INSTITUT FÜR ANGEWANDTE UND
NUMERISCHE MATHEMATIK

TECHNISCHE UNIVERSITÄT WIEN

Report Nr. 113/94

Beschleunigte Algorithmen zur Lösung von singulären Randwertaufgaben

W. Auzinger
P. Trieb
E. Weinmüller

Inhaltsverzeichnis

1 Einleitung

Es gibt eine Vielzahl von Problemstellungen in der Chemie, der Physik und der Mechanik, die durch singuläre Randwertprobleme beschrieben werden. Typischerweise treten singuläre Probleme auf, falls (geometrische) Symmetrien vorhanden sind und deshalb Systeme partieller Differentialgleichungen auf Systeme gewöhnlicher Differentialgleichungen transformiert werden können. Das große Interesse an der numerischen Lösung dieser Systeme ist durch ihren starken Anwendungsbezug erklärt, deshalb gibt es bereits umfangreiche Literatur, in der verschiedene Näherungsmethoden auf ihre Tauglichkeit zur Behandlung singulärer Probleme untersucht werden.

Wir beschäftigen uns in weiterer Folge mit linearen Randwertproblemen zweiter Ordnung mit einer Singularität erster Art:

$$y''(t) - \frac{A_1(t)}{t} y'(t) - \frac{A_0(t)}{t^2} y(t) = f(t), \qquad 0 < t \leq 1, \tag{1.1a}$$

$$B_0 Y(0) + B_1 Y(1) = \beta, \tag{1.1b}$$

wo $Y(t) = (y(t), y'(t))^T$. Hier sind y und f vektorwertige Funktionen der Dimension n, $A_0(t)$ und $A_1(t)$ sind reelle $n \times n$-Matrizen, B_0 und B_1 sind konstante reelle $2n \times 2n$-Matrizen und β ist ein konstanter $2n$-dimensionaler Vektor.

Die numerische Lösbarkeit von skalaren Gleichungen dieses Typs mittels Differenzenverfahren wurde von mehreren Autoren betrachtet, siehe Jamet [7], Natterer [9] sowie Russel und Shampine [12]. Brabston [3] und de Hoog und Weiss [6] haben Systeme von gewöhnlichen Differentialgleichungen erster Ordnung mit einer Singularität des betrachteten Typs untersucht. Das klassische Drei-Punkt Differenzenverfahren zur numerischen Lösung von (1.1) wurde von Weinmüller für den linearen Fall in [14] und für den nichtlinearen Fall in [15] diskutiert, Kollokationsverfahren in Anwendung auf (1.1) wurden von derselben Autorin in [16] betrachtet.

In obigen Verfahren sind hohe Genauigkeitsanforderungen meist mit sehr hohem Aufwand verbunden, eine Tatsache, die speziell in den nichtlinearen Anwendungsbeispielen aus der Mechanik und der Chemie deutlich zu Tage tritt (Beispiele hierzu etwa in Keller und Wolfe [8], Parter, Stein und Stein [10] und Rentrop [11]). *Beschleunigte Algorithmen* besitzen demgegenüber den erheblichen Vorteil, eine sehr genaue Lösung bei vernünftigem Aufwand zu liefern. In dieser Arbeit verwenden wir eine einfache Variante der sogenannten *Iterierten Defekt-Korrektur* (IDeC), die auf einer Idee von Zadunaisky [17] basiert. Zunächst wird mittels Differenzenverfahren eine Basislösung auf einem äquidistanten Gitter auf $[0,1]$ ermittelt, diese wird anschließend durch ein iteratives Verfahren, das in jedem Schritt eine Steigerung des Genauigkeitsniveaus bewirkt, weiter verbessert.

Das Ziel dieser Arbeit besteht darin, die Anwendung dieses Verfahrens zur numerischen Lösung des Problems (1.1) zu studieren. Konvergenzordnung, Genauigkeitsniveau und Fixpunktaspekte werden anhand verschieden „schwieriger" Modellprobleme experimentell untersucht, wobei diese durch unterschiedliche Parameterwahl beliebig unangenehm gestaltet werden können. Die erzielten Ergebnisse werden mit theoretischen Resultaten für nichtsinguläre Randwertprobleme verglichen; signifikante Abweichungen werden diskutiert und dienen als Anlaß für Modifikationen des Verfahrens.

Die Arbeit ist wie folgt organisiert: Der Abschnitt 2 gilt theoretischen Grundlagen. In Abschnitt 2.1 wird die Notation eingeführt und in 2.2 die Ergebnisse der Existenz- und Eindeutigkeitstheorie stetiger Lösungen von (1.1) vorgestellt. Wir diskutieren die Stabilitäts- und Konvergenzresultate der Differenzenmethode für singuläre Probleme in Abschnitt 2.3. Das Prinzip der IDeC-Methode wird in Abschnitt 3 dargelegt und die Konvergenzresultate für klassische Randwertprobleme zweiter Ordnung in Abschnitt 3.3. Die Ergebnisse der numerischen Experimente werden anschließend in Abschnitt 4 präsentiert und kommentiert.

2 Theoretische Grundlagen

2.1 Notationen

Wir bezeichnen mit $\mathbf{C}^n$ den Raum der komplexen Vektoren $x = (\xi_1, \xi_2, \ldots, \xi_n)^T$ der Dimension n und schreiben $|\cdot|$ für die Maximumnorm auf $\mathbf{C}^n$,

$$|x| = |(\xi_1, \xi_2, \ldots, \xi_n)^T| := \max_{1 \le j \le n} |\xi_j|.$$

Mit $C^p[0,1]$ bezeichnen wir sowohl den Raum der auf $[0,1]$ p-mal stetig differenzierbaren vektorwertigen Funktionen als auch den Raum der komplexwertigen Matrizen, deren Einträge p-mal stetig differenzierbar sind. $C^p(0,1]$ wird analog definiert. Für jeden Vektor $y \in C^0[0,1]$ erklären wir die Norm

$$\|y\| := \max_{0 \le t \le 1} |y(t)|.$$

Wir schreiben $C = C[0,1] = C^0[0,1]$ und $C(0,1] = C^0(0,1]$. Für alle $y \in C$ definieren wir den Stetigkeitsmodul als

$$\omega(y;\delta) = \max_{0 \le t \le 1-\delta} |y(t+\delta) - y(t)|.$$

Sei Δ ein äquidistantes Gitter auf $[0,1]$ mit Schrittweite h,

$$\Delta = \Delta_{[0,1]} = \{t_\nu, \nu = 0(1)N | t_\nu = \nu h, t_N = 1\},$$

wo $\nu = 0(1)N$ für $\nu = 0, 1, \ldots, N$ steht. Mit jeder Partition Δ verbinden wir den linearen Raum X_Δ mit den Elementen

$$x_\Delta = (x_0, x_1, \ldots, x_N),$$

wo $x_\nu = (\xi_{\nu 1}, \xi_{\nu 2}, \ldots, \xi_{\nu n})^T \in \mathbf{C}^n$, $\nu = 0(1)N$. Die Norm auf X_Δ wird durch

$$\|x_\Delta\| := \max_{0 \le \nu \le N} |x_\nu|$$

definiert. Mit $R_\Delta : C \to X_\Delta$ bezeichnen wir schließlich die Projektion

$$R_\Delta y = (y(t_0), y(t_1), \ldots, y(t_N)), \qquad R_\Delta y' = (y'(t_0), y'(t_1), \ldots, y'(t_N)).$$

2.2 Analytische Resultate

Die grundlegenden analytischen Eigenschaften von (1.1) wurden, sowohl im linearen wie auch im nichtlinearen Fall, von Weinmüller in [13] untersucht. In diesem Abschnitt fassen wir die Eigenschaften von Lösungen des linearen Problems zusammen, d. h. wir geben die Bedingungen für die Existenz und Eindeutigkeit von stetigen Lösungen an.

