7 Galerkin-Verfahren

Bisher haben wir zur Lösung von AWA sog. „Differenzenverfahren“ betrachtet. In diesem Rahmen stehen mit den Runge-Kutta-Verfahren, den linearen Mehrschrittformeln sowie den durch lokale Extrapolation gewonnenen Formeln Methoden jeder gewünschten Ordnung zur Verfügung, wobei in jedem Zeitschritt lediglich Funktionsauswertungen der rechten Seite $f(t,x)$ erforderlich sind. Mit den A-stabilen impliziten Runge-Kutta-Schemata, den BDF-Verfahren und der extrapolierten impliziten Euler-Formel lassen sich auch *steife* Probleme behandeln. Die dabei in jedem Zeitschritt auftretenden, schlecht konditionierten, nichtlinearen Gleichungssysteme werden mit Hilfe des Newton-Verfahrens gelöst. Es ist daher wünschenswert, dass sich Monotonie-Eigenschaften der AWA auf entsprechende Eigenschaften der Verfahrensfunktion $F(h;t,x,y)$ übertragen. Dies ist bei allgemeinen impliziten Runge-Kutta-Verfahren aber nicht notwendigerweise der Fall. Bei allen betrachteten Differenzenverfahren kann die Zeitschrittweite auf der Basis von heuristischen Schätzungen des Abschneidefehlers gewählt werden. Dahinter steht als „Rechtfertigung“ die *a priori* Fehlerabschätzung mit einer allerdings unbestimmten Fehlerkontante, welche im schlimmsten Fall exponentiell mit der Zeit wachsen kann. Insbesondere bei *steifen* Problemen mit inhärent großen Lipschitz-Konstanten ist diese Begründung ungenügend. Wünschenswert wäre also eine Schrittweitenkontrolle auf der Basis auswertbarer *a posteriori* Fehlerabschätzungen. Dies führt auf die Frage nach der Konstruktion von Einschrittverfahren beliebig hoher Ordnung, welche

- nur Funktionsauswertungen der rechten Seite erfordern,
- zur Integration „steifer“ Probleme geeignet sind,
- dieselben Monotonie-Eigenschaften wie die gegebene AWA besitzen,
- eine *a posteriori* Fehleranalyse mit auswertbaren Fehlerabschätzungen zulässt,
- eine verlässliche Schrittweitenkontrolle erlaubt.

Wir werden im Folgenden mit den sog. „Galerkin[1]-Verfahren“ einen für AWA noch wenig gebräuchlichen Diskretisierungsansatz betrachten, welcher diesen Anforderungen wenigstens im Prinzip gerecht wird.

7.1 Variationelle Formulierung der Anfangswertaufgaben

Die bisher betrachteten Lösungsmethoden für AWA

$$u' = f(t,u)\,, \quad t \geq t_0\,, \quad u(t_0) = u_0\,, \tag{7.1.1}$$

waren alle *Differenzenverfahren*. Bei diesen wird typischerweise die linke Seite in (7.1.1) durch Differenzenquotienten approximiert und die zugehörige Näherungslösung $U_n \approx$

[1] Boris Grigorievich Galerkin (1871-1945): Russischer Bauingenieur und Mathematiker; Prof. in St. Petersburg; Beiträge zur Struktur-Mechanik, insbesondere zur Plattentheorie.

$u(t_n)$ in diskreten Zeitpunkten t_n bestimmt. Wir betrachten nun einen Ansatz, der eine mehr globale Sichtweise hat, das sog. „Galerkin-Verfahren“. Die AWA wird der Einfachheit halber wieder als (global) Lipschitz-stetig angenommen, und die Notation ist so gehalten, dass auch allgemeine d-dimensionale Systeme von Gleichungen erfasst werden, d. h.: Auftretende Funktionen sind gegebenenfalls vektorwertig, und $(\cdot,\cdot)$ bedeutet dann wieder das zugehörige euklidische Skalarprodukt.

Ausgangspunkt ist eine integrale Formulierung der AWA über einem vorgegebenen Zeitintervall $I = [t_0, t_0 + T]$. Für eine Funktion $u \in C^1(I)^d$ mit $u(t_0) = u_0$ ist (7.1.1) äquivalent zu

$$\int_I (u' - f(t,u), \varphi)\, dt = 0 \quad \forall \varphi \in C(I)^d. \tag{7.1.2}$$

Jede (klassische) Lösung von (7.1.1) erfüllt offensichtlich (7.1.2). Die Umkehrung zeigt man etwa, indem für jeden festen Zeitpunkt $t \in I$ als „Testfunktionen“ approximierende Dirac-Funktionen $\delta_\varepsilon(t;\cdot)$ zum Aufpunkt t eingesetzt werden und der Grenzübergang $\varepsilon \to 0$ durchgeführt wird (Fundamentalsatz der Variationsrechnung):

$$0 = \int_I (u_i' - f_i(s,u))\delta_\varepsilon(t;s)\, ds \quad \to_{\varepsilon\to 0} \quad u_i'(t) - f_i(t,u(t)). \tag{7.1.3}$$

Wir verzichten auf die Beweisdetails, da dies für das Folgende nicht wichtig ist. Wenn Missverständnisse ausgeschlossen sind, wird die explizite Erwähnung der t-Abhängigkeit von Funktionen unter dem Integral weggelassen.

Da die Funktionen φ in (7.1.2) beliebig variieren dürfen, nennt man dies auch eine „variationelle Formulierung“ der AWA (7.1.1). Sie besagt geometrisch ausgedrückt, dass das sog. „Residuum“ der Lösung u,

$$R(u) := u' - f(\cdot, u),$$

bzgl. des Skalarprodukts von $L^2(I)^d$ orthogonal zu allen Testfunktionen $\varphi \in C(I)^d$ ist.

7.2 Das „unstetige“ Galerkin-Verfahren

Ein allgemeines Galerkin-Verfahren zur Approximation von (7.1.1) restringiert die Gleichung (7.1.2) auf geeignete, endlich dimensionale Ansatzräume, ganz analog zum Vorgehen etwa beim CG-Verfahren zur Lösung von allgemeinen linearen Gleichungssystemen. Wir betrachten solche Galerkin-Verfahren mit stückweise polynomialen Ansatzfunktionen. Dazu seien

$$t_0 < t_1 < \ldots < t_N = t_0 + T$$

eine Unterteilungen des Integrationsintervalls I in (hier nach links halboffene) Teilintervalle $I_n = (t_{n-1}, t_n]$. Wir setzen wieder

$$h_n = t_n - t_{n-1}, \quad h = \max_{n=1,\ldots,N} h_n.$$

Bzgl. einer solchen Unterteilung $\mathbf{T}_h = \{I_n,\, n = 1, \ldots, N\}$ wird zunächst der Raum $V(I)$ von stückweise glatten Funktionen definiert durch

$$V(I) \,=\, \big\{v : I \to \mathbb{R}^d :\, v(t_0) \in \mathbb{R}^d,\; v_{|I_n} \in C_c^1(I_n)^d\,,\; n = 1, \ldots, N\big\}\;.$$

Dabei bezeichnet $C_c^1(I_n)$ den Raum der auf dem (halb offenen) Intervall I_n stetig differenzierbaren und stetig zum linken Randpunkt t_{n-1} fortsetzbaren Funktionen. Wir wollen die variationelle Formulierung (7.1.2) der AWA (7.1.1) zu einer äquivalenten auf dem Raum $V(I)$ erweitern, welche dann als Grundlage eines Diskretisierungsansatzes auch mit unstetigen Ansatzfunktionen dienen kann. Für Funktionen $v \in V(I)$ werden die folgenden Bezeichnungen eingeführt:

$$v_n^+ = \lim_{t \downarrow t_n} v\,, \qquad v_n^- = \lim_{t \uparrow t_n} v\,, \qquad [v]_n = v_n^+ - v_n^-\,.$$

Gesucht ist nun eine Funktion $u \in V(I)$ mit den Eigenschaften $u(t_0) = u_0^- = u_0$ und

$$\sum_{n=1}^{N} \Big\{ \int_{I_n} (u' - f(t,u), \varphi)\, dt + ([u]_{n-1}, \varphi_{n-1}^+) \Big\} = 0, \tag{7.2.4}$$

für alle $\varphi \in V(I)$. Man überlegt sich leicht, dass diese Formulierung in der Tat äquivalent zu (7.1.2) bzw. zu (7.1.1) ist:
Beweisargument: Durch Wahl von φ_ε mit $\varphi_{\varepsilon,n}^+ = 1$ und $\varphi_\varepsilon(t) \to 0\,(\varepsilon \to 0)$ für $t \neq t_n$ folgt $[u]_n = 0$ für $n \geq 1$, d. h. die Stetigkeit von u bei t_n, sowie $u_0^+ = u_0$ für $n = 0$. Analog folgt das Bestehen der Gleichung $u' = f(\cdot, u)$ auf I_n. Wegen der Stetigkeit von $f(t,x)$ ergibt sich damit wieder, dass u notwendig auch stetig differenzierbar auf I sein muss und folglich mit der Lösung von (7.1.1) übereinstimmt.

Wir gehen noch einen Schritt weiter und integrieren auch noch die Anfangsbedingung in die variationelle Formulierung. Dazu führen wir die folgende Semi-Linearform ein:

$$A(u;\varphi) := \sum_{n=1}^{N} \Big\{ \int_{I_n} (u' - f(t,u), \varphi)\, dt + ([u]_{n-1}, \varphi_{n-1}^+) \Big\} + (u_0^-, \varphi_0^-),$$

welche bzgl. des zweiten Arguments φ (Argument nach dem Semikolon) *linear* ist. Die AWA (7.1.1) ist dann äquivalent zur Bestimmung eines $u \in V(I)$ mit der Eigenschaft

$$A(u;\varphi) = (u_0, \varphi_0^-) \quad \forall \varphi \in V(I). \tag{7.2.5}$$

Zur Diskretisierung von (7.2.5) führen wir nun Teilräume $S_h^{(r)}(I) \subset V(I)$ $(r \in \mathbb{N}_0)$ von stückweise polynomialen Funktionen ein:

$$S_h^{(r)}(I) \,=\, \big\{\varphi \in V(I) :\, \varphi(t_0) \in \mathbb{R}^d,\; \varphi_{|I_n} \in P_r(I_n)^d\,,\; n = 1, \ldots, N\big\}\;.$$

Dabei bezeichnet $P_r(I_n)$ den Raum der Polynome vom Grad kleiner oder gleich r. Man beachte, dass die Ansatzfunktionen $\varphi \in S_h^{(r)}$ i. Allg. unstetig sind und im Anfangszeitpunkt t_0 einen Wert $\varphi(t_0) \neq \lim_{t \downarrow t_0} \varphi(t)$ annehmen können. Der Fall stetiger Ansatzfunktionen wird unten kurz diskutiert werden. Der Galerkin-Ansatz zur Lösung von (7.1.1) besteht nun darin, dass eine Funktion $U \in S_h^{(r)}(I)$ gesucht wird mit den Eigenschaften

$$A(U;\Phi) = (u_0, \Phi_0^-) \quad \forall \Phi \in S_h^{(r)}(I). \tag{7.2.6}$$

Aus offensichtlichem Grund wird dieses Verfahren „unstetiges Galerkin-Verfahren (mit Ansatzgrad r)“ oder kurz „dG(r)-Verfahren“ genannt. Die Lösbarkeit des (endlich dimensionalen) Problems (7.2.6) wird später diskutiert. Zunächst stellen wir fest, dass wegen der zugelassenen Unstetigkeit der Testfunktionen das *globale* Problem auch als ein sukzessives Zeitschrittverfahren geschrieben werden kann. Durch Wahl einer Testfunktion in der Form $\varphi \equiv 0$ auf allen $I_m \neq I_n$ erhält man aus (7.2.6)

$$\int_{I_n} (U', \varphi)\, dt + (U_{n-1}^+, \varphi_{n-1}^+) = \int_{I_n} (f(t,U), \varphi)\, dt + (U_{n-1}^-, \varphi_{n-1}^+)\,, \tag{7.2.7}$$

für alle $\varphi \in P_r(I_n)$. Dabei spielt U_{n-1}^- die Rolle des Anfangswertes auf dem Intervall I_n. Der Anfangswert für $n = 1$ ist natürlich $U_0^- = u_0$. Man beachte, dass die *diskrete* Lösung U an den Stützstellen t_n nicht stetig zu sein braucht. Da die kontinuierliche Lösung u offensichtlich dieselbe variationelle Gleichung (7.2.6) erfüllt,

$$A(u;\Phi) = (u_0, \Phi_0^+) \quad \forall \Phi \in S_h^{(r)}(I),$$

ergibt sich durch Subtraktion die sog. „Galerkin-Orthogonalität“

$$A(u;\Phi) - A(U;\Phi) = 0 \quad \forall \Phi \in S_h^{(r)}(I). \tag{7.2.8}$$

7.2.1 Beispiele

Das scheinbar so andersartige dG(r)-Verfahren besitzt dennoch eine enge Verwandtschaft mit wohl bekannten Differenzenverfahren. Dazu betrachten wir den Fall $d = 1$.

(1) Fall $r=0$: Wir setzen $U_n := U_n^-$ auf $I_n = (t_{n-1}, t_n]$ und (7.2.7) reduziert sich auf

$$U_n - U_{n-1} = \int_{I_n} f(t, U_n)\, dt\,. \tag{7.2.9}$$

Dies ist eine Variante des impliziten Euler-Schemas, welches man durch Approximation des Integrals auf der rechten Seite mit der Boxregel erhält:

$$U_n = U_{n-1} + \int_{I_n} f(t, U_n)\, dt \approx U_{n-1} + h_n f(t_n, U_n).$$

(2) Fall $r=1$: Wir verwenden für $U(t)$ auf I_n die Lagrange-Darstellung

$$U(t) = h_n^{-1}(t - t_{n-1})U_n^- - h_n^{-1}(t - t_n)U_{n-1}^+\,.$$

Setzt man dies in (7.2.7) ein und testet nacheinander mit den Basispolynomen $\varphi \equiv 1$ und $\varphi = h_n^{-1}(t - t_{n-1})$, so ergibt sich für die Funktionswerte U_{n-1}^+ und U_n^- das System

$$U_n^- - U_{n-1}^- = \int_{I_n} f(t,U)\, dt\,, \quad U_n^- - U_{n-1}^+ = 2h_n^{-1} \int_{I_n} f(t,U)(t - t_{n-1})\, dt\,. \tag{7.2.10}$$

Wenn man die Integrale mit der Trapezregel approximiert,

$$\int_{I_n} f(t,U)\,dt \;\approx\; \tfrac{1}{2}h_n\{f(t_{n-1},U_{n-1}^+)+f(t_n,U_n^-)\},$$
$$2h_n^{-1}\int_{I_n} f(t,U)(t-t_{n-1})\,dt \;\approx\; h_n f(t_n,U_n^-)\,,$$

ergibt sich das *implizite* Runge-Kutta-Verfahren

$$U_n^- = U_{n-1}^- + \tfrac{1}{2}h_n(k_1+k_2)\,,\quad k_1 = f(t_{n-1},U_n^- - h_n k_2),\quad k_2 = f(t_n,U_n^-)\,. \tag{7.2.11}$$

Dieses Differenzenverfahren ist von zweiter Ordnung (Nachrechnen als Übungsaufgabe). Wir werden später aber sehen, dass das exakte dG(1)-Verfahren in den diskreten Zeitpunkten t_n sogar von dritter Ordnung ist.

Beide Differenzenverfahren, (7.2.9) und (7.2.10), sind A-stabil. Sie gehören zur Klasse der sog. „subdiagonalen Padé[2] -Verfahren“ und haben angewandt auf das übliche Modellproblem $u'=\lambda u$ die Verstärkungsfaktoren

$$\begin{aligned} &\text{dG(0)-Verfahren:} && \omega(\lambda h) &&= \frac{1}{1-h\lambda}\,,\\ &\text{dG(1)-Verfahren:} && \omega(\lambda h) &&= \frac{1+\frac{1}{3}h\lambda}{1-\frac{2}{3}h\lambda+\frac{1}{6}h^2\lambda^2}\,. \end{aligned}$$

Das Differenzenschema (7.2.11) hat den Verstärkungsfaktor $\omega(\lambda h)=(1-h\lambda+\frac{1}{2}h^2\lambda^2)^{-1}$ und ist folglich A-stabil.

