Ein Algorithmus zur Gittersteuerung bei Kollokationsverfahren für singuläre Randwertprobleme

Winfried Auzinger
Othmar Koch
Wolfgang Polster
Ewa B. Weinmüller

Technical Report

ANUM Preprint No. 21/01

Technische
Universität
Wien

Vienna
University of
Technology

Institute for Applied Mathematics
and Numerical Analysis

Inhaltsverzeichnis

Kapitel 1

Einführung

In dieser Arbeit werden lineare, singuläre Randwertprobleme erster Ordnung der folgenden Form behandelt:

$$z'(t) = \frac{M(t)}{t} z(t) + f(t), \quad t \in (0, 1], \tag{1.1a}$$

$$B_{a2} z(0) + B_{b2} z(1) = \beta_2, \tag{1.1b}$$

$$z \in C[0, 1]. \tag{1.1c}$$

Dabei sind f und z stetige Funktionen in $\mathbb{R}^n$, M ist eine stetige $n \times n$ Matrix, B_{a2}, B_{b2} sind konstante $r \times n$ Matrizen und β_2 ist ein konstanter Vektor in $\mathbb{R}^r$, $r \leq n$. Das Hauptziel dieser Arbeit ist, für die obige Klasse von Problemen Erkenntnisse zu gewinnen, die Basis für die Implementierung eines zuverlässigen Standardcodes sein sollen. Dabei wird das Hauptaugenmerk auf den Entwurf und die Leistungsüberprüfung eines Gittersteuerungsalgorithmus gelegt.

Die Motivation einen Code für singuläre Probleme der Form (1.1) zu entwerfen sind sowohl zahlreiche Anwendungen in der Physik, siehe [6], [9] und [10], der Chemie, [28], und der Mechanik, [7] und [11], als auch weitere Forschungsaktivitäten in verwandten Gebieten, siehe [12], [25], [26] und [27].
Zur numerischen Lösung von singulären Randwertproblemen wird ein Kollokationsverfahren verwendet. Diese Entscheidung basiert auf theoretischen Erkenntnissen und experimentellen Erfahrungen, die die vorteilhaften Eigenschaften dieser Verfahren belegen. Andere Standardmethoden hoher Ordnung, wie explizite Runge-Kutta Methoden oder Mehrschrittverfahren, werden bei der Behandlung der singulären Probleme unzuverlässig, weil sie im Allgemeinen Ordnungsreduktionen zeigen, siehe [19] und [20]. In [4] und [22] wurde gezeigt, dass Schießverfahren sich gut zur Lösung von singulären Randwertproblemen eignen, vorausgesetzt, dass ein sachgemäß gestelltes Anfangs-

wertproblem mit der gleichen Lösung formuliert werden kann. Zur numerischen Lösung der resultierenden Anfangswertprobleme muss man in der Lage sein, eine effiziente Lösungsmethode anzubieten. Leider kann mit dem Schießverfahren nur eine Teilklasse der singulären Randwertprobleme behandelt werden. Das liegt daran, dass nicht zu jedem sachgemäß gestellten singulären Randwertproblem ein äquivalentes, sachgemäß gestelltes Anfangswertproblem existiert, siehe [21].

Wegen ihrer Robustheit und Effizienz gelten die Kollokationsverfahren in Zusammenhang mit regulären Randwertproblemen als attraktive Methoden zur Ermittlung der Lösung. In einem der bestetablierten Standardcodes zur Lösung von Randwertproblemen, `COLSYS` ([1], [2]), wird die Kollokation in Gaußpunkten realisiert. In [20] wurden die Konvergenzeigenschaften von Kollokationsmethoden für eine wichtige, spezielle Klasse[1] von singulären Randwertproblemen untersucht. Für diese Klasse kann die Stufenordnung garantiert werden; Superkonvergenz gilt im Allgemeinen nicht. Daher beschränken wir uns in dieser Arbeit auf äquidistante Verteilung der Kollokationspunkte mit einer geraden Stufenzahl. Für solche Kollokationsverfahren ist es möglich, einen asymptotisch korrekten Fehlerschätzer zu konstruieren, vgl. §3.2. Die für diese Arbeit relevanten Eigenschaften von Kollokationsverfahren werden in §2.2 diskutiert.

Als Basis für die Gittersteuerung wird die Schätzung des globalen Diskretisierungsfehlers herangezogen, weil dazu der lokale Fehler, wegen seiner Unglattheit in der Nähe der Singularität, ungeeignet ist, siehe [16] und [24]. Die Fehlerschätzung basiert auf dem Prinzip der Defektkorrektur ([32]). Der hier vorgeschlagene Schätzer für den globalen Diskretisierungsfehler ist sowohl in den Gitterpunkten, als auch in den *Kollokationspunkten* asymptotisch korrekt, was eine größere Gitterflexibilität zur Folge hat, siehe §3.2.
Das Prinzip der Defektkorrektur kann auch zum Entwurf eines Iterationsverfahrens, *Iterierte Defektkorrektur*, dienen. Diese Methode wurde in [14] bzw. [31] vorgeschlagen, und in [23] wurde gezeigt, dass sie erfolgreich bei der Lösung singulärer Probleme eingesetzt werden kann.

[1] Die Realteile der Eigenwerte von $M(0)$ sind nicht positiv.

1.1 Notationen

Wir bezeichnen mit $\mathbb{C}^n$ den Raum der komplexwertigen Vektoren der Dimension n und mit $|\cdot|$ die dazugehörige Maximumnorm

$$|x| = |(x_1, x_2, \ldots, x_n)| := \max_{1 \le i \le n} |x_i|.$$

In dieser Arbeit ist das Integrationsintervall (o.B.d.A.) $[0, 1]$. Wir schreiben $C_n^p[0, 1]$ für den Raum der komplexen vektorwertigen Funktionen, die p mal stetig auf $[0, 1]$ differenzierbar sind. Für solche Funktionen $y \in C_n^p[0, 1]$ verwenden wir die Maximumnorm

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

Mit $C_{n\times n}^p[0, 1]$ bezeichnen wir den Raum der komplexwertigen $n\times n$ Matrizen, deren Spalten in $C_n^p[0, 1]$ liegen. Häufig wird der untere Index n bzw. $n \times n$ weggelassen. Für $C^0[0, 1]$ schreiben wir $C[0, 1]$.
Für die numerische Behandlung betrachten wir Gitter der Form

$$\Delta := (\tau_0, \tau_1, \ldots, \tau_N), \quad \tau_i \in [0, 1],$$

und bezeichnen

$$h_i := \tau_{i+1} - \tau_i, \quad i = 0, \ldots, N - 1,$$
$$\mathbf{h} := \max_{0 \le i \le N-1} h_i.$$

Wird jedem Gitterpunkt ein Vektor zugeordnet, so bezeichnen wir die Menge dieser Vektoren als *Gittervektoren* und schreiben

$$u_\Delta := (u_0, \ldots, u_N) \in \mathbb{C}^{n(N+1)}.$$

Die Norm der Gittervektoren ist definiert als

$$\|u_\Delta\|_\Delta := \max_{0 \le k \le N} |u_k|.$$

Für eine stetige Funktion $x \in C[0, 1]$ bezeichnet R_Δ die Projektion auf den Raum der Gittervektoren,

$$R_\Delta(x) := (x(\tau_0), \ldots, x(\tau_N)).$$

Es werden überwiegend stückweise äquidistante Gitter der folgenden Form verwendet[2]:

$$\Delta^m := \{t_{i,j} : \ t_{i,j} = \tau_i + j\delta_i, \ i = 0, \ldots, N-1, \ j = 0, \ldots, m+1\}, \\ \delta_i := \tfrac{h_i}{m+1}, \tag{1.2}$$

wobei τ_i die *Gitterpunkte* und $t_{i,j}$ die *Kollokationspunkte* sind. Die Anzahl der Intervalle im Gitter Δ ist $N(\Delta)$, die Anzahl der Intervalle im Gitter Δ^m wird mit

$$N(\Delta^m) = N(\Delta)(m+1)$$

bezeichnet.
Die Gittervektoren u_{Δ^m}, $R_{\Delta^m}(x)$ und $\| \cdot \|_{\Delta^m}$ sind wie vorher spezifiziert zu verstehen. Mit I_n bezeichnen wir die n-dimensionale Einheitsmatrix.

[2]Doppelt auftretende Punkte werden hier nur einmal gezählt.

Kapitel 2

Theoretische Resultate

2.1 Singuläre Randwertprobleme

In diesem Abschnitt formulieren wir die wesentlichen analytischen Eigenschaften eines linearen, singulären Randwertproblems der Form (1.1),

$$\begin{aligned} &z'(t) = \frac{M(t)}{t} z(t) + f(t), \quad t \in (0,1], \\ &B_{a2} z(0) + B_{b2} z(1) = \beta_2, \\ &z \in C[0,1]. \end{aligned}$$

Dabei gilt $f, z \in C_n[0,1]$, $M \in C_{n\times n}[0,1]$ und $B_{a2}, B_{b2} \in \mathbb{R}^{r\times n}, \beta_2 \in \mathbb{R}^r, r \leq n$. Wir formulieren die Randbedingungen die notwendig und hinreichend dafür sind, dass das obige Randwertproblem eine eindeutige Lösung besitzt. Weiters spezifizieren wir die Glattheitseigenschaften dieser Lösung, die hier nicht nur von der Glattheit von f und M abhängen, sondern auch von der Eigenwertstruktur der Matrix $M(0)$. Wir untersuchen zunächst den Fall der konstanten Matrix $M \equiv M(t)$, um die Rolle der Eigenwertstruktur der Koeffizientenmatrix zu illustrieren.
Es sei J die Jordansche Normalform von M, $M = XJX^{-1}$, mit einer regulären Transformationsmatrix X. Wir betrachten also das Problem

$$v'(t) - \frac{J}{t} v(t) = g(t), \tag{2.2}$$

mit $v(t) = X^{-1}z(t)$ und $g(t) = X^{-1}f(t)$. Im ersten Schritt setzen wir voraus, dass $J \in \mathbb{C}^{n\times n}$ aus einem Jordanblock mit dem Eigenwert $\lambda = \sigma + i\rho$ besteht. Für die allgemeine Struktur der Lösung von (2.2) gilt

Lemma 2.1 *Jede Lösung von* (2.2) *hat die Gestalt*

$$v(t) = \Phi(t)c + \Phi(t) \int_1^t \Phi^{-1}(\tau) g(\tau) d\tau,$$

wobei $c \in \mathbb{C}^n$ ein beliebiger Vektor ist und

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

die Matrixlösung[1] *des Anfangswertproblems*

$$\Phi'(t) = \frac{J}{t}\Phi(t), \quad \Phi(1) = I_n, \quad t \in (0, 1]$$

ist.

Für die Matrix t^J gilt

$$t^J = t^\lambda \begin{pmatrix} 1 & \ln(t) & \frac{\ln(t)^2}{2} & \dots & \frac{\ln(t)^{n-1}}{(n-1)!} \\ 0 & 1 & \ln(t) & \dots & \frac{\ln(t)^{n-2}}{(n-2)!} \\ 0 & \ddots & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ln(t) \\ 0 & \dots & \dots & 0 & 1 \end{pmatrix}. \tag{2.3}$$

Im Folgenden wird analysiert unter welchen Bedingungen (2.2) eine eindeutige und stetige Lösung v besitzt. Zusätzlich werden Glattheitseigenschaften von v formuliert. Man erkennt, dass die Glattheit von v nicht nur von f, sondern auch von σ abhängt. Aufgrund der Struktur von (2.3) ist es klar, dass die Fälle $\sigma > 0, \sigma < 0$ und $\lambda = 0$ gesondert behandelt werden müssen[2].

Eigenwerte mit negativem Realteil.

Lemma 2.2 *Sei $\sigma < 0$. Dann gibt es für jedes $g \in C^p[0, 1]$, $p \geq 0$, eine eindeutige Lösung v von (2.2). Weiters gilt*

$$v(t) = t \int_0^1 s^{-J} g(st) ds, \tag{2.4}$$

$v \in C^{p+1}[0, 1]$ und

$$|v(t)| \leq \text{const.} t \|g\|,$$

$$|v'(t)| \leq \text{const.} \|g\|.$$

[1] Die Fundamentalmatrix ist wie folgt definiert: $\Phi(t) = e^{J \ln(t)} := \sum_{i=0}^{\infty} \frac{\ln(t)^i J^i}{i!}$.

[2] Der Fall des rein imaginären Eigenwertes $\lambda = i\rho$ wird ausgeschlossen. Dies führt zu einer Lösung der Form $t^{i\rho} = \cos(\rho \ln t) + i \sin(\rho \ln t)$, die nicht stetig in $t = 0$ ist.

Für $\sigma < 0$ ist die Anfangsbedingung $v(0) = 0$ notwendig und hinreichend für $v \in C[0,1]$.

Eigenwert $\lambda = 0$.

Lemma 2.3 *Sei $\lambda = 0$. Dann gibt es für jedes $g \in C^p[0,1]$, $p \geq 0$, und jede Konstante η eine eindeutige Lösung v von (2.2) mit $v_1(1) = \eta$. Es gilt dann $v \in C^{p+1}[0,1]$, $v_q(0) = 0$, $q = 2, \ldots, n$, und*

$$|v(t)| \leq \text{const.} t(\|g\| + |\eta|),$$

$$|v'(t)| \leq \text{const.} \|g\|.$$

In diesem Fall kann also eine eindeutige, stetige Lösung festgelegt werden, indem man den Wert $v_1(1)$ vorschreibt und $v_q(0) = 0$, $q = 2 \ldots n$, setzt.

Eigenwerte mit positivem Realteil.

Lemma 2.4 *Sei $\sigma > 0$. Dann gibt es für jedes $g \in C^p[0,1]$, $p \geq 0$, und jeden konstanten Vektor $\eta \in \mathbb{C}^n$ eine eindeutige Lösung $v \in C[0,1]$ von (2.2) mit $v(1) = \eta$. Weiters gilt*

1. $0 < \sigma < 1$, $v \in C[0,1] \cap C^1(0,1]$,

$$\begin{aligned} |v(t)| &\leq \text{const.} t^\sigma (1 + |\ln t|^{n-1})(\|g\| + |\eta|), \\ |v'(t)| &\leq \text{const.} t^{\sigma-1}(1 + |\ln t|^{n-1})(\|g\| + |\eta|), \quad 0 \leq t \leq 1, \end{aligned}$$

2. $\sigma = 1$, $v \in C[0,1] \cap C^1(0,1]$,

$$\begin{aligned} |v(t)| &\leq \text{const.} t(1 + |\ln t|^{n})(\|g\| + |\eta|), \\ |v(t)| &\leq \text{const.} (1 + |\ln t|^{n})(\|g\| + |\eta|), \quad 0 \leq t \leq 1, \end{aligned}$$

3. $\sigma > 1$, $v \in C^1[0,1]$,

$$\begin{aligned} |v(t)| &\leq \text{const.} t(\|g\| + |\eta|), \\ |v(t)| &\leq \text{const.} (\|g\| + |\eta|), \quad 0 \leq t \leq 1, \end{aligned}$$

4. (a) $p < \sigma < p + 1$, $v \in C^p[0,1] \cap C^{p+1}(0,1]$,

$$|v^{(p+1)}(t)| \leq \text{const.} t^{\sigma-p-1}(|\ln t|^{n-1} + 1), \quad 0 < t \leq 1,$$

(b) $\sigma = p+1,\ \ v \in C^p[0,1] \cap C^{p+1}(0,1]$,

$$|v^{(p+1)}(t)| \le \text{const.}(|\ln t|^n + 1), \quad 0 < t \le 1,$$

(c) $\sigma > p+1,\ \ v \in C^{p+1}[0,1]$.

In diesem Fall wird die eindeutige Lösung $v \in C[0,1]$ durch die Vorgabe von $v(1)$ festgelegt.

Nun betrachten wir den allgemeinen Fall

$$z'(t) - \frac{M}{t} z(t) = f(t), \tag{2.5}$$

wobei $M \in \mathbb{C}^{n\times n}$.
Sei R die Spektralprojektion auf den Eigenraum M_0 zum Eigenwert 0, und sei S die Spektralprojektion auf den invarianten Unterraum M_+ zu den Eigenwerten mit positiven Realteilen. Wir bezeichnen mit P die Projektion auf $M_+ \oplus M_0$. Mit Q wird die Projektion auf den Komplementärraum von $M_+ \oplus M_0$ bezeichnet,

$$P := R + S, \quad Q := I_n - P.$$

Mit diesen Bezeichnungen lassen sich die Ergebnisse von Lemma 2.2, Lemma 2.3 und Lemma 2.4 wie folgt zusammenfassen:

Satz 2.1 *Es sei $z \in C[0,1]$ eine Lösung von (2.5) mit $f \in C[0,1]$. Dann gilt*

$$Qz(0) = 0 \quad \textit{und} \quad Sz(0) = 0. \tag{2.6}$$

Um die Lösung eindeutig festzulegen, müssen noch zusätzliche Randbedingungen bei $t = 1$ vorgeschrieben werden. Es gilt

Satz 2.2 *Für jedes $f \in C[0,1]$ und für jeden konstanten Vektor $\eta \in \mathbb{R}^n$ gibt es eine eindeutige Lösung $z \in C[0,1]$ von (2.5) die $Pv(1) = P\eta$ erfüllt.*

Weiters folgt aus (2.6), dass $Mz(0) = 0$ gilt.
Wir befassen uns nun mit dem Fall der variablen Koeffizientenmatrix $M(t)$. Für $M \in C^1_{n\times n}[0,1]$ kann mit Hilfe des Banachschen Fixpunktsatzes das folgende Resultat bewiesen werden:

Satz 2.3 *Für jedes $f \in C[0,1]$ und jeden konstanten n-dimensionalen Vektor η hat das lineare Randwertproblem*

$$z'(t) = \frac{M(t)}{t} z(t) + f(t), \quad t \in (0,1], \tag{2.7a}$$

$$Qz(0) = 0, \quad Pz(1) = P\eta, \tag{2.7b}$$

eine eindeutige und stetige Lösung $z(t)$.

Man beachte, dass sich im Fall der variablen Koeffizientenmatrix die Randbedingungen nicht geändert haben.
Aus (2.7b) folgt, dass die Anfangsbedingungen $Qz(0) = 0$ notwendig und hinreichend dafür sind, dass die Forderung (1.1c) erfüllt ist. Dieses Gleichungssystem beinhaltet $n - r$ linear unabhängige Bedingungen für $z(0)$, wobei $r = \dim(\ker(Q))$. Die verbleibenden r Bedingungen müssen in (1.1b) gestellt werden, damit $z \in C[0,1]$ eindeutig festgelegt werden kann. Die Gesamtstruktur der Randbedingungen kann wie folgt zusammengefasst werden:

$$\begin{pmatrix} B_{a1} \\ B_{a2} \end{pmatrix} z(0) + \begin{pmatrix} 0 \\ B_{b2} \end{pmatrix} z(1) = \begin{pmatrix} 0 \\ \beta_2 \end{pmatrix},$$

wobei $B_{a1} \in \mathbb{R}^{(n-r)\times n}$ aus linear unabhängigen Zeilen von Q besteht.
Für weitere technische Details und die Untersuchung des nichtlinearen Problems verweisen wir auf [18].

2.2 Numerische Verfahren

Wie bereits erwähnt, werden in dieser Arbeit *Kollokationsverfahren* zur Lösung von (1.1) verwendet. Diese Verfahren haben für glatte, reguläre Randwertprobleme besonders vorteilhafte Konvergenzeigenschaften; bei entsprechender Wahl der Kollokationspunkte beobachtet man in den Gitterpunkten besonders hohe Konvergenzordnung (Superkonvergenz). Diese Methoden liefern auch für singuläre Probleme zufriedenstellende Ergebnisse.
Zunächst werden *m-stufige Runge-Kutta Verfahren* vorgestellt, da sich die Kollokationsverfahren als Spezialfälle von diesen ergeben.
Wir betrachten ein reguläres lineares Randwertproblem der Form

$$y'(t) = F(t, y(t)) := A(t)y(t) + r(t), \quad t \in [a,b], \tag{2.8a}$$
$$B_a y(a) + B_b y(b) = \gamma, \tag{2.8b}$$

wobei $y, r, \gamma \in \mathbb{R}^n$ und $A, B_a, B_b \in \mathbb{R}^{n\times n}$ gilt.
Ein m-stufiges Runge-Kutta Verfahren für (2.8) auf einem Gitter Δ ist über die folgenden Gleichungen definiert:

$$y_{i+1} = y_i + h_i \sum_{j=1}^{m} \beta_j F_{i,j}, \quad i = 0, \ldots, N(\Delta) - 1, \tag{2.9a}$$

$$F_{i,j} = F(t_{i,j}, y_i + h_i \sum_{l=1}^{m} \alpha_{jl} F_{i,l}), \quad j = 1, \ldots, m, \tag{2.9b}$$

$$B_a y_0 + B_b y_{N(\Delta)} = \gamma. \tag{2.9c}$$

ρ_1	α_{11}	$\dots$	α_{1m}
$\vdots$	$\vdots$		$\vdots$
ρ_m	α_{m1}	$\dots$	α_{mm}
	β_1	$\dots$	β_m

Tabelle 2.1: Runge-Kutta Schema

Dabei sind die Punkte $t_{i,j}$ mit

$$t_{i,j} = \tau_i + h_i\rho_j, \quad 1 \le j \le m, \quad 0 \le i \le N(\Delta) - 1, \tag{2.10}$$

$$0 \le \rho_1 \le \rho_2 \le \cdots \le \rho_m \le 1, \tag{2.11}$$

gegeben. Die Zahlen α_{jl}, β_j und ρ_j sind reelle Parameter und mit y_i ist die Näherungslösung für $y(t_i)$ bezeichnet. Der Index i bezieht sich auf die Gitterpunkte (Teilintervalle), während die Indizes j und l den Punkten innerhalb des jeweiligen Teilintervalls zugeordnet sind. Die Koeffizienten eines Runge-Kutta Verfahrens können, wie in Tabelle 2.1 dargestellt, in einem Tableau angeordnet werden[3]. Ein Runge-Kutta Schema wird als *explizit* bezeichnet, wenn $\rho_1 = 0$ und $\alpha_{jl} = 0$, $j \le l$ gilt, sonst handelt es sich um ein *implizites* Schema. Um (2.9a) und (2.9b) interpretieren zu können führen wir die folgende Schreibweise ein:

$$F_{i,j} = F(t_{i,j}, y_{i,j}), \tag{2.12a}$$

$$y_{i,j} := y_i + h_i \sum_{l=1}^{m} \alpha_{jl} F_{il}. \tag{2.12b}$$

Dabei ist $y_{i,j}$ eine Approximation für $y(t_{i,j})$. Daher ist die Summe in (2.9a) eine Quadraturformel für $\int_{\tau_i}^{\tau_{i+1}} F$ mit Quadraturgewichten $\beta_1, \dots, \beta_m$, und die Summe (2.12b) ist eine Quadraturformel für $\int_{\tau_i}^{t_{i,j}} F$ mit Quadraturgewichten $\alpha_{j1}, \dots, \alpha_{jm}$.
Wenn wir voraussetzen, dass die Genauigkeitsgrade der Quadraturformeln in (2.9a) und (2.12b) p bzw. s sind, dann gilt[4]

$$\int_0^1 \phi(t)dt = \sum_{l=1}^{m} \beta_l \phi(\rho_l), \quad \forall \phi \in \mathbb{P}_p \tag{2.13}$$

[3]Dieses Tableau wird auch *Butcher Array* genannt.
[4]Wir bezeichnen mit $\mathbb{P}_s$ den Raum der Polynome, deren Grad kleiner als s ist.

und

$$\int_0^{\rho_j} \phi(t)dt = \sum_{l=1}^{m} \alpha_{jl}\phi(\rho_l), \quad \forall \phi \in \mathbb{P}_s. \tag{2.14}$$

Im Weiteren fordern wir $p \geq s \geq 1$ und beschränken uns auf Verfahren mit verschiedenen Kollokationspunkten. Die Kollokationsverfahren sind eine Teilmenge der Runge-Kutta Verfahren und werden durch

$$0 \leq \rho_1 < \rho_2 < \cdots < \rho_m \leq 1, \quad s = m \tag{2.15}$$

charakterisiert. Unter dieser Voraussetzung sind bei gegebenen Kollokationspunkten die Quadraturgewichte β_j und α_{jl} eindeutig bestimmt,

$$\beta_j = \int_0^1 L_j(t)dt, \quad \alpha_{jl} = \int_0^{\rho_j} L_l(t)dt.$$

Dabei wird mit $L_j(t)$ die *Lagrangebasis zu den Knoten* ρ_j bezeichnet. Es gilt

$$L_j(t) = \frac{\prod\limits_{\substack{l=0 \\ l\neq j}}^{m} (t - \rho_l)}{\prod\limits_{\substack{l=0 \\ l\neq j}}^{m} (\rho_j - \rho_l)}. \tag{2.16}$$

Um den Zusammenhang zwischen Runge-Kutta Verfahren und Kollokationsverfahren darzustellen setzen wir voraus, dass wir über eine Runge-Kutta Lösung verfügen, das heißt, dass die aus (2.9) berechneten Näherungen y_i und $y_{i,j}$ in allen Gitterpunkten und Kollokationspunkten gegeben sind. Wir definieren daraus wie folgt ein Kollokationspolynom $p(t)$, $t \in [a, b]$. Auf jedem Teilintervall $[\tau_i, \tau_{i+1}]$, $i = 0, \ldots, N(\Delta) - 1$, ist ein Polynom $p_i(t)$ vom Grad m über die Interpolationsbedingungen

$$p_i(\tau_i) = y_i, \quad p_i'(t_{i,j}) = F(t_{i,j}, y_{i,j}), \quad j = 1, \ldots, m, \tag{2.17}$$

spezifiziert. Das, so entstandene, stückweise Polynom erfüllt

$$p_i'(t_{i,j}) = F(t_{i,j}, p_i(t_{i,j})), \ \ i = 0, \ldots, N(\Delta) - 1, \ j = 1, \ldots m, \tag{2.18a}$$
$$p_i(\tau_{i+1}) = p_{i+1}(\tau_{i+1}), \ \ i = 0, \ldots, N(\Delta) - 2, \tag{2.18b}$$
$$B_a p_0(a) + B_b p_{N(\Delta)-1}(b) = \gamma. \tag{2.18c}$$

Man kann zeigen, dass die Verfahren (2.18) und (2.9) mit der Voraussetzung (2.15) bis auf Rechenfehler äquivalent sind.

2.2.1 Konvergenzeigenschaften

Der folgende Satz beschreibt das Konvergenzergebnis für ein Runge-Kutta Verfahren, das auf einem nicht notwendigerweise äquidistanten Gitter Δ mit Kollokationsstellen, die (2.15) erfüllen, ausgeführt wird.

Satz 2.4 *Ein Runge-Kutta Verfahren auf dem Gitter Δ, ist ein Verfahren der Ordnung p für (2.8) mit $A, r \in C^p[a,b]$, d.h.*

$$\begin{aligned} |y_i - y(\tau_i)| &= O(\mathbf{h}^p), \quad 0 \le i \le N(\Delta), \\ |y_{i,j} - y(t_{i,j})| &= O(h_i^{m+1}) + O(\mathbf{h}^p), \quad 1 \le j \le m. \end{aligned} \tag{2.19}$$

Dabei ist zu beachten, dass Superkonvergenz an den Gitterpunkten möglich ist, während an den Kollokationspunkten die Konvergenzordnung $O(\mathbf{h}^{m+1})$ im Allgemeinen nicht verbessert werden kann. Diese Ergebnisse sind in [3] ausführlich diskutiert und bewiesen.
Für das von uns verwendete Kollokationsverfahren, bei dem die m Kollokationspunkte $t_{i,j}$ äquidistant zwischen den Gitterpunkten τ_i verteilt werden, gilt $0 < \rho_1$ und $\rho_m < 1$. Wegen der Singularität dürfen wir nicht zulassen, dass der linke Intervallpunkt ein Kollokationspunkt ist (deshalb $0 < \rho_1$) und für die asymptotische Korrektheit der Fehlerschätzung ist es notwendig, dass der rechte Intervallpunkt kein Kollokationspunkt ist (deshalb $\rho_m < 1$). Für solche Verfahren gilt dann der folgende Satz:

Satz 2.5 *Für ein Kollokationsverfahren das zur Lösung von (2.8) mit $A, r \in C^p[a,b]$ auf dem Gitter Δ^m ausgeführt wird, gilt*

$$\|R_\Delta(p) - R_\Delta(z)\|_\Delta = O(\mathbf{h}^{m+\nu}), \tag{2.20a}$$
$$\|p - z\|_\infty = O(\mathbf{h}^{m+\nu}), \tag{2.20b}$$
$$\|p^{(l)} - z^{(l)}\|_\infty = O(\mathbf{h}^{m+1-l}), \quad l = 1, \ldots, m, \tag{2.20c}$$

wobei $\nu = 0$ für m gerade und $\nu = 1$ für m ungerade.

Für den Beweis des obigen Satzes vgl. [8].
Frühere theoretische Ergebnisse und numerische Experimente lassen vermuten, dass sich dieses Verhalten auf singuläre Randwertprobleme überträgt. In [20] wurden Probleme untersucht, bei denen alle Eigenwerte von $M(0)$ nichtpositiven Realteil haben. In diesem Fall existiert eine eindeutige Kollokationslösung und die Ergebnisse (2.20a) und (2.20b) gelten. Wenn $M(0)$ mehrfachen Eigenwert 0 hat, und der dazugehörige invariante Unterraum nicht mit dem Eigenraum von $M(0)$ identisch ist, können für ungerade m logarithmische Terme in (2.20) auftreten. Da auch der in §3.2 vorgestellte Fehlerschätzer für ungerades m asymptotisch nicht korrekt ist, werden wir

uns in den folgenden Ausführungen auf gerades m beschränken. Das Konvergenzverhalten der Kollokationsverfahren wird an ausgewählten singulären Modellen in §6.1 demonstriert.

