TECHNISCHE UNIVERSITÄT WIEN

VIENNA UNIVERSITY OF TECHNOLOGY

D I P L O M A R B E I T

A posteriori Fehlerschätzer für Zweipunkt-Randwertprobleme mittels Defektkorrektur

Ausgeführt am Institut für
Analysis und Scientific Computing
der Technischen Universität Wien

unter der Anleitung von
Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Winfried Auzinger

durch
Gerhard Kitzler
Währingerstraße 64/7
1090 Wien

Wien, im Oktober 2010

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Motivation

Bei der approximativen Lösung mathematischer Probleme drängt sich in natürlicher Weise die Frage nach der Qualität der Approximation auf. Ist die exakte Lösung u eines Problems bekannt, so lässt sich der Fehler der Approximation p als die Differenz $p - u$ berechnen. Da nun aber numerische Verfahren vor allem dann von Interesse sind, wenn exaktes Lösen nicht möglich ist, ist in diesem Fall auch die Frage nach dem Fehler nicht sofort zu beantworten. Dementsprechend muss man auf Techniken zurückgreifen, bei denen man ohne Kenntnis der exakten Lösung auskommt, um den Fehler zu schätzen. Die Schätzgröße E bezeichnen wir als Fehlerschätzer für den exakten Fehler e. Natürlich drängt sich auch hier wieder die Frage nach der Qualität des Fehlerschätzers auf: Sei zum Beispiel ein Problem P mit exakter Lösung u und approximativer Lösung p gegeben. Für den Fehler e liege außerdem eine Schätzung E vor. Dann können wir natürlich auch den Fehler des Fehlerschätzers $\epsilon := E - e$ betrachten. Naheliegenderweise wird man von der Größe ϵ verlangen, dass sie klein in Bezug auf e bzw. E selbst ist. Im Kontext der numerischen Lösung von Differentialgleichungen existiert die Konvergenzordnung eines Verfahrens, um die Qualität desselbigen zu messen. Bei gegebener Schrittweite h gilt für ein mit Ordnung n konvergentes Verfahren: $\|p - u\| = O(h^n)$. Der Fehler des Fehlerschätzers sollte in diesem Fall zumindest $\|\epsilon\| = O(h^{n+1})$ erfüllen. Ziel dieser Arbeit ist es, einen Fehlerschätzer für Zweipunkt-Randwertprobleme 2-ter Ordnung

$$y''(x) = f\left(x, y(x), y'(x)\right) \quad \text{für alle } x \in [a, b] \tag{1.1}$$

zu konstruieren, der oben erwähnte Eigenschaft erfüllt. Eine zentrale Größe für die Berechnung des Schätzers wird

$$d(x) := p''(x) - f\left(x, p(x), p'(x)\right)$$

darstellen. d bezeichnen wir als den *Defekt* von p bezüglich der Differentialgleichung. Die Idee der Defektkorrektur stammt ursprünglich von Zadunaisky [1] und wurde wenig später von Stetter [2] allgemein beschrieben.

Die theoretische Behandlung der dadurch ermittelten Fehlerschätzer bei gewöhnlichen Differentialgleichungen 1-ter Ordnung wird in [3] vorgenommen, eine **Matlab**-Implementierung wird in [4] beschrieben. Gleichungen 4-ter Ordnung werden in [5] behandelt. Der Schwerpunkt dieser Arbeit liegt zu einem großen Teil auf der praktischen Umsetzung in der Programmierumgebung **Matlab**. Im folgenden Abschnitt werden wir den Problemtyp, der in dieser Arbeit untersucht wird, konkret definieren und Unterschiede zum klassischen Anfangswertproblem aufzeigen.

1.2 Problemstellung

Wir setzen für die im Folgenden auftretenden reellwertigen Funktionen q_1, q_2 sowie g stets die Differentiationsklasse C^1 (zumindest im abgeschlossenen Intervall $[a, b]$) voraus. Sofern an manchen Stellen eine höhere Differentiationsklasse benötigt wird, wird darauf explizit hingewiesen, insbesondere für Verfahren höherer Ordnung. Wir betrachten in dieser Arbeit hauptsächlich Randwertprobleme 2-ter Ordnung mit linearen Randbedingungen:

$$\begin{aligned} y''(x) - f\left(x, y(x), y'(x)\right) &= 0 \quad \text{für alle } x \in [a, b] \\ R_1 \begin{pmatrix} y(a) \\ y(b) \end{pmatrix} + R_2 \begin{pmatrix} y'(a) \\ y'(b) \end{pmatrix} &= \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} \end{aligned} \tag{1.2}$$

Im linearen Fall lässt sich (1.2) speziell als

$$\begin{aligned} y''(x) + q_1(x)y'(x) + q_2(x)y(x) &= g(x) \quad \text{für alle } x \in [a, b] \\ R_1 \begin{pmatrix} y(a) \\ y(b) \end{pmatrix} + R_2 \begin{pmatrix} y'(a) \\ y'(b) \end{pmatrix} &= \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} \end{aligned} \tag{1.3}$$

anschreiben. Eine 2-mal stetig differenzierbare Funktion $u : [a, b] \to \mathbb{R}$, die (1.2) erfüllt, nennen wir eine Lösung des Randwertproblems.

Im Gegensatz zu Anfangswertproblemen hängen Existenz und Eindeutigkeit der Lösung bei Randwertproblemen viel stärker von den Randgleichungen und dem Intervall $[a, b]$ ab. Man setze z.B. $f = f(x, y, y') = -y$. Damit erhält man als Lösung von (1.2) alle Linearkombinationen $c_1u_1(x) + c_2u_2(x)$ der Funktionen u_1 und u_2, wobei $u_1(x) = \cos(x)$ und $u_2(x) = \sin(x)$. Die Randbedingungen $u(0) \neq u(2\pi)$ liefern, wie man leicht nachrechnet, eine unlösbare Problemstellung. Andererseits erhält man durch $u(0) = u(2\pi) = a$ eine Gleichung mit unendlich vielen Lösungen, denn alle Linearkombinationen der Form $a\cos(x) + c\sin(x)$, $c \in \mathbb{R}$ beliebig erfüllen sowohl die Differentialgleichung als auch die Randbedingungen. Bei Anfangswerten $y(a) = y_0$ und $y'(a) = Dy_0$ ergibt sich aus dem Existenz- und Eindeutigkeitssatz von Picard-Lindelöf die eindeutige Lösbarkeit von (1.2). Für theoretische Überlegungen setzen wir im Folgenden die eindeutige Lösbarkeit von (1.2) voraus.

Kapitel 2

Kollokationsverfahren

Für die numerische Behandlung von Randwertproblemen der Form (1.3) bzw. (1.2) gibt es verschiedene Techniken wie z.B. das Schießverfahren, Finite Differenzen, etc.. Wir wollen im Folgenden auf das im Laufe der Diplomarbeit implementierte Kollokationsverfahren genauer eingehen und zunächst dessen Aufbau erklären: Beim Kollokationsverfahren wird die Lösung der Differentialgleichung (1.2) durch stückweise Polynome $p_k : [x_k, x_{k+1}] \to \mathbb{R}$, die gewissen Stetigkeitsbedingungen genügen (zum Grad später) approximiert. Von diesen Teilpolynomen fordert man, dass sie an bestimmten Punkten die Differentialgleichung erfüllen. Diese Vorgehensweise unterscheidet sich von anderen Verfahren, in denen versucht wird, die in der Differentialgleichung auftretenden Ableitungen direkt zu diskretisieren (z.B.: FD-Verfahren).

2.1 Gitter und innere Knoten

Zuerst wählt man eine Zerlegungsfolge $\{x_k,\, k = 0 \dots K\}$ von $[a, b]$, $x_0 < x_1 < \cdots < x_K$ mit $x_0 = a$ und $x_K = b$. Im Folgenden bezeichnen wir die Zerlegungsfolge x_k bzw. die dazugehörigen Intervalle $[x_k, x_{k+1}]$ als *Kollokationsgitter*, bzw. *Kollokationsintervalle*. Zusätzlich zur Zerlegung von $[a, b]$ wählt man eine Zerlegung $\{\zeta_i,\, i = 0 \dots n\}$ von $[0, 1]$ folgendermaßen:

$$\zeta_0 < \zeta_1 < \cdots < \zeta_n \quad \text{mit} \quad \zeta_0 = 0 \quad \text{und} \quad \zeta_n = 1 \tag{2.1}$$

Die Folge ζ_i werden wir als *Kollokationsknotenverteilung* oder kurz *Knotenverteilung* bezeichnen. In jedem Kollokationsintervall $[x_k, x_{k+1}]$ erhält man durch lineare Transformation der Knotenverteilung $n - 1$ sogenannte *innere Kollokationsknoten* $x_{ki}, i = 1 \dots n-1$.

$$\begin{aligned} &\tau_k : \begin{cases} [0,1] & \to & [x_k, x_{k+1}] \\ x & \mapsto & x_k + x h_k \end{cases} \quad \text{mit} \quad h_k = x_{k+1} - x_k \\ &x_{ki} := \tau_k(\zeta_i) \quad k = 0 \dots K-1, \quad i = 1 \dots n-1 \end{aligned} \tag{2.2}$$

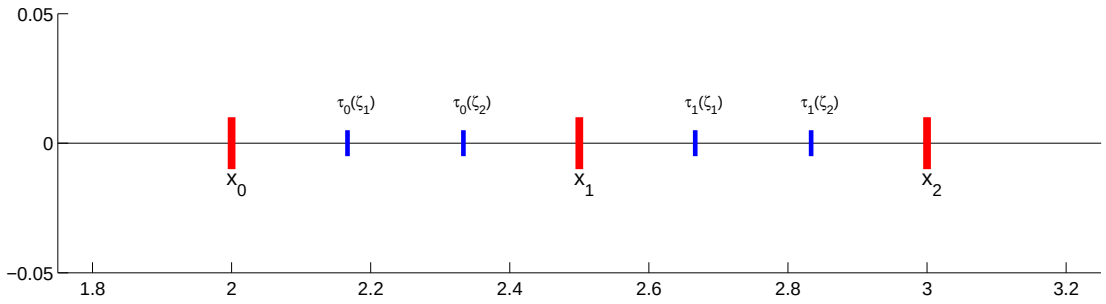

Abbildung 2.1: Die Abbildung zeigt eine äquidistante Zerlegung des Intervalls $[2,3]$ in 2 ($K = 1$) Kollokationsintervalle. Die kleinen Striche markieren die zugehörigen Kollokationsknoten ($n = 3$), die ebenfalls äquidistant verteilt sind.

2.2 Kollokationsgleichungen

Sei $p = \sum_{k=0}^{K-1} p_k$ ein stückweise zusammengesetztes Polynom in folgendem Sinn: Jedes der p_k hat den Träger $[x_k, x_{k+1}]$, außerhalb dieses Intervalls gilt also $p_k \equiv 0$. An allen inneren Kollokationsknoten x_{ki} fordert man nun, dass das entsprechende Teilpolynom p_k (1.3) erfüllt. Das ergibt ein System von $K(n-1)$ Gleichungen für die unbekannten Koeffizienten von p_k:

$$p_k''(x_{ki}) - f\left(x_{ki}, p_k(x_{ki}), p_k'(x_{ki})\right) = 0, \quad i = 1 \ldots n-1,\ k = 0 \ldots K-1 \tag{2.3}$$

Zu diesen $K(n-1)$ sogenannten *inneren Kollokationsbedingungen* benötigt man noch weitere Gleichungen, welche Stetigkeit und stetige Differenzierbarkeit zwischen Teilpolynomen $p_k,\ p_{k+1},\ k = 0 \ldots K-2$ an Übergangsstellen $x_{k+1}\ k = 0, 1, \ldots, K-2$, beschreiben (*Übergangsgleichungen*). Dadurch entstehen weitere $2(K-1)$ Gleichungen. Für die Formulierung von (2.4) verwendet man, dass für die Punkte $x_{kn} = x_{(k+1)0} = x_{k+1}$ gilt.

$$\begin{aligned}
&p_k''(x_{ki}) - f\left(x_{ki}, p_k(x_{ki}), p_k'(x_{ki})\right) = 0, \quad i = 1 \ldots n-1,\ k = 0 \ldots K-1 \\
&p_k(x_{k+1}) - p_{k+1}(x_{k+1}) = 0, \quad k = 0 \ldots K-2 \\
&p_k'(x_{k+1}) - p_{k+1}'(x_{k+1}) = 0, \quad k = 0 \ldots K-2
\end{aligned} \tag{2.4}$$

Ergänzt man dieses System noch um die Randbedingungen

$$R_1 \cdot \begin{pmatrix} y(a) \\ y(b) \end{pmatrix} + R_2 \cdot \begin{pmatrix} y'(a) \\ y'(b) \end{pmatrix} = \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} \tag{2.5}$$

bzw.

$$\begin{aligned}
R_{1_{11}} y(a) + R_{1_{12}} y(b) + R_{2_{11}} y'(a) + R_{2_{12}} y'(b) &= s_1 \\
R_{1_{21}} y(a) + R_{1_{22}} y(b) + R_{2_{21}} y'(a) + R_{2_{22}} y'(b) &= s_2,
\end{aligned} \tag{2.6}$$

ergibt sich ein System von $K(n+1)$ (linearen) Gleichungen für die Koeffizienten der Teilpolynome p_k.

Anm.: Durch diese Anzahl wird der Polynomgrad der Teilpolynome festgelegt: Von den $K(n+1)$ (linearen) Gleichungen gehören genau $n+1$ zum Teilpolynom p_k. Da ein Polynom n-ten Grades durch $n+1$ Koeffizienten bestimmt wird, ergibt sich also bei Teilpolynomgrad n ein (lineares) System mit $n+1$ Gleichungen in $n+1$ Unbekannten. Dementsprechend wählt man für den Grad der Teilpolynome n. Man beachte hierbei noch, dass der Polynomgrad ausschließlich von der Folge ζ_i bzw. genauer deren Länge abhängt.

Anstatt die Polynomkoeffizienten in Monombasis zu ermitteln, wird die Lösung im Folgenden mittels Lagrangepolynomen dargestellt. Für festes k seien $L_i(x) = l_{ki}(x)$ die Lagrangepolynome zu Knotenpunkten x_{ki} $i = 0 \dots n$, also

$$L_i(x) = \prod_{j=0\ j\neq i}^{n} \frac{x - x_{kj}}{x_{ki} - x_{kj}} \tag{2.7}$$

Diese Polynome erfüllen offensichtlich $L_i(x_{ks}) = 0$ für $s \neq i$ und $l_i(x_{ki}) = 1$. Ein Interpolationspolynom zu einer gegebenen Folge von Punkte-Werte-Paaren $(x_{kj}\ , f_j)$ erhält man mittels oben erwähnter Eigenschaften direkt als $\sum_{j=0}^{n} f_j L_j(x)$. Die in dieser Darstellung ermittelten Koeffizienten entsprechen also direkt den Werten des Teilpolynoms p_k an den Kollokationsknoten x_{ki}.

2.3 Polynomableitungen

Um die Ableitungen an Stellen x_{ki} der Teilpolynome p_k darzustellen, machen wir uns folgenden Zusammenhang zunutze:

Lemma 2.3.1 *Sei p ein Polynom n-ten Grades und sei zusätzlich ζ_i wie in (2.1). Dann gibt es Matrizen $\Omega = (\omega_{ij})$ und $\tilde{\Omega} = (\tilde{\omega}_{ij})$ mit der Eigenschaft, dass*

$$\begin{aligned} p'(\zeta_i) &= \sum_{j=0}^{n} \omega_{ij}\, p(\zeta_j), \quad i = 0 \dots n \\ p''(\zeta_i) &= \sum_{j=0}^{n} \tilde{\omega}_{ij}\, p(\zeta_j), \quad i = 0 \dots n \end{aligned}$$

gilt. Eine konkrete Darstellung der Matrizen Ω bzw. $\tilde{\Omega}$ ist durch (2.10) bzw. (2.18) gegeben.

Beweis Zur Folge ζ_i seien l_j die Lagrange Polynome zu den Knoten ζ_j, also $l_i(\zeta) = \prod_{j=0\ j\neq i}^{n} \frac{\zeta-\zeta_j}{\zeta_i-\zeta_j}$. Für ein beliebiges Polynom $p(\zeta) = \sum_{k=0}^{n} \alpha_k\, l_k(\zeta)$ vom Grad n gilt nun

$p(\zeta_i) = \alpha_i$. Für die Ableitung $p'(\zeta)$ berechnen wir zunächst die Ableitungen der $l_i(\zeta)$:

$$l_i'(\zeta) \;=\; \sum_{\substack{k=0\\k\neq i}}^{n} \frac{1}{\zeta_i - \zeta_k} \prod_{\substack{v=0\\v\neq i,k}}^{n} \frac{\zeta - \zeta_v}{\zeta_i - \zeta_v} \tag{2.8}$$

woraus sich für $\zeta = \zeta_i$ sofort $l_i'(\zeta_i) = \sum_{\substack{k=0\\\neq i}}^{n} \frac{1}{\zeta_i - \zeta_k}$ ergibt. Betrachtet man eine Stelle $\zeta = \zeta_j,\ j \neq i$, so erhält man

$$l_i'(\zeta_j) \;=\; \sum_{\substack{k=0\\k\neq i}}^{n} \frac{1}{\zeta_i - \zeta_k} \underbrace{\prod_{\substack{v=0\\v\neq i,k}}^{n} \frac{\zeta_j - \zeta_v}{\zeta_i - \zeta_v}}_{=0 \text{ für } k\neq j} \;=\; \frac{1}{\zeta_i - \zeta_j} \prod_{\substack{v=0\\v\neq i,j}}^{n} \frac{\zeta_j - \zeta_v}{\zeta_i - \zeta_v} \tag{2.9}$$

Zusammensetzen von $p'(\zeta_i)$ liefert letztendlich $p'(\zeta_i) = \sum_{k=0}^{n} \alpha_k\, l_k'(\zeta_i)$, die Matrixeinträge ω_{ij} ergeben sich durch Koeffizientenvergleich:

$$p'(\zeta_i) \;=\; \sum_{k=0}^{n} \alpha_k\, l_k'(\zeta_i) \;=\; \sum_{j=0}^{n} \omega_{ij}\, p(\zeta_j) \;=\; \sum_{j=0}^{n} \omega_{ij}\, \alpha_j$$

Es gilt also $\omega_{ij} = l_j'(\zeta_i)$, d.h.

$$\omega_{ij} \;=\; \begin{cases} \displaystyle\sum_{\substack{k=0\\k\neq i}}^{n} \frac{1}{\zeta_i - \zeta_k} & \text{für } j = i \\[2ex] \displaystyle\frac{1}{\zeta_j - \zeta_i} \prod_{\substack{k=0\\k\neq i,j}}^{n} \frac{\zeta_i - \zeta_k}{\zeta_j - \zeta_k} & \text{für } j \neq i \end{cases} \tag{2.10}$$

Anm.: Betrachtet man $\frac{1}{\omega_{ji}} = (\zeta_i - \zeta_j) \overbrace{\prod_{k\neq i,j} \frac{\zeta_i - \zeta_k}{\zeta_j - \zeta_k}}^{\text{selbes Produkt in } \omega_{ij}}$ erhält man einen Zusammenhang bei Vertauschung der Indizes i und j:

$$\omega_{ij} = \frac{-1}{(\zeta_j - \zeta_i)^2}\, \omega_{ji} \tag{2.11}$$

Für die zweite Ableitung führt ähnliches Vorgehen zum Ziel. In diesem Fall berechnet man:

$$\begin{aligned} l_i''(\zeta) &= \left(\sum_{\substack{k=0\\k\neq i}}^{n} \frac{1}{\zeta_i - \zeta_k} \prod_{\substack{v=0\\v\neq i,k}}^{n} \frac{\zeta - \zeta_v}{\zeta_i - \zeta_v} \right)' \\ &= \sum_{\substack{u=0\\u\neq i}}^{n} \frac{1}{\zeta_i - \zeta_u} \sum_{\substack{k=0\\k\neq u,i}}^{n} \frac{1}{\zeta_i - \zeta_k} \prod_{\substack{v=0\\v\neq u,i,k}}^{n} \frac{\zeta - \zeta_v}{\zeta_i - \zeta_v} \end{aligned} \tag{2.12}$$

Daraus ergibt sich wie oben für $\zeta = \zeta_i$ sofort (das auftretende Produkt ist immer = 1):

$$l_i''(\zeta_i) = \sum_{\substack{u=0 \\ u\neq i}}^{n} \frac{1}{\zeta_i - \zeta_u} \sum_{\substack{k=0 \\ k\neq u,i}}^{n} \frac{1}{\zeta_i - \zeta_k} \tag{2.13}$$

Für $j \neq i$ rechnet man nach:

$$l_i''(\zeta_j) = \sum_{\substack{u=0 \\ u\neq i}}^{n} \frac{1}{\zeta_i - \zeta_u} \sum_{\substack{k=0 \\ k\neq u,i}}^{n} \frac{1}{\zeta_i - \zeta_k} \overbrace{\prod_{\substack{v=0 \\ v\neq u,i,k}}^{n} \frac{\zeta_j - \zeta_v}{\zeta_i - \zeta_v}}^{=0 \text{ für } u,i,k\neq j}, \tag{2.14}$$

also verschwinden in der ersten Summe alle Summanden außer für $u = j$. In der inneren Summe bleibt ebenso nur ein Summand, mit $k = j$. Man erhält zuerst für $u = j$

$$\frac{1}{\zeta_i - \zeta_j} \sum_{\substack{k=0 \\ k\neq j,i}}^{n} \frac{1}{\zeta_i - \zeta_k} \prod_{\substack{v=0 \\ v\neq j,i,k}}^{n} \frac{\zeta_j - \zeta_v}{\zeta_i - \zeta_v} \tag{2.15}$$

und für $k = j$

$$\sum_{\substack{u=0 \\ u\neq i,j}}^{n} \frac{1}{\zeta_i - \zeta_u} \frac{1}{\zeta_i - \zeta_j} \prod_{\substack{v=0 \\ v\neq u,i,j}}^{n} \frac{\zeta_j - \zeta_v}{\zeta_i - \zeta_v} \tag{2.16}$$

$l_i''(\zeta_j)$ errechnet man als Summe der letzten beiden Zeilen zu

$$l_i''(\zeta_j) = \frac{2}{\zeta_i - \zeta_j} \sum_{\substack{k=0 \\ k\neq j,i}}^{n} \frac{1}{\zeta_i - \zeta_k} \prod_{\substack{v=0 \\ v\neq j,i,k}}^{n} \frac{\zeta_j - \zeta_v}{\zeta_i - \zeta_v} \tag{2.17}$$

Wie im Fall der ersten Ableitung erhält man $p''(\zeta_i) = \sum_{k=0}^{n} \alpha_k \, l_k''(\zeta_i)$, und die $\tilde{\omega}_{ij}$ durch Koeffizientenvergleich:

$$p''(\zeta_i) \;=\; \sum_{k=0}^{n} \alpha_k \, l_k''(\zeta_i) \;=\; \sum_{j=0}^{n} \tilde{\omega}_{ij} \, p(\zeta_j) \;=\; \sum_{j=0}^{n} \omega_{ij} \, \alpha_j$$

Also $\tilde{\omega}_{ij} = l_j''(\zeta_i)$, was zu:

$$\tilde{\omega}_{ij} \;=\; \begin{cases} \displaystyle\sum_{\substack{u=0 \\ u\neq i}}^{n} \frac{1}{\zeta_i - \zeta_u} \sum_{\substack{k=0 \\ k\neq u,i}}^{n} \frac{1}{\zeta_i - \zeta_k} & \text{für } j = i \\ \displaystyle\frac{2}{\zeta_j - \zeta_i} \sum_{\substack{k=0 \\ k\neq i,j}}^{n} \frac{1}{\zeta_j - \zeta_k} \prod_{\substack{v=0 \\ v\neq i,j,k}}^{n} \frac{\zeta_i - \zeta_v}{\zeta_j - \zeta_v} & \text{für } j \neq i \end{cases} \tag{2.18}$$

führt. □

Umskalierung auf ein beliebiges Intervall $[x, x+h]$ ergibt:

Proposition 2.3.2 *Für Folgen ν mit $\nu_0 = x$ und $\nu_n = x+h$ existieren Differentiationsmatrizen Ω^h und $\tilde{\Omega}^h$ mit den Eigenschaften aus Satz 2.3.1. Außerdem gelten die Zusammenhänge*

$$\Omega^h = \frac{1}{h}\Omega \qquad \tilde{\Omega}^h = \frac{1}{h^2}\tilde{\Omega}, \tag{2.19}$$

wobei Ω bzw. $\tilde{\Omega}$ die Differentiationsmatrizen für das Standardintervall $[0,1]$ bezeichnen.

