]$*, so nennt man das Verfahren* ***stark S-stabil****.*

Korollar 1. *Es gilt:*

1. *S-stabil* $\Rightarrow$ *A-stabil*
2. *stark S-stabil* $\Rightarrow$ *L-stabil*

Beweis. Setzt man $g \equiv 0$ so folgen die Behauptungen direkt aus den Definitionen. □

Die Umkehrung gilt jedoch nicht. Damit ist S-Stabilität wirklich ein stäkeres Konzept als A- bzw. L-Stabilität.

Wendet man das erweiterte Modellproblem (5) auf das Runge-Kutta Verfahren (3) an, so erhält man:

$$\epsilon_{n+1} = (1 - b^T(A - zI)^{-1}e)\epsilon_n - hG_0 + hb^T(A - zI)^{-1}E(z)(G_1 - zG_2) \tag{8}$$

mit $\epsilon_n = y_n - g(t_n)$, $z = \frac{1}{h\lambda}$ und $hG_0 = g(t_n + h) - g(t_n)$
Seien die c_i der Grösse nach sortiert und sei C' die Menge der verschiedenen c_i mit $|C'| = s' \leq s$
Dann ist die $s \times s'$ Matrix $E(z)$ definiert als:

$$E(z)_{i,j} := \begin{cases} -z & falls\ c_i = c'_j = 0 \\ c_i & falls\ c_i = c'_j \neq 0 \\ 0 & sonst \end{cases} \tag{9}$$

und die Vektoren G_1, G_2 mit s' Komponenten:

$$(G_1)_i := \begin{cases} g'(t_n) & falls\ c'_i = 0 \\ \frac{1}{hc'_i}(g(t_n + hc'_i) - g(t_n)) & sonst \end{cases} \tag{10}$$

$$(G_2)_i := \begin{cases} 0 & falls\ c'_i = 0 \\ g'(t_n + hc'_i) & sonst \end{cases} \tag{11}$$

Man kann (8) zusammenfassen als:

$$\epsilon_{n+1} = \alpha(z)\epsilon_n + h\beta(z, G_0, G_1, G_2) \tag{12}$$

$$\alpha(z) = (1 - b^T(A - zI)^{-1}e) \tag{13}$$

$$\beta(z, G_0, G_1, G_2) = -G_0 + b^T(A - zI)^{-1}E(z)(G_1 - zG_2) \tag{14}$$

wobei $\alpha(z) = R(h\lambda)$ das Stabilitätspolynom und $\beta(z, G_0, G_1, G_2)$ der Abschneidefehler ist.

3 S-Stabilität

Lemma 1. *Sei $\epsilon(z, h, \epsilon_0) = \alpha(z)\epsilon_0 + h\beta(z)$ definiert für $\epsilon_0 \in \mathbb{C}$, $h \in (0, \bar{h}]$ und $z \in R := \{w \in \mathbb{C} | a \leq Re(w) < 0\}$. Dann gibt es ein $0 < h_0 \leq \bar{h}$ mit*

$$|\epsilon(z, h, \epsilon_0)| \leq |\epsilon_0| \tag{15}$$

für alle $\epsilon_0 \neq 0, h \in (0, h_0]$ und alle $z \in R$ genau dann, wenn

1. *$|\alpha(z)| \leq 1$ in R und*
2. *$\frac{\beta(z)}{1-|\alpha(z)|}$ beschränkt in R ist.*

Beweis. " $\Leftarrow$ " (durch Widerspruch)
Sei $|\alpha(z)| \geq 1$ in R dann gibt es ein $\epsilon_0 \neq 0$, so dass $|\epsilon| < |\epsilon_0|$, wenn $\beta(z) \neq 0 \quad \forall h > 0$. Wenn $\beta(z) = 0$ dann ist $|\epsilon| \geq |\epsilon_0|$ für alle $\epsilon_0 \neq 0$.
Sei $\frac{\beta(z)}{1-|\alpha(z)|}$ nicht beschränkt in R, dann gibt es ein $z \in R$ und $\epsilon_0 = 1 - |\alpha(z)| \neq 0$ mit

$$\frac{\beta(z)}{\epsilon_0} = \frac{\beta(z)}{1 - |\alpha(z)|} > K \tag{16}$$

für $K > 0$ beliebig gross. Und damit

$$|\epsilon| = |\epsilon_0||\alpha(z) + h\frac{\beta(z)}{\epsilon_0}| > |\epsilon_0| \tag{17}$$