2.2.2 Implizites Eulerverfahren

Wählt man in (2.10) $m = 1$ und $\rho_1 = 1$, so entsteht ein Runge-Kutta Verfahren mit

$$\alpha_1 = \beta_1 = \int_0^1 1dt = 1.$$

Dieses Kollokationsverfahren heißt *implizites Euler Verfahren* und es hat die Form

$$y_{i+1} = y_i + h_i F(\tau_{i+1}, y_{i+1}), \;\; i = 0, \ldots, N(\Delta) - 1. \tag{2.21}$$

Hier ist der rechte Gitterpunkt ein Kollokationspunkt. Das implizite Euler Verfahren ist ein Verfahren erster Ordnung. Wir verwenden dieses Verfahren zur Konstruktion des Fehlerschätzers, vgl. §3.

Kapitel 3

Schätzung des globalen Diskretisierungsfehlers

Die Methode zur Schätzung des globalen Diskretisierungsfehlers basiert auf dem Prinzip der Defektkorrektur. Sie wurde zum ersten mal von Zadunaisky vorgeschlagen, siehe [32], und von Frank in [13] analysiert. Bei dieser Variante wird das Basisverfahren auf dem gleichen Gitter zweimal ausgeführt. Dabei wird die Näherungslösung des ursprünglichen Problems berechnet und anschließend die Näherungslösung des sogenannten *Nachbarproblems* ermittelt. Die Differenz dieser Lösungen dient als Fehlerschätzung für den globalen Fehler. Diese Information kann auch zur Korrektur der Basislösung verwendet werden, um eine Näherungslösung mit einer höheren Konvergenzordnung zu erhalten. Die darauf aufbauende Iterationsmethode heißt "Iterierte Defektkorrektur"; ihre Eigenschaften im Zusammenhang mit singulären Problemen wurden in [23] analysiert.

In dieser Arbeit wird die Information aus der Fehlerschätzung zur Berechnung eines neuen Gitters herangezogen, wobei man das Prinzip der Gleichverteilung des globalen Fehlers benutzt. Um das aufwändige Basisverfahren nicht zweimal anwenden zu müssen, wird ein billigeres Verfahren in die Schätzprozedur einbezogen. Für diese Variante der Fehlerschätzung wurde in [15] und [31] für reguläre Probleme gezeigt, dass eine geschickte Kombination des Basisverfahrens und des einfacheren Hilfsverfahrens zu einer asymptotisch korrekten Schätzung für den globalen Fehler des Basisverfahrens führt.

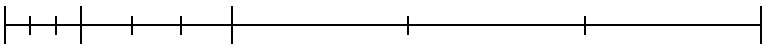


Abbildung 3.1: Ein Gitter, auf dem die klassische Methode anwendbar ist

3.1 Klassische Variante

In der klassischen Variante der Fehlerschätzung benötigen wir ein Gitter Δ, bei dem die Gitterpunkte stückweise äquidistant verteilt sind[1]. Für dieses Gitter setzen wir voraus, dass $N(\Delta) = k\,N_1$ ist, wobei k, $N_1 \in \mathbb{N}$ gilt. Das bedeutet, dass wir für $j = 0, \ldots, N_1 - 1$, k gleiche Schrittweiten

$$h_{kj} = h_{kj+1} = \cdots = h_{k(j+1)-1} =: \bar{h}_j \tag{3.1}$$

wählen. In Abbildung 3.1 ist ein solches Gitter für $k = 3$ und $N_1 = 3$ dargestellt. Die Gitterpunkte in den Teilintervallen sind dabei mit kurzen Strichen markiert. In diesem Abschnitt werden die Gitter ohne Kollokationspunkte dargestellt, die zwischen den kurzen Strichen liegen.
Geht man von einem Startgitter aus, so wird zunächst die Kollokationslösung $R_\Delta(p)$ nach (2.18) bestimmt. Dieser Gittervektor wird komponentenweise mit einer stetigen, stückweisen Polynomfunktion[2] $q(t)$ interpoliert. In jedem Teilintervall wird dabei ein Polynom vom Maximalgrad k durch die Werte in den Punkten $\tau_{kj}, \ldots, \tau_{k(j+1)}$ interpoliert. Mit Hilfe der Funktion $q(t)$ konstruieren wir das Nachbarproblem

$$\begin{aligned} &y'(t) = F(t, y(t)) + d(t), \quad t \in (0, 1], \\ &B_a y(0) + B_b y(1) = \gamma, \end{aligned} \tag{3.2}$$

mit

$$d(t) := q'(t) - F(t, q(t)). \tag{3.3}$$

Laut Konstruktion ist $q(t)$ die exakte Lösung von (3.2). Nun werden die beiden Probleme (2.8) und (3.2) mit einem Verfahren niedriger Ordnung gelöst, um die Näherungslösungen ξ_Δ und π_Δ zu erhalten. In unserem Fall wurde das implizite Eulerverfahren gewählt. Damit sind jeweils die Lösungen

[1] Diese Gitter sind von den in (1.2) vorgestellten Gittern Δ^m zu unterscheiden, da in (3.1) nur die Gitterpunkte τ_i betrachtet werden.

[2] Für das in der Abbildung 3.1 angegebene Gitter wären das in jeder Komponente 3 Polynome vom Grad 3.

der folgenden Gleichungssysteme,

$$\begin{aligned}\frac{\xi_{i+1}-\xi_i}{h_i} &= F(\tau_{i+1},\xi_{i+1}),\\ \frac{\pi_{i+1}-\pi_i}{h_i} &= F(\tau_{i+1},\pi_{i+1})+d(\tau_{i+1}),\end{aligned} \tag{3.4}$$

für $i = 0,\ldots,N(\Delta)-1$ plus Randbedingungen, zu ermitteln. Mit den Randbedingungen sind diese Gleichungssysteme eindeutig lösbar. Um den unbekannten Fehler der Kollokationslösung zu ermitteln überlegt man wie folgt:

$$\begin{aligned}R_\Delta(p)-R_\Delta(z) &= R_\Delta(p)-\xi_\Delta+(\xi_\Delta-R_\Delta(z))\\ &\approx R_\Delta(p)-\xi_\Delta+(\pi_\Delta-R_\Delta(p))\\ &= \pi_\Delta-\xi_\Delta.\end{aligned} \tag{3.5}$$

Die obige Überlegung ist vernünftig, wenn $\xi_\Delta - R_\Delta(z) \approx \pi_\Delta - R_\Delta(p)$. Dafür kann das folgende heuristische Argument angegeben werden: Wenn der Fehler von $R_\Delta(p)-R_\Delta(z)$ klein ist, kann angenommen werden, dass der Defekt $d(t)$ klein ist. Man beachte, dass $R_\Delta(p) = R_\Delta(q)$ gilt. Für reguläre Probleme stellt sich heraus, dass tatsächlich ein asymptotisch korrekter Fehlerschätzer vorliegt, wenn für den Grad k von $q(t)$

$$k \geq m+1$$

gilt. Im Speziellen gilt für $k = m+1$

$$\|(R_\Delta(p)-R_\Delta(z))-(\pi_\Delta-\xi_\Delta)\|_\Delta = O(\mathbf{h}^{m+1}). \tag{3.6}$$

Das Resultat (3.6) folgt aus Ergebnissen in [15] und den in §2 vorgestellten Konvergenzergebnissen für Kollokationsverfahren.
Wendet man dieses Schätzverfahren für singuläre Randwertprobleme (1.1) an, so stellt man experimentell das gleiche Konvergenzverhalten fest. Die numerischen Tests belegen, dass diese Fehlerschätzung robust gegenüber der Singularität bleibt, und eine zuverlässige Information für die in §4 vorgestellte Gittersteuerung liefert.
Die Abbildungen 3.2 und 3.3 beziehen sich auf das Beispiel (5.8), vgl. §5, und demonstrieren wie das adaptierte Gitter der Lösungsstruktur[3] angepasst wird. Abbildung 3.2 zeigt die Daten auf dem äquidistanten Startgitter Δ mit $m = 4$. In der oberen Tabelle der Abbildung sind die wichtigen Kennwerte des Gitters Δ angegeben, wobei h_{min} wie folgt berechnet wird:

$$h_{min} := \min_{i=0,\ldots,N(\Delta)-1} h_i.$$

[3]Man beachte, dass das Gitter in der Nähe der Singularität nicht unnötig fein ist.

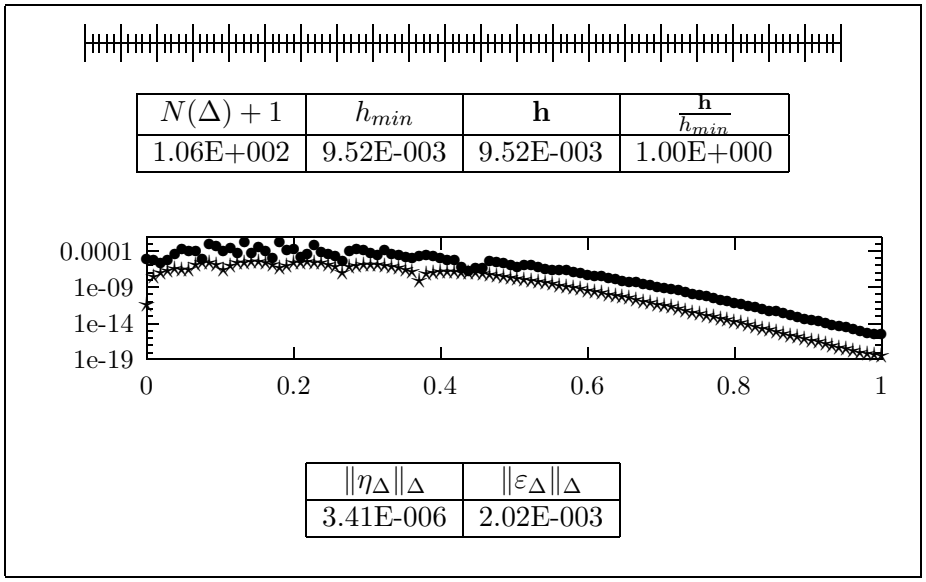


Abbildung 3.2: Klassische Variante: Auswertung auf dem äquidistanten Startgitter

Weiters bezeichnen wir den geschätzten Fehler als

$$\varepsilon_\Delta := \pi_\Delta - \xi_\Delta \tag{3.7}$$

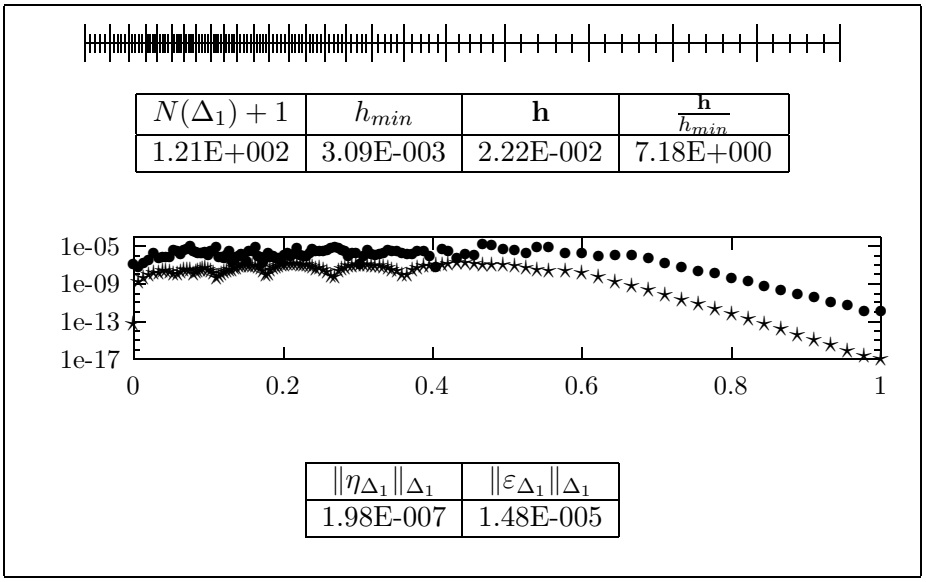
und den exakten globalen Fehler der Kollokationslösung (2.18) mit

$$\eta_\Delta := R_\Delta(p) - R_\Delta(z). \tag{3.8}$$

Die Normen des Fehlers und seiner Schätzung (in jedem Gitterpunkt) sind im Diagramm als logarithmisch skalierte Ordinatenwerte dargestellt und mit folgenden Symbolen bezeichnet:

$$\bullet \doteq |\varepsilon_i| \qquad \star \doteq |\eta_i| \qquad i = 0, \ldots, N(\Delta)$$

Die Abszissenwerte sind die Gitterpunkte des dazugehörigen Gitters. Diese Vereinbarungen gelten, wenn nicht anders angegeben, für alle weiteren Darstellungen der Fehler η_Δ und ε_Δ.
Der geschätzte Fehler ε_Δ liefert eine Grundlage für die Konstruktion eines neuen Gitters Δ_1, vgl. Abbildung 3.3. Man erkennt, dass das Gitter Δ_1 der Lösungsstruktur von (5.8) gut angepasst ist. Die Auswertung auf diesem neuen Gitter, mit nur unwesentlich mehr Gitterpunkten, bringt eine beachtliche



$N(\Delta_1)+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
1.21E+002	3.09E-003	2.22E-002	7.18E+000

$\|\eta_{\Delta_1}\|_{\Delta_1}$	$\|\varepsilon_{\Delta_1}\|_{\Delta_1}$
1.98E-007	1.48E-005

Abbildung 3.3: Klassische Variante: Auswertung auf einem adaptierten Gitter, das aufgrund der Gleichverteilung des globalen Fehlers entstanden ist

Verbesserung des Fehlerniveaus in $\|\eta_{\Delta_1}\|_{\Delta_1}$. Eine ausführliche Beschreibung der Gittersteuerung wird in §4 gegeben.
Bei der klassischen Variante der Fehlerschätzung wird lediglich die Information in den Gitterpunkten τ_i weiterverwendet. Da die Kollokationslösung die gleiche Konvergenzordnung[4] sowohl an den Gitterstellen als auch in den Kollokationspunkten hat, stellt sich die Frage, ob es möglich wäre auch die Information in den Kollokationspunkten zu verwerten und die Fehlerschätzung auf das gesamte Gitter (asymptotisch korrekt) auszudehnen. Dies würde zu kleineren Schrittweiten bei der Anwendung des impliziten Eulerverfahrens führen und damit sowohl das Fehlerniveau dieses Verfahrens verbessern, als auch eine höhere Gitterflexibilität zur Folge haben. Es zeigt sich, dass man so eine modifizierte Strategie angeben kann, sie wird im nächsten Abschitt beschrieben.

3.2 Modifizierte Variante

Eine Möglichkeit die Kollokationspunkte $t_{i,j}$ in die Schätzprozedur einzubeziehen besteht darin, dass man das implizite Eulerverfahren auf dem gesamten Gitter (mit den Kollokationspunkten) ausführt. Damit müssen die

[4]Stufenordnung

folgenden Gleichungssysteme gelöst werden:

$$\frac{\xi_{i,j} - \xi_{i,j-1}}{\delta_i} = F(t_{i,j}, \xi_{i,j}), \tag{3.9a}$$

$$\frac{\pi_{i,j} - \pi_{i,j-1}}{\delta_i} = F(t_{i,j}, \pi_{i,j}) + d(t_{i,j}), \tag{3.9b}$$

$$i = 0, \ldots, N(\Delta) - 1, \quad j = 1, \ldots, m+1,$$

vgl. (1.2). Jetzt ist $q(t)$ eine stückweise Polynomfunktion, die auf jedem Teilintervall $[\tau_i, \tau_{i+1}]$ ein Polynom vom Grad $m+1$ ist, das die Werte der Kollokationslösung an den Stellen $t_{i,j}$ und in den beiden Gitterpunkten τ_i und τ_{i+1} interpoliert. Mit Hilfe von $q(t)$ wird der Defekt $d(t)$ ebenso wie in (3.3) definiert. Die Fehlerschätzung erhält man aus

$$\pi_{\Delta^m} - \xi_{\Delta^m} \approx R_{\Delta^m}(p) - R_{\Delta^m}(z). \tag{3.10}$$

Leider liefert diese direkte Adaptierung der klassischen Vorgangsweise keinen asymptotisch korrekten Fehlerschätzer. Die Begründung dafür ist intuitiv klar:

- Das Polynom $q(t)$ ist nach Konstruktion ein Polynom vom Grad $k = m+1$. Da die Werte von $q(t)$ in $m+2$ Punkten mit den Werten des Polynoms $p(t)$ übereinstimmen, gilt

  $$q(t) \equiv p(t) \tag{3.11}$$

  und deshalb reduziert sich der Grad von $q(t)$ auf m. Da für das Kollokationspolynom $p(t)$,

  $$p'(t_{i,j}) - F(t_{i,j}, p(t_{i,j})) = 0$$

  gilt, verschwindet $d(t)$ an den Kollokationspunkten, vgl. (3.3) und (3.9b). Das bedeutet, dass die nichttriviale Information nur über die Gitterpunkte τ_i übertragen wird.

- Der Beweis dafür, dass der Fehlerschätzer (3.6) asymptotisch korrekt ist, beruht auf der Existenz einer asymptotischen Fehlerentwicklung des globalen Fehlers der Kollokationslösung (siehe [30]). Aus (2.20c) ist jedoch ersichtlich, dass diese Fehlerentwicklung nur in den Gitterpunkten τ_i existieren kann.

Im Folgenden wird der Defekt so modifiziert, dass auf dem Gitter Δ^m eine asymptotisch korrekte Fehlerschätzung angegeben werden kann. Der Defekt $d(t)$ ist nach (3.3) ein Maß dafür, um wieviel $q(t)$ die Bedingungen

$$y'(t_{i,j}) = F(t_{i,j}, y(t_{i,j})),$$
$$i = 0, \ldots, N(\Delta) - 1, \quad j = 1, \ldots, m+1, \tag{3.12}$$

verfehlt. Im Unterschied zu (2.18) ist der rechte Gitterpunkt $\tau_{i+1} = t_{i,m+1}$ hier auch miteinbezogen. Das System (3.12) wird nun in einer Runge-Kutta ähnlichen Form

$$\frac{y_{i,j} - y_{i,j-1}}{\delta_i} = \sum_{k=1}^{m+1} a_{j,k} F(t_{i,k}, y_{i,k}) \tag{3.13}$$

geschrieben. Man beachte, dass in (3.13) die erste Ableitung in gleicher Weise diskretisiert wird wie in (3.9a). Die Koeffizienten $a_{j,k}$ werden so bestimmt, dass die Quadraturformel

$$\frac{1}{t_{i,j+1} - t_{i,j}} \int_{t_{i,j}}^{t_{i,j+1}} \varphi(\tau) d\tau \approx \sum_{k=1}^{m+1} a_{j,k} \varphi(t_{i,k}) \tag{3.14}$$

den Genauigkeitsgrad $m+1$ besitzt. Diese Koeffizienten $a_{j,k}$ ergeben sich zu

$$a_{j,k} = \begin{cases} (m+1)\alpha_{1,k}, & j = 1, \\ (m+1)(\alpha_{j,k} - \alpha_{j-1,k}), & j = 2, \ldots, m+1. \end{cases}$$

Dabei sind $\alpha_{i,k}$ die Einträge im entspechenden Butcher-Array zu dem $(m+1)$-stufigen Runge-Kutta Verfahren, das zu (3.12) äquivalent ist. Man definiert nun den Defekt zur Konstruktion des Nachbarproblems bezüglich des Schemas (3.13),

$$\bar{d}_{i,j} := \frac{p(t_{i,j}) - p(t_{i,j-1})}{\delta_i} - \sum_{k=1}^{m+1} a_{j,k} F(t_{i,k}, p(t_{i,k})). \tag{3.15}$$

Im nächsten Schritt werden die Lösungen, ξ_{Δ^m} und π_{Δ^m}, der folgenden Gleichungssysteme ermittelt:

$$\begin{aligned} \frac{\xi_{i,j} - \xi_{i,j-1}}{\delta_i} &= F(t_{i,j}, \xi_{i,j}), \\ \frac{\pi_{i,j} - \pi_{i,j-1}}{\delta_i} &= F(t_{i,j}, \pi_{i,j}) + \bar{d}_{i,j}, \end{aligned} \tag{3.16}$$

für $i = 0, \ldots, N(\Delta)-1$, $j = 1, \ldots, m+1$ plus Randbedingungen. In [5] wurde gezeigt, dass für reguläre Probleme der Form (2.8) der folgende Satz gilt:

Satz 3.1 *Für ein reguläres, lineares, sachgemäß gestelltes Randwertproblem* (2.8) *gilt*

$$\|(R_{\Delta^m}(p) - R_{\Delta^m}(z)) - (\pi_{\Delta^m} - \xi_{\Delta^m})\|_{\Delta^m} = O(\mathbf{h}^{m+1}). \tag{3.17}$$

Da beim Beweis des obigen Satzes die Linearität von (2.8) keine wesentliche Rolle spielt, gilt das Ergebnis auch für nichtlineare Randwertprobleme. Die Aussage dieses Satzes kann auch auf nicht äquidistante Kollokationspunkte verallgemeinert werden. Bei der Anwendung der modifizierten Methode auf singuläre Randwertprobleme (1.1) konnte die Konvergenzordnung $O(\mathbf{h}^{m+1})$ experimentell bestätigt werden.

3.3 Vergleich der beiden Schätzprozeduren

Es ist zu erwarten, dass die modifizierte Variante der Fehlerschätzung eine zuverlässigere Fehlerschätzung ε_{Δ^m} für den exakten Fehler η_{Δ^m} liefern wird. Die Ergebnisse der Experimente bestätigen diese Erwartungshaltung ausnahmslos, wobei die bemerkenswerte Qualitätssteigerung gegenüber der klassischen Methode an folgenden Beispielen demonstriert wird.
Die Abbildung 3.4 bezieht sich auf das bereits gelöste Beispiel (5.8). Vergleicht man die Abbildungen 3.2 und 3.4, so sieht man, dass die Fehlerschätzung, die mit der modifizierten Variante ermittelt wurde, viel präziser die Größenordnung[5] und den Verlauf des wahren Fehlers wiedergibt. Dabei ist die Anzahl der Gitterpunkte in beiden Fällen gleich 106. Ähnliches ist aus dem Vergleich der Abbildungen 3.3 und 3.4 zu erkennen. Für drei Beispiele sind jetzt die beiden Schätzmethoden direkt verglichen. Dabei sind die zugrundeliegenden Gitter sowohl für die modifizierte, als auch für die klassische Methode angegeben. In den Gittern für die modifizierte Variante werden die Kollokationsstellen $t_{i,j}$ mit kleinen Punkten angedeutet. Für alle Auswertungen gilt $N(\Delta) = 31$ und $m = 4$. Um die Übersicht zu bewahren, werden die Fehler auch für die modifizierte Variante nur an den Gitterpunkten τ_i angegeben[6].

- Beispiel (5.2)

  In Abbildung 3.5 wurde Beispiel (5.2) ausgewertet. Die Fehlerschätzung nach der klassischen Variante ist sehr ungenau und bietet deshalb keine zuverlässige Grundlage für die Gleichverteilung des Fehlers. Die modifizierte Fehlerschätzung ist hingegen sehr zuverlässig und kann als Information bei der Gittersteuerung benutzt werden.

- Beispiel (5.7)

  Die Abbildung 3.6 bezieht sich auf die Auswertung des Beispiels (5.7). Die beobachteten Effekte sind hier ähnlich.

[5] Der geringfügige Unterschied zwischen $\|\eta_{\Delta^4}\|_{\Delta^4}$ und $\|\eta_{\Delta}\|_{\Delta}$ ist darauf zurückzuführen,

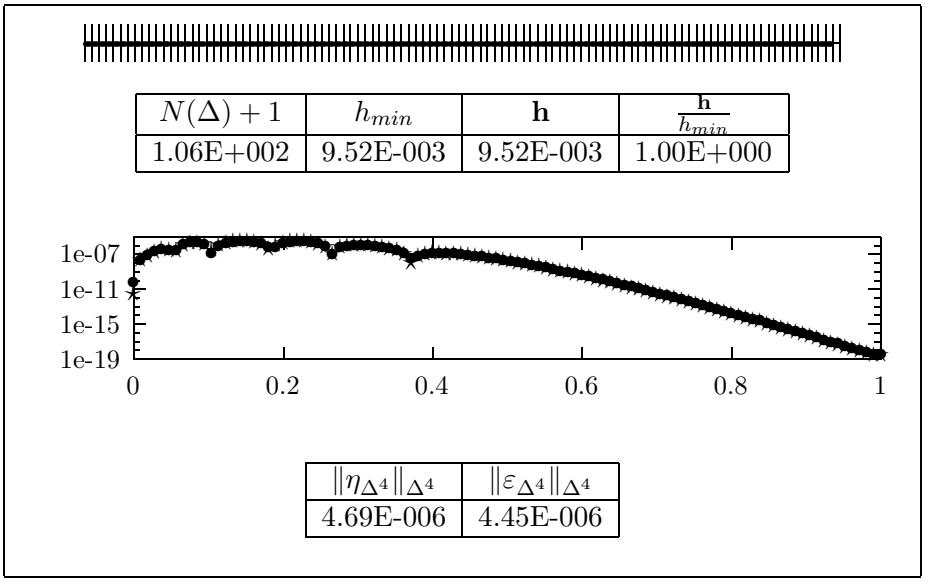


Abbildung 3.4: Modifizierte Variante: Auswertung auf einem äquidistanten Gitter

- Beispiel (5.8)

  Die numerischen Ergebnisse für das Beispiel (5.8) sind in der Abbildung 3.7 zusammengefasst. Hier ist es bemerkenswert, wie extrem die klassische Variante den Fehler in der Nähe der Singularität überschätzt. Die Systemmatrix des impliziten Eulerverfahrens ist in diesem Fall sehr schlecht konditioniert. Diese Erscheinung tritt bei der modifizierten Methode wesentlich seltener auf, da hier grundsätzlich mit feineren Gittern gearbeitet wird. Die Qualität der Fehlerschätzung, die nach der modifizierten Variante ermittelt wurde, ist auch für dieses Beispiel bereits zufriedenstellend.

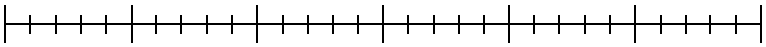
Aufgrund der wesentlich besseren Eigenschaften der modifizierten Variante der Fehlerschätzung, wird im Folgenden nur noch diese Methode für die Gittersteuerung herangezogen. Im nächsten Kapitel beschäftigen wir uns damit, wie die aus der Fehlerschätzung gewonnene Information über den globalen Fehler zur Gittersteuerung verwendet wird. Dabei wird ein neues Gitter angestrebt, auf dem der Fehler η_{Δ^m} gleichverteilt ist.

---

dass bei der Ermittlung von $\|\eta_{\Delta^4}\|_{\Delta^4}$ auch die Kollokationspunkte herangezogen werden.

[6] Das gilt auch für alle weiteren Diagramme, in denen Fehlerverläufe dargestellt sind.