Beweis Wir betrachten die Abbildung $g : [0,1] \to [x, x+h]$ mit $g(y) = x + hy$. Da diese Abbildung bijektiv ist, kann man zu jedem ν_i genau ein Urbild finden, wir bezeichnen dieses mit ζ_i. Für das Polynom $p \circ g$ existiert nach Satz 2.3.1 eine Matrix Ω sodass $(p \circ g)'(\zeta_i)) = \sum_{j=0}^{n} \omega_{ij}\,(p \circ g)(\zeta_i)$ erfüllt ist. Wendet man auf die linke Seite dieser Gleichung die Kettenregel an, erhält man $(p \circ g)'(\zeta_i) = p'(g(\zeta_i))g'(\zeta_i)$. Setzt man beide Gleichungen zusammen erhält man

$$\begin{aligned} p'(g(\zeta_i))g'(\zeta_i) &= \sum_{j=0}^{n} \omega_{ij}\,(p \circ g)(\zeta_i) \Leftrightarrow \\ p'(\nu_i)h &= \sum_{j=0}^{n} \omega_{ij}\,p(\nu_i) \Leftrightarrow \\ p'(\nu_i) &= \sum_{j=0}^{n} \tfrac{1}{h}\omega_{ij}\,p(\nu_i) \end{aligned} \tag{2.20}$$

Nochmaliges Ableiten führt zum behaupteten Zusammenhang bezüglich der zweiten Ableitung.

□

Mithilfe von Proposition 2.3.2 lassen sich nun die Kollokationsgleichungen folgendermaßen formulieren:

$$\begin{aligned} &\frac{1}{{h_k}^2}\sum_{j=0}^{n} \tilde{\omega}_{ij}\,p_k(x_{ki}) - f\left(x_{ki}, p_k(x_{ki}), \frac{1}{h_k}\sum_{j=0}^{n}\omega_{ij}\,p_k(x_{ki})\right) = 0, \\ &i = 1 \ldots n-1,\ k = 0 \ldots K \\ &p_k(x_k) - p_{k+1}(x_k) = 0, \quad k = 0 \ldots K-2 \\ &\frac{1}{h_k}\sum_{j=0}^{n}\omega_{nj}\,p_k(x_k) - \frac{1}{h_{k+1}}\sum_{j=0}^{n}\omega_{0j}\,p_{k+1}(x_k) = 0, \quad k = 0 \ldots K-2. \end{aligned} \tag{2.21}$$

2.3.1 Beispiele

Sei die Folge $\{\zeta_i\}$ gegeben durch $\{0, \frac{1}{2}, 1\}$. Dann gilt für beliebige Polynome vom Grad 2

$$p'(\zeta_i) \;=\; \sum_{j=0}^{2} \omega_{ij}\, p(\zeta_j), \quad i = 0,1,2, \quad \text{mit} \quad \omega_{ij} = \begin{pmatrix} -3 & 4 & 1 \\ -1 & 0 & 1 \\ -1 & -4 & 3 \end{pmatrix} \tag{2.22}$$

sowie

$$p''(\zeta_i) \;=\; \sum_{j=0}^{2} \omega_{ij}\, p(\zeta_j), \quad i = 0,1,2, \quad \text{mit} \quad \tilde{\omega}_{ij} = \begin{pmatrix} 4 & -8 & 4 \\ 4 & -8 & 4 \\ 4 & -8 & 4 \end{pmatrix} \tag{2.23}$$

2.4 Aufbau der Systemmatrix

Im Fall einer linearen Differentialgleichung sind alle Kollokationsgleichungen linear. Dementsprechend lassen sich die Kollokationsgleichungen in der Form $Mp = g$ darstellen, wobei M eine $K(n+1) \times K(n+1)$ Matrix ist und g ein Spaltenvektor der Länge $K(n+1)$. Die Systemmatrix M lässt sich blockweise aufbauen, ein Block entspricht dabei immer einem Kollokationsintervall. Wir werden die Übergangsgleichung für p nach links in die erste Zeile des Matrixblock $M_k \in \mathbb{R}^{(n+1)\times K(n+1)}$, die inneren Kollokationsbedingungen in die Zeilen 2 bis n, und die Übergangsgleichung für p' nach rechts in die letzte Zeile des Blockes M_k schreiben. Im Block M_0 wird statt der Übergangsgleichung nach links die erste Zeile der Randbedingung eingetragen, in M_K ersetzt man die Übergangsbedingung nach rechts durch die zweite Zeile der Randbedingung. Der Aufbau des Gewichtevektors g wird genauso durchführt wie der Aufbau der Systemmatrix M.

2.4.1 Innere Knotenpunkte

Für jedes Teilintervall $[x_k, x_{k+1}]$, $k = 0 \ldots K-1$, erhalten wir $n-1$ Kollokationsgleichungen und 2 Übergangsbedingungen. Im ersten sowie im letzten Intervall muss je eine der beiden Übergangsbedingungen durch eine Randgleichung ersetzt werden. Seien nun q^1_{ki}, $i = 1 \ldots n-1$, die Werte von $q(\tau_k(\zeta_i))$, $i = 1 \ldots n-1$ und Q_{k1} die Diagonalmatrix $\mathrm{diag}(q^1_{k1}, q^1_{k2}, \ldots, q^1_{k(n-1)})$. Ebenso lassen sich die Matrix Q_{k2}, und der Vektor g_k mit $g_{ki} = g(\tau_k(\zeta_i))$, $i = 1 \ldots n-1$ definieren. Wir können die inneren Kollokationsgleichungen im Intervall $[x_k, x_{k+1}]$ nun in Matrixform anschreiben:

$$\frac{1}{h_k^2}\tilde{\Omega} p_k + \frac{1}{h_k} Q_{k1}\, \Omega\, p_k + Q_{k2}\, p_k \;=\; \underbrace{\left(\frac{1}{h_k^2}\tilde{\Omega} + \frac{1}{h_k} Q_{k1}\, \Omega + Q_{k2}\right)}_{:= R_k \in \mathbb{C}^{(n-1)\times(n-1)}} p_k \;=\; g_k$$

Nun lässt sich der k-te $(n+1) \times K(n+1)$-Block der Systemmatrix M (Übergangsbedingungen nicht berücksichtigt!) als

$$\begin{pmatrix} 0 & \dots\dots\dots & 0 & 0 & . & 0 & 0 & \dots\dots\dots & 0 \\ \vdots & \ddots & \vdots & R_{k_{00}} & . & R_{k_{0n}} & \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots & \vdots & & \vdots & \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots & R_{k_{(n-2)0}} & . & R_{k_{(n-2)n}} & \vdots & \ddots & \vdots \\ 0 & \dots\dots\dots & 0 & 0 & . & 0 & 0 & \dots\dots\dots & 0 \end{pmatrix} \tag{2.24}$$

anschreiben.

2.4.2 Übergangsbedingungen

Für die Übergangsgleichungen bezüglich der Ableitungen $p'_k(x_{k+1})$ bzw. $p'_{k+1}(x_{k+1})$ verwendet man wieder die Matrix Ω:

$$\begin{aligned} p'_k(x_{k+1}) &= p'_{k+1}(x_{k+1}),\ k = 0 \dots K-2 \quad \Leftrightarrow \\ \frac{1}{h_k} \sum_{j=0}^{n} \omega_{nj}\, p_{kj} &= \frac{1}{h_{k+1}} \sum_{j=0}^{n} \omega_{0j}\, p_{(k+1)j} \end{aligned} \tag{2.25}$$

Das entspricht $2(K-1)$ weiteren Gleichungen. Wir nehmen die beiden Gleichungen, die dem k-ten Block entsprechen und tragen diese jeweils in die erste und letzte Zeile von (2.24) ein:

$$\begin{pmatrix} 0 & \dots\dots & 0 & 1 & -1 & . & 0 & 0 & \dots\dots\dots & & 0 \\ \vdots & \ddots & & \vdots & R_{k_{00}} & . & R_{k_{0n}} & \vdots & & & \vdots \\ \vdots & \ddots & & \vdots & \vdots & & \vdots & \vdots & & & \vdots \\ \vdots & \ddots & & \vdots & R_{k_{(n-2)0}} & . & R_{k_{(n-2)n}} & 0 & & & 0 \\ 0 & \dots\dots\dots & 0 & h_k^{-1}\omega_{n0} & . & h_k^{-1}\omega_{nn} & h_{k+1}^{-1}\omega_{00} & h_{k+1}^{-1}\omega_{01} & \dots & h_{k+1}^{-1}\omega_{0n} & 0 \end{pmatrix}$$

Der Gewichtevektor g_k auf der rechten Seite muss um eine erste Zeile 0 ergänzt und eine letzte Zeile 0 ergänzt werden.

2.4.3 Randbedingungen

Als letztes werden wir die beiden Gleichungen behandeln, die die Randbedingungen beschreiben:

$$R_1 \begin{pmatrix} y(a) \\ y(b) \end{pmatrix} + R_2 \begin{pmatrix} y'(a) \\ y'(b) \end{pmatrix} = \begin{pmatrix} s_1 \\ s_2 \end{pmatrix}$$

Im ersten Block ersetzt man nun die erste Zeile durch jene Gleichung, die die erste Gleichung der Randbedingung beschreibt. Ebenso ersetzt man im letzten Block die letzte Zeile durch die zweite Gleichung der Randbedingung. Die auftretenden Variablen $y(a)$, $y(b)$, $y'(a)$ sowie $y'(b)$ sind mit p_{00}, $p_{(K-1)n}$, $h_0^{-1} \sum_{k=0}^{n} \omega_{0j}\, p_{0j}$ sowie $h_{K-1}^{-1} \sum_{k=0}^{n} \omega_{nj}\, p_{(K-1)j}$ zu identifizieren.

2.5 Konvergenzaussagen

Betrachtet man die Lösung p von (2.4) – erweitert um 2 Randgleichungen – so gilt bei hinreichender Glattheit von q_1, q_2 und g folgende Konvergenzaussage für die Kollokationslösung p:

Satz 2.5.1 *Für das Randwertproblem (1.2) existiere eine eindeutige Lösung $u \in C^2(a,b)$. Weiters sei eine Knotenverteilung ζ_i, $i = 0 \dots n$, gegeben sowie ein Kollokationsgitter x_k, $k = 0 \dots K$. Für das mittels (2.4) ermittelte stückweise Polynom $p \in C^1(a,b)$ gelten die folgenden Approximationsaussagen:*

$$\begin{aligned} ||e|| &= ||p-u|| = O(h^n) \\ ||e'|| &= ||p'-u'|| = O(h^n) \\ ||e''|| &= ||p''-u''|| = O(h^n), \end{aligned} \tag{2.26}$$

wobei h die Schrittweite $x_k - x_{k-1} \equiv h$ bezeichnet. Die zweite Ableitung des Polynoms p weist an den Stellen x_k Sprungstellen auf, an diesen muss in der Konvergenzaussage also auf eine einseitige Ableitung zurückgegriffen werden. Wählt man eine symmetrische Knotenverteilung ζ_i mit ungerader Länge, so kann man die Ordnung des Verfahrens um 1 verbessern. Dieser Effekt wird als (schwache) Superkonvergenz bezeichnet, man erhält demnach $O(h^{n+1})$ anstatt $O(h^n)$ als Konvergenzordnung.

Der Beweis dieser Aussage wird in dieser Diplomarbeit nicht geführt und findet sich in unter anderem in [6].

Anm. 1: Bei nichtäquidistantem Kollokationsgitter betrachtet man anstelle von h die maximale Intervalllänge, also $h_{nä} = \max\{x_k - x_{k-1} | k = 1 \dots K\}$. Anstelle dieser Variante kann man auch jeden Punkt mit einer Intervalllänge h_k in Verbindung bringen, indem man $h_k := \frac{x_{k+1}-x_{k-1}}{2}$ definiert. h_k stellt somit das arithmetische Mittel der Intervalllängen $x_k - x_{k-1}$ sowie $x_{k+1} - x_k$ dar.

Anm. 2: Für spezielle Kollokationsknoten erhält man an den Endpunkten der Kollokationsintervalle eine noch höhere „Superkonvergenz"-Ordnung, z.B. $O(h^{2n})$ im Fall von Gauß-Knoten (analog wie bei der Gauß-Legendre-Quadratur). Die im Folgenden beschriebene Methode zur Fehlerschätzung zielt jedoch nicht auf eine Schätzung dieser speziellen Fehlerwerte ab.

Kapitel 3

Differenzenschema für Fehler und Fehlerschätzer

In diesem Kapitel wird eine exakte Differenzengleichung für den Fehler $e = p - u$, wobei u die exakte Lösung und p die durch Kollokation ermittelte numerische Lösung von (1.3) bezeichnen, hergeleitet. Hierfür wird zuerst eine Differentialgleichung betrachtet, die der Fehler erfüllt, und durch Anwendung eines Integraloperators sodann in eine Differenzengleichung umgewandelt. Aus dieser Differenzengleichung wird anschließend ein Fehlerschätzer konstruiert, wobei hier die unterschiedlichen Formen von Randbedingungen in den Abschnitten 3.2.1, 3.2.2 sowie 3.2.3 ausführlich behandelt werden. Zuerst werden allerdings die für die folgenden Schritte benötigten ersten sowie zweiten zentralen Differenzenquotienten auf nicht äquidistanten Gittern hergeleitet und der Defekt definiert.

3.1 Vorbereitungen

3.1.1 Defekt der Lösung bezüglich des Randwertproblems

Definition Sei p die durch Kollokation erhaltene Lösung von (1.2). Das Residuum beim Einsetzen in die Differentialgleichung wird als punktweiser Defekt d der Kollokationslösung bezüglich der Differentialgleichung bezeichnet:

$$\begin{aligned} &d(x) := p''(x) - f(x, p(x), p'(x)), \quad \text{bzw. in linearer Schreibweise} \\ &d(x) = p''(x) + q_1(x)p'(x) + q_2(x)p(x) - g(x) \end{aligned} \tag{3.1}$$

Für die Kollokationslösung erkennt man unmittelbar, dass an allen inneren Kollokationsknoten $d(x_{ki}) = 0$ gilt. An diesen Stellen ergibt sich demnach keine verwertbare Information, um einen Fehlerschätzer zu konstruieren. Um hier Abhilfe zu schaffen, wird in Abschnitt 3.2.1 eine integrierte Variante des punktweisen Defektes betrachtet, mit Hilfe derer man einen Fehlerschätzer konstruieren kann.

3.1.2 Zentrale Differenzenquotienten

Um den Fehler $p(x) - u(x)$ schätzen zu können, wird die Differentialgleichung umformuliert. Anstatt der zweiten Ableitung betrachtet man den zweiten zentralen Differenzenquotienten $\delta_h^2(t, f)$,

$$\delta_h^2(t, f) := \frac{f(t-h) - 2f(t) + f(t+h)}{h^2} \tag{3.2}$$

Diese Formel führt allerdings nur bei äquidistanten Intervalllängen $[t-h, t]$ und $[t, t+h]$ zu einer Approximation der zweiten Ableitung, dabei gilt folgende Approximationsaussage:

Proposition 3.1.1 *Sei die Funktion $f \in C^4(a, b)$, dann gilt:*

$$|f''(t) - \delta_h^2(t, f)| = O(h^2). \tag{3.3}$$

Beweis Der Zusammenhang ergibt sich unmittelbar mittels Taylorentwicklung bis zur Ordnung 3,

$$f(t \pm h) = f(t) \pm hf'(t) + \tfrac{h^2}{2}f''(t) \pm \tfrac{h^3}{6}f'''(t) + \tfrac{h^4}{24}f^{IV}(\xi) \tag{3.4}$$

Man beachte, dass es sich dabei um zwei unterschiedliche Zwischenstellen $\xi = \xi^+, \xi^-$ handelt. Bildet man nun den Ausdruck $f''(t) - \delta_h^2(t, f)$, so erkennt man, dass alle Terme mit ungeraden h-Potenzen, einander wegkürzen. Ebenso fallen die h^2-Terme weg, und es bleibt $\frac{h^4}{24h^2}|f^{IV}(\xi^+) + f^{IV}(\xi^-)|$ übrig.

□

Um den 2-ten zentralen Differenzenquotienten für beliebige Intervalllängen zu erhalten, sind die einzelnen Funktionsauswertungen anders zu gewichten:

Satz 3.1.2 *Für Funktionen $f \in C^3(a, b)$ existiert eine Linearkombination aus Funktionswerten $f(t_0),\ f(t),\ f(t_1)$ mit $t_0 < t < t_1$, sodass*

$$\big|a_1 f(t_0) + a_2 f(t) + a_3 f(t_1) - f''(t)\big| = O(h) \tag{3.5}$$

erfüllt ist. Diese Linearkombination wird im Weiteren mit $\delta^2_{h_0,h_1}(t, f)$ bezeichnet.

Beweis Um den Satz zu beweisen, wird mittels Taylorapproximation ein Gleichungssystem für die Koeffizienten $a_i, i = 1 \dots 3$, hergeleitet, welches eine eindeutige Lösung besitzt. Von dieser Lösung rechnet man unmittelbar die im Satz erklärte Konvergenz erster Ordnung nach. Für zwei Intervalle $[t_0, t]$ und $[t, t_1]$ mit $h_0 = t - t_0$ sowie $h_1 = t_1 - t$ betrachtet man den Ausdruck $d^2_{h_0,h_1}(t, f) = a_1 f(t_{i-1}) + a_2 f(t_i) + a_3 f(t_{i+1})$. Taylorentwicklung von f um die Stelle t liefert:

$$f(t - h_0) = f(t) - h_0 f'(t) + \frac{h_0^2}{2}f''(t) - \frac{h_0^3}{6}f'''(\xi^-), \quad \xi^- \in [t_0, t] \tag{3.6}$$

bzw.

$$f(t+h_1) = f(t) + h_1 f'(t) + \frac{h_1^2}{2} f''(t) + \frac{h_1^3}{6} f'''(\xi^+), \quad \xi^+ \in [t, t_1] \tag{3.7}$$

Setzt man diese beiden Ausdrücke in $d_h^2(t,f)$ ein, so erhält man:

$$\begin{aligned} d^2_{h_0,h_1}(t,f) = a_1 \left(f(t) - h_0 f'(t) + \frac{h_0^2}{2} f''(t) - \frac{h_0^3}{6} f'''(\xi^-) \right) + \\ a_2 f(t) + a_3 \left(f(t) + h_1 f'(t) + \frac{h_1^2}{2} f''(t) + \frac{h_1^3}{6} f'''(\xi^+) \right) \end{aligned} \tag{3.8}$$

Für alle weiteren Schritte setzen wir $\alpha = \frac{h_0}{h_1}$ als das Verhältnis zwischen den beiden Intervalllängen fest.
Sortiert man nach Ableitungsgraden von f, so ergibt sich:

$$\begin{aligned} d^2_{h_0,h_1}(t,f) = f(t)\left(a_1 + a_2 + a_3\right) + f'(t)\left(-h_0 a_1 + h_1 a_3\right) + \\ f''(t)\left(\frac{h_0^2 a_1 + h_1^2 a_3}{2}\right) + \left(-\frac{f'''(\xi^-) a_1 h_0^3}{6} + \frac{f'''(\xi^+) a_3 h_1^3}{6} \right) \end{aligned} \tag{3.9}$$

Damit ergeben sich für die Koeffizienten a_1, a_2 und a_3 folgende 3 Gleichungen, die notwendigerweise erfüllt sein müssen, damit $d^2_{h_0,h_1}(t,f) = f''(t) + O(h_1)$ gilt:

$$\begin{aligned} a_1 + a_2 + a_3 &= 0 \\ -h_0 a_1 + h_1 a_3 &= 0 \\ h_0^2 a_1 + h_1^2 a_3 &= 2 \end{aligned} \tag{3.10}$$

Lösen dieses linearen Gleichungssystems führt zu:

$$a_1 = \frac{2}{h_1^2} \frac{1}{\alpha(\alpha+1)} \qquad a_2 = \frac{2}{h_1^2} \frac{-1}{\alpha} \qquad a_3 = \frac{2}{h_1^2} \frac{1}{\alpha+1} \tag{3.11}$$

Setzt man also für ein $f \in C^3(a,b)$

$$\delta^2_{h_0,h_1}(t,f) = \frac{2\left(f(t-h_0) - (\alpha+1) f(t) + \alpha f(t+h_1)\right)}{h_1^2 \alpha(\alpha+1)}, \tag{3.12}$$

so gilt nach obiger Rechnung

$$|f''(t) - \delta^2_{h_0,h_1}(t,f)| \leq \frac{|f'''(\xi^-)| + |f'''(\xi^+)|}{6} (a_1 h_0^3 + a_3 h_1^3), \tag{3.13}$$

was nach einsetzen von a_1 bzw. a_3 zu

$$|f''(t) - \delta^2_{h_0,h_1}(t,f)| \leq \frac{|f'''(\xi^-)| + |f'''(\xi^+)|}{6} \frac{2\alpha h_0 + 2h_1}{\alpha + 1} = O(h_0 + h_1) \tag{3.14}$$

führt. □

Anm.: So wie für die zweite kann auch für die erste Ableitung ein Differenzenquotient angegeben werden, der bei Schrittlängen h_0 und h_1 mit der Ordnung $O(h_1)$ konvergiert, indem man

$$\delta^1_{h_0,h_1}(t,f) = \frac{1}{h_1(\alpha+1)} \left(f(t+h_1) - f(t-h_0)\right) \tag{3.15}$$

setzt. Ähnlich wie oben ergibt sich $O(h)$ Konvergenzverhalten, bzw. im äquidistanten Fall $O(h^2)$ als Konvergenzordnung.

3.1.3 Integraloperatoren zur Diskretisierung der Ableitung

Wir betrachten zuerst den Integraloperator $\mathcal{I}$, definiert als:

$$(\mathcal{I}u)(t_i) := \int_{-1}^{0} K(\zeta) u(t_i + h_{i-1}\zeta)\, d\zeta + \int_{0}^{1} K(\zeta) u(t_i + h_i\zeta)\, d\zeta, \tag{3.16}$$

wobei $K(\zeta) = 1 - |\zeta|$ ist. Mithilfe dieses Operators wird sich eine Differenzenformulierung für die zweite Ableitung einer Funktion f ergeben. Für $f \in C^2(a,b)$ gilt:

Satz 3.1.3 *Für $u \in C^2(t-h, t+h)$ liefert die Anwendung des Operators $\mathcal{I}$ unter der zusätzlichen Voraussetzung $h_0 = h_1 = h$ den zweiten zentralen Differenzenquotienten von u:*

$$\delta^2_{h,h}(t,u) = (\mathcal{I}u'')(t) \tag{3.17}$$

Beweis Um Satz 3.1.3 zu beweisen, hilft partielle Integration: Betrachtet man zuerst den linken Integralanteil, ergibt sich mittels partieller Integration:

$$\int_{-1}^{0} (1-|\zeta|) u''(t+h\zeta) d\zeta = \left.\frac{u'(t+h\zeta)(1+\zeta)}{h}\right|_{-1}^{0} - \int_{-1}^{0} \frac{u'(t+h\zeta)}{h} d\zeta, \tag{3.18}$$

was sich weiter zu

$$\begin{aligned} \int_{-1}^{0} (1-|\zeta|) u''(t+h\zeta) d\zeta &= \frac{u'(t)}{h} - \left.\frac{u(t+h\zeta)(1+\zeta)}{h^2}\right|_{-1}^{0} \\ &= \frac{u'(t)}{h} - \left(\frac{u(t) - u(t-h)}{h^2}\right) \end{aligned} \tag{3.19}$$

umformulieren lässt. Für den rechten Integralanteil ergibt sich auf analoge Art und Weise

$$\int_{-1}^{0}(1-|\zeta|)u''(t+h\zeta)d\zeta = \frac{-u'(t)}{h} + \left(\frac{u(t+h)-u(t)}{h^2}\right) \tag{3.20}$$

Summiert man beide Integralanteile auf, kürzen die beiden Ableitungsterme einander weg und es bleibt

$$\int_{-1}^{1}(1-|\zeta|)u''(t+h\zeta)d\zeta = \left(\frac{u(t+h)+u(t-h)-2u(t)}{h^2}\right), \tag{3.21}$$

also genau der zweite zentrale Differenzenquotient übrig. □

Im Fall von $h_0 \neq h_1$ sind die beiden Integralanteile entsprechend zu gewichten: Setzt man $\mathcal{I}$ als

$$(\mathcal{I}u)(t_i) := l\int_{-1}^{0} K(\zeta)u(t_i+h_{i-1}\zeta)\,d\zeta + r\int_{0}^{1} K(\zeta)u(t_i+h_i\zeta)\,d\zeta, \tag{3.22}$$

fest, so ergeben sich durch Koeffizientenvergleich 4 Bedingungen für die Gewichte l und r.

Satz 3.1.4 *Sei der Operator $\mathcal{I}$ wie in Definition 3.22 definiert und die Gewichte l und r durch $l = \frac{2\alpha}{1+\alpha}$ sowie $r = \frac{2}{1+\alpha}$ gegeben, wobei wie oben $\alpha = \frac{h_0}{h_1}$ gilt. Für 2-mal stetig differenzierbares u liefert die Anwendung von $\mathcal{I}$ den zweiten zentralen Differenzenquotienten von u auf nichtäquidistantem Gitter:*

$$\delta^2_{h_0,h_1}(t,u) = (\mathcal{I}u'')(t) \tag{3.23}$$

Für $h_0 = h_1$ gilt $\alpha = 1$ und damit $l = r = 1$.