Gegeben sei das lineare Randwertproblem

$$y''(t) - \frac{A_1(t)}{t} y'(t) - \frac{A_0(t)}{t^2} y(t) = f(t), \qquad 0 < t \le 1, \tag{2.1a}$$

$$B_0 Y(0) + B_1 Y(1) = \beta, \qquad Y(t) = (y(t), y'(t))^T. \tag{2.1b}$$

Durch die Anwendung der linearen Transformation

$$v(t) = (v_1(t), v_2(t))^T := (y(t), ty'(t))^T$$

auf das System (2.1) erhalten wir das folgende Randwertproblem erster Ordnung:

$$\begin{aligned} v'(t) &= \frac{1}{t}\begin{pmatrix} 0 & I \\ A_0(t) & I + A_1(t) \end{pmatrix} v(t) + t\begin{pmatrix} 0 \\ f(t) \end{pmatrix} = \\ &=: \frac{1}{t} M(t)v(t) + t\mathring{f}(t), \qquad 0 < t \le 1, \end{aligned} \tag{2.2a}$$

$$B_0 Y(0) + B_1 Y(1) = \beta. \tag{2.2b}$$

Betrachten wir nun den Fall $f \in C$ und $A_0(t)$, $A_1(t)$ der Form

$$A_0(t) = A_0 + tC_0(t), \qquad A_1(t) = A_1 + tC_1(t). \tag{2.3}$$

A_0, A_1 bezeichnen hier konstante $n \times n$-Matrizen und C_0, $C_1 \in C$. Mit (2.3) wird das System (2.2) äquivalent zu

$$v'(t) = \frac{1}{t} Mv(t) + \mathring{C}(t)v(t) + t\mathring{f}(t), \qquad 0 < t \le 1, \tag{2.4a}$$

$$B_0 Y(0) + B_1 Y(1) = \beta, \tag{2.4b}$$

wobei

$$M = M(0) = \begin{pmatrix} 0 & I \\ A_0 & I + A_1 \end{pmatrix}, \qquad \mathring{C}(t) = \begin{pmatrix} 0 & 0 \\ C_0(t) & C_1(t) \end{pmatrix}. \tag{2.4c}$$

Um die Struktur der allgemeinen Lösung von (2.1a) zu beschreiben, betrachten wir zunächst das skalare Problem mit konstanten Koeffizienten

$$y''(t) - \frac{a_1}{t} y'(t) - \frac{a_0}{t^2} y(t) = f(t), \qquad 0 < t \le 1, \tag{2.5}$$

a_0, $a_1 \in \mathbb{R}$, und das dazugehörige System erster Ordnung

$$v'(t) = \frac{1}{t} Mv(t) + t\mathring{f}(t), \qquad 0 < t \le 1, \tag{2.6a}$$

wobei

$$M = \begin{pmatrix} 0 & 1 \\ a_0 & 1 + a_1 \end{pmatrix}, \qquad \mathring{f}(t) = \begin{pmatrix} 0 \\ f(t) \end{pmatrix}. \tag{2.6b}$$

Die Lösung dieses Problems läßt sich wie folgt konstruieren: Für die allgemeine Lösung der homogenen Gleichung (2.5), $f(t) \equiv 0$, machen wir den Ansatz

$$y_h(t) := t^\lambda, \quad \lambda \in \mathbb{C}. \tag{2.7}$$

Einsetzen in (2.5) ergibt

$$\begin{aligned} 0 &= \lambda(\lambda-1)t^{\lambda-2} - \frac{a_1}{t}\lambda t^{\lambda-1} - \frac{a_0}{t^2}t^{\lambda} = \\ &= \lambda(\lambda-1)t^{\lambda-2} - a_1\lambda t^{\lambda-2} - a_0 t^{\lambda-2} = \\ &= t^{\lambda-2}(\lambda(\lambda-1) - a_1\lambda - a_0), \qquad 0 < t \le 1, \end{aligned}$$

was offensichtlich äquivalent ist zur Forderung

$$\lambda^2 - \lambda(1+a_1) - a_0 = 0. \tag{2.8}$$

Somit gilt für $y_h(t)$,

$$y_h(t) = \begin{cases} c_1 t^{\lambda_1} + c_2 t^{\lambda_2}, & \text{falls } \lambda_1 \neq \lambda_2, \\ c_1 t^{\lambda} + c_2 t^{\lambda} \ln t, & \text{falls } \lambda_1 = \lambda_2 = \lambda, \end{cases} \tag{2.9}$$

und die allgemeine Lösung von (2.5) ist

$$y(t) = y_h(t) + y_p(t),$$

wobei $y_p(t)$ eine Partikulärlösung der inhomogenen Gleichung (2.5) ist. Wir nützen nun den Zusammenhang zwischen (2.5) und (2.6a), um $y_p(t)$ zu konstruieren. Dazu bemerken wir, daß die Lösungen λ_1, λ_2 von (2.8) auch die Eigenwerte der Matrix M sind:

$$p_M(\lambda) := \det(M - \lambda I) = \begin{vmatrix} -\lambda & 1 \\ a_0 & 1 + a_1 - \lambda \end{vmatrix}.$$

Dies legt den Übergang zur Jordan'schen Normalform J von M nahe. Mit anderen Worten: Wir versuchen, das System (2.6a) zu entkoppeln. Es sei E die 2×2-Matrix der verallgemeinerten Eigenvektoren von M, so gilt $M = EJE^{-1}$. Es sei $w(t) = E^{-1}v(t)$, so ist $w(t)$ die Lösung von

$$w'(t) = \frac{1}{t} J w(t) + t\mathring{g}(t), \tag{2.10}$$

wobei $\mathring{g}(t) = E^{-1}\mathring{f}(t)$. Mit der 2×2-Matrix-Lösung

$$t^J := \exp(J \ln t)$$

des homogenen Systems

$$W_h'(t) = \frac{1}{t} J W_h(t), \qquad 0 < t \le 1, \tag{2.11a}$$

$$W_h(1) = I, \tag{2.11b}$$

kann man die Lösung des homogenen Systems (2.10) wie folgt darstellen:

$$w_h(t) = t^J \tilde{c}, \quad \tilde{c} \in \mathbb{C}^2, \qquad 0 < t \le 1,$$

wobei $\tilde{c}$ ein beliebiger konstanter Vektor ist. Für t^J gilt:

$$J = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \quad \text{und} \quad t^J = \begin{pmatrix} t^{\lambda_1} & 0 \\ 0 & t^{\lambda_2} \end{pmatrix}, \quad \text{falls } \lambda_1 \neq \lambda_2,$$

und

$$J = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix} \quad \text{und} \quad t^J = t^\lambda \begin{pmatrix} 1 & \ln t \\ 0 & 1 \end{pmatrix}, \quad \text{falls } \lambda_1 = \lambda_2 = \lambda,$$

da im skalaren Fall der doppelte Eigenwert λ stets geometrische Vielfachheit 1 hat. Damit ist die allgemeine Lösung des inhomogenen Problems (2.10)

$$\begin{aligned} w(t) &= w_h(t) + w_p(t) = \\ &= t^J \tilde{c} + t^J \int_1^t s^{-J} s \mathring{g}(s)\, ds = \\ &= t^J \tilde{c} + t^J \int_1^t s^{-J} s E^{-1} \mathring{f}(s)\, ds, \end{aligned} \tag{2.12}$$

woraus

$$\begin{aligned} v(t) &= Ew(t) = \underbrace{Et^J E^{-1} E\tilde{c}}_{v_h(t)} + \underbrace{Et^J E^{-1} \int_1^t Es^{-J} E^{-1} s\mathring{f}(s)\, ds}_{v_p(t)} = \\ &= t^M c + t^M \int_1^t s^{-M} s\mathring{f}(s)\, ds, \qquad 0 < t \le 1, \end{aligned} \tag{2.13}$$

mit $c = E\tilde{c} \in \mathbb{C}^2$ beliebig, folgt. Dabei ist

$$Et^J E^{-1} = E \exp(J \ln t) E^{-1} = \exp(EJE^{-1} \ln t) = \exp(M \ln t) =: t^M.$$

Diese Lösung erfüllt die Anfangsbedingung $v(1) = c$ und die Lösung von (2.5) ist $y(t) = v_1(t)$. Man sieht auch sofort, daß die erste Komponente von $v_h(t) = t^M c$ die Struktur von $y_h(t)$ in (2.9) hat. Durch die Vorgabe von $v(1) = \gamma = (\gamma_1, \gamma_2)^T$ kann man die Lösung in (2.13) eindeutig festlegen, d. h. wir haben damit die Lösung $y(t)$ von (2.5) konstruiert, die die Anfangsbedingungen $y(1) = \gamma_1, y'(1) = \gamma_2$ erfüllt. Diese Lösung ist i. a. an der Stelle $t = 0$ unbeschränkt, vgl. (2.9). Es ist aber stets möglich, eine auf $[0, 1]$ stetige Lösung durch richtige Vorgabe der Randbedingungen zu erhalten. Dies wollen wir an dem Spezialfall $\lambda_1, \lambda_2 \in \mathbb{R}$ und anhand der Lösung der homogenen Gleichung präsentieren. Die nachfolgende Tabelle gibt die verschiedenen Eigenwertstrukturen, die für die Stetigkeit der Lösungen notwendigen Anfangsbedingungen, und die stetigen Lösungen wieder.