Gitter Δ für die klassische Methode:



Gitter Δ^4 für das modifizierte Verfahren:

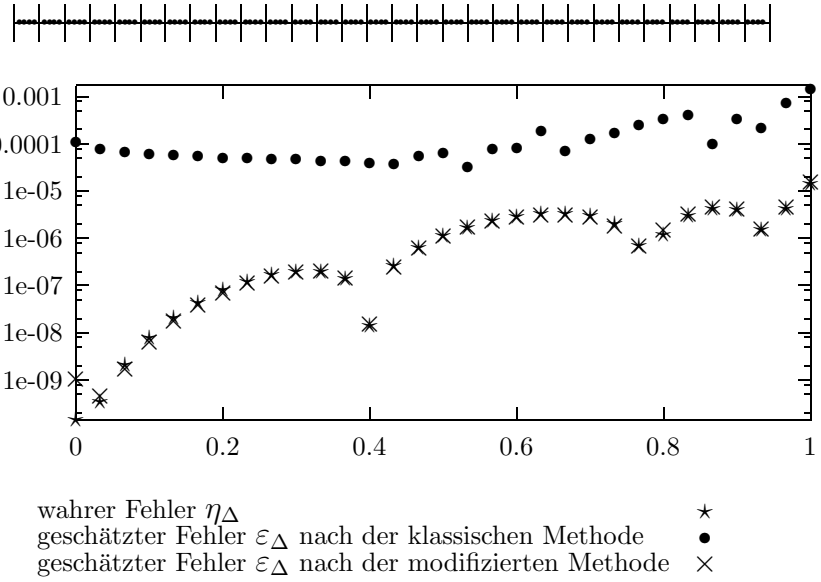
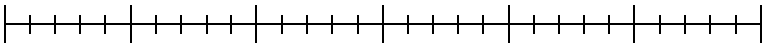
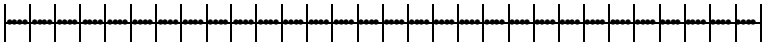


Abbildung 3.5: Gegenüberstellung der Schätzprozeduren für das Beispiel (5.2)

Gitter Δ für die klassische Methode:



Gitter Δ^4 für das modifizierte Verfahren:



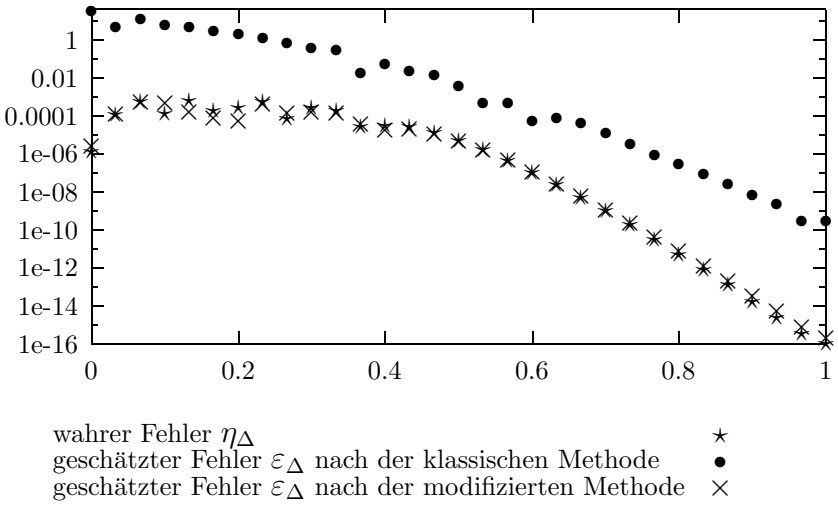
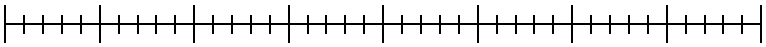
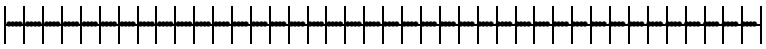


Abbildung 3.6: Gegenüberstellung der Schätzprozeduren für das Beispiel (5.7)

Gitter Δ für die klassische Methode:



Gitter Δ^4 für das modifizierte Verfahren:



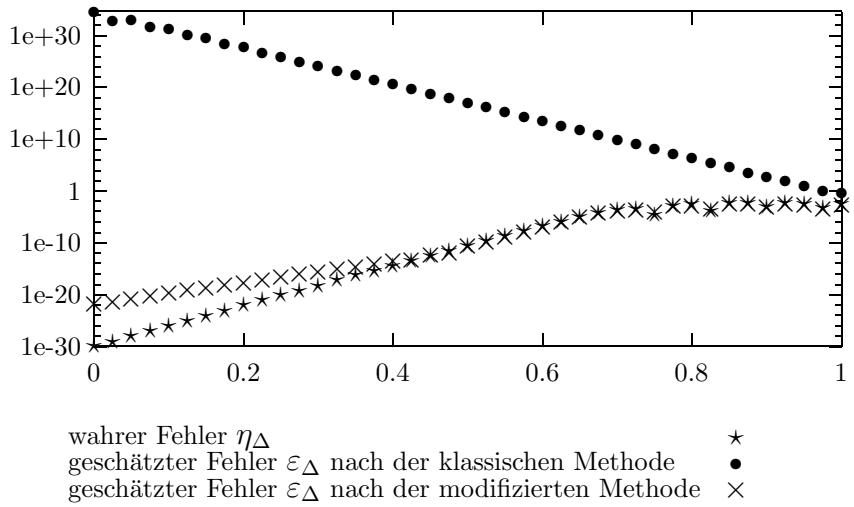


Abbildung 3.7: Gegenüberstellung der Schätzprozeduren für das Beispiel (5.8)

Kapitel 4

Gittersteuerung

4.1 Grundlagen

In den vorhergehenden Kapiteln wurde das Kollokationsverfahren und eine asymptotisch korrekte Schätzung des globalen Fehlers der Kollokationslösung vorgestellt. Ziel dieses Kapitels ist es ein Gitter Δ so zu konstruieren, dass mit möglichst wenigen Intervallen $N(\Delta)$ eine gegebene Toleranzforderung erfüllt wird.
Das führt auf ein *Gittersteuerungsproblem*, das wie folgt formuliert werden könnte: Finde ein Gitter Δ mit möglichst kleiner Intervallanzahl $N(\Delta)$, sodass für die Problemklasse (1.1) und für den Toleranzparameter TOL der Fehler der Kollokationslösung der Forderung $\|\eta_\Delta\|_\Delta < TOL$ genügt.
Es ist klar, dass so ein Gitter Δ im Allgemeinen nicht äquidistant ist. Bei einem Gitter Δ mit dem minimalen $N(\Delta)$, welches das Gittersteuerungsproblem löst, spricht man von einem *optimalen* Gitter. In der Praxis sind solche Gitter entweder nur mit sehr hohem Aufwand zu finden oder gar nicht berechenbar. Wir werden uns daher im Folgenden mit der Suche nach "guten" Gittern befassen - das sind Gitter, deren Intervallanzahl nicht allzusehr von der eines optimalen Gitters abweicht.
Um eine Gittersteuerungsstrategie anwenden zu können, befassen wir uns im Folgenden mit *adaptiven Gittersteuerungen*. Dabei wird das Gitter und die Näherungslösung solange aufgrund von Informationen aus den vorhergehenden Auswertungen neu berechnet, bis die Toleranzforderung erfüllt ist.

4.1.1 Fehlergleichverteilung

Im Folgenden skizzieren wir das wesentliche Konzept der Gleichverteilung einer sogenannten *Monitorfunktion*, das den theoretischen Hintergrund der Gitteranpassung bildet, vgl. [3]. Betrachten wir das analytische Problem

$y'(t) = F(t, y(t))$, $B_a y(a) + B_b y(b) = \gamma$, das mit einem numerischen Verfahren auf dem Gitter Δ gelöst wurde. Der Fehler der Näherungslösung y_i, $i = 0, \ldots, N(\Delta)$, $e_i := |y_i - y(t_i)|$ hängt in der Regel von h_i^m, $m \geq 1$, also nichtlinear von h_i ab. Ist für das neue Gitter N gegeben, so könnte man versuchen, dieses Gitter[1] so zu wählen, dass der Ausdruck

$$\max_{0 \leq i \leq N} |e_i|$$

minimal wird. Diese Aufgabe ist einfacher zu lösen, falls man statt dessen den Ausdruck

$$\max_{0 \leq i \leq N} |d_i|$$

minimiert. Dabei ist d_i ein speziell konstruiertes Fehlermaß mit der Eigenschaft $d_i = h_i \Theta_i$ mit Θ_i, das von h_i unabhängig ist. Diese Funktion Θ_i heißt Monitorfunktion. Man beachte, dass d_i im Gegensatz zu e_i linear von h_i abhängt. Die obigen Überlegungen führen schließlich auf ein Minimierungsproblem

$$\min \left\{ \max_{0 \leq i \leq N} |d_i| : \sum_{i=0}^{N-1} h_i = b - a \right\}, \tag{4.1}$$

wobei a und b die Integrationsgrenzen sind. Die Lösung von (4.1) ist dadurch gegeben, dass alle $|d_i|$ konstant gewählt werden, woraus sich

$$\lambda := \frac{\sum_{i=0}^{N(\Delta)-1} (h_i \Theta_i)}{N}, \quad \tilde{h}_i := \frac{\Theta_i}{\lambda}, \tag{4.2}$$

ergibt.
In dieser Arbeit wird der *globale Fehler* zur Ermittlung von Θ_i herangezogen. Aufgrund der numerischen Ergebnisse aus §6.1 und der Aussagen des Satzes 2.5 werden die Kollokationspunkte in die Berechnungen miteinbezogen. Wir verwenden deshalb die Werte der Fehlerschätzung, $|\varepsilon_i|$, $i = 0, \ldots, N(\Delta_1^m)$, zur Berechnung der Monitorfunktion[2] Θ_i. Um nach (4.2) das neue Gitter berechnen zu können, wird zunächst $N(\Delta_2)$ festgelegt, und anschließend $\lambda = \frac{I}{N(\Delta_2)}$ berechnet, wobei

$$I = \int_a^b \Theta(t) dt \tag{4.3}$$

[1] d.h. die Schrittweiten $\tilde{h}_i$

[2] Die Verwendung von anderen Monitorfunktionen wurde in [29] diskutiert.

gilt. Mit $\Theta(t)$ wird eine Interpolierende von Θ_i bezeichnet. Der Wert λ wird zur Berechnung des neuen Gitters benutzt, wobei

$$\int_{t_{i-1}}^{t_i} \Theta(t)dt = \lambda, \quad i = 1, \ldots, N(\Delta_2)$$

gelten soll.

4.2 Algorithmus zur Gitteranpassung

4.2.1 Grundkonzept

In diesem Kapitel wird der Algorithmus zur Berechnung des angepassten Gitters detailliert beschrieben und seine Wirkungsweise durch numerische Ergebnisse illustriert.
Zur Beurteilung der Güte der Gittersteuerung wird ein Qualitätsmaß eingeführt, anhand dessen entschieden wird, ob ein nicht äquidistantes angepasstes Gitter besser ist als ein äquidistantes Gitter mit gleich vielen Gitterpunkten. Die Abbildung 4.1 bezieht sich auf numerische Berechnungen des Modells (5.8), vgl. §5. Wir arbeiten auf 4 verschiedenen Startgittern, die kohärent verfeinert werden. Die Werte $\|\eta_{\Delta^m}\|_{\Delta^m} = \|R_{\Delta^m}(p) - R_{\Delta^m}(z)\|_{\Delta^m}$ werden in Abhängigkeit von $\mathbf{h}$ dargestellt, wobei beide Achsen logarithmisch skaliert sind. Wegen

$$\|\eta_{\Delta^m}\|_{\Delta^m} = O(\mathbf{h^m}) \quad \text{für} \quad \mathbf{h} \to \mathbf{0}$$

kann das asymptotische Fehlerverhalten für kleine Werte von $\mathbf{h}$ als

$$\|\eta_{\Delta^m}\|_{\Delta^m} \approx c \cdot \mathbf{h}^m \tag{4.4}$$

angenommen werden. Logarithmiert man beidseitig die obere Formel, so erhält man

$$\log \|\eta_{\Delta^m}\|_{\Delta^m} \approx \log c + m \log \mathbf{h}.$$

Die letzte Beziehung bedeutet, dass $\log \|\eta_{\Delta^m}\|_{\Delta^m}$ linear von $\log \mathbf{h}$ abhängt, wobei die Konvergenzordnung m die Steigung der entsprechenden Geraden angibt. Die relative Verbesserung von $\|\eta_{\Delta^m}\|_{\Delta^m}$ ist bei kohärenter Verfeinerung (mit dem Faktor k) von c unabhängig:

$$\begin{aligned}
&\log \|\eta_{\Delta_1^m}\|_{\Delta_1^m} \approx \log c + m \log \mathbf{h_1} \\
&\log \|\eta_{\Delta_2^m}\|_{\Delta_2^m} \approx \log c + m \log \frac{\mathbf{h_1}}{k} \\
&\log \frac{\|\eta_{\Delta_1^m}\|_{\Delta_1^m}}{\|\eta_{\Delta_2^m}\|_{\Delta_2^m}} = \log \|\eta_{\Delta_1^m}\|_{\Delta_1^m} - \log \|\eta_{\Delta_2^m}\|_{\Delta_2^m} \approx m \log k.
\end{aligned} \tag{4.5}$$

Δ	$\|\eta_\Delta\|_\Delta$	$\frac{\mathbf{h}}{h_{min}}$	$Q(\Delta)$
Δ_1	3.73E−07	6,67	1,76
Δ_2	2.60E−05	10	2,3E-03
Δ_3	6.21E−07	1	1
Δ_4	6.76E−05	12,4	9,1E-03

Tabelle 4.1: Vergleich der 4 verschieden verteilten Gitter

Die Verbesserung des Fehlerniveaus ist also von der Lage der gedachten Gerade unabhängig. Deshalb wählen wir das folgende Kriterium dafür, ob ein Gitter "besser ist" als ein anderes mit der gleichen Anzahl der Gitterpunkte: Es sei $N(\Delta_1) = N(\Delta_2)$, dann ist

$$\Delta_1 \text{ "besser verteilt als" } \Delta_2 \Longleftrightarrow Q(\Delta_{1,2}) := \frac{\|\eta_{\Delta_2}\|_{\Delta_2}}{\|\eta_{\Delta_1}\|_{\Delta_1}} > 1. \tag{4.6}$$

Wenn Δ_2 das äquidistante Vergleichsgitter ist, wird der Quotient $\frac{\|\eta_{\Delta_2}\|_{\Delta_2}}{\|\eta_{\Delta_1}\|_{\Delta_1}}$ mit $Q(\Delta_1)$ bezeichnet[3].
Die in Abbildung 4.1 durch Einkreisen hervorgehobenen Punkte haben die gleiche Anzahl von Gitterpunkten (161). Deshalb sind die Auswertungen auf diesen Gittern gleich teuer. Wie das Gitter Δ_2 zeigt, bringt eine schlechte Verteilung der Gitterpunkte ein wesentlich höheres Fehlerniveau. Das Fehlerniveau auf dem Gitter Δ_1 ist besser als das auf dem "gleichteuren" äquidistanten Gitter Δ_3. Intuitiv scheint das Gitter Δ_4 der Lösungstruktur des Problems am besten angepasst zu sein, bringt jedoch in diesem Vergleich das schlechteste Ergebnis. Das könnte daran liegen, dass in diesem Fall ein großes $\mathbf{h}$ einen großen Wert des Quotienten $\frac{\mathbf{h}}{h_{min}} = 12.4$ (und damit eine große Fehlerkonstante c) zur Folge hat. Man beachte, dass große Werte des Quotienten $\frac{\mathbf{h}}{h_{min}}$ sich negativ auf die Stabilität auswirken können.
Die Tabelle 4.1[4] bezieht sich auf die vier Vergleichsgitter, die in der Abbildung 4.1 betrachtet wurden.

[3]Einfache Rechnung zeigt, dass ein nichtäquidantes Gitter Δ_1 genau dann besser verteilt ist als ein anderes nichtäquidistantes Gitter Δ_2, wenn $Q(\Delta_1) > Q(\Delta_2)$ gilt. Dabei werden die Quotienten $Q(\Delta_1)$ und $Q(\Delta_2)$ bezüglich eines äquidistanten Vergleichsgitters mit der gleichen Anzahl von Gitterpunkten berechnet.

[4]Der hier ermittelte Wert von $Q(\Delta_1)$ zeigt nicht das volle Ausmaß der Genauigkeitssteigerung, die durch die Gitteranpassung erreicht werden kann. Das liegt daran, dass die sukzessive kohärente Verfeinerung eines Startgitters mit wenigen Gitterpunkten zu keinem Gitter führt, das der Lösungsstruktur optimal angepasst ist. Für das Beispiel (5.8) ist sogar $Q(\Delta) = 414$ möglich, vgl. §6.6.

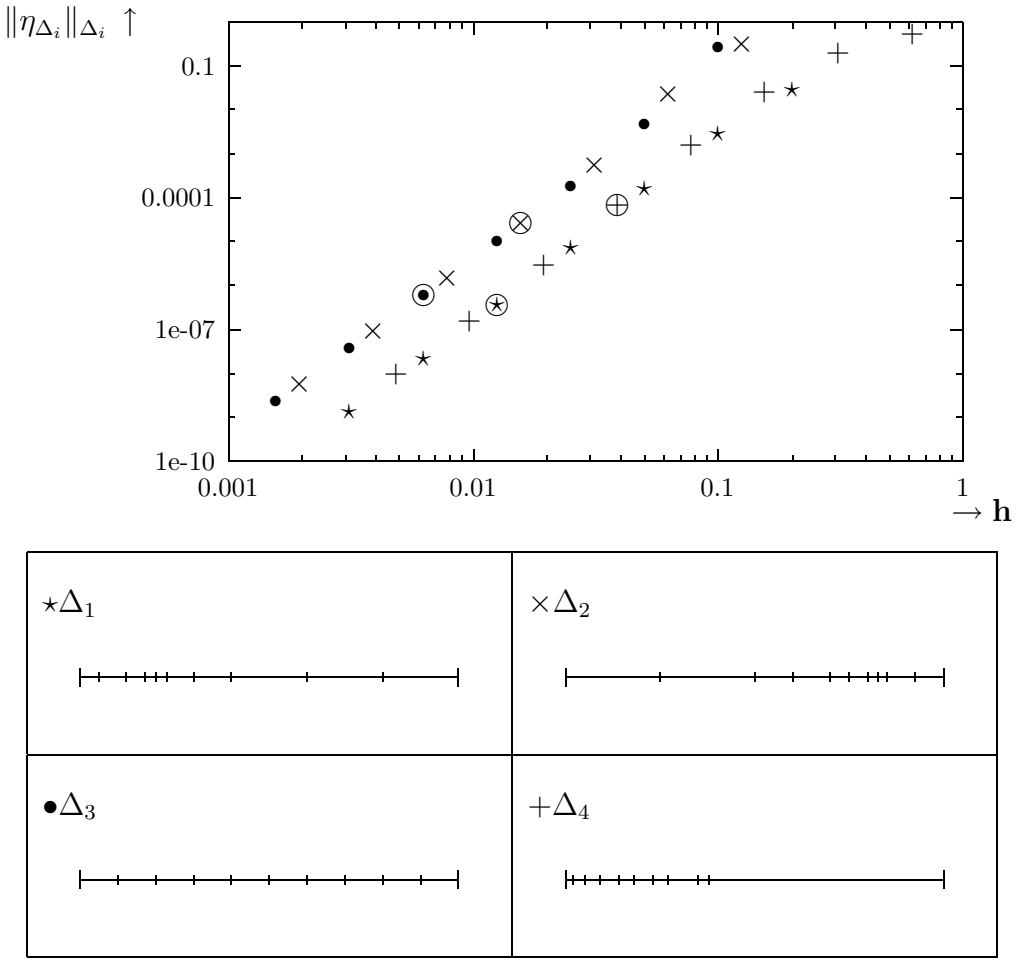


Abbildung 4.1: Gegenüberstellung der Fehlerniveaus bei kohärenter Verfeinerung verschieden verteilter Startgitter

Wir versuchen die unten beschriebene Strategie zur Gitteranpassung zu realisieren:

1. In der Phase 1 wird das äquidistante ***Basisgitter*** berechnet. Die Anforderung an die Güte der Näherungslösung auf diesem Basisgitter ist noch nicht allzu hoch. Es muß aber gewährleistet sein, dass man aus der Fehlerschätzung zuverlässige Information für die Phase 2 gewinnen kann.

2. In der Phase 2 wird das Gitter dem Lösungsverlauf angepasst; dabei wird das Prinzip der ***Gleichverteilung des globalen Fehlers*** näherungsweise realisiert. Es wird angestrebt, dass die Näherungslösung auf diesem neu berechneten Gitter sofort der Toleranzanforderung genügt.

3. Ist dies nicht der Fall so muss das Gitter verfeinert werden, ohne dass die neu gewonnene Strukturinformation verloren geht. Weiters sollte diese Verfeinerung dazu geeignet sein, dass die Toleranz möglichst bald erfüllt wird, da auf den feinen Gittern die Auswertungen teurer sind. Die Phase 3 besteht also aus den ***Gitterverfeinerungen***.

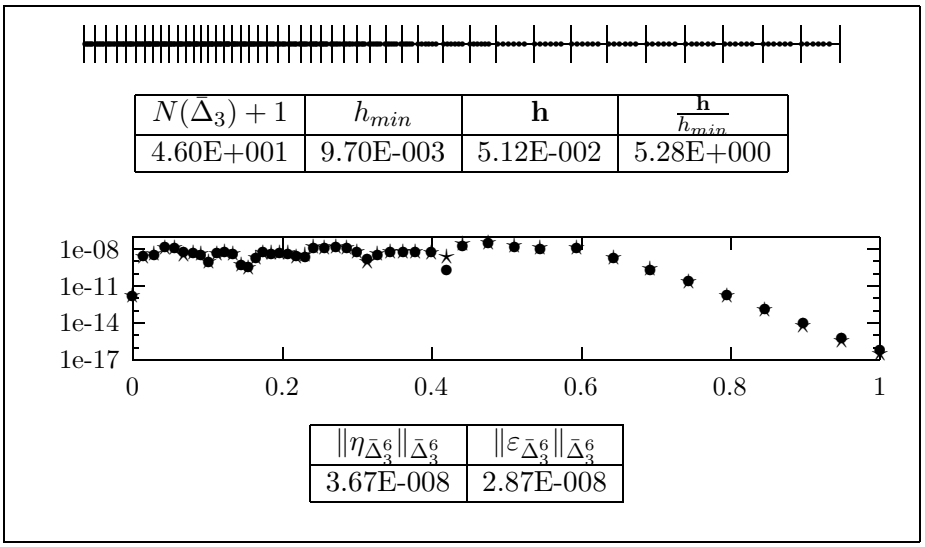
Die Gitter, die während der Realisierung der obigen Strategie entstehen, werden mit $\Delta_{i,j}$ bezeichnet. Dabei gibt der Index $i \in \{1,2,3\}$ Auskunft darüber, in welcher Phase man sich befindet; der Index j gibt an, um das wievielte relevante Gitter es sich in der Phase i handelt. Weiters gibt ein mit Querstrich notiertes Gitter $\bar{\Delta}_{i,j}$ an, dass es sich um ein Gitter handelt, auf dem eine Berechnung der Näherungslösung und ihrer Fehlerschätzung stattfindet. Das jeweils letzte Gitter in der Phase i wird mit $\bar{\Delta}_i$ bezeichnet.

Beispiele

Für das Beispiel (5.8) und[5] rTOL=aTOL=1E−06 wurde mit einem Verfahren der Ordnung m=6 und dem obigen Steuerungsalgorithmus das Gitter $\bar{\Delta}_3^6$, siehe Abbildung 4.2, erzeugt.

Der Fehlerverlauf für das äquidistante Gitter $\bar{\Delta}_a^6$ mit $N(\bar{\Delta}_a) = N(\bar{\Delta}_3)$ wird in der Abbildung 4.3 dargestellt . Man sieht, dass das Gitter $\bar{\Delta}_3^6$ sehr gut verteilt ist; es gilt $Q(\bar{\Delta}_3^6) = 32.70$.

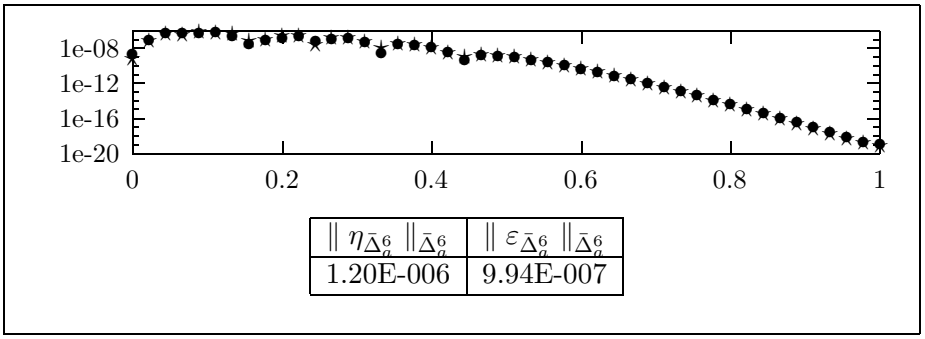
[5]Mit rTOL und aTOL werden die Toleranzen für den relativen bzw. absoluten Fehler bezeichnet.



$N(\bar{\Delta}_3)+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
4.60E+001	9.70E-003	5.12E-002	5.28E+000

$\|\eta_{\bar{\Delta}_3^6}\|_{\bar{\Delta}_3^6}$	$\|\varepsilon_{\bar{\Delta}_3^6}\|_{\bar{\Delta}_3^6}$
3.67E-008	2.87E-008

Abbildung 4.2: Gitter und Auswertungsdaten nach der Gleichverteilung für das Beispiel (5.8), m=6, rTOL=aTOL=1E−06



$\|\eta_{\bar{\Delta}_a^6}\|_{\bar{\Delta}_a^6}$	$\|\varepsilon_{\bar{\Delta}_a^6}\|_{\bar{\Delta}_a^6}$
1.20E-006	9.94E-007

Abbildung 4.3: Auswertung auf dem äquidistanten Gitter $\bar{\Delta}_a^6$, vgl. Abbildung 4.2

$\bar{\Delta}$	$N(\bar{\Delta})+1$	$\|\eta_{\bar{\Delta}^4}\|_{\bar{\Delta}^4}$	$\|\varepsilon_{\bar{\Delta}^4}\|_{\bar{\Delta}^4}$
$\bar{\Delta}_{1,1}$	5	3.27E+03	5.19E+03
$\bar{\Delta}_{1,2}$	22	4.89E+02	7.58E+02
$\bar{\Delta}_{1,3}$	33	1.50E+02	2.95E+02
$\bar{\Delta}_{1,4}$	50	1.72E+01	2.33E+01
$\bar{\Delta}_{2,3}$	54	4.99E−01	4.92E+01
$\bar{\Delta}_{3,1}$	107	3.04E−02	2.99E−02
$\bar{\Delta}_3$	160	5.87E−03	5.84E−03
$\bar{\Delta}_a$	160	9.70E−02	9.9E−02
$\bar{\Delta}_b$	321	5.7E−03	5.74E−03

Tabelle 4.2: Auswertungsdaten für Beispiel (5.4), m=4, aTOL=rTOL=1E−2

Wir verdeutlichen nun die Phasen des Algorithmus anhand von Beispiel[6] (5.4). Die gesamte Folge der Gitter ist in Abbildung 4.4 zu sehen. In der Tabelle 4.2 sind die relevanten Kenngrößen für diese Gitter zusammengefasst, siehe auch Abbildung 4.5. In dieser Abbildung verdeutlichen die vier mit $*$ dargestellten Punkte die Gitter der Phase 1. Das Gitter der Phase 2, auf dem die Auswertung durchgeführt wird, ist mit $\diamond$ dargestellt und die Gitter der Phase 3 werden mit $\bullet$ angedeutet. Die Gitter $\bar{\Delta}_a$ und $\bar{\Delta}_b$, siehe Tabelle 4.2, sind in Abbildung 4.5 mit $\circ$ gekennzeichnet. Es handelt sich um äquidistante Gitter, die mit $\bar{\Delta}_3$ zu vergleichen sind. Für das Gitter $\bar{\Delta}_a$ gilt $N(\bar{\Delta}_a) = N(\bar{\Delta}_3)$, $Q(\bar{\Delta}_3)$= 16.52. Das Gitter $\bar{\Delta}_b$ ist jenes äquidistante Gitter auf dem das Fehlerniveau von $\bar{\Delta}_3$ erreicht wird, $\|\eta_{\bar{\Delta}_b^4}\|_{\bar{\Delta}_b^4} \approx \|\eta_{\bar{\Delta}_3^4}\|_{\bar{\Delta}_3^4}$. Erwähnenswert in diesem Zusammenhang ist die Tatsache, dass die Auswertung auf $\bar{\Delta}_b$ doppelt soviel Rechenaufwand erfordert wie die sieben Auswertungen auf den Gittern in der Abbildung 4.4.

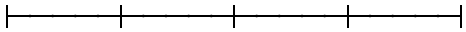
Der "strategische" Aufbau der Gittersteuerung ist in Form eines Flußdiagrammes in Abbildung 4.6 dargestellt. Dabei bedeutet der Block "Auswertung" die Berechnung der Kollokationslösung p_{Δ^m} und der Schätzung des globalen Fehlers ε_{Δ^m} dieser Lösung. Die anderen Blöcke werden in den folgenden Ausführungen detailliert beschrieben.

[6]Hier wurde ein schwieriges Beispiel (sehr unglatte Lösungsstruktur) gewählt, damit alle Phasen des Algorithmus auftreten.