Beweis Addiert man die beiden Integrale, so erhält man nach partieller Integration:

$$(\mathcal{I}u)(t) = \frac{-lu'(t)}{h_0} + \frac{ru'(t)}{h_1} + l\left(\frac{u(t-h_0)-u(t)}{h_0^2}\right) + r\left(\frac{u(t+h_1)-u(t)}{h_1^2}\right) \tag{3.24}$$

Damit die beiden Ableitungsterme wegfallen, muss also $\frac{r}{h_1} = \frac{l}{h_0}$ gelten. Der Koeffizient vor $u(t)$ ist durch $\frac{-2}{h_1^2\alpha}$ gegeben, dementsprechend muss zusätzlich $\frac{l}{h_0^2} + \frac{r}{h_1^2} = \frac{2}{h_1^2\alpha}$ gelten. Berücksichtigt man noch $h_0 = \alpha h_1$ erhält man 2 Gleichungen, durch die l und r bestimmt sind:

$$\begin{aligned} \frac{l}{\alpha h_1} &= \frac{r}{h_1} \Leftrightarrow l = \alpha r \\ \frac{l}{\alpha^2 h_1^2} + \frac{r}{h_1^2} &= \frac{2}{h_1^2\alpha} \Leftrightarrow r + \frac{l}{\alpha^2} = \frac{2}{\alpha} \end{aligned} \tag{3.25}$$

Daraus ergeben sich die beiden Werte

$$l = \frac{2\alpha}{1+\alpha} \text{ und } r = \frac{2}{1+\alpha} \tag{3.26}$$

Es bleibt zu überprüfen, ob mit diesen beiden Werten auch die Bedingungen $r\frac{u(t+h_1)}{h_1^2} = \frac{2u(t+h_1)}{h_1^2(\alpha+1)}$, sowie $l\frac{u(t-h_0)}{h_0^2} = \frac{2u(t-h_0)}{h_1^2(\alpha+1)\alpha}$ erfüllt sind. Dies ist aber der Fall, denn

$$\begin{aligned} r\frac{u(t+h_1)}{h_1^2} &= \frac{2}{1+\alpha}\frac{u(t+h_1)}{h_1^2} = \frac{2u(t+h_1)}{h_1^2(\alpha+1)} \\ l\frac{u(t-h_0)}{h_0^2} &= \frac{2\alpha}{1+\alpha}\frac{u(t-h_0)}{\alpha^2 h_1^2} = \frac{2u(t-h_0)}{h_1^2(\alpha+1)\alpha} \end{aligned} \tag{3.27}$$

□

Korollar 3.1.5 *Die Aussagen von Satz 3.1.4 und Satz 3.1.3 gelten auch für den Fall einer Sprung-Unstetigkeit von u an der Stelle $t = t_i$.*

Beweis offensichtlich □

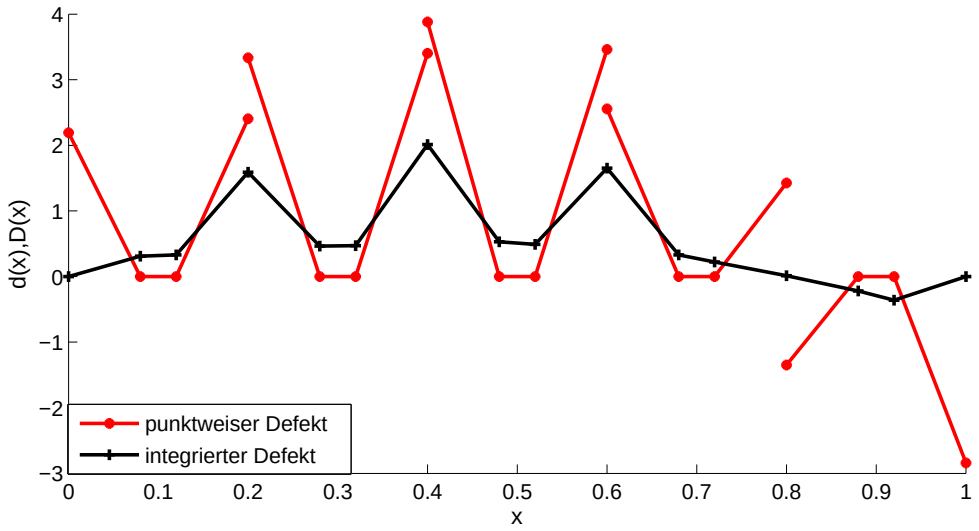

Abbildung 3.1: zeigt den punktweise ausgewerteten Defekt d der Lösung p eines Kollokationsverfahrens sowie die integrierte Version ($\mathcal{I}d$).

Man sieht in Abbildung 3.1 die Unstetigkeit des punktweisen Defekts d an den Gitterpunkten. In der integrierten Version sind diese nicht mehr vorhanden, der Operator $\mathcal{I}$ glättet die Unstetigkeiten im punktweisen Defekt d. Das ist eine Eigenschaft, die alle Integraloperatoren der Form (3.16) mit stetiger Kernfunktion K aufweisen.

3.2 Differenzenschema bei linearer Differentialgleichung

Ziel ist es nun, ein Gleichungssystem für einen Fehlerschätzer E zu finden. Dazu betrachtet man zunächst eine Differenzengleichung für den Fehler e. Mithilfe des im vorigen Abschnitt definierten Integraloperators $\mathcal{I}$ wird die zweite Ableitung des Fehlers in dessen zweiten Differenzenquotienten umgewandelt. Man beachte, dass die Sprungunstetigkeiten der Funktion p'' an den Stellen $x = x_k = x_{k0} = x_{(k-1)n}$, $k = 1 \dots K-1$ aufgrund von Korollar 3.1.5 kein Problem darstellen. Aus der so gefundenen Gleichung für den Fehler konstruiert man anschließend eine Gleichung für den Fehlerschätzer. Bei der Analyse der Abweichung des auf diese Art erklärten Fehlerschätzers ergibt sich auch für diese eine Differenzengleichung. Mithilfe eines Stabilitätsarguments und einer geeigneten Abschätzung der dabei auftretenden Störfunktion S ergibt sich daraus die asymptotische Korrektheit dieses Fehlerschätzers.

Anm.: Die Gleichung für den Fehlerschätzer E resultiert – nach Anwendung des Operators $\mathcal{I}$ – im Wesentlichen aus der Diskretisierung der ersten Ableitung mittels zentraler Differenzenquotienten. Da beim Kollokationsverfahren „2-dimensionale“ Gitter zugrunde liegen (im Sinn einer Abbildung $(k, i) \mapsto x_{ki}$), beim Differenzenverfahren jedoch eindimensionale Gitter vorliegen, werden die Punkte $x_{ki}, k = 0 \dots K-1,\ i = 0 \dots n$ im Folgenden mit den Punkten $t_j, j = 0 \dots K(n-1)+1$ bezeichnet. Bei natürlicher Identifizierung der Punkte ergibt sich folgender Zusammenhang zwischen den Indizes (k, i) und t:

$$\iota := \begin{cases} M \to [0, K(n-1)+1] \\ (k,i) \mapsto j = k(n-1)+i \end{cases}, \tag{3.28}$$

wobei $M = [0, K-1] \times [0, n-1] \cup (K-1, n)$ gilt. Durch diese Wahl des Definitionsbereichs erhält man einen bijektiven Zusammenhang zwischen den Indizes. Im Folgenden werden wir für die Punkte x_{ki} ausschließlich die Notation $t_{\iota(k,i)}$ verwenden.

3.2.1 Dirichlet-Randbedingungen

Zuerst betrachten wir nochmals (1.3) und bemerken, dass für die exakte Lösung u

$$u''(x) = g(x) - q_1(x) - q_2(x) \tag{3.29}$$

gilt. Wendet man den Operator $\mathcal{I}$ auf diese Gleichung an, so erhält man den folgenden Zusammenhang für die exakte Lösung u:

$$\delta^2_{h_0,h_1}(t, u) = (\mathcal{I}u'')(x) = (\mathcal{I}(g - q_1 - q_2))(x) = (\mathcal{I}f(., u, u'))(x) \tag{3.30}$$

Lemma 3.2.1 *Sei p die durch Kollokation gewonnene Approximation der exakten Lösung u von*

$$\begin{aligned} y''(x) - f\left(x, y(x), y'(x)\right) &= 0 \quad \textit{für alle } x \in [a,b] \\ R_1 \begin{pmatrix} y(a) \\ y(b) \end{pmatrix} + R_2 \begin{pmatrix} y'(a) \\ y'(b) \end{pmatrix} &= \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} \end{aligned} \tag{3.31}$$

Sei außerdem die Funktion f_{hom} definiert als

$$f_{hom}(x, y(x), y'(x)) := f(x, y(x), y'(x)) + g(x) \tag{3.32}$$

Der Fehler $e = p - u$ erfüllt dann das diskrete Gleichungssystem

$$\begin{aligned} \delta^2_{h_{i-1},h_i}(t_i, e) &= (\mathcal{I} f_{hom}(., e, e'))(t_i) + (\mathcal{I} d)(t_i) \\ e(a) &= 0 \\ e(b) &= 0 \end{aligned} \tag{3.33}$$

Beweis Der Beweis ergibt sich im Wesentlichen durch geschicktes Umformen: Betrachtet man den zweiten Differenzenquotienten von e, so gilt für diesen

$$\begin{aligned} \delta^2_{h_{i-1},h_i}(t_i, e) &= \delta^2_{h_{i-1},h_i}(t_i, p) - \delta^2_{h_{i-1},h_i}(t_i, u) \\ &= (\mathcal{I} p'')(t_i) - (\mathcal{I} u'')(t_i) \end{aligned}$$

Verwendet man nun (3.30), so erhält man

$$\begin{aligned} \delta^2_{h_{i-1},h_i}(t_i, e) &= (\mathcal{I} p'')(t_i) - (\mathcal{I} f(., u, u'))\,(t_i) \\ &= (\mathcal{I} p'')(t_i) - (\mathcal{I} f(., p, p'))(t_i) + (\mathcal{I} f(., p, p'))(t_i) - (\mathcal{I} f(., u, u'))(t_i) \end{aligned}$$

Berücksichtigt man die Linearität des Integrals ($(\mathcal{I}u + v)(t) = (\mathcal{I}u)(t) + (\mathcal{I}v)(t)$), so erhält man:

$$\delta^2_{h_{i-1},h_i}(t_i, e) = (\mathcal{I} \underbrace{(p'' - f(., p, p'))}_{=d})(t_i) + (\mathcal{I}(f(., p, p') - f(., u, u')))(t_i) \tag{3.34}$$

Da f linear in den letzten beiden Argumenten ist und $p(t) - u(t) = e(t)$, sowie $p'(t) - u'(t) = e'(t)$ gilt, kann man $(\mathcal{I}(f(., p, p') - f(., u, u')))(t_i)$ weiter vereinfachen. Beim Bilden der f-Differenzen fallen die konstanten Terme $g(x)$ weg, und es ergibt sich mit (1.2) und (2.3) folgender Zusammenhang:

$$\delta^2_{h_{i-1},h_i}(t_i, e) + (\mathcal{I}(q_1 e' + q_2 e))(t_i) = (\mathcal{I} d)(t_i) \tag{3.35}$$

Aus (3.32) erhält man $f_{hom}(x, y(x), y'(x)) = q_1(x)y(x) + q_2 y'(x)$, und damit letztendlich

$$\delta^2_{h_{i-1},h_i}(t_i, e) - (\mathcal{I} f_{hom}(., e, e'))(t_i) = (\mathcal{I} d)(t_i) \tag{3.36}$$

Zieht man in erster Linie nur Randbedingungen in Betracht, die ausschließlich die Funktion selbst und nicht ihre Ableitung betreffen, ergibt sich $e(a) = e(b) = 0$: Die Randbedingungen werden in diesem Fall durch ein lineares 2×2 Gleichungssystem mit 2 Unbekannten beschrieben. Dieses lässt sich unter entsprechender Forderung an die beiden Randbedingungen (linear unabhängige Gleichungen) eindeutig lösen und die Lösung beschreibt die Randwerte der Lösungsfunktion u exakt. Dementsprechend gehören in diesem Fall zur Gleichung (3.33) homogene Randbedingungen $e(a) = e(b) = 0$.

□

Aus der Darstellung (3.33) konstruiert man nun einen Fehlerschätzer, indem man $(\mathcal{I}f_{hom}(., e, e'))(t_i)$ durch eine geeignete, diskretisierte Variante $\tilde{F}(., e, e')(t_i)$ ersetzt. Hier bietet sich zum Beispiel eine finite Differenzendiskretisierung an, man setzt also $\tilde{F}(., e, e')(t_i) = f_{hom}(., e(t_i), \delta_{h_{i-1},h_i}(t_i, e))$. Die Inhomogenität $(\mathcal{I}d)(t_i)$ lässt sich numerisch nicht exakt berechnen, dementsprechend werden für den Fehlerschätzer die durch Quadratur ermittelten Werte $Q(\mathcal{I}d)(t_i) := D(t_i)$ verwendet. Mit diesen Überlegungen erhält man folgendes Gleichungssystem um den Fehlerschätzer E zu beschreiben:

$$\begin{aligned} \delta^2_{h_{i-1},h_i}(t_i, E) &= f_{hom}(t_i, E(t_i), \delta^1_{h_{i-1},h_i}(t_i, E)) + D(t_i) \\ E(a) &= 0 \\ E(b) &= 0 \end{aligned} \tag{3.37}$$

wobei $\delta^1_{h_{i-1},h_i}(t, E)$ für den zentralen ersten Differenzenquotienten von E an der Stelle t steht. Mit der Lösung dieses linearen Gleichungssystems erhält man also eine Näherung $E \approx e$ für den exakten Fehler. Es wird sich herausstellen, dass diese Näherung für $e = O(h^n)$ mit Ordnung $n + 1$ gegen e konvergiert, also um eine h-Potenz besser ist als die des Basisverfahrens. Der Beweis dieser Aussage lässt sich in 2 Schritte aufteilen: Zuerst bestimmt man eine Gleichung für den zweiten Differenzenquotienten der Abweichung $\epsilon = E - e$. Aus dieser Differenzengleichung erhält man mittels Stabilitätsargument sodann die gewünschte Ordnung. Wir halten also fest:

Lemma 3.2.2 *Sei e der Fehler, sowie E der aus (3.37) ermittelte Schätzer. Die Abweichung $\epsilon = E - e$ bzw. ihr zweiter Differenzenquotient erfüllen die Differenzengleichung*

$$\begin{aligned} (\mathcal{I}\epsilon'')(t_i) = f_{hom}\left(t_i, \epsilon_i, \frac{\epsilon_{i+1} - \epsilon_{i-1}}{h_i(\alpha + 1)}\right) + f_{hom}\left(t_i, e_i, \frac{e_{i+1} - e_{i-1}}{h_i(\alpha + 1)}\right) - \\ (\mathcal{I}f_{hom}(., e, e'))(t_i) + e_Q(t_i), \end{aligned} \tag{3.38}$$

wobei $e_Q(t_i)$ den Quadraturfehler $D(t_i) - (\mathcal{I}d)(t_i)$ bezeichnet.

Beweis Um Lemma 3.2.2 zu beweisen, helfen geschickte Umformungen weiter:

$$\begin{aligned} (\mathcal{I}\epsilon'')(t_i) &= (\mathcal{I}E'')(t_i) - (\mathcal{I}e'')(t_i) \\ &= f_{hom}\left(t_i, E_i, \frac{E_{i+1} - E_{i-1}}{h_i(\alpha + 1)}\right) + D_i - (\mathcal{I}f_{hom}(., e, e'))(t_i) - (\mathcal{I}d)(t_i) \end{aligned} \tag{3.39}$$

Durch Hinzufügen und Abziehen von $f_{hom}(t_i, e_i, \frac{e_{i+1}-e_{i-1}}{h_i(\alpha+1)})$ ergibt sich:

$$\begin{aligned}(\mathcal{I}\epsilon'')(t_i) = f_{hom}\left(t_i, E_i, \frac{E_{i+1}-E_{i-1}}{h_i(\alpha+1)}\right) - f_{hom}\left(t_i, e_i, \frac{e_{i+1}-e_{i-1}}{h_i(\alpha+1)}\right) + \\ f_{hom}\left(t_i, e_i, \frac{e_{i+1}-e_{i-1}}{h_i(\alpha+1)}\right) - (\mathcal{I}f_{hom}(., e, e'))(t_i) + e_Q(t_i)\end{aligned} \tag{3.40}$$

was man weiters zu

$$\begin{aligned}(\mathcal{I}\epsilon'')(t_i) = f_{hom}\left(t_i, \epsilon_i, \frac{\epsilon_{i+1}-\epsilon_{i-1}}{h_i(\alpha+1)}\right) + f_{hom}\left(t_i, e_i, \frac{e_{i+1}-e_{i-1}}{h_i(\alpha+1)}\right) - \\ (\mathcal{I}f_{hom}(., e, e'))(t_i) + e_Q(t_i)\end{aligned} \tag{3.41}$$

umformt. □

Der Fehlerschätzer und seine Abweichung erfüllen also nach (3.37) und (3.2.2) dieselbe Differenzengleichung mit unterschiedlichen Inhomogenitätsfunktionen. Mithilfe des folgenden Lemmas wird es möglich sein, diesen Zusammenhang auszunützen, indem man das Abschätzen der Abweichung auf ein Abschätzen der auftretenden Inhomogenität reduziert.

Lemma 3.2.3 *Unter den Voraussetzungen, dass $\|q_1(t)\, h\|$ hinreichend klein ist sowie $q_2(t) \geq 0$ gilt, liefert die finite Differenzendiskretisierung eine reguläre Matrix $M_E = M_E(h)$, welche eine Inverse M_E^{-1} besitzt, deren Norm $\|M_E^{-1}\|$ gleichmäßig beschränkt für $h \to 0$ ist. Es existiert also eine von $h > 0$ unabhängige Konstante C_M, sodass $\|M_E^{-1}\| < C_M$ ist. Weiters gilt dann für die Lösung der Differenzengleichungen $M_E x = S$, dass $x = M_E^{-1} S$ und damit die folgende Normabschätzung für x:*

$$\|x\| = \|M_E^{-1} S\| \leq \|M_E^{-1}\|\, \|S\| \tag{3.42}$$

Beweis Standardresultat aus der Numerik von Differentialgleichungen, der Beweis ist in [6] nachzulesen.

Nun ist schließlich alles vorbereitet um die Abweichung $\|\epsilon\|$ abzuschätzen, die Aussage wird als Satz formuliert:

Satz 3.2.4 *Seien e, e' sowie e'' konvergent mit Ordnung n, dann gilt: Die Lösung e des Differenzengleichungssystems (3.36) wird durch die Lösung E von (3.37) mit Ordnung $n+1$ approximiert, es gilt also $\|E - e\| = O(h^{n+1})$.*

Beweis Die finite Differenzendiskretisierung für den Fehlerschätzer E ist nach Lemma 3.2.3 bei hinreichender Glattheit der Funktion q_1 sowie Positivität der Funktion q_2 stabil. In Lemma 3.2.2 hat sich herausgestellt, dass die Abweichung ϵ dieselbe

Differenzengleichung erfüllt, also auch eine – gleichmäßig für $h \to 0$ – beschränkte Inverse besitzt. Man kann den Konvergenzbeweis demnach auf ein Abschätzen der Inhomogenität S zurückführen:

$$S = f_{hom}\left(t_i, e_i, \frac{e_{i+1} - e_{i-1}}{h_i(\alpha+1)}\right) - (\mathcal{I} f_{hom}(., e, e'))(t_i) + e_Q(t_i) \tag{3.43}$$

Aufgrund der Linearität von f_{hom} kann man diese umschreiben in die 3 Terme S_1, S_2 und S_3. Man erhält:

$$\begin{aligned} S &= f_{hom}\left(t_i, e_i, \frac{e_{i+1} - e_{i-1}}{h_i(\alpha+1)}\right) - (\mathcal{I} f_{hom}(., e, e'))(t_i) + e_Q(t_i) \\ &= \underbrace{\left(-q_1(t_i)e_i + (\mathcal{I} q_1 e)(t_i)\right)}_{:=S_1} + \underbrace{\left(-q_2(t_i)\frac{e_{i+1} + e_{i-1}}{h_i(\alpha+1)} + (\mathcal{I} q_2 e')(t_i)\right)}_{:=S_2} + \underbrace{e_Q(t_i)}_{:=S_3} \end{aligned} \tag{3.44}$$

Zuerst wird der Term S_1 behandelt. Um diesen abzuschätzen wird der Integrand zuerst mittels Taylorformel dargestellt:

$$\begin{aligned} (q_1 e)(t_i + h_{i-1}\zeta) &= (q_1 e)(t_i) + h_{i-1}\zeta (q_1 e)'(\xi^-) \quad \xi^- \in [t_{i-1}, t_i] \\ (q_1 e)(t_i + h_i \zeta) &= (q_1 e)(t_i) + h_i \zeta (q_1 e)'(\xi^+) \quad \xi^+ \in [t_i, t_{i+1}] \end{aligned} \tag{3.45}$$

Mithilfe dieser Darstellung ergibt sich bei der Anwendung von $\mathcal{I}$:

$$\begin{aligned} (\mathcal{I}(q_1 e))(t_i) &= \int_{-1}^{0} (1+\zeta)\big((q_1 e)(t_i) + h_{i-1}\zeta (q_1 e)'(\xi^-)\big) d\zeta + \\ &\quad \int_0^1 (1-\zeta)\big((q_1 e)(t_i) + h_i \zeta (q_1 e)'(\xi^+)\big) d\zeta \\ &= (q_1 e)(t_i) \underbrace{\int_{-1}^{1} (1 - |\zeta|) d\zeta}_{=1} + h_{i-1}(q_1 e)'(\xi^-) \underbrace{\int_{-1}^{0} \zeta(1+\zeta) d\zeta}_{=-\frac{1}{6}} + \\ &\quad h_i (q_1 e)'(\xi^+) \underbrace{\int_0^1 \zeta(1-\zeta) d\zeta}_{=-\frac{1}{6}} \\ &= (q_1 e)(t_i) + \frac{-h_{i-1}(q_1 e')(\xi^-)}{6} + \frac{h_i (q_1 e')(\xi^+)}{6} \end{aligned} \tag{3.46}$$

Betrachtet man nun die Differenz $|(\mathcal{I}(q_1 e))(t_i) - (q_1 e)(t_i)|$, so erhält man:

$$|(\mathcal{I}(q_1 e))(t_i) - (q_1 e)(t_i)| \leq \frac{1}{6}\left|h_{i-1}(q_1 e)'(\xi^-)\right| + \frac{1}{6}\left|h_i (q_1 e)'(\xi^+)\right| \tag{3.47}$$

Um die Notation zu vereinfachen sei nun $\tilde{h} = \tilde{h}_i := \frac{h_i+h_{i-1}}{2}$. Unmittelbar erkennt man, dass $\max\{h_{i-1}, h_i\} \leq 2\tilde{h}$ gilt. Für $|(\mathcal{I}(q_1 e))(t_i) - (q_1 e)(t_i)|$ ergibt sich somit:

$$|(\mathcal{I}(q_1 e))(t_i) - (q_1 e)(t_i)| \leq \frac{2\tilde{h}}{6} \left|(q_1 e)'(\xi^-)\right| + \frac{2\tilde{h}}{6} \left|(q_1 e)'(\xi^+)\right| \tag{3.48}$$

Beim Ableiten der Produkte $q_1 e$ erhält man nach der Produktregel $q_1 e' + q_1' e$. Aufgrund der Voraussetzung $q_1 \in C^1[a,b]$, gibt es demnach C_1 mit $|q_1(t)| \leq C_1,\ t \in [a,b]$ sowie C_2 mit $|q_1'(t)| \leq C_2,\ t \in [a,b]$. Verwendet man diese Tatsache um $|(q_1 e)'(\xi^-)|$ nach oben abzuschätzen erhält man $|(q_1 e)'(\xi^\pm)| \leq C_2|e(\xi^\pm)| + C_1|e'(\xi^\pm)| \leq (C_1 + C_2))(|e(\xi^\pm)| + |e'(\xi^\pm)|)$. Man kann die Abschätzungen aus (3.48) nun weiterführen:

$$\begin{aligned} |(\mathcal{I}(q_1 e))(t_i) - (q_1 e)(t_i)| &\leq \frac{\tilde{h}}{3}(C_1 + C_2) \underbrace{\left(|e(\xi^-)| + |e'(\xi^-)| + |e(\xi^+)| + |e'(\xi^+)|\right)}_{\leq K_1\tilde{h}^n + K_2\tilde{h}^n + K_3\tilde{h}^n + K_4\tilde{h}^n} \\ &\leq \frac{4}{3}\tilde{h}^{n+1}(C_1 + C_2)K \end{aligned} \tag{3.49}$$

wobei wir $K = \max_{i=1}^4(K_i)$ gesetzt haben. Damit erhalten wir für S_1 die gewünschte Konvergenz gegen 0 mit Ordnung $n+1$.