7.2.2 Lösbarkeit der Galerkin-Gleichungen

Wir untersuchen zunächst die Wohlgestelltheit der Galerkin-Gleichungen im Fall von L-stetigen (nicht-steifen) AWA.

Satz 7.1 (dG-Verfahren für nicht-steife AWA): *Sei wieder L die globale Lipschitz-Konstante der Funktion $f(t,x)$. Für jedes $r\ge 0$ gibt es eine Konstante $\gamma>0$, so dass die Galerkin-Gleichung (7.2.6) unter der Bedingung $h<\gamma/L$ eine eindeutige Lösung $U\in S_h^{(r)}(I)$ besitzt.*

Beweis: Sei U bis zum Zeitpunkt t_{n-1} berechnet. Der Schritt nach t_n erfordert die Bestimmung von $U_{|I_n}$ aus U_{n-1}^-. Die Lösung $U_{|I_n}$ ist bestimmt durch die Fixpunktgleichung

$$\int_{I_n}(U',\varphi)\,dt + (U_{n-1}^+,\varphi_{n-1}^+) = \int_{I_n}(f(t,U),\varphi)\,dt + (U_{n-1}^-,\varphi_{n-1}^+)\,, \tag{7.2.12}$$

[2]Henri Eugène Padé (1785-1836): Französischer Mathematiker; Prof. in Poitiers und Bordeaux; entwickelte die sog. „Padé-Approximation“, rationale Approximationen der Exponentialfunktion.

für alle $\varphi \in P_r(I_n)$. Wir werden zeigen, dass diese eine Fixpunktabbildung $g : P_r(I_n) \to P_r(I_n)$ definiert (zu festem U_{n-1}^-), welche für $h < \gamma/L$ eine Kontraktion ist. Seien dazu $U = g(\widetilde{U})$ und $V = g(\widetilde{V})$ für zwei beliebige $\widetilde{U}, \widetilde{V} \in P_r(I_n)$. Die Differenzen $W := U - V$ und $\widetilde{W} := \widetilde{U} - \widetilde{V}$ genügen dann der Relation

$$\int_{I_n} (W', \varphi)\, dt + (W_{n-1}^+, \varphi_{n-1}^+) \le L \int_{I_n} \|\widetilde{W}\| \|\varphi\|\, dt \qquad \forall \varphi \in P_r(I_n).$$

Wir setzen hier zunächst $\varphi = W$ und erhalten unter Verwendung der Identität $(W', W) = \frac{1}{2}\frac{d}{dt}\|W\|^2$ und anschließender Integration die Beziehung

$$\|W_n^-\|^2 + \|W_{n-1}^+\|^2 \le 2Lh_n \sup_{I_n} \|\widetilde{W}\| \sup_{I_n} \|W\|\,. \tag{7.2.13}$$

Als nächstes setzen wir $\varphi = (t - t_{n-1})W' \in P_r(I_n)$ und erhalten

$$\int_{I_n} \|W'\|^2 (t - t_{n-1})\, dt \le L \int_{I_n} \|\widetilde{W}\| \|W'\| (t - t_{n-1})\, dt$$

bzw. unter Verwendung der Ungleichung $ab \le \frac{1}{2}a^2 + \frac{1}{2}b^2$:

$$\int_{I_n} \|W'\|^2 (t - t_{n-1})\, dt \le L^2 \int_{I_n} \|\widetilde{W}\|^2 (t - t_{n-1})\, dt \le \tfrac{1}{2} L^2 h_n^2 \sup_{I_n} \|\widetilde{W}\|^2\,. \tag{7.2.14}$$

In Hilfssatz 7.1 werden wir für $W \in P_r(I_n)$ die folgende Abschätzung zeigen:

$$\sup_{I_n} \|W\|^2 \le \kappa^2 \Big\{ \int_{I_n} \|W'\|^2 (t - t_{n-1})\, dt + \|W_n^-\|^2 \Big\}, \tag{7.2.15}$$

mit einer von der Intervallbreite h_n unabhängigen Konstante $\kappa > 0$. Durch Kombination der vorausgehenden Abschätzungen (7.2.13) - (7.2.15) erhalten wir

$$\begin{aligned} \sup_{I_n} \|W\|^2 &\le \kappa^2 \Big\{ \int_{I_n} \|W'\|^2 (t - t_{n-1})\, dt + \|W_n^-\|^2 \Big\} \\ &\le \kappa^2 \Big\{ \tfrac{1}{2} L^2 h_n^2 \sup_{I_n} \|\widetilde{W}\|^2 + 2Lh_n \sup_{I_n} \|\widetilde{W}\| \sup_{I_n} \|W\| \Big\} \end{aligned}$$

und schließlich

$$\sup_{I_n} \|W\| \;\le\; \gamma^{-1} L h_n \sup_{I_n} \|\widetilde{W}\|,$$

mit einer festen Konstante $\gamma > 0$. Hieraus folgt zunächst, dass die Abbildung $g : P_r(I_n) \to P_r(I_n)$ wohl definiert ist, denn für $\tilde{W} = 0$, d. h. für verschwindende rechte Seite in (7.2.12), ist notwendig $W = 0$, was wegen der Linearität der linken Seite in (7.2.12) aufgrund eines fundamentalen Satzes der linearen Algebra die Existenz von $g(\tilde{U}) \in P_r(I_n)$ für jedes Argument $\tilde{U} \in P_r(I_n)$ impliziert. Weiter ist die Abbildung $g(\cdot)$ wie behauptet für $h_n < \gamma/L$ eine Kontraktion. Der Banachsche Fixpunktsatz liefert dann die Existenz einer (eindeutig bestimmten) Lösung der Galerkin-Gleichung (7.2.12).

Q.E.D.

Hilfssatz 7.1 (Diskrete Sobolewsche Ungleichung): *Für Funktionen* $\varphi \in P_r(I_n)^d$ *gilt die diskrete „Sobolewsche*[3] *Ungleichung“*

$$\sup_{I_n} \|\varphi\| \le \kappa \Big(\int_{I_n} \|\varphi'\|^2 (t - t_{n-1})\, dt + \|\varphi_n^-\|^2 \Big)^{1/2}, \tag{7.2.16}$$

mit einer von der Intervallänge h_n *unabhängigen Konstante* $\kappa > 0$ *.*

Beweis: Der Beweis verwendet ein sog. „Skalierungsargument“. Da derartige Argumente in der Analyse von Galerkin-Verfahren häufig vorkommen, wollen wir den Beweis hier vollständig durchführen. Wir betrachten nur den skalaren Fall $d = 1$; die Verallgemeinerung für $d \in \mathbb{N}$ ist dann offensichtlich. Ausgangspunkt ist die Feststellung, dass die Ungleichung (7.2.16) für Polynome $\hat\varphi \in P_r((0,1])$ auf dem Einheitsintervall gilt mit einer Konstante $\hat\kappa > 0$:

$$\sup_{(0,1]} |\hat\varphi| \le \hat\kappa \Big(\int_0^1 |\hat\varphi'|^2\, \hat t\, d\hat t + |\hat\varphi(1)|^2 \Big)^{1/2}. \tag{7.2.17}$$

Dies folgt direkt aufgrund der Äquivalenz aller Normen auf dem endlich dimensionalen Vektorraum $P_r((0,1])$. Man überzeugt sich leicht, dass die Ausdrücke auf der linken und rechten Seite in (7.2.17) in der Tat Normen sind. Wir führen nun eine (affin-lineare) Skalierungstransformation von $(0,1]$ auf I_n ein:

$$\chi : (0,1] \to I_n, \quad t = \chi(\hat t) := t_{n-1} + h_n \hat t, \quad \chi'(\hat t) = h_n.$$

Zu jedem Polynom $\varphi \in P_r(I_n)$ defieren wir damit ein zugehöriges Polynom $\hat\varphi \in P_r((0,1])$ durch

$$\hat\varphi(\hat t) := \varphi(\chi(\hat t)), \quad \hat t \in (0,1].$$

Dann gilt offenbar

$$\hat\varphi'(\hat t) = \varphi'(t)\chi'(\hat t) = \varphi'(t) h_n\,.$$

Mit diesen Bezeichnungen erhalten wir durch Koordinatentransformation die folgenden Beziehungen:

$$\begin{aligned} \sup_{I_n} |\varphi| = \sup_{(0,1]} |\hat\varphi| &\le \kappa \Big(\int_0^1 |\hat\varphi'|^2\, \hat t\, d\hat t + |\hat\varphi(1)|^2 \Big)^{1/2} \\ &= \kappa \Big(\int_{I_n} h_n^2 |\varphi'|^2 h_n^{-1} (t - t_{n-1})\, h_n^{-1} dt + |\varphi_n^-|^2 \Big)^{1/2}. \end{aligned}$$

Dies beweist die Gültigkeit der behaupteten Ungleichung mit derselben Konstante $\kappa = \hat\kappa$ auf dem Intervall I_n . Q.E.D.

[3] Sergei Lvovich Sobolew (1908-1989): Russischer Mathematiker; wirkte zunächst in Leningrad (St. Petersburg) und dann am berühmten Steklov-Institut für Mathematik der Akademie der Wissenschaften in Moskau; fundamentale Beiträge zur Theorie der partiellen Differentialgleichungen, Konzept der verallgemeinerten (distributionellen) Lösung, Sobolew-Räume; beschäftigte sich auch mit numerischen Methoden, numerische Quadratur.

Wir bemerken, dass es eine interessante Übungsaufgabe ist, die Abschätzung (7.2.16) direkt ohne Verwendung des Skalierungsarguments zu beweisen und dabei auch die „beste“ Konstante $\kappa = \kappa(r)$ zu bestimmen. Man beachte, dass (7.2.16) nicht gleichmäßig für $\varphi \in C^1(\bar{I}_n)$ gelten kann.

Als nächstes betrachten wir den Fall „steifer“ AWA unter der zusätzlichen Annahme, dass die rechte Seite $f(t,\cdot)$ L-stetig ist und die *strikte* Monotonieeigenschaft

$$-(f(t,x) - f(t,y), x-y) \geq \gamma \|x-y\|^2, \quad t \in I,\ x,y \in \mathbb{R}^d, \tag{7.2.18}$$

besitzt mit einer Konstante $\gamma > 0$. Dies ist z. B. der Fall für $f(t,x) = A(t)x + b(t)$ mit einer (gleichmäßig bzgl. t) negativ definiten Matrix $A(t)$.

Satz 7.2 (dG-Verfahren für steife AWA): *Die AWA habe eine L-stetige und strikt monotone rechte Seite $f(t,x)$. Die Galerkin-Gleichung (7.2.6) besitzt dann unabhängig von der Schrittweite eine eindeutige Lösung $U \in S_h^{(r)}(I)$, welche mit dem Newton-Verfahren berechnet werden kann.*

Beweis: Ausgangspunkt ist die lokale Galerkin-Gleichung

$$\int_{I_n} (U' - f(t,U), \varphi)\, dt + (U_{n-1}^+, \varphi_{n-1}^+) = (U_{n-1}^-, \varphi_{n-1}^+) \quad \forall \varphi \in P_r(I_n)^d. \tag{7.2.19}$$

Dies ist äquivalent zu einer nichtlinearen Gleichung für $U \in P_r(I_n)^d$. Wir wollen zeigen, dass die zugehörige Abbildung strikt monoton ist. Die (eindeutige) Lösbarkeit der Gleichung (7.2.19) für beliebigen Anfangswert U_{n-1}^- wird dann durch Korollar 1.7 garantiert. Für zwei Funktionen $U, V \in P_r(I_n)$ und ihre Differenz $W := U - V$ gilt unter Verwendung der Monotonieeigenschaft von $f(t,\cdot)$:

$$\begin{aligned}\int_{I_n} (W' - f(t,U) + f(t,V), W)\, dt + \|W_{n-1}^+\|^2 & \\ \geq \int_{I_n} \Big\{ \frac{1}{2}\frac{d}{dt}\|W\|^2 + \gamma \|W\|^2 \Big\}\, dt + \|W_{n-1}^+\|^2 & \\ = \frac{1}{2}\|W_n^-\|^2 + \frac{1}{2}\|W_{n-1}^+\|^2 + \gamma \int_{I_n} \|W\|^2\, dt. &\end{aligned}$$

Da alle Normen auf dem endlich dimensionalen Vektorraum $P_r(I_n)^d$ äquivalent sind, bedeutet dies die behauptete starke Monotonie der Abbildung. Q.E.D.

Bemerkung: Wir bemerken, dass in Satz 7.2 für $r = 0$ und $r = 1$ als Voraussetzung die einfache *Semi-Monotonie* der Funktion $f(t,\cdot)$ ausreicht. In diesem Fall ist nämlich bereits

$$|||W||| := \big(\|W_n^-\|^2 + \|W_{n-1}^+\|^2\big)^{1/2}$$

eine Norm auf $P_r(I_n)^d$, so dass $\gamma = 0$ sein darf. Ob dies auch richtig ist für $r \geq 2$, ist eine (zum Zeitpunkt der Erstellung dieses Textes) noch offene Frage. Ein einfaches Beispiel für eine steife AWA mit nur semi-monotoner rechter Seite ist das 4×4-System

$$u'(t) = Au(t), \quad t \geq 0, \quad u(0) = (1,0,1,1)^T,$$

mit der Matrix

$$A = \begin{pmatrix} -100 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

und der Lösung

$$u_1(t) = e^{-100t}, \;\; u_2(t) = \sin(t), \;\; u_3(t) = \cos(t), \;\; u_4(t) = e^{-t}.$$

7.2.3 Andere Arten von Galerkin-Verfahren

Neben dem impliziten Euler-Schema lassen sich auch einige der anderen bisher betrachteten einfachen Einschrittformeln als Varianten von Galerkin-Verfahren deuten.

a) Verwendet man bei der Herleitung der variationellen Formulierung (7.2.4) ein Punktgitter $\{t_0, ..., t_N\}$ mit nach *rechts* halboffenen Teilintervallen $I_n = [t_{n-1}, t_n)$, so erhält man

$$\sum_{n=1}^{N} \left\{ \int_{I_n} (\tilde{u}' - f(t, \tilde{u}), \varphi)\, dt + ([\tilde{u}]_n, \varphi_n^-) \right\} = 0\,.$$

Ein unstetiger Galerkin-Ansatz mit stückweise konstanten Funktionen ergibt dann die Rekursionsgleichungen

$$[U]_n = \int_{I_n} f(t, U)\, dt\,, \quad 1 \le n \le N\,.$$

Diese entsprechen offensichtlich einem expliziten Einschrittverfahren für die Größen $U_n := U_n^+$, welches im autonomen Fall, $f(t,x) = f(x)$, oder nach numerischer Integration mit der *linksseitigen* Boxregel mit der Polygonzugmethode (explizites Euler-Verfahren)

$$U_n = U_{n-1} + h_n f(t_{n-1}, U_{n-1})$$

übereinstimmt.

b) Ausgehend von der variationellen Formulierung (7.2.4) kann man auf dem Gitter $\{t_0, ..., t_N\}$ stetige Ansätze für U machen. Damit dies aber wieder ein *Zeitschrittverfahren* liefert, welches rekursiv von Zeitlevel zu Zeitlevel abgearbeitet werden kann, müssen die Testfunktionen als *unstetig* gewählt werden. Dies ergibt dann ein sog. „Petrow[4]-Galerkin-Verfahren“ im Gegensatz zum „Galerkin-Verfahren“, bei dem Ansatz- und Testraum dieselben sind. Bei Wahl von stückweise linearen Ansätzen für U und stückweise konstanten Testfunktionen ergibt sich wegen $U_n := U_n^+ = U_n^-$ das Schema

$$\int_{I_n} \{U' - f(t, U)\}\, dt = 0$$

[4]Georgi Iwanowitsch Petrow (1912-1987): Russischer Ingenieur; 1965-1973 Direktor des Instituts für Raumfahrtforschung; Publ.: “Application of the Galerkin method and the problem of flow stability of a viscous liquid” (russ.), Prikl. Mat. Mekh. 4, 36-47 (1947)

bzw.

$$U_n - U_{n-1} + \int_{I_n} f(t, h_n^{-1}(t - t_{n-1})U_n + h_n^{-1}(t_n - t)U_{n-1})\, dt\,.$$

Dieses Schema stimmt für eine lineare, autonome AWA mit $f(t,x) = Ax + b$ oder bei Anwendung der Trapezregel auf das Integral offenbar mit der Trapezformel

$$U_n = U_{n-1} + \tfrac{1}{2}h_n \left\{ f(t_n, U_n) + f(t_{n-1}, U_{n-1}) \right\}\,.$$

überein. Die Resultate der folgenden Abschnitte zur *a priori* und *a posteriori* Fehleranalyse und Schrittweitensteuerung bei den *unstetigen* Galerkin-Verfahren gelten sinngemäß auch für diese *stetigen* Petrow-Galerkin-Verfahren.