Fall 1: $\lambda_1 \neq \lambda_2, \quad y_h(t) = c_1 t^{\lambda_1} + c_2 t^{\lambda_2}$

$\lambda_1 < \lambda_2 < 0:$	$y_h(0) = y_h'(0) = 0,$	$y_h(t) \equiv 0$
$\lambda_1 < 0, \ \lambda_2 = 0:$	$y_h'(t) = 0,$	$y_h(t) = c_2$
$\lambda_1 < 0 < \lambda_2:$	$y_h(0) = 0,$	$y_h(t) = c_2 t^{\lambda_2}$
$\lambda_1 > 0, \ \lambda_2 = 0:$		$y_h(t) = c_1 t^{\lambda_1} + c_2$
$\lambda_1 > \lambda_2 > 0:$		$y_h(t) = c_1 t^{\lambda_1} + c_2 t^{\lambda_2}$

Fall 2: $\lambda_1 = \lambda_2 = \lambda, \quad y_h(t) = c_1 t^\lambda + c_2 t^\lambda \ln t$

$\lambda < 0:$	$y_h(0) = y_h'(0) = 0,$	$y_h(t) \equiv 0$
$\lambda = 0:$	$y_h'(0) = 0,$	$y_h(t) = c_1$
$\lambda > 0:$		$y_h(t) = c_1 t^\lambda + c_2 t^\lambda \ln t$

Aus der obigen Tabelle ist ersichtlich, wie die Stetigkeit der höheren Ableitungen von y_h von der Größe der positiven Eigenwerte von M abhängt.

Diese Vorgangsweise zur Konstruktion einer stetigen Lösung kann man auf ein System mit einer Matrix M, deren Struktur komplizierter ist und wo J aus mehreren Jordanblöcken besteht, erweitern. Auch im Fall variabler Koeffizienten ist es möglich, die entsprechenden Existenz- und Eindeutigkeitsaussagen zu formulieren, vgl. [13].
Aufgrund des Einflusses der Eigenwerte der Matrix M auf das Verhalten der Lösung von (2.4) benützen wir Spektralprojektionen, um die allgemeine stetige Lösung von (2.4a) darzustellen. Mit R bezeichnen wir die Spektralprojektion auf den zum Eigenwert $\lambda = 0$ gehörenden Eigenraum von M, und S sei die Projektion auf den invarianten Unterraum von M, der zu den Eigenwerten mit positivem Realteil gehört. Wir setzen

$$P := R + S, \qquad Q := I - P.$$

Dann gilt der folgende

Satz 2.1 *Für jedes $f \in C$ und für jeden konstanten $2n$-dimensionalen Vektor γ hat das lineare Randwertproblem*

$$v'(t) = \frac{1}{t} M v(t) + \mathring{C}(t)v(t) + t\mathring{f}(t), \qquad 0 < t \leq 1, \tag{2.14a}$$

$$Qv(0) = 0, \qquad Pv(1) = P\gamma, \tag{2.14b}$$

eine eindeutige und stetige Lösung.

Mit $y(t) := v_1(t)$ erhalten wir eine Lösung $y(t)$ von (2.1a), *und $y \in C \cap C^2(0,1]$.*

Das System $Qv(0) = 0$ liefert $q = \text{Rang}[Q]$ linear unabhängige Gleichungen, die für die Stetigkeit von v notwendig sind. Die restlichen $p = \text{Rang}[P]$ linear unabhängigen Gleichungen, die die Lösung v eindeutig festlegen, sind durch $Pv(1) = P\gamma$ gegeben.

Man kann zeigen, dazu nützt man die spezielle Struktur von M, daß man $Qv(0) = 0$ auch äquivalenterweise durch $y(0)$ und $y'(0)$ ausdrücken kann, nämlich als

$$(A_1(0) + A_0(0))y'(0) = 0, \qquad A_0(0)y(0) = 0, \tag{2.15}$$

vgl. [14].

Weiters kann man Bedingungen dafür angeben, daß die Lösung $y(t) = v_1(t)$, die das Randwertproblem (2.14) löst, auch die Randbedingungen (2.1b) erfüllt, siehe dazu [13].

Abschließend formulieren wir noch die Glattheitseigenschaften von y.

Satz 2.2 *Seien f und $C_1 \in C^p$ und $C_0 \in C^{p+1}$, $p \geq 0$. Dann gilt:*

$y \in C^{p+2}$,	*wenn kein Eigenwert von M positiven Realteil hat,*
$y \in C^p \cap C^{p+1}(0,1]$,	*wenn* $p < \sigma_+ \leq p+1$,
$y \in C^{p+1} \cap C^{p+2}(0,1]$,	*wenn* $p+1 < \sigma_+ \leq p+2$,
$y \in C^{p+2}$,	*wenn* $\sigma_+ > p+2$,

wobei σ_+ den kleinsten positiven Realteil der Eigenwerte von M bezeichnet. Weiters ist $y \in C^{p+1} \cap C^{p+2}(0,1]$, *falls* $\sigma_+(=\lambda) = 1$.

Die Bedingungen (2.15) werden in den Anwendungen tatsächlich erfüllt, siehe etwa [2], [8], [11]. In der Schalentheorie ist $A_1(0) = -I$ und $A_0(0) = I$, die Randbedingung lautet daher $y(0) = 0$. Das Spektrum von M ist $\sigma(M) = \{\pm 1\}$. In der Reaktortheorie und in der Astrophysik gilt $A_1(0) = -2I$, $A_0(0) = 0$, $\sigma(M) = \{0, -1\}$ und somit $y'(0) = 0$. In beiden Fällen ist die algebraische Vielfachheit der Eigenwerte gleich n.

2.3 Das numerische Verfahren

Betrachten wir ein Gitter Δ wie im Abschnitt 2.1, so lautet die gewöhnliche Drei-Punkt Diskretisierung für (2.1)

$$\frac{y_{\nu+1} - 2y_\nu + y_{\nu-1}}{h^2} - \frac{A_1(t_\nu)}{t_\nu}\left(\frac{y_{\nu+1} - y_{\nu-1}}{2h}\right) - \frac{A_0(t_\nu)}{t_\nu^2} y_\nu = f(t_\nu), \qquad \nu = 1(1)N-1, \tag{2.16a}$$

$$B_0 Y_0 + B_1 Y_N = \beta, \tag{2.16b}$$

wo $Y_0 = (y_0, y_0')^T$ und $Y_N = (y_N, y_N')^T$. Mit y_0' bzw. y_N' bezeichnen wir dabei die Näherungswerte für $y'(0)$ bzw. $y'(1)$. Ohne Beschränkung der Allgemeinheit nehmen wir an, daß die für die Stetigkeit der Lösung von (2.1) notwendigen Bedingungen durch

$$\tilde{Q} Y_0 = 0 \tag{2.16c}$$

gegeben sind, wo $\tilde{Q}$ eine konstante $q \times 2n$-Matrix mit $q = \text{Rang}[Q]$, vgl. Abschnitt 2.2, ist. Das Gleichungssystem (2.16b) liefert dann noch die fehlenden $p = \text{Rang}[P]$ linear unabhängigen Bedingungen, die die Lösung eindeutig festlegen, $p + q = 2n$. Zur Approximation von $y'(1)$ setzen wir $y_N' = (y_{N+1} - y_{N-1})/2h$ und ergänzen das Differenzenschema (2.16a) durch eine weitere Gleichung für $\nu = N$. Die Wahl von y_0' ist nicht ganz so einfach, da wir den Punkt $t_{-1} = -h$ nicht in das Differenzenschema integrieren können. Ein möglicher Ausweg ist die Verwendung eines Drei-Punkt unsymmetrischen Differenzenquotienten,

$$y_0' = \frac{-y_2 + 4y_1 - 3y_0}{2h}.$$

Falls $y \in C^3$, so ist das ebenso eine $O(h^2)$-Näherung für $y'(0)$.