Um den Ausdruck $S_2 = -q_2(t_i)\frac{e_{i+1}+e_{i-1}}{h_i(\alpha+1)} + (\mathcal{I}q_2 e')(t_i)$ abschätzen zu können ist es sinnvoll den Differenzenquotienten $\frac{e_{i+1}-e_{i-1}}{h_i(\alpha+1)}$ zunächst als Integral darzustellen:

$$\frac{e_{i+1} - e_{i-1}}{h_i(\alpha+1)} = \frac{1}{h_i(\alpha+1)} \int_{t_{i-1}}^{t_{i+1}} e'(t)\, dt \tag{3.50}$$

Mittels der linearen Substitution $t = \frac{t_{i+1}+t_{i-1}}{2} + \frac{h_{i-1}+h_i}{2}\zeta$ kann man die Integration auf das Standardintervall $[-1,1]$ übertragen. Für $-q_2(t_i)\frac{e_{i+1}-e_{i-1}}{h_i(\alpha+1)}$ erhält man:

$$-q_2(t_i)\frac{e_{i+1} - e_{i-1}}{h_i(\alpha+1)} = \frac{-q_2(t_i)}{2} \int_{-1}^{1} e'\left(\frac{t_{i+1}+t_{i-1}}{2} + \frac{h_{i-1}+h_i}{2}\zeta\right) d\zeta \tag{3.51}$$

Nun wird der Integrand $e'\left(\frac{t_{i+1}+t_{i-1}}{2} + \frac{h_{i-1}+h_i}{2}\zeta\right)$ mittels Taylorformel um die Stelle $x_0 = t_i$ entwickelt. Aus $x - x_0 = \frac{t_{i+1}+t_{i-1}}{2} + \frac{h_{i-1}+h_i}{2}\zeta - t_i = \frac{h_i - h_{i-1}}{2} + \frac{h_{i-1}+h_i}{2}\zeta$ erhält (3.51) die Darstellung

$$\begin{aligned} -q_2(t_i)\frac{e_{i+1} - e_{i-1}}{h_i(\alpha+1)} &= \frac{-q_2(t_i)}{2} \int_{-1}^{1} e'(t_i) + \frac{h_i - h_{i-1}}{2} e''(\xi) + \frac{h_i + h_{i-1}}{2}\zeta e''(\xi)\, d\zeta \\ &= \frac{-q_2(t_i)}{2}\Big(2e'(t_i) + (h_i - h_{i-1})e''(\xi)\Big) \end{aligned} \tag{3.52}$$

Für die beiden in $(\mathcal{I}q_2 e')(t_i)$ auftretenden Integrale $\int_{-1}^{0}(1+\zeta)(q_2 e')(t_i + h_{i-1}\zeta)d\zeta$ sowie $\int_0^1(1-\zeta)(q_2 e')(t_i + h_i\zeta)d\zeta$ wird $(q_2 e')(t_i + h_i\zeta)$ bzw. $(q_2 e')(t_i + h_{i-1}\zeta)$ ebenfalls um die

Stelle t_i nach Taylor entwickelt, was zu

$$\begin{aligned}(q_2e')(t_i+h_{i-1}\zeta) &= (q_2e')(t_i)+h_{i-1}\zeta(q_2e')'(\xi^-) \quad \xi^-\in[t_{i-1},t_i]\\ (q_2e')(t_i+h_i\zeta) &= (q_2e')(t_i)+h_i\zeta(q_2e')'(\xi^+) \quad \xi^+\in[t_i,t_{i+1}]\end{aligned} \tag{3.53}$$

führt. Die Integrale erhalten damit man bei der Anwendung von $\mathcal{I}$ die Form:

$$\begin{aligned}(\mathcal{I}(q_2e'))(t_i) &= \int_{-1}^{0}(1+\zeta)\big((q_2e')(t_i)+h_{i-1}\zeta(q_2e')'(\xi^-)\big)d\zeta+\\ &\quad \int_0^1(1-\zeta)\big((q_2e')(t_i)+h_i\zeta(q_2e')'(\xi^+)\big)d\zeta\\ &= (q_2e')(t_i)\underbrace{\int_{-1}^{1}(1-|\zeta|)d\zeta}_{=1}+h_{i-1}(q_2e')'(\xi^-)\underbrace{\int_{-1}^{0}\zeta(1+\zeta)d\zeta}_{=-\frac{1}{6}}+\\ &\quad h_i(q_2e')'(\xi^+)\underbrace{\int_0^1\zeta(1-\zeta)d\zeta}_{=\frac{1}{6}}\end{aligned} \tag{3.54}$$

Berechnet man nun – wie bei der Abschätzung von S_1 – die Differenz $S_2=-q_2(t_i)\frac{e_{i+1}-e_{i-1}}{h_i(\alpha+1)}+(\mathcal{I}(q_2e'))(t_i)$, so erhält man für diese:

$$\begin{aligned}S_2 = &\frac{-q_2(t_i)}{2}\Big(2e'(t_i)+(h_i-h_{i-1})e''(\xi)\Big)+(q_2e')(t_i)-\\ &\frac{1}{6}\,h_{i-1}\,(q_1e')'(\xi^-)+\frac{1}{6}\,h_i\,(q_2e')'(\xi^+)\end{aligned} \tag{3.55}$$

Um auch hier die gewünschte Konvergenzordnung zu sehen, bemerken wir zuerst, dass die beiden Terme $\pm(q_2e)(t_i)$ einander kürzen. Somit erhält man für S_2 die Abschätzung

$$|S_2|\le|q_2e''(t_i)|\,\frac{|h_i-h_{i-1}|}{2}+\frac{1}{6}h_{i-1}\left|(q_1e')'(\xi^-)\right|+\frac{1}{6}h_i\left|(q_2e')'(\xi^+)\right| \tag{3.56}$$

Mit der Festsetzung $\tilde h=\tilde h_i:=\frac{h_i+h_{i-1}}{2}$ erhält man aus (3.56):

$$|S_2|\le 2\tilde h\Big(|(q_2e'')(t_i)|+\tfrac{2}{6}\big(\,|(q_2e')'(\xi^-)|+(q_2e')'(\xi^-)|\,\big)\Big). \tag{3.57}$$

Laut Voraussetzung ist die Funktion $q_2\in C^1[a,b]$. Dementsprechend gibt es wieder $C_1,\ C_2$ mit $|q_2(t)|\le C_1,\ t\in[a,b]$ bzw. $|q_2'(t)|\le C_2,\ t\in[a,b]$. Verwendet man diese Tatsache um die Ableitungen $|(q_2e')'(\xi^\pm)|=|(q_2e'')(\xi^\pm)+(q_2'e')(\xi^\pm)|$ abzuschätzen, so ergibt sich $|(q_2e')'(\xi^\pm)|\le(C_1+C_2)(|e'(\xi^-)|+|e''(\xi^-)|)$. Das führt schließlich zur folgenden Abschätzung:

$$|S_2|\le 2\tilde h\Big(|(C_1e'')(t_i)|+\tfrac{2}{6}(C_1+C_2)\big(\,|e'(\xi^-)|+|e''(\xi^-)|+|e'(\xi^+)|+|e''(\xi^+)|\,\big)\Big) \tag{3.58}$$

Aufgrund der Konvergenz von e, e' und e'' gegen 0 mit Ordnung n kann man diese Terme jeweils durch $K_i\tilde{h}^n, i = 1 \dots 5$ abschätzen und man erhält mit $K := \max_{i=1}^{5}(K_i)$:

$$|S_2| \leq 5\tilde{h}(C_1 + C_2)K\tilde{h}^n = 5(C_1 + C_2)K\tilde{h}^{n+1}, \tag{3.59}$$

womit auch für S_2 Konvergenzordnung $n + 1$ gezeigt ist.

Um den Quadraturfehler hinreichend klein zu halten verwendet man in jedem Kollokationsintervall $[x_k, x_{k+1}]$ interpolatorische Quadratur mit Stützstellen $x_{ki}, i = 0 \dots n$. Dadurch erreicht man bei hinreichender Glattheit des punktweisen Defekts, dass auch der Quadraturfehler $e_Q = O(\tilde{h}^{n+1})$ erfüllt. Insgesamt erhält man damit $|S| = |S_1 + S_2 + S_3| \leq |S_1| + |S_2| + |S_3| = O(\tilde{h}^{n+1})$, und für $\epsilon = E - e$:

$$\|\epsilon\| \leq \|M_E^{-1}\| \, \|S\| \leq CK\tilde{h}^{n+1}. \tag{3.60}$$

Der Fehler des Schätzers besitzt also zumindest Konvergenzordnung $n + 1$, falls der Fehler selbst, sowie seine erste und zweite Ableitung Konvergenzordnung n aufweisen.

□

Anm.: Die Glattheitsforderung an den Defekt ist in der Regel kaum zu realisieren, für eine Beweisführung mit schwächeren Glattheitsforderung kann man z.B. versuchen den Quadraturfehler auf die exakte Lösung zu einzuschränken.

3.2.2 Einseitige Neumannrandbedingung

Betrachtet man Gleichungen mit $R_2 \neq 0$ (siehe 1.3), so muss das Verhalten des Fehlers am Rand genauer untersucht werden. Für eine erste Herleitung eines Differenzenschemas in diesem Fall nehmen wir zuerst vereinfachte Randbedingungen der Form $y(a) = \alpha$ und $y'(b) = \beta$ an. In den folgenden Rechenschritten seien – zwecks übersichtlicherer Notation – $x_{K-1} + h_{K-1}\zeta_{n-1} = t_{N-1}$, $b = t_N$ und $h = b - t_{N-1}$. Um nun das Verhalten des Fehlers am Randpunkt b studieren zu können ist folgender Ansatz hilfreich:

$$\frac{y(t_N) - y(t_{N-1})}{h} = \underbrace{\frac{y(t_N) - y(t_{N-1})}{h} - y'(t_N)}_{\text{als Integral darstellbar?}} + \beta \tag{3.61}$$

Detaillierter sucht man erneut eine Kernfunktion $\tilde{K}(\zeta)$, integrierbar auf $[0, 1]$, sodass

$$\int_0^1 \tilde{K}(\zeta)y''(t_{N-1} + h\zeta)d\zeta = \frac{y(t_N) - y(t_{N-1})}{h} - y'(b) \tag{3.62}$$

gilt, d.h. der erste linksseitige Differenzenquotient kommt als Integral über die zweite Ableitung der Funktion zustande. Taylorentwicklung bis zur Ordnung 1 mit Integraldarstellung des Restgliedes liefert das passende Ergebnis:

$$y(t_{N-1} + h) = y(t_{N-1}) + hy'(t_{N-1}) + \int_{t_{N-1}}^{t_{N-1}+h} (t_{N-1} + h - \nu)y''(\nu)d\nu \tag{3.63}$$

Mittels der linearen Substitution $\nu = t_{N-1} + h\zeta$ ergibt sich $d\nu = h\ d\zeta$. Die Integralgrenzen gehen in 0 bzw. 1 über und man erhält:

$$\begin{aligned} y(t_{N-1}+h) &= y(t_{N-1}) + hy'(t_{N-1}) + \int_0^1 (h-h\zeta)y''(t_{N-1}+h\zeta)hd\zeta \\ &= y(t_{N-1}) + hy'(t_{N-1}) + h^2\int_0^1 (1-\zeta)y''(t_{N-1}+h\zeta)d\zeta \\ &= y(t_{N-1}) + hy'(t_{N-1}) + h^2\ \frac{y'(t_{N-1}+h\zeta)}{h}\Big|_{\zeta=0}^1 - h^2\int_0^1 \zeta y''(t_{N-1}+h\zeta)d\zeta \end{aligned} \tag{3.64}$$

Beim Auswerten an den Grenzen $\zeta = 1$ bzw. $\zeta = 0$ erhält man $hy'(t_{N-1}+h) - hy'(t_{N-1})$ - die beiden Terme $hy'(t_{N-1})$ kürzen einander also. Bringt man $y(t_{N-1})$ und $hy'(t_{N-1}+h)$ auf die andere Seite und dividiert durch h, so erhält man:

$$\frac{y(t_N)-y(t_{N-1})}{h} - y'(t_N) = -h\int_0^1 \zeta y''(t_{N-1}+h\zeta)d\zeta \tag{3.65}$$

Dieses Ergebnis liefert auch schon den gesuchten Operator $\tilde{\mathcal{I}}$: Setzt man

$$(\tilde{\mathcal{I}}y)(t) = -h\int_0^1 \zeta y(t+h\zeta)d\zeta, \tag{3.66}$$

so gilt für 2-mal stetig differenzierbares y:

$$\delta_{h_l}(t,y) = (\tilde{\mathcal{I}}y'')(t) + y'(t) = -h\int_0^1 \zeta y''(t+h\zeta)d\zeta + y'(t), \tag{3.67}$$

wobei $\delta_{h_l}(t,y)$ den linksseitigen Differenzenquotienten von y an der Stelle t zur Intervalllänge h bezeichnet. Das Integral misst also genau den Fehler, der bei Ersetzen von $y'(t)$ durch $\delta_{h_l}(t,y)$ gemacht wird. Dass für den linksseitigen Differenzenquotienten nur Funktionswerte links von t von Interesse sind, spiegelt sich auch hier wider, die Funktion (bzw. genauer: ihre zweite Ableitung) wird schließlich auch nur links von t integriert. Um dieses Ergebnis mit dem Ausgangsproblem in Verbindung zu bringen, führt ähnliches Vorgehen wie im Fall von Dirichletrandbedingungen zum Ziel.

Wie zuvor wird ein exaktes Differenzenschema für den Fehler $e(t) = p(t) - u(t)$ hergeleitet, wobei $p(t)$ die durch das Kollokationsverfahren erhaltene Lösung und $u(t)$ die exakte Lösung von (1.3) bezeichnen. An allen inneren Punkten ergibt sich analog zum Fall von Dirichletrandbedingungen Ergebnis (3.33):

$$\delta^2_{h_{i-1},h_i}(t_i,e) = (\mathcal{I}f_{hom}(.,e,e'))(t_i) + (\mathcal{I}d)(t_i) \tag{3.68}$$

Am rechten Rand b betrachtet man nun den ersten (linksseitigen) Differenzenquotienten von e:

$$\begin{aligned} \delta_{h_l}(b,e) &= \delta_{h_l}(b,p) - \delta_{h_l}(b,u) \\ &= (\tilde{\mathcal{I}}p'')(b) + p'(b) - (\tilde{\mathcal{I}}u'')(b) - u'(b) \end{aligned} \tag{3.69}$$

Aufgrund der speziellen Form der Randbedingungen gilt $p'(b) = u'(b)$, die beiden Terme kürzen einander. Addiert und subtrahiert man $(\tilde{\mathcal{I}}f(.,p,p'))(b)$, so kann man den Differenzenquotienten wieder mittels Defekt darstellen:

$$\begin{aligned}\delta_{h_l}(b,e) &= (\tilde{\mathcal{I}}p'')(b) + (\tilde{\mathcal{I}}f(.,p,p'))(b) - (\tilde{\mathcal{I}}f(.,p,p'))(b) - (\tilde{\mathcal{I}}f(.,u,u'))(b) \\ &= (\tilde{\mathcal{I}}\underbrace{p'' - f(.,p,p')}_{=d})(b) + (\tilde{\mathcal{I}}f_{hom}(.,e,e'))(b)\end{aligned} \tag{3.70}$$

Hier wurde noch berücksichtigt, dass f linear in den beiden letzten Argumenten ist. Um daraus nun eine passende Differenzengleichung für $E \approx e$ zu finden, setzt man:

$$\delta_{h_l}(b,E) = D(b), \tag{3.71}$$

wobei $D(b)$ wiederum die numerisch ausgewertete Variante von $(\tilde{\mathcal{I}}d)(b)$ bezeichnet. Lösen des Systems

$$\begin{aligned}\delta^2_{h_{i-1},h_i}(t_i,E) &= f_{hom}(t_i,E(t_i),\delta^1_{h_{i-1},h_i}(t_i,E)) + D(t_i) \\ \delta_{h_l}(b,E) &= D(b) \\ E(a) &= 0,\end{aligned} \tag{3.72}$$

führt – wie im Fall von Dirichletrandbedingungen – zum Fehlerschätzer E.

Satz 3.2.5 *Die durch (3.72) gegebene Approximation E an den exakten Fehler e liefert einen Schätzer der Ordnung $n+1$, d.h.:*

$$\|\epsilon\| = \|E - e\| = O(h^{n+1}) \tag{3.73}$$

Beweis Der Fehler $\epsilon = E - e$ erfüllt an der Stelle b:

$$\begin{aligned}\delta_{h_l}(b,\epsilon) &= \delta_{h_l}(b,E) - \delta_{h_l}(b,e) \\ &= D(b) - (\tilde{\mathcal{I}}f_{hom}(.,e,e'))(b) - (\tilde{\mathcal{I}}d)(b) \\ &= -(\tilde{\mathcal{I}}f_{hom}(.,e,e'))(b) + e_Q(b),\end{aligned} \tag{3.74}$$

wobei hier wieder – wie im Fall von Dirichletrandbedingungen $e_Q(b)$ den Quadraturfehler bezeichnet.
An allen inneren Punkten erfüllt ϵ (3.41), an der Stelle a gilt wegen $u(a) = p(a) = \alpha$, dass $e(a) = 0$ ist, also $\epsilon(a) = 0$. Verbindet man diese Gleichungen nun mit (3.74), so erhält man ein Gleichungssystem $\tilde{M}_E\epsilon = \tilde{S}$ mit Inhomogenität $\tilde{S}$, wobei $\tilde{S}_0 = 0$, $\tilde{S}_{K(n+1)} = (\tilde{\mathcal{I}}d)(b)$ und $\tilde{S}_j = S_j$ $j = 1 \dots K(n+1)-1$. Da die Inhomogenität $\tilde{S} = S$ an den inneren Punkten in Abschnitt 3.2.1 ausführlich behandelt wurde bleibt nur noch das Verhalten von $\tilde{S}$ an der Stelle b zu studieren. Betrachtet man die letzte Komponente von $\tilde{S}$, erhält man:

$$\begin{aligned}\tilde{S}_{K(n+1)} &= -(\tilde{\mathcal{I}}f_{hom}(.,e,e'))(b) + e_Q(b) \\ &= h\int_0^1 \zeta\big((q_1e)(t+h\zeta) + (q_2e')(t+h\zeta)\big)d\zeta + e_Q(b)\end{aligned} \tag{3.75}$$

Analog zum Fall von Dirichletrandbedingungen schätzen wir q_1 und q_2 mit C_1 bzw. C_2 ab. Außerdem wählt man die Quadratur so, dass für den Quadraturfehler $e_Q(b) = O(h^{n+1})$ gilt. Damit ergibt sich:

$$|\tilde{S}_{K(n+1)}| = h(C_1 + C_2)(|e(\xi^-)| + |e'(\xi^+)|) + O(h^{n+1}) \leq K(C_1 + C_2)h^{n+1}, \tag{3.76}$$

womit auch für den Randpunkt b Konvergenzordnung $n + 1$ gezeigt ist.

□

3.2.3 Allgemeine lineare Randbedingungen

Um das Differenzenschema auf allgemeine lineare Randbedingungen wie in (1.3) zu verallgemeinern, sind nicht mehr viele Schritte nötig. Zuerst wird man allerdings nicht umhinkommen, eine Kernfunktion $\hat{K}$ zu suchen, welche analoge Eigenschaften zur Kernfunktion $\tilde{K}$ in Bezug auf den rechtsseitigen Differenzenquotienten hat. Dazu ist es wiederum am einfachsten, die Funktion $y(t)$ nach Taylor zu entwickeln und die Integralrestgliedformel zu verwenden. Um eine übersichtliche Notation zu gewährleisten, sei dazu $t_0 = a$, $t_1 = \tau_0(\zeta_1)$ und dementsprechend $h = t_1 - t_0$.

$$y(t_0 + h) = y(t_0) + hy'(t_0) + \int_{t_0}^{t_0+h} (t_0 + h - \nu)y''(\nu)d\nu \tag{3.77}$$

Mittels der linearen Substitution $\nu = t_1 - h\zeta$ ergibt sich $d\nu = -h\ d\zeta$. Die Integralgrenzen gehen in 1 bzw. 0 über und man erhält:

$$\begin{aligned} y(t_0 + h) &= y(t_0) + hy'(t_0) + \int_1^0 h\zeta y''(t_1 - h\zeta)(-h)d\zeta \\ &= y(t_0) + hy'(t_0) + h^2 \int_0^1 \zeta y''(t_1 - h\zeta)d\zeta \end{aligned} \tag{3.78}$$

Bringt man $y(t_0)$ und $hy'(t_0 + h)$ auf die andere Seite und dividiert durch h, so erhält man:

$$\frac{y(t_1) - y(t_0)}{h} - y'(t_0) = h \int_0^1 \zeta y''(t_1 - h\zeta)d\zeta \tag{3.79}$$

Dieses Ergebnis liefert – so wie für den linksseitigen Differenzenquotienten – auch schon den gesuchten Operator $\hat{\mathcal{I}}$:

$$(\hat{\mathcal{I}}y)(t) := h \int_0^1 \zeta y(t - h\zeta)d\zeta, \tag{3.80}$$

Mit dieser Definition gilt für 2-mal stetig differenzierbares y:

$$\delta_{h_r}(t, y) = (\hat{\mathcal{I}}y'')(t) + y'(t) = h \int_0^1 \zeta y''(t - h\zeta)d\zeta + y'(t), \tag{3.81}$$

$\delta_{h_r}(t,y)$ bezeichnet hierbei analog zu $\delta_{h_l}(t,y)$ den rechtsseitigen Differenzenquotienten von y an der Stelle t zur Intervalllänge h. Auch hier kann das Integral als der Fehler gedeutet werden, der bei Ersetzen von $y'(t)$ durch $\delta_{h_r}(t,y)$ gemacht wird. Betrachtet man analog zu oben $\delta_{h_r}(a,e)$ speziell im Fall von Randbedingungen der Form $y'(a)=\alpha$ und $y(b)=\beta$, so ergibt sich wegen $p'(a)=u'(a)=\alpha$:

$$\begin{aligned}\delta_{h_r}(a,e) &= \delta_{h_r}(a,p)-\delta_{h_r}(a,u)\\ &= (\hat{\mathcal{I}}p'')(a)+p'(a)-(\hat{\mathcal{I}}u'')(a)-u'(a)\\ &= (\hat{\mathcal{I}}p'')(a)+(\hat{\mathcal{I}}f(.,p,p'))(a)-(\hat{\mathcal{I}}f(.,p,p'))(a)-(\hat{\mathcal{I}}f(.,u,u'))(a)\\ &= (\hat{\mathcal{I}}\underbrace{p''-f(.,p,p')}_{=\hat{d}})(a)+(\hat{\mathcal{I}}f_{hom}(.,e,e'))(a)\end{aligned} \tag{3.82}$$

Der daraus konstruierte Fehlerschätzer E sieht im Prinzip genauso aus wie im Fall der linksseitigen Randableitung, der Nachweis der Konvergenz funktioniert ebenfalls genauso. Aus diesem Grund wird die Herleitung an dieser Stelle nicht explizit durchgeführt.

Betrachtet man nun Randbedingungen der Form

$$R_1\begin{pmatrix}y(a)\\y(b)\end{pmatrix}+R_2\begin{pmatrix}y'(a)\\y'(b)\end{pmatrix}=\begin{pmatrix}s_1\\s_2\end{pmatrix}, \tag{3.83}$$

so erhält man, unter der Annahme, dass sowohl die Kollokationslösung p als auch die exakte Lösung u die Randbedingungen erfüllen:

$$R_1\begin{pmatrix}e(a)\\e(b)\end{pmatrix}+R_2\begin{pmatrix}e'(a)\\e'(b)\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix}, \tag{3.84}$$

wobei hier wieder $e=p-u$ ist. Man kann in diesem Fall (3.70) bzw. (3.82) NICHT direkt übernehmen. Bei der Herleitung dieser beiden Gleichungen wurde nämlich die spezielle Form der Randbedingungen berücksichtigt, da $p'(b)=u'(b)=\beta$ bzw. $p'(a)=u'(a)=\alpha$ verwendet wurde. Im Fall allgemeiner linearer Randbedingungen ist dies aber nicht sichergestellt, anstatt (3.70) und (3.82) verwendet man die folgende, verallgemeinerte Version dieser beiden Gleichungen:

$$\begin{aligned}\delta_{h_r}(a,e) &= (\hat{\mathcal{I}}d)(a)+(\hat{\mathcal{I}}f_{hom}(.,e,e'))(a)+e'(a)\\ \delta_{h_l}(b,e) &= (\tilde{\mathcal{I}}d)(b)+(\tilde{\mathcal{I}}f_{hom}(.,e,e'))(b)+e'(b)\end{aligned} \tag{3.85}$$

Die Herleitung funktioniert genauso wie in (3.70) bzw. (3.82) gezeigt wird, man hat nur $p'(b)-u'(b)$ und $p'(a)-u'(a)$ nicht 0, sondern $e'(b)$ bzw. $e'(a)$ zu setzen. Dementsprechend erhält man $e'(a)=\delta_{h_r}(a,e)-(\hat{\mathcal{I}}d)(a)-(\hat{\mathcal{I}}f_{hom}(.,e,e'))(a)$ sowie $e'(b)=\delta_{h_l}(b,e)-(\tilde{\mathcal{I}}d)(b)-(\tilde{\mathcal{I}}f_{hom}(.,e,e'))(b)$. Setzt man diese Ausdrücke für $e'(a)$ bzw. $e'(b)$ in (3.84) ein, so erhält man:

$$R_1\begin{pmatrix}e(a)\\e(b)\end{pmatrix}+R_2\begin{pmatrix}\delta_{h_r}(a,e)-(\hat{\mathcal{I}}d)(a)-(\hat{\mathcal{I}}f_{hom}(.,e,e'))(a)\\ \delta_{h_l}(b,e)-(\tilde{\mathcal{I}}d)(b)-(\tilde{\mathcal{I}}f_{hom}(.,e,e'))(b)\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix} \tag{3.86}$$

Bringt man nun noch den konstanten Term auf die rechte Seite, so erhält man als exaktes Gleichungssystem des Fehlers am Rand:

$$R_1 \begin{pmatrix} e(a) \\ e(b) \end{pmatrix} + R_2 \begin{pmatrix} \delta_{h_r}(a,e) - (\hat{\mathcal{I}} f_{hom}(.,e,e'))(a) \\ \delta_{h_l}(b,e) - (\tilde{\mathcal{I}} f_{hom}(.,e,e'))(b) \end{pmatrix} = R_2 \begin{pmatrix} (\hat{\mathcal{I}} d)(a) \\ (\tilde{\mathcal{I}} d)(b) \end{pmatrix} \tag{3.87}$$

Das Gleichungssystem für den Fehlerschätzer lautet in diesem Fall:

$$R_1 \begin{pmatrix} E(a) \\ E(b) \end{pmatrix} + R_2 \begin{pmatrix} \delta_{h_r}(a,E) \\ \delta_{h_l}(b,E) \end{pmatrix} = R_2 \begin{pmatrix} (\hat{\mathcal{I}} d)(a) \\ (\tilde{\mathcal{I}} d)(b) \end{pmatrix} \tag{3.88}$$

Durch Bilden der Differenzen von (3.88) und (3.87) erhält man folgendes Gleichungssystem für den Fehler ϵ des Schätzers:

$$R_1 \begin{pmatrix} \epsilon(a) \\ \epsilon(b) \end{pmatrix} + R_2 \begin{pmatrix} \delta_{h_r}(a,\epsilon) + (\hat{\mathcal{I}} f_{hom}(.,e,e'))(a) \\ \delta_{h_l}(b,\epsilon) + (\tilde{\mathcal{I}} f_{hom}(.,e,e'))(b) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{3.89}$$

Bringt man den konstanten Term auf die rechte Seite, so ergibt sich:

$$R_1 \begin{pmatrix} \epsilon(a) \\ \epsilon(b) \end{pmatrix} + R_2 \begin{pmatrix} \delta_{h_r}(a,\epsilon) \\ \delta_{h_l}(b,\epsilon) \end{pmatrix} = \begin{pmatrix} -(\hat{\mathcal{I}} f_{hom}(.,e,e'))(a) \\ -(\tilde{\mathcal{I}} f_{hom}(.,e,e'))(b) \end{pmatrix} \tag{3.90}$$

Mit einem Stabilitätsargument wie in den vorangegangenen Abschnitten reicht es auch hier aus, die Norm des Vektors $\begin{pmatrix} -(\hat{\mathcal{I}} f_{hom}(.,e,e'))(a) \\ -(\tilde{\mathcal{I}} f_{hom}(.,e,e'))(b) \end{pmatrix}$ abzuschätzen, um die Konvergenz von ϵ an den Stellen a und b zu zeigen. Diese hat sich jedoch schon als $O(h^{n+1})$ herausgestellt, damit liegt auch hier Konvergenz der Ordnung $n+1$ vor. Für die Stabilität des Differenzenverfahrens bei gemischten Randbedingungen siehe [6].