7.3 A priori Fehleranalyse

Die dG(r)-Verfahren lassen a priori Fehlerabschätzungen zu, welche eine ähnliche Struktur wie diejenigen für Differenzenverfahren haben. Sie stellen aber etwas geringere Anforderungen an die Regularität der exakten Lösung. Wir wollen diesen wichtigen Punkt zunächst anhand eines ganz einfachen Spezialfalles diskutieren. Betrachtet werde die triviale skalare AWA

$$u'(t) = f(t)\,,\ t \in I = [0,1]\,, \quad u(0) = 0\,, \tag{7.3.20}$$

deren Lösung $u(t)$ gerade die Stammfunktion der rechten Seite $f(t)$ über dem jeweiligen Intervall $[0,t]$ ist. Die dG(0)-Verfahren nimmt angewendet auf diese AWA die folgende Gestalt an:

$$U_n^- = U_{n-1}^+ = \int_{I_n} f(t)\, dt + U_{n-1}^-\,. \tag{7.3.21}$$

Die exakte Lösung $u(t)$ erfüllt offenbar dieselbe Gleichung,

$$u_n = \int_{I_n} f(t)\, dt + u_{n-1},$$

so dass der Fehler $e = u - U$ zu den Zeiten t_n^- verschwindet: $e_0^- = e_1^- = \ldots = e_N^- = 0$. In den Zwischenpunkten $t \in I_n$ gilt

$$|e(t)| = \Big| \int_t^{t_n} e'\, dt \Big| = \Big| \int_t^{t_n} u'\, dt \Big| \le h_n \sup_{I_n} |u'|\,, \quad t \in I_n.$$

Dies impliziert die globale Fehlerabschätzung

$$\sup_I |e| \le \max_{1 \le n \le N} \{ h_n \sup_{I_n} |u'| \}\,, \tag{7.3.22}$$

welche für stückweise konstante Approximation $(r = 0)$ hinsichtlich der Konvergenzordnung und den Regularitätsanforderungen an die Lösung u sicherlich optimal ist. Vergleicht man dieses Resultat mit dem entsprechenden für das implizite Euler-Verfahren (für die vorliegende semi-monotone AWA),

$$\sup_I |e| \le \tfrac{1}{2} \sum_{n=1}^{N} h_n^2 \sup_{I_n} |u''| \le \tfrac{1}{2} T \max_{1\le n\le N} \{h_n \sup_{I_n} |u''|\}\,, \tag{7.3.23}$$

so fallen zwei Unterschiede auf:

- Der Fehler in (7.3.23) wächst linear mit der Zeit T.
- Die Abschätzung (7.3.23) erfordert eine Schranke für die zweite Ableitung u''.

Dieser Unterschied ist verfahrenstypisch und tritt auch im Fall allgemeinerer AWA auf. Im betrachteten trivialen Fall wäre die Folgerung für die dG(0)-Verfahren, dass die Auswertung der Integrale in (7.3.21) statt mit der Box-Regel (wie beim impliziten Euler-Verfahren) besser mit einer Quadraturformel 2. Ordnung, z. B. der Mittelpunktsregel, erfolgen sollte:

$$U_n = h_n f(t_{n-1/2}) + U_{n-1}\,, \quad t_{n-1/2} = \tfrac{1}{2}(t_n + t_{n-1})\,.$$

Für dieses Differenzenverfahren erhält man die Fehlerdarstellung

$$e_n = e_{n-1} + \int_{t_{n-1}}^{t_n} f(t)\,dt - h_n f(t_{n-1/2}) = e_{n-1} + \tfrac{1}{24} h_n^3 f''(\zeta_n)\,.$$

Hieraus folgt die Abschätzung

$$\sup_I |e| \le \max_{1\le n\le N} \big\{h_n \sup_{I_n} |u'| + \tfrac{1}{24}\, T\, h_n^2 \sup_{I_n} |f''|\big\}\,. \tag{7.3.24}$$

Der bei Intergration über große Zeitintervalle dominante Zeitfaktor T kann hier durch die erhöhte Potenz der Schrittweite h_n kompensiert werden. Dies macht Sinn, wenn die höheren Ableitungen der Lösung nicht deutlich größer als die niedrigen sind. Dies ist natürlich im vorliegenden Spezialfall kein allzu großer Fortschritt, aber das zugrunde liegende Prinzip ist auch in viel allgemeineren Situationen wirksam und führt zu neuartigen Ansätzen bei der Kontrolle des Diskretisierungsfehlers. Insbesondere erscheint in der Abschätzung (7.3.24) der reine Verfahrensfehler (Approximation der Zeitableitung) separiert vom Quadraturfehler bei der Auswertung der rechten Seite $f(t,\cdot)$. Beim Differenzenverfahren werden dagegen beide Fehleranteile vermischt.

Wir geben nun a priori Abschätzungen für den Diskretisierungsfehler der dG(r)-Verfahren an.

Satz 7.3 (A priori Fehler - nicht-steif): *Im Fall einer allgemeinen L-stetigen AWA gilt für das dG(r)-Verfahren für hinreichend kleine Schrittweiten $h_n < \gamma/L$ (siehe Satz 7.1) die a priori Fehlerabschätzung*

$$\sup_I \|e\| \le K \max_{1\le n\le N} \big\{h_n^{r+1} \sup_{I_n} \|u^{(r+1)}\|\big\}, \tag{7.3.25}$$

mit einer im allg. exponentiell von L und T abhängigen Konstante $K = K(L,T)$.

Beweis: Wir führen den Beweis in mehreren Schritten.
(i) Wir betrachten als Übung zunächst den einfachsten Fall $r = 0$. Subtraktion der Gleichungen für u und U,

$$u_n = \int_{I_n} f(t, u(t))\, dt + u_{n-1},$$
$$U_n^- = U_{n-1}^+ = \int_{I_n} f(t, U(t))\, dt + U_{n-1}^-,$$

ergibt für die Differenz $e_n := u_n - U_n^-$:

$$\begin{aligned} e_n &= \int_{I_n} \{f(t, u(t)) - f(t, U(t))\}\, dt + e_{n-1} \\ &= \int_{I_n} \{f(t, u(t)) - f(t, u_n) + f(t, u_n) - f(t, U(t))\}\, dt + e_{n-1}. \end{aligned}$$

Damit folgt

$$\begin{aligned} \|e_n\| &= \int_{I_n} \|f(t, u(t)) - f(t, u_n)\|\, dt + \int_{I_n} \|f(t, u_n) - f(t, U_n^-)\|\, dt + \|e_{n-1}\| \\ &\le Lh_n \sup_{t \in I_n} \|u(t) - u(t_n)\| + Lh_n \|e_n\| + \|e_{n-1}\|, \end{aligned}$$

und weiter durch rekursive Anwendung dieser Ungleichung und Beachtung von $e_0 = 0$:

$$\|e_n\| \le L \sum_{m=1}^{n} h_m \|e_m\| + L \sum_{m=1}^{n} h_m \sup_{t \in I_m} \|u(t) - u(t_m)\|.$$

Die diskrete Gronwollsche Ungleichugn liefert dann die Abschätzung

$$\|e_n\| \le e^{\gamma L(t_n - t_0)} L \sum_{m=1}^{n} h_m \sup_{t \in I_m} \|u(t) - u(t_m)\|,$$

bzw.

$$\begin{aligned} \sup_{t \in I_n} \|e\| &\le \sup_{t \in I_n} \|u(t) - u_n\| + \|e_n\| \\ &\le \sup_{t \in I_n} \|u(t) - u_n\| + e^{\gamma L(t_n - t_0)} L \sum_{m=1}^{n} h_m \sup_{t \in I_m} \|u(t) - u(t_m)\|. \end{aligned}$$

Mit Hilfe der Abschätzung

$$\|u(t) - u(t_m)\| = \Big\| \int_t^{t_m} u'(t)\, dt \Big\| \le \int_t^{t_m} \|u'(t)\| \le h_m \sup_{I_m} \|u'\|.$$

erhalten wir also

$$\sup_{t \in I_n} \|e\| \le K(L, T) \max_{1 \le m \le n} \big\{ h_m \sup_{I_m} \|u'\| \big\}.$$

Dies impliziert die behauptete Fehlerabschätzung für $r = 0$.

(ii) Wir wenden uns jetzt dem allgemeinen Fall $r \geq 0$ zu. Der exakten Lösung u wird eine Approximierende $\bar{U} \in S_h^{(r)}(I)$ zugeordnet durch die Vorschrift

$$\bar{U}_n^- = u(t_n), \quad \int_{I_n} (\bar{U} - u) q \, dt = 0 \quad \forall q \in P_{r-1}(I_n)^d, \quad n = 1, ..., N.$$

Wir werden unten in Hilfssatz 7.2 zeigen, dass diese Approximierende lokal auf jedem Teilintervall I_n eindeutig bestimmt ist ($r+1$ Bestimmungsgleichungen für $r+1$ freie Parameter) und dass die Fehlerabschätzung gilt:

$$\sup_{I_n} \|u - \bar{U}\| \leq c_I \, h_n^{r+1} \sup_{I_n} \|u^{(r+1)}\|, \tag{7.3.26}$$

mit einer sog. „Interpolationskonstante“ $c_I > 0$. Für diese Approximierende $\bar{U}$ gilt nach Konstruktion mit beliebigem $\varphi \in P_r(I_n)$ die folgende Beziehung (beachte $\varphi' \in P_{r-1}(I_n)$)

$$\begin{aligned} \int_{I_n} (\bar{U}', \varphi) \, dt + (\bar{U}_{n-1}^+, \varphi_{n-1}^+) &= - \int_{I_n} (\bar{U}, \varphi') \, dt + (\bar{U}_n^-, \varphi_n^-) \\ &= - \int_{I_n} (u, \varphi') \, dt + (u_n, \varphi_n^-) \\ &= \int_{I_n} (u', \varphi) \, dt + (u_{n-1}, \varphi_{n-1}^+) \\ &= \int_{I_n} (f(t,u), \varphi) \, dt + (u_{n-1}, \varphi_{n-1}^+). \end{aligned} \tag{7.3.27}$$

(iii) Wir wollen eine rekursive Abschätzung für den intervallweisen Fehlerterm $E_n := \sup_{I_n} \|e\|^2$ herleiten. Dazu wird aufgespalten $e := \bar{e} + \eta$ mit $\bar{e} := u - \bar{U}$ und $\eta := \bar{U} - U$. Im Hinblick auf (7.3.26) genügt es, die Fehlerkomponente η abzuschätzen. Wir bedienen uns dazu der „diskreten“ Sobolewschen Ungleichung (7.2.16) für $\eta \in P_r(I_n)$:

$$\sup_{I_n} \|\eta\|^2 \leq \kappa^2 \Big\{ \int_{I_n} \|\eta'\|^2 (t - t_{n-1}) \, dt + \|\eta_n^-\|^2 \Big\}. \tag{7.3.28}$$

Durch Subtraktion der Gleichungen (7.2.7) für U und (7.3.27) für $\bar{U}$ erhalten wir

$$\int_{I_n} (\eta', \varphi) \, dt + (\eta_{n-1}^+, \varphi_{n-1}^+) = \int_{I_n} (f(t,u) - f(t,U), \varphi) \, dt + (\eta_{n-1}^-, \varphi_{n-1}^+),$$

für beliebiges $\varphi \in P_r(I_n)$. Wir wählen nun $\varphi := \eta$ und erhalten

$$\int_{I_n} (\eta', \eta) \, dt + \|\eta_{n-1}^+\|^2 = \int_{I_n} (f(t,u) - f(t,U), \eta) \, dt + (\eta_{n-1}^-, \eta_{n-1}^+),$$

bzw. bei Beachtung von $(\eta', \eta) = \frac{1}{2} \frac{d}{dt} \|\eta\|^2$ und anschließender Integration über I_n:

$$\tfrac{1}{2} \|\eta_n^-\|^2 + \tfrac{1}{2} \|\eta_{n-1}^+\|^2 \leq L \int_{I_n} \|e\| \|\eta\| \, dt + \tfrac{1}{2} \|\eta_{n-1}^-\|^2 + \tfrac{1}{2} \|\eta_{n-1}^+\|^2.$$

Dies ergibt

$$\|\eta_n^-\|^2 \le 2L \int_{I_n} \|e\|\|\eta\|\,dt + \|\eta_{n-1}^-\|^2,$$

und nach rekursiver Anwendung für $n, n-1, ..., 1$ und Beachtung von $\eta_0^- = 0$:

$$\|\eta_n^-\|^2 \le 2L \sum_{\nu=1}^{n} \int_{I_\nu} \|e\|\|\eta\|\,dt\,. \tag{7.3.29}$$

Als nächstes setzen wir $\varphi := \eta'(t-t_{n-1}) \in P_r(I_{n-1})$ und erhalten

$$\begin{aligned}\int_{I_n} \|\eta'\|^2(t-t_{n-1})\,dt &= \int_{I_n} (f(t,u)-f(t,U),\eta')(t-t_{n-1})\,dt \le L\int_{I_n} \|e\|\|\eta'\|(t-t_{n-1})\,dt \\ &\le L\Big(\int_{I_n} \|e\|^2(t-t_{n-1})\,dt\Big)^{1/2}\Big(\int_{I_n} \|\eta'\|^2(t-t_{n-1})\,dt\Big)^{1/2}\end{aligned}$$

bzw.

$$\int_{I_n} \|\eta'\|^2(t-t_{n-1})\,dt \le L^2 \int_{I_n} \|e\|^2(t-t_{n-1})\,dt\,. \tag{7.3.30}$$

Kombination der Abschätzungen (7.3.28), (7.3.29) und (7.3.30) ergibt

$$\sup_{I_n} \|\eta\|^2 \le \kappa^2 L^2 h_n^2 \sup_{I_n} \|e\|^2 + 2\kappa^2 L \sum_{\nu=1}^{n} h_\nu \sup_{I_\nu} \|e\|\|\eta\|\,.$$

Unter Verwendung der Abschätzung (7.3.26) für $\bar e = u - \bar U$ erschließen wir hieraus

$$\begin{aligned}\sup_{I_n} \|e\|^2 &\le 2\kappa^2 L^2 h_n^2 \sup_{I_n} \|e\|^2 + 2c_I h_n^{2r+2} \sup_{I_n} \|u^{(r+1)}\|^2 \\ &+ \kappa^2 L \sum_{\nu=1}^{n} h_\nu \big\{3 \sup_{I_\nu} \|e\|^2 + 2c_I h_\nu^{2r+2} \sup_{I_\nu} \|u^{(r+1)}\|^2\big\} \\ &\le 5\kappa^2 L \sum_{\nu=1}^{n} h_\nu \sup_{I_\nu} \|e\|^2 + 4\kappa^2 L c_I \sum_{\nu=1}^{n} h_\nu^{2r+3} \sup_{I_\nu} \|u^{(r+1)}\|^2,\end{aligned}$$

wobei o.B.d.A. $\kappa^2 L \ge 1$ und $Lh_n \le 1$ angenommen wurde. Auf der Basis dieser Beziehung liefert nun die *diskrete* Gronwallsche Ungleichung (2.1.11) unter der Voraussetzung $5\kappa^2 L h_\nu < 1$, dass

$$\sup_{I_n} \|e\|^2 \le \exp\big(\sum_{\nu=0}^{n} \sigma_\nu 5\kappa^2 L h_\nu\big) 4\kappa^2 L c_I \sum_{\nu=1}^{n} h_\nu^{2r+3} \sup_{I_\nu} \|u^{(r+1)}\|^2, \tag{7.3.31}$$

wobei $\sigma_\nu = (1 - 5\kappa^2 L h_\nu)^{-1}$. Dies impliziert die behauptete a priori Fehlerabschätzung mit der Konstante $\gamma_0 := 1/(5\kappa^2)$. Q.E.D.