Sei d_0 die Dimension des größten Jordanblockes, der zum Eigenwert $\lambda = 0$ von M gehört. Bezeichnen wir mit $\lambda_+ = \sigma_+ + i\kappa_+$ denjenigen Eigenwert von M, der den kleinsten positiven Realteil σ_+ besitzt, so steht d_+ für die Dimension des größten Jordanblockes zu λ_+. Es gilt der folgende

Satz 2.3 *Nehmen wir an, daß das homogene Randwertproblem* (2.1) *nur die triviale Lösung besitzt. Dann hat das Differenzenschema* (2.16a) *mit den Randbedingungen* (2.16b) *und* (2.16c) *eine eindeutige Lösung für jedes f und jedes β, sofern h nur genügend klein gewählt ist, und es gilt die folgende Abschätzung:*

$$\|y_\Delta\| \leq \text{const}\{|\beta| + \|f_\Delta\|\}.$$

Sei y_Δ eine Lösung von (2.16).

(a) Falls f, $C_1 \in C$ und $C_0 \in C^1$, dann gilt

$$\|y_\Delta - R_\Delta y\| \leq \begin{cases} \text{const}\{h^{\sigma_+}|\ln h|^{d_+-1} + \omega(f,h)\}, & 0 < \sigma_+ \leq 1, \\ \text{const}\{h|\ln h|^{d_+-1} + \omega(f,h)\}, & 1 < \sigma_+ < 2, \\ \text{const}\{h|\ln h|^{d_+} + \omega(f,h)\}, & \sigma_+ = 2, \\ \text{const}\{h + \omega(f,h)\}, & \sigma_+ > 2 \text{ oder } S = 0. \end{cases}$$

(b) Falls f, $C_1 \in C^2$ und $C_0 \in C^3$, dann gilt

$$\|y_\Delta - R_\Delta y\| \leq \begin{cases} \text{const } h^{\sigma_+}|\ln h|^{d_+-1}, & 0 < \sigma_+ < 2, \\ \text{const } h^2(|\ln h|^{d_+} + |\ln h|^{d_0-1}), & \sigma_+ = 2, \\ \text{const } h^2|\ln h|^{d_0-1}, & \sigma_+ > 2 \text{ oder } S = 0. \end{cases}$$

In beiden Fällen hängt die Konstante zwar von den Problemdaten (etwa $A_0(t), A_1(t)$), nicht aber von h ab. Falls $\lambda = 0$ nicht zum Spektrum von M gehört, muß man $|\ln h|^{d_0-1}$ durch 1 ersetzen.

Den Beweis findet man in [14], die analoge Aussage für den nichtlinearen Fall in [15]. In beiden Arbeiten wird anhand von Modellproblemen gezeigt, daß die Abschätzungen scharf sind und daß die derart prognostizitierte Konvergenzgeschwindigkeit tatsächlich beobachtet werden kann. Hat kein Eigenwert von M positiven Realteil und sind die Jordanblöcke, die zu $\lambda = 0$ gehören, diagonal, so kann man die „klassische" Konvergenzgeschwindigkeit $O(h^2)$ erwarten.

3 Die Iterierte Defekt-Korrektur (IDeC)

Nach den Ergebnissen des vorherigen Abschnittes läge es nun nahe, zur Erzielung einer Lösung höherer Genauigkeit einfach die Anzahl der Gitterpunkte zu erhöhen und damit die Schrittweite h zu verringern. Tatsächlich stoßen wir damit aber bald an die Grenzen des zur Verfügung stehenden Arbeitsspeichers; außerdem muß man mitunter sehr lange Rechenzeiten einkalkulieren, um ein gewünschtes Genauigkeitsniveau zu erreichen. Diese Methode ist also in der Praxis nur eingeschränkt einsetzbar. Eine Alternative besteht darin, das gesamte Verfahren mit Differenzenquotienten höherer Ordnung durchzuführen. Dies führt jedoch zu einer stärkeren Kopplung im Gleichungssystem und damit zu einer größeren Bandbreite, was sich wiederum negativ auf die Performance auswirkt.

Diese Nachteile lassen sich bei iterativen Verfahren weitgehend vermeiden, da man, bis auf Ausnahmen an den Rändern, durchwegs mit Blockdiagonalmatrizen arbeiten kann. Die Konvergenzordnung erhöht sich dabei sukzessive in jedem Iterationsschritt, was ein respektables Genauigkeitsniveau in vernünftiger Zeit liefert. Als Beispiel wären hier etwa Extrapolationsmethoden zu nennen. Wir konzentrieren uns jedoch auf ein Verfahren, das auf Schätzungen des globalen Diskretisierungsfehlers basiert und Iterierte Defekt-Korrektur genannt wird.

3.1 Schätzung des globalen Diskretisierungsfehlers

Die Grundlage der Methode stellt die Idee dar, die von Zadunaisky in [17] zur Schätzung des globalen Diskretisierungsfehlers bei der numerischen Lösung von Differentialgleichungen mittels Differenzenverfahren vorgeschlagen wurde.

Zur Illustration betrachten wir die skalare Randwertaufgabe

$$y''(t) = f(t, y(t)), \qquad a \le t \le b, \tag{3.1a}$$

$$y(a) = \alpha, \quad y(b) = \beta, \tag{3.1b}$$

mit der exakten Lösung $z(t)$. Sei $\mathbb{B} = ([a,b] \times \mathbb{R}) \subset \mathbb{R}^2$. Für f mögen auf $\mathbb{B}$ folgende Annahmen erfüllt sein:

$$f \in C^\infty(\mathbb{B}), \tag{3.2a}$$

$$f \text{ ist Lipschitz-stetig mit der Lipschitz-Konstanten } L, \tag{3.2b}$$

$$\frac{\partial}{\partial y} f(t,y) \ge 0. \tag{3.2c}$$

Obwohl die Resultate auch unter schwächeren Voraussetzungen an die Differenzierbarkeit von f ihre Gültigkeit behalten, nehmen wir der Einfachheit halber an, daß sämtliche Ableitungen von f existieren.

Verwenden wir zur numerischen Lösung von (3.1) das klassische Drei-Punkt Differenzenverfahren auf dem äquidistanten Gitter

$$\Delta_{[a,b]} = \{t_\nu, \nu = 0(1)N | t_\nu = a + \nu h, t_N = b\}, \tag{3.3}$$

so erhalten wir mit $y_{\Delta_{[a,b]}} = (y_0, \ldots, y_N)$ das Gleichungssystem

$$y_0 - \alpha = 0, \tag{3.4a}$$

$$\frac{y_{\nu-1} - 2y_\nu + y_{\nu+1}}{h^2} - f(t_\nu, y_\nu) = 0, \qquad \nu = 1(1)N-1, \tag{3.4b}$$

$$y_N - \beta = 0. \tag{3.4c}$$

Die Lösungswerte ζ_ν, $\nu = 0(1)N$, von (3.4) stellen nun die ersten Approximationswerte für $z(t_\nu)$ dar.

Den daraus resultierenden globalen Diskretisierungsfehler $\zeta_\nu - z(t_\nu)$ schätzt Zadunaisky nun auf folgende Art:

Die ζ_ν-Werte werden durch ein Polynom $P_h(t)$ mit $P_h(t_\nu) = \zeta_\nu$ interpoliert; der Index h soll dabei anzeigen, daß das Polynom von der Schrittweite h abhängt, die in (3.3) bzw. (3.4) zur numerischen Lösung von (3.1) herangezogen wurde. In diesem Einführungsabschnitt arbeiten wir nur mit einem einzigen interpolierenden Polynom auf $[a, b]$. Später werden wir, um zu hohe Polynomgrade und die damit verbundenen Nachteile zu vermeiden, auf stückweise Interpolation zurückgreifen. Betrachten wir nun das folgende neue Randwertproblem:

$$y''(t) = f(t, y(t)) + P_h''(t) - f(t, P_h(t)), \qquad a \le t \le b, \tag{3.5a}$$

$$y(a) = \alpha, \quad y(b) = \beta. \tag{3.5b}$$

Da sich (3.5) von (3.1) nur durch den Störungsterm

$$d_h(t) := P_h''(t) - f(t, P_h(t)), \tag{3.6}$$

den sogenannten *Defekt*, unterscheidet, bezeichnen wir es als „Nachbarproblem" (NP) von (3.1). Offensichtlich hat die Randwertaufgabe (3.5) eine eindeutige Lösung, nämlich $P_h(t)$. Obwohl wir die exakte Lösung kennen, lösen wir (3.5) wieder numerisch, und zwar mit derselben Methode und derselben Schrittweite h, wie wir das schon für das Randwertproblem (3.1) gemacht haben. Auf diese Weise erhalten wir Näherungswerte π_ν für $P_h(t_\nu)$, $\nu = 0(1)N$. Mit diesen können wir nun den bekannten globalen Diskretisierungsfehler $\pi_\nu - P_h(t_\nu) = \pi_\nu - \zeta_\nu$, als Schätzung für den unbekannten Fehler $\zeta_\nu - z(t_\nu)$ verwenden.