Beispiel (5.4): aTOL=rTOL=1E−2, m=4

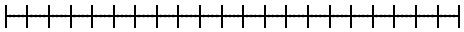
Basisgitterberechnungen

$\bar{\Delta}_{1,1}$:



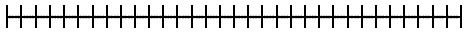
$N(\bar{\Delta}_{1,1})+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
5.00E+000	2.50E-001	2.50E-001	1.00E+000

$\bar{\Delta}_{1,2}$:



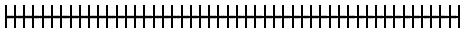
$N(\bar{\Delta}_{1,2})+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
2.20E+001	4.76E-002	4.76E-002	1.00E+000

$\bar{\Delta}_{1,3}$:



$N(\bar{\Delta}_{1,3})+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
3.30E+001	3.12E-002	3.12E-002	1.00E+000

$\bar{\Delta}_{1,4}$:



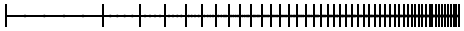
$N(\bar{\Delta}_{1,4})+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
5.00E+001	2.04E-002	2.04E-002	1.00E+000

Gleichverteilung

$\Delta_{2,1}$:

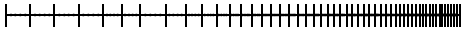
$N(\Delta_{2,1})+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
5.00E+001	5.74E-003	2.14E-001	3.74E+001

$\Delta_{2,2}$:



$N(\Delta_{2,2})+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
5.00E+001	5.74E-003	2.14E-001	3.74E+001

$\bar{\Delta}_{2,3}$:



$N(\bar{\Delta}_{2,3})+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
5.40E+001	5.74E-003	5.62E-002	9.80E+000

Gitterverfeinerungen

$\bar{\Delta}_{3,1}$:

$N(\bar{\Delta}_{3,1})+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
1.07E+002	2.87E-003	2.81E-002	9.80E+000

$\bar{\Delta}_{3,2}$:

$N(\bar{\Delta}_{3,2})+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
1.60E+002	1.91E-003	1.87E-002	9.80E+000

Abbildung 4.4: Die während des Steuerungsalgorithmus auftretenden Gitter für das Beispiel (5.4)

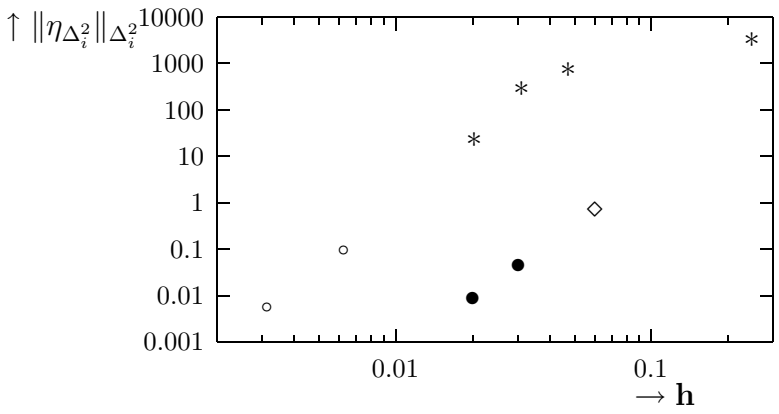


Abbildung 4.5: Beispiel (5.4), m=4, aTOL=rTOL=1E−2

4.2.2 Einfluß der Toleranzen

Hier erklären wir, wie die in Abbildung 4.6 auftretenden Abfragen "Genauigkeit erreicht?" realisiert werden. Der Benutzer gibt die beiden Toleranzparameter aTOL und rTOL an und erwartet, dass der absolute und der relative Fehler der Näherungslösung die, durch diese Parameter vorgeschriebenen, Genauigkeitsforderungen erfüllt. Dies motiviert die folgende (lokale) Toleranzabfrage:

$$|\varepsilon_i| < aTOL + rTOL|p_i|, \quad i = 0, \ldots, N(\Delta^m), \tag{4.7}$$

die sich für rTOL=0 zu

$$|\varepsilon_i| < aTOL, \quad \forall i = 0, \ldots, N(\Delta^m),$$

und für aTOL=0 zu

$$|\varepsilon_i| < rTOL|p_i|, \quad \forall i = 0, \ldots, N(\Delta^m)$$

reduziert. Der Indexlauf hier bedeutet, dass die Abfrage sowohl in den Gitterpunkten als auch in den Kollokationspunkten durchgeführt wird. Die Abfrage (4.7) wurde in der Praxis folgendermaßen realisiert:

$$TOL_q := \max_{i=0,\ldots,N(\Delta^m)} \frac{|\varepsilon_i|}{aTOL + rTOL|p_i|}, \quad TOL_q < 1. \tag{4.8}$$

4.2.3 Basisgitterberechnungen

Die Aufgabe der Phase 1 des Gittersteuerungsalgorithmus ist es, ein äquidistantes Basisgitter $\bar{\Delta}_1$ so zu finden, dass der Fehler $\varepsilon_{\bar{\Delta}_1^m}$ eine zuverlässige Information für die mögliche Gleichverteilung des globalen Fehlers liefert. Man

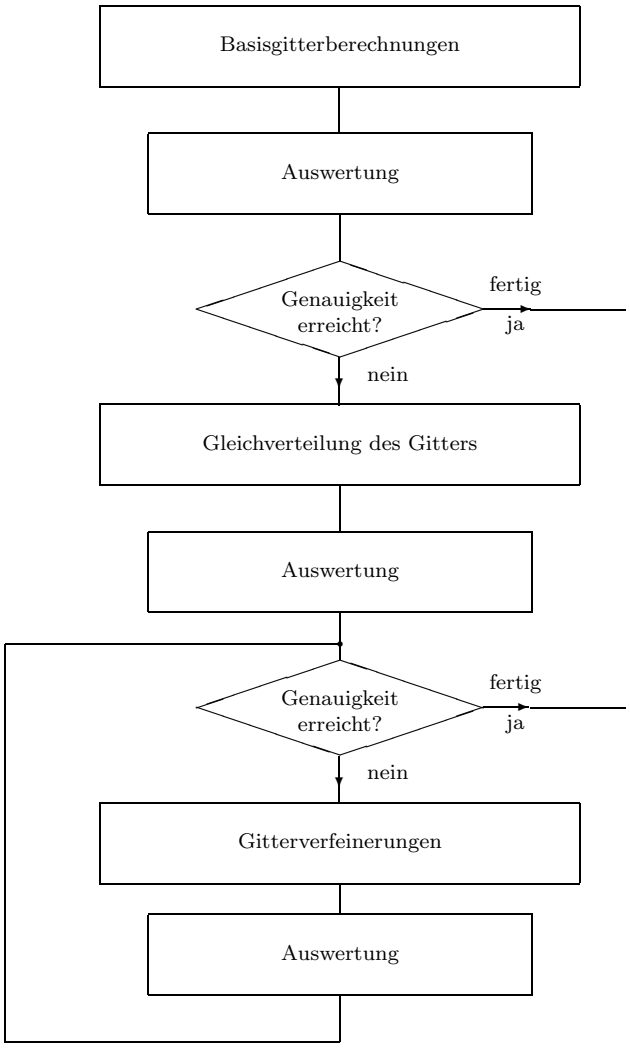


Abbildung 4.6: Schematischer Aufbau des Gittersteuerungsalgorithmus

aTOL \ m	2	4	6	8
1E−1	3	1	1	1
1E−2	10	3	2	1
1E−3	31	5	3	2
1E−4	100	10	4	3
1E−5	316	17	6	4
1E−6	1000	31	10	5
1E−7	3162	56	14	7
1E−8	10000	100	21	10
1E−9	31622	177	31	13
1E−10	100000	316	46	17
1E−11	316227	562	68	23
1E−12	1000000	1000	100	31

Tabelle 4.3: Werte für $N(\bar{\Delta}_{1,1}) + 1$ nach (4.9)

versucht diese Information auf relativ groben Gittern, mit wenig Rechenaufwand, bereitzustellen. In dieser Phase gehen wir von einem äquidistanten Startgitter aus, das kohärent verfeinert wird.
Wir setzen $c = 1$ in (4.4) und berechnen die Schrittweite des äquidistanten Startgitters gemäß

$$h_{ini} = \sqrt[m]{aTOL}.$$

Die Anzahl der Gitterpunkte wird wie folgt festgelegt:

$$N(\bar{\Delta}_{1,1}) + 1 = \left\lfloor \frac{1}{h_{ini}} \right\rfloor, \tag{4.9}$$

vgl. Tabelle 4.3.
An dieser Stelle muß man bedenken, dass man nur den Wert aTOL als Benutzervorgabe zur Verfügung hat und für die praktische Durchführung des Algorithmus nicht nur h_{ini} bzw. $N(\bar{\Delta}_{1,1})$, sondern auch m festzulegen ist. Setzt man glatte Daten des analytischen Problems voraus, so kann man an eine automatische Berechnung vom m denken.

- Automatische Festlegung von m (als Funktion der Toleranz):

  Aus Effizienzgründen wird bei strenger Toleranzanforderung ein Verfahren höherer Ordnung vorzuziehen sein, da im Allgemeinen ein gewisses Fehlerniveau mit so einem Verfahren auf gröberen Gittern er-

reicht wird. Die folgende Tabelle wurde zunächst aufgrund von intuitiver Einschätzung vorgeschlagen und anschließend durch numerische Experimente bestätigt.

aTOL	m
$\in [1E\text{-}1, 1E\text{-}2]$	2
$\in (1E\text{-}2, 1E\text{-}4]$	4
$\in (1E\text{-}4, 1E\text{-}6]$	6
$\in (1E\text{-}6, 1E\text{-}12]$	8

Die Daten und Ergebnisse dieser Experimente werden in §6.3 detailliert diskutiert. Die daraus resultierenden Werte für $N(\bar{\Delta}_{1,1}) + 1$ sind in Tabelle 4.4 hervorgehoben. Mit dem dieser Tabelle entnommenen Wert $N(\bar{\Delta}_{1,1})$, legt man das äquidistante Startgitter $\bar{\Delta}_{1,1}$ wie folgt fest:

$$\bar{\Delta}_{1,1} = \left\{\tau_i : \tau_i = i\delta,\, i = 0, \ldots, N(\bar{\Delta}_{1,1})\right\}, \quad \delta = \frac{1}{N(\bar{\Delta}_{1,1})}. \tag{4.10}$$

- Manuelle Eingabe von m:

  In Situationen wo die Daten des analytischen Problems unglatt sind, und damit im Allgemeinen auch seine Lösung, relativiert sich die Aussage über die Effizienz der Verfahren hoher Ordnung in Zusammenhang mit strengen Toleranzanforderungen. Im Fall von unglatten Lösungsstrukturen wird für Testzwecke die Ordnung m manuell vorgegeben.

Für manche Konstellationen von aTOL und m treten in der Tabelle 4.3 unvernünftig kleine bzw. große Werte für $N(\bar{\Delta}_{1,1}) + 1$ auf. Deshalb wählen wir aus praktischen Überlegungen $5 \leq N(\bar{\Delta}_{1,1}) + 1 \leq 100$, vgl. Tabelle 4.4. Durch Einfügen der äquidistant verteilten Kollokationspunkte entsteht das Gitter $\bar{\Delta}^m_{1,1}$. Damit ist das Startgitter festgelegt und es wird darauf die Kollokationslösung und ihre Fehlerschätzung berechnet. Nun fällt die Entscheidung, ob man dieses Gitter akzeptiert oder verfeinern muss. Gilt

$$\|\varepsilon_{\bar{\Delta}^m_{1,1}}\|_{\bar{\Delta}^m_{1,1}} \leq \|p_{\bar{\Delta}^m_{1,1}}\|_{\bar{\Delta}^m_{1,1}}, \tag{4.11}$$

d.h. im Allgemeinen

$$\|\varepsilon_{\bar{\Delta}^m_{1,j}}\|_{\bar{\Delta}^m_{1,j}} \leq \|p_{\bar{\Delta}^m_{1,j}}\|_{\bar{\Delta}^m_{1,j}}, \tag{4.12}$$

so wird das Gitter akzeptiert und der Algorithmus tritt in die Phase 2, andernfalls muss das Startgitter verfeinert werden. Obwohl in diesem Fall die

aTOL \ m	2	4	6	8
1E−1	**5**	5	5	5
1E−2	**10**	5	5	5
1E−3	31	**5**	5	5
1E−4	100	**10**	5	5
1E−5	100	17	**6**	5
1E−6	100	31	**10**	5
1E−7	100	56	14	**7**
1E−8	100	100	21	**10**
1E−9	100	100	31	**13**
1E−10	100	100	46	**17**
1E−11	100	100	68	**23**
1E−12	100	100	100	**31**

Tabelle 4.4: Werte für $N(\bar{\Delta}_{1,1}) + 1$, die im Testcode verwendet wurden

Fehlerschätzung noch nicht zuverlässig[7] ist, der relative Fehler ist größer als 1, versucht man die Information über die Lösung zu nutzen, um die neue Schrittweite zu berechnen.
Nach [17] kann die neue Schrittweite wie folgt festgelegt werden:

$$den = \Big(\frac{1}{\max(|t_0|, |t_{N(\bar{\Delta}^m_{1,1})}|)}\Big)^m + \|F\|^m, \tag{4.13a}$$

$$h = \Big(\frac{TOL}{den}\Big)^{\frac{1}{m}}. \tag{4.13b}$$

Dabei ist F die rechte Seite in (1.1a) und die Idee ist, $\|F\|$ als Maß der Unglattheit des Problems anzusehen. Die Werte von

$$F(t) = \frac{1}{t}M(t)z(t) + f(t)$$

können näherungsweise durch Einsetzen der Kollokationslösung ermittelt werden,

$$F_i := \frac{1}{\tau_i}M(\tau_i)p(\tau_i) + f(\tau_i), \quad i = 1, \ldots, N(\Delta_{1,1}).$$

Für F_0 setzen wir

$$F_0 \approx \frac{p(\tau_1) - p(\tau_0)}{\tau_1 - \tau_0}, \ \tau_0 = 0.$$

[7] Die Testbeispiele zeigen, dass auf groben Gittern die Fehlerschätzung unzuverlässig ist und das oft vorhandene gute Fehlerniveau nicht richtig wiedergibt.

Beispielhaft seien hier, in Situationen wo Basisgitterverfeinerungen notwendig sind, die Werte für $\|F_{\bar{\Delta}_{1,1}}\|_{\bar{\Delta}_{1,1}}$ angeführt:

Beispiel (5.4), m=6, aTOL=1E−7, rTOL=1E−7	$\|F_{\bar{\Delta}_{1,1}}\|_{\bar{\Delta}_{1,1}} = 3,395 \cdot 10^4$
Beispiel (5.7), m=6, aTOL=1E−7, rTOL=1E−7	$\|F_{\bar{\Delta}_{1,1}}\|_{\bar{\Delta}_{1,1}} = 145,24$
Beispiel (5.10), m=6, aTOL=1E−7, rTOL=1E−7	$\|F_{\bar{\Delta}_{1,1}}\|_{\bar{\Delta}_{1,1}} = 71,81$

Angesichts dieser Werte setzen wir $den = \|F\|^m$ und errechnen die Anzahl der Punkte in dem neuen äquidistanten Basisgitter als

$$N_{ref} = \left\lfloor \frac{\|F_{\bar{\Delta}_{1,1}}\|_{\bar{\Delta}_{1,1}}}{\sqrt[m]{aTOL}} \right\rfloor . \tag{4.14}$$

N_{ref} gibt die Gesamtanzahl der Punkte im neuen Gitter an, es inkludiert sowohl die Gitterpunkte als auch die Kollokationspunkte. Der Wert von N_{ref} entspricht ungefähr dem dazugehörigen Wert aus der Tabelle 4.3 multipliziert mit dem Wert $\|F_{\bar{\Delta}_{1,1}}\|_{\bar{\Delta}_{1,1}}$. Ist der Fehler der numerischen Lösung groß, so kann das daran liegen, dass die exakte Lösung $z(t)$ "unglatt" ist, also ihre Ableitungen große Werte annehmen. Bei solchen $z(t)$ wird $\|F_{\bar{\Delta}_{1,1}}\|_{\bar{\Delta}_{1,1}}$ sehr groß sein und deshalb würde das neue Gitter extrem viele Gitterpunkte beinhalten. Um den daraus resultierenden großen Rechenaufwand zu vermeiden, wird die Anzahl der Punkte im verbesserten Basisgitter $\bar{\Delta}^m_{1,2}$ nach oben mit $5(N(\bar{\Delta}^m_{1,1}) + 1)$ beschränkt. Nach unten wird die Punkteanzahl von $\bar{\Delta}^m_{1,2}$ mit $2(N(\bar{\Delta}^m_{1,1}) + 1)$ beschränkt. Schließlich ergibt sich

$$N(\bar{\Delta}^m_{1,2}) = \left\lceil \frac{\max\Big(2(N(\bar{\Delta}^m_{1,1}) + 1), \min(N_{ref}, 5(N(\bar{\Delta}^m_{1,1}) + 1)\Big) - 1}{m+1} \right\rceil (m+1). \tag{4.15}$$

Die Beschränkungen von $N(\bar{\Delta}^m_{1,2})$ nach oben und nach unten haben sich in der Praxis gut bewährt, siehe §6.2.
Mit Hilfe der Formel (4.15) wird nun das neue äquidistante Gitter $\bar{\Delta}^m_{1,2}$ festgelegt.
Ist (4.12) nach der Auswertung auf $\bar{\Delta}^m_{1,2}$ nicht erfüllt, so muss weiter verfeinert werden. Man versucht daher, sich zu einem äquidistanten Gitter $\bar{\Delta}^m_{1,j}$ "hinzutasten" auf dem (4.12) gilt. Dazu werden sukzessive $\lceil \frac{N(\bar{\Delta}_{1,j})}{2} \rceil$ Intervalle hinzugefügt,

$$N(\bar{\Delta}_{1,j}) = N(\bar{\Delta}_{1,j-1}) + \left\lceil \frac{N(\bar{\Delta}_{1,j-1})}{2} \right\rceil, \quad j \geq 3. \tag{4.16}$$

Der erste Index j, für den (4.12) gilt, liefert den globalen Fehler $\varepsilon_{\bar{\Delta}^m_{1,j}}$, der für die Gleichverteilung herangezogen wird. Man schreibt dann

$$\bar{\Delta}^m_1 := \bar{\Delta}^m_{1,j}.$$

Bei dieser Vorgangsweise muss darauf geachtet werden, dass $N(\bar{\Delta}_{1,j})$ nicht zu groß wird. Daher fordert man

$$N(\bar{\Delta}_{1,j}) < K(m),$$

wobei $K(m)$ eine von m abhängige Konstante ist. In unserem Testcode ist $K(m) = \frac{10^4}{m+1}$, woraus $N(\bar{\Delta}^m_{1,j}) < 10^4$ folgt. Die Phase 1 der Gitterverfeinerungsstrategie wird in der Abbildung 4.7 dargestellt.

Beispiele

Das erste Modellproblem ist die Aufgabe (5.11) mit Eingabedaten

$$m{=}6,\ \text{aTOL}{=}\text{rTOL}{=}1\text{E}{-}5.$$

Dabei soll ein für glatte Lösungsstrukturen typisches Verhalten demonstriert werden, wo die Toleranz sofort auf dem Basisgitter nach (4.9) erfüllt wird, siehe Abbildung 4.8. Der in den Abbildungen auftretende Wert TOL_g wird nach (4.24) berechnet.
In den folgenden 2 Beispielen sind Basisgitterverfeinerungen notwendig.
Beispiel (5.10) mit den Eingabedaten

$$m{=}8,\ \text{aTOL}{=}\text{rTOL}{=}1\text{E}{-}8.$$

Zuerst wird das Basisgitter $\bar{\Delta}^8_{1,1}$ nach (4.9) berechnet. Die Auswertung und das Gitter sind in Abbildung 4.9 dargestellt. Das Gitter $\bar{\Delta}^8_{1,2}$ mit N_{ref} nach (4.14) und die Daten der Auswertung sind in der Abbildung 4.10 zu finden. Man sieht, dass die Auswertung auf diesem Gitter $\bar{\Delta}^8_{1,2} = \bar{\Delta}^8_1$ die geforderte Güte aufweist.
Beispiel (5.4) mit den Eingabedaten

$$m{=}4,\ \text{aTOL}{=}\text{rTOL}{=}1\text{E}{-}2.$$

Die Abbildungen 4.11, 4.12, 4.13 und 4.14 zeigen, wie das Basisgitter insgesamt drei mal verfeinert werden muss. Diese Gitter sind bereits aus der Abbildung 4.4 bekannt. Die Anzahl der Teilintervalle des Startgitters, $N(\bar{\Delta}_{1,1}) = 4$, wird zunächst etwa verfünffacht, $N(\bar{\Delta}_{1,2}) = 21$, und nachher zwei mal mit dem Faktor 1.5 vergrößert.
Abschließend wird in der Abbildung 4.15 ein Beispiel diskutiert, bei dem der sehr seltene Fall auftritt, dass die globale Prüfung der Qualität gemäß (4.12) keine gute Information für die Gleichverteilung liefert. In diesem Fall würde das Gitter $\bar{\Delta}_2$ in der Nähe des linken Intervallrandes zu fein ausfallen.

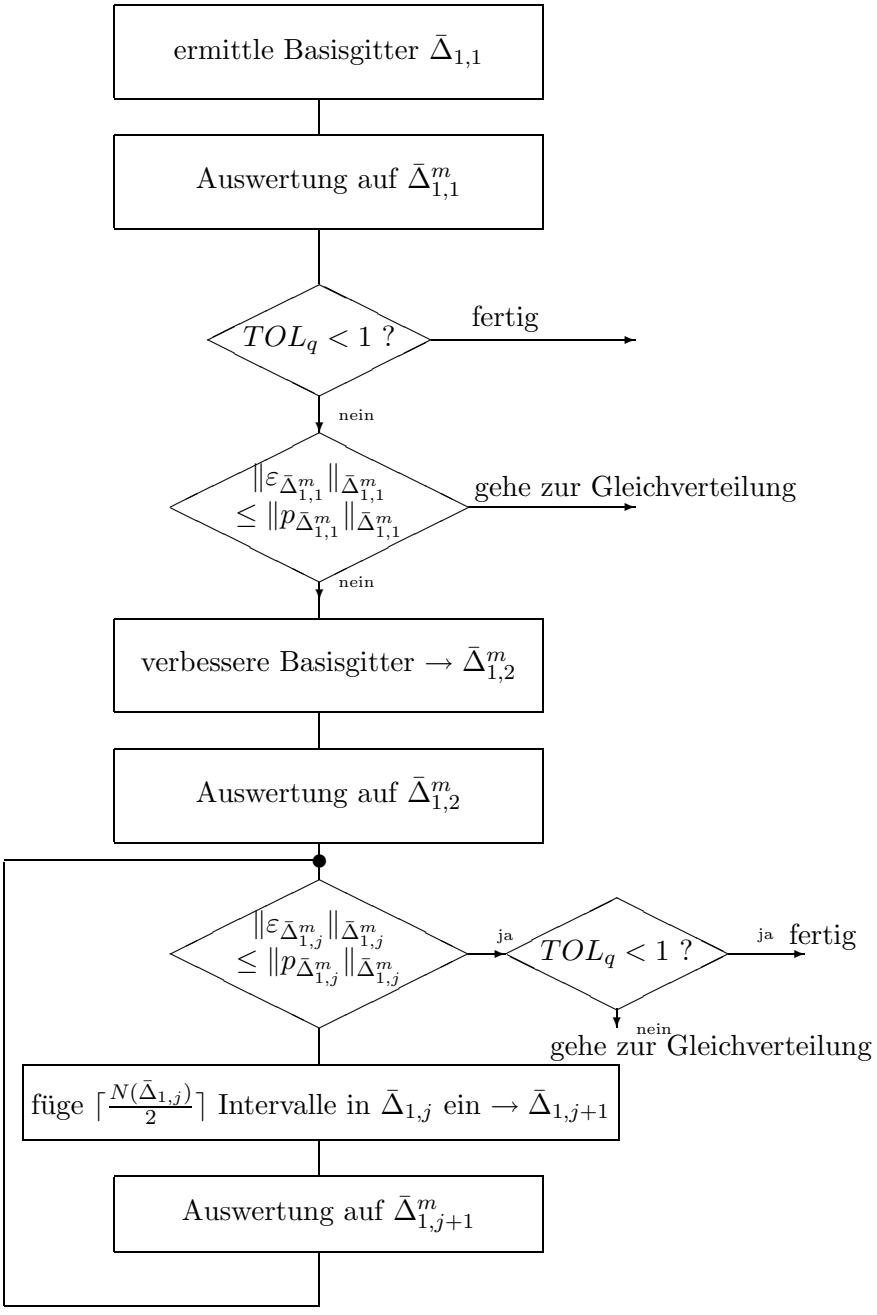
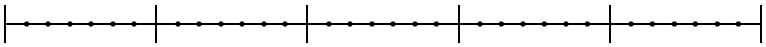
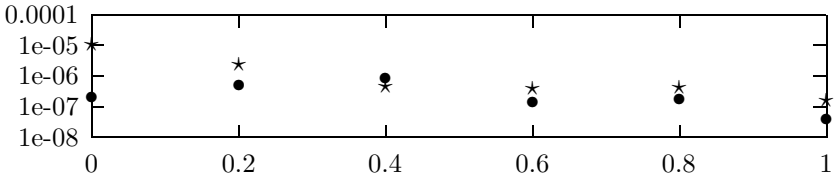


Abbildung 4.7: Schematischer Aufbau der Phase 1

Auswertung auf dem Gitter $\bar{\Delta}^6_{1,1}$:



$N(\bar{\Delta}_{1,1})+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
6.00E+000	2.00E-001	2.00E-001	1.00E+000

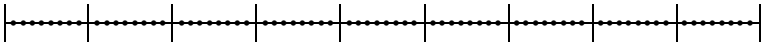


$\|\eta_{\bar{\Delta}^6_{1,1}}\|_{\bar{\Delta}^6_{1,1}}$	$\|\varepsilon_{\bar{\Delta}^6_{1,1}}\|_{\bar{\Delta}^6_{1,1}}$	TOL_q	TOL_g
1.09E-005	8.74E-007	7.74E-002	1.13E-005

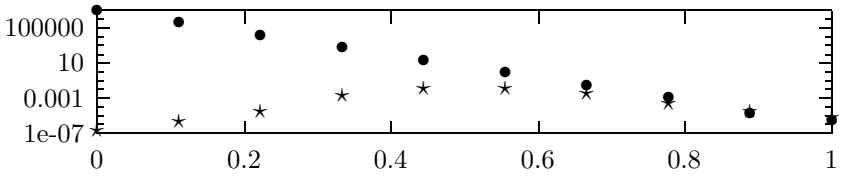
Da $TOL_q < 1$ gilt, ist die Toleranzforderung bereits nach der 1. Auswertung erfüllt.

Abbildung 4.8: Auswertung auf dem Basisgitter, Beispiel (5.11), m=6, aTOL=rTOL=1E−5

Auswertung auf dem Gitter $\bar{\Delta}^8_{1,1}$:



$N(\bar{\Delta}_{1,1})+1$	h_{min}	**h**	$\frac{\mathbf{h}}{h_{min}}$
1.00E+001	1.11E-001	1.11E-001	1.00E+000

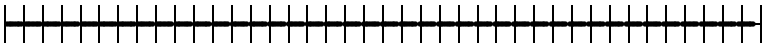


$\|\eta_{\bar{\Delta}^8_{1,1}}\|_{\bar{\Delta}^8_{1,1}}$	$\|\varepsilon_{\bar{\Delta}^8_{1,1}}\|_{\bar{\Delta}^8_{1,1}}$	TOL_q	TOL_g
1.58E-002	9.41E+006	9.41E+014	1.00E-008

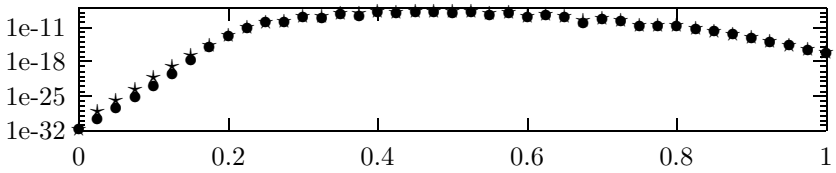
Die Qualität auf dem Gitter $\bar{\Delta}^8_{1,1}$ ist noch nicht ausreichend.

Abbildung 4.9: Auswertung auf dem Basisgitter, Beispiel (5.10), m=8, aTOL=rTOL=1E−8

Hier ist $N_{ref} = 360$, $\min(360, 410) = 360$ und wegen (4.15) gilt für die Anzahl der Gitterpunkte von $\bar{\Delta}_{1,2}$, $N(\bar{\Delta}_{1,2}) = \lceil \frac{\max(164,360)-1}{9} \rceil = 40$. Auswertung auf dem Gitter $\bar{\Delta}^8_{1,2}$:



$N(\bar{\Delta}_{1,2})+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
4.10E+001	2.50E-002	2.50E-002	1.00E+000

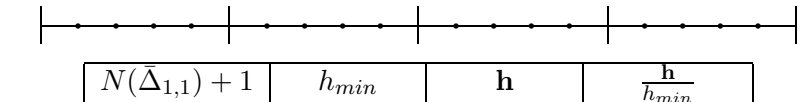


$\|\eta_{\bar{\Delta}^8_{1,2}}\|_{\bar{\Delta}^8_{1,2}}$	$\|\varepsilon_{\bar{\Delta}^8_{1,2}}\|_{\bar{\Delta}^8_{1,2}}$	TOL_q	TOL_g
2.40E-008	1.77E-008	7.00E-001	2.00E-008

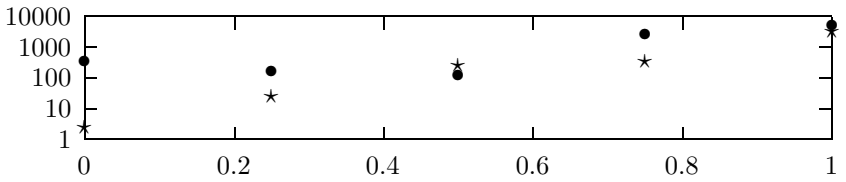
Die Qualität der Auswertung auf dem Gitter $\bar{\Delta}^8_{1,2}$ ist ausreichend.

