Spezialverfahren für Anfangswertprobleme

Bernhard Schmitt

Sommersemester 2007

Inhaltsverzeichnis

1 Einleitung

Bei Anfangswertproblemen hat die unabhängige Variable in der Regel die Bedeutung der Zeit und wird daher mit t bezeichent. Das Anfangswertproblem wird in folgender Weise formuliert,

$$u'(t) = f(t, u(t)),\ t \in [t_a, t_e), \quad u(t_a) = u_0, \tag{1.0.1}$$

mit einer Funktion $f(t, y)$, $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$. Die Verwendung eines halboffenen Intervalls erlaubt den Fall $t_e = \infty$. Themenübersicht:

1. Grenzen der Standardverfahren
2. Steife Anfangswertprobleme
3. Oberklasse der Standardverfahren: Allgemeine Lineare Methoden (ALM/GLM)
4. Erhaltungssätze bei Differentialgleichungen: *Geometrische* Verfahren
5. Parallelisierung: Iterative Verfahren, Peer-Methoden

In der Numerik von Differentialgleichungen wurden die wichtigsten Standardverfahren behandelt, mit denen viele Anfangswertprobleme tatsächlich effizient und verlässlich gelöst werden können. Aus der Praxis oder aufgrund der Weiterentwicklung von Computern kommen aber immer wieder neue Anforderungen an die Numerik. So bekommen die Standardverfahren Schwierigkeiten, wenn die bei der Analyse verwendeten Größen zu große Werte annehmen, etwa die Lipschitzkonstante L und die Intervalllänge $t_e - t_a$. Viele Differentialgleichungen (aus Chemie und Physik) weisen z.B. extrem große Lipschitzkonstanten L auf trotz glatter Lösungen u. Das einfachste Testbeispiel ist die DGl $y'(t) = \lambda(y(t) - g(t)) + g'(t)$ mit $\lambda < 0$ und $L = |\lambda| \gg 1$, wo sich jede Lösung y sehr schnell der Funktion $g(t)$ nähert ($\to$ steife AWPe). Bei anderen Anwendungen (Molekularphysik, Himmelsmechanik) ist man an vergleichsweise sehr langen Zeitintervallen interessiert und will dabei Erhaltungssätze (Energie) möglichst exakt einhalten, um trotz des unvermeidlichen Fehlerwachstums physikalisch sinnvolle Ergebnisse zu erhalten ($\to$ geometrische Verfahren). Hier spielen auch neue Erkenntnisse zur *Dynamik* von numerischen Verfahren eine Rolle: Bei längeren Zeitintervallen ist es praktisch unmöglich, ein bestimmtes AWP genau zu lösen. Man weiß aber, dass man ein benachbartes Problem gut löst, wenn das Verfahren die Dynamik der Dgl einigermaßen korrekt modelliert.

Ganz andere Anforderungen kommen aus der aktuellen Rechnerentwicklung. Wegen der physikalischen Grenzen (Lichtgeschwindigkeit) können Höchstleistungen nur noch durch Parallelverarbeitung erreicht werden, dies gilt mittlerweile bis hinab auf PC-Ebene (Multi-Core-Prozessoren). Unglücklicherweise arbeiten aber alle Standardverfahren sequentiell, keines erlaubt eine Parallelisierung in der Methode. Denn jede(r) einzelne Stufe/Schritt bei den Runge-Kutta und Mehrschrittverfahren kann erst ausgeführt werden, wenn die davorliegenden abgeschlossen

sind. Einen Ausweg bietet teilweise der Übergang zu einer übergeordneten Verfahrensklasse, der der *Allgemeinen Linearen Methoden* (ALM, General Linear Methods), welche sowohl Ein- als auch Mehrschrittverfahren umfaßt. Vor Einführung dieser Methoden ist es sinnvoll sich mit der ersten Problemklasse zu beschäftigen.

2 Steife Anfangswertprobleme

2.1 Problematik

In der Praxis trifft man oft auf Differentialgleichungen, deren Lösungen sehr unterschiedliche *Zeitskalen* besitzen, wobei sowohl sehr schnell, als auch langsam veränderliche Lösungen auftreten. Dies ist, z.B., bei mehrstufigen chemischen Reaktionen der Fall, wenn die Einzelreaktionen sehr unterschiedliche Zeitkonstanten besitzen. Aber auch bestimmte Approximationsverfahren für parabolische, partielle Differentialgleichungen führen auf gewöhnliche Anfangswertprobleme dieser Art mit großen Dimensionen n. Eine ausführlichere Darstellung findet sich in den Büchern von Hairer/Wanner-2 und Strehmel/Weiner.

Ein recht allgemeiner Ansatz (vgl. Numerik-2B) zur Konstruktion von Integrationsverfahren für das APW (1.0.1) ist zunächst die Approximation in diskreten Zeitschritten von $t_{m-1} \geq t_a$ nach $t_m = t_{m-1} + h_{m-1}$, $m \geq 1$. Dann integriert man beide Seiten der Dgl über dieses Intervall und erhält die Integralgleichung

$$u(t_{m-1}+h) = u(t_{m-1}) + h\int_0^1 f(t_{m-1}+hx, u(t_{m-1}+hx))\,dx. \tag{2.1.1}$$

Durch Approximation des Integrals mit Quadraturformeln kann man Einschritt- oder Mehrschrittverfahren erzeugen. Bei der Klasse der Runge-Kutta-(Einschritt-)Verfahren wählt man Quadraturknoten $c_i \in [0,1]$ und approximiert die unbekannten Zwischenwerte ("Stufen") durch einfache Näherungen. Ein *Runge-Kutta-Verfahren* mit s Stufen lautet daher

$$\begin{aligned} k_i &= \quad f(t_{m-1}+hc_i, y_{m-1}+h\sum_{j=1}^{s} a_{ij}k_j), \quad i=1,\ldots,s, \\ y_m &= \quad y_{m-1}+h\sum_{j=1}^{s} b_j k_j, \end{aligned} \tag{2.1.2}$$

und wird kompakt durch das Butcher-Tableau seiner Koeffizienten beschrieben

$$\begin{array}{c|c} \mathbf{c} & A \\ \hline & b^{\mathsf{T}} \end{array} = \begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array}$$

Damit die triviale Dgl $y' = 1$ exakt integriert wird, ist $c_i = \sum_{j=1}^{s} a_{ij}$ ($\mathbf{c} = A\mathbb{1}$) zu wählen. Wenn A eine strikt untere Dreieckmatrix ist, lassen sich die Stufen der Reihe nach berechnen,

das Verfahren ist dann *explizit*. Das einfachste explizite Verfahren, das in jedem expliziten RK-Verfahren wegen $c_1 = 0$ als erste Stufe steckt, ist das *Euler-Verfahren*

$$k_1 = f(t_{m-1}, y_{m-1}) \Rightarrow \quad y_m = y_{m-1} + hf(t_{m-1}, y_{m-1}). \tag{2.1.3}$$

Die Analyse dieser Verfahren führt über die Begriffe *Konsistenz* und *Stabilität* zur *Konvergenz*. Die Stabilität ist bei RK-Verfahren immer gegeben, wenn die rechte Seite f einer Lipschitzbedingung

$$\|f(t,y) - f(t,v)\| \le L\|y - v\| \quad \forall y, v \in \mathbb{R}^n \tag{2.1.4}$$

erfüllt. Allerdings ist diese Bedingung für viele wichtige Problemklassen unrealistisch, zur Problematik steifer Probleme werden zwei typische Beispiele betrachtet.

Beispiel 2.1.1 Bei singulär gestörten Dgln, etwa im $\mathbb{R}^2$ für $u(t) = (v(t), w(t))$ ist ein Teil der Ableitungen mit einem sehr kleinen Vorfaktor $\varepsilon > 0$ versehen, etwa wie in

$$\begin{aligned} v' &= f(v,w) \\ \varepsilon w' &= g(v,w). \end{aligned} \tag{2.1.5}$$

Bei Division durch ε bekommt g also einen großen Vorfaktor. Dazu wird folgende lineare Dgl mit $\varepsilon = 1999/999000 \doteq 0.002$ betrachtet

$$\begin{aligned} v' &= -v + w \\ \varepsilon w' &= v - 2w \end{aligned} \iff u'(t) = Gu(t) := \begin{pmatrix} -1 & 1 \\ \frac{1}{\varepsilon} & -\frac{2}{\varepsilon} \end{pmatrix} u(t), \tag{2.1.6}$$

mit Anfangswert $u(0) = u_0 := (1000/999, -999/1999) \cong (1, -1/2)^\mathsf{T}$. Da die Eigenwerte der Koeffizientenmatrix G die Werte $\lambda_1 = -999/1999 \cong -\frac{1}{2}$ und $\lambda_2 = -1000$ sind, lautet die exakte Lösung dieses Problems

$$u(t) = \begin{pmatrix} e^{\lambda_1 t} + \frac{1}{999} e^{-1000t} \\ \frac{1000}{1999} e^{\lambda_1 t} - e^{-1000t} \end{pmatrix}. \tag{2.1.7}$$

Bis auf eine sehr kurze Startphase $0 \le t \le 40/1000$ verhält sich die Lösung wie die Funktion $\binom{1}{1000/1999} e^{\lambda_1 t}$, trivialerweise gilt auch $u(t) \to 0,\ t \to \infty$. Wird das Problem mit dem (expliziten) Euler-Verfahren und konstantem h behandelt, erhält man die numerische Lösung aus der Rekursion

$$y_m = (I + hG) y_{m-1} \quad \Rightarrow \quad y_m = \begin{pmatrix} 1 \\ \frac{1000}{1999} \end{pmatrix} (1 + \lambda_1 h)^m - \begin{pmatrix} \frac{-1}{999} \\ 1 \end{pmatrix} (1 - 1000h)^m.$$

Hier bekommt man das gewünschte Verhalten $y_m \to 0$ $(m \to \infty)$ offensichtlich nur unter der Bedingung $|1 - 1000h| < 1$, d.h., unter der starken Schrittweiteneinschränkung

$$h < \frac{2}{1000} = \frac{1}{500}.$$

Im anderen Fall wächst y_m oszillierend stark an. Obwohl also (z.B.) $e^{\lambda_2 t} \le 10^{-8}$ für $t \ge 0.02$ ist und die Lösung für solche Argumente vollkommen glatt aussieht, auch durch einen Polygonzug mit größeren Schrittweiten h gut approximiert werden kann, erzwingt das Vorhandensein des Eigenwerts $\lambda_2 = -1000$ im Problem die Verwendung kleiner Schrittweiten *bei diesem Verfahren*.

Beim *impliziten* Euler-Verfahren wird das Integral in der zur Dgl äquivalenten Integralgleichung (2.1.1)

$$u(t_{m-1}+h) = u(t_{m-1}) + h\int_0^1 f(t_{m-1}+xh, u(t_{m-1}+xh))\,dx$$

durch den Funktionswert am *rechten* Rand approximiert, $c_1 = a_{11} = 1$. Bei diesem Verfahren,

$$\left.\begin{array}{rcl} k_1 &=& f(t_m, y_{m-1}+hk_1) \\ y_m &=& y_{m-1}+hk_1 \end{array}\right\} \iff \quad y_m - hf(t_m, y_m) = y_{m-1} \tag{2.1.8}$$

sind k_1 bzw. y_m nur implizit definiert, müssen also durch Auflösung eines (nicht-)linearen Gleichungssystems bestimmt werden. Im linearen Beispiel (2.1.6) führt dies auf die Beziehungen

$$y_m - hGy_m = y_{m-1} \iff (I-hG)y_m = y_{m-1} \iff y_m = (I-hG)^{-1}y_{m-1}.$$

Mit Hilfe der Eigenwertzerlegung von G kann man auch hier wieder das Ergebnis für konstante Schrittweite h explizit bestimmen,

$$y_m = \begin{pmatrix} 1 \\ \frac{1000}{1999} \end{pmatrix} \frac{1}{(1-\lambda_1 h)^m} - \begin{pmatrix} \frac{-1}{999} \\ 1 \end{pmatrix} \frac{1}{(1+1000h)^m}.$$

Da jetzt $1/(1-\lambda_1 h) < 1$ und $1/(1+1000h) < 1$ gilt *für alle* $h > 0$, ist zumindestens das asymptotische Verhalten der Lösung u reproduziert, $y_m \to 0,\ m \to \infty$ für alle Schrittweiten. ■

Die beiden Beispielverfahren unterscheiden sich also ganz wesentlich in ihren Stabilitätseigenschaften. Nur das implizite Verfahren muß sich bei der Schrittweitenwahl ausschließlich nach den Genauigkeitsanforderungen richten. Eine informelle Definition von *steifen AWPen* ist die, dass implizite Verfahren effizienter sind als explizite. Bei singulär gestörten Problemen (2.1.5) kann man die einleitende Bemerkung zur Dynamik erläutern. Unter gewissen Voraussetzungen an g und für kleines $0 < \varepsilon \ll 1$ bewegt sich jede Lösung sehr schnell in die Nähe der durch (den Grenzfall $\varepsilon = 0$) $g(v,w) = 0$ definierte Kurve ("Mannigfaltigkeit" für $n > 1$), diese ist *attraktiv* bzw. stabil.

Beispiel 2.1.2 Linienmethode bei parabolischen Gleichungen: Wenn eine Bakterienpopulation, die in einem Reagenzglas nach einem Wachstumsgesetz $U'(t) = g(U(t))$ wachsen würde, auf einer (eindimensionalen) wäßrigen Membran mit Ortsvariable $x \in [0,1]$ lebt, kommt die Diffusion hinzu. Das Wachstum hängt dann von Zeit und Ort ab und wird dann durch die Reaktions-Diffusions-Gleichung

$$U_t(t,x) = \alpha U_{xx} + g(U), \quad U_x(0) = U_x(1) = 0,\ U(x,0) = b(x),$$

modelliert. Die Randbedingungen bei $x \in \{0,1\}$ beschreiben den fehlenden Austausch dort. Als einfache Ortsdiskretisierung kann man über das x-Intervall ein Gitter $x_i = (i-\frac{1}{2})\xi$, $i = 1,\ldots,n$, $\xi = 1/n$, legen und die Ableitung U_{xx} durch den zweiten Differenzenquotienten ersetzen:

$$U_{xx}(t,x) = \frac{1}{\xi^2}\Big(U(t,x-\xi) - 2U(t,x) + U(t,x+\xi)\Big) + O(\xi^2). \tag{2.1.9}$$

Die Randbedingung $U_x(0) = 0$ kann man mit dem gleichen Fehler $O(\xi^2)$ durch die Symmetriebedingung $U(-\frac{1}{2}\xi) = U(\frac{1}{2}\xi)$ ersetzen, und man eliminiert damit den Wert $U(-\frac{1}{2}\xi)$, der im Differenzenquotienten (2.1.9) an der Stelle x_1 auftritt. Analoges gilt am rechten Rand. Unter Vernachlässigung der Fehler ($\rightarrow$ Konvergenzanalyse) bekommt man für die Näherungsfunktionen $u_i(t) \cong U(t, x_i)$ das folgende System von *gewöhnlichen*(!) Dgln

$$u'(t) = \begin{pmatrix} u_1' \\ u_2' \\ u_3' \\ \vdots \\ u_{n-1}' \\ u_n' \end{pmatrix} = \frac{\alpha}{\xi^2} \begin{pmatrix} -1 & 1 & & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ & & & & 1 & -1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{n-1} \\ u_n \end{pmatrix} + \begin{pmatrix} g(u_1) \\ g(u_2) \\ g(u_3) \\ \vdots \\ g(u_{n-1}) \\ g(u_n) \end{pmatrix}$$

Für eine genaue Approximation ist n groß zu wählen und daher auch $1/\xi^2 = n^2$. Schon für $g \equiv 0$ kann daher die Lipschitzkonstante $L = 4\alpha n^2$ sehr groß sein. Man kann aber leicht zeigen, dass die Matrix negativ semidefinit ist, der größte Eigenwert ist 0 (zum EV $\mathbb{1}$), der kleinste $\cong -L = -4\alpha n^2$. Diskretisierungen parabolischer Dgln sind eine der wichtigsten Beispiele für steife AWPe, hier für $\alpha \cong 1$. In der Praxis sind aber auch Diffusionskonstanten $\alpha \cong 10^{-5}$ realistisch, das System ließe sich dann bis $n \cong 300$ durchaus noch mit expliziten Verfahren lösen.

2.2 Absolute Stabilität

Zur Überprüfung wichtiger Eigenschaften von Verfahren betrachtet man das Verhalten der numerischen Lösungen im Vergleich zur exakten Lösung bei bestimmten *Testproblemen*. Die einfachste Testgleichung ist die skalare, linear-homogene

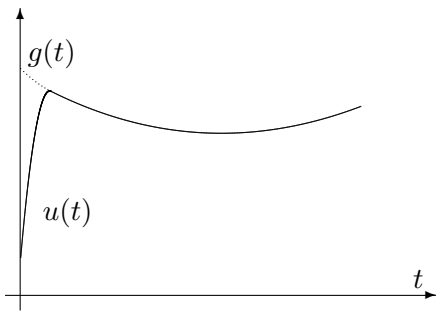
$$u'(t) = \lambda u(t), \quad u(0) = 1, \quad \lambda \in \mathbb{C}, \quad \operatorname{Re}\lambda \leq 0, \tag{2.2.1}$$

die aber über die Eigenvektor-Zerlegung auch Aussagen über Systeme der Form (2.1.6) gestattet. Keine ihrer Lösungen wächst, da $|u(t+x)| = |e^{\lambda x} u(t)| \leq |u(t)| \; \forall x \geq 0$. Eine inhomogene Variante der linearen Testgleichung ist die *Prothero-Robinson-Gleichung* mit einer glatten Funktion $g \in C^1[0, \infty)$,

$$u'(t) = \lambda(u(t) - g(t)) + g'(t), \quad t \geq 0. \tag{2.2.2}$$

Nach Konstruktion ist $u_I = g$ eine spezielle inhomogene Lösung und die allgemeine Lösung von (2.2.2) ist daher

$$u(t) = e^{\lambda t}(u(0) - g(0)) + g(t), \quad t \geq 0.$$



Für $\operatorname{Re}\lambda \ll -1$ bewegt sich diese extrem schnell auf die Funktion $g(t)$ zu (transiente Phase), danach wird sie glatt und kann dann numerisch mit großen Schrittweiten approximiert werden. Das ist aber nur mit impliziten Verfahren möglich.