Zadunaisky gibt folgende heuristische Begründung für sein Vorgehen an: Da $P_h(t)$ die Näherungswerte ζ_ν für $z(t_\nu)$ interpoliert, stellt $P_h(t)$ eine gewisse Approximation der exakten Lösung $z(t)$ von (3.1) auf $[a, b]$ dar und $P_h''(t)$ eine solche für $z''(t)$. Daraus folgt, daß der Defekt $d_h(t)$ „klein" sein wird. Die Randwertaufgaben (3.1) und (3.5) sind deshalb benachbarte mathematische Probleme, die, wenn sie mit demselben numerischen Algorithmus gelöst werden, bezüglich Struktur und Größenordnung des Fehlers ähnlich sein werden.

In [17] verwendet Zadunaisky diese Methode zur Schätzung des globalen Diskretisierungsfehlers bei der Lösung von Anfangswertproblemen mittels Runge-Kutta-Verfahren und gibt dafür nur die oben skizzierte heuristische Motivation.

3.2 Die Methode der Iterierten Defekt-Korrektur

Wir stellen nun die IDeC in dem Rahmen vor, wie sie von Frank in [4] eingeführt wurde.

Die in Abschnitt 3.1 beschriebene Methode liefert eine mit korrektem Vorzeichen versehene Schätzung $\pi_\nu - P_h(t_\nu)$, $\nu = 0(1)N$, für den globalen Diskretisierungsfehler. Damit erhalten wir aus

$$\overset{(1)}{\zeta}_\nu := \zeta_\nu - (\pi_\nu - P_h(t_\nu)) \tag{3.7}$$

eine verbesserte Näherung für $z(t_\nu)$, denn aus der in Abschnitt 3.1 hergeleiteten Relation

$$\pi_\nu - P_h(t_\nu) \approx \zeta_\nu - z(t_\nu)$$

folgt sofort

$$z(t_\nu) = \zeta_\nu - (\zeta_\nu - z(t_\nu)) \approx \zeta_\nu - (\pi_\nu - P_h(t_\nu)) = \overset{(1)}{\zeta}_\nu. \tag{3.8}$$

Es liegt nun natürlich nahe, dieses Vorgehen zu iterieren: Die durch (3.7) erhaltenen $\overset{(1)}{\zeta}_\nu$-Werte werden durch ein passendes Polynom $\overset{(1)}{P}_h(t)$ interpoliert, und analog zu (3.5) kann man ein neues Nachbarproblem mit der exakten Lösung $\overset{(1)}{P}_h(t)$ konstruieren. Löst man die Randwertaufgabe nun wieder mit dem gleichen Algorithmus und der gleichen Schrittweite h (Lösung: $\overset{(1)}{\pi}_\nu$), so erhält man eine neue Fehlerschätzung $\overset{(1)}{\pi}_\nu - \overset{(1)}{P}_h(t_\nu)$ und so weiter.

Da das neue NP

$$y''(t) = f(t, y(t)) + \overset{(1)}{P}''_h(t) - f(t, \overset{(1)}{P}_h(t)), \qquad a \le t \le b, \tag{3.9a}$$

$$y(a) = \alpha, \quad y(b) = \beta. \tag{3.9b}$$

„näher" bei (3.1) liegt als (3.5) ist zu erwarten, daß die Approximation von $z(t_\nu)$ durch die neuen Näherungswerte

$$\overset{(2)}{\zeta}_\nu := \zeta_\nu - (\overset{(1)}{\pi}_\nu - \overset{(1)}{P}_h(t_\nu)) \tag{3.10}$$

besser ist als die durch $\overset{(1)}{\zeta}_\nu$. Es werden ja die besseren Näherungswerte $\overset{(1)}{\zeta}_\nu$ interpoliert und deshalb wird der neue Defekt $\overset{(1)}{d}_h(t)$ kleiner sein als $d_h(t)$, vgl. (3.6).

In Abschnitt 3.3 werden wir die Gültigkeit dieser Überlegungen nachweisen. Unter Verwendung von stückweiser Interpolation mit passendem Polynomgrad läßt sich zeigen, daß obige Iteration für das Differenzenverfahren (3.4) in jedem Schritt eine Verbesserung des Fehlers der Größenordnung $O(h^2)$ ergibt. Von zentraler Bedeutung ist dabei die Existenz der asymptotischen Entwicklung des globalen Diskretisierungsfehlers der Näherungslösung ζ_ν, $\nu = 0(1)N$, von (3.4):

$$\zeta_\nu - z(t_\nu) = \sum_{k=1}^{K} e_{2k}(t_\nu) h^{2k} + O(h^{2K+2}), \tag{3.11}$$

mit Funktionen $e_{2k}(t)$, $k = 1(1)K$, die von h unabhängig sind. Für ein abgeändertes Differenzenverfahren, das eine Reduktion wie $O(h^4)$ liefert, siehe Frank [4]; in diesem Fall ist $e_2(t) \equiv 0$.

Man erhält also immer bessere Approximationen $\overset{(j)}{\zeta}_\nu$ für $z(t_\nu)$, $\nu = 0(1)N$, indem man sukzessive die (verbesserten) Schätzungen für den globalen Diskretisierungsfehler von den ursprünglichen ζ_ν-Werten abzieht:

$$\begin{aligned} \overset{(1)}{\zeta}_\nu &:= \zeta_\nu - (\pi_\nu - P_h(t_\nu)), & \overset{(2)}{\zeta}_\nu &:= \zeta_\nu - (\overset{(1)}{\pi}_\nu - \overset{(1)}{P}_h(t_\nu)), \\ & & \overset{(3)}{\zeta}_\nu &:= \zeta_\nu - (\overset{(2)}{\pi}_\nu - \overset{(2)}{P}_h(t_\nu)), \quad \ldots \end{aligned} \tag{3.12}$$

Dieses Vorgehen bezeichnen wir als *Iterierte Defekt-Korrektur.*

Der Vorteil der IDeC liegt in ihrer großen numerischen Effizienz. Jeder Iterationsschritt benötigt sehr wenig Aufwand, insbesondere viel weniger, als zur Berechnung der ursprünglichen ζ_ν-Werte notwendig gewesen ist. Um die ζ_ν-Werte zu erhalten, muß das $N+1$-dimensionale nichtlineare Gleichungssystem (3.4) gelöst werden, etwa unter Verwendung eines Newton-Verfahrens. Man kann sich leicht davon überzeugen (siehe Frank [4]), daß

1. der Vektor $\zeta = (\zeta_0, \ldots, \zeta_N)$ der ersten Näherung für $R_\Delta z$ bei entsprechend kleiner Schrittweite h einen derart guten Startvektor für das Newton-Verfahren zur Lösung sämtlicher Nachbarprobleme darstellt, daß man nur sehr wenige Newton-Schritte benötigt, und daß
2. ein Pseudo-Newton-Verfahren völlig ausreicht, d.h. man kann dieselbe Jacobi-Matrix, ausgewertet im Vektor ζ, für alle Newton-Schritte beibehalten.

Eines der nützlichsten Merkmale des Verfahrens besteht also darin, daß *alle* linearen Gleichungssysteme, die im Laufe der Iterierten Defekt-Korrektur gelöst werden müssen, dieselbe Koeffizientenmatrix aufweisen.