Hilfssatz 7.2 (Interpolationssatz): *Einer Funktion* $u \in C^{r+1}(\bar{I}_n)$ *wird durch die Vorschrift*

$$\bar{U}_n^- = u(t_n), \quad \int_{I_n} (\bar{U} - u) q \, dt = 0 \quad \forall q \in P_{r-1}(I_n)^d, \tag{7.3.32}$$

eindeutig eine Interpolierende $\bar{U} \in P_r(I_n)$ *zugeordnet. Für diese gilt die Fehlerabschätzung*

$$\sup_{I_n} \|u - \bar{U}\| \le c_I h_n^{r+1} \sup_{I_n} \|u^{(r+1)}\|, \tag{7.3.33}$$

mit einer von h_n *unabhängigen sog. „Interpolationskonstante“* $c_I = c_I(r)$.

Beweis: (i) Auf dem Intervall I_n haben wir $r+1$ lineare Bedingungen (7.3.32) zur Bestimmung der ebenfalls $r+1$ freien Parameter in $\bar{U} \in P_r(I_n)$. Das resultierende lineare Gleichungssystem besitzt eine eindeutige Lösung, da jedes Polynom $p \in P_r(I_n)$ mit den Eigenschaften

$$p_n^- = 0, \quad \int_{I_n} pq \, dt = 0 \quad \forall q \in P_{r-1}(I_n)^d.$$

notwendig das Nullpolynom $p \equiv 0$ ist. Um dies einzusehen, nehmen wir an, dass $p \not\equiv 0$ ist. Seien $\{\tau_i, i = 1, ..., m\}$ die Nullstellen von p im Innern von I_n mit *ungerader* Vielfachheit. Dann gilt

$$\int_{I_n} p(t) \prod_{i=1}^{m} (t - \tau_i) \, dt = \int_{I_n} \tilde{p}(t) \prod_{i=1}^{m} (t - \tau_i)^{2s} \, dt > 0$$

mit einem Polynom $\tilde{p}$ ohne Nullstelle in I_n . Da nun p orthogonal zu allen Polynomen in P_{r-1} ist, muss also zwangsläufig $m \ge r$ sein. Zusammen mit $\tau_{r+1} := t_n$ hat also p genau $r+1$ (dann notwendig einfache) Nullstellen, und es ergibt sich der Widerspruch $p \equiv 0$.
(ii) Mit demselben Argument wie unter (i) erhalten wir, dass die Fehlerfunktion $u - \bar{U} \in C^{r+1}(I_n)$ in I_n mindestens $r+1$ (einfache) Nullstellen besitzt. Betrachten wir nun das Nullpolynom $p \equiv 0$ als Lagrange-Interpolierende in $P_r(I_n)$ von $u - \bar{U}$ in diesen Punkten, so liefert die allgemeine Fehlerabschätzung für die Lagrange-Interpolation die folgende Abschätzung:

$$\sup_{I_n} \|u - \bar{U}\| \le \frac{1}{(r+1)!} h_n^{r+1} \sup_{I_n} \|(u - \bar{U})^{(r+1)}\| = \frac{1}{(r+1)!} h_n^{r+1} \sup_{I_n} \|u^{(r+1)}\| .$$

Dies impliziert die Behauptung mit der Konstante $c_I := \frac{1}{(r+1)!}$. Q.E.D.

Für das dG(r)-Verfahren lässt sich mit einer ganz anderen als der in Satz 7.3 verwendeten Beweistechnik zeigen, dass der Fehler in den diskreten Zeitpunkten t_n sogar mit der besseren Ordnung $O(h^{r+2})$ konvergiert. Dieser sog. „Superapproximationseffekt“ kann natürlich nicht auf dem ganzen Intervall I gelten, da allgemeine Funktionen durch Polynome vom Grad r global höchstens mit der Ordnung $O(h^{r+1})$ approximiert werden können.

Satz 7.4 (Superkonvergenz): *Das dG(r)-Verfahren ist „superkonvergent“ in den Stützpunkten t_n, d. h.: Für den Fehler $e := u - U$ gilt:*

$$\max_{1\le n\le N} \|e(t_n)\| \le K \max_{1\le n\le N} \big\{ h_n^{r+2} \sup_{I_n} \|u^{(r+1)}\| \big\} + K \max_I \|e\|^2, \tag{7.3.34}$$

mit einer i. Allg. exponentiell von L und T abhängigen Konstante $K = K(L,T)$.

Beweis: Zum Beweis bedienen wir uns eines sog. „(diskreten) Dualitätsarguments“. Sei wieder $e := u - U$ gesetzt.

(i) Zunächst schreiben wir

$$f(t,u) - f(t,U) = \int_0^1 \frac{d}{ds} f(t, U+se)\, ds = \int_0^1 f_x(t,U+se)e\, ds =: B(t)e\,.$$

Für die Semi-Linearform $A(\cdot;\cdot)$ von oben gilt damit

$$\begin{aligned} A(u;V) - A(U;V) &= \sum_{n=1}^N \int_{I_n} \big\{ (e',V) - (f(t,u) - f(t,U), V)\, ds \big\}\, dt \\ &\quad + \sum_{n=2}^N ([e]_{n-1}, V_{n-1}^+) + (e_0^+, V_0^+) \\ &= \sum_{n=1}^N \int_{I_n} \big\{ (e',V) - (B(t)e, V) \big\}\, dt + \sum_{n=2}^N ([e]_{n-1}, V_{n-1}^+) + (e_0^+, V_0^+)\,. \end{aligned}$$

Die rechte Seite definiert nun offenbar eine von u und U abhängige Bilinearform

$$L(u,U)(W,V) := \sum_{n=1}^N \int_{I_n} \big\{ (W',V) - (B(t)W, V) \big\}\, dt + \sum_{n=2}^N ([W]_{n-1}, V_{n-1}^+) + (W_0^+, V_0^+).$$

Mit dieser gilt wegen der Galerkin-Orthogonalität:

$$L(u,U)(e,V) = A(u;V) - A(U;V) = 0, \quad V \in S_h^{(r)}(I).$$

Für die Form $L(u,u)(W,V)$ erhalten wir mit

$$\int_0^1 f_x(t, u+se)\, ds = f_x(t,u)$$

durch partielle Integration und Umordnung von Termen:

$$\begin{aligned} L(u,u)(W,V) &= \sum_{n=1}^N \int_{I_n} (W' - f_x(t,u)W, V)\, dt + \sum_{n=2}^N ([W]_{n-1}, V_{n-1}^+) + (W_0^+, V_0^+) \\ &= \sum_{n=1}^N \int_{I_n} -(W, V' - f_x(t,u)^* V)\, dt - \sum_{n=1}^{N-1} (W_{n-1}^-, [V]_{n-1}) + (W_N^-, V_N^-). \end{aligned}$$

(ii) Sei nun $Z \in S_h^{(r)}(I)$ die Lösung des diskreten variationellen Problems

$$L(u,u)(V,Z) = (V_N^-, e_N^-) \quad \forall V \in S_h^{(r)}(I). \tag{7.3.35}$$

Dies ist die dG(r)-Approximation des linearen „dualen Problems“ („Rückwärtsproblem“)

$$z' + f_x(t,u)^* z = 0, \;\; t_N \geq t \geq 0, \quad z(t_N) = e_N^-. \tag{7.3.36}$$

Diese AWA ist wohl-gestellt, da sie nach der Transformation $s = t_N - t$ und $\tilde{z}(s) := z(t)$ in die wohl-gestellte „Vorwärtsaufgabe“

$$\tilde{z}' - f_x(s,u)^* \tilde{z} = 0, \;\; 0 \leq s \leq t_N, \quad \tilde{z}(0) = e_N^-,$$

übergeht. Für die diskrete „duale Lösung“ Z gilt die a priori Abschätzung

$$\sup_I \|Z\| + \int_I \|Z'\| \, dt \leq c_{S,h} \|e_N^-\|. \tag{7.3.37}$$

mit einer sog. „Stabilitätskonstante“ $c_{S,h}$, die i. Allg. exponentiell von Schranken für $f_x(t,x)$, d. h. L, und T abhängt.

(iii) Wir setzen $\eta := \bar{U} - U$, wobei $\bar{U} \in S_h^{(r)}(I)$ wieder die oben eingeführte Approximation der exakten Lösung u ist. Mit dieser gilt konstruktionsgemäß $e_N^- = \eta_N^-$. Wir setzen nun $V := \eta := \bar{U} - U$ in (7.3.35) und $\bar{e} := u - \bar{U}$. Dann ergibt sich mit Hilfe der Definitionsgleichung von Z und der Galerkin-Orthogonalität für U:

$$\begin{aligned}
\|e_N^-\|^2 &= L(u,u)(\eta, Z) \\
&= L(u,U)(\eta, Z) - (L(u,U) - L(u,u))(\eta, Z) \\
&= L(u,U)(e, Z) - L(u,U)(\bar{e}, Z) - (L(u,U) - L(u,u))(\eta, Z) \\
&= -L(u,U)(\bar{e}, Z) - (L(u,U) - L(u,u))(\eta, Z) \\
&= -L(u,u)(\bar{e}, Z) + (L(u,U) - L(u,u))(\bar{U} - u, Z) - (L(u,U) - L(u,u))(\eta, Z) \\
&= -L(u,u)(\bar{e}, Z) - (L(u,U) - L(u,u))(e, Z).
\end{aligned}$$

Den zweiten Term auf der rechten Seite schätzen wir nun wie folgt ab:

$$\begin{aligned}
|(L(u,U) - L(u,u))(e,Z)| &\leq \sum_{n=1}^N \int_{I_n} \Big| \int_0^1 (f_x(t, U + se)e, Z) \, ds - (f_x(t,u)e, Z) \Big| \, dt \\
&\leq LT \sup_I \|e\|^2 \sup_I \|Z\| \leq c_{S,h} LT \sup_I \|e\|^2 \|e_N^-\| \, .
\end{aligned}$$

Zur weiteren Behandlung des ersten Terms auf der rechten Seite verwenden wir die Projektionseigenschaften von $\bar{e}$ wie folgt:

$$\begin{aligned}
L(u,u)(\bar{e}, Z) &= \sum_{n=1}^N \int_{I_n} -(\bar{e}, Z' - f_x(t,u)^* Z) \, dt - \sum_{n=1}^{N-1} (\bar{e}_{n-1}^-, [Z]_{n-1}) + (\bar{e}_N^-, Z_N^-) \\
&= \sum_{n=1}^N \int_{I_n} (\bar{e}, f_x(t,u)^* Z - P_0 f_x(t,u)^* Z) \, dt.
\end{aligned}$$

Hieraus folgt dann mit Hilfe der üblichen Fehlerabschätzungen für $\bar{e}$ und P_0:

$$\begin{aligned}|L(u,u)(\bar{e},Z)| &\le \sum_{n=1}^{N} \int_{I_n} \|\bar{e}\| \|f_x(t,u)^*Z - P_0 f_x(t,u)^*Z\| \, dt \\ &\le c_I \sum_{n=1}^{N} h_n^{r+1} \sup_{I_n} \|u^{(r+1)}\| \, h_n \int_{I_n} \|(f_x(t,u)^*Z)'\| \, dt \\ &\le c_I \max_{1\le n\le N} \left\{ h_n^{r+1} \sup_{I_n} \|u^{(r+2)}\| \right\} \sum_{n=1}^{N} \int_{I_n} \|(f_x(t,u)^*Z)'\| \, dt.\end{aligned}$$

Durch Ausdifferenzieren sehen wir, dass auf jedem Zeilintervall I_n gilt:

$$\|(f_x(t,u)^*Z)'(t)\| \;\le\; c(u)\{\|Z(t)\| + \|Z'(t)\|\}.$$

Mit der obigen a priori Abschätzung für Z folgt daher

$$|L(u,u)(\bar{e},Z)| \;\le\; c_I c_{S,h} \max_{1\le n\le N} \left\{ h_n^{r+2} \sup_{I_n} \|u^{(r+1)}\| \right\} \|e_N^-\|.$$

Kombination der bis hierhin abgeleiteten Abschätzungen ergibt nun

$$\|e_N^-\| \;\le\; c_I c_{S,h} \max_{1\le n\le N} \left\{ h_n^{r+2} \sup_{I_n} \|u^{(r+1)}\| \right\} + c_{S,h} L T \sup_{I} \|e\|^2.$$

Q.E.D.

Bemerkung 7.1: Mit Hilfe einer Verfeinerung des vorangehenden Beweises lässt sich zeigen, dass das dG(r)-Verfahren in den diskreten Zeitpunkten sogar eine noch höhere Konvergenzordnung besitzt:

$$\max_{1\le n\le N} \|e(t_n)\| \le K \max_{1\le n\le N} \left\{ h_n^{2r+1} \sup_{I_n} \|u^{(r+1)}\| \right\} + K \max_{I} \|e\|^2, \tag{7.3.38}$$

Dies bedeutet für das dG(0)-Verfahren (implizites Euler-Verfahren) noch keine Verbesserung gegenüber der Konvergenzordnung $m = 1$, doch beim dG(1)-Verfahren ergibt sich bereits die Ordnung $m = 3$.

7.4 A posteriori Fehlerschätzung und Schrittweitensteuerung

7.4.1 Allgemeines zur a posteriori Fehleranalyse

Wir wollen die allgemeine Vorgehensweise bei der a posteriori Fehleranalyse für Galerkin-Verfahren zunächst in dem einfacheren Rahmen algebraischer Gleichungssysteme herleiten. Wir beginnen mit der approximativen Lösung linearer Probleme. Mit (regulären) Matrizen A, $A_h \in \mathbb{R}^{n\times n}$ und Vektoren b, $b_h \in \mathbb{R}^n$ seien die Gleichungssysteme

$$Ax = b, \qquad A_h x_h = b_h$$

betrachtet. Zur Abschätzung des Fehlers $e_h := x - x_h$ kann man sich des „Abschneidefehlers“ $\tau_h := A_h x - b_h$ oder des „Residuums“ $\rho_h := b - Ax_h$ bedienen. Für ersteres gilt

$$A_h e_h = A_h x - A_h x_h = A_h x - b_h = \tau_h$$

und folglich

$$\|e_h\| \le \|A_h^{-1}\| \, \|\tau_h\|. \tag{7.4.39}$$

Dies entspricht der typischen Vorgehensweise bei der *a priori* Fehleranalyse von Differenzenverfahren für AWA. Die resultierende Fehlerschranke basiert auf Abschätzungen für den (unbekannten) Abschneidefehler τ_h sowie der Stabilität des „diskreten“ Operators A_h, d. h. auf Schranken der Form $\|A_h^{-1}\| \le c_{S,h}$ mit gleichmäßig bzgl. des Parameters h beschränkten Konstanten $c_{S,h}$. Diese Beziehung kann zum Nachweis der Konvergenz $\|e_h\| \to 0$ $(h \to 0)$ verwendet werden. Sie erlaubt aber nur mit starken Einschränkungen eine *a posteriori* Fehlerschätzung.

Mit dem Residuum ρ_h gilt

$$Ae_h = Ax - Ax_h = b - Ax_h = \rho_h$$

und folglich

$$\|e_h\| \le \|A^{-1}\| \, \|\rho_h\|. \tag{7.4.40}$$

In diesem Fall ist der Fehler abgeschätzt durch das auswertbare Residuum mit der Stabilitätskonstante $c_S := \|A^{-1}\|$ des ungestörten Operators A. Hieraus lässt sich meist keine *a priori* Information über die Konvergenz der Approximation für $h \to 0$ ableiten, wohl aber eine *a posteriori* Fehlerabschätzung für die berechnete Näherung x_h. Dazu ist eine Schätzung für die Stabilitätskonstante c_S erforderlich, welche i. Allg. kaum mit brauchbarer Genauigkeit analytisch zur Verfügung steht. Die Bestimmung der Stabilitätskonstante c_S kann aber numerisch vorgenommen werden. Dazu nehmen wir an, dass die Abschätzung einer Fehlerkomponente $|e_{h,i}|$ oder allgemeiner eines linearen Funktionalwerts

$$J(e_h) = (e_h, j)$$

mit einem Vektor $j \in \mathbb{R}^n$ gewünscht ist. Die Abschätzung der Norm $\|e_h\|$, lässt sich mit der Setzung $j := \|e_h\|^{-1} e_h$ in diesen Rahmen einordnen. Mit der Lösung $z \in \mathbb{R}^n$ des sog. „dualen“ Problems

$$A^* z = j, \tag{7.4.41}$$

gebildet mit der „Adjungierten“ (hier der Transponierten) A^* von A, gilt dann

$$J(e_h) = (e_h, j) = (e_h, A^* z) = (Ae_h, z) = (\rho_h, z)$$

und folglich die „gewichtete“ Fehlerabschätzung

$$|J(e_h)| \le \sum_{i=1}^{n} |z_i| \, |\rho_{h,i}|. \tag{7.4.42}$$

In dieser Beziehung beschreiben die Gewichte $\omega_i := |z_i|$ die Auswirkung einer Reduzierung der einzelnen Residuenkomponenten ρ_i (durch Verbesserung der Approximation) auf die Zielgröße $J(e_h)$. Zur Auswertung dieser Abschätzung muss die „duale“ Lösung z aus (7.4.41) bestimmt werden. Der dazu erforderliche Aufwand entspricht etwa dem der Lösung des eigentlichen Problems, d. h.: Zielorientierte *a posteriori* Fehlerschätzung ist kostspielig.