Runge-Kutta-Verfahren (2.1.2) führen beim Testproblem (2.2.1) auf die skalaren Gleichungen

$$\begin{aligned} y_m &= y_{m-1} + h\sum_{j=1}^{s} b_j k_j, \\ k_i &= \lambda y_{m-1} + h\lambda \sum_{j=1}^{s} a_{ij} k_j \quad \Longleftrightarrow \quad (I - zA)k = \lambda y_{m-1}\mathbb{1} \end{aligned}$$

wobei $z := h\lambda$, $k := (k_1, .., k_s)^\mathsf{T}$, $A = (a_{ij})$, $\mathbb{1} := (1, .., 1)^\mathsf{T}$ gesetzt wurde. Analog sei $b = (b_1, .., b_s)^\mathsf{T}$. Aus dieser Beziehung ergibt sich die Rekursion $y_{m-1} \to y_m$ in geschlossener Form

$$y_m = y_{m-1} + hb^\mathsf{T}k = \Big(1 + zb^\mathsf{T}(I - zA)^{-1}\mathbb{1}\Big)y_{m-1} = \varphi(z)y_{m-1}, \tag{2.2.3}$$

$$\varphi(z) := 1 + zb^\mathsf{T}(I - zA)^{-1}\mathbb{1}. \tag{2.2.4}$$

Einschrittverfahren führen also bei (2.2.1) auf eine Beziehung $y_m = \varphi(h\lambda)y_{m-1}$, die insbesondere nur vom Produkt $h\lambda = z$ abhängt. Für die exakte Lösung u gilt dagegen $u(t_m) = e^{h\lambda}u(t_{m-1})$ und beim Vergleich kommt es also i.w. auf die Vorfaktoren $\varphi(z) \cong e^z$ (?!) an. In (2.2.3) ist $\varphi(z)$ eine rationale Funktion der komplexen Variabeln z, die sogenannte *Stabilitätsfunktion*. Für explizite RK-Verfahren wurden diese schon in der Numerik 2B berechnet:

Verfahren:	expl. Euler	Runge u.Heun (Num2B, Nr.2/3)	klass. Runge-Kutta (Nr.4)	impl. Euler
$\varphi(z) =$	$1 + z$	$1 + z + \frac{1}{2}z^2$	$1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4$	$\frac{1}{1-z}$

Offensichtlich sind die Stabilitätsfunktionen der expliziten Verfahren ganze rationale Funktionen, also Polynome. Für explizite Verfahren ist ja A eine strikt untere Dreiecksmatrix, also nilpotent mit $A^s = 0$. Daher ist die Neumannreihe ein Polynom,

$$\varphi(z) = 1 + zb^\mathsf{T}(I - zA)^{-1}\mathbb{1} = 1 + \sum_{j=0}^{s-1}(b^\mathsf{T}A^j\mathbb{1})z^{j+1}.$$

Wenn ein Verfahren mit Ordnung p konvergiert, bildet die Stabilitätsfunktionen natürlich eine rationale Approximationen an die exakte Lösung der Testgleichung, die e-Funktion mit $\varphi(z) = e^z + \mathcal{O}(z^{p+1}), z \to 0$.

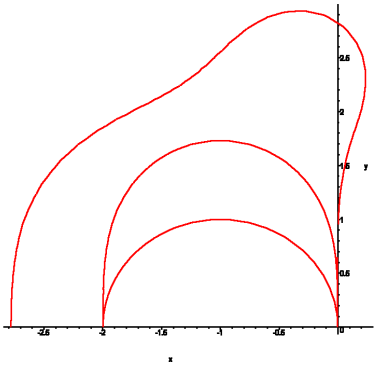
Bei der Testgleichung (2.2.1), welche nur beschränkte Lösungen besitzt, interessieren diejenigen Argumentwerte $z = h\lambda$, für die die numerische Lösung y ebenfalls nicht wächst.

Definition 2.2.1 *Ein Einschrittverfahren besitze die Stabilitätsfunktion φ, d.h. bei Anwendung auf (2.2.1) führe es auf die Beziehung $y_m = \varphi(h\lambda)y_{m-1}$. Das Verfahren heißt* absolut stabil *bei $z \in \mathbb{C}$, falls $|\varphi(z)| \le 1$ ist. Die Menge*

$$S := \{z \in \mathbb{C} : \ |\varphi(z)| \le 1\} \tag{2.2.5}$$

heißt Stabilitätsbereich *des Verfahrens.*

Zur stabilen Integration von (2.2.1) muß bei einem gegebenen Verfahren die Schrittweite h so gewählt werden, dass $h\lambda \in S$ gilt, da sonst unsinnige, wachsende Lösungen auftreten können. Die Graphik zeigt die Stabilitätsbereiche (mit $\operatorname{Im} z \geq 0$) für explizites Euler-, Heun- und Runge-Kutta-Verfahren, von innen nach außen. Für die in $\mathbb{R}$ liegenden *Stabilitätsintervalle* $S \cap \mathbb{R}$ der Beispielverfahren gilt



Verfahren	Intervall
expl. Euler	$[-2, 0]$
Runge u.Heun	$[-2, 0]$
klass.Runge-Kutta	$[-2.78, 0]$
impl. Euler	$(-\infty, 0]$

Die in Beispiel 2.1.1 für das explizite Euler-Verfahren gefundene Schrittweiten-Einschränkung $h \leq 2/(-\lambda)$ entspricht also der Bedingung $h\lambda \in [-2, 0] = S \cap \mathbb{R}$. Von allen erwähnten Verfahren kommt bei Anwendung auf Probleme der Form (2.1.6) nur das implizite Eulerverfahren ohne Schrittweitenbeschränkung aus, da dessen Stabilitätsbereich die gesamte negative reelle Achse umfaßt. Da bei expliziten Verfahren die Stabilitätsfunktion φ ein Polynom ist, mit $|\varphi(z)| \to \infty,\ z \to \infty$, besitzen diese Verfahren nur beschränkte Stabilitätsbereiche.

Daher sind zur Lösung *steifer* Probleme, bei denen schnell und langsam ausklingende Lösungen auftreten können, nur implizite Verfahren sinnvoll, da nur diese sich ausschließlich nach Genauigkeitsforderungen richten müssen. Bei expliziten Verfahren dagegen wird die Schrittweite nicht durch die gewünschte Genauigkeit bestimmt, sondern (sehr restriktiv) durch Stabilitätsschwächen. Eine automatische Schrittweitensteuerung erzeugt hier eine um den Grenzfall schwankende Schrittweitenfolge (vgl. Demo-Beispiel). In Anlehnung an das Lösungsverhalten der Dgl (2.2.1) zeichnet man Verfahren mit analogen Dämpfungseigenschaften aus.

Definition 2.2.2 *Ein Einschrittverfahren heißt* A-stabil*, wenn gilt*

$$S \supseteq \mathbb{C}_- := \{z \in \mathbb{C} :\ Re\, z \leq 0\}, \quad d.h., \quad |\varphi(z)| \leq 1\ \forall z \in \mathbb{C}_-.$$

Das Verfahren heißt L-stabil*, wenn zusätzlich* $\varphi(-\infty) = 0$ *ist.*

Da die Stabilitätsfunktionen von RK-Verfahren rational sind, gilt natürlich $\varphi(-\infty) = \varphi(\infty)$. Bei L-stabilen Verfahren bleiben die Lösungen nicht nur beschränkt für $\lambda \to -\infty$, sondern sie konvergieren wie die exakte Lösung gegen null. Dies hat den praktischen Vorteil, dass Fehler in Eigenräumen zu negativen Eigenwerten ($\operatorname{Re}\lambda \ll 0$) stark gedämpft werden. Offensichtlich ist von den betrachteten Verfahren nur das implizite Eulerverfahren A-stabil, da hier die Bedingung $|\varphi(z)| = |\frac{1}{1-z}| \leq 1 \iff 1 \leq |1-z|^2 = 1 - 2\operatorname{Re} z + |z|^2$ für $\operatorname{Re} z \leq 0$ sicher erfüllt ist. Außerdem ist es offensichtlich auch L-stabil. Der folgende Satz enthält ein algebraisches Kriterium für die A-Stabilität von Runge-Kutta-Verfahren, welches später eingesetzt wird.

Satz 2.2.3 *Ein hinreichendes Kriterium für die A-Stabilität des Verfahrens (2.1.2) ist, dass die folgende Matrix B regulär und beide genannten positiv semidefinit sind:*

$$B := diag(b_i), \quad BA + A^{\mathsf{T}}B - bb^{\mathsf{T}}.$$

Beweis Zunächst werden einige Umformulierungen des Problems vorgenommen. Als erstes wird die (Cayley-) Transformation

$$\mathcal{T} : w \mapsto v = \frac{1-w}{1+w} \iff w = \frac{1-v}{v+1} \quad v, w \in \mathbb{C}$$

betrachtet, sie ist ihre eigene Inverse. Die Bedingungen

$$|v| \le 1 \iff 1 - 2\mathrm{Re}\, w + |w|^2 = |1-w|^2 \le |1+w|^2 = 1 + 2\mathrm{Re}\, w + |w|^2 \iff \mathrm{Re}\, w \ge 0$$

sind äquivalent, die rechte Halbebene $\mathbb{C}_+$ wird durch $\mathcal{T}$ auf den Einheitskreis abgebildet. Für die Stabilitätsfunktion $\varphi(z) = 1 + b^{\mathsf{T}}(\frac{1}{z}I - A)^{-1}\mathbb{1}$, $z \neq 0$, des RK-Verfahrens kann daher statt der Bedingung $|\varphi| \le 1$ auch

$$0 \le \mathrm{Re}\, \frac{1-\varphi}{1+\varphi} = \mathrm{Re}\, \frac{-b^{\mathsf{T}}(\frac{1}{z}I - zA)^{-1}\mathbb{1}}{2 + b^{\mathsf{T}}(\frac{1}{z}I - A)^{-1}\mathbb{1}} \tag{2.2.6}$$

untersucht werden. Dazu wird jetzt die um eine Rang-1-Matrix abgeänderte Matrix $A' = A - \frac{1}{2}\mathbb{1}b^{\mathsf{T}}$ eingeführt, hier folgt mit der Rang-1-Formel und $M := (\frac{1}{z}I - A)$

$$\frac{1}{2}b^{\mathsf{T}}(\frac{1}{z}I - A')^{-1}\mathbb{1} = \frac{1}{2}b^{\mathsf{T}}(\frac{1}{z}I - A + \frac{1}{2}\mathbb{1}b^{\mathsf{T}})^{-1}\mathbb{1} = \frac{1}{2}b^{\mathsf{T}}M^{-1}\mathbb{1} - \frac{1}{2}\frac{b^{\mathsf{T}}M^{-1}\mathbb{1}b^{\mathsf{T}}M^{-1}\mathbb{1}}{2 + b^{\mathsf{T}}M^{-1}\mathbb{1}} = \frac{b^{\mathsf{T}}M^{-1}\mathbb{1}}{2 + b^{\mathsf{T}}M^{-1}\mathbb{1}}.$$

Dies ist gerade das Negative von (2.2.6) und für reguläres B gilt mit $b = B\mathbb{1}$, dass

$$|\varphi(z)| \le 1 \forall z \in \mathbb{C}_- \iff 0 \le \mathrm{Re}\, b^{\mathsf{T}}(A' - \frac{1}{z}I)^{-1}\mathbb{1} = \mathrm{Re}\, b^{\mathsf{T}}(BA - \frac{1}{2}bb^{\mathsf{T}} - \frac{1}{z}B)^{-1}b \,\forall z \in \mathbb{C}_-.$$

Der letzte Ausdruck ist eine quadratische Form (mit $x = b \in \mathbb{R}^n$), hinreichend für die Nichtnegativität ist die reelle Definitheit der auftretenden Matrix. Dazu gilt für $M \in \mathbb{C}^{s\times s}$, dass

$$\mathrm{Re}\, x^* M^{-1} x \ge 0 \forall x \in \mathbb{C}^s \quad \Leftarrow \quad 2\mathrm{Re}\, y^* M y = y^*(M + M^*)y \ge 0 \forall y \in \mathbb{C}^s, \tag{2.2.7}$$

im konkreten Fall daher

$$BA + AB^{\mathsf{T}} - bb^{\mathsf{T}} - 2(\mathrm{Re}\, \frac{1}{z})B \text{ pos.semi-definit } \forall \mathrm{Re}\, z \le 0.$$

Dies entspricht der Behauptung. ∎

Der Begriff A-Stabilität bezieht sich auf die extrem einfache, skalare Testgleichung (2.2.1). Da man allgemein beim Vergleich von 2 Lösungen $u' = f(.,u)$, $v' = f(.,v)$ über den Mittelwertsatz aber eine Linearisierung

$$u'(t) - v'(t) = F'(t)(u(t) - v(t)), \quad F'(t) := \int_0^1 f_y(t, v(t) + r(u(t) - v(t)))\, dr$$

durchführen kann, bekommt man über die Eigenwertbetrachtung von F' notwendige Bedingungen für eine stabile Integration von Problemen mit Eigenwerten von F' in der linken komplexen Halbebene (z.B. parabolische Systeme). Es gibt aber sogar eine nichtlineare Variante der A-Stabilität. Sie betrifft AWPe mit rechten Seiten f, die eine *einseitige Lipschitzbedingung* erfüllen:

$$(y-v)^{\mathsf{T}}(f(t,y)-f(t,v)) \le \mu\|y-v\|_2^2 \quad \forall y,v \in \mathbb{R}^n, \tag{2.2.8}$$

mit $\mu \in \mathbb{R}$. Insbesondere ist hier $\mu \le 0$ möglich. Im linearen Fall $f(t,y) = My+g$ reduziert sich diese Bedingung auf die reelle(!) Definitheit $w^{\mathsf{T}}Mw \le \mu\|w\|_2^2$ mit $w = y-v$, in der analog zu (2.2.7) nur die symmetrische Matrix $\frac{1}{2}(M+M^{\mathsf{T}})$ eine Rolle spielt. Das folgende Lemma besagt, dass sich für $\mu \le 0$ zwei Lösungen der Dgl mit rechter Seite f nicht voneinander entfernen als Verallgemeinerung der Aussage $|e^{t\lambda}| \le 1$, $\mathrm{Re}\,\lambda t \le 0$ für die Lösungen von $u' = \lambda u$.

Lemma 2.2.4 *Es seien $u, v \in C^1[t_a, t_e)$ zwei Lösungen der Dgl $y' = f(t,y)$ mit einer rechten Seite f, die eine einseitige Lipschitzbedingung (2.2.8) erfüllt. Dann gilt*

$$\|u(t)-v(t)\|_2 \le e^{\mu(t-t_a)}\|u(t_a)-v(t_a)\|_2, \quad t_a \le t < t_e.$$

Beweis Die Funktion $d(t) = e^{-2\mu t}\|u(t)-v(t)\|_2^2$ ist differenzierbar. Für ihre Ableitung gilt

$$\begin{aligned} d'(t) &= -2\mu e^{-2\mu t}\|u(t)-v(t)\|_2^2 + 2e^{-2\mu t}(u(t)-v(t))^{\mathsf{T}}(u'(t)-v'(t)) \\ &= -2\mu e^{-2\mu t}\|u(t)-v(t)\|_2^2 + 2e^{-2\mu t}(u(t)-v(t))^{\mathsf{T}}(f(t,u(t))-f(t,v(t))) \\ &\le 2(\mu-\mu)e^{-2\mu t}\|u(t)-v(t)\|_2^2 = 0. \end{aligned}$$

Also wächst die Funktion nie, $d(t) - d(t_a) = \int_{t_a}^t d'(t)dt \le 0$. ∎

Ein analoges Verhalten numerischer Verfahren heißt B-Stabilität:

Definition 2.2.5 *Ein Verfahren für Anfangswertprobleme heißt* B-stabil*, wenn es bei Anwendung auf ein Problem mit einseitiger Lipschitzbedingung (2.2.8) und $\mu \le 0$ Näherungen y_m, v_m erzeugt mit*

$$\|y_m - v_m\|_2 \le \|y_{m-1} - v_{m-1}\|_2, \quad m \ge 1.$$

Dies ist natürlich eine sehr starke Bedingung, das implizite Eulerverfahren erfüllt sie aber. Denn für 2 Lösungsfolgen y_m, v_m dieses Verfahrens (2.1.8) gilt

$$\left.\begin{aligned} y_m - hf(t_m,y_m) &= y_{m-1} \\ v_m - hf(t_m,v_m) &= v_{m-1} \end{aligned}\right\} \Rightarrow y_m - v_m - h(f(t_m,y_m)-f(t_m,v_m)) = y_{m-1} - v_{m-1},$$

und durch Multiplikation mit $(y_m - v_m)^{\mathsf{T}}$ folgt aus (2.2.8), dass

$$\begin{aligned} (1-h\mu)\|y_m - v_m\|_2^2 &\le \\ \|y_m - v_m\|_2^2 - h(y_m-v_m)^{\mathsf{T}}(f(t_m,y_m)-f(t_m,v_m)) &= (y_m-v_m)^{\mathsf{T}}(y_{m-1}-v_{m-1}) \\ &\le \|y_m - v_m\|_2\|y_{m-1}-v_{m-1}\|_2. \end{aligned}$$

Für $h\mu < 1$ erhält man eine zu Lemma 2.2.4 analoge Schranke, wobei aber die Exponentialfunktion im Vorfaktor durch $1/(1-h\mu) = \varphi(h\mu)$, also die Stabilitätsfunktion des Eulerverfahrens zu ersetzen ist:

$$\|y_m - v_m\|_2 \leq \frac{1}{1-h\mu}\|y_{m-1} - v_{m-1}\|_2.$$

Für $\mu \leq 0$ bestätigt dies also die B-Stabilität aus Definition 2.2.5. Diese starke Stabilitätseigenschaft hat man nur bei wenigen Verfahren unter anderem in der folgenden Klasse.