3.3 Algorithmen und theoretische Resultate für nichtsinguläre Randwertprobleme

Wie bereits im vorigen Abschnitt erwähnt wurde, wollen wir stückweise Interpolation anstatt eines einzigen interpolierenden Polynoms $P_h(t)$ auf ganz $[a,b]$ verwenden. Um dabei mit einem fixen Polynomgrad m arbeiten zu können, teilen wir $[a,b]$ in n gleichgroße Teilintervalle $I_1, I_2, \ldots, I_n$, $n \in \mathbb{N}$, wo

$$I_i := [t^{i-1}, t^i], \quad t^i = a + imh, \qquad i = 1(1)n; \quad t^0 = a, t^n = b. \tag{3.13}$$

Die Forderung $t^0 = a, t^n = b$ legt die Schrittweite h in (3.3) mit

$$h = \frac{b-a}{n \cdot m}, \qquad (n \cdot m = N) \tag{3.14}$$

eindeutig fest und es ergibt sich als neues äquidistantes Gitter

$$\Delta_{[a,b]} = \{t_\nu, \nu = 0(1)n \cdot m | t_\nu = a + \nu h, t_{n \cdot m} = b\} \tag{3.15}$$

mit $t^i = t_{im}$, $i = 1(1)n$.

Wir bezeichnen mit $P_{ih}(t)$ das Polynom, das die zum Teilintervall I_i gehörigen Punkte $(t_{(i-1)m}, \zeta_{(i-1)m}), \ldots, (t_{im}, \zeta_{im})$ interpoliert. Die Interpolierende $P_h(t)$ auf $[a,b]$ definieren wir nun durch die Relation

$$P_h(t) := P_{ih}(t) \quad \text{für} \quad t \in I_i, \quad i = 1(1)n. \tag{3.16}$$

Die rechte Seite $f_h(t,y)$ des Nachbarproblems im 0-ten Iterationsschritt (vgl. (3.5) und (3.6)) wird ebenso stückweise durch

$$\begin{aligned} d_h(t) &:= d_{ih}(t) = P''_{ih}(t) - f(t, P_{ih}(t)), \\ f_h(t,y) &:= f_{ih}(t,y) = f(t,y) + d_{ih}(t), \end{aligned} \qquad t \in I_i, \quad i = 1(1)n \tag{3.17}$$

definiert, wobei zu beachten ist, daß $d_h(t)$ an den Stellen t^i, $i = 1(1)n-1$, unstetig ist. Analog wird bei den Defekten $\overset{(j)}{d}_h(t)$ und den rechten Seiten $\overset{(j)}{f}_h(t,y)$ der weiteren Nachbarprobleme in den Schritten $j = 1,2,\ldots$, vorgegangen.

Das mittels (3.17) beschriebene „zusammengesetzte“ NP

$$y''(t) = f_h(t, y(t)), \qquad a \le t \le b, \tag{3.18a}$$

$$y(a) = \alpha, \quad y(b) = \beta, \tag{3.18b}$$

hat in dieser Form keine eindeutige Lösung mehr, da an den Knotenpunkten $t^i = t_{im}$, $i = 1(1)n-1$, keine Randbedingungen vorgeschrieben sind. Dies führt natürlich zum Scheitern des Verfahrens, da die Eindeutigkeit der Lösung $P_h(t)$ einer der wesentlichen Bestandteile von Zadunaisky's Idee ist (vgl. Abschnitt 3.1). Das unter (3.16) definierte zusammengesetzte Polynom $P_h(t)$ löst zwar offensichtlich das System (und weist somit an den Knotenpunkten t^i Sprungstellen in der ersten Ableitung auf), die numerische Lösung von (3.18) ist jedoch nicht in eindeutiger Weise möglich.

Um Eindeutigkeit zu erhalten, betrachten wir deshalb zunächst folgendes Randwertproblem für ein $i \in \{1,\ldots,n\}$:

$$\begin{aligned} y''(t) &= f_{ih}(t, y(t)), \\ y(t^{i-1}) = \gamma_{i-1}, &\quad y(t^i) = \gamma_i, \end{aligned} \qquad t \in I_i, \quad \gamma_{i-1}, \gamma_i \in \mathbb{R}. \tag{3.19}$$

Da $d_{ih}(t)$ nicht von y abhängt, sind die Voraussetzungen zur eindeutigen Lösbarkeit erfüllt. Wenn wir nun die reellen Zahlen $\gamma_1, \ldots, \gamma_{n-1}$ vorschreiben und dabei $\gamma_0 = \alpha, \gamma_n = \beta$ berücksichtigen (vgl. (3.18b)), so ist die Lösung der Randwertaufgabe (3.19) für jedes I_i eindeutig bestimmt; wir bezeichnen sie mit $z_i(t)$, $i = 1(1)n$. Die auf $[a,b]$ stetige Funktion $z(t) := z_i(t)$, $t \in I_i$, ist somit die eindeutige Lösung des zusammengesetzten Randwertproblems

$$y''(t) = f_h(t, y(t)), \qquad a \le t \le b, \tag{3.20a}$$

$$y(a) = \alpha, \quad y(t^i) = \gamma_i, \ i = 1(1)n-1, \quad y(b) = \beta. \tag{3.20b}$$

Dieses NP hat also eine Mannigfaltigkeit von Lösungen, die durch die Parameter $\gamma_1, \ldots, \gamma_{n-1}$ beschrieben wird, und $P_h(t)$ ist ein Element dieser Mannigfaltigkeit. Man

muß demnach $n-1$ Parameter festlegen; $z(t)$ weist dabei ebenso wie $P_h(t)$ an den Knotenpunkten Sprungstellen in der ersten Ableitung auf. In Anlehnung an Zadunaisky erweist es sich jedoch als günstiger, nicht die Werte der Lösung γ_i, $i = 1(1)n-1$ vorzuschreiben, sondern die Sprunghöhe φ_i der ersten Ableitung an den Stellen t^i, $i = 1(1)n-1$. Betrachten wir also die folgende Klasse von Randwertaufgaben:

In der Menge der stetigen und stückweise analytischen Funktionen

$$\Theta = \{y(t) | y(t) \in C[a,b], y(t) := y_i(t) \text{ für } t \in I_i, \text{ wo } y_i(t) \in C^\infty[I_i], i = 1(1)n\} \quad (3.21)$$

suchen wir nach jenem Element $z(t) = z_i(t), t \in I_i$, das die folgenden Gleichungen erfüllt:

$$y_i''(t) = f_i(t, y_i(t)), \qquad t \in I_i, \ i = 1(1)n, \qquad (3.22a)$$
$$y(a) = \alpha, \quad y(b) = \beta, \qquad (3.22b)$$
$$y_{i+1}'(t^i) - y_i'(t^i) = \varphi_i, \qquad i = 1(1)n-1. \qquad (3.22c)$$

Die rechten Seiten $f_i(t,y)$ auf $\mathbb{B}_i := (I_i \times \mathbb{R}) \subset \mathbb{R}^2$ sollen dabei denselben Bedingungen (3.2) genügen wie f auf $\mathbb{B}$; φ_i bezeichnet die Sprunghöhe der Ableitung im Knotenpunkt $t^i = t_{im}$.

Klarerweise ist (3.1) nur ein Spezialfall von (3.22) mit $\varphi_i = 0$, $i = 1(1)n$ und denselben rechten Seiten für alle $\mathbb{B}_i$. Ebenso sind die einzelnen Nachbarprobleme, die auf den Teilintervallen I_i des Gitters $\Delta_{[a,b]}$ laufen, vom Typ (3.22). Mit den Sprunghöhen

$$\varphi_{ih} := P_{i+1,h}'(t^i) - P_{ih}'(t^i)$$

und den rechten Seiten $f_{ih}(t,y)$, wie sie in (3.17) eingeführt wurden, hat der 0-te Schritt der Iteration die folgende Gestalt:

$$y_i''(t) = f_{ih}(t, y_i(t)), \qquad t \in I_i, \ i = 1(1)n, \qquad (3.23a)$$
$$y(a) = \alpha, \quad y(b) = \beta, \qquad (3.23b)$$
$$y_{i+1}'(t^i) - y_i'(t^i) = \varphi_{ih}, \qquad i = 1(1)n-1. \qquad (3.23c)$$

Die eindeutige Lösung ist jetzt natürlich das in (3.16) definierte zusammengesetzte Polynom $P_h(t)$, und die Iterierte Defekt-Korrektur ist somit auch für stückweise Interpolation wohldefiniert.