Wir wollen nun die eben beschriebene Vorgehensweise zur Ableitung von *a posteriori* Fehlerabschätzungen für lineare Funktionale des Fehlers für *nichtlineare* Gleichungssysteme verallgemeinern. Mit zwei stetig differenzierbaren Abbildungen $A(\cdot)$, $A_h(\cdot) : \mathbb{R}^n \to \mathbb{R}^n$ und Vektoren b, $b_h \in \mathbb{R}^n$ seien die Systeme

$$A(x) = b, \qquad A_h(x_h) = b_h, \tag{7.4.43}$$

betrachtet. Die Abbildung $A(\cdot)$ sei differenzierbar mit Jacobi-Matrix $A'(x)$. Wir nehmen an, dass die beiden Probleme in (7.4.43) eindeutige Lösungen haben. Wir wollen wieder den Approximationsfehler $e_h := x - x_h$ mit Hilfe des berechenbaren Residuums $\rho_h := b - A(x_h)$ abschätzen. Nach dem Fundamentalsatz der Differential- und Integralrechnung gilt für beliebiges $y \in \mathbb{R}^n$:

$$\begin{aligned}(\rho_h, y) = (A(x) - A(x_h), y) &= \int_0^1 \frac{d}{ds}(A(x_h + se_h), y)\, ds \\ &= \int_0^1 (A'(x_h + se_h)e_h, y)\, ds = (B_h e_h, y),\end{aligned}$$

mit der von x und x_h abhängigen Matrix

$$B_h = B(x, x_h) := \int_0^1 A'(x_h + se_h)\, ds\,.$$

Ist wieder ein Funktionalwert $J(e_h) = (e_h, j)$ abzuschätzen, wird nun das *linearisierte* duale Problem

$$B_h^* z = j \tag{7.4.44}$$

verwendet. Mit seiner Lösung z gilt dann nach Konstruktion:

$$J(e_h) = (e_h, j) = (e_h, B_h^* z) = (B_h e_h, z) = (\rho, z),$$

bzw. genau wie im linearen Fall:

$$|J(e_h)| \le \eta := \sum_{i=1}^{n} |z_i|\, |\rho_i|. \tag{7.4.45}$$

Zur Auswertung dieser Abschätzung wäre nun wieder das duale Problem zu lösen. Dies ist aber nicht ohne weiteres möglich, da die Matrix B_h von der unbekannten Lösung x abhängt. Es liegt nahe, hier pragmatisch einfach x durch die berechnete Approximation x_h zu ersetzen. Dies führt wegen

$$B_h \approx \tilde{B}_h := B(x_h, x_h) = \int_0^1 A'(x_h)\, ds = A'(x_h).$$

auf das approximative duale Problem

$$A'(x_h)^* \tilde{z} = j\,.$$

Um den Lösungsaufwand weiter zu reduzieren wird auch nur eine Approximation dazu gelöst:

$$A_h'(x_h)^* \tilde{z}_h = j_h. \tag{7.4.46}$$

Mit der resultierenden approximativen dualen Lösung $\tilde{z}_h$ wird dann die Fehlerschätzung verwendet:

$$|J(e_h)| \approx \tilde{\eta} := |(\rho_h, \tilde{z}_h)| = \sum_{i=1}^{n} \tilde{\omega}_i |\rho_i|, \quad \tilde{\omega}_i := |\tilde{z}_{h,i}|. \tag{7.4.47}$$

Die Frage nach der Verlässlichkeit der Approximation $\tilde{\eta} \approx \eta$ kann i. Allg. nur heuristisch beantwortet werden. Zunächst gilt die gestörte Fehleridentität

$$\begin{aligned} J(e_h) &= (e, j) = (e_h, \tilde{B}_h^* \tilde{z}) = (\tilde{B}_h e_h, \tilde{z}) \\ &= ((\tilde{B}_h - B_h)e_h, \tilde{z}) + (B_h e_h, \tilde{z}) \\ &= ((\tilde{B}_h - B_h)e_h, \tilde{z}) + (\rho_h, \tilde{z}) \\ &= ((\tilde{B}_h - B_h)e_h, \tilde{z}) + (\rho_h, \tilde{z} - \tilde{z}_h) + (\rho_h, \tilde{z}_h). \end{aligned}$$

Mit der Lipschitz-Konstante L' von $A'(\cdot)$ gilt

$$\|\tilde{B}_h - B_h\| = \| \int_0^1 \{A'(x_h) - A'(x_h + s e_h)\}\, ds \| \le \tfrac{1}{2} L' \|e_h\|.$$

Dies ergibt die Abschätzung

$$|J(e_h)| \le \tfrac{1}{2} L' \|e_h\|^2 + \|\rho_h\| \|\tilde{z} - \tilde{z}_h\| + \tilde{\eta}.$$

Die beiden ersten, quadratischen Terme können bei „gutartigen“ Problemen als klein gegenüber dem zu schätzenden Fehler angenommen werden, so dass $\tilde{\eta}$ den wesentlichen Anteil des Schätzers darstellt.

7.4.2 Realisierung für das dG-Verfahren

Die a posteriori Fehleranalyse beim dG(r)-Verfahren bedient sich des eben für algebraische Gleichungssysteme skizzierten Ansatzes. Dieser wird im Folgenden Schritt für Schritt entwickelt werden. Wir verwenden wieder die abkürzende Bezeichnung U für die approximierende Lösung der AWA, welche bestimmt ist durch $U_0^- = u_0$ und

$$\sum_{n=1}^{N} \left\{ \int_{I_n} (U' - f(t,U), \varphi)\, dt + ([U]_{n-1}, \varphi_{n-1}^+) \right\} = 0 \qquad \forall \varphi \in S_h^{(r)}(I)\,. \tag{7.4.48}$$

Bei Verwendung der *Semilinearform*

$$A(U)(V) := \sum_{n=1}^{N} \int_{I_n} (U' - f(t,U), V)\, dt + \sum_{n=2}^{N} ([U]_{n-1}, V_{n-1}^+) + (U_0^+, V_0^+)$$

erhält dies die kompakte Gestalt

$$A(U)(\varphi) \;=\; (u_0, V_0^+) \quad \forall \varphi \in S_h^{(r)}(I)\,. \tag{7.4.49}$$

Durch Vergleich mit der entsprechenden Gleichung für die exakte Lösung u erhält man die Beziehung (Galerkin-Orthogonalität)

$$A(u)(\varphi) - A(U)(\varphi) = 0 \quad \forall \varphi \in S_h^{(r)}(I)\,. \tag{7.4.50}$$

Wir wollen eine a posteriori Abschätzung für den Fehler $e = u - U$ zum Zeitpunkt t_N herleiten. Die Rolle des *Residuums* der diskreten Lösung U übernimmt das Funktional $\rho(U)(\cdot)$, welches durch folgende Relation definiert ist:

$$\rho(U)(\varphi) := (u_0, \varphi_0^+) - A(U)(\varphi)\,, \quad \varphi \in V(I)\,.$$

Offenbar ist $\rho(U)(\varphi) = 0 \;\; \forall \varphi \in S_h^{(r)}(I)$, d. h.: Das Residuum ist in einem gewissen Sinne *orthogonal* zum diskreten Ansatzraum.

Wir wollen als nächstes das duale Problem zur Abschätzung des Endzeitfehlers e_N^- herleiten. Die Beziehungen

$$\begin{aligned} A(u)(\varphi) - A(U)(\varphi) = &\sum_{n=1}^{N} \int_{I_n} (e' - \{f(t,u) - f(t,U)\}, \varphi)\, dt \\ &+ \sum_{n=2}^{N} ([e]_{n-1}, \varphi_{n-1}^+) + (e_0^+, \varphi_0^+) \end{aligned}$$

und (f_x bezeichnet wieder die Jacobi-Matrix von $f(t,x)$ bzgl. der Variablen x.)

$$f(t,u) - f(t,U) \;=\; \int_0^1 f_x(t, U + se)e\, ds \;=: B(t)e$$

legen die Verwendung der folgenden Bilinearform nahe:

$$L(u,U)(v,\varphi) \;:=\; \sum_{n=1}^{N} \int_{I_n} (v' - Bv, \varphi)\, dt + \sum_{n=2}^{N} ([v]_{n-1}, \varphi_{n-1}^+) + (v_0^+, \varphi_0^+)\,.$$

Die zugehörige adjungierte Form $L^*(u,U;v,\varphi) := L(u,U;\varphi,v)$ erhält man durch partielle Integration in der Form

$$L^*(u,U)(v,\varphi) = \sum_{n=1}^{N} \int_{I_n} (-v' - B^*v, \varphi)\, dt - \sum_{n=1}^{N-1} ([v]_n, \varphi_n^-) + (v_N^-, \varphi_N^-)$$

mit der adjungierten Matrix B^* von B. Im Rahmen des Dualitätsarguments wird nun eine Funktion $z \in V(I)$ gesucht mit den Eigenschaften $z_N^+ = \|e_N^-\|^{-1} e_N^-$ und

$$L^*(u,U)(z,\varphi) = (z_N^+, \varphi_N^-) \quad \forall \varphi \in V(I)\,. \tag{7.4.51}$$

Dieses in der Zeit rückwärts laufende Problem ist linear und von derselben Struktur wie die bisher betrachteten Standardprobleme. Die zugehörige AWA lautet

$$z'(t) + B^*(t)z(t) = 0\,, \quad t_0 \le t \le t_N\,, \quad z(t_N) = \|e_N^-\|^{-1} e_N^-\,, \tag{7.4.52}$$

und kann durch die einfache Variablentransformation $t \to t_N + t_0 - t$ in eine vorwärtslaufende überführt werden. Ihre eindeutige Lösbarkeit folgt daher aus dem allgemeinen Existenzssatz für lineare AWA.

Mit der *dualen* Lösung z erhält man nun durch Wahl der Testfunktion $\varphi := e$ in (7.4.51) die Fehlerdarstellung

$$\|e_N^-\| = L^*(u,U;z,e) = L(u,U)(e,z) = A(u)(z) - A(U)(z) = \rho(U)(z)\,.$$

Die Orthogonalitätsbeziehung (7.4.50) erlaubt es nun, auf der rechten Seite eine beliebige Approximation $Z \in S_h^{(r)}(I)$ von z einzuschieben. Wir erhalten damit

$$\|e_N^-\| = \rho(U)(z - Z)\,, \quad Z \in S_h^{(r)}(I). \tag{7.4.53}$$

Das Residuum auf der rechten Seite muss nun explizit ausgewertet werden. Wir haben i. Allg. die Darstellung (beachte $U_0^- := u_0$)

$$\begin{aligned}\rho(U)(z-Z) &= (u_0 - U_0^+, (z-Z)_0^+) - \sum_{n=1}^N \int_{I_n} (R(U), z-Z)\,dt - \sum_{n=2}^N ([U]_{n-1}, (z-Z)_{n-1}^+) \\ &= -\sum_{n=1}^N \left\{ \int_{I_n} (R(U), z-Z)\,dt + ([U]_{n-1}, (z-Z)_{n-1}^+) \right\},\end{aligned}$$

mit dem punktweisen Residuum $R(U) := U' - f(t,U)$. Wir wollen nun geeignete Kandidaten für die Approximierenden $Z \in S_h^{(r)}(I)$ diskutieren.

(i) Eine natürliche Wahl ist die orthogonale Projektion $P_r z \in S_h^{(r)}(I)$ von z auf $S_h^{(r)}(I)$ bzgl. des L^2-Skalarprodukts (sog. „L^2-Projektion“). Diese ist (eindeutig) bestimmt durch die intervallweise Orthogonalitätseigenschaft

$$\int_{I_n} (z - P_r z, q)\,dt = 0 \quad \forall q \in P_r(I_n)^d. \tag{7.4.54}$$

Wir werden unten in Hilfssatz 7.3 Fehlerabschätzungen für die L^2-Projektion herleiten. Damit erhält man unter Ausnutzung der Orthogonalitätseigenschaften von $Z = P_r z$:

$$\begin{aligned}\rho(U)(z-P_r z) = -\sum_{n=1}^N \Big\{ &\int_{I_n} (R(U) - P_r R(U), z - P_r z)\,dt \\ &+ ([U]_{n-1}, (z-P_r z)_{n-1}^+) \Big\}.\end{aligned} \tag{7.4.55}$$

(ii) Eine naheliegende Alternative zur Wahl $Z = P_r z$ ist die modifizierte L^2-Projektion $Z = \tilde{P}_r z \in S_h^{(r)}$, welche für $r \geq 1$ durch

$$(\tilde{P}_r z)_{n-1}^+ = z(t_{n-1}), \quad \int_{I_n} (z - \tilde{P}_r z, q)\, dt \;=\; 0 \quad \forall q \in P_{r-1}(I_n)^d .$$

definiert ist. Im Fall $r = 0$ wird nur die Interpolationsbedingung bei t_{n-1} wirksam. Für diese Projektion haben wir bereits Fehlerabschätzungen in Hilfssatz 7.2 hergeleitet (Tatsächlich wird in Hilfssatz 7.2 eine Variante dieser Projektion mit der Interpolationsvorschrift $(\tilde{P}_r z)_n^- = z(t_n)$ betrachtet.). Unter Verwendung der Projektionseigenschaften von $Z = \tilde{P}_r z$ und Beachtung von $(z - \tilde{P}_r z)_{n-1}^+ = 0$ erhalten wir nun:

$$\rho(U)(z - \tilde{P}_r z) \;=\; -\sum_{n=1}^{N} \int_{I_n} (R(U) - P_{r-1} R(U), z - \tilde{P}_r z)\, dt. \tag{7.4.56}$$

Die beiden Residuumsdarstellungen (7.4.55) und (7.4.56) sind natürlich im Hinblick auf die Identität (7.4.53) äquivalent.

Aus den Residumsdarstellungen (7.4.55) bzw. (7.4.56) gewinnt man nun unmittelbar eine a posteriori Fehlerabschätzung für $\|e_N^-\|$. Wir fassen dieses Resultat ausgehend von der einfacheren Darstellung (7.4.56) in einem Satz zusammen.