Abbildung 4.10: Auswertung auf einem verfeinerten Basisgitter (unter Einbeziehung von N_{ref}), Beispiel (5.10), m=8, aTOL=rTOL=1E−8

Auswertung auf dem Gitter $\bar{\Delta}^4_{1,1}$:



$N(\bar{\Delta}_{1,1})+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
5.00E+000	2.50E-001	2.50E-001	1.00E+000

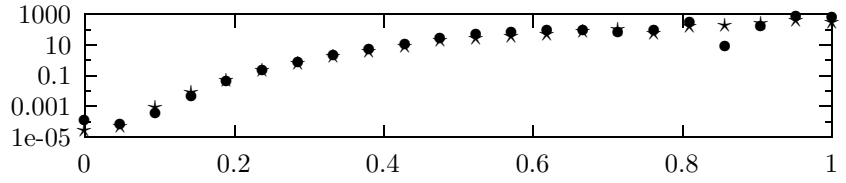


$\|\eta_{\bar{\Delta}^4_{1,1}}\|_{\bar{\Delta}^4_{1,1}}$	$\|\varepsilon_{\bar{\Delta}^4_{1,1}}\|_{\bar{\Delta}^4_{1,1}}$	TOL_q	TOL_g
3.27E+003	5.19E+003	1.01E+004	3.45E-002

Abbildung 4.11: Auswertung auf dem Startgitter, Beispiel (5.4), m=4, aTOL=rTOL=1E−2

Auswertung auf dem Gitter $\bar{\Delta}^4_{1,2}$:

$N(\bar{\Delta}_{1,2})+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
2.20E+001	4.76E-002	4.76E-002	1.00E+000

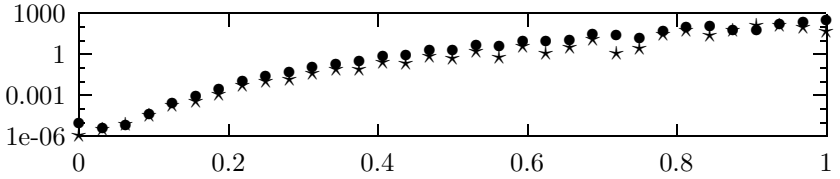


$\|\eta_{\bar{\Delta}^4_{1,2}}\|_{\bar{\Delta}^4_{1,2}}$	$\|\varepsilon_{\bar{\Delta}^4_{1,2}}\|_{\bar{\Delta}^4_{1,2}}$	TOL_q	TOL_g
4.89E+002	7.58E+002	1.01E+003	2.67E-001

Abbildung 4.12: Auswertung auf dem verfeinerten Basisgitter, Beispiel (5.4), m=4, aTOL=rTOL=1E−2

Durch das Einfügen von $\lceil \frac{N(\bar{\Delta}_{1,2})}{2} \rceil$ Intervallen ergibt sich $\bar{\Delta}_{1,3}$:

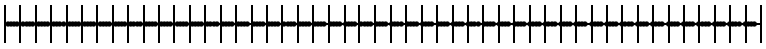
$N(\bar{\Delta}_{1,3})+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
3.30E+001	3.12E-002	3.12E-002	1.00E+000



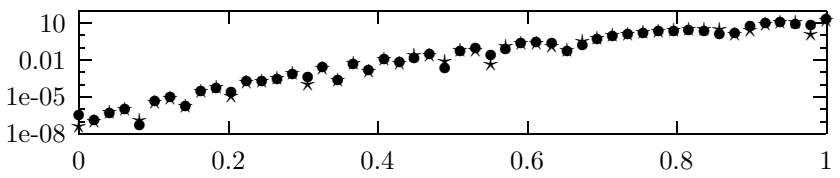
$\|\eta_{\bar{\Delta}^4_{1,3}}\|_{\bar{\Delta}^4_{1,3}}$	$\|\varepsilon_{\bar{\Delta}^4_{1,3}}\|_{\bar{\Delta}^4_{1,3}}$	TOL_q	TOL_g
1.50E+002	2.95E+002	4.21E+003	6.77E-002

Abbildung 4.13: Auswertung auf dem verfeinerten Basisgitter, Beispiel (5.4), m=4, aTOL=rTOL=1E−2

Durch das Einfügen von $\lceil \frac{N(\bar{\Delta}_{1,3})}{2} \rceil$ Intervallen ergibt sich $\bar{\Delta}_{1,4}$:



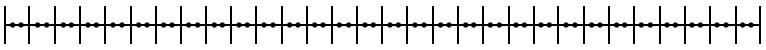
$N(\bar{\Delta}_{1,4})+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
5.00E+001	2.04E-002	2.04E-002	1.00E+000



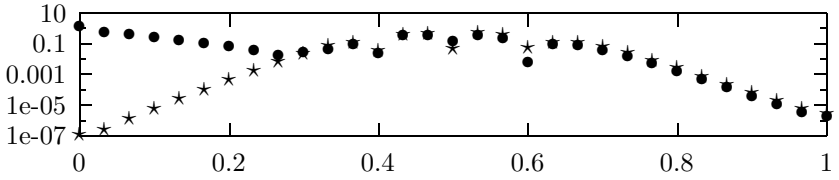
$\|\eta_{\bar{\Delta}^4_{1,4}}\|_{\bar{\Delta}^4_{1,4}}$	$\|\varepsilon_{\bar{\Delta}^4_{1,4}}\|_{\bar{\Delta}^4_{1,4}}$	TOL_q	TOL_g
1.72E+001	2.33E+001	4.08E+001	1.76E-001

Abbildung 4.14: Auswertung auf dem verfeinerten Basisgitter, Beispiel (5.4), m=4, aTOL=rTOL=1E−2

Auswertung auf dem Gitter $\bar{\Delta}^2_{1,1}$:



$N(\bar{\Delta}_{1,1})+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
3.10E+001	3.33E-002	3.33E-002	1.00E+000



$\|\eta_{\bar{\Delta}^2_{1,1}}\|_{\bar{\Delta}^2_{1,1}}$	$\|\varepsilon_{\bar{\Delta}^2_{1,1}}\|_{\bar{\Delta}^2_{1,1}}$	TOL_q	TOL_g
5.39E-001	1.34E+000	1.34E+003	1.00E-003

Abbildung 4.15: Auswertung auf dem Startgitter, Beispiel (5.10), m=2, aTOL=rTOL=1E−3

4.2.4 Verteilung der Gitterpunkte

Ist die Phase 1 erfolgreich abgeschlossen, ohne dass die Toleranzforderung erfüllt ist, so erfüllt $\varepsilon_{\bar{\Delta}_1^m}$ die Bedingung (4.12). Wie in [3] und in §4.1 beschrieben, wird diese Schätzung des globalen Fehlers für die Gleichverteilung herangezogen, wobei zunächst die Information aus $\varepsilon_{\bar{\Delta}_1^m}$ in eine Monitorfunktion übergeleitet wird (siehe [29]). Dazu wird

$$\Theta_i := \Theta(\varepsilon_i) = |\varepsilon_i|^{\frac{1}{m}}, \quad i = 0, \ldots, N(\bar{\Delta}_1^m),$$

berechnet.
Die numerischen Ergebnisse in §6.4 legen nahe, die Variation von $\Theta_{\bar{\Delta}_1^m}$ in einem gewissen Ausmaß durch Glättung zu beschränken, ohne dass der wesentliche Informationsgehalt von $\Theta_{\bar{\Delta}_1^m}$ verloren geht. Dazu wird ein $s_a \in \{0, \ldots, N(\bar{\Delta}_1^m)\}$ gewählt und der geglättete Gittervektor $\bar{\Theta}_i$ wie folgt errechnet:

$$\bar{\Theta}_i = \Psi(\Theta(\varepsilon_i), s_a) := \tag{4.17}$$
$$\frac{1}{\min(N(\bar{\Delta}_1^m)+1, i+s_a) - \max(0, i-s_a) + 1} \sum_{k=\max(0,i-s_a)}^{\min(N(\bar{\Delta}_1^m)+1, i+s_a)} \Theta_k.$$

Es wird also jedes Θ_i über seine linken und rechten s_a Nachbarn gemittelt[8]. Da die Anzahl der Intervalle $N(\bar{\Delta}_1^m)$ variiert, wird der Glättungsfaktor `SMOOTHING_FACTOR` relativ angegeben. Das bedeutet, dass der Benutzer den Parameterwert `SMOOTHING_FACTOR` in Prozent vorzuschreiben hat. Die numerischen Experimente in §6.4 lassen erkennen, dass keine Glättung ($s_a = 0$) schlechte Ergebnisse liefert. Daher wird die Anzahl der Nachbarn, die einzubeziehen sind, mit 2 nach unten beschränkt. Der Wert s_a wird gemäß

$$s_a = \min\left(2, \left\lfloor (N(\bar{\Delta}_1^m)+1) \cdot \frac{\texttt{SMOOTHING\_FACTOR}}{100} \right\rfloor\right) \tag{4.18}$$

berechnet. Aufgrund der numerischen Erfahrungen wurde `SMOOTHING_FACTOR`=5% als Standardwert gewählt.
Beim Glätten verliert man möglicherweise die Information über große Werte von $\Theta_{\bar{\Delta}_1^m}$. Diesem Verlust wird mit Hilfe des Parameters `RECOVER_PEAKS` entgegengesteuert. Ist der Parameterwert `RECOVER_PEAKS` gleich 1 gesetzt, so definieren wir

$$\tilde{\Theta}_i := \max(\Theta_i, \bar{\Theta}_i), \quad i = 0, \ldots, N(\bar{\Delta}_1^m).$$

[8]Man beachte, dass die Glättung nicht von der Verteilung der Gitterpunkte abhängt.

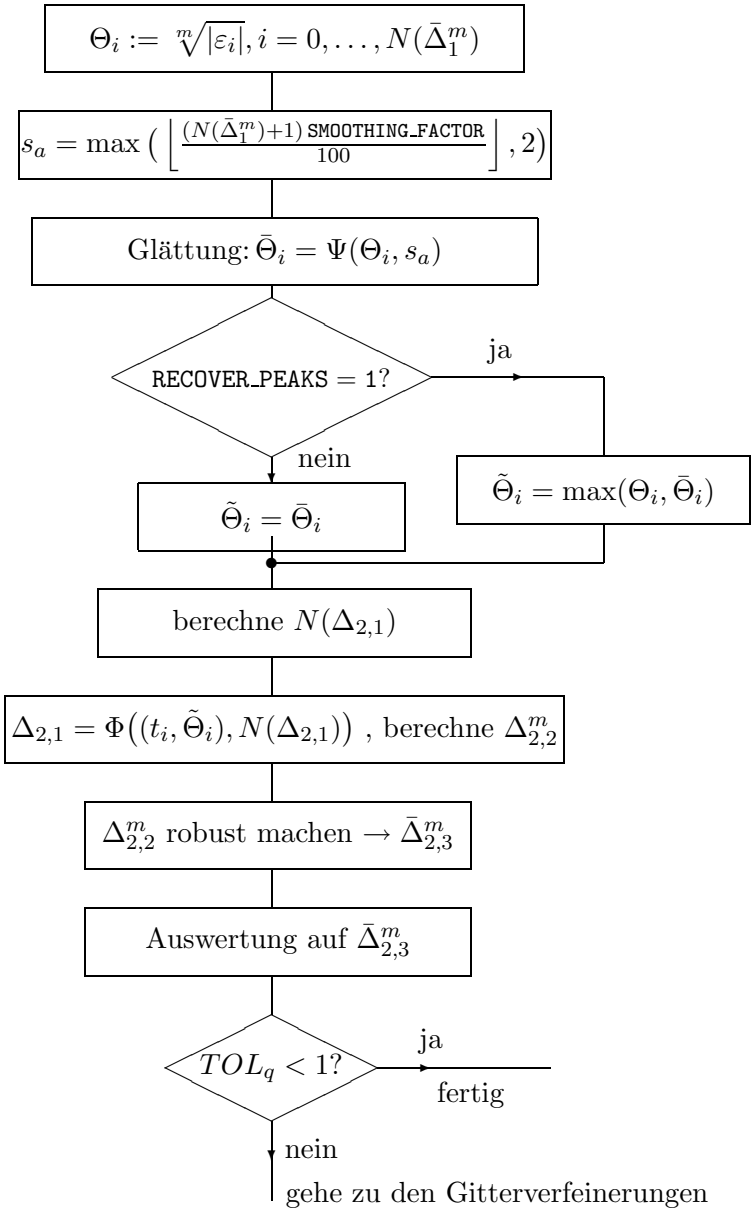


Abbildung 4.16: Schematischer Aufbau der Phase 2

Wenn RECOVER_PEAKS=0 gilt, so setzen wir

$$\tilde{\Theta}_i := \bar{\Theta}_i, \quad i = 0, \ldots, N(\bar{\Delta}_1^m).$$

Das bedeutet, dass im ersten Fall die Ausreisser nach oben voll berücksichtigt werden, während im zweiten Fall die geglätteten Werte benutzt werden.

Beispiele

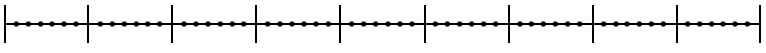
Die Auswirkungen der beiden neu eingeführten Parameter SMOOTHING_FACTOR und RECOVER_PEAKS werden an Beispielen demonstriert. Zuerst wird das Beispiel (5.8) mit den Eingabedaten

$$m{=}6,\ \text{aTOL}{=}\text{rTOL}{=}1\text{E}{-}6,\ \texttt{SMOOTHING\_FACTOR}{=}5\%$$

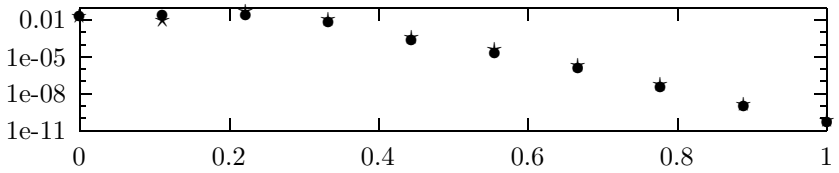
getestet, wobei die Werte des Parameters RECOVER_PEAKS variieren. In der Abbildung 4.17 ist die Auswertung auf dem Grundgitter $\bar{\Delta}_1^6$ dokumentiert. Die Daten dieser Auswertung werden für die Gleichverteilung benutzt, wobei RECOVER_PEAKS gleich 1 bzw. 0 gesetzt wird.
Die Abbildungen 4.18 und 4.19 beziehen sich auf die Wahl RECOVER_PEAKS=1. Das sich nach der Gleichverteilung ergebende Gitter[9] $\Delta_{2,1}^6$ ist in der Abbildung 4.18 dargestellt. Wir bezeichnen dieses Gitter mit Hilfe des hochgestellten Symbols m, $\Delta_{2,1}^m$, um hervorzuheben, dass es sich hier um verschobene Gitterpunkte und Kollokationspunkte handelt, die alle in der Zeichnung angedeutet werden. Diese Notation stimmt nicht ganz mit der Definition (1.2) überein, wo stückweise äquidistante Lage der Kollokationspunkte angenommen wird. Das für die Auswertung taugliche stückweise äquidistante Gitter $\bar{\Delta}_2^6$ und die dazugehörige Auswertung ist in Abbildung 4.19 dokumentiert. Die geforderte Toleranz wird hier sofort erreicht.

[9] Später werden wir genau erklären, wie die Konstruktion von $\Delta_{2,1}^m$ und $\bar{\Delta}_2^m$ durchgeführt wird.

Auswertung auf dem Gitter $\bar{\Delta}^6_{1,1}$:



$N(\bar{\Delta}_{1,1})+1$	h_{min}	**h**	$\frac{\mathbf{h}}{h_{min}}$
1.00E+001	1.11E-001	1.11E-001	1.00E+000



$\|\eta_{\bar{\Delta}^6_{1,1}}\|_{\bar{\Delta}^6_{1,1}}$	$\|\varepsilon_{\bar{\Delta}^6_{1,1}}\|_{\bar{\Delta}^6_{1,1}}$	TOL_q	TOL_g
5.32E-002	2.69E-002	2.11E+004	1.02E-006

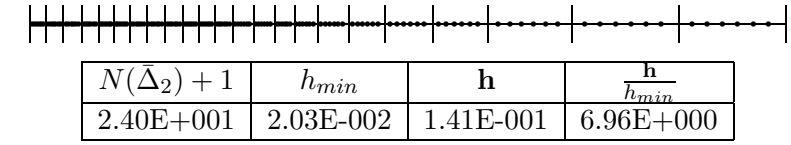
Abbildung 4.17: Basisgitter für Beispiel (5.8)

Das Gitter $\Delta^6_{2,1}$:

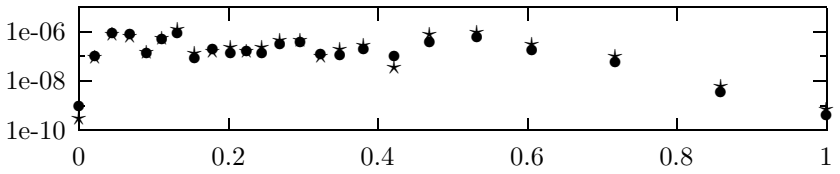
$N(\Delta_{2,1})+1$	h_{min}	**h**	$\frac{\mathbf{h}}{h_{min}}$
2.30E+001	2.03E-002	2.83E-001	1.39E+001

Abbildung 4.18: Verteiltes Gitter mit `RECOVER_PEAKS`=1 für Beispiel (5.8)

Auswertung auf dem Gitter $\bar{\Delta}_2^6$:



$N(\bar{\Delta}_2)+1$	h_{min}	**h**	$\frac{\mathbf{h}}{h_{min}}$
2.40E+001	2.03E-002	1.41E-001	6.96E+000



$\|\eta_{\bar{\Delta}_2^6}\|_{\bar{\Delta}_2^6}$	$\|\varepsilon_{\bar{\Delta}_2^6}\|_{\bar{\Delta}_2^6}$	TOL_q	TOL_g
1.40E-006	1.09E-006	8.31E-001	1.00E-006

Auf dem Gitter $\bar{\Delta}_2^6$ wird die Toleranzanforderung sofort erfüllt.

Abbildung 4.19: Auswertung auf dem verteilten Gitter mit `RECOVER_PEAKS`=1 für Beispiel (5.8)

Das Gitters $\Delta_{2,1}^6$:

$N(\Delta_{2,1})+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
2.10E+001	2.43E-002	2.92E-001	1.20E+001

Abbildung 4.20: Verteiltes Gitter mit RECOVER_PEAKS=0 für Beispiel (5.8)

Die Abbildungen 4.20 und 4.21 beziehen sich auf das selbe Beispiel mit RECOVER_PEAKS=0. In Abbildung 4.21 ist zu erkennen, dass die Auswertung auf dem Gitter $\bar{\Delta}_2^6$ die geforderte Toleranz noch nicht erreicht hat und deshalb weitere Gitterverfeinerung notwendig wäre.

Um die Wirkung des Parameters SMOOTHING_FACTOR zu demonstrieren, wird Beispiel (5.3) mit den Eingabedaten

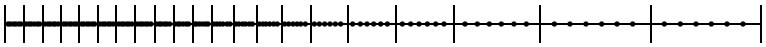
$$m=2,\ \text{aTOL}=\text{rTOL}=1\text{E}{-3},\ \texttt{RECOVER\_PEAKS}=1$$

getestet. In Abbildung 4.22 ist die Auswertung auf dem verfeinerten Basisgitter $\bar{\Delta}_{1,2}^2 = \bar{\Delta}_1^2$ zu sehen. Mit diesen Auswertungsdaten wird das Gitter $\Delta_{2,1}$ mit SMOOTHING_FACTOR=5% und mit SMOOTHING_FACTOR=0% erzeugt. Für SMOOTHING_FACTOR=5% ist das Gitter $\Delta_{2,1}^2$ in Abbildung 4.23 dargestellt, das Gitter $\bar{\Delta}_2^2$ und die zugehörige Auswertung ist in Abbildung 4.24 zu sehen.

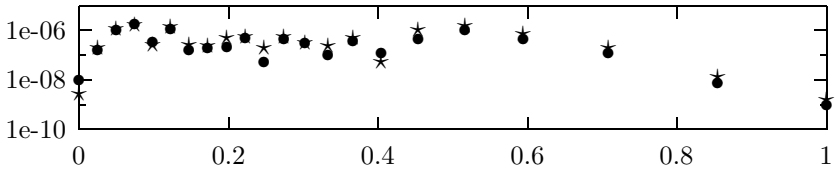
Für SMOOTHING_FACTOR=0% ist das Gitter $\Delta_{2,1}^2$ in Abbildung 4.25 dargestellt, das Gitter $\bar{\Delta}_2^2$ und die zugehörige Auswertung ist in Abbildung 4.26 dokumentiert.

Vergleicht man die Abbildungen 4.23 und 4.25, so sieht man, dass die Gitter $\Delta_{2,1}^2$ dem Fehlerverlauf, der in Abbildung 4.22 zu sehen ist, gut anpasst sind. Allerdings weist das Gitter ohne Glättung (SMOOTHING_FACTOR=0%) eine größere Variation der Schrittweiten auf. Dieses Gitter "pulsiert" stärker als das Gitter zu den geglätteten Daten (SMOOTHING_FACTOR=5%).

Auswertung auf dem Gitter $\bar{\Delta}_2^6$:



$N(\bar{\Delta}_2)+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
2.20E+001	2.43E-002	1.46E-001	5.99E+000



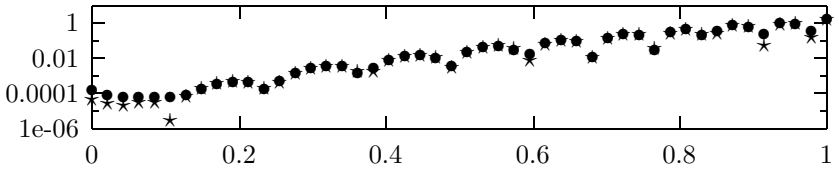
$\|\eta_{\bar{\Delta}_2^6}\|_{\bar{\Delta}_2^6}$	$\|\varepsilon_{\bar{\Delta}_2^6}\|_{\bar{\Delta}_2^6}$	TOL_q	TOL_g
1.94E-006	1.94E-006	1.78E+000	1.03E-006

Die Toleranz wurde nicht erreicht, es sind noch weitere Gitterverfeinerungen notwendig.

Abbildung 4.21: Auswertung auf dem verteilten Gitter mit `RECOVER_PEAKS`=0 für Beispiel (5.8)

Auswertung auf dem Gitter $\bar{\Delta}^2_{1,2} = \bar{\Delta}^2_1$:

$N(\bar{\Delta}_{1,2})+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
4.80E+001	2.13E-002	2.13E-002	1.00E+000



$\|\eta_{\bar{\Delta}^2_{1,2}}\|_{\bar{\Delta}^2_{1,2}}$	$\|\varepsilon_{\bar{\Delta}^2_{1,2}}\|_{\bar{\Delta}^2_{1,2}}$	TOL_q	TOL_g
1.59E+000	1.68E+000	1.95E+001	1.98E-002

Abbildung 4.22: Auswertung auf dem Gitter $\bar{\Delta}^2_1$ für Beispiel (5.3)

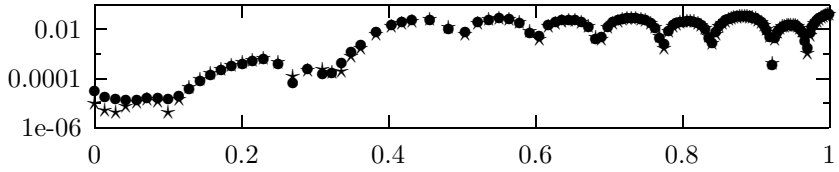
Das Gitter $\Delta^2_{2,1}$ ergibt sich nach der Glättung mit $s_a = 7$:

$N(\Delta_{2,1})+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
9.70E+001	2.60E-003	2.30E-001	8.84E+001

Abbildung 4.23: Das Gitter $\Delta^2_{2,1}$ mit `SMOOTHING_FACTOR`=5% für Beispiel (5.3)

Das Gitter $\bar{\Delta}_2^2$:

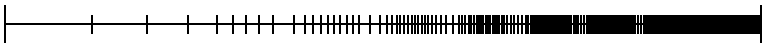
$N(\bar{\Delta}_2)+1$	h_{min}	**h**	$\frac{\mathbf{h}}{h_{min}}$
1.20E+002	2.60E-003	2.54E-002	9.78E+000



$\|\eta_{\bar{\Delta}_2^2}\|_{\bar{\Delta}_2^2}$	$\|\varepsilon_{\bar{\Delta}_2^2}\|_{\bar{\Delta}_2^2}$	TOL_q	TOL_g
4.31E-002	4.43E-002	2.25E+000	1.21E-002

Abbildung 4.24: Auswertung auf dem Gitter $\bar{\Delta}_2^2$ mit `SMOOTHING_FACTOR`=5% für Beispiel (5.3)

Das Gitter $\Delta_{2,1}^2$ ergibt sich nach der Glättung mit $s_a = 0$:

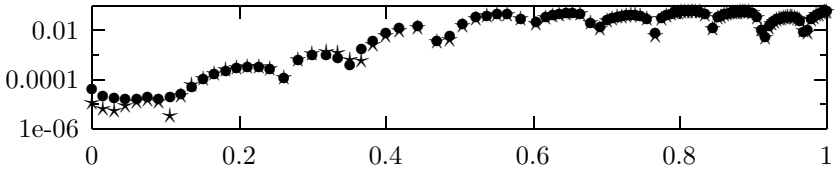


$N(\Delta_{2,1})+1$	h_{min}	**h**	$\frac{\mathbf{h}}{h_{min}}$
8.50E+001	2.60E-003	2.42E-001	9.30E+001

Abbildung 4.25: Das Gitter $\Delta_{2,1}^2$ mit `SMOOTHING_FACTOR`=0% für Beispiel (5.3)

Das Gitter $\bar{\Delta}_2^2$:

$N(\bar{\Delta}_2)+1$	h_{min}	**h**	$\frac{\mathbf{h}}{h_{min}}$
1.08E+002	2.60E-003	2.60E-002	9.97E+000



$\|\eta_{\bar{\Delta}_2^2}\|_{\bar{\Delta}_2^2}$	$\|\varepsilon_{\bar{\Delta}_2^2}\|_{\bar{\Delta}_2^2}$	TOL_q	TOL_g
6.35E-002	6.42E-002	3.68E+000	1.32E-002

Abbildung 4.26: Auswertung auf dem Gitter $\bar{\Delta}_2^2$ mit `SMOOTHING_FACTOR`=0% für Beispiel (5.3)

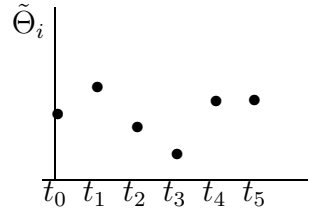


Abbildung 4.27: Punktmenge $\tilde{\Theta}_i$

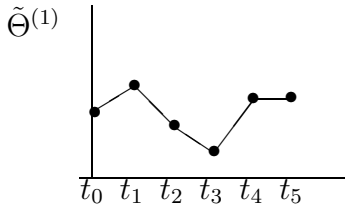


Abbildung 4.28: Funktion $\tilde{\Theta}^{(1)}$ ist die stückweise lineare Interpolierende der Punktmenge $\tilde{\Theta}_i$

Um Gleichverteilung der Monitorfunktion durchführen zu können wird eine Approximation für das Integral

$$\int_0^1 \tilde{\Theta}(\tau)d\tau \tag{4.19}$$

benötigt, wobei von der Funktion $\tilde{\Theta}$ nur die diskreten Werte $\tilde{\Theta}_i =: \tilde{\Theta}(t_i)$ bekannt sind, vgl. Abbildung 4.27. Wir approximieren die analytische Funktion $\tilde{\Theta}$, indem wir uns die Werte $\tilde{\Theta}_i$ durch eine stetige stückweise lineare Funktion $\tilde{\Theta}^{(1)}$ interpoliert denken, siehe Abbildung 4.28.
Weiters definieren wir die Funktion $\tilde{\Theta}^{(2)}$ wie folgt:

$$\tilde{\Theta}^{(2)}(t) := \begin{cases} \tilde{\Theta}_j & : \quad t = t_j, \quad j = 0 \ldots \Delta_1^m, \\ \frac{\tilde{\Theta}_j + \tilde{\Theta}_{j-1}}{2} & : \quad \text{sonst}, \quad j = 1 \ldots \Delta_1^m. \end{cases} \tag{4.20}$$

Eine schematische Darstellung dieser Funktion ist in Abbildung 4.29 zu sehen. Es gilt

$$I := \int_0^1 \tilde{\Theta}^{(1)}(\tau)d\tau = \int_0^1 \tilde{\Theta}^{(2)}(\tau)d\tau \tag{4.21}$$

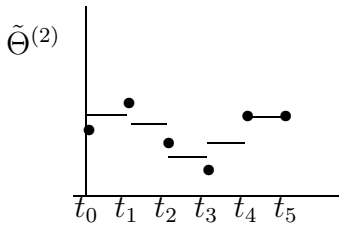


Abbildung 4.29: Schematische Darstellung von $\tilde{\Theta}^{(2)}$

und

$$I = \sum_{k=0}^{N(\bar{\Delta}_1^m)-1} \frac{\tilde{\Theta}_{k+1} + \tilde{\Theta}_k}{2}(t_{k+1} - t_k)$$

$$= h \cdot \Big(\sum_{k=1}^{N(\bar{\Delta}_1^m)-1} \frac{\tilde{\Theta}_k}{2} + \frac{\tilde{\Theta}_{N(\bar{\Delta}_1^m)}}{2} + \sum_{k=1}^{N(\bar{\Delta}_1^m)-1} \frac{\tilde{\Theta}_k}{2} + \frac{\tilde{\Theta}_0}{2} \Big)$$

$$= h \Big[\frac{\tilde{\Theta}_0}{2} + \frac{\tilde{\Theta}_{N(\bar{\Delta}_1^m)}}{2} + \sum_{k=1}^{N(\bar{\Delta}_1^m)-1} \tilde{\Theta}_k \Big]. \tag{4.22}$$

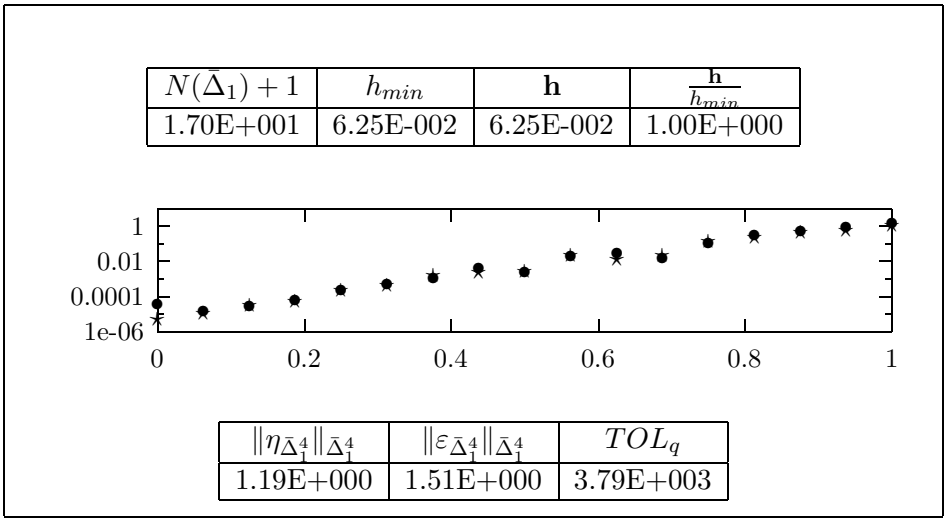
Dies bedeutet, dass wir das Integral (4.19) mit Hilfe der *Trapezregel* approximieren.
Anschliessend wird ein neues Gitter $\Delta_{2,1}^m = (\bar{t}_0, \ldots, \bar{t}_{\bar{N}-1})$ berechnet, wobei man zuerst die Anzahl der Gitterpunkte $\bar{N}$ festlegen muss. Dabei soll möglichst gewährleistet sein, dass auf diesem neuen Gitter die Toleranz sofort erfüllt wird. Diese Überlegung motiviert die folgende Wahl:

$$\bar{N} \geq (N(\bar{\Delta}_1^m) + 1) \frac{I}{\sqrt[m]{TOL_g}}. \tag{4.23}$$

Man sieht, dass mit wachsendem Wert von I, der die Schwierigkeit des Problems beschreibt, auch $\bar{N}$ wächst, d.h. die Gitter feiner werden.
Hier ist mit TOL_g ein *globaler* Toleranzparameter gemeint. (Die Toleranzabfrage bezog sich bis jetzt auf die Überprüfung der lokalen Güte der Näherungslösung.) Anhand des Beispiels (5.3) wurden drei Varianten der Wahl dieses globalen Toleranzparameters getestet. Die Eingabedaten waren:

$$m{=}4,\ \text{aTOL}{=}\text{rTOL}{=}1\text{E}{-}5,$$
$$\texttt{SMOOTHING\_FACTOR}{=}5\%,\ \texttt{RECOVER\_PEAKS}{=}1.$$

Das Ausgangsgitter für alle Varianten war das Gitter $\bar{\Delta}_1^4$, das am Ende der Phase 1 zur Verfügung steht, vgl. 4.30.