3.3 Differenzenschema bei nichtlinearer Differentialgleichung

3.3.1 Dirichlet-Randbedingungen

Im nichtlinearen Fall muss die Herleitung des Fehlerschätzers ein wenig abgeändert werden. Alle Schritte funktionieren im Prinzip genauso, man kann nur die Linearität von f nicht mehr ausnützen. Dementsprechend lassen sich die Differenzengleichungen nicht immer in der Form $\delta^2 e = f_{hom}(e, \delta e)$ schreiben. Für den Fehler $e = p - u$ gilt stattdessen:

$$\begin{aligned} \delta^2_{h_{i-1},h_i}(t_i,e) &= (\mathcal{I}p'')(t_i) - (\mathcal{I}u'')(t_i) \\ &= (\mathcal{I}p'')(t_i) - (\mathcal{I}f(.,u,u'))(t_i) \\ &= (\mathcal{I}p'')(t_i) - (\mathcal{I}f(.,p,p'))(t_i) + (\mathcal{I}f(.,p,p'))(t_i) - (\mathcal{I}f(.,u,u'))(t_i), \end{aligned} \tag{3.91}$$

wobei für die erste Gleichheit verwendet wurde, dass $u(t)$ die exakte Lösung von (1.2) ist, die dritte Zeile resultiert aus hinzufügen und abziehen des Terms $(\mathcal{I}f_p)(t_i)$. Berücksichtigt man die Linearität des Integrals ($(\mathcal{I}u + v)(t) = (\mathcal{I}u)(t) + (\mathcal{I}v)(t)$), so erhält man

$$\delta^2_{h_{i-1},h_i}(t_i, e) = (\mathcal{I}d)(t_i) + (\mathcal{I}(f(., p, p') - f(., u, u')))(t_i) \tag{3.92}$$

In diesem Fall darf also $f(., p, p') - f(., u, u')$ nicht zu $f(., e, e')$ vereinfacht werden, man löst in diesem Fall ($\delta E_j := \delta_{h_{i-1},h_i}(., E_j)$ $j = 1, 2$ – um die Übersicht zu wahren):

$$\begin{aligned} \delta^2_{h_{i-1},h_i}(t_i, E_1) &= (\mathcal{I}f(., E_1, \delta E_1))(t_i) + (\mathcal{I}d)(t_i) \\ \delta^2_{h_{i-1},h_i}(t_i, E_2) &= (\mathcal{I}f(., E_2, \delta E_2))(t_i) \end{aligned} \tag{3.93}$$

Die Differenz $E := E_1 - E_2$ erfüllt offensichtlich

$$\delta^2_{h_{i-1},h_i}(t_i, E) = (\mathcal{I}d)(t_i) + (\mathcal{I}(f(., E_1, \delta E_1') - f(., E_2, \delta E_2)))(t_i) \tag{3.94}$$

und durch E ist im nichtlinearen Fall ein Fehlerschätzer gegeben. Der Beweis für die asymptotische Korrektheit wird im nichtlinearen Fall nicht geführt und kann in [7] nachgelesen werden. Die Idee beruht im Wesentlichen auf der Linearisierung von f um die exakte Lösung von (1.2). Die dabei auftretenden Fehler lassen sich insgesamt durch h^{n+1} abschätzen, womit man ein Ergebnis wie in Satz 3.2.4 erhält.
Die Konvergenzordnung des so erklärten Fehlerschätzers ergibt sich in numerischen Beispiele ebenso wie im linearen Fall zu $n + 2$, wobei mit n die Ordnung des Basisverfahrens bezeichnet sei.

3.4 Anmerkungen

3.4.1 Fehlerschätzer bei anderen Basisverfahren

Die Herleitung des exakten Differenzenschemas, sowie Abschätzungen und Konvergenzaussagen berücksichtigt nur an zwei Stellen, dass es sich bei p um ein Kollokationspolynom handelt. In den Abschätzungen für die Konvergenzordnung benötigt man die Konvergenz des Fehlers e sowie seiner ersten und zweiten Ableitung mit Ordnung n, das wird durch Verwendung einer Kollokationslösung sichergestellt. Bei der numerischen Auswertung der Integrale über den Defekt verwendet man außerdem die Tatsache, dass der punktweise Defekt d an inneren Kollokationsknoten gleich 0 ist.

Sei nun $u = \{u_0, u_1 \dots u_N\}$ Lösung eines Differenzenverfahrens auf dem Gitter $a = t_0 < t_1 < \dots < t_N = b$. Hermite Interpolation nach der Vorschrift

$$p_0(t_0) = u_0, \qquad p_0(t_1) = u_1, \qquad p_0(t_2) = u_2 \tag{3.95}$$

liefert ein Polynom p_0 im Intervall $[t_0, t_1]$. Mit den weiteren Forderungen

$$p_i(t_i) = u_i, \qquad p_i(t_{i+1}) = u_{i+1}, \qquad p_i'(t_i) = p_{i-1}'(t_i) \tag{3.96}$$

erhält man nun Teilpolynome p_i, jeweils mit Träger in $[t_i, t_{i+1}]$. Das zusammengesetzte Polynom $p = \sum_{i=0}^{N-1} p_i$ weist die Eigenschaften (u sei die exakte Lösung der zugrunde liegenden Differentialgleichung) $p - u = O(h^2)$, $p' - u' = O(h^2)$ sowie $p'' - u'' = O(h)$ besitzt. Zu diesem stückweisen Polynom lässt sich analog zum Kollokationsfall ein Fehlerschätzer E konstruieren. Das Problem bei der Konvergenzuntersuchung wird jedoch die reduzierte Konvergenzordnung der zweiten Ableitung darstellen, das Produkt he'' kann dementsprechend nur mit h^2 und nicht mit h^3 abgeschätzt werden, was zu Ordnung 2 des Fehlers ϵ des Schätzers führen würde. Praktische Versuche ergaben dennoch Ordnung 4. Die Behandlung von Neumannrandbedingungen verkompliziert sich hier, da man beim Lösen der Randgleichungen einen Diskretisierungsfehler macht. In diesem Fall müsste man noch den Defekt in den Randbedingungen mitberücksichtigen um auch für Neumannrandbedingungen einen Fehlerschätzer konstruieren zu können. Eine weitere Variante Neumannrandbedingungen exakt zu erfüllen, wäre durch eine Abwandlung der Interpolationsvorschrift gegeben: Durch Hermiteinterpolation im letzten Teilintervall lässt sich erreichen, dass das zusammengesetzte Polynom p die Randbedingungen ebenfalls exakt erfüllt.

3.4.2 Empirische Konvergenzordnungen

Im Fall von Dirichletrandbedingungen ergaben numerische Versuche (Abschnitt 4) für den Fehlerschätzer immer die Konvergenzordnung $n+2$, wobei n die Ordnung des Verfahrens bezeichnet. Die in dieser Arbeit vorgestellten Ergebnisse ergaben jedoch nur Konvergenzordnung $n + 1$. Die Idee hinter diesen Beweisen war die auftretende Inhomogenität $\|S\|$ (siehe Lemma 3.2.3) mit h^{n+1} abschätzen zu können, daraus ergab sich die gewünschte Ordnung. Um auf Ordnung $n + 2$ schließen zu können muss das Produkt $\|M_E^{-1}S\|$ näher betrachtet werden (siehe [7]).

Im Fall von gemischten Randbedingungen ergab sich mit der hier vorgestellten Variante nicht Konvergenzordnung $n + 2$. Die einseitigen Differenzenquotienten konvergieren in den Randpunkten ja nicht mit Ordnung 2, sondern nur mit Ordnung 1 (Dies liegt nicht an der Tatsache, dass es sich um Randpunkte handelt, als vielmehr daran, dass die einseitigen Differenzenquotienten eben nur mit Ordnung 1 konvergieren). Um dieses Problem zu umgehen, besteht die Möglichkeit die einseitigen Differenzenquotienten geeignet zu ersetzen. Anstatt δ_{h_l} und δ_{h_l} betrachtet man eine Linearkombination aus 3 Punkten:

$$f'(x) \approx a_1 f(x_0) + a_2 f(x_1) + a_3 f(x), \tag{3.97}$$

wobei $x_0 = x_1 - h_0$ und $x_1 = x - h_1$ gilt. Die Koeffizienten ergeben sich wie bei der Herleitung der zentralen Differenzenquotienten durch Taylorentwicklung. Im Gegensatz zu den einseitigen Differenzenquotienten konvergiert diese Diskretisierung (3-Punktregel) mit Ordnung 2. Setzt man nun in obiger Gleichung $f = e$, so erhält man (mit Hilfe einer neuen Kernfunktion) eine Differenzengleichung für den exakten Fehler am Rand. Der daraus errechnete Fehlerschätzer ergab in praktischen Anwendungen analog zum

Dirichletfall die Konvergenzordnung $n + 2$.

3.4.3 Verbesserung der Lösung

Mithilfe des erhaltenen Fehlerschätzers bietet sich die Möglichkeit, die erhaltene Approximationslösung p zu verbessern. Mithilfe von $e = p - u$ und $\epsilon = E - e$ erhält man:

$$e = p - u \Leftrightarrow E - \epsilon = p - u \Leftrightarrow \epsilon = E - p + u \tag{3.98}$$

Demnach stellt $p - E$ wegen $\|u - (P - E)\| = \|\epsilon\| = O(h^{n+1})$ eine verbesserte Approximation an die exakt Lösung u dar. Hierfür liegt jedoch kein Fehlerschätzer mehr vor, in der Praxis stellt sich demnach in natürlicher Weise die Frage nach der Wahl eines guten Fehlerschätzers oder einer verbesserten Konvergenz der Lösung.

Kapitel 4

Numerische Ergebnisse

Dieses Kapitel soll der Präsentation numerischer Ergebnisse dienen. Das erste Beispiel wird ein wenig ausführlicher behandelt, um das Verhalten des Fehlerschätzers bei unterschiedlichen Kollokationsgraden, Randbedingungen, etc. zu untersuchen. Ebenso wollen wir an diesem Beispiel auch den Einsatz des Fehlerschätzers bei anderen Basisverfahren aufzeigen. In den nachfolgenden Tabellen bezeichnet die Spalte n die Anzahl an Kollokationsintervallen. Spalten mit der Bezeichnung O_i beinhalten die beobachtete Konvergenzordnung. ϵ_i bzw. F_i bezeichnen letztendlich die ermittelten Fehlerwerte. Die Kollokationsgitterpunkte für die i-te Zeile der Tabellen wurden durch Halbierung gewonnen: Bei gegebenen Gitter mit $s+1$ Teilintervallen $[x_i, x_{i+1}]$, $i = 0 \dots s-1$ wurden die neuen Gitterpunkte durch Halbierung der Teilintervalle gewonnen: $\tilde{x}_i = \frac{x_i + x_{i+1}}{2}, i = 0 \dots s-1$. Im nächsten Schritt liegen also doppelt soviele Intervalle vor (siehe Abb. (4.1)). Das erste, für diesen Prozess verwendete Gitter werden wir im Folgenden als Grundgitter bezeichnen. Der Lösungsbereich $[a, b]$ ist – wenn nicht explizit darauf hingewiesen wird – immer mit dem Intervall $[0, 1]$ zu identifizieren, die Probleme werden also auf einem Standardintervall gelöst.
Mittels der in den Abschnitten 3.2.1 und 3.2.2 gefundenen Abschätzungen erwartet man für den Fehlerschätzer E Konvergenz der Ordnung $n+1$, wobei n die Konvergenzordnung des Verfahrens bezeichnet (genauer: die Konvergenzordnung der Funktion, der ersten sowie zweiten Ableitung). Die präsentierten Tabellen zeigen, dass die Konvergenz des Schätzers sogar noch um eine Ordnung besser ist, in nahezu allen Fällen erhält man Konvergenz der Ordnung $n+2$. Fehlertabellen und Grafiken wurden mit den im Anhang gelisteten **Matlab**funktionen erstellt.

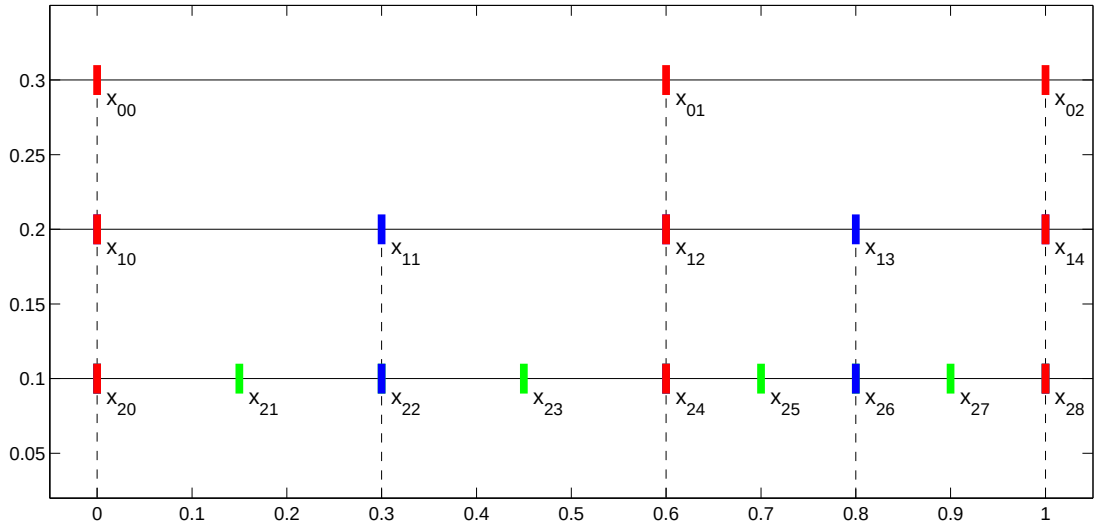

Abbildung 4.1: zeigt die sukzessive Halbierung der Intervalle $[x_i, x_{i+1}]$ zum Grundgitter $\{0, \frac{6}{10}, 1\}$.

4.1 Beispiel 1

Als erstes betrachten wir ein lineares, reguläres Problem mit glatter Lösung und glatten Koeffizientenfunktionen.

$$y''(x) = \frac{1}{2}y'(x) + \frac{1}{2}y(x) - \frac{1}{2}(1+6x)e^x \tag{4.1}$$

4.1.1 Dirichlet-Randbedingungen

Mit den Randbedingungen $y(0) = 0$ und $y(1) = 0$ erhält man als exakte Lösung von (4.1) die Funktion $y(x) = x(1-x)e^x$:

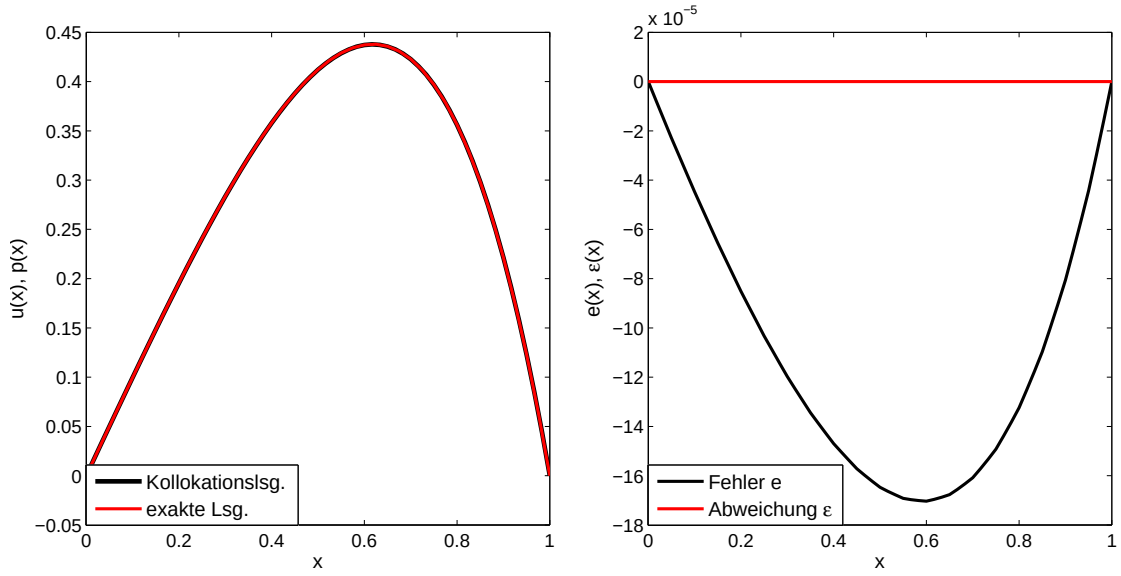

Abbildung 4.2: zeigt (links) die exakte Lösung von (4.1), sowie eine Kollokationslösung. Rechts sind der exakte Fehler sowie die Abweichung des Schätzers dargestellt.

Für die Erstellung von Abbildung 4.2 wurde $\zeta = \{0, \frac{1}{2}, 1\}$ gesetzt und äquidistante Schrittweite $h = \frac{1}{20}$ gewählt. Die rechte Abbildung zeigt vor allem die Qualität des Fehlerschätzers: Betrachtet man dessen Abweichung im Vergleich mit den exakten Fehlerwerten, so erkennt man optisch keinen Unterschied zur 0-Linie, die Werte sind also erheblich kleiner. Für die erste Fehlertabelle betrachten wir 3 verschiedene Folgen ζ_i:

- **Kollokationsgrad 2:** $\zeta_1 = \{0, \frac{1}{2}, 1\}$ aufgrund der Symmetrie der Folge bzgl. $\frac{1}{2}$ und der ungeraden Anzahl innerer Knotenpunkte erwartet man als Ordnung des Verfahrens in diesem Fall 2 statt 1.
- **Kollokationsgrad 3:** $\zeta_2 = \{0, \frac{1}{3}, \frac{2}{3}, 1\}$ Obwohl es sich um eine bzgl. $\frac{1}{2}$ symmetrische Folge handelt ist keine „Superkonvergenz“ zu erwarten, da eine gerade Anzahl innerer Kollokationsknoten vorliegt. Die Ordnung des Verfahrens ist sollte somit gleich 2 sein.
- **Kollokationsgrad 4:** $\zeta_3 = \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$ Eine ebenfalls symmetrische Folge ζ_i wird Konvergenzordnung 4 liefern.

n	F_1	O_1	F_2	O_2	F_3	O_3
2	3.12e-2		1.69e-2		3.77e-4	
4	7.73e-3	2.01	4.14e-3	2.03	2.35e-5	4.00
8	1.93e-3	2.00	1.03e-3	2.01	1.47e-6	4.00
16	4.82e-4	2.00	2.57e-4	2.00	9.18e-8	4.00
32	1.20e-4	2.00	6.43e-5	2.00	5.74e-9	4.00
64	3.01e-5	2.00	1.61e-5	2.00	3.56e-10	4.01
128	7.53e-6	2.00	4.02e-6	2.00	1.79e-11	4.32
256	1.88e-6	2.00	1.01e-6	2.00	1.77e-11	0.02
512	4.70e-7	2.00	2.51e-7	2.00	1.41e-10	2.99
1024	1.18e-7	2.00	6.27e-8	2.00	1.97e-10	0.49

Tabelle 4.1: zeigt Fehlerwerte der Kollokation sowie die daraus resultierenden Ordnungen. Die Spalten mit Bezeichnung $F_i, i = 1 \dots 3$ beschreiben die Fehlerwerte an der Stelle $x_0 = \frac{1}{2}$. Die mit O_1, O_2 sowie O_3 bezeichneten Spalten zeigen die betreffenden Ordnungen.

n	ϵ_1	O_1	ϵ_2	O_2	ϵ_3	O_3
2	1.32e-4		4.63e-5		1.44e-7	
4	8.15e-6	4.02	2.87e-6	4.01	2.21e-9	6.02
8	5.08e-7	4.00	1.79e-7	4.00	3.44e-11	6.01
16	3.17e-8	4.00	1.12e-8	4.00	4.99e-13	6.11
32	1.98e-9	4.00	7.00e-10	4.00	6.17e-14	3.02
64	1.24e-10	4.00	4.39e-11	3.99	6.19e-13	-3.33
128	7.67e-12	4.01	3.85e-12	3.51	1.83e-12	-1.57
256	1.18e-12	2.70	4.81e-14	6.32	9.54e-12	-2.38
512	3.45e-13	1.77	1.56e-11	-8.34	4.33e-11	-2.18
1024	3.58e-13	-0.05	1.57e-11	-0.01	1.37e-10	-1.66

Tabelle 4.2: zeigt Fehlerwerte des Schätzers sowie die daraus resultierenden Ordnungen desselbigen. Die Spalten mit Bezeichnung $\epsilon_i, i = 1 \dots 3$ beschreiben die Fehlerwerte an der Stelle $x_0 = \frac{1}{2}$. Die mit O_1, O_2 sowie O_3 bezeichneten Spalten zeigen die betreffenden Ordnungen.