Satz 7.5 (A posteriori Fehler): *Die AWA (7.1.1) sei (global) Lipschitz-stetig. Dann gilt für das dG(r)-Verfahren die lokale a posteriori Fehlerabschätzung*

$$\|e_N^-\| \leq \sum_{n=1}^{N} \Big\{ \sup_{I_n} \|R(U) - P_{r-1} R(U)\| \int_{I_n} \|z - \tilde{P}_r z\|\, dt \Big\} \tag{7.4.57}$$

mit der modifizierten L^2-Projektion $Z = \tilde{P}_r z$ und der L^2-Projektion P_{r-1} auf den Ansatzraum $S_h^{(r-1)}$. Hieraus folgt die etwas gröbere globale a posteriori Fehlerabschätzung

$$\|e_N^-\| \leq c_S c_I \max_{1 \leq n \leq N} \big\{ h_n^{r+1} \sup_{I_n} \|R(U) - P_{r-1} R(U)\| \big\} \tag{7.4.58}$$

mit der Interpolationskonstante c_I aus der Abschätzung

$$\int_{I_n} \|z - \tilde{P}_r z\|\, dt \;\leq\; c_I h_n^{r+1} \int_{I_n} \|z^{(r+1)}\|\, dt \tag{7.4.59}$$

und der Stabilitätskonstante c_S aus der Abschätzung

$$\int_I \|z^{(r+1)}\|\, dt \;\leq\; c_S . \tag{7.4.60}$$

Ausgehend von der ersten Fehlerdarstellung (7.4.55) erhalten wir die zu (7.4.57) analoge Fehlerabschätzung

$$\|e_N^-\| \;\leq\; c_S c_I \max_{1 \leq n \leq N} \big\{ h_n^{r+1} \sup_{I_n} \|R(U) - P_r R(U)\| + h_n^r \|[U]_{n-1}\| \big\} \tag{7.4.61}$$

mit der Interpolationskonstante c_I aus der Abschätzung

$$\max\Big\{\int_{I_n}\|z-P_rz\|\,dt,\ h_n\|(z-P_rz)^+_{n-1}\|\Big\}\ \le\ c_I h_n^{r+1}\int_{I_n}\|z^{(r+1)}\|\,dt \tag{7.4.62}$$

und derselben Stabilitätskonstante c_S wie in (7.4.60). Eine überschlagsweise Analyse zeigt, dass der Term $h_n^{r+1}\|R(U)-P_rR(U)\|$ im Allgemeinen um mindestens eine h_n-Potenz kleiner ist als der Term $h_n^r\|[U]_{n-1}\|$. Dies legt es nahe, aus Kostengründen die folgene vereinfachte Fehlerschätzung zu verwenden:

$$\|e_N^-\|\ \approx\ c_S c_I \max_{1\le n\le N}\big\{h_n^r\|[U]_{n-1}\|\big\}\,. \tag{7.4.63}$$

Zur Auswertung der Fehlerschranken (7.4.57) bzw. (7.4.61) benötigen wir möglichst gute Abschätzungen für die Interpolationen P_rz sowie $\tilde P_rz$, d. h. für die Interpolationskonstanten c_I.

Hilfssatz 7.3 (Interpolation): *Für die Projektionen $P_rz\in P_r(I_n)$ sowie $\tilde P_rz\in P_r(I_n)$ gelten die Abschätzungen*

$$\max\Big\{\int_{I_n}\|z-P_rz\|\,dt, h_n\|(z-P_r)^+_{n-1}\|\Big\}\ \le\ c_I h_n^{r+1}\int_{I_n}\|z^{(r+1)}\|\,dt\,, \tag{7.4.64}$$

bzw.

$$\int_{I_n}\|z-\tilde P_rz\|\,dt\ \le\ c_I h_n^{r+1}\int_{I_n}\|z^{(r+1)}\|\,dt\,, \tag{7.4.65}$$

mit der Interpolationskonstante $c_I=\frac{1}{(r+1)!}$.

Beweis: Wir führen den Beweis konstruktiv, um eine explizite Schranke für die Interpolationskonstante c_I zu erhalten. Offenbar genügt es, die Behauptung im skalaren Fall $d=1$ zu zeigen. Wir verwenden zunächst die Orthogonalitätseigenschaften von $z-P_rz$, um zu begründen, dass $z-P_rz$ im Innern von I_n mindenstens $r+1$ paarweise verschiedene Nullstellen haben muss. Die Argumentation verläuft dabei analog zu der im Beweis von Hilfssatz 7.2. In diesen Nullstellen $\{\tau_0,...,\tau_r\}$ wird also z durch das Polynom $P_rz\in P_r(I_n)$ interpoliert. Die allgemeine Fehlerdarstellung für die Lagrange-Interpolation lautet

$$(z-P_rz)(t)=z[s_0,\cdots,s_n,t]\prod_{i=0}^{r}(t-s_i)\,,$$

mit der „dividierten Differenz“ $z[s_0,\cdots,s_r,t]$ in integraler Darstellung

$$\begin{aligned}z[s_0,\cdots,s_r,t]=\int_0^1\int_0^{\tau_1}\cdots\int_0^{\tau_{r-1}}\int_0^{\tau_r} z^{(r+1)}(s_0+\tau_1(s_1-s_0)+\cdots\\ +\tau_r(s_r-s_{r-1})+\tau(t-s_r))\,d\tau d\tau_r\cdots d\tau_2 d\tau_1\,.\end{aligned}$$

Integration über I_n ergibt dann

$$\int_{I_n} |(z - P_r z)(t)|\, dt \;\le\; h_n^{r+1} \int_{I_n} |z[s_0, \cdots, s_r, t]|\, dt \le \frac{1}{(r+1)!} h_n^{r+1} \int_{I_n} |z^{r+1}|\, dt\,.$$

Dies sieht man wie folgt. Vertauschung der Integrationsreihenfolge ergibt zunächst

$$\begin{aligned}\int_{I_n} |z[s_0, \cdots, s_r, t]|\, dt \le \int_0^1 \int_0^{\tau_1} \cdots \int_0^{\tau_{r-1}} \int_0^{\tau_r} \Big(\int_{I_n} |z^{(r+1)}(s_0 + \tau_1(s_1 - s_0) + \cdots \\ + \tau_r(s_r - s_{r-1}) + \tau(t - s_r))|\, dt \Big)\, d\tau d\tau_r \cdots d\tau_2 d\tau_1\,.\end{aligned}$$

Wir setzen nun

$$\xi := s_0 + \tau_1(s_1 - s_0) + \cdots + \tau_r(s_r - s_{r-1}) + \tau(t - s_r)$$

und finden wegen $d\xi = \tau dt$:

$$\int_{I_n} |z^{(r+1)}(s_0 + \tau_1(s_1 - s_0) + \cdots + \tau_r(s_r - s_{r-1}) + \tau(t - s_r))|\, dt \le \tau \int_{I_n} |z^{(r+1)}(\xi)|\, d\xi\,.$$

Dies impliziert

$$\begin{aligned}\int_{I_n} |z[s_0, \cdots, s_r, t]|\, dt &\le \int_0^1 \int_0^{\tau_1} \cdots \int_0^{\tau_{r-1}} \int_0^{\tau_r} \tau\, d\tau d\tau_r \cdots d\tau_2 d\tau_1 \int_{I_n} |z^{(r+1)}(\xi)|\, d\xi \\ &\le \frac{1}{(r+1)!} h_n^{r+1} \int_{I_n} |z^{r+1}|\, dt\,.\end{aligned}$$

Das Argumentation für die Projektion $\tilde{P}_r z$ verläuft ganz analog. Q.E.D.

Die Abschätzung (7.4.57) bezieht sich auf den Fehler zum Endzeitpunkt $t_N = t_0 + T$. Natürlich folgt hieraus auch eine entsprechende Abschätzung über das ganze Integrationsintervall I, wenn man den jeweiligen Zeitpunkt t_n als Endpunkt des Intervalls $[t_0, t_n]$ betrachtet und (7.4.57) sinngemäß anwendet.

7.4.3 Auswertung der a posteriori Fehlerabschätzung

Während die Interpolationskonstante c_I sehr präzise angegeben werden kann, ist die Bestimmung der Stabilitätskonstante c_S i. Allg. schwierig. Dies hat im Wesentlichen drei Ursachen:

- Die Koeffizientenmatrix B^* des dualen Problems hängt explizit von der (unbekannten) exakten Lösung u ab.
- Die Inhomogenität des dualen Problems ist gerade der (unbekannte) Fehler $z_N^- = e_N^- \|e_N^-\|^{-1}$.

– Selbst bei Kenntnis der exakten Matrix B^* sowie der rechten Seite e_N^- würde eine direkte Abschätzung mit Hilfe der Struktureigenschaften des Problems, d. h. der Funktion f_x, in der Regel eine zu grobe Schranke für c_S liefern.

Wir wollen nun die Aussage von Satz 7.5 zunächst für den einfachsten Fall mit $r = 0$, d. h. das dG(0)-Verfahren, konkretisieren. Dazu notieren wir zunächst als Spezialfall von Hilfssatz 7.3 die folgende Interpolationsabschätzung:

$$\|(z - P_0 z)_{n-1}^+\| \le h_n \int_{I_n} \|z'\| \, dt \,, \tag{7.4.66}$$

d. h.: Die Interpolationskonstante ist in diesem Fall $c_I = 1$. Zur Bestimmung der zugehörigen Stabilitätskonstante c_S stellen wir folgenden Hilfssatz bereit.

Hilfssatz 7.4 (Duale Stabilität): *Für die duale lineare AWA (7.4.51) gilt die a priori Abschätzung*

$$\int_I \|z'\| \, dt \le \beta e^\beta \,, \quad \beta = \int_I \|B^*(t)\| \, dt \,. \tag{7.4.67}$$

Ist die Funktion $f(t,x)$ strikt monoton,

$$-(f(t,x) - f(t,x'), x - x) \;\ge\; \gamma \|x - x'\|^2 \,, \quad x, x' \in \mathbb{R}^d \,,$$

so gilt

$$\int_I \|z'\| \, dt \le \min\{T, 2/\gamma\} \sup_I \|B^*\| \,. \tag{7.4.68}$$

Beweis: Zunächst gilt offensichtlich

$$\int_I \|z'\| \, dt = \int_I \|B^* z\| \, dt \le \beta \sup_I \|z\| \,. \tag{7.4.69}$$

Zur weiteren Abschätzung der rechten Seite schreiben wir

$$z(t) = v(t_N) + \int_t^{t_N} B^* z \, ds$$

und erhalten

$$\|z(t)\| \le \|z(t_N)\| + \int_t^{t_N} \|B^*\| \, \|z\| \, ds \,.$$

Anwendung des Gronwallschen Lemmas (diesmal rückwärts in der Zeit) ergibt

$$\|z(t)\| \le \exp\big(\int_t^{t_N} \|B^*\| \, ds \big) \, \|z(t_N)\|$$

bzw.

$$\sup_I \|z(t)\| \le e^\beta \|z(t_N)\| \,.$$

Kombiniert mit (7.4.69) ergibt dies die erste Behauptung. Sei nun die Funktion $f(t,x)$ als monoton angenommen. Damit wird auch die Matrix B^* definit im Sinne

$$-(B^*y,y) = \int_0^1 -(y, f_x(t,U+se)y)\,ds \;\geq\; \gamma\|y\|^2\,.$$

Multiplikation von (7.4.52) mit $-e^{\gamma(t_N-t)}z$ ergibt

$$-\frac{1}{2}\frac{d}{dt}\left\{e^{\gamma(t_N-t)}z^2\right\} - \frac{1}{2}\gamma e^{\gamma(t_N-t)}z^2 - e^{\gamma(t_N-t)}(B^*z,z) \;=\; 0$$

und folglich nach Integration von t nach t_N

$$\|z(t)\| \leq e^{-\gamma(t_N-t)/2}\|z(t_N)\|\,.$$

Wir kombinieren dies mit (7.4.69) und finden

$$\int_I \|z'\|\,dt \leq \sup_I \|B^*\| \int_I e^{-\gamma(t_N-t)/2}\,dt\,\|z(t_N)\| \leq \min\{T, 2/\gamma\} \sup_I \|B^*\|\,\|z(t_N)\|\,,$$

was den Beweis vervollständigt. Q.E.D.

Die Schranken $c_S = \beta e^\beta$ bzw. $c_S = \min\{T,2/\gamma\}\sup_I\|B^*\|$ aus den a priori Abschätzungen (7.4.67) bzw. (7.4.68) sind i. Allg. für praktische Zwecke viel zu grob. Zur Auswertung der a posteriori Abschätzung

$$\|e_N^-\| \leq c_S \max_{1\leq n\leq N} h_n \|[U]_{n-1}\| \tag{7.4.70}$$

sollte man daher c_S numerisch zu schätzen suchen. Dies könnte etwa nach folgender Strategie geschehen:

1. Die Koeffizientenmatrix $B(t)$ wird approximiert durch

$$B(t) = \int_0^1 f_x(t,U+se)\,ds \;\approx\; f_x(t,U)\,. \tag{7.4.71}$$

   Dies ist gerechtfertigt, da bei zunehmender Approximationsgenauigkeit in Folge der Gitteranpassung der Fehler e immer kleiner wird (garantiert durch die a priori Fehlerabschätzung).

2. Der Fehler e_N^- wird durch die Differenz der Lösungen zu zwei aufeinander folgenden Gittern approximiert:

$$e_N^- \;\approx\; \tilde{e}_N^- \;:=\; U^{h'} - U^h\,, \quad \tilde{z}_N := \|\tilde{e}_N^-\|^{-1}\tilde{e}_N^-\,. \tag{7.4.72}$$

   Dies ist gerechtfertigt, wenn das h'-Gitter bereits sehr fein ist und somit nur heuristisch (und gegebenenfalls durch den Erfolg) begründbar. Im Fall, dass man nur an einer einzelnen Fehlerkomponente $e_{N,i}^-$ interessiert ist, kann dagegen mit dem festen Anfangswert $z_N := (\delta_j^{(i)})_{j=1}^N$ gearbeitet werden.

3. Das duale Hilfsproblem

$$\tilde{z}'(t) + f_x(t,U)^*\tilde{z}(t) = 0, \quad t_0 \le t \le t_N, \quad \tilde{z}(t_N) = z_N^+, \tag{7.4.73}$$

wird numerisch auf dem aktuellen (oder aus Sparsamkeitsgründen möglicherweise auch auf einem gröberen) Gitter gelöst. Aus der resultierenden approximierenden Lösung $Z \in S_h^{(r)}(I)$ wird dann der duale Fehler $z - Z$ oder direkt die globale Stabilitätskonstante c_S geschätzt:

$$\int_I \|z'\|\,dt \;\approx\; \sum_{n=1}^{N} h_n \|[Z]_{n-1}\| \;=:\; \tilde{c}_S\,. \tag{7.4.74}$$

Für die dG(r)-Verfahren höherer Ordnung, $r \ge 1$, werden zur Auswertung der höheren Ableitungen $z^{(r+1)}$ der dualen Lösung diese direkt mit Hilfe von entsprechenden Differenzenquotienten oder (bei nicht zu komplizierten Differentialgleichungen) unter Ausnutzung der Rekursion $z' = -B^*z$ aus der gerechneten Näherung Z geschätzt. Für das dG(1)-Verfahren ergibt sich somit z. B. die approximative Fehlerschätzung

$$\|e_N^-\| \approx \frac{1}{2}\sum_{n=1}^{N} h_n^2 \|[U]_{n-1}\|\,\|B^*[Z]_{n-1} + B^{*\prime}Z_n^-\|\,. \tag{7.4.75}$$

Dieser ganze Prozess sollte während der fortschreitenden Gitterverfeinerung (und damit einhergehender Fehlerreduktion) iteriert werden. Den Einfluss der unter (1) - (3) beschriebenen Approximationsschritte kann man prinzipiell auch *a posteriori* kontrollieren. Dies wird jedoch im Detail sehr kompliziert, so dass man sich meist mit einer einfachen „ad-hoc-Kontrolle“ der Entwicklung des Schätzers bei fortschreitender Gitterverfeinerung begnügt. Wir wollen zum Abschluss dieser Diskussion noch eine *a priori* Abschätzung für den Fehler durch den Linearisierungsschritt (1) angeben.

Satz 7.6 (Linearisierung im dualen Problem): *Sei $\tilde{z} \in V(I)$ die (eindeutige) Lösung des linearisierten dualen Problems*

$$\tilde{z}'(t) + f_x(t,U)^*\tilde{z}(t) = 0, \quad t_0 \le t \le t_N, \quad \tilde{z}(t_N) = z_N^+, \tag{7.4.76}$$

wobei etwa $z_N^+ := \|e_N^-\|^{-1}e_N^-$ zur Abschätzung des Endzeitfehlers gewählt ist. Damit gilt die gestörte a posteriori Fehlerdarstellung

$$\|e_N^-\| = R(U;\tilde{z}-Z) + \mathcal{O}(\sup_I \|e\|^2)\,, \tag{7.4.77}$$

mit dem Residuum $R(U;\cdot)$ von U wie oben und einem beliebigen $Z \in S_h^{(r)}(I)$.