2.3 Implizite Runge-Kutta-Verfahren

Wenn die Koeffizientenmatrix $A \in \mathbb{R}^{s\times s}$ eines Runge-Kutta-Verfahrens nicht strikt untere Dreickstruktur hat, ist es ein implizites ("IRK") Verfahren, bei dem die Stufen k_i durch Lösung von Gleichungssystemen bestimmt werden müssen. Bei vollbesetzter Koeffizientenmatrix A hat dieses System die Größe $s \cdot n$, ein mehrfaches der Raumdimension!

Bei den IRK-Verfahren gibt es eine Klasse, bei der gleichzeitig sehr hohe Konvergenzordnungen und gute Stabilitätseigenschaften aufeinandertreffen. Diese Verfahren ergeben sich aus Anwendung der Gauß-Quadraturformeln auf die Integralgleichung (2.1.1). Dabei sind die Stützstellen c_i in (2.1.2) Gaußknoten, d.h. die Nullstellen des m-ten Legendre-Polynoms (vgl. Numerik 2B). Alle Koeffizienten sind Integrale der zugehörigen Lagrangepolynome

$$L_i(t) := \prod_{\substack{j=1\\ j\neq i}}^{s} \frac{t-c_j}{c_i-c_j},\ a_{ij} := \int_0^{c_i} L_j(t)\,dt,\ b_i := \int_0^1 L_i(t)\,dt, \quad i,j = 1,\ldots,s. \tag{2.3.1}$$

Die b-Koeffizienten sind die bekannten Gauß-Quadraturgewichte $b_i > 0$, $i = 1,\ldots,s$. Das einfachste Gauß-IRK-Verfahren gehört zum Knoten $c_1 = \frac{1}{2}$ (Rechteck-/Mittelpunktregel), es lautet

$$k_1 = f(t_{m-1} + \frac{h}{2}, y_{m-1} + \frac{h}{2}k_1), \quad y_m = y_{m-1} + hk_1.$$

Durch Elimination von k_1 läßt es sich in einem impliziten Schritt formulieren

$$y_m - y_{m-1} = hf\Big(t_{m-1} + \frac{h}{2}, \frac{1}{2}(y_{m-1} + y_m)\Big), \tag{2.3.2}$$

die zugehörige Stabilitätsfunktion ist

$$\varphi(z) = \frac{1+z/2}{1-z/2} = 1 + z + \frac{z^2}{2} + \frac{z^3}{4(1-z/2)}. \tag{2.3.3}$$

Die Entwicklung nach z-Potenzen zeigt, dass die Ordnung (nicht größer als) zwei ist.

Satz 2.3.1 *Für beliebiges $s \in \mathbb{N}$ sind die Gauß-Runge-Kutta-Verfahren A-stabil und B-stabil. Die Konsistenzordnung des s-stufigen Verfahrens ist $2s$.*

Die Aussage zur A-Stabilität folgt aus Satz 2.2.3, da man bei den Gaußverfahren tatsächlich

$$BA + A^{\mathsf{T}}B - bb^{\mathsf{T}} = 0$$

nachprüft. Die Nullmatrix ist natürlich semidefinit. Die gleiche Bedingung garantiert auch die nichtlineare B-Stabilität. Die Beweise (auch zur Konvergenzordnung $2s$) werden nicht geführt, da die Gauß-Runge-Kutta-Verfahren zwar exzellente theoretische Eigenschaften besitzen, aber nur geringe praktische Bedeutung haben (s.u.). Die Gauß-IRK-Verfahren haben wegen der symmetrischen Lage der Knoten, $c_{s+1-i} = 1 - c_i$ einen Zeit-symmetrischen Aufbau, bei einer Vorzeichenänderung im Testproblem $y' + \lambda y = 0$ bekommt man die gleichen Näherungsfolge rückwärts in der Zeit, vgl. (2.3.2). Daher gilt für die Stabilitätsfunktion

$$\varphi(-z) = \frac{1}{\varphi(z)} \iff \varphi(z)\varphi(-z) \equiv 1 \Rightarrow \quad |\varphi(ix)| = 1, x \in \mathbb{R}. \tag{2.3.4}$$

Im Beispiel (2.3.3) ist dies offensichtlich. Diese Eigenschafte bedeutet einen *Erhaltungssatz*, für rein imaginäres $\lambda \in i\mathbb{R}$ wird die Invarianz $|u(t)| \equiv |u_0|$ der exakten Lösung im Verfahren erhalten, $|y_m| \equiv |u_0| \forall m$. Solche Eigenschaften werden in §4 intensiver betrachtet.

Die (Gauß-) IRK-Verfahren besitzen aber auch gravierende Nachteile. So gilt die Ordnungsaussage $2s$ nicht mehr bei sehr steifen Problemen ("Ordnungsreduktion" bei (2.2.2) für $\operatorname{Re}\lambda \to -\infty$). Zudem sind sie sehr teuer, da man in jedem Verfahrensschritt nichtlineare Gleichungssysteme der vielfachen Größe sn, z.B. mit dem Newtonverfahren, lösen muß. Es gibt verschiedene Ansätze zur effektiveren Implementierung von IRK-Verfahren allgemein. Diese schränken aber entweder die Struktur ein, und damit leider auch die erreichbare Ordnung auf $s+1$, oder besitzen andere Nachteile.

Bei sehr schwierigen, steifen Problemen kann es doch erforderlich sein, IRK-Verfahren einzusetzen. Dabei ist aber die Eigenschaft (2.3.4) hinderlich, da symmetrische Verfahren wegen $|\varphi(\infty)| = 1$ nicht L-stabil sind (die beiden Eigenschaften sind unvereinbar). Man verwendet sogenannte Radau-Knoten, wo $c_s = 1$ festliegt und die restlichen so festgelegt werden, dass das Knotenpolynom noch orthogonal ist,

$$\omega_s \perp \Pi_{s-2}, \quad \omega(t) = (t - c_1) \cdots (t - c_{s-1})(t - 1).$$

Die Konsistenzordnung der Radau-Verfahren ist $2s - 1$, alle sind L-stabil. Für $s = 3$ bekommt man immerhin Ordnung 5 und beim Code RADAU5 von Hairer/Wanner wird durch Transformationen die Größe der simultan zu lösenden Systeme auf $2n$ vermindert. Das Verfahrenstableau ist

$$\begin{array}{c|c} c & A \\ \hline & b^{\mathsf{T}} \end{array} = \begin{array}{c|ccc} \frac{4-\sqrt{6}}{10} & \frac{88-7\sqrt{6}}{360} & \frac{296-169\sqrt{6}}{1800} & \frac{-2+3\sqrt{6}}{225} \\ \frac{4+\sqrt{6}}{10} & \frac{296+169\sqrt{6}}{1800} & \frac{88+7\sqrt{6}}{360} & \frac{-2-3\sqrt{6}}{225} \\ 1 & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \\ \hline & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \end{array}$$

und die Stabilitätsfunktion

$$\varphi(z) = \frac{1 + \frac{2}{5}z + \frac{1}{20}z^2}{1 - \frac{3}{5}z + \frac{3}{20}z^2 - \frac{1}{60}z^3} = e^z + O(z^6), \ z \to 0.$$

2.4 Linear-implizite Verfahren

Die Verfahren in diesem Abschnitt werden nur für autonome Dgln $f(t,y) = f(y)$ behandelt. Eine effiziente Alternative zu IRK-Verfahren sind *Linear-implizite Runge-Kutta-Verfahren*, etwa *Rosenbrock-Wanner-Verfahren* (ROW). Zur Motivation werde die Lösung der Gleichung (2.1.8) beim impliziten Euler-Verfahren betrachtet. Der erste Schritt der Newton-Iteration mit Startwert $y_m^{[0]} := y_{m-1}$ und $T_m := \frac{\partial f}{\partial y}(y_{m-1})$ ergibt

$$y_m^{[1]} = y_{m-1} + h \underbrace{(I - hT_m)^{-1} f(y_{m-1})}_{=:k_1}.$$

Man kann nun diese Vorschrifft als eigenständiges Verfahren studieren in Bezug auf Stabilität und Konsistenz. Bei mehrstufigen Verfahren ist dabei die Verwendung zusätzlicher Parameter γ, γ_{ij} sinnvoll. Die Verfahrensvorschrift als *ROW-Methode* lautet dann (mit der Matrix $T := \frac{\partial f}{\partial y}(y_{m-1})$),

$$\begin{aligned} (I - h\gamma T)k_i &= f\Big(y_{m-1} + h\sum_{j=1}^{i-1} \alpha_{ij} k_j\Big) + hT \sum_{j=1}^{i-1} \gamma_{ij} k_j, \quad i = 1, \ldots, s, \qquad (2.4.1) \\ y_m &= y_{m-1} + h \sum_{j=1}^{s} b_j k_j. \end{aligned}$$

Damit die Größe der zu lösenden Systeme auf $n \times n$ beschränkt bleibt, sind $(\alpha_{ij})_{i,j=1}^s$ und $\Gamma = (\gamma_{ij})_{i,j=1}^s$ strikt untere Dreieckmatrizen, die Anzahl der Parameter einer s-stufigen Methode ist daher $s^2 + 1$. Die Stabilität diskutiert man bei $f(y) = \lambda y$ mit $T = \lambda$, das ROW-Verfahren entspricht hier einem *diagonal-impliziten* RK-Verfahren mit Koeffizienten $a_{ij} = \alpha_{ij} + \gamma_{ij}$, $a_{ii} = \gamma$. Die Stabilitätsfunktionen solcher Methoden besitzt nur einen reellen Pol $1/\gamma$, es gibt aber A- und L-stabile Verfahren:

Satz 2.4.1 *Die Stabilitätsfunktion einer ROW-Methode (2.4.1) mit $T = \lambda$ ist*

$$\varphi(z) = 1 + \sum_{k=1}^{s} b^{\mathsf{T}} \beta^{k-1} \mathbb{1} \left(\frac{z}{1-\gamma z}\right)^k, \; z \in \mathbb{C}, \quad \beta = (\alpha_{ij} + \gamma_{ij})_{ij=1}^s.$$

Beweis Mit $z = h\lambda$ lauten die Stufengleichungen in (2.4.1) $(1-\gamma z)k = \lambda y \mathbb{1} + z\beta k$. Mit $\varrho(\beta) = 0$ liefert dies

$$\begin{aligned} \varphi(z) &= 1 + zb^{\mathsf{T}}\Big((1-\gamma z)I - z\beta\Big)^{-1}\mathbb{1} = 1 + \frac{z}{1-\gamma z} b^{\mathsf{T}}\Big(I - \frac{z}{1-\gamma z}\beta\Big)^{-1}\mathbb{1} \\ &= 1 + \sum_{j=0}^{s-1} b^{\mathsf{T}} \beta^j \mathbb{1} \left(\frac{z}{1-\gamma z}\right)^{j+1}. \quad \blacksquare \end{aligned}$$

Der Grund für die Verwendung eines einheitlichen γ-Wertes in allen Stufen von (2.4.1) ist, dass dann überall die gleiche Matrix $(I - h\gamma T)$ auftritt und daher pro Schritt $y_{m-1} \to y_m$ nur eine

einzige LR- oder QR-Zerlegung erforderlich wird, sowie s Auflösungen mit unterschiedlichen rechten Seiten. Tatsächlich muß auch nur diese QR-/LR-Zerlegung gespeichert werden, denn T wird auf der rechten Seite von (2.4.1) nicht noch einmal benötigt. Man implementiert nämlich die Stufen in der Form

$$\Big(I - h\gamma T\Big)\Big(k_i + \frac{1}{\gamma}\sum_{j=1}^{i-1}\gamma_{ij}k_j\Big) = f\Big(y_{m-1} + h\sum_{j=1}^{i-1}\alpha_{ij}k_j\Big) + \frac{1}{\gamma}\sum_{j=1}^{i-1}\gamma_{ij}k_j. \tag{2.4.2}$$

Bei großen Raumdimensionen, wie sie bei Semidiskretisierung von parabolischen Dgln auftreten, kann diese LR-Zerlegung den größten Anteil im Gesamtaufwand ausmachen. Dies berücksichtigen die sog. *W-Methoden*, bei denen in (2.4.1) T eine beliebige Matrix sein kann, die f' nur so weit approximiert, um das Verfahren stabil zu machen. So kann man etwa die LR-Zerlegung von f' nur alle paar Schritte neu berechnen. Oder man betrachtet das System (2.4.2) mit der exakten Matrix $(I - h\gamma f')$, löst es aber mit Iterationsverfahren nur näherungsweise. Dann ist die tatsächlich verwendete Matrix T nicht einmal explizit bekannt. Der Unterschied zwischen W- und ROW-Methoden zeigt sich in den Ordnungsbedingungen. Bei Einschrittverfahren betrachtet man den lokalen Fehler (oBdA in t_0) bei Start in der exakten Lösung $u(t_0) = y(t_0)$ und vergleicht die Taylorentwicklung der Lösung

$$u(t_1) = u(t_0 + h) = u_0 + \sum_{k=1}^{p}\frac{h^k}{k!}u^{(k)}(t_0) + \dots \tag{2.4.3}$$

mit der der Näherung y_1 aus (2.4.1), betrachtet als Funktion von h. Da $y_0 = u(t_0) = u_0$ und im autonomen Fall $u'(t) \equiv f(u(t))$ gilt, folgen für die in (2.4.1) auftretenden höheren Ableitungen die Darstellungen

$$\begin{aligned}
u_0' &= f,\\
u_0'' &= f'u' = f'f,\\
u_0^{(3)} &= f''u'u' + f'u'' = f''ff + (f')^2 f,\\
u_0^{(4)} &= f'''fff + 3f''f(f'f) + f'f''ff + (f')^3 f,
\end{aligned}$$

usw., wobei f und seine Ableitungen in u_0 ausgewertet werden. Die hier auftretenden Ausdrücke, z.B. $f''f(f'f)$ (f'' ist eine bilineare Abbildung und operiert auf f und $f'f$), heißen elementare Differentiale und man muß zum Erreichen einer bestimmten Ordnung die Vorfaktoren aller Differentiale in beiden Entwicklungen angleichen. Die Entwicklung von $y_1(h)$ und $k_i = k_i(h)$ bekommt man mit $\gamma_{ii} := \gamma$ rekursiv aus

$$k_i = f\Big(u_0 + h\sum_{j=1}^{i-1}\alpha_{ij}k_j\Big) + hT\sum_{j=1}^{i}\gamma_{ij}k_j = f + \sum_{\ell=1}\frac{h^\ell}{\ell!}f^{(\ell)}\cdot\Big(\sum_{j=1}^{i-1}\alpha_{ij}k_j\Big)^\ell + hT\sum_{j=1}^{i}\gamma_{ij}k_j.$$

Mit den Abkürzungen $c_i = \sum_j \alpha_{ij}$ (vgl. RK-Verfahren), $\beta_{ij} = \alpha_{ij} + \gamma_{ij}$, $\beta_i = \sum_j \beta_{ij}$ bekommt man die Entwicklung

$$u(t_0 + h) - y_1(h) \;=\; h(1 - \sum_i b_i)f + h^2(\frac{1}{2} - \sum_i b_i c_i)(f' - T)f + h^2(\frac{1}{2} - \gamma - \sum_i b_i\beta_i)Tf$$

$$+h^3(\frac{1}{3}-\sum_i b_i c_i^2)f''ff+h^3(\frac{1}{6}-\sum b_i\alpha_{ij}c_j)(f'-T)^2f$$
$$+h^3(\frac{1}{6}-\frac{\gamma}{2}-\sum b_i\alpha_{ij}\beta_j)(f'-T)Lf+h^3(\frac{1}{6}-\frac{\gamma}{2}-\sum b_i\beta_{ij}c_j)T(f'-T)f$$
$$+h^3(\frac{1}{6}-\gamma-\gamma^2-\sum b_i\beta_{ij}\beta_j)T^2f+O(h^4)$$

Zur Konstruktion von Verfahren etwa der globalen Ordnung 3 sind die Verfahrenskoeffizienten so zu wählen, dass alle Klammerterme verschwinden. Durch die Nichtlinearität dieser Bedingungen kann die Lösbarkeit praktisch nur durch eine Konstruktion gezeigt werden. Hier zeigt sich auch klar der Unterschied zwischen ROW- und W-Methoden. Für $T = f'(u_0)$ fallen viele Terme sowieso weg, die Anzahl der *Ordnungsbedingungen* ist viel geringer. Ein Kompromiss zwischen beiden Ansätzen ist die Neuzerlegung in größeren Abständen mit

$$T = f_y(y_{m-1}) + O(h). \tag{2.4.4}$$

Anzahl der Ordnungsbedingungen

Ordnung	1	2	3	4
Anz. Verfahrensparameter	2	5	10	17
ROW-Methode	1	2	4	8
$T = f_y(y_{m-1}) + O(h)$	1	2	5	11
W-Methode	1	3	8	21

Beispiel 2.4.2 Für eine 2-stufige W-Methode der Ordnung 2 sind wegen $c_1 = 0$, $c_2 = \alpha_{21}$ die drei Bedingungen

$$b_1 + b_2 = 1,\ b_2c_2 = \frac{1}{2},\ b_2\beta_{21} = b_2(c_2 + \gamma_{21}) = \frac{1}{2} - \gamma$$

zu erfüllen. Die letzte kann zu $b_2\gamma_{21} = -\gamma$ verkürzt werden. Mit $b_2 = \frac{1}{2c_2}$ folgen $b_1 = 1 - \frac{1}{2c_2}$ und $\gamma_{21} = -2\gamma c_2$. Bei einer Matrix nach (2.4.4) kommen für Ordnung 3 die beiden Bedingungen $\frac{1}{3} = b_2c_2^2 = \frac{1}{2}c_2$ und $\sum_{i,j} b_i\beta_{ij}\beta_j = \frac{1}{6} - \gamma - \gamma^2$ hinzu, die mit $c_2 = \frac{2}{3}$ und $\gamma = \frac{1}{6}(3 + \sqrt{3})$ auch erfüllt werden können (da die Matrix $\beta^2 = 0$ ist).