In Tabelle 4.1 kann man die empirisch beobachteten Ordnungen der unterschiedlichen Varianten des Verfahrens ablesen. Die Spalte mit Bezeichnung F_1 bezeichnet die Fehlerwerte an der Stelle $\frac{1}{2}$. Durchgehend erkennt man, dass sich die Fehlerwerte bei verdoppelter Kollokationsknotenanzahl auf $\frac{1}{4}$ reduzieren. Dieses Verhalten kann man auch anhand der Spalte O_1 sehen: Die Werte bezeichnen die empirisch beobachtete Konvergenzordnung der Fehlerwerte aus der Spalte F_1. Ordnung 2 bedeutet – entsprechend der ersten Spalte – dass sich die Fehlerwerte bei Halbierung der Schrittweite um

$(\frac{1}{2})^2$ reduzieren würden. Ordnung 2 resultiert in diesem Fall aus der symmetrischen Anordnung der Werte von ζ.
Die beiden Spalten F_2 sowie O_2 repräsentieren entsprechende Daten zur Kollokationsfolge $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$. Beim Vergleich der Spalten F_1 und F_2 fällt auf, dass die Fehlerwerte in der Spalte F_2 nur ungefähr halb so groß sind. Wie die Werte aus Spalte F_1 reduzieren sich auch hier die Fehlerwerte bei Halbierung der Schrittweite um $\frac{1}{4}$. Die Konvergenzordnungen O_2 sind dementsprechend wieder 2. Für die letzten beiden Spalten wurde Kollokation der Ordnung 4 gewählt, was aufgrund der symmetrisch gewählten Kollokationsfolge ζ_3 zu Konvergenzordnung 4 führt (Spalte O_3). Die Fehlerwerte in der Spalte F_3 reduzieren sich in diesem Fall bei Halbierung der Schrittweite sogar um $\frac{1}{16}$. Im Gegensatz zu den ersten beiden Kollokationsfolgen lässt sich das Verhalten des Fehlers in diesem Fall nicht über die ganze Tabelle hin beobachten. Ab Zeile 8 bzw. 256 Kollokationsintervallen kommt es zur Stagnation der Fehlerwerte. Die Gitterpunkte in Verbindung mit den zusätzlichen inneren Knoten sind ab diesem Zeitpunkt zu klein um mit Lagrangepolynomen befriedigende Ergebnisse zu erzielen. Die Werte des Fehlers beginnen in den letzten beiden Zeilen sogar wieder zu wachsen. Der Grund hierfür sind Rundungseffekte, die durch andere Implementierung teilweise in den Griff zu bekommen sind.
In Tabelle 4.2 werden für dieselben 3 Kollokationsfolgen die Fehlerwerte des Fehlerschätzers dargestellt. In der Spalte mit Bezeichnung ϵ_1 sind die Fehlerwerte des Schätzers zur Kollokationsfolge ζ_1 aufgelistet, in der Spalte O_1 die entsprechenden Ordnungen. Diese betragen bis zur achten Zeile immer 4, ab diesem Punkt kommen allerdings – wie in den letzten beiden Spalten von Tabelle 4.1 – Rundungsfehler zum Tragen und bewirken eine reduzierte Konvergenzordnung. Im letzten Schritt nimmt der Fehler wieder zu.
In den Spalten ϵ_2 sowie O_2 sind die Fehler des Schätzers und die daraus resultierenden empirischen Ordnungen zur Kollokationsfolge ζ_2 aufgelistet. Analog zur Tabelle 4.1 erkennt man, zwar dieselbe Konvergenzordnung wie bei Verwendung von ζ_1, die Fehlerwerte sind jedoch prinzipiell um einen Faktor $\frac{1}{2}$ kleiner. Dies führt dazu, dass Rundungsfehler schon eine Zeile früher zu Tragen kommen (Zeile 7), sich dort allerdings noch nicht gravierend auswirken. Die Konvergenzordnung ist in diesem Schritt um $\frac{1}{2}$ reduziert. Für die dritte Kollokationsfolge ζ_3 betrachte man die Spalten ϵ_3 sowie O_3. Die in der letzten Spalte ersichtliche Konvergenz der Ordnung 6 lässt sich nur in den ersten 3 Schritten beobachten, die Ordnung reduziert sich sehr bald auf 3, ab dieser Zeile (32 Kollokationsintervalle) werden die Fehlerwerte wieder größer, im Fall von Kollokationsgrad 4 bricht der Fehlerschätzer also bei 16 Kollokationsintervallen im Standardintervall $[0, 1]$ zusammen. Die Fehlerwerte befinden sich hierbei allerdings schon im Bereich von 5e-13.

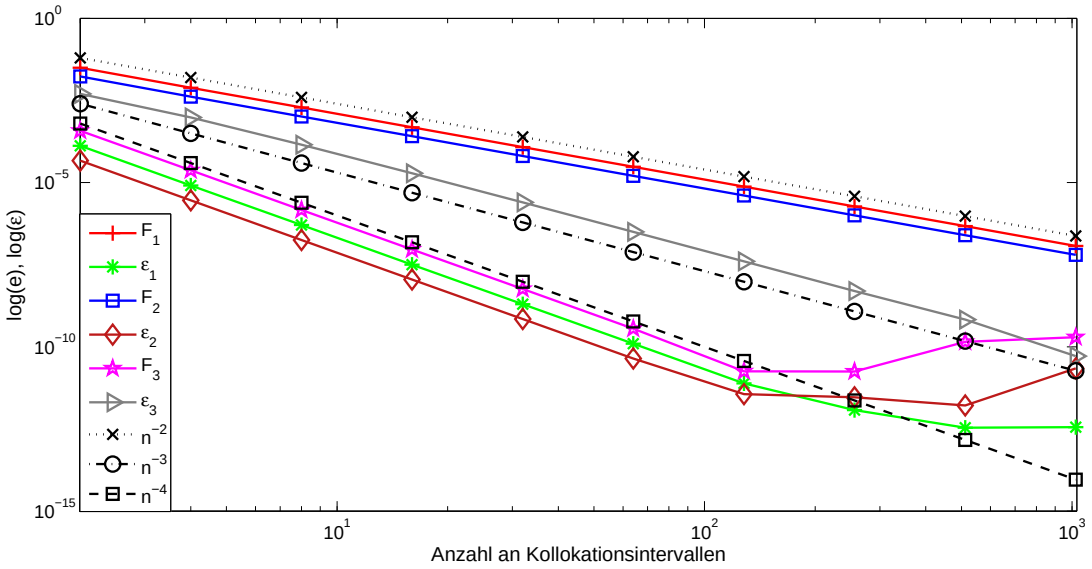

Abbildung 4.3: zeigt analog zu Tabelle 4.1 und Tabelle 4.2 die Fehler der einzelnen Varianten des Verfahrens, sowie die Fehler der daraus resultierenden Fehlerschätzer.

In Abbildung 4.3 sind nochmals die Fehlerwerte aus den ersten beiden Tabellen dargestellt. Aufgrund der logarithmischen Skalierung der Achsen ergibt sich für die Fehlerkurve die Form einer Geraden, wobei der Anstieg gleichzeitig die Ordnung bestimmt. Die rote Kurve entspricht den Fehlerwerten bei Kollokation zweiter Ordnung, die grüne dem entsprechenden Fehlerschätzer. Man erkennt mithilfe der n^{-2} bzw. der n^{-4}-Geraden, ebenso wie in den beiden Tabellen, dass der Fehler mit Ordnung 2 konvergiert, der Fehler des Schätzers mit Ordnung 3. Variante 2 liefert im Prinzip diesselben Ordnungen, die blaue Gerade, die den Fehler bei Variante 2 darstellt, ist ebenso parallel zur n^{-2}-Geraden. Auch die zugehörige Gerade für den Fehlerschätzer (rosa) ist parallel zur n^{-4}-Linie. Man erkennt allerdings, dass der Fehler nur bis zu einer bestimmten Intervallanzahl einem Geradenverlauf entspricht. Bei Kollokation vierter Ordnung treten auch schon beim Verfahren selbst signifkante Rundungsfehler auf, dementsprechend ist auch hier der Verlauf der türkisen Linie nur bis zu einer bestimmten Intervallanzahl mit einer Geraden zu identifizieren. Die Fehler des Fehlerschätzers lassen sich in diesem Fall sogar nur über weit weniger als 100 Kollokationsintervalle mit einer Geraden identifizieren.

Nachdem also bei äquidistanten Intervalllängen keine Probleme auftreten, soll an Beispiel 1 auch demonstriert werden, dass im Prinzip „beliebige“ Gitter vorgegeben werden können. Dafür seien die folgenden Kollokationsfolgen gepaart mit nicht äquidistanten Grundgittern gegeben:

- **Kollokationsgrad 3:** $\zeta_1 = \{0, \frac{1}{3}, \frac{4}{5}\, 1\}$ führt zu Konvergenz zweiter Ordnung, „Superkonvergenz“ im Sinn von regelmäßigen inneren Knoten tritt hier nicht

auf, da nicht einmal die Anzahl innerer Kollokationsknoten passend wäre. Für das sukzessive zu halbierende Grundgitter wählen wir die Folge $\{0, \frac{1}{3}, 1\}$

- **Kollokationsgrad 4:** $\zeta_2 = \{0, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, 1\}$ eine nicht-äquidistante aber dennoch symmetrische Folge wird wieder „Superkonvergenz“ mit sich bringen. Als zugehöriges Startgitter sei wieder $\{0, \frac{1}{3}, 1\}$ gegeben.
- **Kollokationsgrad 4:** $\zeta_3 = \{0, \frac{1}{4}, \frac{2}{3}, \frac{3}{4}, 1\}$ liefert die erwartete Ordnung 3 für die Kollokationslösung, da keine symmetrischen Knoten vorliegen. Wieder sei als Startgitter sei $\{0, \frac{1}{3}, 1\}$ festgesetzt.

In den folgenden Fehlertabellen ist der maximale Fehler über alle Gitterpunkte dargestellt, eine konkrete Auswertungsstelle x_0 liegt demnach nicht vor:

n	F_1	O_1	F_2	O_1	F_3	O_3
2	2.15e-2		2.67e-3		2.53e-3	
4	5.19e-3	2.05	1.69e-4	3.98	1.89e-4	3.74
8	1.15e-3	2.17	1.08e-5	3.97	1.71e-5	3.47
16	2.76e-4	2.06	6.76e-7	4.00	1.68e-6	3.34
32	6.70e-5	2.04	4.22e-8	4.00	1.81e-7	3.22
64	1.65e-5	2.02	2.63e-9	4.00	2.09e-8	3.12
128	4.10e-6	2.01	1.60e-10	4.03	2.49e-9	3.06
256	1.02e-6	2.01	1.99e-11	3.01	2.80e-10	3.16
512	2.55e-7	2.00	4.70e-11	1.24	5.93e-11	2.24
1024	6.36e-8	2.00	6.32e-10	3.75	3.12e-10	2.40

Tabelle 4.3: zeigt Ergebnisse zu Beispiel 1, diesmal mit nichtäquidistantem Kollokationsgitter. Die Spalten $F_i, i = 1 \dots 3$ entsprechen auch hier dem Verfahrensfehler, Spalten mit Bezeichnung $O_i, i = 1 \dots 3$ repräsentieren die entsprechenden Ordnungen.

Man erkennt anhand von Tabelle 4.3 sehr gut, dass die erwarteten Ordnungen auch bei nicht äquidistantem Gitter erreicht werden. Die erste Variante mit Kollokationsgrad 3 liefert demnach Konvergenz zweiter Ordnung, für die Varianten vierter Ordnung erhält man im symmetrischen Fall sogar Konvergenzordnung 4 (Spalten F_2 sowie O_2). ζ_3 führt – wie ζ_1 zu Konvergenz dritter Ordnung. Die Fehlerwerte brechen bei Werten im Bereich von 10e-10 zusammen, und kommt es zu einer Reduktion der Ordnung.

n	ϵ_1	O_1	ϵ_2	O_1	ϵ_3	O_3
2	3.20e-4		3.14e-5		2.62e-5	
4	1.44e-5	4.47	6.30e-7	5.64	5.85e-7	5.49
8	5.37e-7	4.75	1.11e-8	5.83	1.12e-8	5.70
16	2.12e-8	4.66	1.84e-10	5.91	2.59e-10	5.44
32	1.02e-9	4.38	2.97e-12	5.95	7.01e-12	5.21
64	5.52e-11	4.21	1.36e-12	1.12	1.06e-12	2.73
128	3.98e-12	3.79	4.17e-12	-1.61	4.09e-12	-1.95
256	4.96e-12	-0.32	3.38e-11	-3.02	1.72e-11	-2.07
512	1.55e-11	-1.65	1.33e-10	-1.98	4.86e-11	-1.50
1024	6.01e-11	-1.95	3.53e-10	-1.40	2.30e-10	-2.24

Tabelle 4.4: zeigt die Abweichung des Fehlerschätzers aus Beispiel 1. Analog bezeichnen wieder Spalten mit Bezeichnung $O_i, i = 1 \dots 3$ die empirische Ordnung, Spalten mit Bezeichnung $\epsilon_i, i = 1 \dots 3$ zeigen die Abweichung von E.

4.1.2 Neumann-Randbedingungen

Nachdem bei Gleichungen zweiter Ordnung die Randbedingungen auch Ableitungswerte beinhalten können, und das Differenzenschema für den Fehlerschätzer hierfür entsprechend modifiziert werden muss, wollen wir auch hier mittels numerischer Ergebnisse die Vorgehensweise rechtfertigen: Randbedingungen der Form $y(0) = 0$ und $y'(1) = -e$, ergeben dieselbe exakte Lösung wie bei obiger Aufgabenstellung. Für die nachfolgende Fehlertabelle betrachten wir einerseits den klassischen einseitigen Differenzenquotienten und andererseits die in (3.97) vorgestellte 3-Punktregel.

n	F	O_F	ϵ_1	O_{ϵ_1}	ϵ_2	O_{ϵ_2}
2	1.03e-1		1.65e-1		1.82	
4	2.47e-2	2.07	3.96e-2	2.06	6.86e-3	8.05
8	6.20e-3	1.99	7.99e-3	2.31	3.66e-4	4.23
16	1.55e-3	2.00	1.39e-3	2.52	2.24e-5	4.03
32	3.88e-4	2.00	2.15e-4	2.69	1.40e-6	4.00
64	9.71e-5	2.00	3.05e-5	2.82	8.79e-8	3.99
128	2.43e-5	2.00	4.09e-6	2.90	5.48e-9	4.00
256	6.07e-6	2.00	5.30e-7	2.95	2.14e-10	4.68
512	1.52e-6	2.00	6.78e-8	2.97	2.63e-10	-0.30
1024	3.91e-7	1.96	1.27e-8	2.42	4.09e-9	-3.96

Tabelle 4.5: zeigt die Unterschiede zwischen den beiden Varianten, Neumann Randbedingungen zu behandeln. F bezeichnet den Verfahrensfehler, O_F die Ordnung. ϵ_i und O_{ϵ_i} stehen für den Fehler des Schätzers bzw. dessen Konvergenzordnung.

In Tabelle 4.5 sind die Fehlerwerte des Kollokationsverfahrens, sowie die Fehler der beiden Fehlerschätzer abzulesen. Die Kollokationsfolge ζ ist in diesem Fall durch $\{0,\ \frac{1}{2},\ 1\}$ gegeben, die Auswertungsstelle x_0 durch $x_0 = \frac{1}{2}$. Diese äquidistante Zerlegung liefert wie im Fall von Dirichlet-Randbedingungen Konvergenzordnung 1, was in der Spalte mit Bezeichnung O_F zu sehen ist. Die Fehlerwerte des Verfahrens (F) reduzieren sich dementsprechend von Zeile zu Zeile um $\frac{1}{4}$. Da auch bei 1024 Kollokationsintervallen noch keine zu kleinen Fehlerwerte auftreten, lässt sich Ordnung 2 über die ganze Tabelle hin beobachten. Die mittleren beiden Spalten ϵ_1 und O_{ϵ_1} zeigen Fehlerwerte und die daraus resultierende Ordnung des Schätzers, der in Abschnitt 3.2.1 diskutiert wurde. Die Konvergenzordnung ist im Gegensatz zu Dirichletrandbedingungen nicht um 2 besser, als die des Verfahrens an sich, sondern beträgt in diesem Fall nur noch 3. (Spalte O_{ϵ_1}). Die Fehlerwerte (ϵ_1) reduzieren sich dementsprechend jeweils um einen Faktor $\frac{1}{8}$. Fehlerwerte für den Schätzer der in (3.97) vorgestellte wurde und eine verbesserte Variante Randbedingungen mit Ableitung zu behandeln darstellt finden sich in den letzten beiden Spalten. Hier kann man in der Spalte O_{ϵ_2} wieder Konvergenz der Ordnung $2 + 2 = 4$ beobachten, was sogar dazu führt, dass bei 256 Intervallen die Ordnung nicht mehr eingehalten wird, die Fehlerwerte sogar wieder größer werden.

4.1.3 Differenzenverfahren

Wir wollen nun ein letztes Mal auf (4.1) eingehen. Wie schon in Abschnitt 3.4 erwähnt lässt sich das Prinzip des Fehlerschätzers auf andere Basisverfahren übertragen, wir wollen das am Beispiel der finiten Differenzen demonstrieren:

n	F	O_F	ϵ	O_ϵ
2	2.42e-2		1.24e-2	
4	6.18e-3	1.97	6.52e-5	7.57
8	1.55e-3	1.99	4.32e-6	3.92
16	3.88e-4	2.00	2.76e-7	3.97
32	9.71e-5	2.00	1.74e-8	3.99
64	2.43e-5	2.00	1.09e-9	3.99
128	6.07e-6	2.00	6.86e-11	4.00
256	1.52e-6	2.00	4.24e-12	4.02
512	3.80e-7	2.00	4.06e-13	3.39
1024	9.49e-8	2.00	9.68e-14	2.07

Tabelle 4.6: zeigt Fehlerwerte des Differenzenverfahrens und des Fehlerschätzers, sowie die daraus resultierenden Ordnungen. F bezeichnet die Fehlerwerte an der Stelle $x_0 = \frac{1}{2}$, O_F die daraus resultierende Ordnung. Die anderen beiden Spalten sind genauso zu verstehen.

Auch in Tabelle 4.6 erkennt man, dass die Konvergenz des Fehlerschätzers um 2 Ordnungen besser ist als die des Basisverfahrens. Wie oben bricht die Konvergenzordnung des Schätzers aufgrund von Rundungsfehlern zusammen. Mit der Beweisvariante, die in dieser Arbeit vorgestellt wurde, lässt sich dieses Ergebnis allerdings auch nicht für Ordnung $n + 1 = 3$ herleiten, da die Konvergenzordnung der zweiten Ableitung schon um 1 reduziert ist und damit nur noch 1 beträgt.

Anm.: Dieses lineare Beispiel wurde in der Form Matrix · Vektor implementiert, die Konditionszahlen der zugrundeliegende Matrizen verhielten sich bei allen bisher gezeigten Varianten wie K^3, wobei K die Anzahl an Kollokationsintervallen beschreibt. Die Konditionszahl des Kollokationsverfahrens war prinzipiell immer größer, als die des Hilfsverfahrens, da bei letzterem wirklich eine Tridiagonalmatrix vorliegt, die Übergangsgleichungen bei der Kollokation außerdem mehr Aufwand darstellen. Die mittels Beispiel 1 produzierten Ergebnisse zeigen die schon zuvor angesprochene, um 2 höhere Konvergenzordnung des Fehlerschätzers. Um dieses Ergebnis beweisen zu können, sind allerdings viel feinere Betrachtungen des zugrundeliegenden diskreten Gleichungssystems nötig, welche in dieser Arbeit nicht durchgeführt wurden (siehe [7]).

4.2 Beispiel 2

Als zweites Beispiel soll eine nichtlineare Gleichung behandelt werden. Da allgemeine Randbedingungen für diesen Problemtyp nicht implementiert wurden, werden nur Tabellen zu unterschiedlichen Kollokationsknoten vorgestellt:

$$\begin{aligned} y''(x) &= 1 - (y')^2 \\ y(0) &= \tfrac{1}{2} \\ y(1) &= \tfrac{1}{4} \end{aligned} \tag{4.2}$$

Die exakte Lösung dieser Gleichung ist durch $y(x) = xe^x$ gegeben und sieht folgendermaßen aus:

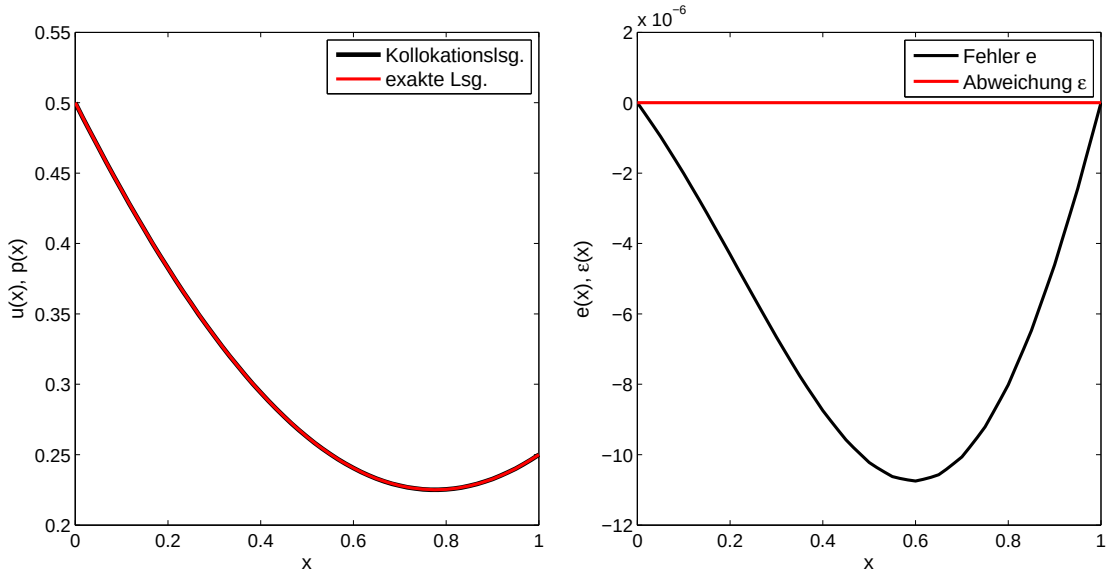

Abbildung 4.4: zeigt im linken Bild die zu (4.2) gehörige exakte, sowie Kollokationslösung. Rechts sind wieder der Fehler sowie die Abweichung ϵ zu erkennen.

Ähnlich wie beim ersten Beispiel wurden unterschiedliche Folgen ζ getestet, diesmal allerdings hauptsächlich asymmetrisch bezüglich $\frac{1}{2}$:

- **Kollokationsgrad 2:** $\zeta_1 = \{0, \frac{1}{3}, 1\}$ führt zu Konvergenz erster Ordnung, „Superkonvergenz“ im Sinn von regelmäßigen inneren Knoten tritt nicht auf. Die Konvergenz des Fehlerschätzers ist allerdings auch bei nichtäquidistanter Folge ζ ist auch hier wieder um 2 besser als die des Basisverfahrens.
- **Kollokationsgrad 3:** $\zeta_2 = \{0, \frac{1}{4}, \frac{3}{5}, 1\}$ Wie zu erwarten erhält man Konvergenzordnungen 2 bzw. 4.
- **Kollokationsgrad 4:** $\zeta_3 = \{0, \frac{1}{5}, \frac{1}{2}, \frac{3}{4}, 1\}$ liefert die erwartete Ordnung 3 für die Kollokationslösung, der Fehlerschätzer weist in diesem Fall sogar eine um 3 verbesserte Konvergenzordnung auf.

n	F_1	O_1	F_2	O_2	F_3	O_3
2	5.38e-3		8.13e-4		5.96e-5	
4	2.48e-3	1.12	2.38e-4	1.77	5.31e-6	3.49
8	1.17e-3	1.09	6.33e-5	1.91	5.44e-7	3.29
16	5.61e-4	1.05	1.63e-5	1.96	6.08e-8	3.16
32	2.75e-4	1.03	4.13e-6	1.98	7.14e-9	3.09
64	1.36e-4	1.02	1.04e-6	1.99	8.64e-10	3.05
128	6.76e-5	1.01	2.61e-7	1.99	1.05e-10	3.04
256	3.37e-5	1.00	6.54e-8	2.00	1.03e-11	3.35

Tabelle 4.7: zeigt Fehlerwerte der Kollokation sowie die daraus resultierenden Ordnungen. Die Spalten mit Bezeichnung $F_i, i = 1 \dots 3$ beschreiben die Fehlerwerte an der Stelle $x_0 = \frac{1}{2}$. Die mit O_1, O_2 sowie O_3 bezeichneten Spalten entsprechen den Ordnungen.

n	ϵ_1	O_1	ϵ_2	O_2	ϵ_3	O_3
2	4.79e-5		9.78e-6		2.29e-7	
4	5.35e-6	3.16	7.97e-7	3.62	3.29e-9	6.12
8	4.86e-7	3.46	5.44e-8	3.87	5.00e-11	6.04
16	4.51e-8	3.43	3.54e-9	3.94	7.47e-13	6.07
32	4.53e-9	3.32	2.26e-10	3.97	1.17e-14	6.00
64	4.92e-10	3.20	1.42e-11	3.99	3.58e-14	-1.62
128	5.66e-11	3.12	7.96e-13	4.15	6.14e-13	-4.10
256	5.52e-12	3.36	1.64e-12	-1.05	7.04e-14	3.13

Tabelle 4.8: zeigt Fehlerwerte des Schätzers sowie die daraus resultierenden Ordnungen desselbigen. Die Spalten mit Bezeichnung $\epsilon_i, i = 1 \dots 3$ beschreiben die Fehlerwerte an der Stelle $x_0 = \frac{1}{2}$. Die mit O_1, O_2 sowie O_3 bezeichneten Spalten zeigen die betreffenden Ordnungen.

In Tabelle 4.7 sind die Fehlerwerte der Kollokationslösung dargestellt. Aufgrund der asymmetrischen Kollokationsfolgen liegt keine „Superkonvergenz“ mehr vor, die Ordnungen verhalten sich wie erwartet. Tabelle 4.8 zeigt die zu Tabelle 4.7 gehörigen Abweichungen ϵ des Fehlerschätzers E. Man erkennt, dass diese – wie im linearen Fall – durchgehend um 2 höher als die Verfahrensordnungen sind, bei Kollokationsgrad 4 erkennt man allerdings sogar eine um 3 verbesserte Ordnung. Aufgrund der daraus resultierenden kleinen Fehlerwerte bricht die Ordnung schon in der sechsten Zeile zusammen und es kommt zu einer Stagnation der Abweichung ϵ, die Werte sind hier aber schon in einer Größenordnung von 10e-14.