Beweis: Die variationelle Formulierung des gestörten dualen Problems (7.4.76) lautet

$$A'(U;\varphi,\tilde{z}) = (\varphi_N^+, \tilde{z}_N^+) \quad \forall\varphi \in V(I), \tag{7.4.78}$$

mit der 1. Ableitung von $A(\cdot;\cdot)$ bei U:

$$A'(U;\varphi,v) \;:=\; -\sum_{n=1}^{N}\int_{I_n}(\varphi, v' + f_x(t,U)^* v)\,dt - \sum_{n=1}^{N-1}(\varphi_n^-, [v]_n,) + (\varphi_N^-, v_N^-)\,.$$

Die 2. Ableitung von $A(\cdot;\cdot)$ am Argument η hat die Form

$$A''(\eta;\psi,\varphi,v) \;:=\; -\sum_{n=1}^{N}\int_{I_n}(\varphi, f_{xx}(t,\eta)^*\psi v)\,dt$$

mit der 2. Ableitung f_{xx} von f nach dem Argument x:

$$f_{xx} := \Big(\frac{\partial^2 f_i(t,x)}{\partial x_j \partial x_k}\Big)_{i,j,k=1}^{d}.$$

Durch Taylor-Entwicklung bis zum Restglied 2. Stufe erhalten wir

$$A(u;\tilde z) \;=\; A(U;\tilde z) - A'(U;e,\tilde z) - \int_0^1 A''(U+se;e,e,\tilde z)(s-1)\,ds$$

bzw.

$$A'(U;e,\tilde z) \;=\; R(U;\tilde z) + \int_0^1 A''(U+se;e,e,\tilde z)(s-1)\,ds$$

mit dem Fehler $e := u - U$. Wählen wir in der variationellen Formulierung des dualen Problems(7.4.78) als Testfunktion $\varphi := e$, so ergibt sich

$$\|e_N^-\| = A'(U;e,\tilde z) = R(U;\tilde z) + \int_0^1 A''(U+se;e,e,\tilde z)(s-1)\,ds.$$

Hieraus folgt mit Hilfe der Galerkin-Orthogonalität die Behauptung. Q.E.D.

Bemerkung: Eine verfeinerte Schrittweitenwahl ist auf Basis der „gewichteten“ a posteriori Fehlerabschätzung

$$\|e_N^-\| \le \sum_{n=1}^{N}\Big\{\sup_{I_n}\|R(U)-P_{r-1}R(U)\|\int_{I_n}\|z-\tilde P_r z\|\,dt\Big\} \tag{7.4.79}$$

möglich. Dabei sind die „Residuen“ $\rho_n := \|R(U)-P_{r-1}R(U)\|$ direkt auswertbar, während die „Gewichte“

$$\omega_n := \int_{I_n}\|z-\tilde P_r z\|\,dt$$

aus einer numerisch berechneten diskreten dualen Lösung approximiert werden müssen.

7.4.4 Adaptive Schrittweitenwahl beim dG(0)-Verfahren

Die Strategie zur adaptiven Schrittweitensteuerung beim dG(0)-Verfahren sieht dann wie folgt aus. Sei eine Fehlertoleranz $\mathrm{TOL} \gg \varepsilon$ (Maschinengenauigkeit) vorgegeben. Beginnend mit einem (möglicherweise äquidistanten) Gitter T_0 mit Schrittweitenvektor $(h_n^{(0)})_{n=1}^{N_0}$ werden auf einer Folge von sukzessiv verfeinerten Gittern T_k, $k = 0, 1, 2, \ldots$, mit Schrittweitenfolgen $(h_n^{(k)})_{n=1}^{N_k}$ Näherungslösungen $U^{(k)} \in S_k^{(r)}(I)$ erzeugt, so dass nach K Schritten gilt

$$\|u(t_N) - U_N^{(K)}\| \leq \mathrm{TOL}\,. \tag{7.4.80}$$

Im k-ten Verfeinerungsschritt wird mit der erreichten Lösung $U = U^{(k)}$ auf dem Gitter T_k zunächst das duale Problem (7.4.51) näherungsweise gelöst, wobei als Startwert die Differenz $Z_N^+ = \|(U^{(k-1)} - U^{(k)})_N^-\|^{-1}(U^{(k-1)} - U^{(k)})_N^-$ genommen wird. Mit der (diskreten) dualen Lösung Z wird eine approximative Stabilitätskonstante c_S bestimmt zu

$$c_S := \sum_{n=1}^{N} \|[Z]_{n-1}\|\,. \tag{7.4.81}$$

Die Interpolationskonstante, in diesem Fall $c_I = 1$, wird als im Voraus bestimmt angenommen. Damit wird dann der aktuelle Fehler geschätzt etwa durch die a posteriori Abschätzung (7.4.63):

$$\|e_N^-\| \;\approx\; c_S c_I \max_{1\leq n\leq N} \left\{h_n^r \|[U]_{n-1}\|\right\}. \tag{7.4.82}$$

Um eine gesicherte Abschätzung für $\sup_I \|u - U\|$ zu erhalten, muss dieser *lokale* Prozess für jeden Zwischenzeitpunkt $t_n \in I$ separat durchgeführt werden, was sehr aufwendig werden kann.

Das Kriterium zur lokalen Verfeinerung der Gitterweite ist nun, inwieweit die lokalen Beiträge der Intervalle I_n zum Gesamtfehler jeweils noch über der gegebenen Toleranz liegen. Dazu setzt man

$$\rho_n := \|h_n^{-1}[U]_{n-1}\|$$

und fragt bei den Zeitschritten $t_{n-1} \to t_n$ jeweils ab, ob

$$\begin{aligned}
&(a) \quad c_S c_I h_n \rho_n > \mathrm{TOL}\,,\\
&(b) \quad \tfrac{1}{4}\,\mathrm{TOL} < c_S c_I h_n \rho_n \leq \mathrm{TOL}\,,\\
&(c) \quad c_S c_I h_n \rho_n \leq \tfrac{1}{4}\,\mathrm{TOL}\,.
\end{aligned}$$

Im Fall (a) wird die Schrittweite h_n halbiert, im Fall (b) beibehalten und im Fall (c) verdoppelt. Nach Erreichen des Endzeitpunkts $t_0 + T$ wird mit der gerade berechneten Näherung U erneut das duale Problem gelöst und eine verbesserter Wert für die Stabilitätskonstante c_S bestimmt. Mit diesem wird der ganze Gitteranpassungszyklus dann wiederholt. Dieser Prozess führt nach endlich vielen Schritten zu einer Äquilibrierung

der lokalen Fehlerindikatoren $c_S c_I h_n \rho_n$ über dem Intervall I und zur Reduzierung des Gesamtfehlers unter die Toleranz TOL:

$$c_S c_I \min_n \{h_n \rho_n\} \cong \|u(t_N) - U_N\| \cong c_S c_I \max_n \{h_n \rho_n\} \cong \text{TOL}\,,$$

wobei das erreichte Endgitter dann in gewissem Sinne „optimal“, d. h. sparsamst ist.

7.4.5 Vergleich zwischen dG- und Differenzen-Verfahren

Wir wollen die wesentlichen Unterschiede der beschriebenen residuen-basierten Fehler- und Schrittweitenkontrolle beim Galerkin- Verfahren zur traditionellen Schrittweitensteuerung bei Differenzen- Verfahren diskutieren. Dazu beschränken wir uns wieder auf den einfachsten Fall, nämlich das dG(0)-Verfahren angewendet für eine autonome AWA

$$u'(t) = f(u), \quad t \geq 0, \quad u(0) = u_0.$$

Wir betrachten das skalare Beispiel $f(x) = x^2$ mit $u_0 = 1$, wobei die exakte Lösung

$$u(t) = \frac{1}{1-t}$$

für $t \to 1$ singulär wird. Die Lipschitz-Konstante von $f(\cdot)$ entlang dieser Lösung verhält sich für $0 \leq t_n < 1$ wie $L(t_n) = 2(1-t_n)^{-1}$:

$$|f(x) - f(y)| = |x+y||x-y| \leq 2\max\{|x|,|y|\}|x-y|\,.$$

Für diesen Fall ist das dG(0)-Verfahren

$$([U]_{n-1}, \varphi) = \int_{I_n} (f(U), \varphi)\, dt \quad \forall \varphi \in P_0(I_n), \quad n \geq 1, \quad U_0^- = u_0,$$

äquivalent zum impliziten Euler-Verfahren

$$U_n = U_{n-1} + h_n f(U_n), \quad n \geq 1, \quad U_0 = u_0,$$

wenn man jeweils $U_n := U_n^-$ setzt und berücksichtigt, dass $U_n^- = U_{n-1}^+$.

dG(0)-Verfahren: Die Schrittweitenkontrolle beim dG(0)-Verfahren basiert in diesem einfachen Fall auf der a posteriori Fehlerschätzung

$$|e_N^-| \approx c_S c_I \max_{1 \leq n \leq N} \{|[U]_{n-1}|\}$$

mit der Interpolationskonstante $c_I = 1$ und der Stabilitätskonstante c_S aus der a priori Abschätzung für das duale Problem

$$\int_{t_0}^{t_N} |z'|\, dt \leq c_S\,.$$

Das Kriterium für die Wahl der lokalen Schrittweite ist

$$h_n = \frac{\mathrm{TOL}}{c_I c_S \rho_n}, \quad \rho_n = h_n^{-1} |[U]_{n-1}|. \tag{7.4.83}$$

Nehmen wir an, dass die verwendeten Schrittweiten bereits klein genug sind, so dass

$$h_n^{-1} |[U]_n| \approx h_n^{-1} |u_n - u_{n-1}| \approx \sup_{I_n} |u'|,$$

so wird

$$h_n \approx \left(c_S \sup_{I_n} |u'| \right)^{-1} \mathrm{TOL}.$$

Bei dem Testbeispiel ist die Stabilitätskonstante bestimmt durch das duale Problem (nach Linearisierung um die exakte Lösung u)

$$z'(t) = -\frac{2}{1-t} z(t), \quad 1 > t_n \geq 0, \quad z(t_N) = 1,$$

mit der Lösung

$$z(t) = \exp\left(2 \int_t^{t_n} \frac{ds}{1-s} \right) z(t_N) = \left(\frac{1-t}{1-t_n} \right)^2 z(t_N),$$

so dass

$$c_S \leq (1-t_N)^{-2}.$$

Dies ergibt unter Beachtung der Beziehung $\sup_{I_n} |u'| \approx (1-t_n)^{-2}$ und der Schrittweitenformel

$$h_n \approx (1-t_n)^2 (1-t_N)^2 \, \mathrm{TOL}, \tag{7.4.84}$$

als Konsequenz ein gleichmäßiges Fehlerverhalten

$$\sup_{[0,t_N]} |e| \approx \mathrm{TOL}, \quad 0 \leq t_N < 1.$$

Zur Abschätzung des Rechenaufwands zur Erzielung dieser Genauigkeit (ohne Berücksichtigung des zusätzlichen Aufwands zur Fehlerkontrolle) bestimmen wir die Anzahl N von Zeitschritten zur Erreichung des Endzeitpunkts t_N aus der Formel

$$N = \sum_{n=1}^{N} h_n h_n^{-1} \;\approx\; \frac{1}{\mathrm{TOL}} \frac{1}{(1-t_N)^2} \sum_{n=1}^{N} \frac{h_n}{(1-t_n)^2} \;\approx\; \frac{1}{(1-t_N)^3} \frac{1}{\mathrm{TOL}}. \tag{7.4.85}$$

Bei Annäherung an die Singularität erhöht sich der Aufwand für feste Fehlertoleranz TOL also kubisch mit dem reziproken Abstand zu $t = 1$.

Euler-Verfahren: Die übliche Schrittweitenkontrolle beim (impliziten) Euler-Verfahren basiert auf der a priori Fehlerabschätzung ($e_n = u(t_n) - U_n$, $e_0 = 0$)

$$|e_N| \leq K(t_N) \sum_{n=1}^{N} h_n |\tau_n(u)| \tag{7.4.86}$$

mit dem lokalen Abschneidefehler

$$|\tau_n(u)| = |h_n^{-1}(u_n - u_{n-1}) - f(u_n)| \leq \tfrac{1}{2} h_n \sup_{I_n} |u''|,$$

und einer Konstanten $K(t_N)$, die sich im allgemeinen Fall wie

$$K(t_N) \;\approx\; \exp\left(\int_0^{t_N} L(t)\,dt\right) \;\approx\; \exp\left(2\int_0^{t_N} \frac{dt}{1-t}\right) \;\approx\; \frac{1}{(1-t_N)^2}$$

verhält, wobei $L(t)$ wieder die Lipschitz-Konstante von $f(\cdot)$ entlang der exakten Lösung $u(t)$ ist. Sei TOL die zu garantierende Fehlertoleranz. Der lokale Abschneidefehler $|\tau_n|$ wird mit Hilfe eines Extrapolationsschritts über das Intervall I_n geschätzt. Nimmt man an, dass diese Schätzung exakt ist, so ergibt sich die neue Schrittweite dann gemäß

$$h_n \;\approx\; \frac{\mathrm{TOL}}{K(t_N)\sup_{I_n}|u''|}. \tag{7.4.87}$$

Hier muss die stark wachsende Stabilitätskonstante $K(t_n)$ auf jeden Fall mit berücksichtigt werden, um eine krasse Unterschätzung des Fehlers zu verhindern. Wegen $\sup_{I_n}|u''| \approx (1-t_n)^{-3}$ führt diese Schrittweitenkontrolle auf

$$h_n \;\approx\; (1-t_N)^2(1-t_n)^3\,\mathrm{TOL}. \tag{7.4.88}$$

Wir wollen aber nicht vergessen, dass i. Allg. diese Konstante nicht bekannt ist, und man üblicherweise rein heuristisch $K(t_N) = 1$ setzen würde. Vergleich der Schrittweitenformel (7.4.88) mit der entsprechenden Formel (7.4.84) für das dG(0)-Verfahren zeigt, dass erstere bei Annäherung von $t_n \to 1$ eine stärkere Reduzierung von h_n erzeugt, was sich natürlich auch in einem deutlich höheren numerischen Aufwand niederschlägt:

$$N \;\approx\; \frac{1}{(1-t_N)^2}\sum_{n=1}^{N} h_n \frac{1}{(1-t_n)^3}\,\frac{1}{\mathrm{TOL}^1} \;\approx\; \frac{1}{(1-t_N)^4}\frac{1}{\mathrm{TOL}}\,.$$

Im Gegensatz zur residuen-basierten Fehlerkontrolle des Galerkin-Verfahrens wächst hier der Aufwand mit der vierten Potenz des reziproken Abstandes zu $t = 1$. Dieser Effekt ließe sich durch eine Verfeinerung der Schrittweitenstrategie, die aber in den üblichen ODE-Codes nicht üblich ist, vermeiden. Dazu bestimmt man die lokale Schrittweite anstatt aus (7.4.87) gemäß der Formel

$$h_n^2 \;\approx\; \frac{\mathrm{TOL}}{N\,K(t_N)\sup_{I_n}|u''|}. \tag{7.4.89}$$

Dies ist wegen der Verwendung der (am Zeitpunkt t_n noch unbekannten) Gesamtanzahl N der Gitterpunkte eine *implizite* Strategie, die eigentlich keine vorwärtsschreitende Wahl der lokalen Schrittweiten h_n erlaubt und in jedem (globalen) Verfeinerungsschritt hinsichtlich der tatsächlichen Größe von N nachiteriert werden müsste. In der Praxis würde man aber, wenn die Verfeinerung in jedem Schritt nicht zu abrupt erfolgt, einfach den

Wert für N von der vorausgehenden Verfeinerungsstufe verwenden. Nach dieser Strategie ergibt sich die Schrittweitenformel

$$h_n \approx \left(\frac{\mathrm{TOL}}{N\,K(t_N)\sup_{I_n}|u''|}\right)^{1/2} \approx \left(\frac{\mathrm{TOL}}{N}\right)^{1/2}(1-t_N)(1-t_n)^{3/2}\,,$$

sowie die zugehörige Schrittzahl

$$N \approx \left(\frac{N}{\mathrm{TOL}}\right)^{1/2}\frac{1}{1-t_N}\sum_{n=1}^{N}h_n\frac{1}{(1-t_n)^{3/2}} \approx \left(\frac{N}{\mathrm{TOL}}\right)^{1/2}\frac{1}{(1-t_N)^{3/2}}\,.$$

bzw.

$$N \approx \frac{1}{\mathrm{TOL}}\frac{1}{(1-t_N)^3}\,.$$

Es ergibt sich nunmehr die gleiche Aufwandschätzung wie beim Galerkin-Verfahren. Dazu ist aber beim Differenzenverfahren eine hinreichend genaue Schätzung (durch Zusatzrechnung) der Abschneidefehler τ_n erforderlich, während beim Galerkin-Verfahren nur die leicht berechenbaren Residuen ρ_n eingehen.