Man kann auch bei W/ROW-Methoden eingebettete Verfahren studieren zur Schrittweitensteuerung, z.B., existieren verschiedene 4(3)-Paare von ROW-Methoden. Diese sind für Probleme mit Eigenwerten nahe der imaginären Achse oder für mittlere Genauigkeitsforderungen konkurrenzfähig zu den anschließend behandelten BDF-Verfahren. Die linear-impliziten Methoden besitzen nicht die herausragenden Stabilitätseigenschaften der Impliziten Runge-Kutta-Verfahren. Allerdings ist dabei zu beachten, dass die IRK-Verfahren B-Stabilität nur bei exakter Lösung der übergroßen nichtlinearen Stufensysteme besitzen, was die Verfahren noch teurer macht.

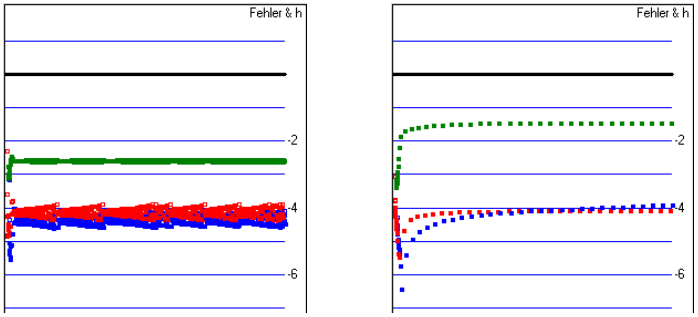
Beispiel 2.4.3 Für $\mu = 800$ hat das AWP zur Dgl

$$\begin{pmatrix} u_1' \\ u_2' \end{pmatrix} = \mu(1 - u_1^2 - u_2^2)\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} + \begin{pmatrix} -u_2 \\ u_1 \end{pmatrix}$$

Lösungen, die alle sehr schnell auf den Einheitskreis zu- und ihn dann entlanglaufen. Für $u(0) = (\frac{1}{2}, 0)^\mathsf{T}$ ist die Lösung

$$u(t) = \frac{1}{\sqrt{1 + c^{-2\mu t}}} \begin{pmatrix} \cos t \\ \sin t \end{pmatrix}, \ c = 3.$$

Die Steifheit des Problems erkennt man beim Einsatz von Einschrittverfahren daran, dass aufgrund der Steuerung Schrittweite und Fehlerschätzung schnell schwanken. So benötigt bei $tol = 10^{-4}$ das Verfahren DOPRI5 3881 Schritte, die ROW-Methode GRK4T dagegen nur 317 Schritte.



Abschließend folgt wieder ein Hinweis auf verfügbare Software. Verschiedene 4-stufige ROW-Methoden sind im Programm ROS4 zusammengefaßt. Ein neueres, linear-implizites Verfahren RODAS von Hairer/Wanner (1995?) erfüllt zusätzliche Ordnungsbedingungen für singulär gestörte Dgln.

Das Programm ROWMAP von Weiner/S./Podhaisky (1997) nutzt spezielle Iterationsverfahren zur Lösung der Stufen (2.4.2), sogenannte Krylov-Verfahren. Diese besitzen hier zwei wichtige Vorteile. Zunächst ist ihre Konvergenz besonders gut bei großen Eigenwerten, bei denen die Inverse $(I - h\gamma f')^{-1}$ für die Stabilität der Integration besonders wichtig ist. Außerdem nutzt diese Iteration nur Informationen der Matrix f' aus Matrix-Vektor-Multiplikationen

$$f'(y)v = \frac{1}{\epsilon}(f(y + \epsilon v) - f(y)) + O(\epsilon), \tag{2.4.5}$$

die in der angegebenen Weise durch Differenzen approximiert werden können. Daher sind sie besonders nutzerfreundlich, denn zur Problembeschreibung ist nur **ein** Unterprogramm für $f(.)$, aber keines zusätzlich für $f'(.)$ zu implementieren ("Matrix-freies" Verfahren).

2.5 Implizite Mehrschrittverfahren

Bei Mehrschrittverfahren greift man bei der Approximation des Integral in (2.1.1) vor allem auf schon bekannte, frühere Werte $f(y_{m-j})$, $j \geq 1$, zurück, d.h. auf Quadraturknoten links von $[0, 1]$. Dies ist der Ansatz der Adams-Verfahren. In Verallgemeinerung der Grundstruktur verwendet man auch Lösungswerte y_{m-j} und erhält die allgemeine Form der linearen Mehrschrittverfahren ($\alpha_0 = 1$)

$$\sum_{j=0}^{r} \alpha_j y_{m-j} = h \sum_{j=0}^{r} \beta_j f(t_{m-j}, y_{m-j}), \quad m \geq r. \tag{2.5.1}$$

Bei den expliziten MS-Verfahren waren aus Stabilitätsgründen (1.Ordnungsbarriere von Dahlquist) nur Methoden sinnvoll mit genau 2 nichtverschwindenen Koeffizienten α_j. Für steife Probleme hat sich dagegen im praktischen Einsatz eine Verfahrensklasse bewährt mit einer gegenteiligen Festlegung. Sie besitzt genau einen nichtverschwindenden Koeffizienten $\beta_0 \neq 0$ und die Koeffizienten α_j gehören zum linksseitigen (in Zeitrichtung rückwärtigen) Differenzenquotienten mit

$$\frac{1}{h}\sum_{j=0}^{r} \alpha_j v(-jh) = \beta_0 v'(0) + \mathcal{O}(h^r),\ h \to 0, \tag{2.5.2}$$

für $v \in C^r[-rh, 0]$. Die Koeffizienten sind

r	$j=$	1	2	3	4	β_0	δ_m
1	$\alpha_j =$	-1				1	90^o
2	$3\alpha_j =$	-4	1			$\frac{2}{3}$	90^o
3	$11\alpha_j =$	-18	9	-2		$\frac{6}{11}$	86^o
4	$25\alpha_j =$	-48	36	-16	3	$\frac{12}{25}$	73^o

Das Verfahren ist implizit, in der Form ($\alpha_0 = 1$)

$$y_m - h\beta_0 f(t_m, y_m) = -\sum_{j=1}^{r} \alpha_j y_{m-j}, \quad m \geq r, \tag{2.5.3}$$

erkennt man, dass in jedem Schritt ein i.a. nichtlineares $n \times n$-System für y_m zu lösen ist wie beim impliziten Eulerverfahren (2.1.8), welches das einfachste BDF-Verfahren mit $r = 1$ ist. Den Konsistenzfehler bekommt man durch Einsetzen der exakten Lösung $u(t)$ und mit $f(t_m, u(t_m)) = u'(t_m)$ gibt (2.5.2) diesen genau an mit $v = u$. Daher besitzt die r-stufige Version dieser BDF-Verfahren (backward difference formula) Ordnung r. Die Stabilitätseigenschaften bei steifen Problemen sind allerdings weniger gut als bei impliziten Einschrittverfahren. Bei der Testgleichung (2.2.1) reduzieren sich BDF-Verfahren auf die Differenzengleichung ($z := h\lambda$)

$$(1 - \beta_0 z)y_m + \sum_{j=1}^{r} \alpha_j y_{m-j} = 0, \quad m \geq r. \tag{2.5.4}$$

Der Lösungsansatz $y_m = C\xi^m$ liefert $0 \stackrel{!}{=} C\Big((1-\beta_0 z)\xi^m + \sum_{j=1}^{r} \alpha_j \xi^{m-j}\Big) = C\xi^{m-r} p_r(\xi, z)\, \forall m$ mit dem *charakteristischen Polynom*

$$p_r(\xi; z) = (1 - \beta_0 z)\xi^r + \alpha_1 \xi^{r-1} + \ldots + \alpha_r.$$

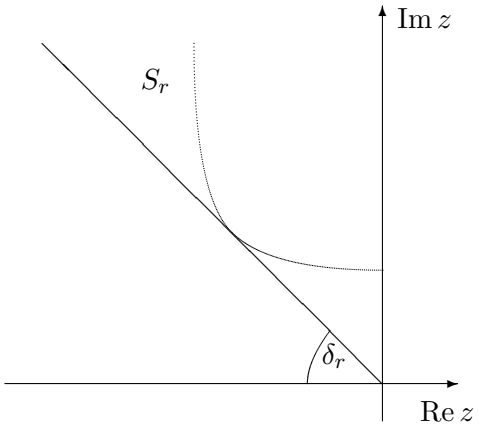
Dessen Nullstellen führen also auf spezielle Lösungen von (2.5.4) und man bekommt die allgemeine Lösung in der Form $y_m = \sum_{j=1}^{r} C_j \xi_j^m$, wenn p_r genau r verschiedene Nullstellen ξ_j besitzt. Bei mehrfachen Nullstellen ist die Darstellung durch Polynomanteile $m^l \xi^m$ zu modifizieren. In Analogie zu den Einschrittverfahren kann man die Menge dieser Nullstellen $\{\xi_j\} = \{\xi_j(z)\}_{j=1}^{r}$ als (mehrwertige) Stabilitätsfunktion des Mehrschrittverfahrens interpretieren. Insbesondere nennt man ein Mehrschrittverfahren *absolut stabil* an der Stelle $z \in \mathbb{C}$, wenn gilt

$$p_r(\xi; z) = 0 \quad \Rightarrow \quad |\xi| \leq 1 \ (\xi \text{ einfach für } |\xi| = 1). \tag{2.5.5}$$

Im Unterschied zu Einschrittverfahren, die für $z = \lambda = 0$ wegen $\varphi(0) = 1$ stabil sind, ist aber schon die Nullstabilität, (2.5.5) bei $z = 0$, eine nichttriviale Anforderung, die aber von den BDF-Verfahren erfüllt wird (für $r \leq 6$). Die weitergehende A-Stabilität, mit (2.5.5) für jedes $z \in \mathbb{C}_-$, ist aber überzogen, es gibt eine zweite *Ordnungsbarriere* von Dahlquist:

Satz 2.5.1 *Die Ordnung eines A-stabilen Mehrschrittverfahrens ist höchstens* 2.

Daher können die Stabilitätsbereiche S_r der BDF-Verfahren höherer Ordnung $r \geq 2$ nicht mehr die ganze Halbebene $\mathbb{C}_-$ umfassen, in der obigen Tabelle ist der (halbe) Öffnungswinkel δ_r des größten in S_r gelegenen Sektors angegeben. Der Winkel ist relevant, da sich $z = h\lambda$ bei Änderung von h auf einem Strahl bewegt. Für $r > 6$ sind die Verfahren unbrauchbar, da dann nicht einmal mehr die negative reelle Achse vollständig in S enthalten ist. Auch für $r = 6$ ist der Winkel $\delta_6 = 17.8^o$ schon sehr klein (vgl. Diagramm).



Die bisherigen Aussagen bezogen sich auf äquidistante Gitter. Allerdings können BDF-Verfahren auch in der Nordsieckform

$$\sum_{k=1}^{r} \frac{1}{k} \nabla^k y_m = h f(t_m, y_m) \tag{2.5.6}$$

(mit festen Koeffizienten) dargestellt werden, die Ansatzpunkt für Schrittweiten- und Ordnungssteuerung ist. Dabei bezeichnet ∇^k die Rückwärtsdifferenzen, die durch $\nabla y_m = y_m - y_{m-1}$, $\nabla^k y_m = \nabla^{k-1} y_m - \nabla^{k-1} y_{m-1}$ definiert sind. Dabei approximiert $\nabla^k y_m \cong h^k y^{(k)}(t_m)$ und man kann diese Werte mit Vorfaktoren $(h'/h)^k$ an eine geänderte Schrittweite h' anpassen. Durch eine unterschiedliche Anzahl von Summanden kann die Ordnung geändert werden. So kann eine Anlaufrechnung (für die Startwerte $y_1, \ldots, y_{r-1}$) entfallen, wenn man zuerst mit kleinen Schrittweiten das implizite Eulerverfahren benutzt und dann die Ordnung schrittweise erhöht, solange dies sinnvoll ist (Indikatoren?!). Eine effiziente und robuste Implementierung ist bei Mehrschrittverfahren aber sehr schwierig.

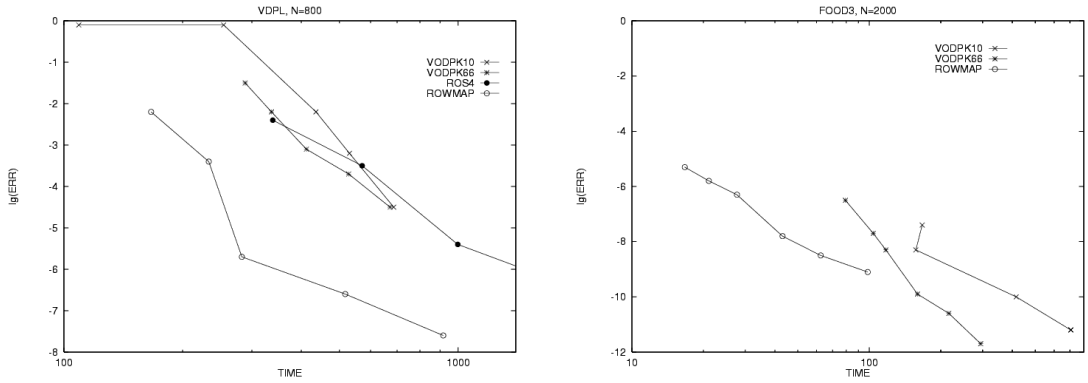
Frei verfügbare Software mit impliziten Mehrschrittverfahren wurde v.a. von Hindmarsh e.a. geschrieben unter den Namen EPISODE (1977), LSODE (1980), VODE. Die Idee für den Einsatz von Krylov-Iterationsverfahren stammt von Gear und Saad (1983) und wurde von Brown/Byrne/Hindmarsh in dem Programm VODPK umgesetzt. Mit sehr viel zusätzlichen Funktionalitäten wird eine C++ -Implementierung unter dem Namen *Sundials* angeboten. Das folgende Beispiel vergleicht den Einsatz der Mehr- und Einschrittverfahren.

Beispiel Die folgenden Diagrammen aus der ROWMAP-Arbeit von 1997 zeigen die Effizienz der Codes ROS4, ROWMAP und VODPK bei zwei bekannten gewöhnlichen Dgl-Systemen, die

durch Diffusion zu parabolischen Problemen werden. Die Probleme sind die vanderPol-Gleichung

$$(u_1)_t = \frac{1}{20}\Delta u_1 + u_2,\ (u_2)_t = \Delta u_2 + 100(1-u_1^2)u_2 - u_1$$

in 2 Orts-Dimensionen und ein Räuber-Beute-Modell FOOD3 in 3 Orts-Dimensionen.



3 Allgemeine Lineare Methoden

In der Form von Allgemeinen Linearen Methoden (*general linear methods*, GLM) können unter anderem Ein- und Mehrschrittverfahren einheitlich beschrieben werden. Da diese Form sehr abstrakt ist, weil z.B. die Bedeutung der verwendeten Variablen offen bleibt, werden hier vor allem Darstellungen und Techniken behandelt, und nur wenige konkrete Ergebnisse.

3.1 Definition und Beispiele

Ein- und Mehrschrittverfahren für das AWProblem (1.0.1) können durch Angabe ihrer Verfahrensfunktion $\Phi_h(t,\cdots)$ beschrieben werden. Der wichtigste Unterschied zwischen beiden Verfahrensklassen besteht darin, dass bei Mehrschrittverfahren Φ_h mehr als ein y-Argument besitzt und Einschrittverfahren (nach außen unsichtbare) interne Stufen verwenden. Mit Schrittweite $h>0$ lautet ein Zeitschritt jeweils

$$\begin{array}{lll} y_m = & \Phi_h(t_{m-1}, y_{m-1}) & \text{Einschritt-Verfahren} \\ y_m = & \Phi_h(t_{m-1}, y_{m-1}, \ldots, y_{m-r}) & \text{Mehrschritt-Verfahren} \end{array} \tag{3.1.1}$$

Für eine vollständige Formulierung müssen aber noch die internen Stufen berücksichtigt werden. Zunächst werden die Mehrschrittverfahren als Einschrittverfahren umgeschrieben. Dazu läßt man ein Fenster der Länge r über die Folge $(y_m)_{m\geq 0}$ gleiten durch die Definition

$$y^{[m]} := \begin{pmatrix} y_{m-r+1} \\ \vdots \\ y_m \end{pmatrix} \in \mathbb{R}^{rn}, \tag{3.1.2}$$

d.h. $y_i^{[m]} = y_{m-r+i}$. Damit wird aus (3.1.1b) die "einstufige" Rekursion

$$y^{[m]} = \bar{\Phi}_h(y^{[m-1]}) = \begin{pmatrix} y_{m-r+1} \\ \vdots \\ y_{m-1} \\ \Phi_h(t_{m-1}, y_{m-1}, \ldots, y_{m-r}) \end{pmatrix}.$$

In dieser Schreibweise stellt sich der Unterschied zu Einschrittverfahren so dar, dass statt eines Wertes y_m jeweils $r \geq 1$ Werte $y^{[m]}$ mitgeführt werden. Bei der Bedeutung der Komponenten sollte man sich aber nicht auf (3.1.2) festlegen, es kann z.B. sinnvoller sein, als Teilvektoren $y_i^{[m]}$ Taylorkoeffizienten der Näherung mitzuführen ("Nordsieck-Form", s.u.). Zusätzlich können jetzt (wie bei RK-Verfahren) noch s *interne* Stufen

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_s \end{pmatrix} \in \mathbb{R}^{sn}, \quad \text{mit} \quad F(Y) = \begin{pmatrix} f(Y_1) \\ \vdots \\ f(Y_s) \end{pmatrix} \in \mathbb{R}^{sn} \tag{3.1.3}$$

berücksichtigt werden. Da sie "privat", also von außen nicht sichtbar sind, werden die Stufen Y nicht zusätzlich mit dem Schrittindex m versehen. Zur kompakten Beschreibung wird das von den Runge-Kutta-Verfahren bekannte Butcher-Tableau um zwei zusätzliche Matrizen erweitert,

$$\left(\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right) \in \mathbb{R}^{(s+r)\times(s+r)}. \tag{3.1.4}$$

Dazu gehört die Verfahrensvorschrift

$$\begin{aligned} Y_i &= h\sum_{j=1}^{s} a_{ij} f(Y_j) + \sum_{j=1}^{r} u_{ij} y_j^{[m-1]}, \quad i = 1, \ldots, s, \\ y_i^{[m]} &= h\sum_{j=1}^{s} b_{ij} f(Y_j) + \sum_{j=1}^{r} v_{ij} y_j^{[m-1]}, \quad i = 1, \ldots, r. \end{aligned} \tag{3.1.5}$$

Das Verfahren heißt *explizit*, wenn A eine strikt untere Dreieckmatrix ist, ansonsten *implizit*. In der zweiten Zeile treten keine neuen Funktionsauswertungen auf, die Werte $f(Y_j) = k_j$ entsprechen bei Runge-Kutta-Verfahren den Stufenwerten. Mit den Block-Vektoren aus (3.1.2,3.1.3) ergibt sich aber auch die kompakte Form

$$\begin{aligned} Y &= hAF(Y) + Uy^{[m-1]} \\ y^{[m]} &= hBF(Y) + Vy^{[m-1]} \end{aligned} \qquad \text{für } n = 1.$$

Auch für größere Raumdimensionen $n > 1$ ist die Schreibweise (3.1.5) bequem, aber nicht korrekt schon alleine wegen der unpassenden Matrixdimensionen. An dieser Stelle ist folgende Matrixverknüpfung hilfreich (oft auch Kronecker-Produkt genannt).