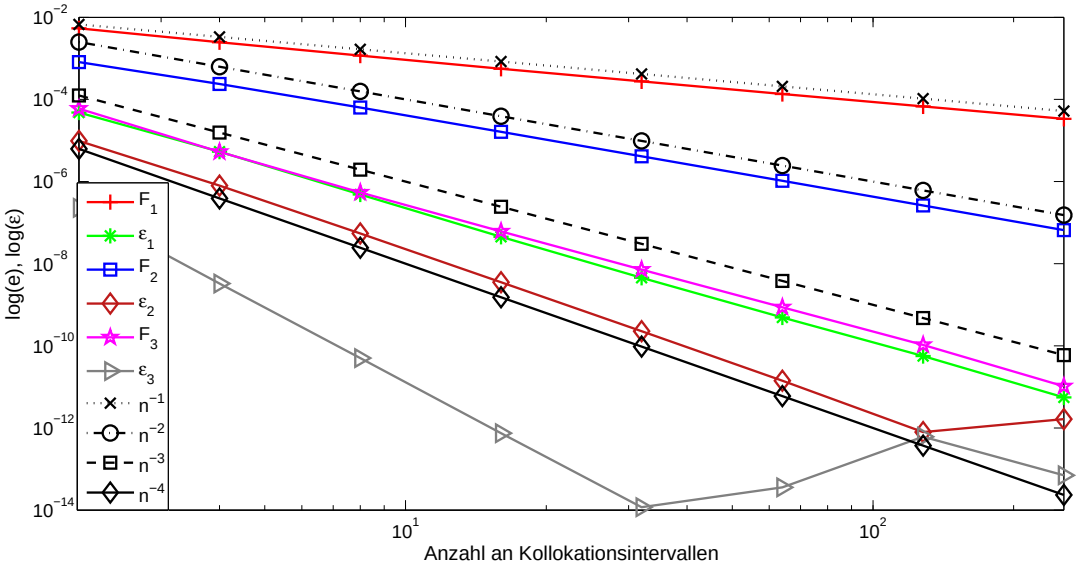

Abbildung 4.5: zeigt – analog zu den Tabellen 4.7 und 4.8 – die Fehler der einzelnen Varianten des Verfahrens, sowie die Abweichungen der daraus resultierenden Fehlerschätzer.

In Abbildung 4.5 ist der Fehlerverlauf aus den obigen beiden Tabellen logarithmisch dargestellt. Die zur Kollokation erster Ordnung gehörigen Geraden (rot und grün) weisen Ordnung 1 bzw. 3 auf. Die durchgehend beobachtbaren Ordnungen 1 bzw. 3 entsprechen perfekt dem Verlauf einer Geraden, was auch bei Kollokation der Ordnung 2 schön zu verfolgen ist. (blaue und rosa Geraden). Bei Kollokation der Ordnung 3 bricht der Geradenverlauf der Abweichung des Fehlerschätzers – analog zur Tabelle – rasch ab und die Fehlerwerte werden wieder größer.

4.3 Beispiel 3

Das nächste Beispiel wird die Funktionalität dieses Fehlerschätzers bei singulären Randwertproblemen zeigen. Unter einem singulären Randwertproblem verstehen wir in diesem Fall eine Gleichung der Form (1.3), wobei für die Koeffizientenfunktionen $q_1(t) = \frac{1}{t}\tilde{q}_1(t)$, $q_2(t) = \frac{1}{t^2}\tilde{q}_2(t)$ gilt. Die Lösung wird als glatte Funktion angenommen, daraus ergibt sich schließlich die rechte Seite. Ein beliebtes Modellproblem, welches z.B. auch in [8] verwendet wird, ist durch folgende Gleichung gegeben:

$$y''(x) + \frac{1}{x}y'(x) - \frac{\mu^2}{x^2} = g(x), \tag{4.3}$$

mit

$$g(x) \equiv cx^{k-2}e^{-\alpha x}(k^2 - \mu^2 - \alpha x(1+2k)) + \alpha^2 cx^k e^{-\alpha x} \tag{4.4}$$

wobei $c \equiv \left(\frac{\alpha}{k}\right)^k e^k$ gilt.

4.3.1 Dirichletrandbedingungen

Das Problem hat mit den Randbedingungen $y(0) = 0$ sowie $y(1) = ce^{-\alpha}$ die exakte und eindeutige Lösung $y(x) = cx^k e^{-\alpha x} \in C^\infty(0,1)$. Für die im Folgenden präsentierten Fehlertabellen seien außerdem noch die Parameter $\alpha = 8$, $k = 4$ sowie $\mu = 1$. In der folgenden Abbildung wird die Lösungsfunktion graphisch veranschaulicht:

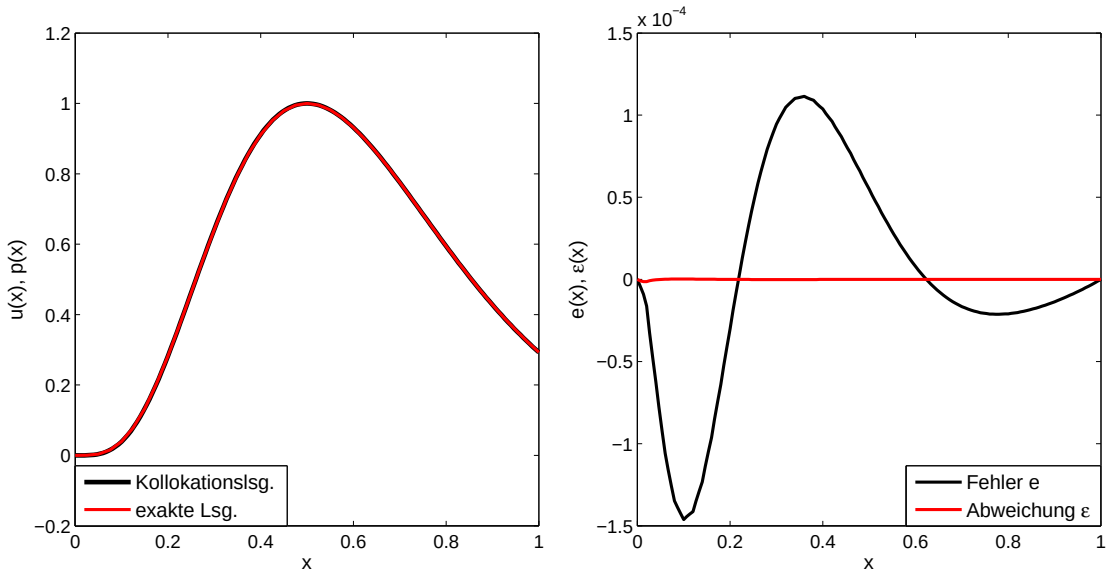

Abbildung 4.6: zeigt im linken Bild die Kollokationslösung sowie die exakte Lösung von (4.3). Im rechten Bild werden der Fehler sowie die Abweichung des Fehlerschätzers dargestellt. Für diese Abbildung wurde die Folge $\zeta = \{0, \frac{1}{2}, 1\}$ gesetzt, das Kollokationsgitter wurde äquidistant mit einer Schrittweite von $h = \frac{1}{50}$ gewählt.

Für die folgenden Fehlertabellen wurden 3 verschiedene Folgen ζ verwendet, wobei wieder unterschiedliche Kollokationsgrade untersucht wurden, das Grundgitter ist in allen 3 Fällen durch $\{0, \frac{1}{2}, 1\}$:

- **Kollokationsgrad 2:** $\zeta_1 = \{0, \frac{1}{2}, 1\}$: Auf Grund der Symmetrie von $\frac{1}{2}$ bezüglich 0 erwartet man Konvergenz der Ordnung 2.
- **Kollokationsgrad 3:** $\zeta_2 = \{0, \frac{1}{3}, \frac{2}{3}, 1\}$: ebenfalls eine symmetrische (sogar äquidistante Folge) an Werten von ζ wird aufgrund der geraden Anzahl innerer Punkte zu Konvergenzordnung 2 führen.
- **Kollokationsgrad 4:** $\zeta_3 = \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$: Man erwartet aufgrund der Symmetrie von ζ_3 Konvergenz der Ordnung 4.

n	F_1	O_1	F_2	O_1	F_3	O_3
2	4.20e-1		9.54e-2		5.80e-2	
4	2.38e-2	4.14	1.04e-2	3.20	4.73e-4	6.94
8	3.69e-3	2.69	2.34e-3	2.15	2.91e-5	4.02
16	7.96e-4	2.21	5.49e-4	2.09	1.27e-6	4.52
32	1.91e-4	2.06	1.35e-4	2.03	6.41e-8	4.31
64	4.71e-5	2.02	3.35e-5	2.01	3.70e-9	4.12
128	1.18e-5	2.00	8.37e-6	2.00	2.27e-10	4.03
256	2.94e-6	2.00	2.09e-6	2.00	2.08e-11	3.45
512	7.34e-7	2.00	5.23e-7	2.00	2.70e-11	0.38
1024	1.83e-7	2.00	1.31e-7	2.00	1.08e-10	2.00

Tabelle 4.9: zeigt Fehlerwerte Kollokation zur Problemstellung (4.3). Die Spalten mit Bezeichnung $O_i, i = 1 \dots 3$ beschreiben den Fehler an der Stelle $x_0 = \frac{1}{2}$. Die mit O_1, O_2 sowie O_3 bezeichneten Spalten zeigen die entsprechenden Ordnungen.

n	ϵ_1	O_1	ϵ_2	O_1	ϵ_3	O_3
2	1.42e-1		1.37e-2		1.15e-2	
4	3.29e-3	5.43	3.45e-4	5.31	3.20e-5	8.49
8	1.57e-4	4.39	2.60e-5	3.73	2.86e-6	3.48
16	9.99e-6	3.97	1.63e-6	4.00	8.60e-8	5.05
32	6.36e-7	3.97	9.42e-8	4.11	1.92e-9	5.48
64	4.00e-8	3.99	5.51e-9	4.10	3.77e-11	5.67
128	2.50e-9	4.00	3.31e-10	4.06	5.16e-13	6.19
256	1.56e-10	4.00	2.04e-11	4.02	2.38e-14	4.44
512	9.74e-12	4.00	2.32e-12	3.13	2.25e-12	-6.56
1024	6.47e-13	3.91	4.27e-12	-0.88	5.60e-12	-1.32

Tabelle 4.10: zeigt die zu Tabelle 4.9 gehörenden Abweichungen der Schätzer sowie die daraus resultierenden Ordnungen derselbigen. Die Spalten mit Bezeichnung $\epsilon_i, i = 1 \dots 3$ beschreiben die Abweichung an der Stelle $x_0 = \frac{1}{2}$. Die Spalten O_1, O_2 sowie O_3 bezeichnen die zugehörigen Ordnungen.

Beim singulären Problem (4.3) erhält man für die Kollokationslösung – entsprechend dem nichtsingulären Fall – wieder die erwarteten Konvergenzordnungen, auch die „Superkonvergenz“ tritt auf (Spalten F_1 sowie O_1). Auch für Kollokation vierter Ordnung erhält man aufgrund der symmetrischen inneren Kollokationsknoten $\{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$ Ordnung 4. Für die zu Kollokation zweiter und dritter Ordnung gehörenden Fehlerschätzer erkennt man auch bei singulärer Gleichung die um 2 verbesserte Ordnung. Für den Schätzer zu Kollokationsgrad 4 erhält man allerdings keine um 2 verbesserte Ordnung. Hier liegt die Ordnung zwischen 5 und 6.

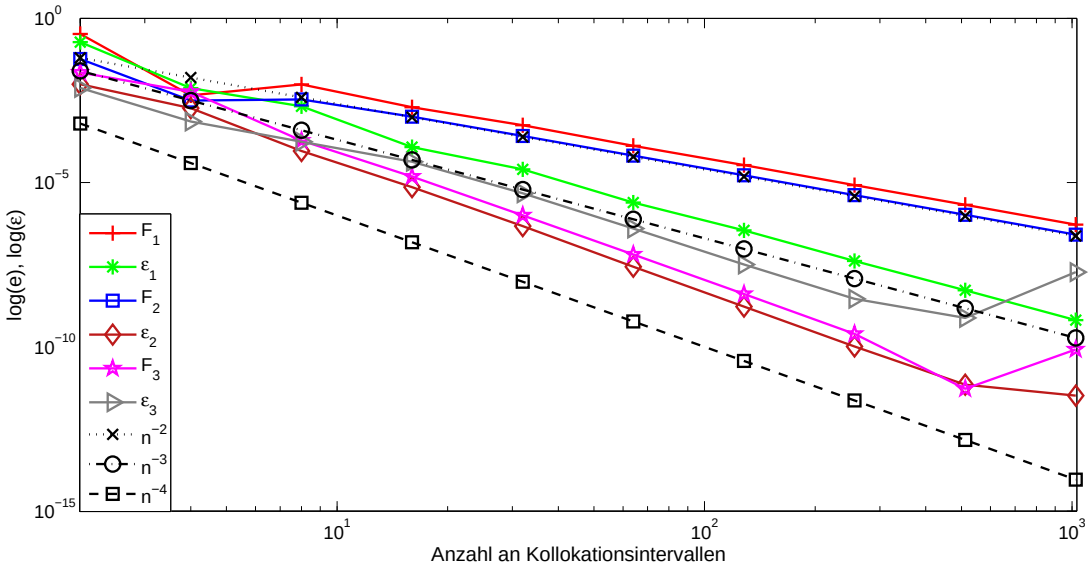

Abbildung 4.7: zeigt die Fehlerwerte aus den Tabellen 4.9 und 4.10, doppelt logarithmisch skaliert.

4.3.2 Allgemeine lineare Randbedingungen

Da die in Abschnitt 4.3 behandelte Gleichung linear ist, kann auch hier der Fehlerschätzer bei allgemeinen linearen Randbedingungen getestet werden: Randbedingungen der Form $y(a) + y'(a) =$ sowie $y'(b)$ ergaben folgende Tabelle (für die Knotenverteilung wählen wir $\zeta = \{0, \frac{1}{3}, 1\}$, das Grundgitter setzen wir mit $\{0, \frac{1}{5}, \frac{1}{2}, 1\}$ fest.):

n	F	O_F	ϵ_1	O_{ϵ_1}	ϵ_2	O_{ϵ_2}
3	2.91e-2		1.12e-1		6.92e-2	
6	7.27e-3	2.00	4.16e-3	4.75	9.68e-4	6.16
12	1.73e-3	2.07	4.80e-4	3.12	1.40e-4	2.80
24	4.28e-4	2.02	7.55e-5	2.67	1.17e-5	3.57
48	1.06e-4	2.01	6.64e-6	3.51	8.24e-7	3.83
96	2.67e-5	2.00	4.84e-7	3.78	5.22e-8	3.98
192	6.66e-6	2.00	3.88e-8	3.64	2.96e-9	4.14
384	1.67e-6	2.00	4.83e-9	3.01	1.79e-10	4.05
768	4.16e-7	2.00	6.03e-10	3.00	1.36e-11	3.71
1536	1.04e-7	2.00	7.71e-11	2.97	3.21e-12	2.09

Tabelle 4.11: zeigt die Fehlerwerte des Kollokationsverfahren bei allgemeinen Randbedingungen. ϵ_1 beschreibt die Abweichung des Schätzers, der mittels einseitiger Differenzenquotienten konstruiert wurde, ϵ_2 beschreibt die Werte jenes Schätzers der mit der 3-Punktregel konstruiert wurde.

Auch bei singulärer Gleichung arbeitet der Fehlerschätzer wünschenswert. Bei Kollokation der Ordnung 3 ist – unabhängig von der Symmetrie von ζ – keine Superkonvergenz gegeben, demnach liefert das Verfahren Ordnung 2. Wie im Fall der regulären (4.1) liefert der mittels einseitiger Differenzenquotienten konstruierte Schätzer Konvergenzordnung 3. Für den mittels 3-Punktregel konstruierten Schätzer gilt auch in diesem Fall $\|\epsilon\| = O(h^{n+2} = O(h^4)$.

4.4 Beispiel 4

Beispiel 4 wird ein singulär gestörtes Problem zeigen, wobei die Gleichung aus Beispiel 2 als Grundlage dienen wird, die linke Seite allerdings noch mit ϵ multipliziert wird:

$$\begin{aligned} \epsilon y''(x) &= 1 - (y')^2 \\ y(0) &= \tfrac{1}{2} \\ y(1) &= \tfrac{1}{4} \end{aligned} \tag{4.5}$$

Für $\epsilon \to 0$ ergibt sich eine stückweise lineare Lösung mit einem Knick an einer Stelle die von den Randdaten abhängt. Für $\epsilon = \frac{1}{100}$ sieht die exakt Lösung folgendermaßen aus:

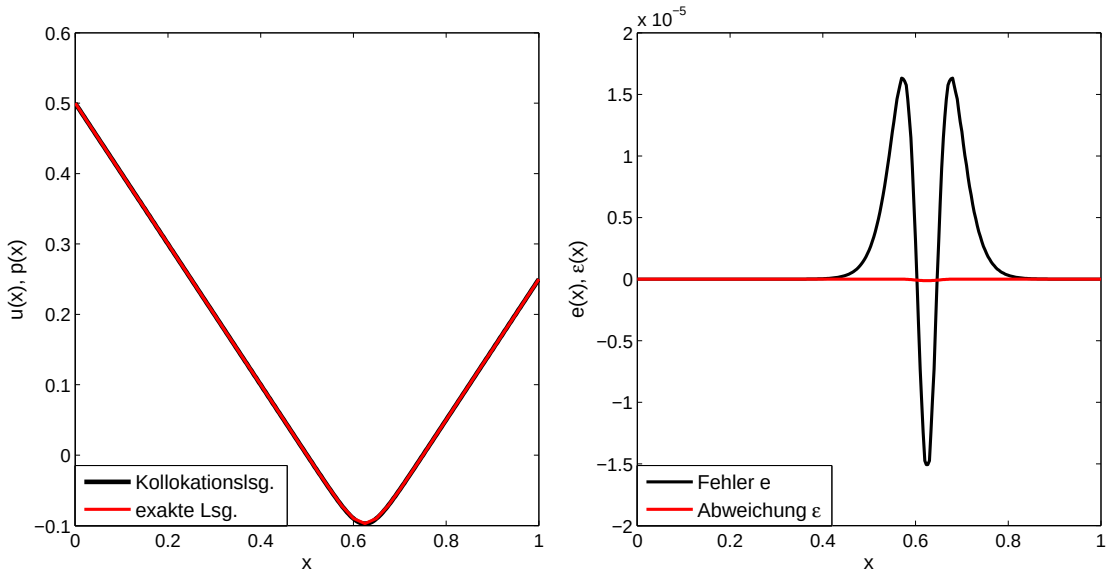

Abbildung 4.8: zeigt im linken Bild die Kollokationslösung sowie die exakte Lösung von (4.5). Im rechten Bild werden der Fehler sowie die Abweichung des Fehlerschätzers dargestellt. Für diese Abbildung wurde die Folge $\zeta = \{0, \frac{1}{2}, 1\}$ gesetzt, das Kollokationsgitter wurde äquidistant mit einer Schrittweite von $h = \frac{1}{100}$ gewählt.

Für die im Folgenden präsentierten Fehlertabellen wurde ϵ mit $\frac{1}{10}$ gewählt, da für $\epsilon = \frac{1}{100}$ keine vernünftigen Werte zu beobachten waren. Wir rechnen mit folgenden Knotenverteilungen, das Startgitter sei für alle 3 Fälle durch $\{0, \frac{1}{4}, 1\}$ gegeben:

- $\zeta_1 = \{0, \frac{1}{3}, 1\}$
- $\zeta_2 = \{0, \frac{1}{2}, 1\}$
- $\zeta_2 = \{0, \frac{1}{4}, \frac{3}{4}, 1\}$

n	F_1	O_1	F_2	O_1	F_3	O_3
2	2.25e-1		4.43e-1		8.84e-2	
4	1.20e-1	0.91	5.90e-2	2.91	1.54e-2	2.52
8	2.05e-2	2.55	1.73e-2	1.77	2.46e-3	2.65
16	6.45e-3	1.67	3.23e-3	2.42	2.88e-4	3.09
32	2.82e-3	1.19	7.72e-4	2.06	6.17e-5	2.23
64	1.26e-3	1.16	1.90e-4	2.02	1.39e-5	2.14
128	5.98e-4	1.07	4.73e-5	2.01	3.39e-6	2.04
256	2.90e-4	1.04	1.18e-5	2.00	8.41e-7	2.01

Tabelle 4.12: zeigt die zu die Fehlerwerte zu den oben genannten Knotenverteilungen ζ_i. Die Spalten mit Bezeichnung $F_i, i = 1 \dots 3$ beschreiben den maximalen Fehler über alle Gitterpunkte. Die Spalten O_1, O_2 sowie O_3 bezeichnen die aus diesen Fehlern resultierenden Ordnungen.

n	ϵ_1	O_1	ϵ_2	O_1	ϵ_3	O_3
2	1.40e-1		7.77e-1		1.17e-1	
4	2.80e-2	2.32	2.92e-2	4.73	4.22e-3	4.79
8	1.87e-2	0.58	4.59e-3	2.67	2.72e-4	3.95
16	7.04e-4	4.73	2.28e-4	4.33	3.09e-5	3.14
32	6.92e-5	3.35	1.25e-5	4.19	2.17e-6	3.84
64	7.17e-6	3.27	7.42e-7	4.08	1.32e-7	4.04
128	8.19e-7	3.13	4.57e-8	4.02	8.16e-9	4.02
256	9.79e-8	3.06	2.85e-9	4.01	5.08e-10	4.01

Tabelle 4.13: zeigt die zu Tabelle 4.7 gehörenden Abweichungen der Schätzer sowie die daraus resultierenden Ordnungen derselbigen. Die Spalten mit Bezeichnung $\epsilon_i, i = 1 \dots 3$ beschreiben die Abweichung an der Stelle $x_0 = \frac{1}{3}$. Die Spalten O_1, O_2 sowie O_3 bezeichnen die zugehörigen Ordnungen.

Für den Fall $\epsilon = \frac{1}{10}$ erhält man noch gut verfolgbare Ordnungen. Für die Variante mit ζ_1 erhält man Konvergenz erster Ordnung. Demnach sind die Fehlerwerte bei 256 Kollokationsintervall erst im Bereich von 10e-4. Für den entsprechenden Fehlerschätzer erhält man Ordnung 3. Bei der zweiten Variante erhält man – aufgrund des

symmetrisch gewählten inneren Punktes $\frac{1}{2}$ wieder Konvergenz der Ordnung 2, bzw. entsprechend dazu Ordnung 4 für den Fehlerschätzer. Für die Kollokationsvariante mit 2 nichtsymmetrischen, nichtäquidistanten inneren Punkten ergibt sich Ordnung 2, speziell beim Basisverfahren ist der Fehler allerdings bei 256 Teilintervallen um einen Faktor 10^{-2} kleiner im Gegensatz zur – ebenfalls Ordnung 2 aufweisenden – symmetrischen, „superkonvergenten“ Variante. Für den Fehlerschätzer erhält man wiederum Ordnung $n + 2 = 4$.

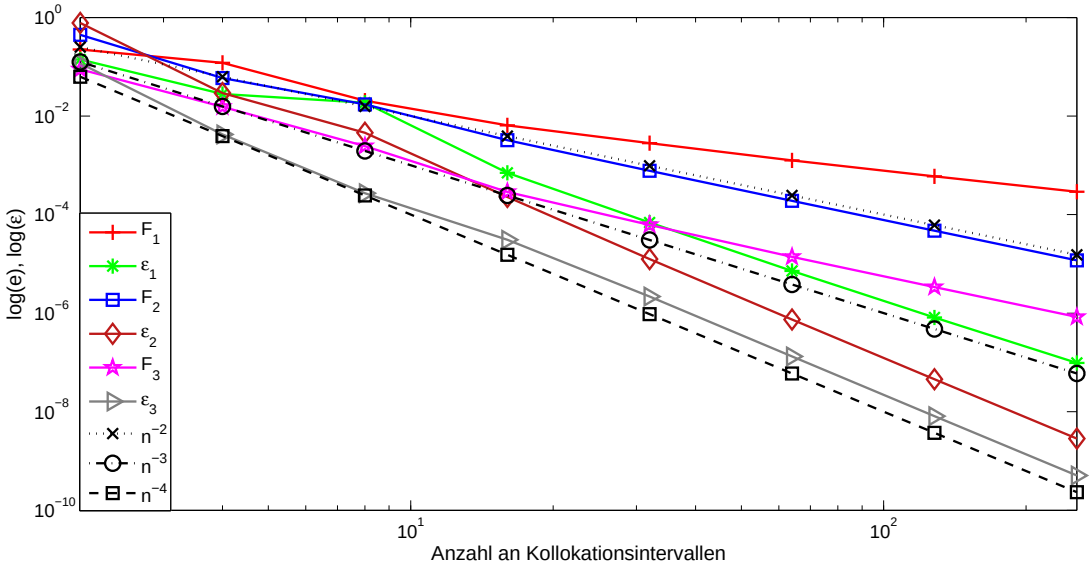

Abbildung 4.9: zeigt wieder – in Analogie zu den übrigen Beispielen – die Fehlerwerte des Verfahrens sowie des Schätzers zu (4.5).