für $h > 0$.
" $\Rightarrow$ "
Sei $\frac{\beta(z)}{1-|\alpha(z)|} < K$ für alle $z \in R$ mit festem $K > 0$. Dann gilt

$$|\epsilon| \leq |\alpha(z)||\epsilon| + h|\beta(z)| = |\epsilon_0| - \underbrace{(1 - |\alpha(z)|)}_{>0} \underbrace{(\epsilon_0 - h\frac{|\beta(z)}{1 - |\alpha(z)|})}_{>0} < |\epsilon_0| \tag{18}$$

für $|\alpha(z)| < 1$ und $h \in (0, h_0]$ mit $h_0 = min\{\bar{h}, \frac{|\epsilon_0|}{K}\}$. □

Korollar 2. *Ein Runge-Kutta Verfahren ist A-stabil genau dann, wenn* $|\alpha(z)| < 1$ *für alle* z *mit* $Re(z) < 0$.

Beweis. Setze $g \equiv 0$. □

Korollar 3. *Ein Runge-Kutta Verfahren ist S-stabil genau dann, wenn es A-stabil ist und* $\frac{\beta(z,G_0,G_1,G_2)}{1-|\alpha(z)|}$ *beschränkt in R (wie oben) ist für alle beschränkten* g *und* g' *auf* $[x_n, x_n + h]$.

Beweis. Setzt man $\beta(z) = \beta(z, G_0, G_1, G_2)$ so folgt die Behauptung direkt aus der Definition von S-Stabilität, Korollar (2) und Lemma (1). □

Definiere:

$$\alpha_0 := \lim_{|z|\to 0} \alpha(z) = 1 - \lim_{|z|\to 0} b^T (A - zI)^{-1} e \overset{\text{DIRK}}{=} 1 - b^T A^{-1} e \tag{19}$$

$${b_0}^T := \lim_{|z|\to 0} b^T (A - zI)^{-1} E(z) \tag{20}$$

$$\tag{21}$$

für $Re(z) < 0$. Man nennt ein Runge-Kutta Verfahren "stiffly accurate" $\Leftrightarrow \lim_{|z|\to 0} \beta(z, G_0, G_1, G_2) = 0$.

Lemma 2. *Ein DIRK Verfahren mit* $c_s = 1$ *und* $a_{si} = b_i \quad \forall i = 1, \ldots, s$ *ist stiffly accurate und L-stabil. Wobei* a_{ij} *Elemente aus der Matrix* A *und* b *die Gewichte aus (3) bzw. (4) sind.*

Beweis. Zunächst wird gezeigt, dass ${b_0}^T = (0, \ldots, 0, 1)$.
${b_0}^T = \lim_{|z|\to 0} \underbrace{b^T (A - zI)^{-1}}_{\to (0,\ldots,0,1)} E(z) = (0, \ldots, 0, 1)$ wegen

$$E(z) \to \begin{pmatrix} * & \ldots\ldots & * \\ \vdots & \ddots & \vdots \\ * & \ldots\ldots & * \\ 0 & \ldots \quad 0 & 1 \end{pmatrix} \tag{22}$$

Die letzte Zeile von $E(z)$ hat deshalb diese Gestalt, weil der letzte Knoten $c_s = 1$.
$\lim_{|z|\to 0} b^T (A - zI)^{-1} = b^T A^{-1} = (0, \ldots, 0, 1)$ da dies der letzten Zeile von AA^{-1} entspricht.
Und damit ist $\lim_{|z|\to 0} \beta(z, G_0, G_1, G_2) = -G_0 + (0, \ldots, 0, 1) G_1 = 0$
L-stabil:
$\alpha_0 = 1 - b^T A^{-1} e = 1 - (0, \ldots, 0, 1)(1, \ldots, 1) = 0.$

□

Satz 1. *Ein A-stabiles Runge-Kutta Verfahren ist genau dann S-stabil, wenn* $|\alpha_0| < 1$ *und* b_0 *endlich ist.*