Bemerkung: Ein alternativer Zugang zur a posteriori Fehlerschätzung beim Differenzenverfahren verwendet ein „diskretes“ Dualitätsargument ähnlich dem „kontinuierlichen“ Dualitätsargument beim Galerkin-Verfahren. Ausgangspunkt ist wieder die linearisierte Fehlergleichung

$$e_n = e_{n-1} + h_n f'(U_n)e_n + h_n\tau_n(u) + h_n O(e_n^2). \tag{7.4.90}$$

Sei Z_n die Lösung des rückwärts laufenden Euler-Schemas

$$Z_{n-1} = Z_n + h_n f'(U_n)Z_{n-1}, \quad 0 \le t_n < t_N, \tag{7.4.91}$$

mit Startwert Z_N. Damit gilt

$$\begin{aligned}(e_n, Z_n) &= (e_n, Z_n - Z_{n-1}) + (e_n - e_{n-1}, Z_{n-1}) + (e_{n-1}, Z_{n-1})\\ &= -h_n(e_n, f'(U_n)Z_{n-1}) + h_n(f'(U_n)e_n, Z_{n-1})\\ &\quad + h_n(\tau_n(u) + \mathcal{O}(e_n^2), Z_{n-1}) + (e_{n-1}, Z_{n-1}),\end{aligned}$$

and nach Summation für $1 \le n \le N$:

$$(e_N, Z_N) = (e_0, Z_0) + \sum_{n=1}^{N} h_n(\tau_n(u) + \mathcal{O}(e_n^2), Z_{n-1}).$$

Für $Z_N := e_N\|e_N\|^{-1}$ und $e_0 = 0$ erhalten wir

$$\|e_N\| \le c_{S,k}\sum_{n=1}^{N} h_n\big\{\|\tau_n(u)\| + \mathcal{O}(\|e_n\|^2)\big\}, \tag{7.4.92}$$

mit der *diskreten* Stabilitätskonstante $c_{S,k} := \max_{0\le n\le N-1} \|Z_n\|$. Vernachlässigung des quadratischen Fehlerterms ergibt dann

$$\|e_N\| \approx c_{S,k} \sum_{n=1}^{N} h_n \|\tau_n(U_n)\|, \tag{7.4.93}$$

mit der Approximation $\tau_n(U_n) \approx \tau_n(u)$ für den Abschneidefehler. Hier ist die i. Allg. zu pessimistische *a priori* Fehlerkonstante $K(t_N)$ ersetzt durch die *a posteriori* Stabilitätskonstante $c_{S,k}$. Die letztere wird in der Praxis aus der berechneten dualen Lösung Z_n zu einer Approximation des Startwerts $Z_N = e_N \|e_N\|^{-1}$ berechnet. Die auf der Abschätzung (7.4.93) basierende Schrittweitenkontrolle ist wieder *implizit* wie die für das Galerkin-Verfahren, liefert aber brauchbare Fehlerabschätzungen.

Für beide Verfahrensvarianten, Differenzen- oder Galerkin-Verfahren, kann die Effizienz der Rechnung durch Verwendung einer stärker lokalisierten Adaptionsstrategie noch weiter verbessert werden. Ausgangspunkt ist die „gewichtete“ a posteriori Fehlerabschätzung (7.4.79)

$$\|e_N^-\| \le \sum_{n=1}^{N} h_n^2 \rho_n \omega_n$$

mit den „Residuen“ und „Gewichten“

$$\rho_n := \sup_{I_n} |R(U)|, \quad \omega_n := h_n^{-1} \int_{I_n} |z - \tilde{P}_0 z|\, dt,$$

oder

$$\rho_n := h_n^{-1} |[U]_{n-1}|, \quad \omega_n := \frac{1}{2} \int_{I_n} |z'|\, dt.$$

Wie schon vorher berechnen wir $\rho_n \approx \sup_{I_n} |u'| \approx (1-t_n)^{-2}$ und

$$\omega_n = \frac{1}{2(1-T)^2} \int_{I_n} (1-t)\, dt \;\approx\; h_n \frac{1-t_n}{(1-T)^2}.$$

Basierend auf der obigen „gewichteten“ a posteriori Fehlerabschätzung ergibt sich die Schrittweitenwahl

$$h_n = \Big(\frac{\mathrm{TOL}}{N \rho_n \omega_n}\Big)^{1/2} \approx (1-t_n)^{1/2} (1-T) \frac{\mathrm{TOL}^{1/2}}{N^{1/2}}$$

Folglich wird

$$N \;=\; \sum_{n=1}^{N} h_n h_n^{-1} \;\approx\; \frac{N^{1/2}}{(1-T)\mathrm{TOL}^{1/2}} \sum_{n=1}^{N} \frac{h_n}{(1-t_n)^{1/2}} \;\approx\; \frac{N^{1/2}}{(1-T)\mathrm{TOL}^{1/2}},$$

und schließlich

$$N \approx \frac{1}{(1-T)^2} \frac{1}{\mathrm{TOL}}.$$

Dieses Beispiel zeigt deutlich, dass selbst bei sehr einfachen AWA nur durch sehr sorgfältige Berücksichtigung der Regularitäts- und Stabilitätseigenschaften des zugrunde liegenden Problems eine zuverlässige und effiziente Fehlerschätzung und Schrittweitensteuerung möglich ist. Der Galerkin-Ansatz hat dabei klare konzeptionelle Vorteile gegenüber dem Differenzenverfahren, auch wenn beide hier bei diesen extrem einfachen Beispielen bei richtiger Anwendung der Theorie auf ähnliche Ergebnisse führen.

7.5 Übungsaufgaben

Aufgabe 7.1: Das dG(1)-Verfahren nimmt angewendet auf die AWA

$$u'(t) = f(t, u(t)), \quad t \geq t_0, \quad u(t_0) = u_0, ,$$

die folgende Gestalt an:

$$U_n^- = \int_{I_n} f(t, U)\, dt + U_{n-1}^-, \quad U_n^- - U_{n-1}^+ = \frac{2}{h_n} \int_{I_n} f(t, U)(t - t_{n-1})\, dt,$$

für $n = 1, \ldots, N$. Dabei ist für $n \geq 2$ der Wert U_{n-1}^- vom vorausgehenden Intervall $I_{n-1} = (t_{n-2}, t_{n-1}]$ her bekannt, und U_n^-, U_{n-1}^+ sind die beiden zu bestimmenden Werte auf dem Intervall $I_n = (t_{n-1}, t_n]$. Für $n = 1$ ist $U_0^- := u^0$ der gegebene Anfangswert der Anfangswertaufgabe bei t_0.

i) Man konkretisiere dieses Verfahren mit der Setzung $y_n := U_n^-$ als implizites Einschrittverfahren für eine lineare, *autonome* AWA mit $f(t, x) = Ax + b$.

ii) Welche Ordnung hat dieses Verfahren?

iii) Wie sieht das zugehörige Stabilitätsintervall aus?

Aufgabe 7.2: Im Text wurde für Funktionen $v \in P_r(I_n), I_n = (t_{n-1}, t_n]$, mit Hilfe eines Transformationsarguments die diskrete „Sobolewsche Ungleichung“

$$\sup_{t \in I_n} |v(t)| \leq \kappa \Big(\int_{I_n} |v'(t)|^2 (t - t_{n-1})\, dt + |v_n^-|^2 \Big)^{1/2}$$

bewiesen. Man zeige mit derselben Argumentation die Variante

$$\sup_{t \in I_n} |v(t)| \leq \kappa \left(\int_{I_n} |v'(t)|\, dt + h_n^{-1} \Big| \int_{I_n} v(t)\, dt \Big| \right).$$

Welche von den beiden Abschätzungen gilt auch gleichmäßig für Funktionen $v \in C^1\big(\bar{I}_n\big)$? Man gebe eine Begründung.

Aufgabe 7.3: Im Text wurde gezeigt, wie auch das explizite Euler-Verfahren (Polygonzugmethode) über einen „unstetigen“ Galerkin-Ansatz gewonnen werden kann.

a) Man reproduziere die Herleitung des betreffenden expliziten unstetigen Galerkin-Verfahrens $\mathrm{dG}_{\exp}(0)$.

b) Man zeige, dass dieses $\mathrm{dG}_{\exp}$-Verfahren eine ähnliche a priori Fehlerabschätzung zulässt wie sein implizites Gegenstück dG(0):

$$\sup_{t\in I}|(u-U)(t)| \le \big(1+L_f T e^{L_f T}\big)\max_{1\le n\le N}\big\{h_n \sup_{t\in I_n}|u'(t)|\big\}.$$

c) Man stelle das entsprechende $\mathrm{dG}_{\exp}(1)$-Verfahren auf. Lässt sich hieraus durch Einsatz geeigneter Quadraturformeln ein explizites Differenzenschema zweiter Ordnung ableiten?

Aufgabe 7.4: Sei I_n eines der (halboffenen) Teilintervalle von $I=[t_0,t_0+T]$ mit Länge $h_n=t_n-t_{n-1}$. Die sog. L^2-Projektion $\pi_r u\in P_r(I_n)$ einer Funktion $u\in C(\bar I_n)$ auf den Polynomraum $P_r(I_n)$ ist definiert durch

$$\int_{I_n}\big\{u(t)-\pi_r u(t)\big\}\varphi(t)\,dt=0 \quad \forall\varphi\in P_r(I_n).$$

Für $u\in C^{r+1}(\bar I_n)$ gilt hierbei die Fehlerabschätzung

$$\sup_{t\in I_n}|(u-\pi_r u)(t)| \le \frac{1}{(r+1)!}h_n^{(r+1)}\sup_{t\in I_n}|u^{(r+1)}(t)|.$$

(i) Man beweise diese Aussage für die Spezialfälle $r=0,1$.

(ii) Man zeige, dass Funktionen $f\in C(\bar I_n)$ mit der Eigenschaft

$$\int_{I_n} f(t)\varphi(t)\,dt=0 \quad \forall\varphi\in P_r(I_n),$$

mindestens $r+1$ reellwertige Nullstellen besitzen, und beweise damit die obige Fehlerabschätzung für alle $r\in\mathbb{N}_0$.

Aufgabe 7.5 (praktische Aufgabe): Man approximiere die AWA

$$u'(t)=u(t)^2,\quad 0\le t<1,\quad u(0)=1,$$

mit der (singulären) Lösung $u(t)=(1-t)^{-1}$ mit Hilfe des „unstetigen“ dG(1)- sowie des „stetigen“ cG(1)-Verfahrens. Die a priori Fehlerabschätzungen aus dem Text garantieren für beide Verfahren das Konvergenzverhalten $\max_I|U-u|=\mathcal{O}(h^2)$. Man überprüfe die Vorhersage, dass für das dG(1)-Verfahren in den diskreten Zeitgitterpunkten t_n die höhere Konvergenzordnung $|(U-u)(t_n)|=\mathcal{O}(h^3)$ besteht.

Aufgabe 7.6: Betrachtet werde die AWA

$$u'(t)=u(t)^2,\quad 0\le t\le T<1,\quad u(0)=1,$$

mit der (singulären) Lösung $u(t)=(1-t)^{-1}$. Man konkretisiere hierfür die im Text angegebene a priori Fehlerabschätzung für das DG(0)-Verfahren und vergleiche sie mit dem entsprechenden Resultat für das implizite Euler-Verfahren. Wie wächst in beiden Fällen der Aufwand (gemessen in der Anzahl der Auswertungen von f) in Abhängigkeit von der vorgegebenen Toleranz bei Annäherung an die kritische Stelle $T\to 1$?

Aufgabe 7.7 (Praktische Aufgabe): Man approximiere die 3-d, steife AWA

$$u'(t) = Au(t), \quad t \geq 0, \quad u(0) = (1, 0, -1)^T,$$

mit der Systemmatrix

$$A = \begin{pmatrix} -21 & 19 & -20 \\ 19 & -21 & 20 \\ 40 & -40 & -40 \end{pmatrix}$$

und der Lösung

$$\begin{aligned} u_1(t) &= \frac{1}{2}e^{-2t} + \frac{1}{2}e^{-40t}\{\cos(40t) + \sin(40t)\}, \\ u_2(t) &= \frac{1}{2}e^{-2t} - \frac{1}{2}e^{-40t}\{\cos(40t) + \sin(40t)\}, \\ u_3(t) &= -\frac{1}{2}e^{-40t}\{\cos(40t) - \sin(40t)\} \end{aligned}$$

mit Hilfe (a) der Trapezregel und (b) des cG(1)-Verfahrens. Zu berechnen ist der Wert $u(2)$ auf 6 wesentliche Dezimalstellen. Man versuche, in beiden Fällen möglichst sparsam zu arbeiten. Mit welchem Verfahren lässt sich diese Aufgabe am effizientesten, d. h. in geringster Zeit, lösen?

Aufgabe 7.8: Das cG(1)-Verfahren verwendet auf einer Zerlegung des Lösungsintervalls $I = [t_0, t_0 + T]$ stetige, stückweise lineare Ansatzfunktionen und unstetige, stückweise konstante Testfunktionen.

Man führe mit Hilfe der Vorgehensweise aus dem Text eine a posteriori Fehleranalyse für das cG(1)-Verfahren durch. Abgeschätzt werden soll der Endzeitfehler $\|U_N - u(t_N)\|$. Zur Vereinfachung konzentriere man sich auf den Fall einer linearen autonomen AWA,

$$u'(t) = Au(t) + b, \quad t \geq 0, \quad u(0) = u_0,$$

für welche das cG(1)-Verfahren äquivalent zur „Trapezregel“ ist.

Aufgabe 7.9: Man betrachte die approximative Lösung der autonomen AWA

$$u'(t) = u(t)^2, \quad 0 \leq t \leq T < 1, \quad u(0) = 1,$$

mit der (singulären) Lösung $u(t) = (1-t)^{-1}$ mit Hilfe a) des impliziten Euler-Verfahrens und b) des dG(0)-Verfahrens. Es soll der Wert $u(T)$, $T = t_N$, berechnet werden unter Verwendung einer adaptiven Schrittweitensteuerung. Was ist der asymptotische Aufwand $N = N(\mathrm{TOL}, T)$ der beiden Verfahren in Abhängigkeit von der vorgegebenen Fehlertoleranz bei Annäherung $T \to 1$?

a) Für das implizite Euler-Verfahren

$$U_n = U_{n-1} + h_n U_n^2, \quad n \geq 1, \quad U_0 = 1.$$

folgt aus der für allgemeine Einschrittverfahreb im Text hergeleiteten a priori Fehlerabschätzung die Fehlerschranke

$$|e_N| \leq K(T) \sum_{n=1}^{N} h_n \tau_n(u), \qquad \tau_n(u) = h_n \tau_n^0(u) + \mathcal{O}(h_n^2),$$

Dies induziert die (implizite) Schrittweitenstrategie

$$h_n \;\approx\; \left(\frac{\mathrm{TOL}}{N\, K(T) |\tau_n^0(u)|} \right)^{1/2}.$$

b) Die Schrittweitenwahl in der dG(0)-Methode basiert auf der „gewichteten“ a posterori Fehlerabschätzung

$$|e_N| \;\leq\; c_I \sum_{n=1}^{N} h_n \rho_n \omega_n,$$

und folgt der (impliziten) Strategie

$$h_n \approx \frac{\mathrm{TOL}}{N \rho_n \omega_n}.$$

Aufgabe 7.10 (Praktische Aufgabe): Man löse die AWA

$$u'(t) = u(t)^2, \quad 0 \leq < 1, \quad u(0) = 1,$$

mit der (singulären) Lösung $u(t) = (1-t)^{-1}$ mit Hilfe des dG(0)- und des impliziten Euler-Verfahrens bei Verwendung der jeweiligen Strategien zur Schrittweitensteuerung aus Augabe 7.9. Die einzuhaltende Fehlertoleranz ist $\varepsilon = 10^{-3}$. Man versuche, möglichst weit an die singuläre Stelle $t = 1$ heranzurechnen. Wie entwickeln sich bei den beiden Verfahren die Schrittweiten für $t \to 1$?