$N(\bar{\Delta}_1)+1$	h_{min}	**h**	$\frac{\mathbf{h}}{h_{min}}$
1.70E+001	6.25E-002	6.25E-002	1.00E+000

$\|\eta_{\bar{\Delta}_1^4}\|_{\bar{\Delta}_1^4}$	$\|\varepsilon_{\bar{\Delta}_1^4}\|_{\bar{\Delta}_1^4}$	TOL_q
1.19E+000	1.51E+000	3.79E+003

Abbildung 4.30: Das Ausgangsgitter $\bar{\Delta}_1^4$ vor der Gleichverteilung

- "Konservative Variante": Hier ist

$$TOL_g := aTOL.$$

Falls rTOL $\neq 0$ gilt, liefert diese Variante den größten Wert von $\bar{N}$. Daraus folgt mit TOL_g=1E−5

$$I = 0.33508,\ (N(\bar{\Delta}_1^4)+1)\frac{I}{\sqrt[m]{TOL_g}} = 482.65 \ \Rightarrow\ \bar{N} = 486.$$

Die Daten des Gitters $\bar{\Delta}_2^4$ und der dazugehörigen Auswertung sind der Abbildung 4.31 zu entnehmen.

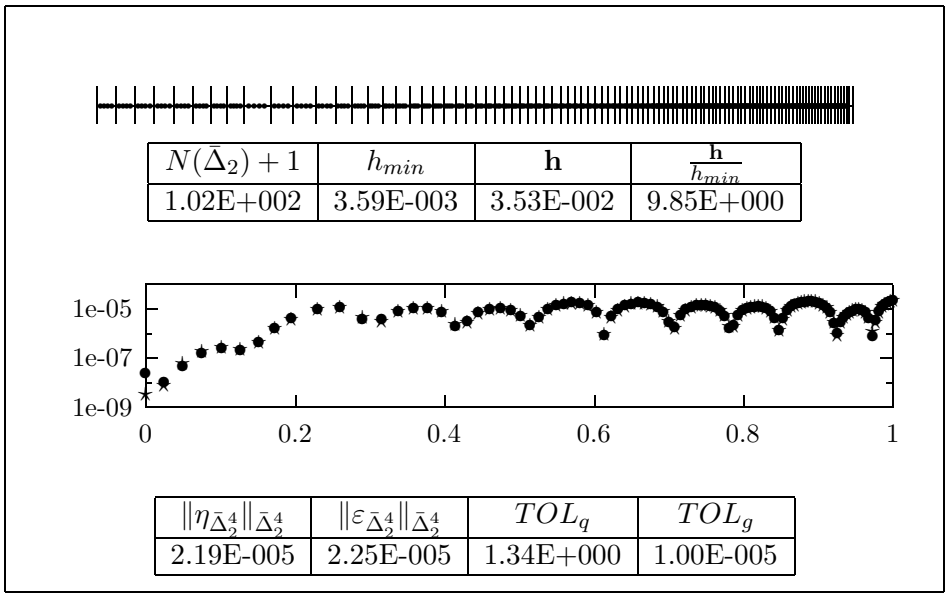
- "Grobe Variante": Bei dieser Variante werden sowohl der absolute als auch der relative Toleranzparameter berücksichtigt. Die Größe der Näherungslösung geht in einer globalen Weise über $\|p_{\bar{\Delta}_1^4}\|_{\bar{\Delta}_1^4}$ ein. Wir setzen

$$TOL_g := aTOL + \|p_{\bar{\Delta}_1^4}\|_{\Delta_1^4}\, rTOL.$$

Folglich ist mit TOL_g=4.91E−4

$$I = 0.33508,\ (N(\bar{\Delta}_1^4)+1)\frac{I}{\sqrt[m]{TOL_g}} = 182.3244 \ \Rightarrow\ \bar{N} = 186.$$

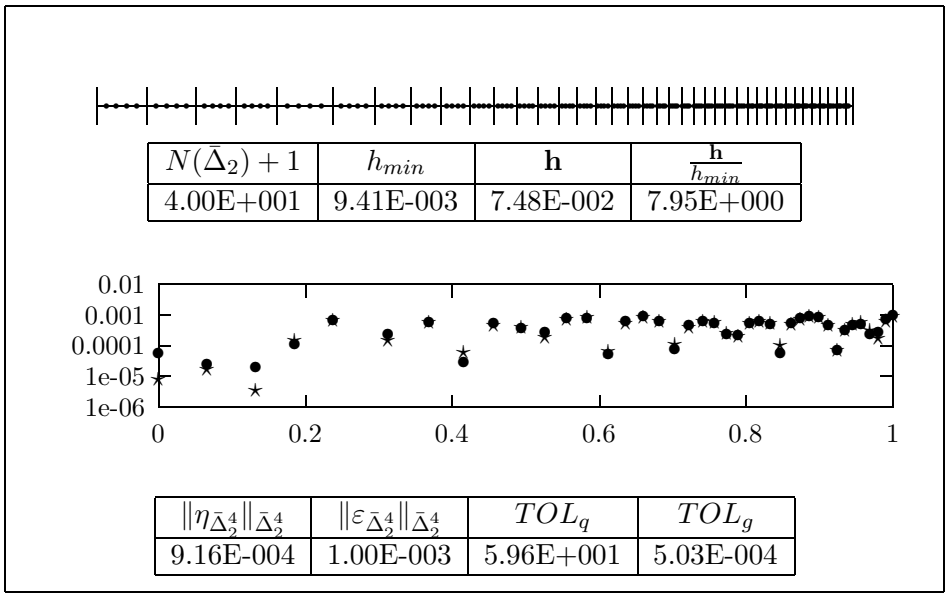
Die Daten des Gitters $\bar{\Delta}_2^4$ und der dazugehörigen Auswertung sind in der Abbildung 4.32 zusammengefasst.



$N(\bar{\Delta}_2)+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
1.02E+002	3.59E-003	3.53E-002	9.85E+000

$\|\eta_{\bar{\Delta}_2^4}\|_{\bar{\Delta}_2^4}$	$\|\varepsilon_{\bar{\Delta}_2^4}\|_{\bar{\Delta}_2^4}$	TOL_q	TOL_g
2.19E-005	2.25E-005	1.34E+000	1.00E-005

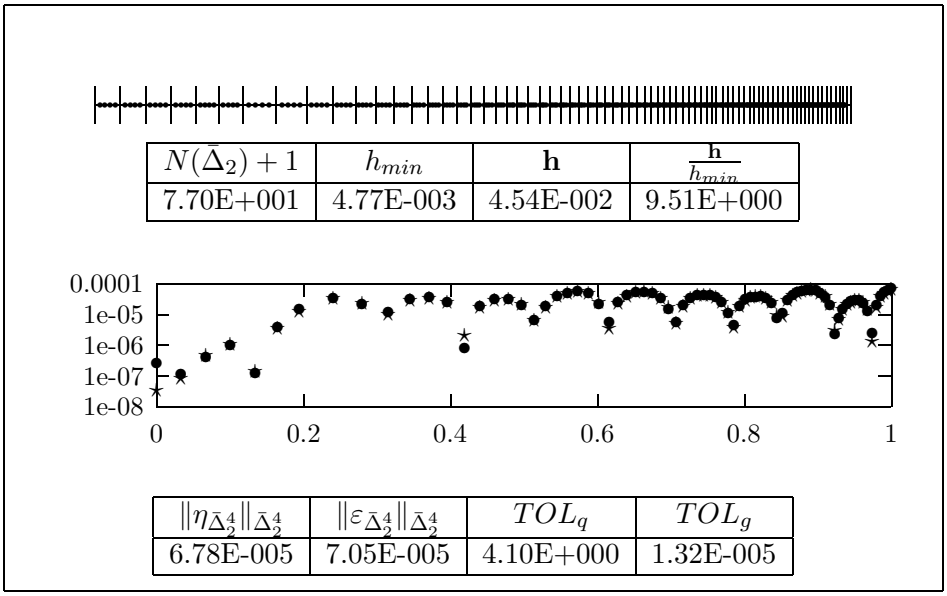
Abbildung 4.31: Das Gitter $\bar{\Delta}_2^4$, das mit der konservativen Variante berechnet wurde



$N(\bar{\Delta}_2)+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
4.00E+001	9.41E-003	7.48E-002	7.95E+000

$\|\eta_{\bar{\Delta}_2^4}\|_{\bar{\Delta}_2^4}$	$\|\varepsilon_{\bar{\Delta}_2^4}\|_{\bar{\Delta}_2^4}$	TOL_q	TOL_g
9.16E-004	1.00E-003	5.96E+001	5.03E-004

Abbildung 4.32: Das Gitter $\bar{\Delta}_2^4$, das mit der groben Variante berechnet wurde



$N(\bar{\Delta}_2)+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
7.70E+001	4.77E-003	4.54E-002	9.51E+000

$\|\eta_{\bar{\Delta}_2^4}\|_{\bar{\Delta}_2^4}$	$\|\varepsilon_{\bar{\Delta}_2^4}\|_{\bar{\Delta}_2^4}$	TOL_q	TOL_g
6.78E-005	7.05E-005	4.10E+000	1.32E-005

Abbildung 4.33: Das Gitter $\bar{\Delta}_2^4$ das mit der Kompromissvariante berechnet wurde

- "Kompromißvariante":

  Bei dieser Variante wird ein Wert für $\bar{N}$, der zwischen den Werten der vorhergehenden Varianten liegt, errechnet,

  $$TOL_g := aTOL + |p_j|\, rTOL, \tag{4.24}$$

  wobei der Index j jener Index ist für den

  $$TOL_q = \max_{i=0,\ldots,N(\bar{\Delta}_1^m)} \frac{|\varepsilon_i|}{aTOL + rTOL|p_i|} = \frac{|\varepsilon_j|}{aTOL + rTOL|p_j|} \tag{4.25}$$

  gilt.

  Mit TOL_g=3.15E−5 ergibt sich

  $$I = 0.33508,\ (N(\bar{\Delta}_1^4)+1)\frac{I}{\sqrt[m]{TOL_g}} = 362.3761 \ \Rightarrow\ \bar{N} = 366.$$

  Die Daten des Gitters $\bar{\Delta}_2^4$ und der dazugehörigen Auswertung sind in der Abbildung 4.33 angeführt.

Wir haben die oben beschriebenen Varianten verglichen und die Kompromißvariante als die Standardrealisierung gewählt. Die Auswertung auf dem

groben Gitter (grobe Variante) ist zwar sehr billig, oft jedoch ist $\bar{N}$ zu klein, um die Forderung (4.7) sofort zu erfüllen.
Die Auswertung auf dem feinen Gitter (konservative Variante) hat den schwerwiegenden Nachteil, dass bei schwierigen Modellen der Gesamtaufwand sehr groß werden kann, da eventuelle nachfolgende Gitterverfeinerungen sehr teuer werden.
In der Kompromißvariante findet man einen guten Mittelweg zwischen den beiden anderen Vorgangsweisen. In vielen Fällen wird die geforderte Toleranz sofort unterschritten, bei weiteren Auswertungen sind diese in den meisten Fällen nicht zu teuer.
Mit (4.23) und TOL_g nach (4.24) wird $\bar{N}$ so festgelegt, dass das daraus resultierende Gitter für die Kollokation geeignet ist,

$$N(\Delta_{2,1}^m) + 1 = \bar{N} :=$$
$$\max\left(N(\bar{\Delta}_1^m) + 1, \left\lceil \frac{(N(\bar{\Delta}_1^m)+1)\frac{I}{\sqrt[m]{TOL_g}}-1}{m+1} \right\rceil (m+1) + 1\right). \tag{4.26}$$

Die Anzahl $\bar{N}$ wird nach unten mit $N(\bar{\Delta}_1^m) + 1$ beschränkt.
Nun läßt sich die zentrale Formel für die Gleichverteilung formulieren[10]: Bei gegebenem $\bar{t}_{j-1}$ wähle $\bar{t}_j$ so, dass

$$\int_{\bar{t}_{j-1}}^{\bar{t}_j} \tilde{\Theta}^{(2)}(\tau)\, d\tau = \frac{I}{\bar{N}-1}, \quad j = 1, \ldots, \bar{N}-1 \tag{4.27}$$

gilt. Die Forderung (4.27) bedeutet, dass der Anteil des Integrals von $\tilde{\Theta}^{(2)}$ zwischen zwei benachbarten Gitterpunkten konstant sein soll. Die Eindeutigkeit eines solchen Gitters ist gewährleistet, weil der Integrand positiv ist und $I \neq 0$ gilt.
Die Eingabedaten des Gleichverteilungsalgorithmus Φ sind die Stützpunkte $(t_i, \tilde{\Theta}_i)$ und die Anzahl der Intervalle des neuen Gitters $\bar{N}$. Das Resultat des Gleichverteilungsalgorithmus ist das nach (4.27) gleichverteilte Gitter. Die Gitteranpassung wird schematisch wie folgt zusammengefasst:

$$\Delta_{2,1} = \Phi\big((t_i, \tilde{\Theta}_i), \bar{N}\big),$$
$$\big((t_i, \tilde{\Theta}_i), \bar{N}\big) \overset{(4.20)}{\longrightarrow} (\tilde{\Theta}^{(2)}(t), \bar{N}) \overset{(4.22),(4.27)}{\longrightarrow} (\bar{t}_0, \ldots, \bar{t}_{\bar{N}-1}) =: \Delta_{2,1}. \tag{4.28}$$

[10]Man beachte, dass in (4.27) statt $\tilde{\Theta}^{(2)}$ auch andere $\tilde{\Theta}_i$ interpolierende Funktionen verwendet werden können.

Das entstehende Gitter $\Delta_{2,1}^m = (\bar{t}_0, \ldots, \bar{t}_{\bar{N}-1})$ ist nun für die Auswertung tauglich zu machen. Dazu werden die Gitterpunkte wie folgt festgelegt:

$$\begin{aligned}\Delta_{2,2} &= \{\bar{t}_j : j = 0 \vee j \equiv 0 \mod (m+1)\} \\ &= \{\bar{t}_0, \bar{t}_{m+1}, \bar{t}_{2(m+1)}, \ldots, \bar{t}_{\bar{N}-1}\} \\ &= \{\tau_0, \tau_1, \tau_2, \ldots, \tau_{\frac{\bar{N}-1}{m+1}}\} \\ &= \{\tau_0, \tau_1, \tau_2, \ldots, \tau_{N(\Delta_{2,2})}\}. \end{aligned} \tag{4.29}$$

In jedem Teilintervall von $\Delta_{2,2}$ werden dann m äquidistante Kollokationspunkte eingefügt, wodurch das Gitter $\Delta_{2,2}^m$ entsteht. Auf diesem Gitter könnte schon eine Auswertung erfolgen. Es stellt sich jedoch oft heraus, dass die Auswertung auf so einem Gitter unvorteilhaft ist. Dazu betrachten wir das Beispiel (5.3) mit den Eingabeparametern

$$m{=}2,\ \text{aTOL}{=}\text{rTOL}{=}1\text{E}{-}3,$$

$$\texttt{SMOOTHING\_FACTOR}{=}5\%,\ \texttt{RECOVER\_PEAKS}{=}1.$$

Die Nachteile, die sich bei der Auswertung auf dem Gitter $\Delta_{2,2}^2$ ergeben, sind in der Abbildung 4.34 zu sehen. Sehr auffällig ist die große Variation der Schrittweiten, die große Werte von $\frac{\mathbf{h}}{h_{min}}$ zur Folge hat. Vergleicht man die Abbildungen 4.24 und 4.34, so fällt auf, dass auf dem neuen Gitter die Fehlerschätzung sehr unzuverlässig geworden ist und der Wert von TOL_q stark zugenommen hat. Eine mögliche Ursache dieser Effekte ist darin zu suchen, dass in dem neu verteilten Gitter der Quotient $\frac{\mathbf{h}}{h_{min}}$ im Prinzip beliebig groß werden kann.
Um diesem Verhalten entgegenzuwirken und die Auswertung auf dem verteilten Gitter möglichst "robust" zu machen, wird der Quotient $\frac{\mathbf{h}}{h_{min}}$ nach oben beschränkt,

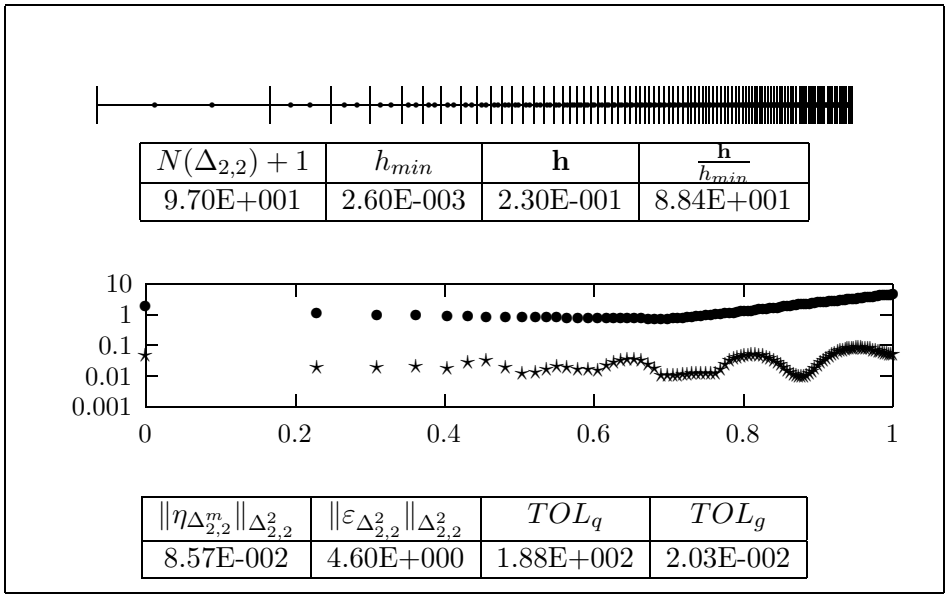
$$\frac{\mathbf{h}}{h_{min}} \leq 10. \tag{4.30}$$

Diese Schranke wird noch in §6.5 eingehend diskutiert.
Ein Gitter, das die Beschränkung (4.30) erfüllt, kann wie folgt konstruiert werden:

Algorithmus 4.1 *Erzeugt ein Gitter, für das* $\frac{\mathbf{h}}{h_{min}} \leq 10$ *gilt*[11].

- *Schritt 1: Ermittle* $\frac{\mathbf{h}}{h_{min}}$

[11] Es führt zu keinen Verwechslungen, wenn wir hier das Gitter als Menge von Punkten interpretieren.



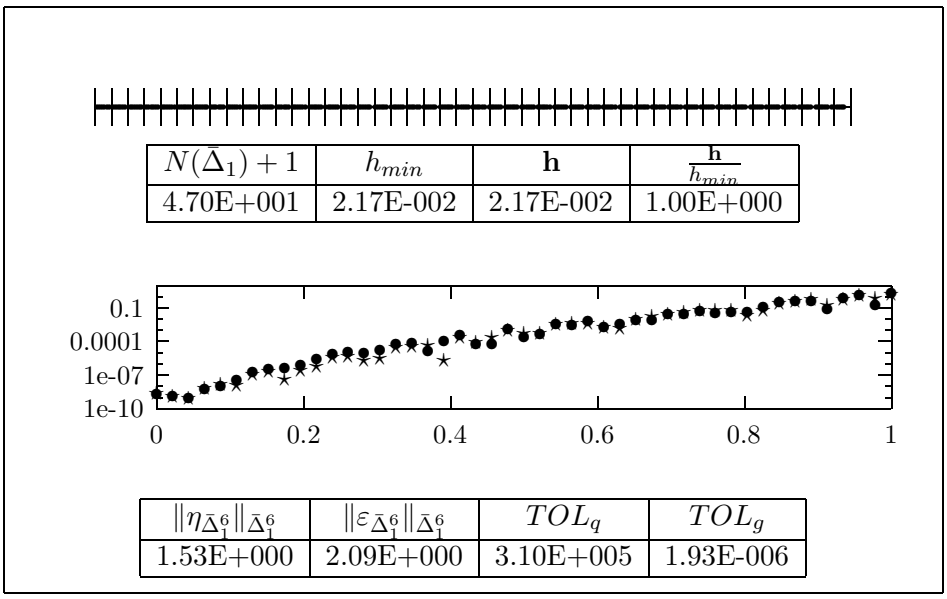
$N(\Delta_{2,2})+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
9.70E+001	2.60E-003	2.30E-001	8.84E+001

$\|\eta_{\Delta^m_{2,2}}\|_{\Delta^2_{2,2}}$	$\|\varepsilon_{\Delta^2_{2,2}}\|_{\Delta^2_{2,2}}$	TOL_q	TOL_g
8.57E-002	4.60E+000	1.88E+002	2.03E-002

Abbildung 4.34: Auswertung auf einem nicht stabilisierten Gitter

- *Schritt 2: Solange* $\frac{\mathbf{h}}{h_{min}} > 10$ *gilt*
  - *finde* $j \in 0, \ldots, N(\Delta)-1$, *sodass* $h_j = h_{max}$,
  - $\Delta \cup \frac{\tau_j+\tau_{j+1}}{2} \rightarrow \Delta$,
  - *gehe zu Schritt 1.*

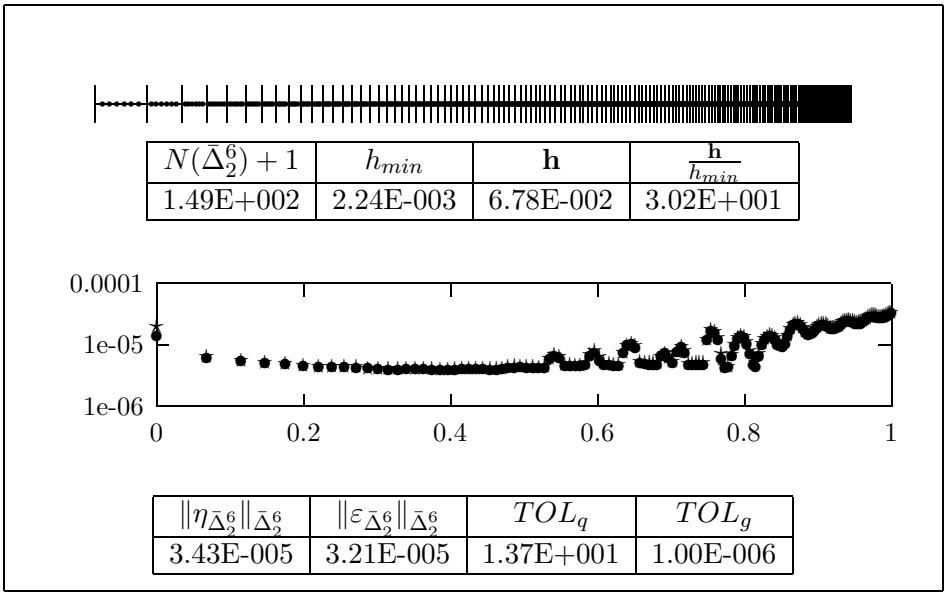
Es wird also solange in der Mitte des jeweils längsten Intervalls ein Gitterpunkt eingefügt, bis das Gitter die Forderung $\frac{\mathbf{h}}{h_{min}} \leq 10$ erfüllt. Die Schleife wird so oft durchlaufen, wie ein Gitterpunkt eingefügt wird. Die Anzahl dieser Schleifendurchläufe ist nach oben mit 1000 beschränkt. Als Ergebnis des Algorithmus 4.1 erhalten wir das Gitter $\Delta_{2,3}$. Durch das Hinzufügen der Kollokationspunkte ergibt sich das für die Auswertung taugliche Gitter $\bar{\Delta}^m_{2,3} =: \bar{\Delta}^m_2$. In den Abbildungen 4.35 und 4.36 wird der typische Verlauf der Gitteranpassung illustriert. Man erkennt, dass das Ziel, den Fehlerverlauf durch die Gleichverteilung auf ungefähr konstantes Niveau zu bringen, erreicht wurde.



$N(\bar{\Delta}_1)+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
4.70E+001	2.17E-002	2.17E-002	1.00E+000

$\|\eta_{\bar{\Delta}_1^6}\|_{\bar{\Delta}_1^6}$	$\|\varepsilon_{\bar{\Delta}_1^6}\|_{\bar{\Delta}_1^6}$	TOL_q	TOL_g
1.53E+000	2.09E+000	3.10E+005	1.93E-006

Abbildung 4.35: Auswertung auf dem Basisgitter am Ende der Phase 1, Beispiel (5.4) m=6, aTOL=rTOL=1E−6



$N(\bar{\Delta}_2^6)+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
1.49E+002	2.24E-003	6.78E-002	3.02E+001

$\|\eta_{\bar{\Delta}_2^6}\|_{\bar{\Delta}_2^6}$	$\|\varepsilon_{\bar{\Delta}_2^6}\|_{\bar{\Delta}_2^6}$	TOL_q	TOL_g
3.43E-005	3.21E-005	1.37E+001	1.00E-006

Abbildung 4.36: Auswertung nach der Gleichverteilung am Ende der Phase 2, Beispiel (5.4) m=6, aTOL=rTOL=1E−6

Algorithmische Details

Wir diskutieren jetzt eingehend zwei wichtige Bestandteile des Algorithmus zur Gitteranpassung.