Beweis. $\frac{1}{1-|\alpha(z)|}$ ist beschränkt in R genau dann, wenn $|\alpha_0| < 1$. Für beschränkte g und g' ist $\beta(z, G_0, G_1, G_2)$ beschränkt genau dann, wenn b_0 endlich ist. Damit folgt die Behauptung aus Korollar (3). □

Satz 2. *Ein S-stabiles Runge-Kutta Verfahren ist stark S-stabil genau dann, wenn es L-stabil und stiffly accurate ist.*

Beweis. Für starke S-Stabilität benötigt man nur noch $\lim_{|z|\to 0} \epsilon_{n+1} = 0$ zu zeigen.
$\epsilon_{n+1} \to 0 \quad \Leftrightarrow \quad \alpha(z) \to 0 \quad$ und $\quad \beta(z, G_0, G_1, G_2) \to 0$
Ersteres ist genau dann der Fall, wenn das Verfahren L-stabil ist und zweiteres genau dann, wenn es stiffly accurate ist. □

4 Verfahrensherleitung

Mit $p \in \mathbb{N}$ wird ab sofort die Ordnung des Verfahrens bezeichnet.

$$T := diag(c_1, \ldots, c_s) \in \mathbb{R}^{s \times s} \tag{23}$$
$$e := (1, \ldots, 1)^T \in \mathbb{R}^s \tag{24}$$
$$b := (b_1, \ldots, b_s)^T \in \mathbb{R}^s \tag{25}$$

Für ein Runge-Kutta Verfahren mit Ordnung $p \geq x$ gelten die Gleichungen (25.x)

$$(b, e) = 1 \tag{25.1}$$

$$(b, Te) = \frac{1}{2} \qquad (b, Ae) = \frac{1}{2} \tag{25.2}$$

$$(b, T^2e) = \frac{1}{3} \qquad (b, TAe) = \frac{1}{3} \qquad (b, ATe) = \frac{1}{6} \qquad (b, A^2e) = \frac{1}{6}$$

$$(b, T^3e) = \frac{1}{4} \qquad (b, TATe) = \frac{1}{8} \qquad (b, AT^2e) = \frac{1}{12} \qquad (b, A^2Te) = \frac{1}{24} \tag{25.3}$$
$$(b, T^2Ae) = \frac{1}{4} \qquad (b, TA^2e) = \frac{1}{8} \qquad (b, ATAe) = \frac{1}{12} \qquad (b, A^3e) = \frac{1}{24}$$

$$(b, T^3Ae) = \frac{1}{5} \qquad (b, TATAe) = \frac{1}{15} \qquad (b, TA^3e) = \frac{1}{30} \qquad (b, A^2T^2e) = \frac{1}{60}$$
$$(b, T^4e) = \frac{1}{5} \qquad (b, TAT^2e) = \frac{1}{15} \qquad (b, TA^2Te) = \frac{1}{30} \qquad (b, A^2TAe) = \frac{1}{60} \tag{25.4}$$
$$(b, T^2ATe) = \frac{1}{10} \qquad (b, AT^3e) = \frac{1}{20} \qquad (b, ATATe) = \frac{1}{40} \qquad (b, A^3Te) = \frac{1}{120}$$
$$(b, T^2A^2e) = \frac{1}{10} \qquad (b, AT^2Ae) = \frac{1}{20} \qquad (b, ATA^2e) = \frac{1}{40} \qquad (b, A^4e) = \frac{1}{120} \tag{25.5}$$

Lemma 3. *Für ein DIRK Verfahren mit* $(s, p) = (4, 5)$ *und* $\delta := ATe - \frac{1}{2}T^2e \neq 0$ *gilt*

$$A^Tb = (I - T)b \tag{26}$$
$$A^TTb = \frac{1}{2}(I - T^2)b \tag{27}$$

Beweis. b, Tb, T^2b und A^Tb sind orthogonal zu $\delta \neq 0$, denn

$$\begin{array}{rcrcrcl}
(b, ATe - \frac{1}{2}T^2e) & = & (b, ATe) - \frac{1}{2}(b, T^2e) & = & \frac{1}{6} - \frac{1}{2}\frac{1}{3} & = & 0 \\
(Tb, ATe - \frac{1}{2}T^2e) & = & (Tb, ATe) - \frac{1}{2}(Tb, T^2e) & = & \frac{1}{8} - \frac{1}{2}\frac{1}{4} & = & 0 \\
(T^2b, ATe - \frac{1}{2}T^2e) & = & (T^2b, ATe) - \frac{1}{2}(T^2b, T^2e) & = & \frac{1}{10} - \frac{1}{2}\frac{1}{5} & = & 0 \\
(A^Tb, ATe - \frac{1}{2}T^2e) & = & (A^Tb, ATe) - \frac{1}{2}(A^Tb, T^2e) & = & \frac{1}{24} - \frac{1}{2}\frac{1}{12} & = & 0
\end{array}$$