Kapitel 5

Anhang

5.1 Aufbau des Lösers

In diesem Kapitel wird der im Zuge dieser Diplomarbeit verfasste **Matlab**-Code vorgestellt. Die folgenden Funktionen wurden erstellt:

- **Basisverfahren:**
  - kollokation
  - check_settings
  - check_zeta
  - dx_matrizen
  - calculate_defect
- **Defektberechnung:**
  - calculate_defect
  - integral_matrizen
  - integral_koeffizienten_randableitung
  - integral_koeffizienten_randableitung_3punkt_links
  - integral_koeffizienten_randableitung_3punkt_rechts
  - integral_matrizen_singular
- **Hilfsverfahren:**
  - differenzen_verfahren
  - check_settings

Ein Funktionsaufruf ist ähnlich der obigen Aufzählung gestaltet: Die Funktion *kollokation* ruft anfangs *check_settings* sowie *check_zeta* auf, um Benutzereingaben zu überprüfen und eventuelle Zeilenvektoren auf Spaltenvektoren zu transformieren. Mittels den aus *dx_matrizen* stammenden Koeffizienten zur Ableitungsberechnung wird die Differentialgleichung gelöst und die Funktion *calculate_defect* aufgerufen. Abhängig von den Randbedingungen und deren Diskretisierung sowie einer Singularität der Koeffizientenfunktionen im Randpunkt a werden in dieser Funktion unterschiedliche Koeffizienten benötigt, um die interpolatorische Quadratur des Defektes zu bewerkstelligen. Der daraus resultierende Defektvektor stellt die rechte Seite für das im Anschluss aufgerufene Differenzenverfahren *differenzen_verfahren* dar. Auch in dieser Funktion wird die Benutzereingabe überprüft, da diese Funktion auch als eigenständiger Löser zur Verfügung stehen soll.

5.2 Basisverfahren

```
function [Loesung,Punktevektor,eps,Kondition,...
          Kondition_eps,System_matrix,rechte_Seite]...
           = kollokation(s,q,g,R1,R2,SR,Gitter,zeta,rb,singular)
% Inputargumente checken:
[SR,Gitter]=check_settings(SR,Gitter);
zeta=check_zeta(zeta);
% Die Transformationsabbildung tr:
% Jedes x aus[0,1] wird auf y in S_j+1-S_j abgebildet
tr = @(x,j) x * ( Gitter(j+1)-Gitter(j) ) + Gitter(j);
% Notwendige Werte initialisieren:
h=diff(Gitter);                    %...Intervalllaengen h_k
K=length(Gitter)-1;                %...Anzahl an Intervallen
n=length(zeta);                    %...Polynomgrad
System_matrix = sparse(K*n,K*n);   %...Systemmatrix initialisieren
Punktevektor=zeros(K*n,1);         %...Knoten initialisieren
% Berechnungen starten
% Zuerst werden die beiden Zusammenhangsmatrizen berechnet:
[dot,dotdot]=dx_matrizen(zeta);
% Die Systemmatrix besteht aus K verschiedenen n x K*n Bloecken ->
%  Schleife k=1:K
for k=1:K
    % Berechne Systemmatrix Block R_k:
    % Die verschobenen zeta Werte sind im Vektor x_k gespeichert
    x_k=tr(zeta,k);
    [S,Q,G]=Funktionsauswertungen(x_k,s,q,g);
    Punktevektor((k-1)*n+1:k*n,1)=x_k;
    % Die Matrix R_k ergibt sich nach umformen der Gleichung zu:
```

```
28      R_k = 1 / h(k)^2 * dotdot + 1/h(k) * S * dot + Q;
29      % wobei die Matrizen S und Q Diagonalmatrizen mit Werten
30      % entsprechend den Funktionen der Gleichung sind.
31      % Nun werden die Systemmatrix und die rechte Seite
32      % als n x K*n bzw n x 1 Block zusammengebaut:
33      temp=zeros(n,K*n);
34      temp_rs=zeros(n,1);
35      % Kollokationbedingungen im k-ten Intervall
36      % verwenden fuer die Punkte x_1..x_n-1:
37      temp(2:n-1,(k-1)*n+1:k*n)=R_k(2:n-1,:);
38      temp_rs(2:n-1,1)=G(2:n-1);
39      % Im Falle k = 1 muss die erste Zeile durch die erste Zeile der
40      % Randbedingung ersetzt werden, da hier kein Uebergang nach links
41      % stattfinden kann
42      if k==1
43          % Randbedingung 1 fuer y'
44          temp(1,[1:n,(K-1)*n+1:K*n])=[R2(1,1) * 1/h(1)  * dot(1,:),...
45                                       R2(1,2) * 1/h(K-1)*dot(n,:)];
46          % Randbedingung 1 fuer y
47          temp(1,[1,K*n])=[temp(1,1)+R1(1,1),temp(1,K*n)+R1(1,2)];
48          % Randbedingung auf rechter Seite
49          temp_rs(1,1)=SR(1);
50      else
51          %Uebergangsbedingung p nach links
52          h_mean=mean([h(k-1),h(k)]);
53          temp(1,(k-1)*n:(k-1)*n+1)= 1/h_mean^2 * [1 -1];
54          temp_rs(1,1)=0;
55      end
56      % Im Falle k = K muss die letzte Zeile durch die zweite Zeile der
57      % Randbedingung erstetzt werden, da hier kein Uebergang nach rechts
58      % stattfinden kann
59      if k==K
60          % Randbedingung 2 fuer y'
61          temp(n,[1:n,(K-1)*n+1:K*n])= [R2(2,1) * 1/h(1)  * dot(1,:),...
62                                        R2(2,2) * 1/h(K-1)* dot(n,:)];
63          % Randbedingung 2 fuer y
64          temp(n,[1,K*n])=[temp(n,1)+R1(2,1),temp(n,K*n)+R1(2,2)];
65          % Randbedingung auf rechter Seite
66          temp_rs(n,1)=SR(2);
67      else
68          % Uebergangsbedingung p' nach rechts
69          h_mean=mean([h(k+1),h(k)]);
```

```
70          temp(n,(k-1)*n+1:(k+1)*n) = 1/h_mean * [ 1/h(k)   * dot(n,:),...
71                                                   -1/h(k+1)* dot(1,:)];
72          temp_rs(n,1)=0;
73      end
74      % Der naechste Block im n x K*n Format ist fertig berechnet ->
75      % Systemmatrix und ebenso die rechte Seite aktualisieren:
76      System_matrix((k-1)*n+1:k*n,:)=temp;
77      rechte_Seite((k-1)*n+1:k*n,1)=temp_rs;
78      clear temp
79  end
80  % Die Loesung in Matlab erhaelt man mit \:
81  Loesung=System_matrix \ rechte_Seite;
82  % Aufpassen, an den Stuetzstellen liegen die Polynomwerte doppelt vor!
83  Kondition=condest(System_matrix);
84  [D_removed,d]=...
85    calculate_defect(Loesung,s,q,g,R2,zeta,Gitter,rb,singular);
86  % Die doppelt belegten Knoten streichen:
87  Punktevektor(n:n:(K-1)*n)=[];
88  % Doppelpunkte der Loesung streichen
89  Loesung(n:n:(K-1)*n)=[];
90  rechte_Seite_diff=@(x) D_removed(Punktevektor==x);
91  [eps,Kondition_eps]=
92  differenzen_verfahren(s,q,rechte_Seite_diff,...
93                        R1,R2,[D_removed(1),D_removed(end)],...
94                        Punktevektor,rb);
95
96  function [S,Q,G]=Funktionsauswertungen(x,s,q,g)
97  % wird waehrend der Hauptfunktion verwendet um die Punktauswertungen
98  % an den Funktionen s,q und g zu taetigen.
99      n=length(x);
100     temp_q=zeros(n,1);
101     temp_g=temp_q;
102     temp_s=temp_q;
103     for i=1:n
104         temp_q(i)=q(x(i));
105         temp_g(i)=g(x(i));
106         temp_s(i)=s(x(i));
107     end
108     S = diag(temp_s);
109     Q = diag(temp_q);
110     G = temp_g;
```

Der Quellcode der Funktionen *check_settings* sowie *check_zeta* wird an dieser Stelle nicht gezeigt, da diese Funktionen – dem Namen entsprechend – bei Falscheingaben des Benutzer lediglich eine Fehlermeldung zurückliefern.
Die folgende Funktion realisiert die in Lemma (2.3.1) eingeführten Differentiationsmatrizen Ω, sowie $\tilde{\Omega}$

```
function [omega_dot,omega_dot_dot]=dx_matrizen(zeta)
zeta=check_zeta(zeta);
% Polynomgrad festsetzen (der Polynomgrad ist eigentlich n-1,
% wegen der Matlab Indizierung beginnend bei 1)
n=length(zeta);
% Matrix erzeugen:
omega_dot=zeros(n);
for i=1:n
    % Summenindex...Vektor der die Summationsindizes enthaelt:
    % [1,2,...,i-1,i+1,...,n]
    Summenindex=1:n;
    Summenindex(i)=[];
    omega_dot(i,i)=0;
    for s=Summenindex
        omega_dot(i,i) = omega_dot(i,i) + ( x(i)-x(s) )^(-1);
    end
    for j=1:i-1
        % Summenindex...Vektor der die Summationsindizes enthaelt:
        % [1,2,j-1,j+1,...,i-1,i+1,...,n]
        Produktindex=Summenindex;
        Produktindex(j)=[];
        omega_dot(i,j)=1;
        for s=Produktindex
            omega_dot(i,j) = omega_dot(i,j) * ( x(i) - x(s) ) /...
                                              ( x(j) - x(s) );
        end
        omega_dot(i,j) = omega_dot(i,j) * ( x(j) - x(i) )^(-1);
        omega_dot(j,i) = -(omega_dot(i,j) * ( x(j) - x(i) )^2)^(-1);
    end
end
omega_dot_dot=omega_dot^2;
```

Hier bemerkt man in der letzten Zeile, dass die Matrix zur Beschreibung der 2-ten Ableitung (`omega_dot_dot`) als Quadrat der Matrix `omega_dot` zustande kommt. Dieses Vorgehen ist dadurch gerechtfertigt, dass ja die Ableitung eines Polynoms n-ten Grades wieder als Polynom n-ten Grades aufgefasst werden kann.

5.3 Defektberechnung

```
function [D_removed,d] =
calculate_defect(Loesung,s,q,g,R2,zeta,...
                 Gitter,randableitung,singular)

tr = @(x,j) x * ( Gitter(j+1)-Gitter(j) ) + Gitter(j);
h=diff(Gitter);                    %.....Intervalllaengen h_k
K=length(Gitter)-1;                %.....Anzahl an Intervallen
n=length(zeta);                    %.....Polynomgrad
[dot,dotdot]=dx_matrizen(zeta);
% Vektor der aufintegrierten Defekte initialisieren:
D_removed=zeros(1,K*(n-1)+1);

% Die folgenden Matrizen werden benoetigt, um die auftretenden
% Integrale in der Taylorentwicklung auszuwerten (linker sowie
% rechter Anteil separat zu behandeln da eine unterschiedliche
% Gewichtung dieser Anteile benoetigt wird.)
[int_links,int_rechts]=integral_matrizen(zeta);

for i=1:K
    % Die aktuellen kollokationsknoten (im i-ten Intervall):
    zeta_temp=tr(zeta,i);
    if i~=1
        % Wenn moeglich, auch den vorletzten Knoten aus dem (i-1)-ten
        % Intervall auswaehlen:
        zeta_temp_before=tr(zeta(end-1),i-1);
    elseif i==1
        zeta_temp_before=[];
    end
    if i~=K
        % Wenn moeglich, auch den zweiten Knoten aus dem (i+1)-tem
        % Intervall auswaehlen:
        zeta_temp_next=tr(zeta(2),i+1);
    elseif i==K
        zeta_temp_next=[];
    end
    % Um die Werte des Defektes zu berechnen werden die Loesung,
    % deren erste und zweite Ableitung  sowie die
    % Funktionsauswertungen von s,q,g:
    loes_temp=Loesung((i-1)*n+1:i*n);
    loes_temp_str=dot*loes_temp*1/h(i);
    loes_temp_strstr=dotdot*loes_temp*1/h(i)^2;
```

```
42      [S,Q,G]=Funktionsauswertungen(zeta_temp,s,q,g);
43      d((i-1)*n+1:i*n)=loes_temp_strstr+S*loes_temp_str+Q*loes_temp-G;
44      % Um numerische Ungenauigkeiten zu beseitigen:
45      d((i-1)*n+2:i*n-1)=0;
46      % Nun werden die Werte von d mithilfe der Matrizen D_links
47      % und D_rechts aufintegriert:
48      if singular && i==1
49          [int_links_sing,int_rechts_sing]=...
50            integral_matrizen_singular(zeta);
51          D_links=int_links_sing*d(2:n)';
52          D_rechts=int_rechts_sing*d(2:n)';
53      else
54          D_links=int_links*d((i-1)*n+1:i*n)';
55          D_rechts=int_rechts*d((i-1)*n+1:i*n)';
56      end
57      % zeta wird nun um die 2 Stuetzstellen erweitert, um die benoetigten
58      % Quotienten alpha_i = h_i-1 / h_i berechnen zu koennen
59      zeta_temp=[zeta_temp_before;zeta_temp;zeta_temp_next];
60      % Lokale Schrittweiten h_local:
61      h_local=[zeta_temp;0]-[0;zeta_temp];
62      % Erster und letzter Wert gehoeren gestrichen:
63      h_local([1,end])=[];
64      % Die Quotientenfolge alpha wird in Matrixschreibweise berechnet,
65      % um warnings zu vermeiden, wird der Divisor nicht um 0
66      % ergaenzt wie oben, sondern um 1 ergaenzt
67      alpha_local=[h_local;0]./[1;h_local];
68      % ersten und letzten Wert streichen:
69      alpha_local([1,end])=[];
70      % Im ersten und letzten Intervall muessen noch die Quotienten 1
71      % hinzugefuegt werden:
72      if i==1
73          alpha_local=[1;alpha_local];
74      elseif i==K
75          alpha_local=[alpha_local;1];
76      end
77
78      % Die linken Integralanteile sind nun mit dem Faktor 2/(1+alpha) zu
79      % gewichten, die rechten mit 2*alpha / (1+alpha)
80      D_links=D_links.*(2./(1+alpha_local));
81      D_rechts=D_rechts.*(2*alpha_local./(1+alpha_local));
82      % Die gewichteten Integrale addieren
83      D((i-1)*n+1:i*n)=D_links+D_rechts;
```

```
84      % und in den Vektor D_removed eintragen:
85      D_removed((i-1)*n+(2-i):i*n-(i-1))=...
86        D_removed((i-1)*n+(2-i):i*n-(i-1))+...
87        D((i-1)*n+1:i*n);
88      % Der Vektor D enthaelt den linken und rechten Anteil an
89      % Knotenstellen separat (doppelte Punkte), in D_removed sind auch
90      % diese Werte summiert
91  end
92  if singular
93      d(1)=0;
94      d(end)=0;
95  end
96  if randableitung==2
97      I=tr(zeta,K);
98      int_rand=(I(end)-I(1))*...
99        integral_koeffizienten_randableitung(zeta,[0,1]);
100     D_removed(1)  = R2(1,2)*sum(int_rand.*d((i-1)*n+1:i*n)')+...
101                     R2(1,1)*sum(int_rand.*d(1:n)');
102     D_removed(end)= R2(2,2)*sum(int_rand.*d((i-1)*n+1:i*n)')+...
103                     R2(2,1)*sum(int_rand.*d(1:n)');
104 elseif randableitung==3
105     I=tr(zeta,K);
106     [intr_links_rand,intr_rechts_rand]=...
107     integral_koeffizienten_randableitung_verbessert_rechts(zeta,[0,1]);
108     [intl_links_rand,intl_rechts_rand]=...
109     integral_koeffizienten_randableitung_verbessert_links(zeta,[0,1]);
110     D_removed(1) =...
111       R2(1,2)*(sum((I(end)-I(1))*intr_links_rand.*d((i-1)*n+1:i*n)')+...
112       sum((I(end)-I(1))*intr_rechts_rand.*d((i-1)*n+1:i*n)'))+...
113       R2(1,1)*(sum((I(end)-I(1))*intl_links_rand.*d(1:n)')+...
114       sum((I(end)-I(1))*intl_rechts_rand.*d(1:n)'));
115     D_removed(end) =...
116       R2(2,2)*(sum((I(end)-I(1))*intr_links_rand.*d((i-1)*n+1:i*n)')+...
117       sum((I(end)-I(1))*intr_rechts_rand.*d((i-1)*n+1:i*n)'))+...
118       R2(2,1)*(sum((I(end)-I(1))*intl_links_rand.*d(1:n)')+...
119       sum((I(end)-I(1))*intl_rechts_rand.*d(1:n)'));
120 end
121
122 function [S,Q,G]=Funktionsauswertungen(x,s,q,g)
123 % wird waehrend der Hauptfunktion verwendet um die Punktauswertungen
124 % an den Funktionen s,q und g zu taetigen.
125     n=length(x);
```

```
126     temp_q=zeros(n,1);
127     temp_g=temp_q;
128     temp_s=temp_q;
129     for i=1:n
130         temp_q(i)=q(x(i));
131         temp_g(i)=g(x(i));
132         temp_s(i)=s(x(i));
133     end
134     S = diag(temp_s);
135     Q = diag(temp_q);
136     G = temp_g;
```

Von den in der Funktion `calculate_defect` verwendeten Funktionen `integral_matrizen`, `integral_matrizen_singular`, etc. wird nur die Standardvariante als Code präsentiert, da die anderen sehr ähnlich sind.

```
1  function [B_alpha,B_beta]=integral_matrizen(zeta)
2
3  zeta=check_zeta(zeta);
4  n=length(zeta);
5  % Intervalllaengen h:
6  h=diff(zeta);
7  B_alpha=zeros(length(zeta),length(zeta));
8  B_beta =zeros(length(zeta),length(zeta));
9
10 for i=2:n-1
11     % fuer jeden inneren Punkte von zeta entsteht ein Gleichungssystem
12     % der Dimension n x n:
13     rs_alpha=zeros(1,n);
14     rs_beta =zeros(1,n);
15     A=zeros(n,n);
16     % Aufstellen der Systemmatrix A und der Vektoren fuer die
17     % rechte Seite:
18     for k=1:n
19         A(k,:)=zeta.^(k-1);
20         rs_alpha(k)=-zeta(i-1) * ( zeta(i  )^(k) - zeta(i-1)^(k) ) *...
21                  1 / (k) + ( zeta(i)^(k+1)   - zeta(i-1)^(k+1))/(k+1);
22         rs_beta(k)=  zeta(i+1) * ( zeta(i+1)^(k) - zeta(i)^(k)   ) *...
23                  1 / (k) - ( zeta(i+1)^(k+1) - zeta(i)^(k+1))  /(k+1);
24     end
25     % Die Koeffizienten, um die entsprechenden Integrale zu erhalten
26     % werden in B_alpha bzw. B_beta abgespeichert:
27     B_alpha(i,:)=(A \ rs_alpha')'/h(i-1)^2;
```

```
    B_beta(i,:)= (A \ rs_beta')'/h(i)^2;
end
% Fuer den ersten Punkt gibt es nur den rechten Anteil:
i=1;
for k=1:n
    A(k,:)=zeta.^(k-1);
    rs_beta(k)=zeta(i+1)*( zeta(i+1)^(k)-zeta(i)^(k))/k -...
              (zeta(i+1)^(k+1) - zeta(i)^(k+1))/(k+1);
end
B_beta(i,:)= (A \ rs_beta')'/h(i)^2;

% Fuer den letzten Punkt gibt es nur den linken Anteil:
i=n;
for k=1:n
    A(k,:)=zeta.^(k-1);
    rs_alpha(k)=-zeta(i-1) * ( zeta(i)^(k) - zeta(i-1)^(k) ) /k +...
                           ( zeta(i)^(k+1) - zeta(i-1)^(k+1))/(k+1);
end
B_alpha(i,:)=(A \ rs_alpha')'/h(i-1)^2;
```

5.4 Hilfsverfahren

```
function [Loesung,Kondition,System_matrix,l]=...
        differenzen_verfahren(s,q,g,R1,R2,SR,Gitter,rb_discretization)
[SR,Gitter]=check_settings(SR,Gitter);
n=length(Gitter);               % Anzahl an Gleichungen bzw. Unbekannten
System_matrix=sparse(n,n);      % System_matrix initialisieren
l=zeros(n,1);                   % rechte Seite initialisieren.
h=diff(Gitter);                 % Die Intervalllaengen h_k.
for i=2:n-1
    % Berechne zuerst das Verhaeltnis zwischen den
    % beiden Intervalllaengen (h_{i-1} und h_i):
    alpha=h(i-1)/h(i);
    % Die folgenden drei Eintraege ergeben sich durch
    % Diskretisierung der Gleichung mittels Differenzenquotienten:
    System_matrix(i,i-1:i+1)=...
    [2*(h(i)^2*(alpha+1)*alpha )^(-1)-s(Gitter(i))*(h(i)*(1+alpha))^(-1),...
    -2*(h(i)^2*alpha)^(-1)+q(Gitter(i)),...
     2*(h(i)^(2)*(alpha+1))^(-1)+s(Gitter(i))*( h(i)*(1+alpha) )^(-1)];
    % die rechte Seite wird punktweise ausgewertet:
    l(i)=g(Gitter(i));
end
```

```
21 if rb_discretization==3
22     % Berechne die Ableitung an a:
23     x=[Gitter(1),Gitter(2),Gitter(3)];
24     h=[x(2)-x(1),x(3)-x(1)];
25     a3=h(1)/h(2)*1/(h(1)-h(2));
26     a2=h(2)/h(1)*1/(h(2)-h(1));
27     a1=-a3-a2;
28     ystrich_a =  [a1,a2,a3];
29     % Berechne die Ableitung an b:
30     x=[Gitter(n-2),Gitter(n-1),Gitter(n)];
31     h=[x(3)-x(1),x(3)-x(2)];
32     b1=h(2)/h(1)*1/(h(1)-h(2));
33     b2=h(1)/h(2)*1/(h(2)-h(1));
34     b3=-b1-b2;
35     ystrich_b =  [b1,b2,b3];
36     % Randbedingungen in System_matrix hinzufuegen:
37     System_matrix(1,[1:3,n-2:n])=...
38       [R1(1,1)*[1 0 0] + R2(1,1)*ystrich_a,...
39        R1(1,2)*[0 0 1] + R2(1,2)*ystrich_b];
40     System_matrix(n,[1:3,n-2:n])=...
41       [R1(2,1)*[1 0 0] + R2(2,1)*ystrich_a,...
42        R1(2,2)*[0 0 1] + R2(2,2)*ystrich_b];
43 else
44     % Berechne die Ableitung an a:
45     ystrich_a =  1/h(1)*[-1,1];
46     % Berechne die Ableitung an b:
47     ystrich_b =  1/h(end)*[-1,1];
48     % Randbedingungen in System_matrix hinzufuegen:
49     System_matrix(1,[1:2,n-1:n])=System_matrix(1,[1:2,n-1:n])+...
50       [R1(1,1)*[1 0]+R2(1,1)*ystrich_a,...
51        R1(1,2)*[0 1]+R2(1,2)*ystrich_b];
52     System_matrix(n,[1:2,n-1:n])=System_matrix(n,[1:2,n-1:n])+...
53       [R1(2,1)*[1 0]+R2(2,1)*ystrich_a,...
54        R1(2,2)*[0 1]+R2(2,2)*ystrich_b];
55 end
56 % zuletzt werden die Randbedingungen in die rechte Seite eingebaut:
57 l(1)=SR(1);
58 l(n)=SR(2);
59
60 % Die Loesung erhaelt man in Matlab mittels \ Operator:
61 Loesung=System_matrix \ l;
62 Kondition=condest(System_matrix);
```

Die nichtlineare Variante des Codes wird an dieser Stelle nicht mehr präsentiert, die Idee soll aber kurz erläutert werden: Um nichtlineare Gleichungen behandeln zu können benötigt man in erster Linie – wie im linearen Fall – die auszuwertende Funktion $f(x, y(x), y'(x))$, die sich nun nicht mehr in Matrixschreibweise implementieren lässt. Um große Gleichungssysteme dennoch lösen zu können war es notwendig die benötigte Gitterfunktion F_h zu implementieren, um diese einem Newtonverfahren zu übergeben. Zu diesem Zweck wird die Gitterfunktion F_h automatisiert in einer eigenen Datei implementiert, wodurch für den Benutzer kein zusätzlicher Aufwand entsteht.

Literaturverzeichnis

[1] P.E. Zadunaisky. On the estimation of errors propagated in the numerical integration of ODEs. *Numer. Math.*, 27:21–39, 1976. 1.1

[2] H.J. Stetter. The defect correction principle and discretization methods. *Numer. Math.*, 29:425–443, 1978. 1.1

[3] W. Auzinger, O. Koch, and E. Weinmüller. Efficient collocation schemes for singular boundary value problems. *Numer. Algorithms*, 31:5–25, 2002. 1.1

[4] W. Auzinger, G. Kneisl, O. Koch, and E. Weinmüller. SBVP 1.0 - A MATLAB solver for singular boundary value problems. ANUM Preprint No2/02, Vienna University of Technology. 1.1

[5] Lukas Exl. A-Posteriori Fehlerschätzer für Differentialgleichungen höherer Ordnung. Diplomarbeit, TU Wien, 2010. 1.1

[6] U. Ascher, R.M.M. Mattheij, and R.D. Russell. *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations.* Prentice-Hall, Englewood Cliffs, NJ, 1988. 2.5, 3.2.1, 3.2.3

[7] Amir Saboor Bagherzadeh. Defect-based error estimation for higher order differential equations. Dissertation, TU Wien, 2010. 3.3.1, 3.4.2, 4.1.3

[8] J. Cash, G. Kitzhofer, O. Koch, G. Moore, and E. Weinmüller. Numerical Solution of Singular Two Point BVPs. *Numer. Math.*, 4:129–149, 2009. 4.3