Definition 3.1.1 *Zu zwei Matrizen $A = (a_{ij}) \in \mathbb{R}^{p\times s}$, $B \in \mathbb{R}^{m\times n}$ wird das* direkte Produkt *$A \otimes B \in \mathbb{R}^{(pm)\times(sn)}$ definiert durch*

$$(A \otimes B)_{im+k,jn+l} := a_{ij} b_{kl}, \quad \begin{cases} 1 \leq i \leq p, & 1 \leq j \leq s, \\ 1 \leq k \leq m, & 1 \leq l \leq n. \end{cases}$$

Das Ergebnis ist die Blockmatrix

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \dots & a_{1s}B \\ \vdots & \vdots & & \vdots \\ a_{p1}B & a_{p2}B & \dots & a_{ps}B \end{pmatrix} \in \mathbb{R}^{(pm)\times(sn)}.$$

Mit $s = n = 1$ gilt diese Definition auch für Vektoren.

Solche Matrizen treten auf, wenn bei Block-Vektoren Y (3.1.3) die Matrix B auf allen Blöcken gleichartig operiert und dann Linearkombinationen aller Blöcke mit den Elementen von $A \in \mathbb{R}^{p\times s}$ durchgeführt werden. Dies entspricht genau der Situation im Allgemeinen Linearen Verfahren (3.1.5) und führt auf deren kompakte Schreibweise

$$\begin{aligned} Y = &\quad h(A \otimes I)F(Y) + (U \otimes I)y^{[m-1]}, \\ y^{[m]} = &\quad h(B \otimes I)F(Y) + (V \otimes I)y^{[m-1]}. \end{aligned} \tag{3.1.6}$$

Diese Gestalt dient zunächst einmal einer einheitlichen Analyse von Verfahren, kann für die Implementierung aber unhandlich sein, wie man an den folgenden Beispielen sehen kann.

Beispiel 3.1.2

(a) Nach der einleitenden Motivation ist bei den Runge-Kutta-Verfahren $r = 1$ und A hat die alte Bedeutung. Weiter ist $B = b^{\mathsf{T}}$, $U = \mathbb{1} \in \mathbb{R}^s$, $V = 1 \in \mathbb{R}$.

(b) Für Mehrschrittverfahren kann man verschiedene Darstellungen (3.1.5) konstruieren. Bei den Adams-Bashforth-Verfahren $y_m = y_{m-1} + h\sum_{j=1}^{r} \beta_j f(y_{m-j})$ treten rechts alte f-Werte auf und man ist versucht diese unter die internen Werte $F(Y)$ einzuordnen. Besser ist es allerdings, diese f-Werte über

$$y^{[m-1]} = (y_{m-1}^{\mathsf{T}}, hf(y_{m-2})^{\mathsf{T}}, \dots, hf(y_{m-r})^{\mathsf{T}})^{\mathsf{T}} \in \mathbb{R}^{rn}$$

weiterzureichen. Die Struktur (3.1.6) kann man durch Duplizierung von Variablen und Gleichungen erhalten. Zunächst ist $Y_1 = y_{m-1} = y_1^{[m-1]}$ einzuführen, damit man die neue Funktionsauswertung in $y_m = Y_2 = y_{m-1} + h\beta_1 f(Y_1) + h\sum_{j=2}^{r} \beta_j f(y_{m-j})$ unterbringt. Mit $s = 2$ bekommt man für das A-B-Verfahren dann das Butcher-Tableau

$$\left(\begin{array}{cc|ccccc} 0 & 0 & 1 & 0 & \dots & & \\ \beta_1 & 0 & 1 & \beta_2 & \dots & \beta_{r-1} & \beta_r \\ \hline \beta_1 & 0 & 1 & \beta_2 & \dots & \beta_{r-1} & \beta_r \\ 1 & 0 & 0 & 0 & \dots & & 0 \\ & & 0 & 1 & \dots & & 0 \\ & & & & \ddots & & \vdots \\ & & & & & 1 & 0 \end{array}\right), \quad s = 2.$$

(c) Interessant ist auch die Betrachtung des Prädiktor-Korrektor-Verfahrens aus einem Adams-Bashforth und -Moulton-Schritt (z.B. $r = 2$):

$$\begin{aligned} \tilde{y}_m &:= y_{m-1} + h\beta_1 f(y_{m-1}) + h\beta_2 f(y_{m-2}) \\ y_m &:= y_{m-1} + h\mu_0 f(\tilde{y}_m) + h\mu_1 f(y_{m-1}) + h\mu_2 f(y_{m-2}), \end{aligned}$$

$\beta^\mathsf{T} = \frac{1}{2}(0, 3, -1)$, $\mu^\mathsf{T} = \frac{1}{12}(5, 8, -1)$. Dies ist eine echte ALMe, da offensichtlich $\tilde{y}_m$ eine rein interne Stufe ist und 2 neue f-Auswertungen pro Schritt anfallen. Analog zum vorherigen Beispiel ist $s = 3$, das Gesamtverfahren schreibt sich in ALM-Form daher

$$\begin{array}{rccccc} Y_1 = & & & & y_1^{[m-1]} & \\ Y_2 = & h\beta_1 f(Y_1) & & & +y_1^{[m-1]} & +\beta_2 y_2^{[m-1]} \\ Y_3 = & h\mu_1 f(Y_1) & +h\mu_0 f(Y_2) & +0 & +y_1^{[m-1]} & +\mu_2 y_2^{[m-1]} \\ \hline y_1^{[m]} = & h\mu_1 f(Y_1) & +h\mu_0 f(Y_2) & +0 & +y_1^{[m-1]} & +\mu_2 y_2^{[m-1]} \\ y_2^{[m]} = & & hf(Y_2) & & & \end{array}$$

Das *direkte Produkt* aus Definition 3.1.1 hat einige schöne Eigenschaften ($\rightarrow$ Übungen), die wichtigste ist die Produktregel (bei passenden Dimensionen)

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD),$$

welche man direkt nachrechnet. Sie hat Konsequenzen für die Eigenwerte von $A \otimes B$ aber auch für die häufig auftretenden Matrizen der Form $A \otimes I + I \otimes B$, da dessen beide Summanden kommutieren:

$$(A \otimes I)(I \otimes B) = A \otimes B = (I \otimes B)(A \otimes I). \tag{3.1.7}$$

Ein solches System tritt etwa auf, wenn eine unbekannte Matrix $X \in \mathbb{R}^{n\times s}$ aus der Sylvester-Gleichung $XA^\mathsf{T} + BX = C$ mit $A \in \mathbb{R}^{s\times s}$, $B \in \mathbb{R}^{n\times n}$, bestimmt werden soll. Durch "Stapeln" der Spalten von X in einem langen Vektor $\vec{x} \in \mathbb{R}^{sn}$ mit $\vec{x}_{in+j} = X_{ij}$ bekommt man ein gewöhnliches Lineares System:

$$XA^\mathsf{T} + BX = C \quad \Longleftrightarrow \quad (A \otimes I + I \otimes B)\vec{x} = \vec{c}.$$

Auch für Matrixnormen gibt es einfache Regeln, etwa

$$\|A \otimes B\|_p = \|A\|_p \|B\|_p, \quad p \in \{1, 2, \infty\}. \tag{3.1.8}$$

Die Fälle $p \in \{1, \infty\}$ sind dabei trivial,

3.2 Stabilität

Bei der einfachen Testgleichung $u' = \lambda u$, $\lambda \in \mathbb{C}$, reduziert sich das Verfahren mit $z := h\lambda$ auf die Gleichungen

$$\left.\begin{aligned} Y &= zAY + Uy^{[m-1]} \\ y^{[m]} &= zBY + Vy^{[m-1]} \end{aligned}\right\} \Longleftrightarrow \quad y^{[m]} = \Big(V + zB(I - zA)^{-1}U\Big)y^{[m-1]}$$

durch Elimination der internen Variablen Y. Führt man das Verfahren aus über viele Schritte, müssen daher die Potenzen des auftretenden Matrix-Vorfaktors gleichmäßig beschränkt bleiben.

Definition 3.2.1 *Für das Verfahren (3.1.5) heißt*

$$M(z) := V + zB(I - zA)^{-1}U \in \mathbb{C}^{r\times r}, \quad z \in \mathbb{C},$$

die Stabilitätsmatrix. Das Verfahren heißt an der Stelle $z \in \mathbb{C}$ (absolut) stabil, wenn

$$\varrho(M(z)) \le 1, \text{ und Eigenwerte mit Betrag 1 einfach sind.} \tag{3.2.1}$$

Die elementarsten Anforderungen an die ALM stellt das triviale AWP $u' = 0$, $u_0 = 1$. Hier sollte die Lösung $u \equiv 1$ durch die Verfahrensvorschrift $Y = Uy^{[m-1]}$, $y^{[m]} = Vy^{[m-1]}$ stabil reproduziert werden. In diesem Zusammenhang ist also $Y = \mathbb{1}$ zu fordern, während die Bedeutung des Vektors $y^{[m]}$ von der Verfahrens-Konstruktion abhängt. Dazu verwendet man folgende Begriffe

Definition 3.2.2 *Die ALM (3.1.5) heißt* präkonsistent*, wenn ein $w \in \mathbb{R}^s$ existiert mit*

$$Uw = \mathbb{1}, \quad Vw = w.$$

Das Verfahren heißt nullstabil*, wenn $\varrho(V) \le 1$ ist mit (3.2.1).*

Da bei einer präkonsistenten ALM w ein (Rechts-) Eigenvektor von V zum Eigenwert eins ist, ist $\varrho(V) < 1$ ausgeschlossen. Nullstabilität erzwingt umgekehrt, dass w einziger Eigen- (und Haupt-) Vektor zum EW 1 ist. Für weitergehende Konsistenzuntersuchungen muß man die Bedeutung der Variablen in $y^{[m]}$ kennen, um diese mit (Ableitungen) der Lösung vergleichen zu können. Dies ist im abstrakten Rahmen hier nicht möglich.

Die Unbestimmtheit in der Formulierung wird auch klar, wenn man Transformationen des Verfahrens (3.1.6) betrachtet. Bei den Auswertungen $F(Y)$ müssen die Y_i Funktionswerte der Lösungen sein, hier ist keine Umformulierung möglich. Dagegen ist die Bedeutung der mitgeführten Werte $y^{[m]}$ offen, und mit einer regulären $r \times r$-Matrix W könnte man auch Linearkombinationen, also die Vektoren $w^{[m]} = (W \otimes I)y^{[m]}$ weitergeben. Hier sei an die Multiplikationsregel (3.1.7) des direkten Produkts erinnert. Man bekommt damit das Verfahren

$$\begin{aligned} Y = &\quad h(A \otimes I)F(Y) + ((UW^{-1}) \otimes I)w^{[m-1]}, \\ w^{[m]} = (W \otimes I)y^{[m]} = &\quad h((WB) \otimes I)F(Y) + ((WVW^{-1}) \otimes I)w^{[m-1]}. \end{aligned}$$

mit dem Butchertableau

$$\left(\begin{array}{c|c} A & UW^{-1} \\ \hline WB & WVW^{-1} \end{array}\right) = \begin{pmatrix} I & 0 \\ 0 & W \end{pmatrix} \left(\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right) \begin{pmatrix} I & 0 \\ 0 & W^{-1} \end{pmatrix}. \tag{3.2.2}$$

Die Methoden zu den beiden Tableaus in (3.2.2) sind offenbar äquivalent.

4 Geometrische Integrationsverfahren

4.1 Invarianten

Ein grundlegendes Prinzip in der Physik ist die Energieerhaltung. Bei Differentialgleichungen bedeutet das oft, dass bestimmte Funktionale der Lösung unverändert bleiben, $\eta(u(t)) \equiv \eta(u(0))$. Bei der numerischen Lösung von AWPen sind Fehler zwar unvermeidlich, für spezielle, wichtige Funktionale η existieren aber Verfahren, bei denen η auch für die numerische Approximation unverändert bleibt und dadurch gerade bei langen Zeitintervallen $[t_a, t_e)$ unphysikalische Situationen vermeidet. Es werden jetzt nur noch autonome Probleme

$$u' = f(u), \quad u(0) = u_0 \in \mathbb{R}^n, \tag{4.1.1}$$

($t_a = 0$ oBdA) behandelt. Hier liefert die Zeitverschiebung $u(t - t^*)$ einer Lösung $u(t)$ wieder eine (mit anderem Anfangswert). Daher verbindet man mit der Dgl auch den Begriff eines *Flusses* $u_0 \mapsto u(t; u_0) = \phi^t(u_0)$ (vgl. Numerik-2B) der einen Punkt $u_0 \in \mathbb{R}^n$ "mitreißt" und zur Stelle $u(t; u_0)$ transportiert. Insbesondere gilt auch $\phi^t(\phi^s(u_0)) = \phi^{t+s}(u_0)$. Das einfachste Beispiel zur linearen, homogenen Dgl $u' = Au$ mit konstanter Matrix A ist die lineare Abbildung $u_0 \mapsto \phi^t(u_0) = e^{tA}u_0$. Eine mögliche Invarianz eines solchen Flusses ist die Flächenerhaltung, ein Anfangsgebiet wird durch den Fluß zwar verzerrt, die Fläche ändert sich aber nicht ($\rightarrow$ "geometrische" Verfahren).

Definition 4.1.1 *Eine nicht-konstante Funktion $\eta : \mathbb{R}^n \to \mathbb{R}$ wird* erstes Integral *der Dgl (4.1.1) genannt, wenn gilt*

$$\eta'(v)f(v) = grad\,\eta(v)f(v) = 0 \quad \forall v \in \mathbb{R}^n.$$

Für Lösungen des AWPs (4.1.1), für das eine Invariante η existiert, folgt direkt aus der Kettenregel

$$\frac{d}{dt}\eta(u(t)) = \eta'(u(t))u'(t) = \eta(u(t))f(u(t)) \equiv 0 \,\forall t,$$

daher ist $\eta(u(t)) \equiv \eta(u_0)$ konstant, die Bahn $u(t)$ bewegt sich auf der *Mannigfaltigkeit* $\{y \in \mathbb{R}^n : \eta(y) = 0\}$. Der Begriff "erstes Integral" kommt daher, dass man im $\mathbb{R}^2$ mit der Kenntnis von η die Lösungen implizit bestimmt hat.

Beispiel 4.1.2 (Lotka-Volterra-Gleichungen) Ein einfaches Modell für eine Population aus Räubern $v \geq 0$ (Füchse, Haie) und Beutetieren $w \geq 0$ (Hasen, Fische) ist

$$\begin{array}{rl|l} v' = & \alpha vw - \beta v & \textit{Geburtenrate proport. zur Beute, konstante Sterberate} \\ w' = & \gamma w - \delta vw & \textit{Sterberate proport. zu Räubern, konstante Vermehrung} \end{array}$$

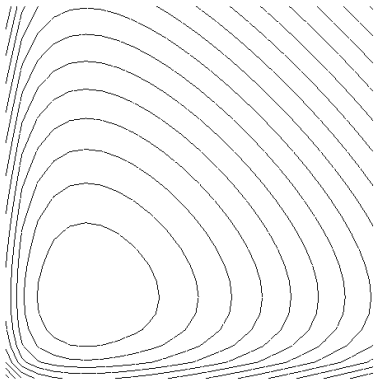
Eine Lösung der Form $v(w)$ bekommt man durch Division der beiden Gleichungen,

$$\frac{v'}{w'} = \frac{dv}{dw} = \frac{v(\alpha w - \beta)}{(\gamma - \delta v)w} \quad \Longleftrightarrow \quad \frac{\gamma - \delta v}{v}\,dv = \frac{\alpha w - \beta}{w}\,dw$$

Die rechte Version zeigt die Dgl mit getrennten Variabeln, Lösungen ergeben sich daher implizit aus der Gleichung

$$\eta(v,w) = \gamma \ln v - \delta v + \beta \ln w - \alpha w = const.$$

Lösungen $(v(t), w(t))$ sind zeitperiodisch, bilden im Phasenraum $\mathbb{R}^2_+$ geschlossene Bahnen und bewegen sich längs der Höhenlinien $\eta(v(t), w(t)) = const$ des *ersten Integrals* η, die rechts für $\alpha = \ldots = \gamma = 1$ gezeigt werden. ■



Beispiel 4.1.3 (Mathematisches Pendel) Die Dgl $w'' + \alpha^2 w = 0$, $\alpha > 0$, ist mit $v = w'$ äquivalent zum System

$$v' = -\alpha^2 w, \quad w' = v \quad \Longleftrightarrow \quad \begin{pmatrix} v' \\ w' \end{pmatrix} = \begin{pmatrix} 0 & -\alpha^2 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} v \\ w \end{pmatrix}.$$

Wie im letzten Beispiel bekommt man

$$0 = v\,dv + \alpha^2 w\,dw \;\Rightarrow\; \eta(u,v) = \frac{1}{2}(v^2 + \alpha^2 w^2) = const,$$

was die bekannten Lösungen $(v, w) = (\alpha\cos(\alpha t + \beta), \sin(\alpha t + \beta))$ offensichtlich erfüllen. Jetzt soll η aber auch für Ergebnisse numerischer Verfahren überprüft werden. Beim expliziten Eulerverfahren (2.1.3) bekommt man

$$y_m = y_{m-1} + hf(y_{m-1}) = y_{m-1} + h \begin{pmatrix} 0 & -\alpha^2 \\ 1 & 0 \end{pmatrix} y_{m-1} = \begin{pmatrix} 1 & -h\alpha^2 \\ h & 1 \end{pmatrix}^m u_0.$$