Also gibt es wegen $s = 4$ Konstanten $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \in \mathbb{R}$ nicht alle 0 mit

$$\lambda_1 b + \lambda_2 Tb + \lambda_3 T^2 b + \lambda_4 A^T b = 0 \tag{28}$$

Multipliziert man diese Gleichung jeweils mit e, Te, T^2e erhält man folgendes lineare Gleichungssystem:

$$\lambda_1 + \frac{1}{2}\lambda_2 + \frac{1}{3}\lambda_3 + \frac{1}{2}\lambda_4 = 0$$
$$\frac{1}{2}\lambda_1 + \frac{1}{3}\lambda_2 + \frac{1}{4}\lambda_3 + \frac{1}{6}\lambda_4 = 0$$
$$\frac{1}{3}\lambda_1 + \frac{1}{4}\lambda_2 + \frac{1}{5}\lambda_3 + \frac{1}{12}\lambda_4 = 0$$

falls $\lambda_1 = 0$ folgt $\lambda_2 = \lambda_3 = \lambda_4 = 0$, also setze $\lambda_1 = 1$ und löse das lineaere Gleichungssystem:
$\lambda_1 = 1; \quad \lambda_2 = -1; \quad \lambda_3 = 0; \quad \lambda_4 = 1$ Also:

$$\begin{aligned} b + Tb + A^T b &= 0 \\ \Leftrightarrow (I - T)b &= A^T b \end{aligned}$$

Analog sind $b, TA^T b, A^T Tb$ und $T^2 b$ orthogonal zu $\delta \neq 0$

$$\begin{array}{ccccccc} (b, ATe - \frac{1}{2}T^2 e) & = & (b, ATe) - \frac{1}{2}(b, T^2 e) & = & \frac{1}{6} - \frac{1}{2}\frac{1}{3} & = & 0 \\ (TA^T b, ATe - \frac{1}{2}T^2 e) & = & (TA^T b, ATe) - \frac{1}{2}(TA^T b, T^2 e) & = & \frac{1}{40} - \frac{1}{2}\frac{1}{20} & = & 0 \\ (A^T Tb, ATe - \frac{1}{2}T^2 e) & = & (A^T Tb, ATe) - \frac{1}{2}(A^T T, T^2 e) & = & \frac{1}{30} - \frac{1}{2}\frac{1}{15} & = & 0 \\ (T^2 b, ATe - \frac{1}{2}T^2 e) & = & (T^2 b, ATe) - \frac{1}{2}(T^2 b, T^2 e) & = & \frac{1}{10} - \frac{1}{2}\frac{1}{5} & = & 0 \end{array}$$

Also gibt es wieder Konstanten $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \in \mathbb{R}$ nicht alle 0 mit

$$\lambda_1 b + \lambda_2 TA^T b + \lambda_3 A^T Tb + \lambda_4 A^2 b = 0 \tag{29}$$

Multipliziere die Gleichung mit e, Ae und $A^2 e$ und erhalte das lineare Gleichungssystem:

$$\begin{aligned} \lambda_1 + \frac{1}{6}\lambda_2 + \frac{1}{3}\lambda_3 + \frac{1}{3}\lambda_4 = 0 \\ \frac{1}{2}\lambda_1 + \frac{1}{12}\lambda_2 + \frac{1}{8}\lambda_3 + \frac{1}{4}\lambda_4 = 0 \\ \frac{1}{6}\lambda_1 + \frac{1}{40}\lambda_2 + \frac{1}{30}\lambda_3 + \frac{1}{10}\lambda_4 = 0 \end{aligned}$$

Für $\lambda_1 = 0$ erhält man als Lösung $\lambda_2 = \lambda_3 = \lambda_4 = 0$. Also setze $\lambda_1 = 1$ und erhalte $\lambda_2 = 0; \quad \lambda_3 = -2; \quad \lambda_4 = -1$. Also insgesamt

$$\begin{aligned} b - 2A^T Tb - T^2 b &= 0 \\ \Leftrightarrow 2A^T Tb &= b - T^2 b \\ \Leftrightarrow A^T Tb &= \frac{1}{2}(I - T^2)b. \end{aligned}$$

□

Satz 3. *Es gibt kein DIRK Verfahren mit* $(s, p) = (4, 5)$.