Um die Randwertprobleme des Typs (3.23) mittels Differenzenverfahren lösen zu können, müssen wir zuerst die Bedingungen $y_{i+1}'(t^i) - y_i'(t^i) = \varphi_{ih}$ für die Sprunghöhe an den Knotenpunkten diskretisieren, was mittels zentraler Differenzenquotienten

$$\begin{aligned} y_i'(t^i) = y_i'(t_{im}) &= \frac{y_i(t_{im+1}) - y_i(t_{im-1})}{2h} \\ y_{i+1}'(t^i) = y_{i+1}'(t_{im}) &= \frac{y_{i+1}(t_{im+1}) - y_{i+1}(t_{im-1})}{2h} \end{aligned} \qquad i = 1(1)n-1 \qquad (3.24)$$

realisiert wird. Die Größen $y_i(t_{im+1})$ und $y_{i+1}(t_{im-1})$, die ja wegen $t_{im+1} \in I_{i+1}$ bzw. $t_{im-1} \in I_i$ nicht definiert sind, sind durch das Einführen zusätzlicher Gleichungen

im Differenzenschema (vgl. Abschnitt 2.3) wieder eliminierbar, und wir erhalten die diskretisierte Sprungbedingung

$$\frac{y_{im-1} - 2y_{im} + y_{im+1}}{h^2} - \frac{1}{2}(f_{ih}(t_{im}, y_{im}) + f_{i+1,h}(t_{im}, y_{im})) - \frac{\varphi_{ih}}{h} = 0 \qquad (3.25)$$

mit $(y_1, \ldots, y_{n\cdot m}) = y_{\Delta_{[a,b]}}$, $\Delta_{[a,b]}$ wie in (3.15).

Damit können wir nun das Differenzenschema zur Lösung der Nachbarprobleme vom Typ (3.23) direkt angeben:

$$\frac{y_{\nu-1} - 2y_\nu + y_{\nu+1}}{h^2} - f_{ih}(t_\nu, y_\nu) = 0, \qquad (i-1)m < \nu < im, \quad i = 1(1)n, \qquad (3.26a)$$

$$y_0 = \alpha, \quad y_{n\cdot m} = \beta, \qquad (3.26b)$$

$$\frac{y_{im-1} - 2y_{im} + y_{im+1}}{h^2} - \frac{1}{2}(f_{ih}(t_{im}, y_{im}) + f_{i+1,h}(t_{im}, y_{im})) - \frac{\varphi_{ih}}{h} = 0, \quad i = 1(1)n-1. \qquad (3.26c)$$

Im Sinne Zadunaisky's müssen die ursprüngliche Randwertaufgabe (3.1) und die Nachbarprobleme (3.23) mit demselben Algorithmus und derselben Schrittweite h gelöst werden. Der obige Algorithmus reduziert sich für $\varphi_{ih} = 0$, $i = 1(1)n-1$, und $f_{ih} = f$, $i = 1(1)n$, auf das anfänglich eingeführte Schema (3.4).

Für das asymptotische Verhalten des absoluten Fehlers in Abhängigkeit von h gilt nun folgender Satz (vgl. Abschnitt 3.2):

Satz 3.1 *Sei der Polynomgrad der stückweise Interpolierenden $m = 3 + 2r, r \in \mathbb{N}$. Dann gilt unter Verwendung der Algorithmen* (3.4) *bzw.* (3.26) *zur Lösung von* (3.1)*:*

$$\begin{aligned} &\overset{(0)}{\zeta}_\nu - z(t_\nu) = O(h^2), \\ &\overset{(1)}{\zeta}_\nu - z(t_\nu) = O(h^4), \\ &\overset{(2)}{\zeta}_\nu - z(t_\nu) = O(h^6), \quad \ldots, \quad \overset{(r)}{\zeta}_\nu - z(t_\nu) = O(h^{2+2r}). \end{aligned} \qquad (3.27)$$

Weitere Schritte führen zu keiner Erhöhung der Konvergenzordnung mehr.

Der vollständige Beweis wurde von Frank in [5] veröffentlicht, wobei die asymptotische Entwicklung (3.11) entscheidend eingeht.

Zusammenhänge zu Kollokationsmethoden

Ist das zu lösende Randwertproblem von der Bauart (3.1), sodaß Satz 3.1 gültig ist, so ist nach dem r-ten Iterationsschritt keine Verbesserung der Konvergenzordnung mehr zu erzielen, es sei denn, man würde Interpolationspolynome mit einem höheren Grad verwenden. Von besonderem Interesse ist auch eine Charakterisierung des Fixpunktes der Defektkorrektur-Iteration. Dieser wird zwar (abgesehen von einfachen Sonderfällen, vgl. [1]) nicht nach endlich vielen Schritten erreicht; seine Kenntnis liefert jedoch wesentliche Informationen über die mit der Iteration maximal erreichbare Genauigkeit.

Angenommen, die IDeC-Iteration steht nach dem r-ten Schritt am Fixpunkt, dann gilt

$$\overset{(r)}{\zeta}_\nu = \overset{(r+1)}{\zeta}_\nu = \zeta_\nu - (\overset{(r)}{\pi}_\nu - \overset{(r)}{P}_h(t_\nu)), \qquad \nu = 0(1)n \cdot m.$$

Wegen $\overset{(r)}{P}_h(t_\nu) = \overset{(r)}{\zeta}_\nu$ ist das äquivalent zu $\overset{(r)}{\pi}_\nu = \zeta_\nu$. Nun ist $\overset{(r)}{\pi}_\nu$ die mittels Algorithmus (3.26) erhaltene numerische Lösung der Randwertaufgabe (vgl. (3.5) bzw. (3.9))

$$y''(t) = f(t, y(t)) + \overset{(r)}{P}''_h(t) - f(t, \overset{(r)}{P}_h(t)), \qquad a \le t \le b,$$
$$y(a) = \alpha, \quad y(b) = \beta,$$

und ζ bekommen wir als numerische Lösung von (vgl. (3.1))

$$y''(t) = f(t, y(t)), \qquad a \le t \le b,$$
$$y(a) = \alpha, \quad y(b) = \beta.$$

Offensichtlich ist die Forderung $\overset{(r)}{\pi}_\nu = \zeta_\nu$ daher gleichbedeutend mit verschwindendem Defekt an den Gitterpunkten. An den Knotenpunkten $t^i = t_{im}$, $i = 1(1)n-1$, muß daher

$$\begin{aligned}
\overset{(r)}{d}_h(t^i) &= \frac{1}{2}(\overset{(r)}{P}''_{ih}(t^i) + \overset{(r)}{P}''_{i+1,h}(t^i)) - \frac{1}{2}(f(t^i, \overset{(r)}{P}_{ih}(t^i)) + \\
&\quad + f(t^i, \overset{(r)}{P}_{i+1,h}(t^i))) + \frac{1}{h}(\overset{(r)}{P}'_{i+1,h}(t^i) - \overset{(r)}{P}'_{ih}(t^i)) = \\
&= \frac{1}{2}(\overset{(r)}{P}''_{ih}(t^i) + \overset{(r)}{P}''_{i+1,h}(t^i)) - f(t^i, \overset{(r)}{P}_{ih}(t^i)) + \\
&\quad + \frac{1}{h}(\overset{(r)}{P}'_{i+1,h}(t^i) - \overset{(r)}{P}'_{ih}(t^i)) = 0 \qquad (3.28a)
\end{aligned}$$

gelten; für die anderen Stellen t_ν, $(i-1)m < \nu < im$, $i = 1(1)n$, ergibt sich

$$\overset{(r)}{d}_h(t_\nu) = \overset{(r)}{P}''_h(t_\nu) - f(t_\nu, \overset{(r)}{P}_h(t_\nu)) = 0. \qquad (3.28b)$$

Offensichtlich kann man den so definierten Fixpunkt des IDeC-Verfahrens als die Lösung eines gewissen Kollokationsverfahrens deuten. Dieses Verfahren unterscheidet sich geringfügig von den klassischen Kollokationsmethoden; wie in [1] gezeigt wurde, ist es stabil und besitzt die Konsistenzordnung $O(h^{m-1})$.