Bezeichnet man (der Einfachheit halber) $y_m = (v_m, w_m)^\mathsf{T}$ bekommt man durch direktes Nachrechnen

$$\begin{aligned} \eta(y_m) &= (v_{m-1} - h\alpha^2 w_{m-1})^2 + \alpha^2(hv_{m-1} + w_{m-1}^2) = (1 + h^2\alpha^2)(v_{m-1}^2 + \alpha^2 w_{m-1}) \\ &= (1 + h^2\alpha^2)\eta(y_{m-1}). \end{aligned}$$

Für diese Näherungen ist η keinesfalls invariant, der Wert wächst über alle Grenzen mit $\eta(y_m) = (1 + h^2\alpha^2)^m \eta(u_0) \to \infty\,(m \to \infty)$. Beim impliziten Eulerverfahren (2.1.1) ist dagegen

$$y_m = y_{m-1} + hf(y_m) \;\Longleftrightarrow\; \begin{pmatrix} 1 & h\alpha^2 \\ -h & 1 \end{pmatrix} y_m = y_{m-1}$$

und es gilt analog

$$(1 + h^2\alpha^2)\eta(y_m) = \eta(y_{m-1}) \Rightarrow \quad \eta(y_m) = \frac{1}{(1 + h^2\alpha^2)^m}\,\eta(u_0) \to 0\ (m \to \infty),$$

das System "verliert Energie". Zuletzt wird die einstufige symmetrische Mittelpunktregel (2.3.2) (Gauß-IRK-Verfahren) betrachtet, im linearen Beispiel lautet sie

$$\begin{pmatrix} 1 & \frac{h}{2}\alpha^2 \\ -\frac{h}{2} & 1 \end{pmatrix} y_m = \begin{pmatrix} 1 & -\frac{h}{2}\alpha^2 \\ \frac{h}{2} & 1 \end{pmatrix} y_{m-1} \;\Rightarrow\quad y_m = \frac{1}{1 + h^2\alpha^2/4} \begin{pmatrix} 1 & -\frac{h}{2}\alpha^2 \\ \frac{h}{2} & 1 \end{pmatrix}^2 y_{m-1}.$$

Man kann die erste Version als einen halben expliziten, gefolgt von einem halben impliziten Eulerschritt interpretieren. Daher heben sich die vorher beobachteten Effekte auf, es gilt

$$\eta(y_m) = \frac{1}{(1+h^2\alpha^2/4)^2}(1+h^2\alpha^2/4)^2\,\eta(y_{m-1}) \equiv \eta(y_{m-1}),$$

dieses Verfahren bewahrt (wegen seiner Symmetrie) die Invarianz des ersten Integrals. ∎

Bevor numerische Verfahren genauer untersucht werden, werden wichtige Strukturen diskutiert.

Eine wichtige Klasse von Problemen mit Invarianten sind *Hamilton-Systeme* in Räumen gerader Dimension $n = 2d \in 2\mathbb{N}$. Tatsächlich sind es Systeme von 2 Variablen $u^\mathsf{T} = (v^\mathsf{T}, w^\mathsf{T})^\mathsf{T}$, die aus einer stetig diffbaren skalaren *Hamilton-Funktion* $H(v,w) = H(u)$ hergeleitet werden. Mit der unitären und schiefsymmetrischen Matrix

$$J := \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \otimes I_d = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}, \quad J^\mathsf{T} = J^{-1} = -J,$$

gibt es folgende Schreibweisen des Hamilton-Systems

$$\left.\begin{array}{rl} v' = & -\frac{\partial H}{\partial w}^\mathsf{T}(v,w) \\ w' = & \frac{\partial H}{\partial v}^\mathsf{T}(v,w) \end{array}\right\} \quad \Longleftrightarrow \quad u' = J^\mathsf{T}\nabla H(u), \quad u = \begin{pmatrix} v \\ w \end{pmatrix}. \tag{4.1.2}$$

Da der Gradient einer skalaren Funktion ein Zeilenvektor ist, wird die Transposition durch die Bezeichnung $\nabla H := (\operatorname{grad} H)^\mathsf{T} = (H')^\mathsf{T}$ umgangen. Außerdem vertauscht die Matrix J die beiden Teile von $\operatorname{grad} H = (H_v, H_w)$ (mit einer Vorzeichenänderung), daher sind die beiden Formen in (4.1.2) identisch.

Satz 4.1.4 *Zu einer stetig diffbaren Funktion $H \in C^1(\mathbb{R}^{2d})$ sei $u(t)^\mathsf{T} = (v(t)^\mathsf{T}, w(t)^\mathsf{T}) \in C^1(\mathbb{R})$ Lösungskurve des Problems (4.1.2). Dann ist H ein erstes Integral von (4.1.2), es gilt*

$$H(u(t)) \equiv H(u(0)) \quad \forall t \in \mathbb{R}.$$

Beweis Wie oben gilt nach der Kettenregel

$$\frac{d}{dt}H(v(t),w(t)) = H_v v'(t) + H_w w'(t) \overset{(4.1.2)}{=} -H_v H_w^\mathsf{T} + H_w H_v^\mathsf{T} = 0.$$

Da mit H und u auch $H(u)$ stetig diffbar ist, folgt die Behauptung aus dem Hauptsatz. ∎

Umgekehrt gilt, dass ein Standardproblem $u' = f(u)$ mit $u \in C^1$ genau dann ein Hamiltonsystem ist, wenn $Jf'(u)$ überall symmetrisch ist.

Das Beispiel 4.1.3 betraf schon ein Hamilton-System, es folgen weitere Hamilton-Funktionen für verschiedene Anwendungen, dabei bedeutet durchgängig $v(t) = w'(t)$:

1. Physikalisches Pendel $v' = w'' = -\alpha^2 \sin w$,

   $$H(v,w) = \frac{1}{2}v^2 - \alpha^2 \cos w.$$

   H beschreibt die Gesamtenergie aus kinetischer Energie $K(v) = \frac{1}{2}v^2$ und potenzieller (Lage-) Energie $P(w) = -\alpha^2 \cos w$.

2. Mehrkörperproblem der Himmelsmechanik, n Massen m_i mit Orten und Geschwindigkeiten $w_i, v_i \in \mathbb{R}^3$:
$$H(v,w) = \frac{1}{2}\sum_{i=1}^{n}\frac{\|v_i\|_2^2}{m_i} - G\sum_{i=2}^{n}\sum_{j=1}^{i-1}\frac{m_i m_j}{\|w_i - w_j\|_2}.$$
Auch hier sind die beiden Bestandteile kinetische und potenzielle Energie.

3. In der Molekulardynamik treten ähnliche Energiefunktionale auf, $w_i, v_i \in \mathbb{R}^3$ sind Atom-Positionen bzw. Geschwindigkeiten, und mit Potentialfunktionen V_{ij} ist die Hamilton-Funktion
$$H(v,w) = \frac{1}{2}\sum_{i=1}^{n}\frac{\|v_i\|_2^2}{m_i} - \sum_{i=2}^{n}\sum_{j=1}^{i-1}V_{ij}(\|w_i - w_j\|_2),$$

In allen 3 Beispielen ist $H(v,w) = K(v) + P(w)$ die Summe der kinetischen Energie K und der potenziellen $P(w)$, die Hamilton-Funktion ist *separabel*. $K(v)$ ist eine quadratische Funktion und die Ableitung $H_v^\mathsf{T} = \nabla K = M^{-1}v$ daher durch eine konstante Matrix gegeben, die Inverse der Massenmatrix $M = \mathrm{diag}(m_i)$. Hier reduziert sich das Hamilton-System (4.1.2) auf

$$v' = -\nabla P(w), \quad w' = M^{-1}v, \tag{4.1.3}$$

das durch Differentiation auf ein kleineres ("konservatives", ohne w') System zweiter Ordnung reduziert werden kann,
$$w'' = -M^{-1}\nabla P(w).$$

4.2 Symplektische Verfahren

Für numerische Verfahren gilt nur ausnahmsweise, dass das erste Integral exakt bewahrt wird. Es gibt aber eine einfachere, geometrische Erhaltungseigenschaft von Hamiltonsystemen, die sich auf Verfahren übertragen läßt, nämlich die Invarianz des Flächeninhalts 2-dimensionaler Flächen. Im $\mathbb{R}^2$ ($d = 1$) ist die Fläche eines Parallelograms, das von zwei Vektoren $u^{(i)} = (v_i, w_i)^\mathsf{T}$, $i = 1, 2$, aufgespannt wird, gerade

$$\det\begin{pmatrix} v_1 & v_2 \\ w_1 & w_2 \end{pmatrix} = v_1 w_2 - v_2 w_1 = {u^{(1)}}^\mathsf{T}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}u^{(2)}.$$

Abbildungen, die den 2D-Flächeninhalt im $\mathbb{R}^{2d}$ erhalten, nennt man *symplektisch*.

Definition 4.2.1 *Eine lineare Abbildung $A : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ heißt* symplektisch*, wenn gilt*

$$A^\mathsf{T} J A = J.$$

Eine auf einer offenen Menge $U \subseteq \mathbb{R}^{2d}$ differenzierbare Abbildung $g : U \to \mathbb{R}^{2d}$ heißt symplektisch*, wenn ihre Ableitung g' überall symplektisch ist,*

$$g'(u)^\mathsf{T} J g'(u) = J \quad \forall u \in U. \tag{4.2.1}$$

Der oben definierte Fluß $y \mapsto u(t;y) = \phi^t(y)$ eines Hamiltonsystems ist symplektisch.

Satz 4.2.2 *Sei $U \subseteq \mathbb{R}^{2d}$ offen und $H \in C^2(U)$. Dann ist für festes $t \geq 0$ die Abbildung $g := \phi^t$ symplektisch.*

Beweis Beim Hamiltonsystem ist die rechte Seite $f = J^\mathsf{T}\nabla H$. Die Ableitung des Flusses

$$\frac{\partial \phi^t}{\partial y} =: \Psi(t)$$

nach dem Anfangswert y erfüllt (vgl. Numerik 2B) die Variationsgleichung

$$\Psi' = f'(\phi^t(y))\Psi = J^\mathsf{T}\nabla^2 H(\phi^t(y))\Psi.$$

Dabei ist die Hessematrix $\nabla^2 H(\phi^t(y)) =: U(t,y)$ symmetrisch. Mit $\phi^0(y) = y$ ist $\Psi(0) = I$, also gilt $\Psi(0)^\mathsf{T} J\Psi(0) = J$. Dieser Ausdruck bleibt unverändert, mit $JJ^\mathsf{T} = I$, $J^2 = -I$ sieht man, dass

$$\frac{d}{dt}(\Psi^\mathsf{T} J\Psi) \quad = \quad (\Psi')^\mathsf{T} J\Psi + \Psi J\Psi' = \Psi^\mathsf{T} U J^2\Psi + \Psi^\mathsf{T} JJ^\mathsf{T} U\Psi = \Psi^\mathsf{T}(-U+U)\Psi = 0.$$

Daher gilt auch für $g' = \Psi(t)$ Eigenschaft (4.2.1). ∎

Die Eigenschaft (4.2.1) gilt tatsächlich auch für bestimmte numerische Verfahren, wobei man die Abbildung $g:\ u_0 \mapsto y_m$ (schrittweise) zu untersuchen hat.

Satz 4.2.3 *Die implizite Mittelpunktregel (2.3.2) ist symplektisch.*

Beweis Die Verfahrensfunktion $\Phi(y)$ mit $y_m = \Phi(y_{m-1})$ ist hier implizit definiert durch

$$0 = \Phi(y) - y - hf(\frac{1}{2}(\Phi(y)+y)).$$

Aus der Kettenregel folgt (mit $z := \frac{1}{2}(\Phi(y)+y)$) durch ableiten nach y, dass gilt

$$0 = \Phi' - I - \frac{h}{2}f'(z)(\Phi'+I) \iff (I - \frac{h}{2}f'(z))\Phi' = (I + \frac{h}{2}f'(z)).$$

Wichtig ist hier, dass $f'(z) = J^\mathsf{T}\nabla H(z)$ beide Male an der selben Stelle ausgewertet wird, die geklammerten Matrizen sind vertauschbar und daher gilt wieder mit $\nabla^2 H(z) =: U$, dass

$$(\Phi')^\mathsf{T} J\Phi' = J \iff (I + \frac{h}{2}UJ)J(I + \frac{h}{2}J^\mathsf{T}U) = (I - \frac{h}{2}UJ)J(I - \frac{h}{2}J^\mathsf{T}U)$$

$$\iff J - \frac{h}{2}U + \frac{h}{2}U - \frac{h^2}{4}UJ^\mathsf{T}U = J + \frac{h}{2}U - \frac{h}{2}U - \frac{h^2}{4}UJ^\mathsf{T}U$$

ist wegen $JJ^\mathsf{T} = I = -J^2$. ∎

Damit kennt man ein symplektisches Verfahren der Ordnung 2. Man kann sogar zeigen, dass ein Runge-Kutta-Verfahren (2.1.2) genau dann symplektisch ist, wenn für seine Koeffizienten gilt

$$b_i a_{ij} + b_j a_{ji} - b_i b_j = 0\ \forall i,j, \quad \text{d.h. } BA + A^\mathsf{T} B - bb^\mathsf{T} = 0.$$

Da diese Bedingung gerade für die Gauß-RK-Verfahren erfüllt ist (vgl. Satz 2.3.1), sind diese symplektisch bei Ordnung $2s$. Es sei aber an ihren hohen Rechenaufwand erinnert.

Partitionierte Verfahren Hamiltonsysteme weisen durch die Aufteilung in zwei Teile für v, w eine ausgeprägte Struktur auf und man kann daher für die Teilsysteme unterschiedliche Verfahren betrachten. Beim *Eulerverfahren* etwa läßt sich die in Beispiel 4.1.3 beobachtete Asymmetrie dadurch ausbügeln, dass man jeweils einen expliziten und einen impliziten Schritt macht. Dies führt auf die beiden *symplektischen* Eulerverfahren

$$(a)\left\{\begin{array}{rcl} v_m = & v_{m-1} - h\nabla_w H(v_m, w_{m-1}) \\ w_m = & w_{m-1} + h\nabla_v H(v_m, w_{m-1}) \end{array}\right. \quad (b)\left\{\begin{array}{rcl} v_m = & v_{m-1} - h\nabla_w H(v_{m-1}, w_m) \\ w_m = & w_{m-1} + h\nabla_v H(v_{m-1}, w_m) \end{array}\right. \tag{4.2.2}$$

Bei einer separablen Hamiltonfunktionen $H(v, w) = K(v) + P(w)$ sind beide Verfahren sogar explizit, bei Variante (b) ist dabei zunächst w_m zu berechnen.

Satz 4.2.4 *Die symplektischen Eulerverfahren (4.2.2) sind symplektisch.*

Beweis (für separierte H.-Funktion) Analog zum Vorgehen bei der Mittelpunktregel in Satz 4.2.3 berechnet man für die Ableitung der Verfahrensfunktion $\Phi' = \partial(v_m, w_m)/\partial(v_{m-1}, w_{m-1})$ von Verfahren (a) die Beziehung

$$\begin{pmatrix} I + hH_{wv}^{\mathsf{T}} & 0 \\ -hH_{vv} & I \end{pmatrix} \Phi' = \begin{pmatrix} I & -hH_{ww} \\ 0 & I + hH_{wv} \end{pmatrix}.$$

Bei separierter Hamilton-Funktion verschwinden die gemischten Ableitungen und Φ' kann direkt angegeben werden. Damit gilt tatsächlich (4.2.1), denn

$$(\Phi')^{\mathsf{T}} J \Phi' = \underbrace{\begin{pmatrix} -hH_{vv} & I \\ -I + h^2 H_{ww} H_{vv} & -hH_{ww} \end{pmatrix}}_{(\Phi')^{\mathsf{T}} J} \underbrace{\begin{pmatrix} I & -hH_{ww} \\ hH_{vv} & I - h^2 H_{vv} H_{ww} \end{pmatrix}}_{\Phi'} = J.$$

Der Beweis für (4.2.2(b)) geht analog. ∎

Eine weitere *partitionierte* (Runge-Kutta-) Methode ist

$$\begin{aligned} v_m = & \underbrace{v_{m-1} - \frac{h}{2}\nabla P(w_{m-1})}\, - \tfrac{h}{2}\nabla P(w_m), \\ w_m = & \; w_{m-1} + h\nabla K(\underbrace{v_{m-1} - \frac{h}{2}\nabla P(w_{m-1})}). \end{aligned}$$

Man sieht aber sofort, dass es günstiger ist, das markierte Argument von ∇K als Hilfswert zu benutzen und erhält das Störmer-Verlet-Verfahren:

$$\left.\begin{array}{rcl} v_{m-1/2} = & v_{m-1} - \frac{h}{2}\nabla P(w_{m-1}) \\ w_m = & w_{m-1} + h\nabla K(v_{m-1/2}) \\ v_m = & v_{m-1/2} - \frac{h}{2}\nabla P(w_m) \end{array}\right\} \iff \left\{\begin{array}{rcl} w_m = & w_{m-1} + h\nabla K(v_{m-1/2}) \\ v_{m+1/2} = & v_{m-1/2} - h\nabla P(w_m) \end{array}\right.$$

Die rechte Version zeigt den einheitlichen symmetrischen Aufbau als *explizites* Verfahren auf *versetzten* Gittern. Bei kinetischer Energie $K(v) = \frac{1}{2}v^\mathsf{T}M^{-1}v$, (4.1.3), können die v-Näherungen eliminiert werden. Dies führt auf das Zweischrittverfahren

$$w_{m+1} - 2w_m + w_{m+1} = -h^2 M^{-1}\nabla P(w_m).$$

Das Störmer-Verlet-Verfahren ist symplektisch, da es sich aus den beiden (zueinander adjungierten) symplektischen Eulerverfahren (4.2.2) zusammensetzt (o.Beweis). Es wurde von C. Störmer für astronomische Berechnungen und von L. Verlet für die Molekulardynamik entwcikelt und wird in Anwendungen sehr häufig benutzt.