Der Glättungsalgorithmus:

Die Summe in (4.17) muss nicht in jedem Schritt, $j = 0, \ldots, N(\Delta_1^m)$, berechnet werden: Mit $N = N(\Delta_1^m)$ gilt

$$\bar{\Theta}_j = \Psi(\Theta(\varepsilon_j), s_a) := \frac{1}{\min(N, j+s_a) - \max(0, j-s_a) + 1} \sum_{k=\max(0,j-s_a)}^{\min(N,j+s_a)} \Theta_k. \tag{4.31}$$

Fall 1: Befindet man sich am Anfang des Integrationsintervalls, d.h. gilt

$$\begin{aligned} j - s_a < 0 \quad &\wedge \quad j + 1 + s_a \leq N \quad \Rightarrow \\ j + 1 - s_a \leq 0 \quad &\wedge \quad j + s_a < N, \end{aligned}$$

dann ist

$$\begin{aligned} \bar{\Theta}_{j+1} &= \frac{1}{\min(N, j+1+s_a) - \max(0, j+1-s_a) + 1} \sum_{k=\max(0,j+1-s_a)}^{\min(N,j+1+s_a)} \Theta_k \\ &= \frac{1}{j+2+s_a} \sum_{k=0}^{j+1+s_a} \Theta_k \\ \text{und} \quad \bar{\Theta}_j &= \frac{1}{j+1+s_a} \sum_{k=0}^{j+s_a} \Theta_k, \\ \bar{\Theta}_{j+1} &= \frac{1}{j+2+s_a} \left[\sum_{k=0}^{j+s_a} \Theta_k + \Theta_{j+1+s_a} \right] \\ &= \frac{j+1+s_a}{j+2+s_a} \bar{\Theta}_j + \frac{1}{j+2+s_a} \Theta_{j+1+s_a}. \end{aligned} \tag{4.32}$$

Fall 2a: Gilt

$$\begin{aligned} j - s_a \geq 0 \quad &\wedge \quad j + 1 + s_a \leq N \quad \Rightarrow \\ j + 1 - s_a > 0 \quad &\wedge \quad j + s_a < N, \end{aligned}$$

dann setzt man

$$\begin{aligned}\bar{\Theta}_{j+1} &= \frac{1}{2s_a+1}\sum_{k=j+1-s_a}^{j+1+s_a}\Theta_k = \frac{1}{2s_a+1}\left[\sum_{k=j-s_a}^{j+s_a}\Theta_k + \Theta_{j+1+s_a} - \Theta_{j-s_a}\right] \\ &= \bar{\Theta}_j + \frac{1}{2s_a+1}\big[\Theta_{j+1+s_a} - \Theta_{j-s_a}\big]. \end{aligned} \tag{4.33}$$

Fall 2b: Gilt hingegen

$$\begin{aligned} j - s_a < 0 \quad &\wedge \quad j+1+s_a > N \Rightarrow \\ j+1-s_a \le 0 \quad &\wedge \quad j+s_a \ge N, \end{aligned}$$

dann ergibt sich

$$\bar{\Theta}_{j+1} = \frac{1}{N-0+1}\sum_{k=0}^{N}\Theta_k = \bar{\Theta}_j. \tag{4.34}$$

Man beachte, dass sich die Fälle 2a und 2b ausschließen.
Fall 3: Befindet man sich am Ende des Integrationsintervalls, d.h. gilt

$$\begin{aligned} j - s_a \ge 0 \quad &\wedge \quad j+1+s_a > N \Rightarrow \\ j+1-s_a > 0 \quad &\wedge \quad j+s_a \ge N \end{aligned}$$

so ergeben analoge Berechnungen

$$\bar{\Theta}_{j+1} = \frac{N-j+s_a+1}{N-j+s_a}\bar{\Theta}_j - \frac{1}{N-j+s_a}\Theta_{j-s_a}. \tag{4.35}$$

Insgesamt hat man also für ein $j \in \{0 \dots N-1\}$:

$$\bar{\Theta}_0 = \frac{1}{s_a+1}\sum_{k=0}^{s_a}\Theta_k,$$

$N > 2s_a + 1$:

$$\bar{\Theta}_{j+1} = \begin{cases} j \in \{1,\dots,s_a-1\} & : \ \frac{j+1+s_a}{j+2+s_a}\bar{\Theta}_j + \frac{1}{j+2+s_a}\Theta_{j+1+s_a} \\ j \in \{s_a,\dots,N-s_a-1\} & : \ \bar{\Theta}_j + \frac{1}{2s_a+1}[\Theta_{j+1+s_a} - \Theta_{j-s_a}] \\ j \in \{N-s_a,\dots,N\} & : \ \frac{N-j+s_a+1}{N-j+s_a}\bar{\Theta}_j - \frac{1}{N-j+s_a}\Theta_{j-s_a} \end{cases}$$

$s_a \le N \le 2s_a + 1$:

$$\bar{\Theta}_{j+1} = \begin{cases} j \in \{1,\dots,s_a-1\} & : \ \frac{j+1+s_a}{j+2+s_a}\bar{\Theta}_j + \frac{1}{j+2+s_a}\Theta_{j+1+s_a} \\ j \in \{s_a,\dots,N-s_a-1\} & : \ \bar{\Theta}_j \\ j \in \{N-s_a,\dots,N\} & : \ \frac{N-j+s_a+1}{N-j+s_a}\bar{\Theta}_j - \frac{1}{N-j+s_a}\Theta_{j-s_a} \end{cases}$$

$N < s_a$:

$$\bar{\Theta}_{j+1} = \bar{\Theta}_j.$$

Durch die obige Rekursion kann der Berechnungsaufwand des Glättungsalgorithmus stark reduziert werden.

Der Verteilungsalgorithmus:

Hier sind die wesentlichen Schritte des Verteilungsalgorithmus angeführt:

$$\Delta_{2,1}^m = \Phi\big((t_i, \tilde{\Theta}_i), \bar{N}\big)$$

$$\big((t_i, \tilde{\Theta}_i), \bar{N}\big) \overset{(4.20)}{\longrightarrow} (\tilde{\Theta}^{(2)}(t), \bar{N}) \overset{(4.22),(4.27)}{\longrightarrow} (\bar{t}_0, \ldots, \bar{t}_{\bar{N}-1}) =: \Delta_{2,1}. \quad (4.36)$$

Die Aufgabe des Verteilungsalgorithmus ist die Berechnung der neuen Gitterpunkte nach (4.27). Seien die Punkte t_i des Gitters Δ_1^m mit $i \in \{0, \ldots, N(\Delta_1^m)\}$ indiziert. Zur Indizierung der Punkte $\bar{t}_j$ des Gitters $\Delta_{2,1}^m$ wählen wir $j \in \{0, \ldots, N(\Delta_{2,1}^m)\}$. Auf dem alten Gitter führen wir für die "Flächen" unter der Treppenfunktion $\tilde{\Theta}^{(2)}$ und für die Schrittweiten die folgende Notation ein:

$$A_i := \int_{t_{i-1}}^{t_i} \Theta^{(2)}(\tau) d\tau = \frac{\tilde{\Theta}_{i-1} + \tilde{\Theta}_i}{2} h_i, \quad (4.37)$$

$$h_i = t_i - t_{i-1} \quad i = 1, \ldots, N(\Delta_1^m). \quad (4.38)$$

Die Abbildung 4.37 zeigt schematisch den Ablauf des Gleichverteilungsalgorithmus. Die rechteckigen Balken stellen die Treppenfunktion $\Theta^{(2)}(t)$ dar. Die darin dargestellten zusammenhängenden schraffierten Bereiche stellen gleichgroße Flächen dar. Wie zu sehen ist, kann ein Intervall $[\bar{t}_j, \bar{t}_{j+1}]$ keinen, einen oder mehrere Punkte t_i enthalten. Eine direkte Implementation des Gleichverteilungsalgorithmus muss daher die einzelnen Flächen A_i von einem Punkt $\bar{t}_j$ aus summieren, bis der Wert $\frac{I}{N(\Delta_{2,1}^m)}$ überschritten ist. Dann ist nämlich das Intervall $[t_i, t_{i+1}]$ gefunden, in dem $\bar{t}_{j+1}$ liegt. Anschliessend muss nur noch der exakte Wert für $\bar{t}_{j+1}$ errechnet werden.
Aus diesem Konzept heraus entsteht folgender Algorithmus:

Algorithmus 4.2

- *Schritt 1: Initialisiere*

$$j = 1, \quad A_h = 0, \quad l = 1, \quad t_0 = \bar{t}_0.$$

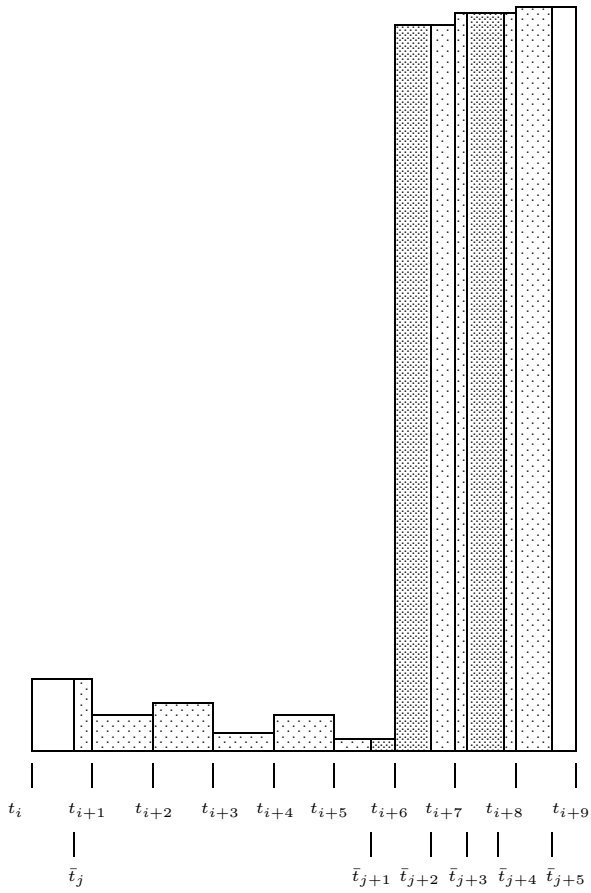


Abbildung 4.37: Schematische Darstellung zum Prinzip der Gleichverteilung

- *Schritt 2: Suche das Intervall in dem $\bar{t}_j$ liegt, also suche ein k sodass*[12]

$$A_h + \sum_{i=l}^{k} A_i \leq \frac{I}{N(\Delta_{2,1}^m)} < A_h + \sum_{i=l}^{k+1} A_i$$

gilt. Der neue Gitterpunkt befindet sich im Intervall $[t_k, t_{k+1}]$, die Fläche A_{k+1} wird geteilt.

- *Schritt 3: Berechne $\bar{t}_j$ aus*

$$\frac{I}{N(\Delta_{2,1}^m)} = A_h + \sum_{i=l}^{k} A_i + (\bar{t}_j - t_k)\frac{\tilde{\Theta}_{k+1} + \tilde{\Theta}_k}{2}.$$

- *Schritt 4: Berechne die neuen Initialwerte*

$$A_h = A_{k+1} - (\bar{t}_j - t_k)\frac{\tilde{\Theta}_{k+1} + \tilde{\Theta}_k}{2}; \quad l = k + 2; \quad j = j + 1.$$

- *Schritt 5: Wenn*

$$j < N(\Delta_{2,1}^m) - 1$$

gehe zu Schritt 2, sonst ist man fertig.

[12]Hier ist auch die leere Summe, z.B. $k = 0$ für das erste Intervall zugelassen. Weiters muss man beachten, dass der Index i nicht den Wert von $N(\bar{\Delta}_1^m)$ übersteigt.

4.2.5 Gitterverfeinerungen

Wenn die Auswertung auf dem Gitter $\bar{\Delta}_2^m$ die geforderte Güte nach (4.8) nicht erreicht, muss ein feineres Gitter erzeugt werden, ohne dass die neu gewonnene Strukturinformation auf dem Gitter $\bar{\Delta}_2^m$ verloren geht.
Nach den Überlegungen in §4.2.1 kann man mit Hilfe der Konvergenzordnung die Anzahl der Gitterpunkte so berechnen, dass man erwarten kann, die Toleranz zu unterschreiten. Der dazu notwendige Grad der Verfeinerung läßt sich wie folgt festlegen, vgl. (4.4):
Unser Ziel ist es, jenen Faktor zu berechnen, mit dem man die Länge der Intervalle $[\tau_{i-1}, \tau_i]$, $i = 1, \ldots, N(\bar{\Delta}_2)$ verkleinern muss, um das Fehlerniveau TOL_g zu erreichen. Dabei ist TOL_g der nach (4.24) berechnete globale Toleranzparameter. Folglich sucht man ein k, sodass

$$TOL_g \approx c \left(\frac{\mathbf{h}}{k}\right)^m \tag{4.39}$$

gilt mit

$$k = \frac{N(\bar{\Delta}_{3,1})}{N(\bar{\Delta}_2)}. \tag{4.40}$$

Durch Einsetzen von $\|\eta_{\bar{\Delta}_2^m}\|_{\bar{\Delta}_2^m} = c \cdot \mathbf{h}^m$ in (4.39) folgt

$$k \approx \sqrt[m]{\frac{\|\eta_{\bar{\Delta}_2^m}\|_{\bar{\Delta}_2^m}}{TOL_g}}. \tag{4.41}$$

Ersetzt man den globalen Fehler durch seine Fehlerschätzung, so ergibt sich die Anzahl der Punkte im neuen Gitter aus (4.40),

$$N(\bar{\Delta}_{3,1}^m) \approx N(\bar{\Delta}_{3,1}^m)_I := \left\lfloor N(\bar{\Delta}_2^m) \sqrt[m]{\frac{\|\varepsilon_{\bar{\Delta}_2^m}\|_{\bar{\Delta}_2^m}}{TOL_g}} \right\rfloor. \tag{4.42}$$

Für das gesuchte Intervallverhältnis $k \in \mathbb{N}$ gilt ungefähr

$$k \approx k_I := \frac{\left\lfloor N(\bar{\Delta}_2^m) \sqrt[m]{\frac{\|\varepsilon_{\bar{\Delta}_2^m}\|_{\bar{\Delta}_2^m}}{TOL_g}} \right\rfloor}{N(\bar{\Delta}_2^m)}. \tag{4.43}$$

Für $k = \lfloor k_I \rfloor$, $k > 1$, ergibt sich damit eine einfache Konstruktion des neuen Gitters:

$$\begin{aligned} &\bar{\Delta}_{3,1} = \bar{\Delta}_2 \cup \left\{\tau_{ij} : \tau_{ij} = \tau_i + j\frac{h_i}{k}\right\}, \\ &i = 0, \ldots, N(\bar{\Delta}_2) - 1, \quad j = 1, \ldots, k-1. \end{aligned} \tag{4.44}$$

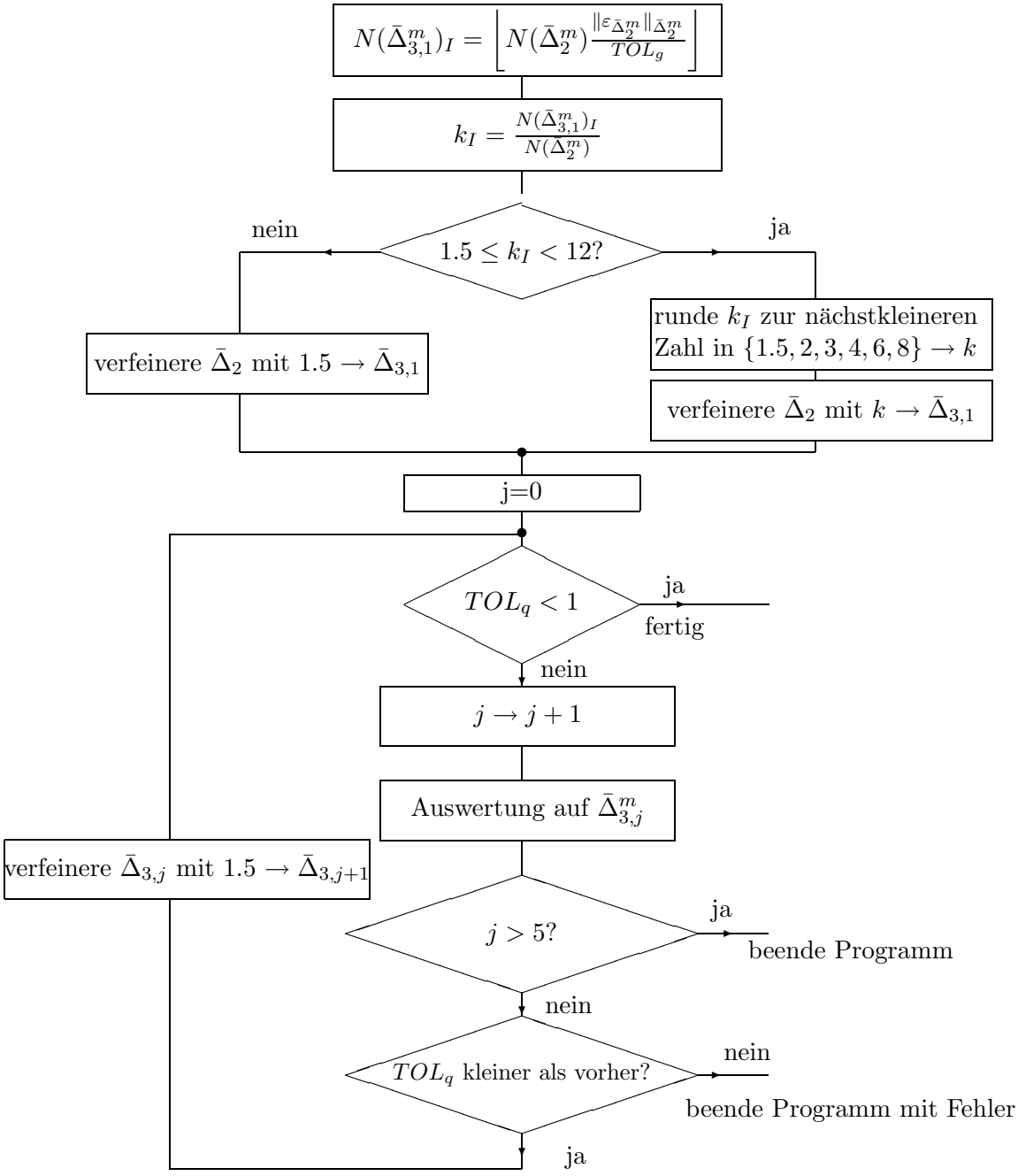
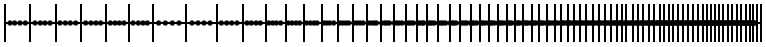
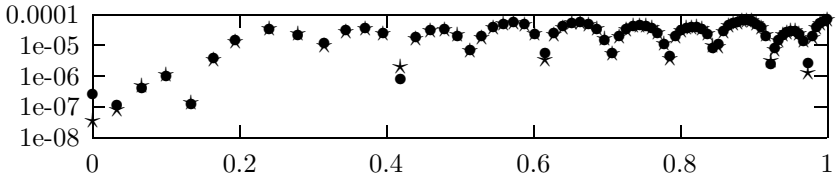


Abbildung 4.38: Schematischer Ablauf der Gitterverfeinerungsphase

Auswertung auf dem Gitter $\bar{\Delta}_2^4$:



$N(\bar{\Delta}_2)+1$	h_{min}	**h**	$\frac{\mathbf{h}}{h_{min}}$
7.70E+001	4.77E-003	4.54E-002	9.51E+000



$\|\eta_{\bar{\Delta}_2^4}\|_{\bar{\Delta}_2^4}$	$\|\varepsilon_{\bar{\Delta}_2^4}\|_{\bar{\Delta}_2^4}$	TOL_q	TOL_g
6.78E-005	7.05E-005	4.10E+000	1.32E-005

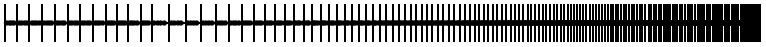
Abbildung 4.39: Das Gitter $\bar{\Delta}_2^4$ und die dazugehörige Auswertung vor der Verfeinerung, Beispiel (5.3), m=4, aTOL=rTOL=1E−5

Wie das folgende Beispiel zeigt, siehe Abbildungen 4.39 und 4.40, liefert die Verfeinerungsstufe $k = 2$ oft ein zu feines Gitter, sodass die Toleranz stark unterschritten wird. Dieser Fall liegt dann vor, wenn bei der Auswertung nach der Gleichverteilung die geforderte Toleranz knapp nicht erreicht wurde. Den obigen Abbildungen liegt das Beispiel (5.3) mit den Eingabedaten

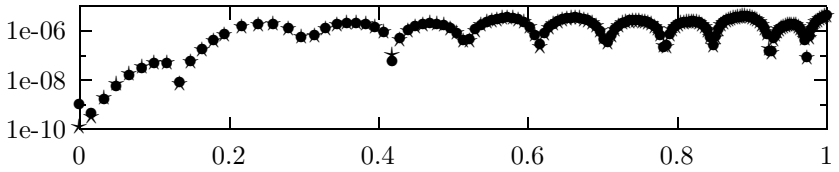
$$m\text{=}4,\ \text{aTOL=rTOL=1E}-5$$

zugrunde. Aus den Daten der Abbildung 4.39 erkennt man, dass das Fehlerniveau knapp die Toleranzforderung verfehlt. In Abbildung 4.40 ist zu sehen, dass durch Verdoppeln der Intervallanzahl die Toleranz weit unterschritten wird. Dieser Mangel kann durch Einführung einer Zwischenstufe mit $k = 1.5$ behoben werden. Um die Verfeinerung mit dem Faktor $k = 1.5$ zu realisieren werden aus 2 Intervallen 3 gemacht. Wenn die Anzahl der Intervalle im Gitter

Auswertung auf dem Gitter $\bar{\Delta}^4_{3,1}$:



$N(\bar{\Delta}^m_{3,1})+1$	h_{min}	**h**	$\frac{\mathbf{h}}{h_{min}}$
1.53E+002	2.38E-003	2.27E-002	9.51E+000



$\|\eta_{\bar{\Delta}^m_{3,1}}\|_{\bar{\Delta}^m_{3,1}}$	$\|\varepsilon_{\bar{\Delta}^m_{3,1}}\|_{\bar{\Delta}^m_{3,1}}$	TOL_q	TOL_g
4.23E-006	4.30E-006	2.58E-001	1.32E-005

Abbildung 4.40: Das Gitter $\bar{\Delta}^4_{3,1}$ und die dazugehörige Auswertung nach der Verfeinerung mit $k = 2$, Beispiel (5.3), m=4, aTOL=rTOL=1E−5

$\bar{\Delta}_2$ ungerade ist, muss ein zusätzliches Intervall eingefügt werden, wobei das Verhältnis $\frac{\mathbf{h}}{h_{min}}$ nicht größer werden darf. Wir wählen das neue Gitter wie folgt:

$$N(\bar{\Delta}_2) \text{ ungerade} \Rightarrow$$
$$\Delta_h = \bar{\Delta}_2 \cup \frac{\tau_j + \tau_{j+1}}{2}, \quad h_j = \mathbf{h}, \tag{4.45}$$

d.h. ein neuer Gitterpunkt wird in der Mitte des längsten Intervalls eingefügt. Das Gitter mit einer geraden Intervallanzahl kann mit $k = 1.5$ verfeinert werden, indem 2 Intervalle zusammengefaßt werden und an ihrer Stelle 3 gleich lange Intervalle erzeugt werden.
Wir betrachten noch einmal das vorige Beispiel. Die Auswertung auf dem Gitter $\bar{\Delta}_2$ war bereits in Abbildung 4.39 zu sehen. Im Vergleich zur Abbildung 4.40 ist jetzt in Abbildung 4.41 zu erkennen, dass durch Einführung der Verfeinerung mit dem Faktor $k = 1.5$ mit einer billigeren Auswertung die Toleranz unterschritten werden konnte.
Um den Wert k zu erhalten, wird der nach (4.43) berechnete Wert von $1.5 \leq k_I < 12$ zum nächstkleineren Wert aus der Menge $\{1.5, 2, 3, 4, 6, 8\}$ gerundet. Das Gitter $\bar{\Delta}_2$ wird dann nach (4.44) verfeinert, um $\bar{\Delta}_{3,1}$ zu erhalten. Für $0 < k_I < 1.5$ setzen wir $k = 1.5$. Wird $k \geq 12$, so nimmt man an, dass die Fehlerschätzung sehr unzuverlässig ist und wiederholt die Berechnungen auf einem neuen Gitter, das mit dem Faktor $k = 1.5$ verfeinert wurde. Als Illustration betrachten wir das Beispiel (5.7) mit den Eingabedaten

$$m{=}4, \text{ aTOL}{=}\text{rTOL}{=}1\text{E}{-}4.$$

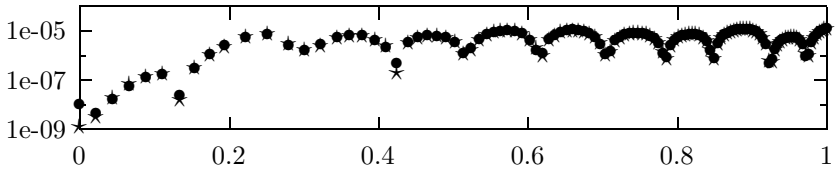
In der Abbildung 4.42 ist die Auswertung auf dem Gitter $\bar{\Delta}_2$ dargestellt. Trotz der Maßnahmen, die die Robustheit auf diesem Gitter erhöhen, tritt ein sehr großer Fehler $\|\varepsilon_{\bar{\Delta}_2^m}\|_{\bar{\Delta}_2^m}$ auf. Damit ist $k > 12$. Wir lasten diesen großen Wert nicht der schlechten Verteilung des Gitters an, sondern interpretieren ihn als Folge von einem zu großen $\mathbf{h}$. In diesem Beispiel liefert die 1.5-fache Verfeinerung sofort ein zufriedenstellendes Ergebnis, siehe Abbildung 4.43.
Das so entstandende Gitter $\bar{\Delta}_{3,1}$ erfüllt nach der Auswertung oft bereits die Toleranzanforderung. Der Algorithmus ist dann beendet und es wird die verbesserte Basislösung,

$$p_i - \varepsilon_i, \quad i = 0, \ldots, N(\bar{\Delta}_{3,1}^m) \tag{4.46}$$

ausgegeben mit der Information, dass der Algorithmus ohne Fehler beendet wurde.
Wenn die Auswertung noch nicht die geforderte Güte erreicht, so wird angenommen, dass die Toleranzanforderung knapp verfehlt wurde. Das Gitter $\bar{\Delta}_{3,1}$ wird dann mit dem Faktor $k = 1.5$ verfeinert und man erhält das

Auswertung auf dem Gitter $\bar{\Delta}^4_{3,1}$:

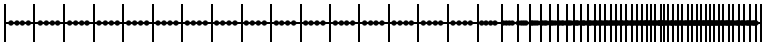
$N(\Delta_{3,1})+1$	h_{min}	**h**	$\frac{h_{max}}{h_{min}}$
1.15E+002	3.18E-003	2.83E-002	8.91E+000



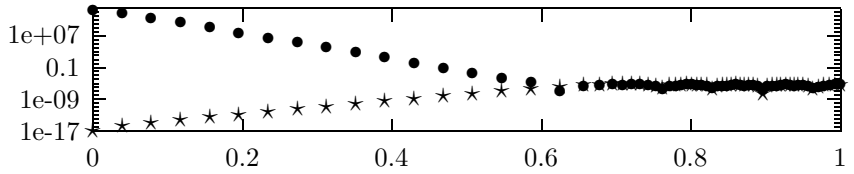
$\|\eta_{\bar{\Delta}^4_{3,1}}\|_{\bar{\Delta}^m_{3,1}}$	$\|\varepsilon_{\bar{\Delta}^m_{3,1}}\|_{\bar{\Delta}^m_{3,1}}$	TOL_q	TOL_g
1.31E-005	1.34E-005	7.94E-001	1.32E-005

Abbildung 4.41: Das Gitter $\bar{\Delta}^4_{3,1}$ und die dazugehörige Auswertung nach der Verfeinerung mit $k = 1.5$, Beispiel (5.3), m=4, aTOL=rTOL=1E−5

Auswertung auf dem Gitter $\bar{\Delta}_2^4$:



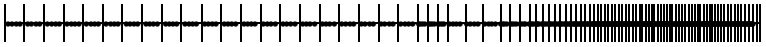
$N(\bar{\Delta}_2)+1$	h_{min}	**h**	$\frac{h_{max}}{h_{min}}$
6.20E+001	5.20E-003	3.91E-002	7.51E+000



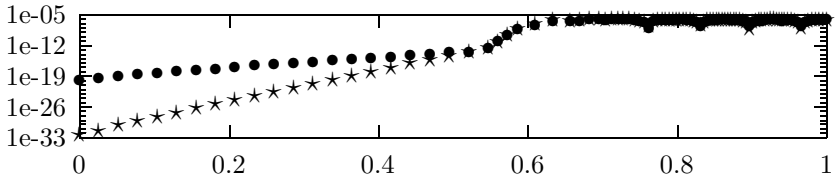
$\|\eta_{\bar{\Delta}_2^4}\|_{\bar{\Delta}_2^4}$	$\|\varepsilon_{\bar{\Delta}_2^4}\|_{\bar{\Delta}_2^4}$	TOL_q	TOL_g
8.92E-006	4.41E+013	4.41E+017	1.00E-004

Abbildung 4.42: Das Gitter $\bar{\Delta}_2^4$ und die dazugehörige Auswertung, Beispiel (5.7), m=4, aTOL=rTOL=1E−4

Auswertung auf dem Gitter $\bar{\Delta}_3^4$:



$N(\bar{\Delta}_3)+1$	h_{min}	**h**	$\frac{\mathbf{h}}{h_{min}}$
9.40E+001	3.58E-003	2.61E-002	7.28E+000



$\|\eta_{\bar{\Delta}_3^4}\|_{\bar{\Delta}_3^4}$	$\|\varepsilon_{\bar{\Delta}_3^4}\|_{\bar{\Delta}_3^4}$	TOL_q	TOL_c
1.55E-006	1.08E-006	9.97E-003	1.00E-004

Abbildung 4.43: Das Gitter $\bar{\Delta}_{3,1}^4$ und die dazugehörige Auswertung nach der Verfeinerung mit $k = 1.5$, Beispiel (5.7), m=4, aTOL=rTOL=1E−4

$\bar{\Delta}_{3,2}$. Wenn nach der Auswertung auf diesem Gitter die Toleranz wieder nicht erfüllt ist, wird diese Vorgangsweise sukzessive wiederholt. Bei strengen Toleranzanforderungen kann es vorkommen, dass die geforderte Güte nicht erreicht werden kann. Dieser Effekt kann durch Rechenfehler hervorgerufen werden und äußert sich dadurch, dass der Wert TOL_q nicht abnimmt. In diesem Fall wird das Programm mit einer entsprechenden Fehlermeldung beendet. In dieser Phase kann das Gitter maximal fünf mal verfeinert (immer mit dem Faktor $k = 1.5$) werden. Ist dann die Toleranz noch nicht erfüllt, so wird das Programm mit einer Fehlermeldung beendet, eine entsprechende Information wird dem Benutzer zu Verfügung gestellt. Auch in den Fällen, wo das Programm nicht ordnungsgemäß beendet wurde, wird die verbesserte Basislösung ausgegeben.