4 Numerische Ergebnisse

Der Einfachheit halber beschränken wir uns auf den skalaren Fall, da er weniger rechenzeit– und speicherplatzintensiv ist, aber dennoch alle charakteristischen Eigenschaften der singulären Probleme aufweist. Insbesondere werden anhand von verschiedenen Modellbeispielen folgende Fragestellungen beleuchtet:

1. Wie wirkt sich die Eigenwertstruktur der Matrix M (siehe Abschnitt 2.2, speziell ab (2.5)) auf das Lösungsverhalten aus ? Die Modelle sind dabei meist so gewählt, daß innerhalb eines Beispiels durch geeignete Parameterwahl verschiedene Eigenwertsituationen diskutiert werden können.

2. Welches asymptotische Verhalten des Fehlers bei klein werdender Schrittweite h ist zu beobachten ? Kann das Rundungsfehlerniveau erreicht werden ? Besonderes Interesse gilt dabei im Hinblick auf Kollokationsmethoden natürlich dem Fixpunktverhalten.

3. Inwiefern stimmt das Verhalten der Konvergenzordnung mit den Resultaten für nichtsinguläre Randwertprobleme überein ? Wir beobachten insbesondere, ob sich die Ordnungen $O(h^2), O(h^4), \ldots$ einstellen und ob eine Erhöhung der Konvergenzordnung auch nach dem 4. Schritt bzw. über das Niveau von $O(h^{2+2r}) = O(h^8)$ hinaus auftritt. Ist in jedem Fall eine Stabilisierung des Ordnungsniveaus feststellbar ?

4. Was kann über den Defekt nach den jeweiligen Korrekturschritten ausgesagt werden ? Hier achten wir speziell auf Glattheits– und asymptotische Eigenschaften bei klein werdendem h.

Die Resultate werden sowohl graphisch als auch in tabellarischer Form präsentiert.

Beschreibung der Tabellen und Graphiken

Die Tabellen geben in Abhängigkeit vom jeweiligen Iterationsschritt s, $s = 0(1)9$, folgende Werte wieder:

- h: Die Schrittweite $h = \dfrac{1}{n \cdot m} = \dfrac{1}{n \cdot 9}$ wird durch Verdoppelung von n sukzessive halbiert, das feinste Gitter ist mit $h = \dfrac{1}{576}$ festgelegt. Typische Startwerte für n sind $n = 1$ bzw. $n = 2$.

- $e_h(t)$ bzw. $\|e_h\|$: Mit $e_h(t_\nu) := \zeta_\nu - z(t_\nu)$ bezeichnen wir den absoluten Fehler der Näherungslösung ζ_ν nach dem s-ten Defektkorrekturschritt an der Stelle t_ν, $\nu = 0(1)N$. $\|e_h\|$ steht somit für die Maximumnorm dieses Fehlers, vgl. Abschnitt 2.1. Der Index h soll die Abhängigkeit von der Schrittweite h andeuten.

- t_e: Wird die Norm $\|e_h\|$ angegeben, so bezeichnet t_e die Stelle, an der das Maximum angenommen wird: $\|e_h\| = |e_h(t_e)|$.

- p ist die erzielte Konvergenzordnung beim Übergang zur halben Schrittweite: Aus dem Ansatz

$$\begin{aligned} e_h(t) \quad (\text{bzw. } \|e_h\|) &= c \cdot h^p \\ e_{\frac{h}{2}}(t) \quad (\text{bzw. } \|e_{\frac{h}{2}}(t)\|) &= c \cdot (\frac{h}{2})^p \end{aligned}$$

mit reeller Errorkonstanten c ergibt sich für den festen Punkt t_ν

$$p = \frac{\ln |e_h(t_\nu)/e_{\frac{h}{2}}(t_\nu)|}{\ln 2}$$

bzw. bezüglich der Maximumnorm

$$p = \frac{\ln(\|e_h\|/\|e_{\frac{h}{2}}\|)}{\ln 2}.$$

Klarerweise ist die Berechnung der Konvergenzordnung mit Hilfe der Maximumnormen des globalen Fehlers besonders präzise, wenn das Maximum an derselben Stelle t_e angenommen wird, da andernfalls die Errorkonstanten verschieden sind. Dies ist beim Lesen der entsprechenden Tabellen zu berücksichtigen.

- $e_h(t)/h^p$ bzw. $\|e_h\|/h^p$: Gemäß obigem Ansatz streben diese Größen für klein werdendes h gegen den Wert der Errorkonstanten c.

- $d_h(t)$ bzw. $\|d_h\|$, t_d: Mit

$$d_h(t) := P_h''(t) - \frac{a_1}{t} P_h'(t) - \frac{a_0}{t^2} P_h(t) - f(t), \qquad 0 \le t \le 1,$$

bezeichnen wir wie schon in Abschnitt 3 den Defekt der Näherungslösung nach dem jeweiligen Iterationsschritt. $\|d_h\|$ und t_d definieren wir analog zu den entsprechenden Größen des absoluten Fehlers. Zur Berechnung von $d_h(t)$ und somit $\|d_h\|$ werden nur die Werte an den Gitterstellen $t = t_\nu$, $\nu = 0(1)N$, herangezogen.

- $d_h(t)/h^p$ bzw. $\|d_h\|/h^p$: Da hier mit p die Konvergenzordnung des absoluten Fehlers übernommen wird, haben diese Werte keine unmittelbare theoretische Entsprechung. In ihrem numerischen Verhalten ähneln sie jedoch klarerweise den oben erklärten Errorkonstanten, sodaß sie für Vergleichszwecke ihre Berechtigung haben. Wir werden sie auch als „Pseudo-Defektkonstanten" bezeichnen. Auch hier können für unterschiedliche Schrittweiten die Defektmaxima an verschiedenen Stellen angenommen werden.

Neben der exakten Lösung plotten wir für die ersten vier Iterationsschritte zusätzlich den Verlauf des absoluten Fehlers und seiner Fehlerschätzung sowie den Defekt auf $[0, 1]$. Die Zeichnungen sind dabei alle logarithmisch skaliert und optimal an den jeweiligen Wertebereich angepaßt, wodurch der Skala am linken Bildrand besonderes Augenmerk geschenkt werden muß. Der Defekt nach einem Korrekturschritt wird stets für zwei verschiedene Gitter dargestellt: ein relativ grobes mit 19 Gitterpunkten und ein feines mit 289 Punkten.

Implementierungen

Die Realisierung des beschriebenen Verfahrens erfolgte in Fortran 77 auf einer IBM ES/9000 der Universität Wien (genaue Modellbezeichnung: IBM 9021-720, Maschinenzahlenbereich etwa 10^{-78} bis 10^{75}) mit vierfacher Genauigkeit ($eps \approx 10^{-30}$). Da die Entwicklung eines qualitativ hochwertigen Programmcodes nicht vordergründliches Ziel dieser Arbeit war, werden wir programmiertechnische Erläuterungen nur anführen, wenn sie zum Verständnis der nachfolgenden Experimente unmittelbar notwendig sind. Im Laufe der Ergebnisdiskussion eingebrachte Erweiterungen des numerischen Verfahrens werden direkt in Abschnitt 4 an den entsprechenden Stellen besprochen.

Parameterwahl

Zur Interpolation auf dem unter (3.15) eingeführten äquidistanten Gitter $\Delta_{[a,b]}$ verwenden wir Lagrange-Interpolation. Da wir zu Vergleichszwecken möglichst viele Formalvoraussetzungen von Satz 3.1 übernehmen wollen, behalten wir den Polynomgrad $m = 3 + 2r$, $r \in \mathbb{N}$ für die stückweise Interpolierenden P_{ih} bei. In einer Anzahl verschiedener numerischer Experimente hat sich der Wert $r = 3$ als optimaler Kompromiß zwischen großer Genauigkeit durch ein möglichst feines Gitter einerseits und niedrigem Polynomgrad zur Vermeidung des Überschwingens des interpolierenden Polynoms andererseits erwiesen. Der verwendete Polynomgrad beträgt somit $m = 9$.

Gemäß Satz 3.1 für nichtsinguläre Randwertprobleme ist die Anzahl der Defektkorrekturschritte, die zu einer Erhöhung der Konvergenzordnung führen sollten, durch $r + 1 = 4$ gegeben. Das bedeutet, daß die Iteration aus dem Basisschritt und drei Defektkorrekturschritten besteht.

In unseren Testreihen führen wir weitere Iterationsschritte durch, um das Verhalten des IDeC-Verfahren über diese Schranke hinaus zu studieren. Nach 10 Iterationsschritten wird die Iteration jedoch in jedem Fall abgebrochen.