Die Bedeutung der Symplektizität für Energieerhaltung in der Hamilton-Funktion erschließt sich erst durch eine "Rückwärtsanalyse" (sehr aufwändig). Es läßt sich damit zeigen, dass symplektische Verfahren ein AWP mit gestörter Hamilton-Fkt $\tilde{H}(u) = H(u) + O(h^p)$ exakt lösen, wenn p die Ordnung des Verfahrens ist. Dadurch wächst der Fehler in H nicht wie bei anderen Verfahren stark mit der Zeit, sondern bleibt beschränkt für exponentiell lange Zeiten $mh \leq e^{h_0/h}$.

5 Parallele Verfahren für Anfangswertprobleme

In aller Regel wird ein hoher Aufwand bei AWPen, welcher erst eine Parallelisierung interessant macht, durch eine große Dimension n des Problems verursacht. In dieser Situation hat man folgende Parallelisierungsansätze:

1. Parallelisierung im System:
   Viele Dgln großer Dimension besitzen einen regelmäßigen Aufbau (vgl. Beisp. 2.1), den man dazu nutzen kann, die Auswertung der rechten Seite f und/oder die Lösung von Stufengleichungen auf viele ($p \leq n$) Prozessoren zu verteilen. Dieser Ansatz ist aber problemabhängig und vom Anwender zu implementieren.

2. Parallelisierung in der Methode:
   Die AWP-Methode besitzt eine strukturelle Parallelität mit "black-box"-Auswertung der rechten Seite f. Der Ansatz hat (zunächst) einen geringen Parallelisierungsgrad, welcher aber durch zusätzliche Parallelisierung beim f-Aufruf (Punkt 1) vergrößert werden kann.

Die beiden Ansätze 1. und 2. schließen sich also gegenseitig nicht aus, sondern können evtl. ergänzend eingesetzt werden. Im folgenden wird der problemunabhängige Ansatz 2 diskutiert.

Um tatsächlich eine Leistungssteigerung bei Parallelrechnung (auf p Prozessoren) zu erreichen, ist eine Zerlegung des Gesamtaufwands in p voneinander unabhängige Aufgaben erforderlich, welche auch annähernd gleichen Rechenbedarf besitzen ("Lastverteilung"). Diese Ziele können beim Ansatz 2. durch ein geeignetes Design erreicht werden, erfordern beim Ansatz 1. dagegen viel Kreativität beim Anwender (aber mit größerem p).

5.1 Parallelansätze mit Standardverfahren

Bei der Konstruktion von Parallelverfahren kann man sich an den bekannten Verfahrensklassen orientieren:

- Einschritt-Verfahren: die teilweise Entkopplung von Stufen bei RK-Verfahren führt nicht zum Ziel, da dann elementare Differentiale für die Ordnungsbedingungen fehlen. Bei IRK-Verfahren kann man die Lösung des $(sn) \times (sn)$ teilweise entkoppeln, oder Iterationsverfahren mit entkoppelten Systemen betrachten.
- Mehrschrittverfahren: auch hier wurde Parallelisierung meist über entkoppelte Iterationen versucht
- Allgemeine Lineare Verfahren: Hier gibt es Unterklassen mit eingebauter Parallelität, z.B. die DIMSIM-Klasse von Butcher.

Im Folgenden wird eine neue Verfahrensklasse behandelt, die bei Testrechnungen sehr konkurrenzfähig mit etablierten Standardverfahren ist. Sie gehört zwar auch in die Klasse der ALMen, läßt sich aber einfacher direkt formulieren.

5.2 Peer - Zweischritt - Methoden

Das namensgebende Kennzeichen dieser Methoden ist, dass sie zwar wie Runge-Kutta-Verfahren pro Zeitschritt $s \geq 1$ Stufenlösungen verwenden, dass diese aber nicht wie dort nur Hilfsarbeit zur Konstruktion einer besonders guten Schlußlösung y_m leisten ("Master-Slave-Struktur"), sondern alle gleich gute Approximationen an die Lösung darstellen (Team-, Partner-, Peer-Struktur). Für nicht-steife Probleme berechnet eine *explizite parallele Peer-Methode* zu paarweise verschiedenen Knoten c_i, $i = 1, \ldots, s$, jeweils Stufenlösungen $Y_{mi} \cong u(t_{mi})$, $t_{mi} = t_m + h_m c_i$, durch

$$Y_{mi} = \sum_{j=1}^{s} b_{ij} Y_{m-1,j} + h_m \sum_{j=1}^{s} a_{ij} f(t_{m-1,j}, Y_{m-1,j}),\ i = 1, \ldots, s, \tag{5.2.1}$$

$$\iff \quad Y_m = (B \otimes I) Y_{m-1} + h_m (A \otimes I) F(Y_{m-1}). \tag{5.2.2}$$

Die Matrizen $A = (a_{ij})$, $B = (b_{ij}) \in \mathbb{R}^{s \times s}$ enthalten die Koeffizienten des Verfahrens. In der zweiten Variante wurden die Stufen zu Blockvektoren der Form (3.1.3) zusammengefaßt wie bei den Allgemeinen Linearen Methoden (3.1.5). Da in der Peer-Methode auf der rechten Seite nur Werte Y_{m-1} aus dem vorhergehenden Zeitschritt auftreten, können alle s Stufen offensichtlich auf s Prozessoren *parallel* abgearbeitet werden. Durch die Hinzunahme aktueller Stufen $h_m \sum_{j=1}^{i-1} q_{ij} f(Y_{mj})$ bekäme man sequentielle Zeischritt-Verfahren. Die Peer-Verfahren (5.2.2) sind auch ALMn, in der Schreibweise mit Butcher-Tableau haben sie die Form

$$\left(\begin{array}{c} Y_m \\ \hline y^{[m]} \end{array}\right) = \left(\begin{array}{c} Y_m \\ \hline hF_m \end{array}\right) = \left(\begin{array}{c|cc} Q & A & B \\ \hline Q & A & B \\ I & 0 & 0 \end{array}\right) \left(\begin{array}{c} hF(Y_m) \\ \hline Y_{m-1} \\ hF_{m-1} \end{array}\right),$$

(parallel für $Q = 0$) wenn man Stufenlösungen und Funktionsauswertungen nochmal in $y^{[m]}$ unterbringt, die ALM-Schreibweise ist aber sehr redundant.

Für steife Probleme kann man die implizite Peer-Zweischritt-Methode

$$Y_{mi} - h_m\gamma_i f(Y_{mi}) = \sum_{j=1}^{s} b_{ij} Y_{m-1,j}, \quad i = 1, \ldots, n, \tag{5.2.3}$$

als Ausgangspunkt betrachten. Ihr (redundantes) Butcher-Tableau zu $y^{[m]} = Y_m$ ist

$$\left(\begin{array}{c|c} G & B \\ \hline G & B \end{array}\right).$$

Da $G = \mathrm{diag}(\gamma_i)$ Diagonalmatrix ist, ist das Verfahren *parallel*. Man kann wie (2.4.1) aber nur einen Newtonschritt mit einer Näherungsmatrix $T \cong f'$ betrachtet und einem Prädiktor $\tilde{Y}_{mi} := \frac{1}{\gamma_i}\sum_j a_{ij} Y_{m-1,j}$. Durch Herausziehen der Summe aus $f(\tilde{Y}) \to \sum_j a_{ij} f(Y_{m-1,j})$ bekommt man folgende linear-implizite Variante, $i = 1, \ldots, s$:

$$\begin{aligned} (I - h_m\gamma_i T_m)Y_{mi} &= \sum_{j=1}^{s}(b_{ij}I - h_m a_{ij} T_m)Y_{m-1,j} + h_m \sum_{j=1}^{s} a_{ij} f(t_{m-1,j}, Y_{m-1,j}) \\ &= \sum_{j=1}^{s} b_{ij} Y_{m-1,j} + h_m \sum_{j=1}^{s} a_{ij}(f(t_{m-1,j}, Y_{m-1,j}) - T_m Y_{m-1,j}), \end{aligned} \tag{5.2.4}$$

bzw. mit Gesamt-Vektoren

$$(I - h_m G \otimes T_m)Y_m = (B \otimes I - h_m A \otimes T_m)Y_{m-1} + h_m(A \otimes I)F(Y_{m-1}). \tag{5.2.5}$$

Offensichtlich geht (5.2.5) in das explizite Verfahren (5.2.1) über für $T_m = 0$. Es sei schon hier darauf hingewiesen, dass wegen der Zweischritt-Struktur die Koeffizienten von der Schrittweitenfolge abhängen müssen, daher wären alle Matrizen mit einem Schrittindex zu versehen: A_m, B_m, G_m.

Stabilität: Bei der Testgleichung $y' = \lambda y$, $\mathrm{Re}\,\lambda \le 0$, und mit der Wahl $T = \mu$ gehört zum Verfahren (5.2.5) die Rekursion

$$(I - h_m\mu G)Y_m = (B_m + h_m(\lambda - \mu)A_m)Y_{m-1}.$$

Daraus bekommt man in den beiden interessanten Fällen $\mu = 0$ (explizite Verfahren) und $\mu = \lambda$ zu $z := h_m\lambda$ die Stabilitätsmatrizen

$$\begin{aligned} M_m(z) &= B_m + zA_m, && \text{für } \mu = 0, \\ M_m(z) &= (I - zG_m)^{-1}B_m, && \text{für } \mu = \lambda. \end{aligned} \tag{5.2.6}$$

Die Stabilitätsmatrix im zweiten Fall entspricht der des impliziten Verfahrens (5.2.3). Insbesondere sind im Ansatz (5.2.4,5.2.5) die Parameter so abgestimmt, dass mit $M(\infty) = 0$ eine maximale Dämpfung steifer Fehleranteile bewirkt wird. In beiden Fällen gilt in (5.2.6) $M_m(0) = B_m$,

sodass für die Nullstabilität die Forderung $\varrho(B_m) \leq 1$ aus Definition 3.2.2 erforderlich ist. Präkonsistenz führt dabei auf den Eigenwert 1 zum Eigenvektor $\mathbb{1} = B\mathbb{1}$. Leider ist die Bedingung $\varrho(B_m) \leq 1$ nur notwendig, tatsächlich sind wenige, handhabbare Kriterien bekannt, die die Nullstabilität garantieren. Diese erfordert nämlich Folgendes.

Definition 5.2.1 *Die Peer-Methode (5.2.4) heißt* nullstabil*, wenn alle Produkte*

$$B_m B_{m-1} \cdots B_{k+1} B_k, \quad m \geq k \in \mathbb{N},$$

der Matrixfamilie $\{B_k\}_{k\geq 1}$ gleichmäßig beschränkt sind.

Zum Erreichen der Nullstabilität wird die Gestalt der Koeffizientenmatrizen wie im folgenden Satz eingeschränkt. Die Voraussetzungen sind trivialerweise erfüllt bei konstantem $B_m \equiv B$ mit einer strikt stabilen Matrix B. Etwas allgemeiner ist folgendes Kriterium.

Satz 5.2.2 *Für die Matrixfamilie $\{B_m\}_{m\geq 1}$, gebe es eine gemeinsame Basistransformation $U \in \mathbb{R}^{s\times s}$ und eine Konstante $0 \leq q < 1$, so dass*

$$U^{-1} B_m U =: \tilde{B}_m = \begin{pmatrix} 1 & * & \cdots & * \\ & \lambda_{m,2} & \cdots & * \\ & & \ddots & \vdots \\ & & & \lambda_{m,s} \end{pmatrix} \quad \forall m \in \mathbb{N}$$

gilt mit $\|\tilde{B}_m\|_\infty \leq K$, $|\lambda_{m,i}| \leq q$, $i = 2, \ldots, s$, $m \in \mathbb{N}$. Dann sind Matrixprodukte gleichmäßig beschränkt, es existiert eine Konstante K' mit

$$\|B_m B_{m-1} \cdots B_{k+1} B_k\|_\infty \leq K' \quad \forall m \geq k \in \mathbb{N}.$$

Beweis Zunächst gilt (oBdA $k = 1$) $\|B_m B_{m-1} \cdots B_2 B_1\| \leq cond(U) \|\tilde{B}_m \tilde{B}_{m-1} \cdots \tilde{B}_2 \tilde{B}_1\|$. Nach Voraussetzung ist jede der Matrizen elementweise beschränkt durch

$$|\tilde{B}_m| \leq \bar{B} = \begin{pmatrix} 1 & b & \cdots & b \\ & q & \cdots & b \\ & & \ddots & \vdots \\ & & & q \end{pmatrix} = \begin{pmatrix} 1 & b\mathbb{1}^\mathsf{T} \\ & R \end{pmatrix},$$

wobei $0 \leq b \leq K$, $R \geq 0$, $\varrho(R) = q < 1$ ist. Daher existiert $\mathbb{1}^\mathsf{T} \sum_{j=0}^\infty R^j \mathbb{1} = \mathbb{1}^\mathsf{T} (I - R)^{-1} \mathbb{1} =: K_1$. Da die Zeilensummennorm betragsmonoton ist, gilt $\|\tilde{B}_m \tilde{B}_{m-1} \cdots \tilde{B}_2 \tilde{B}_1\| \leq \|\bar{B}^m\| = \|\bar{B}^m \mathbb{1}\|_\infty$. Man verifiziert leicht, dass für $m \in \mathbb{N}$ gilt

$$\bar{B}^m = \begin{pmatrix} 1 & b\mathbb{1}^\mathsf{T} \\ & R \end{pmatrix}^m \leq \begin{pmatrix} 1 & b\mathbb{1}^\mathsf{T}(I-R)^{-1} \\ 0 & R^m \end{pmatrix} \leq \begin{pmatrix} 1 & b\mathbb{1}^\mathsf{T}(I-R)^{-1} \\ 0 & \sum_{j=0}^\infty R^j \end{pmatrix} = \begin{pmatrix} 1 & b\mathbb{1}^\mathsf{T}(I-R)^{-1} \\ 0 & (I-R)^{-1} \end{pmatrix}.$$

Die Norm ist daher beschränkt durch $\|\bar{B}^{m+1}\mathbb{1}\|_\infty \leq \max\{1 + KK_1, K_1\}$ und die Behauptung folgt mit $K' = cond(U) \max\{1 + KK_1, K_1\}$. ∎

Da die Konsistenzuntersuchungen für die expliziten und die impliziten Verfahren unterschiedliche Konsequenzen haben, werden beide jetzt getrennt behandelt.

5.3 Explizite Peer-Methoden

Wegen der Zweischritt-Struktur sind Konsistenzbetrachtungen der Peer-Verfahren einfach, man betrachtet das Residuum beim Einsetzen der Lösung u unter Berücksichtigung von $u'(t) = f(t, u(t))$ für $i = 1, \ldots, s$:

$$h_m \Delta_{mi} := u(t_m + h_m c_i) - \sum_{j=1}^{s} b_{ij} u(t_{m-1} + h_{m-1} c_j) - h_m \sum_{j=1}^{s} a_{ij} u'(t_{m-1} + h_{m-1} c_j). \quad (5.3.1)$$

Offensichtlich treten hier zwei Schrittweiten auf und bei Taylorentwicklung um den Punkt t_{m-1} tritt das Schrittweitenverhältnis

$$\sigma_m := \frac{h_m}{h_{m-1}} \quad (5.3.2)$$

dieser beiden auf. Die Abhängigkeit von diesem Verhältnis σ wird eine große Rolle spielen. Bei Entwicklung um t_{m-1} ist $t_m + h_m c_i = t_{m-1} + h_{m-1}(1 + \sigma_m c_i)$ und man erhält

$$h_m \Delta_{mi} \;=\; \sum_{k=0}^{q-1} \frac{h_{m-1}^k}{k!} u^{(k)}(t_{m-1}) \Big((1 + \sigma c_i)^k - \sum_{j=1}^{s} b_{ij} c_j^k - k\sigma \sum_{j=1}^{s} a_{ij} c_j^{k-1} \Big) + O(h^q)$$

Bei der letzten Summe wurde eine Indexverschiebung durchgeführt, sie fehlt für $k = 0$. An dieser Entwicklung liest man folgende Ordnungsbedingungen ab, wobei qs die Anzahl der Bedingungen ist,

$$\mathbf{AB(q)}: \quad (1 + \sigma c_i)^k - \sum_{j=1}^{s} b_{ij} c_j^k - k\sigma \sum_{j=1}^{s} a_{ij} c_j^{k-1} = 0, \; k = 0, \ldots, q-1. \quad (5.3.3)$$

In der folgenden Konvergenzaussage bezeichnet $H = \max\{h_m : t_a \le t_m < t_e\}$.

Satz 5.3.1 *Das Peer-Verfahren sei* nullstabil *nach Definition 5.2.1 und* konsistent *mit Ordnung q durch Erfüllen von $\mathbf{AB(q+1)}$. Außerdem seien genaue Startwerte $Y_{0i} - u(t_{0i}) = O(H^q)$ gegeben. Rechte Seite f und Lösung u seien genügend glatt, außerdem gelte $\|A_m\| \le K \, \forall m$. Dann konvergieren alle Peer-Näherungen mit Ordnung q gegen die Lösung*

$$Y_{mi} - u(t_{mi}) = O(H^q), \quad i = 1, \ldots, s, \; t_0 \le t_m \le t_e.$$

Bem. Die einheitliche Konvergenzaussage für alle Stufenlösungen war Grund für die Bezeichnung "Peer". Im Gegensatz zu Einschrittverfahren kann hier durch einfache Interpolation für jedes t eine $O(h^q)$-Approximation berechnet werden.

Der Beweis geht ähnlich wie bei Einschrittverfahren, allerdings darf man nicht zu früh zu Schranken übergehen, sondern man muß zuerst Produkte $B_m \cdots B_k$ sammeln, und dann abschätzen, vgl. [WJSP].