Beweis. (durch Widerspruch, man nimmt an es gäbe solch ein Verfahren)
Unter der Voraussetzung $a_{11} = c_1$ (was keine Einschränkung ist , wegen Konsistenzbedingung) gilt $\delta = ATe - \frac{1}{2}T^2 e \neq 0$ da sonst $a_{11} = 0$ ist und somit kein DIRK Verfahren. Also gelten die Gleichungen (26) und (27) aus Lemma (3).
Betrachtet man die 4. Komponente von Gleichung (26)

$$\begin{aligned} a_{44} b_4 &= (1 - c_4) b_4 \\ \Leftrightarrow a_{44} &= 1 - c_4 \end{aligned}$$

sonst hätte das Verfahren nur 3 Stufen ($s = 3$). Die 4. Komponene der Gleichung (27) ergibt

$$\begin{aligned} a_{44} c_4 b_4 &= \frac{1}{2}(1 - c_4^2) b_4 \\ \Leftrightarrow (1 - c_4) c_4 &= \frac{1}{2}(1 - c_4^2) \\ \Leftrightarrow c_4^2 - 2c_4 + 1 &= 0 \end{aligned}$$

Also $c_4 = 1$ und damit $a_{44} = 0$ also kein DIRK Verfahren. Widerspruch. □

Satz 4. *Es gibt genau 2 stark S-stabile SDIRK Verfahren mit* $(s,p)=(2,2)$. *Diese sind von der Form*

$$\begin{array}{cc|c} \alpha & 0 & \alpha \\ 1-\alpha & \alpha & 1 \\ \hline 1-\alpha & \alpha & \end{array}$$

mit $\alpha = 1 \pm \frac{1}{2}\sqrt{2}$. *Dieses Verfahren wird auch Alexander Verfahren genannt.*

Beweis. Man überzeugt sich, dass dieses Verfahren von Ordnung 2 ist. Weiterhin hat das Stabilitätspolynom seine Singularität bei $\alpha > 0$ also ist $R(h\lambda)$ analytisch in der linken komplexen Halbebene und es gilt $|R(iy)| \leq 1 \quad \forall y \in \mathbb{R}$. Also ist das Verfahren nach dem Maximumprinzip A-stabil. Nach den vorangegangenen Bemerkungen hat ein stark S-stabiles SDIRK Verfahren notwendig folgende Form

$$\begin{array}{cc|c} \alpha & 0 & c \\ 1-\alpha & \alpha & 1 \\ \hline 1-\alpha & \alpha & \end{array}$$

wobei nur noch c zu bestimmen ist. Aus den Gleichungen von (25.2) erhält man

$$\begin{array}{rcccl} (b,Te) & = & \frac{1}{2} & = & (b,Ae) \\ \Leftrightarrow (1-\alpha)\alpha+\alpha & = & \frac{1}{2} & = & (1-\alpha)c+\alpha \end{array} \quad \text{also } c=\alpha \text{ und } \alpha^2-2\alpha+\frac{1}{2}=0. \qquad \square$$

Satz 5. *Es gibt genau ein stark S-stabiles SDIRK Verfahren mit* $(s,p)=(3,3)$. *Es ist von der Form*

$$\begin{array}{ccc|c} \alpha & 0 & 0 & \alpha \\ c_2-\alpha & \alpha & 0 & c_2 \\ b_1 & b_2 & \alpha & 1 \\ \hline b_1 & b_2 & \alpha & \end{array}$$

wobei α *Nullstelle von* $x^3-3x^2+\frac{3}{2}x-\frac{1}{6}=0$ *mit* $\alpha \in (\frac{1}{6},\frac{1}{2})$ *ist.* $c_2=\frac{1+\alpha}{2}, \quad b_1=-\frac{6\alpha^2-16\alpha+1}{4}$ $b_2=\frac{6\alpha^2-20\alpha+5}{4}$

Beweis. Analog wie oben überzeugt man sich, dass das Verfahren die Ordnung $p=3$ hat. Mit dem selben Argument ist es auch A-stabil genau dann, wenn $\alpha \in (\frac{1}{6},\frac{1}{2})$. Weiterhin hat ein stark S-stabiles SDIRK Verfahren die notwendige Form

$$\begin{array}{ccc|c} \alpha & 0 & 0 & c_1 \\ \beta & \alpha & 0 & c_2 \\ b_1 & b_2 & \alpha & 1 \\ \hline b_1 & b_2 & \alpha & \end{array}$$

Für das Stabilitätspolynom $R(h\lambda)$ muss folgendes gelten

$$\begin{aligned} \frac{\tau_2(h\lambda)^2+\tau_1 h\lambda+1}{(1-\alpha h\lambda)^3} &= e^{h\lambda}+\mathcal{O}((h\lambda)^4) \\ \tau_2(h\lambda)^2+\tau_1 h\lambda+1 &= (1-\alpha h\lambda)^3 e^{h\lambda}+\mathcal{O}((h\lambda)^4) \\ \tau_2(h\lambda)^2+\tau_1 h\lambda+1 &= (1-3h\lambda\alpha+3h^2\alpha^2\lambda^2-h^3\alpha^3\lambda^3)(\sum_{n=0}^{\infty}\tfrac{x^n}{n!})+\mathcal{O}((h\lambda)^4) \\ \tau_2(h\lambda)^2+\tau_1 h\lambda+1 &= (h\lambda)^3(-\alpha^3+3\alpha^2-\tfrac{3}{2}\alpha+\tfrac{1}{6})+(h\lambda)^2(3\alpha^2-3\alpha+\tfrac{1}{2})+(h\lambda)(-3\alpha+1)+1+\mathcal{O}((h\lambda)^4) \end{aligned}$$

$\Rightarrow \alpha$ ist Nullstelle von $x^3-3x^2+\frac{3}{2}x-\frac{1}{6}=0$.
Jetzt wird gezeigt, dass $Ae=Te$
Beweis durch Widerspruch: Wenn $Ae-Te\neq 0$ dann sind b, Tb und A^Tb orthogonal zu $Ae-Te$ (wie in Lemma (3)) Also gibt es $\lambda_1,\lambda_2,\lambda_3 \in \mathbb{R}$ nicht alle 0 mit

$$\lambda_1 b+\lambda_2 A^T b+\lambda_3 Tb=0 \tag{30}$$

Bilde jeweils Skalarprodukt mit e und Ae

$$\begin{aligned}\lambda_1 + \frac{1}{2}\lambda_2 + \frac{1}{2}\lambda_3 &= 0\\ \frac{1}{2}\lambda_1 + \frac{1}{6}\lambda_2 + \frac{1}{3}\lambda_3 &= 0\end{aligned}$$

$\lambda_1 = 0 \Rightarrow \lambda_2 = \lambda_3 = 0$. Also setze $\lambda_1 = 1 \Rightarrow \lambda_2 = -1; \quad \lambda_3 = -1$

$$\begin{aligned}A^T b &= (I - T)b\\ \Rightarrow (A^T b)_3 &= ((I - T)b)_3\\ \Leftrightarrow \alpha^2 &= 0\end{aligned}$$

Was einem explizitem Verfahren enstspricht. Also gilt was zu zeigen war $Ae = Te$.
$\Rightarrow$ das Verfahren hat die Form

$$\begin{array}{ccc|c} \alpha & 0 & 0 & \alpha \\ c_2 - \alpha & \alpha & 0 & c_2 \\ b_1 & b_2 & \alpha & 1 \\ \hline b_1 & b_2 & \alpha & \end{array}$$

Jetzt sind noch c_2, b_1, b_2 zu bestimmen. Die Gewichte b lassen sich über die Lagrange-Interpolationspolynome bestimmen. Das liegt daran, da jedes Runga-Kutta Verfahren einer Quadraturformel entspricht.

$$b_3 = \alpha = \int_0^1 \frac{(x-\alpha)(x-c_2)}{(1-\alpha)(1-c_2)}\mathrm{d}x \tag{31}$$

$\Rightarrow c_2 = \frac{1+\alpha}{2}$

$$b_1 = \int_0^1 \frac{(x-c_2)(x-1)}{(\alpha-c_2)(\alpha-1)}\mathrm{d}x = -\frac{6\alpha^2 - 16\alpha + 1}{4} \tag{32}$$

$$b_2 = \int_0^1 \frac{(x-\alpha)(x-1)}{(c_2-\alpha)(c_2-1)}\mathrm{d}x = \frac{6\alpha^2 - 20\alpha + 5}{4} \tag{33}$$

□

Satz 6. *Es gibt kein stark S-stabiles SDIRK Verfahren mit* $(s, p) = (4, 4)$.