Interessant ist natürlich die erreichbare Ordnung der Verfahren. Man sieht schnell, dass Ordnung s über $\mathbf{AB(s+1)}$ erreichbar ist. Denn $\mathbf{AB(1)}$ entspricht gerade der Präkonsistenz $B_m \mathbb{1} = \mathbb{1}$, die restlichen Bedingungen lassen sich zu Matrix-Gleichungen zusammenfassen. Dabei

treten insbesondere die Vandermonde-Matrix V und die Pascalmatrix P der Binomialkoeffizienten auf:

$$V = \left(c_i^{j-1}\right)_{i,j=1}^{s} = \begin{pmatrix} 1 & c_1 & \dots & c_1^{s-1} \\ \vdots & & & \vdots \\ 1 & c_s & \dots & c_s^{s-1} \end{pmatrix},\ P = \left(\binom{j-1}{i-1}\right)_{i,j=1}^{s} = \begin{pmatrix} 1 & 1 & 1 & \cdots \\ & 1 & 2 & \dots \\ & & 1 & \dots \\ & & & \ddots \end{pmatrix}. \quad (5.3.4)$$

Außerdem werden noch Diagonalmatrizen eingeführt:

$$S := \operatorname{diag}(1, \sigma, \dots, \sigma^{s-1}), \quad D = \operatorname{diag}(1, 2, \dots, s), \quad C = \operatorname{diag}(c_1, \dots, c_s).$$

Diese treten auf bei $(1+\sigma c_i)^{j-1} = \sum_{k=1}^{j-1} c_i^{k-1}\sigma^{k-1}\binom{j-1}{k-1} = (VSP)_{ij}$. Dann ist $AB(s+1)$ äquivalent mit folgender Matrixgleichung. Die Zuordnung zur elementweisen Formulierung wird dabei angedeutet.

$$\begin{array}{cccc} \sum_{j=1}^s b_{ij} = 1 \quad \text{sowie} & (1+\sigma c_i)^k & = \sum_{j=1}^s b_{ij} c_j^k & +\sigma \sum_{j=1}^s a_{ij} c_j^{k-1} k, \quad k = 1, \dots, s \\ B\mathbb{1} = \mathbb{1} & (I + \sigma C)VSP & = BCV & +\sigma AVD. \end{array}$$

Hierdurch sind also die beiden "freien Matrizen" A und B gekoppelt und es kann jede durch die andere ausgedrückt werden. Gravierende Einschränkungen gibt es aber nur für die Matrix B über die Nullstabilität, vgl. Satz 5.2.2, und daher ist es sinnvoll, B möglichst einfach festzulegen, etwa durch

$$B_m \equiv B \text{ fest}, \ B\mathbb{1} = \mathbb{1}, \ \text{mit Eigenwerten } \lambda_1 = 1, \ \lambda_2 = \dots = \lambda_s = 0. \quad (5.3.5)$$

Diese Festlegung läßt noch viele Freiheitsgrade, sie bedeutet nur, dass $B - \mathbb{1}v^\mathsf{T}$ ($v \in \mathbb{R}^n$) nilpotent ist. Damit ergibt sich die andere Matrix (in Abhängigkeit von σ_m!) als

$$\sigma_m A_m = \Big((I + \sigma_m C)VS_mP - BCV\Big)D^{-1}V^{-1}. \quad (5.3.6)$$

Bei Schrittweitensteuerung ist die Matrix A_m also in jedem Schritt (aus vorberechneten Anteilen wie $PD^{-1}V^{-1}$ etc) neu zu berechnen. Der Aufwand $\mathrm{O}(s^3)$ ist bei großen Problemen geringfügig gegenüber dem Gesamtaufwand

Fehlerschätzung: Die Inverse V^{-1} löst das Interpolationsproblem $\sum_{j=1}^s c_i^{j-1} a_j \stackrel{!}{=} y_i$, $i = 1, \dots, s$: $a = V^{-1}y$. Daher ist $a_s = y[c_1, \dots, c_s]$ die höchste dividierte Differenz zu den Daten y. und

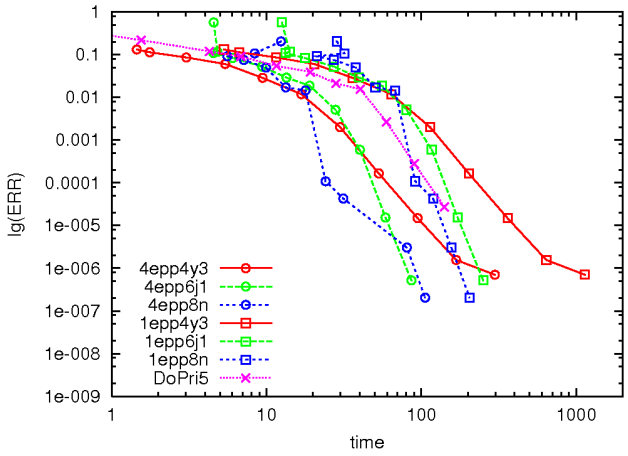
$$e_s^\mathsf{T} D^{-1}V^{-1}F(Y_{m-1}) \cong \frac{1}{s} e_s^\mathsf{T} V^{-1}(u'(t_{m-1,i}))_{i=1}^s \cong \frac{1}{s!} u^{(s)}(t_m)$$

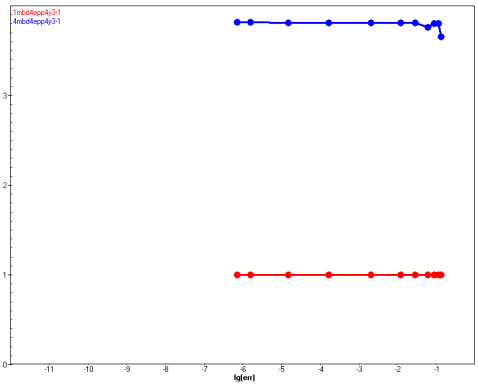
enthält eine Approximation für die s-te Ableitung. Diese kann man, wie bei eingebetteten Verfahren üblich (Schätzung durch Verfahren niederer Ordnung), als Ersatz für die Ableitung $u^{(s+1)}$ im lokalen Fehler Δ_m verwenden. Damit kann also eine Fehlerschätzung zur Schrittweitensteuerung implementiert werden.

Verfahrenssuche: Mit (5.3.6) bleiben mit den Knoten c_i und der Matrix B nach (5.3.5) viele freie Parameter zur Verfahrenskonstruktion. Weitere Zielkriterien für ein Verfahren sind der

Stabilitätsbereich bzw. das Stabilitätsintervall $\{x \in \mathbb{R} : \varrho(M(x)) \leq 1\}$ und die Fehlerkonstante in $\Delta_m \doteq K h_m^s u^{(s+1)}(t_m)$. Außerdem sollten die Koeffizienten B und A nicht "zu große" Normen annehmen, da Rundungsfehler mit $\|B\|, h\|A\|$ multipliziert werden.

Beispiel 5.3.2 Die folgenden Graphiken zeigen die Effizienzdiagramme (Genauigkeit/Rechenzeit bei $TOL = 10^{-2}..10^{-12}$) von Testrechnungen mit 4 Prozessoren bei einem Mehrkörperproblem im $\mathbb{R}^3$ mit 400 Massen. Da die Berechnung der vielen Gravitationskräfte sehr aufwändig ist, dominiert die f-Auswertung den Aufwand und die drei expliziten Peer-Verfahren (mit $s = 4, 6, 8$) erreichen einen fast optimalen Speed-up. Im linken Diagramm sind die Peer-Verfahren auf 4 Prozessoren mit Kreisen markiert, die auf einem mit Quadraten, das (nicht-parallele) Vergleichsverfahren Dopri5 mit Kreuzen. Das rechte Diagramm zeigt den Speed-up des 4-stufigen Verfahrens EPP4y abhängig von der Genauigkeit (ca 3.8).





5.4 Peer-Zweischritt-W-Methoden

Um für die Implementierung bei steifen AWPen möglichst viel Freiheit zu haben, ist das Ziel bei der Konstruktion der linear-impliziten Verfahren (5.2.5) eine von der Wahl der Matrix T unabhängige Konvergenz wie bei W-Methoden, vgl. §2.4. Dazu gruppiert man Terme mit und ohne T getrennt und läßt in der Taylorentwicklung beide bis zu einer bestimmten Ordnung verschwinden. Dies führt natürlich zu einer größeren Anzahl an Ordnungsbedingungen. Der Defekt der exakten Lösung u in (5.2.4) ist hier

$$\begin{aligned} h_m \Delta_m &= u(t_{mi}) - \sum_{j=1}^{s} b_{ij} u(t_{m-1,j}) - h_m \sum_{j=1}^{s} a_{ij} u'(t_{m-1,j}) \\ &\quad - h_m T_m (\gamma_i u(t_{mi}) - \sum_{j=1}^{s} a_{ij} u(t_{m-1,j})). \end{aligned}$$

Dabei entspricht der Anteil in der ersten Zeile dem Konsistenzfehler der expliziten Verfahren. Neu hinzu kommen jetzt aber die Bedingungen, die in der mit T_m multiplizierte Klammer einen

entsprechenden Fehler erzeugen. Wieder durch Taylorentwicklung um t_{m-1} erhält man

$$\mathbf{\Gamma(q)}: \quad \gamma_i(1+\sigma c_i)^k - \sum_{j=1}^{s} a_{ij} c_j^k = 0, \quad k = 0, \ldots, q-1. \tag{5.4.1}$$

Hier bedeutet q wieder die Anzahl der Bedingungen, zu beachten ist aber, dass $\mathbf{\Gamma(q)}$ wegen des Vorfaktors $h_m T_m$ einen anderen Beitrag zu Δ_m liefert.

Lemma 5.4.1 *Bei genügend glatter Lösung gilt mit* $\mathbf{AB(q+1)}$ *und* $\mathbf{\Gamma(q)}$, $q \geq 1$, *dass der lokale Fehler* $\Delta_m = O(h_{m-1}^q)$ *ist.*

Für vergleichbare Ordnungen hat man also fast die doppelte Anzahl an Ordnungsbedingungen als im expliziten Fall, daher kann auch nur Ordnung $s-1$ problemlos erreicht werden. Für $q = s$ faßt man $\mathbf{\Gamma(s)}$ wieder zu einer Matrixgleichung zusammen, $GVSP = AV$, welche die Koeffizientenmatrix A abhängig von σ_m und G festlegt:

$$A_m = G\Theta_m, \quad \Theta_m := VS_mPV^{-1}. \tag{5.4.2}$$

Die Matrix Θ_m bewirkt die Polynomextrapolation von den Stützstellen $t_{m-1,i}$ ins nächste Intervall auf t_{mi}. Mit $\mathbf{AB(s)}$ ist dann aber auch die Matrix B i.w. festgelegt und hängt ebenfalls von σ_m ab. Zur Darstellung der Bedingungen

$$\mathbf{AB(s)}: \quad \underbrace{(1+\sigma c_i)^{k-1}}_{(VSP)_{ij}} = \sum_{j=1}^{s} b_{ij} \underbrace{c_j^{k-1}}_{v_{ij}} - \sigma \sum_{j=1}^{s} a_{ij} c_j^{k-2}(k-1), \quad k = 1, \ldots, s,$$

wird noch die Schiebematrix $F_0 := (\delta_{i-1,j})$ benötigt. Damit lautet $\mathbf{AB(s)}$:

$$B_mV = VS_mP - \sigma_m A_m VDF_0^{\mathsf{T}} \overset{\Gamma(s)}{=} VSP + \sigma_m GVS_mPDF_0^{\mathsf{T}}.$$

Man rechnet leicht nach, dass die Matrix DF_0^{T} mit der Pascalmatrix kommutiert, schöner ist aber die Folgerung aus der Eigenschaft

$$P = e^{DF_0^{\mathsf{T}}},$$

einer Polynom-Taylorentwicklung. Damit kann Θ_m nach rechts herausgezogen werden und es gilt

$$\mathbf{AB(s)}, \mathbf{\Gamma(s)} \;\Rightarrow\; A_m = G\Theta_m,\ B_m = (I-GE)\Theta_m,\ E := VDF_0^{\mathsf{T}}V^{-1}. \tag{5.4.3}$$

Leider hängt also jetzt die Matrix B_m vom Schrittindex ab und die Nullstabilität nach Definition 5.2.1 wird ein nichttriviales Problem. Es läßt sich aber eine Verfahrensklasse angeben, für die das Kriterium von Satz 5.2.2 anwendbar ist. Dabei berücksichtigt man, dass bei fast allen Matrizen in (5.4.3) eine Ähnlichkeitstransformation mit der Vandermonde-Matrix V auftritt. Macht man diese rückgängig durch Betrachtung der Matrizen $\tilde{B} = V^{-1}BV$, etc., bekommt man statt (5.4.3) die Darstellung

$$\tilde{B}_m = (I - \tilde{G}DF_0^{\mathsf{T}})S_mP, \quad \tilde{A}_m = \tilde{G}S_mP. \tag{5.4.4}$$

Da P eine obere Dreieckmatrix ist, ist $\tilde{B}_m$ ebenfalls eine, sofern $\tilde{G}$ eine Hessenbergmatrix ist, da F_0^{T} die Spalten um eins nach rechts verschiebt. Für das weitere ist Folgendes hilfreich.

Lemma 5.4.2 *Es sei* $\varphi(t) = \prod_{i=1}^{s}(t - c_i) = \sum_{j=0}^{s} \phi_j t^j$. *Dann gilt*

$$V^{-1}CV = F := \begin{pmatrix} 0 & 0 & \dots & -\phi_0 \\ 1 & 0 & \dots & -\phi_1 \\ & \ddots & & \vdots \\ & & 1 & -\phi_{s-1} \end{pmatrix}.$$

Beweis Die Multiplikation CV bewirkt eine Linksverschiebung der Spalten von V. Die letzte Spalte kann wegen $c_i^s = -\sum_{j=0}^{s-1} \phi_j c_i^j$ aus V rekonstruiert werden. Daher gilt $CV = VF$ und somit die Behauptung. ∎

Da die Frobeniusmatrix F Hessenbergform hat, liefert der Ansatz

$$\gamma_i = g_0 + g_1 c_i,\ i = 1, \dots, s \iff G = g_0 I + g_1 C \iff \tilde{G} = g_0 I + g_1 F \tag{5.4.5}$$

mit Parametern g_0, g_1 offensichtlich eine Hessenbergmatrix $\tilde{G}$. Bei $\tilde{B}$ in (5.4.4) entsteht so wegen $FDF_0^{\mathsf{T}} = \mathrm{diag}(0, 1, \dots, s-1) =: \hat{D}$ tatsächlich die obere Dreieckmatrix

$$\tilde{B} = (I - (g_0 I + g_1 F) D F_0^{\mathsf{T}}) S_m P = (I - g_1 \hat{D} - g_0 D F_0^{\mathsf{T}}) S_m P$$

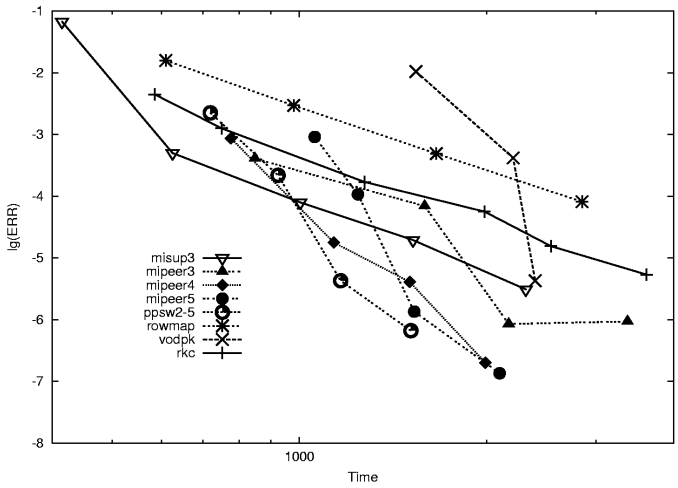
mit der Hauptdiagonalen $\tilde{b}_{ii} = \sigma^{i-1}(1 - (i-1)g_1)$, $i = 1, \dots, s$. Für nicht zu großes $\sigma < \sigma_{\sup}$ können durch Wahl von $g_1 > 0$ die Voraussetzungen von Satz 5.2.2 mit $U = V$ erfüllt werden. Die folgende Tabelle zeigt diese Werte

s	3	4	5	6	7	8
σ_{sup}	2.414	1.678	1.444	1.330	1.262	1.218
g_1	0.5858	0.4039	0.3075	0.2481	0.2078	0.1788

Weitere Details zu dieser Verfahrensklasse finden sich in [SWP05].

Zur parallelen Implementierung: wie bei W-Methoden könnte man versuchen, bei einheitlichem γ eine LR-Zerlegung von $(I - h\gamma T_m)$ zu berechnen und danach die s Stufenlösungen durch parallele Auflösung zu bestimmen. Da die LR-Zerlegung aber schlecht parallelisierbar ist, sind vor allem Iterationsverfahren (Krylov-) interessant, wobei dann auch unterschiedliche γ_i unschädlich sind. Im folgenden Beispiel wurden daher die s Stufensysteme in (5.2.4) unabhängig voneinander parallel mit Krylovverfahren gelöst.

Beispiel 5.4.3 Im Effizienzdiagramm werden Ergebnisse zu einem schwierigen Strahlungs-Diffusionsproblem (parabolisches System) auf einem Gitter mit 100×100 Punkten gezeigt. Es vergleicht u.a. drei parallele linear-implizite Peer-Verfahren *mipeer** nach Satz **??** mit Stufenzahlen $s = 3, 4, 5$ und die in §2 beschriebenen Methoden ROWMAP und VODPK und zeigt die Überlegenheit der Peer-Verfahren. Für scharfe Toleranzen sind allerdings die Ergebnisse eines lateren Peer-Verfahrens *ppsw2-5* noch besser, welches ohne Strukureinschränkungen an B konstruiert wurde [MOL], dessen Nullstabilität aber nur für $\sigma_m \equiv 1$ gesichert ist.



Literatur zu §5:

SW04 B.A. Schmitt, R. Weiner, *Parallel two-step W-methods with peer variables*, SIAM J. Numer. Anal. 42, 265-282.

SWP05 B.A. Schmitt, R. Weiner, H. Podhaisky, *Multi-implicit peer two-step methods for parallel time integration*, BIT 45 (2005), 197-217.

WJSP R. Weiner, S. Jebens, B. Schmitt, H. Podhaisky, *Explicit parallel two-step peer methods*, erscheint in Comp.Math. Applcs.

Index