Beweis. (durch Widerspruch)
Das Verfahren hat notwendig die Form

$$\begin{array}{cccc|c} \alpha & 0 & 0 & 0 & c_1 \\ \beta & \alpha & 0 & 0 & c_2 \\ \gamma & \delta & \alpha & 0 & c_3 \\ b_1 & b_2 & b_3 & \alpha & 1 \\ \hline b_1 & b_2 & b_3 & \alpha & \end{array}$$

Das Stabilitätspolynom lautet damit

$$R(h) = \frac{1 + \tau_1 h + \tau_2 h^2 + \tau_3 h^3}{(1 - \alpha h)^4} = e^h + \mathcal{O}(h^5) \tag{34}$$

Analog wie im Beweis von Satz (5) ist α Nullstelle von $f(x) = x^4 - 4x^3 + 3x^2 - \frac{2}{3}x + \frac{1}{24}$. Es gilt $|R(iy)| \leq 1 \quad \forall y \in \mathbb{R} \Leftrightarrow \alpha \in (\frac{1}{2}, \frac{3}{5})$. Da A eine untere Dreiecksmatrix ist, gilt $(A - \alpha I)^4 = 0$ und damit

$$\mu := b^T A^4 e = b^T(4\alpha A^3 - 6\alpha^2 A^2 + 4\alpha^3 A - \alpha^4)e = \frac{1}{24} - \frac{1}{2}\alpha + 2\alpha^2 - 2\alpha^3 \tag{35}$$

weiter ist $\mu \neq 0$ da $f(x)$ irreduzibel über $\mathbb{Q}[X]$ ist. Wie in Satz (5) wird jetzt $Ae = Te$ durch Widerspruch bewiesen. Wenn $Ae - Te \neq 0$ dann sind $b, A^T b, Tb$ und $(A^T)^2 b$ orthogonal zu $Ae - Te$. Also gibt es $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \in \mathbb{R}$ nicht alle 0 mit

$$\lambda_1 b + \lambda_2 A^T b + \lambda_3 Tb + \lambda_4 (A^T)^2 b = 0 \tag{36}$$

Bilde jeweils Skalarprodukt mit e, Ae und $A^2 e$ und erhalte

$$\begin{aligned} \lambda_1 + \frac{1}{2}\lambda_2 + \frac{1}{2}\lambda_3 + \frac{1}{6}\lambda_4 &= 0 \\ \frac{1}{2}\lambda_1 + \frac{1}{6}\lambda_2 + \frac{1}{3}\lambda_3 + \frac{1}{24}\lambda_4 &= 0 \\ \frac{1}{6}\lambda_1 + \frac{1}{24}\lambda_2 + \frac{1}{8}\lambda_3 + \mu\lambda_4 &= 0 \end{aligned}$$

$\lambda_1 = 0 \Rightarrow \lambda_2 = \lambda_3 = \lambda_4 = 0$ also setze $\lambda_1 = 1$ und löse das linere Gleichungssystem
$\Rightarrow \lambda_2 = -1; \quad \lambda_3 = -1; \quad \lambda_4 = 0$
$\Rightarrow A^T b = (I - T)b$ und damit ist die 4. Komponente $\alpha^2 = 0$ Widerspruch. Also gilt $Ae = Te$ und das Verfahren ist von der Form

$$\begin{array}{ccccc|c} \alpha & 0 & 0 & 0 & c_1 \\ c_2 - \alpha & \alpha & 0 & 0 & c_2 \\ c_2 - \beta - \alpha & \beta & \alpha & 0 & c_3 \\ b_1 & b_2 & b_3 & \alpha & 1 \\ \hline b_1 & b_2 & b_3 & \alpha & \end{array} \tag{37}$$

1.Fall: die Knoten sind paarweise verschieden
Die Gewichte b sind eindeutig bestimmt sobald die Knoten fest gewählt wurden. Aus dem Gewicht $b_4 = \alpha$ kann man c_2 in Abhängigkeit von c_3 bestimmen (wie in (31)):

$$\alpha = \frac{\frac{1}{4} - \frac{1}{3}(\alpha + c_2 + c_3) + \frac{1}{2}(\alpha c_2 + \alpha c_3 + c_2 c_3) - \alpha c_2 c_3}{(1 - \alpha)(1 - c_2)(1 - c_3)} \tag{38}$$

$\Rightarrow$ das Verfahren ist durch die Wahl von β und entweder c_2 oder c_3 eindeutig bestimmt. Setzt man das Tableau (37) jeweils in

$$(b, ATe) = \frac{1}{6}, \quad (b, TATe) = \frac{1}{8}, \quad (b, AT^2 e) = \frac{1}{12}$$

ein so erhält man die 3 Gleichungen

$$b_3(c_2 - \alpha)\beta = \frac{1}{6} - \frac{3}{2}\alpha + 3\alpha^2 - \alpha^3 \tag{39}$$

$$b_3 c_3 (c_2 - \alpha)\beta = \frac{1}{8} - \frac{7}{6}\alpha + \frac{5}{2}\alpha^2 - \alpha^3 \tag{40}$$

$$b_3(c_2^2 - \alpha^2)\beta = \frac{1}{8} - \frac{4}{3}\alpha + \frac{7}{2}\alpha^2 - 2\alpha^3 \tag{41}$$

$\Rightarrow b_3 \neq 0$, da sonst α einem Polynom vom Grad 3 genügt. (39) und (40) bestimmen c_3 eindeutig in Abhängigkeit von α, jedoch liefert dann (41) und (38) widersprüchliche Werte für c_2. Also kann es kein stark S-stabiles DIRK Verfahren mit paarweise verschiedenen Knoten geben.
2.Fall: die Knoten sind nicht paarweise verschieden
Dann muss es notwendig 3 verschiedene Knoten geben. Also gibt es 3 Fälle:

i) c_2 oder $c_3 = \alpha$

ii) c_2 oder $c_3 = 1$

iii) $c_2 = c_3$ beide ungleich $1, \alpha$

Zuerst werden i) und iii) zum Widerspruch geführt.
Sei $\phi(x) := \frac{1+x}{2}$ die stetige Transformation von $[-1, 1]$ auf das Intervall $[0, 1]$. Dann sind in allen 3 Fällen die Knoten von der Form $\phi(r_i), \quad i = 1, 2, 3$ wobei r_i Nullstellen eines Polynoms vom Grad 3 und orthogonal zur konstanten Funktionen auf $[-1, 1]$ ist.
$\Rightarrow$ Der dritte Knoten ist damit $\frac{1-2\alpha}{2(1-3\alpha)}$ da 2 Knoten gleich α und 1 sind.

$$\Rightarrow b_4 = \frac{1 - 6\alpha + 6\alpha^2}{6(1-\alpha)(1-4\alpha)} = \alpha$$

Was jedoch ein Widespruch ist, da α nun einem Polynom vom Grad 3 genügt. Also sind i) und iii) unmöglich.
Sei also $c_3 = 1$ wie in Fall ii). Dann sind (39) und (40) gleich und man erhält $\alpha = \frac{1}{2}$ oder $\alpha = \frac{1}{6}$ Widerspruch zur A-Stabilität.
Gilt $c_2 = 1$ dann sind (39) und (41) gleich und man erhält wieder $\alpha = \frac{1}{2}$ oder $\alpha = \frac{1}{6}$. Also gibt es kein stark S-stabiles DIRK Verfahren mit $(s, p) = (4, 4)$. □

Literatur

[Ale77] ALEXANDER, Roger: Diagonally Implicit Runge-Kutta Methods for Stiff O.D.E.'s. In: *SIAM Journal on Numerical Analysis* 14 (1977), December, S. 1006–1021

[Rob74] ROBINSON, A. Prothero A.: On the Stability of One-Step Methods for Solving Stiff Systems of Ordinary Differential Equations. In: *Mathematics of Computation* 28 (1974), January, Nr. 125, S. 145–162