Kapitel 5

Beispielsammlung

Bei der Konstruktion der Testbeispiele wurde versucht ein möglichst allgemeines Spektrum von M abzudecken.
Zunächst werden Beispiele mit einem positiven und einem negativen Eigenwert von M beschrieben.

- Beispielreihe 200

$$\begin{aligned} v'(t) = \frac{1}{t}\begin{pmatrix} 0 & 1 \\ 2 & 6 \end{pmatrix} v(t) + \begin{pmatrix} 0 \\ -t\sin(k^2t^2)(4k^4t^4+10) \end{pmatrix}, \; t \in (0,1], \\ \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} v(0) + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} v(1) = \begin{pmatrix} 0 \\ \sin(k^2) \end{pmatrix}, \end{aligned} \tag{5.1}$$

mit exakter Lösung

$$v(t) = \begin{pmatrix} t^2\sin(k^2t^2) \\ 2t^2(\sin(k^2t^2) + t^2k^2\cos(k^2t^2)) \end{pmatrix}.$$

Die Eigenwerte von M sind:

$$\lambda_1 = 3 + \sqrt{11}, \quad \lambda_2 = 3 - \sqrt{11}.$$

Durch Änderung von k kann die Glattheit von $v(t)$ gesteuert werden. Dieses Beispiel wurde mit 3 verschiedenen Werten für k getestet:

- k=2:

$$v(t) = \begin{pmatrix} t^2\sin(4t^2) \\ 2t^2(\sin(4t^2) + 4t^2\cos(4t^2)) \end{pmatrix}, \tag{5.2}$$

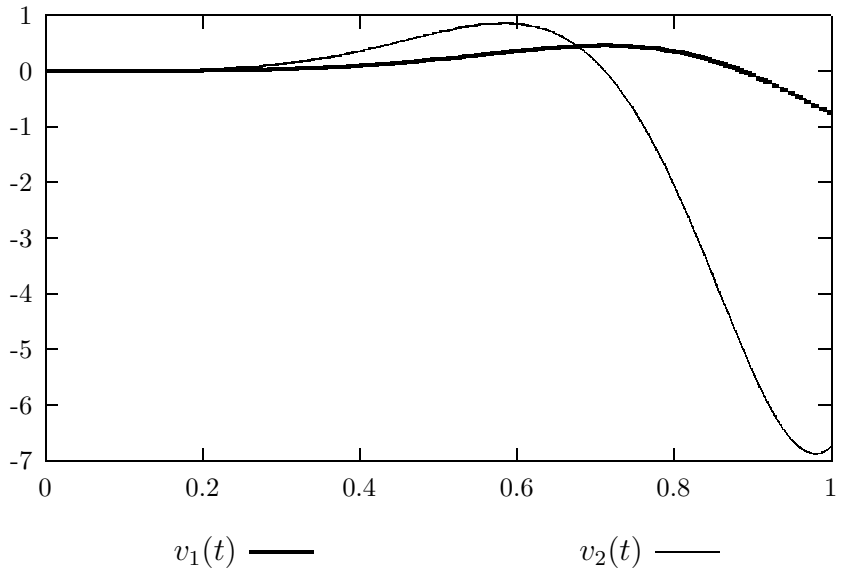


Abbildung 5.1: Graph der Lösung von Beispiel (5.2)

- k=5:

$$v(t) = \begin{pmatrix} t^2 \sin(25t^2) \\ 2t^2(\sin(25t^2) + 25t^2 \cos(25t^2)) \end{pmatrix}, \tag{5.3}$$

- k=10:

$$v(t) = \begin{pmatrix} t^2 \sin(100t^2) \\ 2t^2(\sin(100t^2) + 100t^2 \cos(100t^2)) \end{pmatrix}, \tag{5.4}$$

- Beispielreihe 100

$$v'(t) = \frac{1}{t} \begin{pmatrix} 0 & 1 \\ \mu^2 + \alpha^2 t^2 & 0 \end{pmatrix} v(t) + \begin{pmatrix} 0 \\ ct^{k-1} e^{-\alpha t}(k^2 - \mu^2 - \alpha t(1+2k)) \end{pmatrix},$$
$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} v(0) + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} v(1) = \begin{pmatrix} 0 \\ ce^{-\alpha} \end{pmatrix}, \quad t \in (0,1], \tag{5.5}$$

mit der exakten Lösung

$$v(t) = \begin{pmatrix} v_1(t) \\ v_2(t) \end{pmatrix} = \begin{pmatrix} ct^k e^{-\alpha t} \\ ct^k e^{-\alpha t}(k - \alpha t) \end{pmatrix}. \tag{5.6}$$

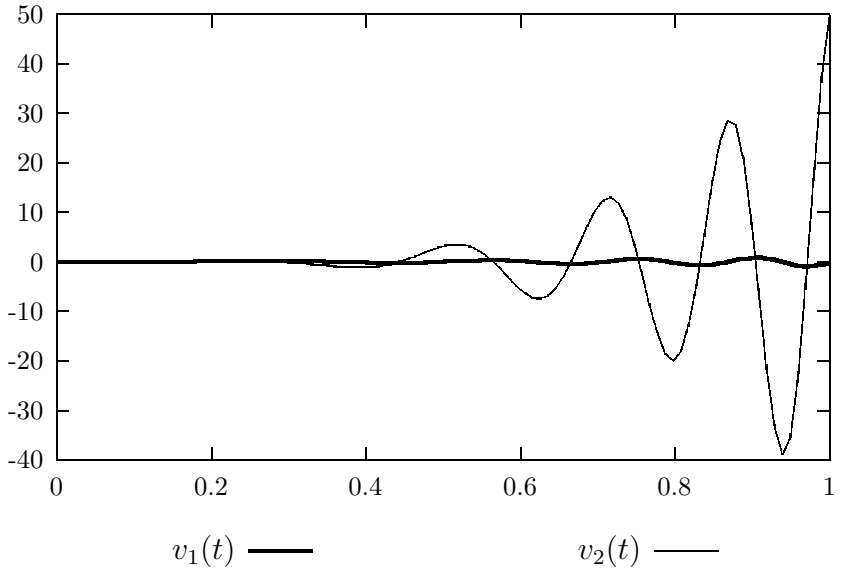


Abbildung 5.2: Graph der Lösung von Beispiel (5.3)

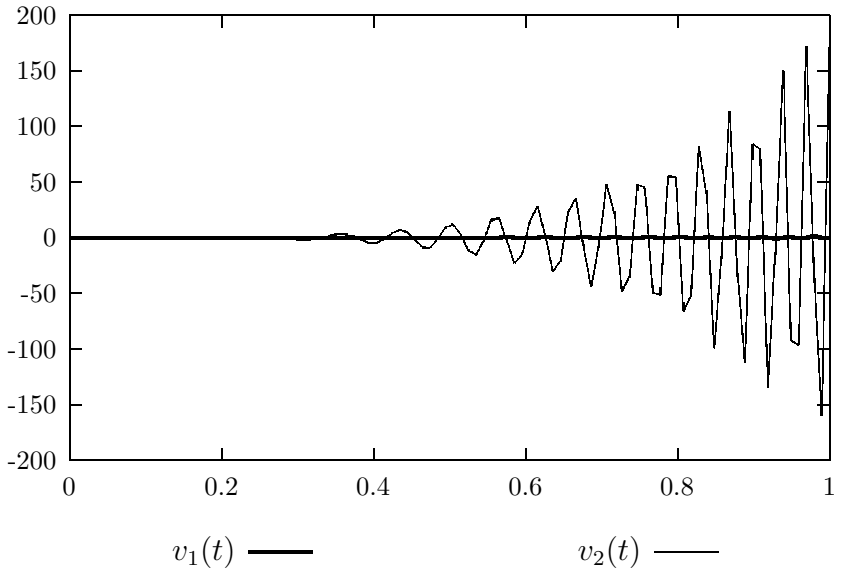


Abbildung 5.3: Graph der Lösung von Beispiel (5.4)

Wie aus der Lösungsstruktur (5.6) zu erkennen ist, ist c eine Skalierungskonstante. Diese ist so gewählt, dass $\|v_1\|_\infty = 1$ gilt. Es ergibt sich

$$c = \left(\frac{\alpha}{k}\right)^k \cdot e^k.$$

Mit dem Parameter μ kann die Eigenwertstruktur von M beeinflusst werden. Mit $\mu = 1$ ergibt sich $\lambda_1 = 1$ und $\lambda_2 = -1$. Die eindeutige Lösung $z(t) \in C[0,1]$ ist nicht von μ abhängig.

In [16] ist eine andere Parametrisierung dieses Beispiels zu finden. Mit Hilfe von

$$p_1 = \frac{k}{\alpha}, \quad p_2 = \frac{\sqrt{k}}{\alpha}$$

gewinnt man weitere Information über die Lösungsstruktur: p_1 ist der Abszissenwert des Maximums von v_1 und p_2 kann als Maß für die Breite des entstehenden "Peaks" gedeutet werden. Da man dadurch die "Unglattheit" der Lösung und den dazugehörigen Abszissenwert steuern kann, wurde dieses Beispiel mit verschiedenen Parametern getestet:

- Beispiel 1001

μ	p_1	p_2	k	α
1	0.9	0.05	324	360

$$v(t) = \begin{pmatrix} ct^{324}e^{-360t} \\ ct^{324}e^{-360t}(324 - 360t) \end{pmatrix} \tag{5.7}$$

- Beispiel 1002

μ	p_1	p_2	k	α
1	0.2	0.05	16	80

$$v(t) = \begin{pmatrix} ct^{16}e^{-80t} \\ ct^{16}e^{-80t}(16 - 80t) \end{pmatrix} \tag{5.8}$$

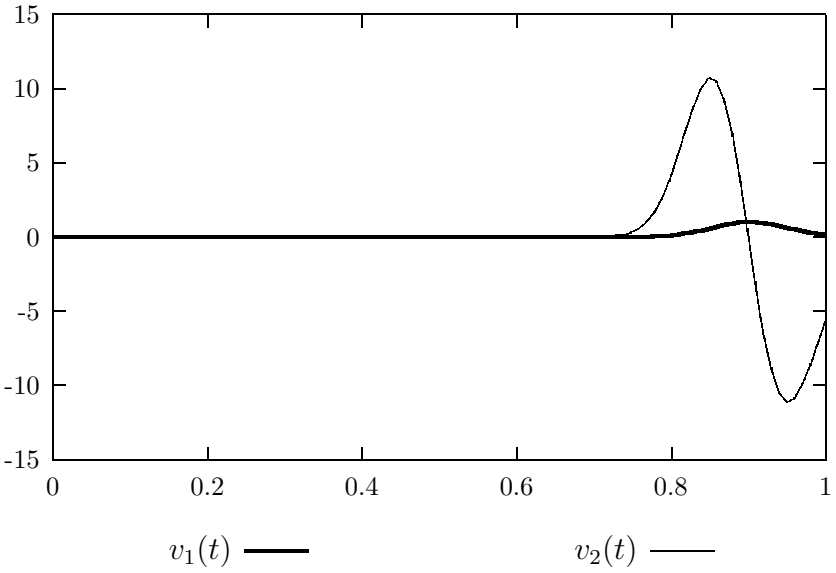


Abbildung 5.4: Graph der Lösung von Beispiel (5.7)

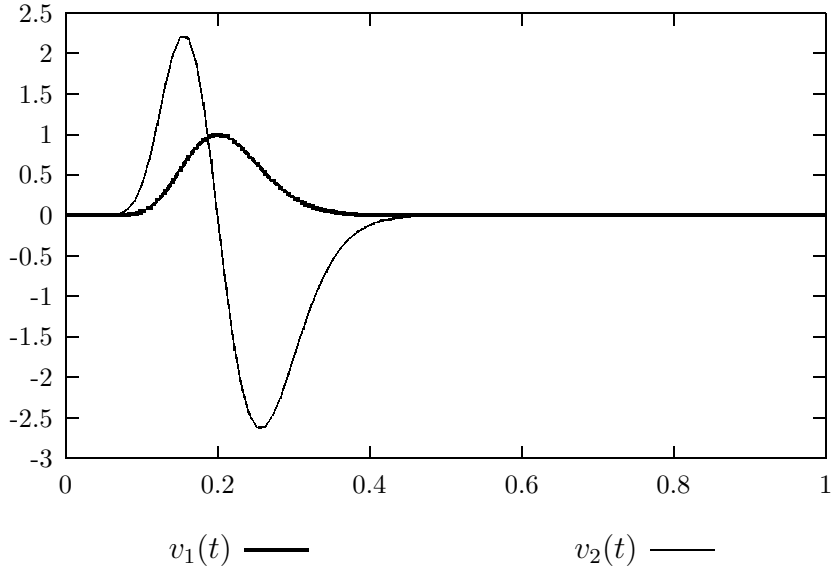


Abbildung 5.5: Graph der Lösung von Beispiel (5.8)

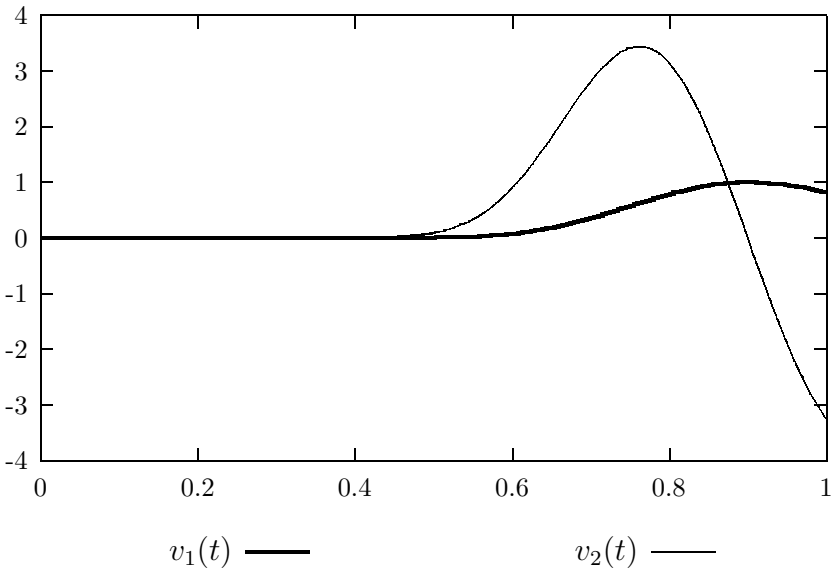


Abbildung 5.6: Graph der Lösung von Beispiel (5.9)

- Beispiel 1004

μ	p_1	p_2	k	α
1	0.9	0.15	36	40

$$v(t) = \begin{pmatrix} ct^{36}e^{-40t} \\ ct^{36}e^{-40t}(36 - 40t) \end{pmatrix} \tag{5.9}$$

- Beispiel 1005

μ	p_1	p_2	k	α
1	0.5	0.05	100	200

$$v(t) = \begin{pmatrix} ct^{100}e^{-200t} \\ ct^{100}e^{-200t}(100 - 200t) \end{pmatrix} \tag{5.10}$$

- Beispiel 1003

  Bei diesem Beispiel wurde durch die Wahl von $\mu = 0$ die Eigenwertstruktur von M verändert. Es gilt hier $\lambda_1 = \lambda_2 = 0$,

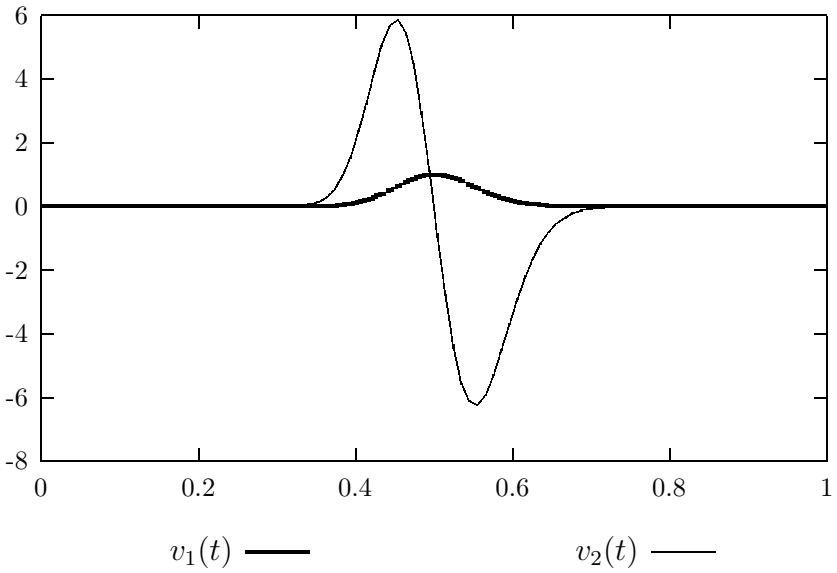


Abbildung 5.7: Graph der Lösung von Beispiel (5.10)

μ	p_1	p_2	k	α
0	1.5	0.6124	6	4

$$v(t) = \begin{pmatrix} ct^6 e^{-4t} \\ ct^6 e^{-4t}(6-4t). \end{pmatrix} \tag{5.11}$$

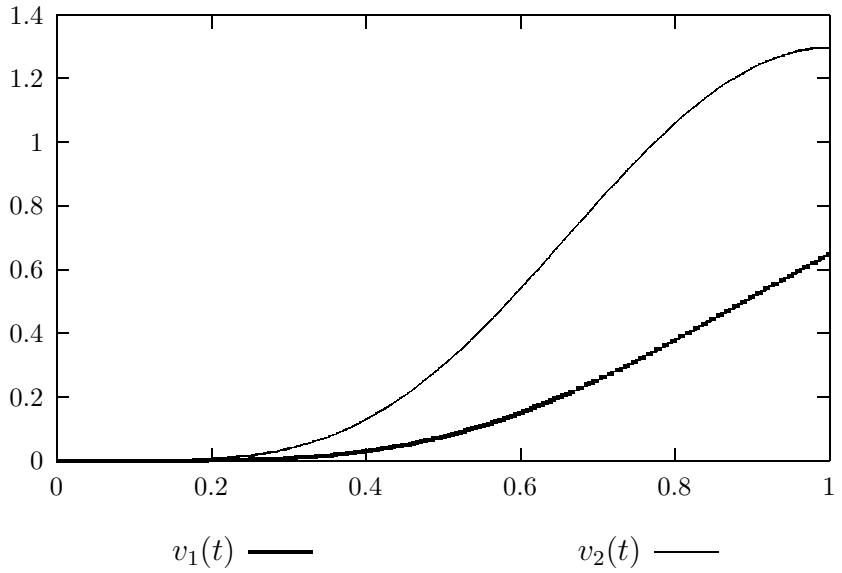


Abbildung 5.8: Graph der Lösung von Beispiel (5.11)

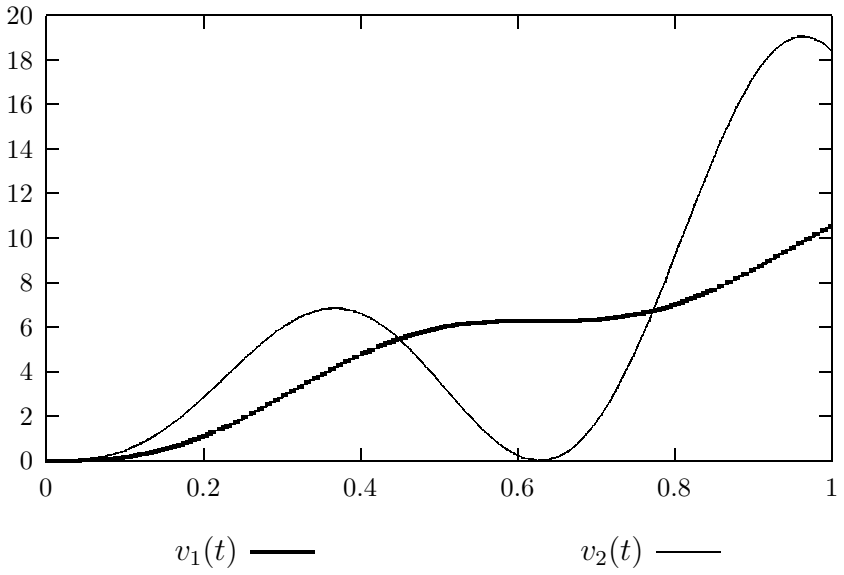


Abbildung 5.9: Graph der Lösung von Beispiel (5.12)

Es folgt ein Beispiel, bei dem ein positiver Eigenwert, und ein Eigenwert $\lambda = 0$ auftritt.

- Beispiel 15

$$v'(t) = \frac{1}{t}\begin{pmatrix} 0 & 1 \\ -100t^2 & 2 \end{pmatrix} v(t) + 10\begin{pmatrix} 0 \\ -1000t^2 + 10\cos(10t) - 10 \end{pmatrix},$$
$$\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} v(0) + \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} v(1) = \begin{pmatrix} 10 - \sin(10) \\ 0 \end{pmatrix},$$
$$t \in (0, 1], \tag{5.12}$$

mit exakter Lösung

$$v(t) = \begin{pmatrix} 10t - \sin(10t) \\ 10t(1 - \cos(10t)) \end{pmatrix}.$$

Die Eigenwerte von $M(0)$ sind

$$\lambda_1 = 0, \quad \lambda_2 = 2.$$

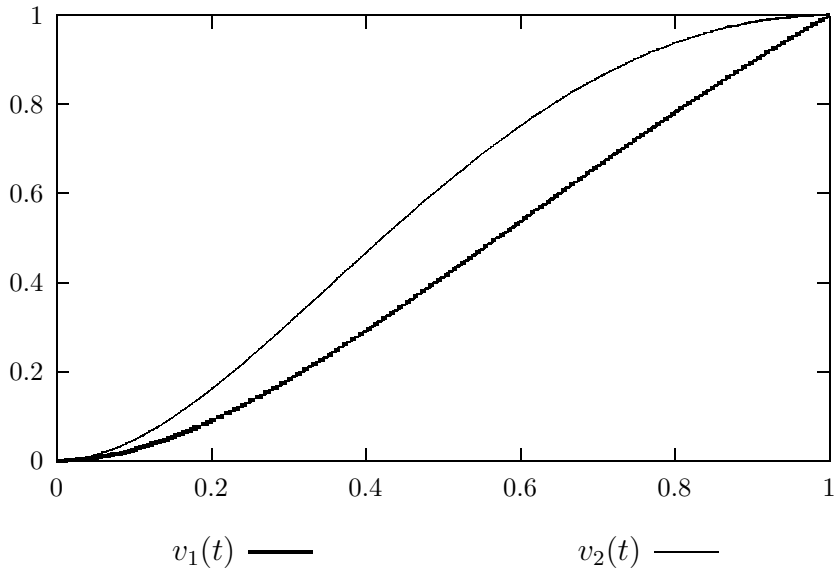


Abbildung 5.10: Graph der Lösung von Beispiel (5.13)

Im folgenden Beispiel hat $M(0)$ nur positive Eigenwerte.

- Beispiel 17e

$$v'(t) = \frac{1}{t}\begin{pmatrix} 0 & 1 \\ -6+t^2 & 5 \end{pmatrix} v(t), \quad t \in (0,1],$$
$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} v(0) + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} v(1) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \tag{5.13}$$

mit exakter Lösung

$$v(t) = \begin{pmatrix} t^2 e^{1-t} \\ (2-t)t^2 e^{1-t} \end{pmatrix}.$$

Die Eigenwerte von $M(0)$ lauten

$$\lambda_1 = 2, \quad \lambda_2 = 3.$$

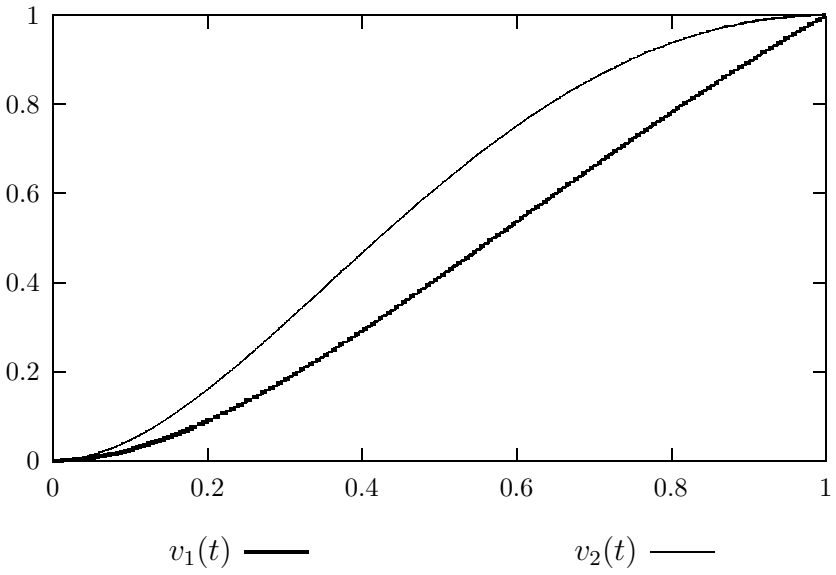


Abbildung 5.11: Graph der Lösung von Beispiel (5.14)

Es folgt ein Beispiel, bei dem $M(0)$ nur negative Eigenwerte hat.

- Beispiel 18e

$$v'(t) = \frac{1}{t}\begin{pmatrix} 0 & 1 \\ -2-8t+t^2 & -3 \end{pmatrix} v(t) + t\begin{pmatrix} 0 \\ 12te^{1-t} \end{pmatrix},\ t \in (0,1],$$
$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} v(0) + \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} v(1) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad (5.14)$$

  mit exakter Lösung

$$v(t) = \begin{pmatrix} t^2e^{1-t} \\ (2-t)t^2e^{1-t} \end{pmatrix}.$$

  Die Eigenwerte von $M(0)$ sind

$$\lambda_1 = -1, \quad \lambda_2 = -2.$$

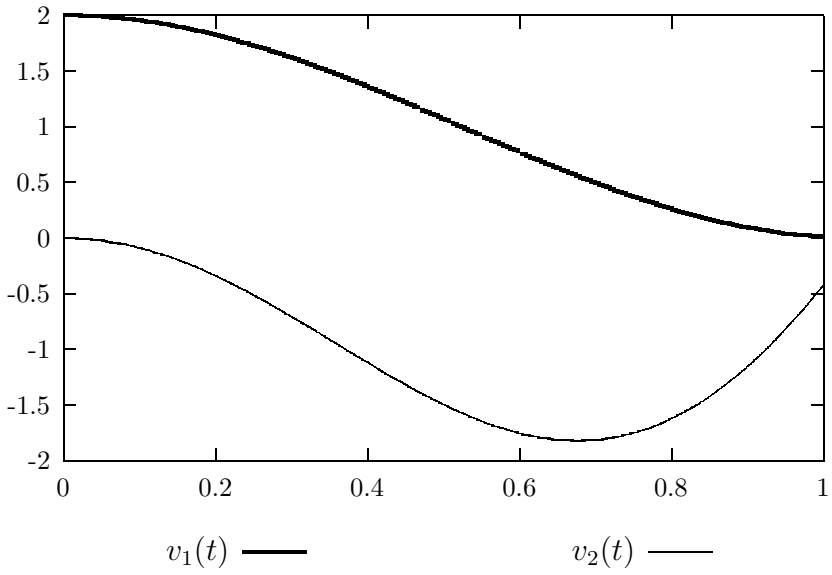


Abbildung 5.12: Graph der Lösung von Beispiel (5.15)

Im nächsten Beispiel tritt ein negativer Eigenwert und ein Eigenwert $\lambda = 0$ auf.

- Beispiel 35b

$$v'(t) = \frac{1}{t}\begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix} v(t) + \begin{pmatrix} 0 \\ -9t\cos(3t) - 6t\sin(3t) \end{pmatrix}, \; t \in (0,1],$$
$$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} v(0) + \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} v(1) = \begin{pmatrix} 1+\cos(3) \\ 0 \end{pmatrix}, \qquad (5.15)$$

  mit exakter Lösung

$$v(t) = \begin{pmatrix} 1+\cos(3t) \\ t(-3\sin(3t)) \end{pmatrix}.$$

  Die Eigenwerte von $M(0)$ sind

$$\lambda_1 = 0, \quad \lambda_2 = -1.$$

Das nächste Beispiel hat 2 Eigenwerte $\lambda = 0$ und 2 negative Eigenwerte.

- Beispiel 9a

$$v'(t) = \frac{1}{t}\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & -\frac{1}{2} & 0 \\ 0 & 0 & 0 & -\frac{1}{2} \end{pmatrix} v(t) + t \begin{pmatrix} 0 \\ 0 \\ e^{2t}(5+11t+4t^2) - e^{-2t} \\ e^{-2t}(5-11t+4t^2) \end{pmatrix},$$
$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} v(0) + \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} v(1) = \begin{pmatrix} 3+e^2 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$
$$t \in (0,1], \tag{5.16}$$

mit exaxter Lösung

$$v(t) = \begin{pmatrix} 3 + t^2 e^{2t} \\ t^2 e^{-2t} \\ 2t^2 e^{2t}(1+t) \\ 2t^2 e^{-2t}(1-t) \end{pmatrix}.$$

Die Eigenwerte von $M(0)$ sind hier

$$\lambda_1 = 0, \quad \lambda_2 = 0, \quad \lambda_3 = -\frac{1}{2}, \quad \lambda_4 = -\frac{1}{2}.$$

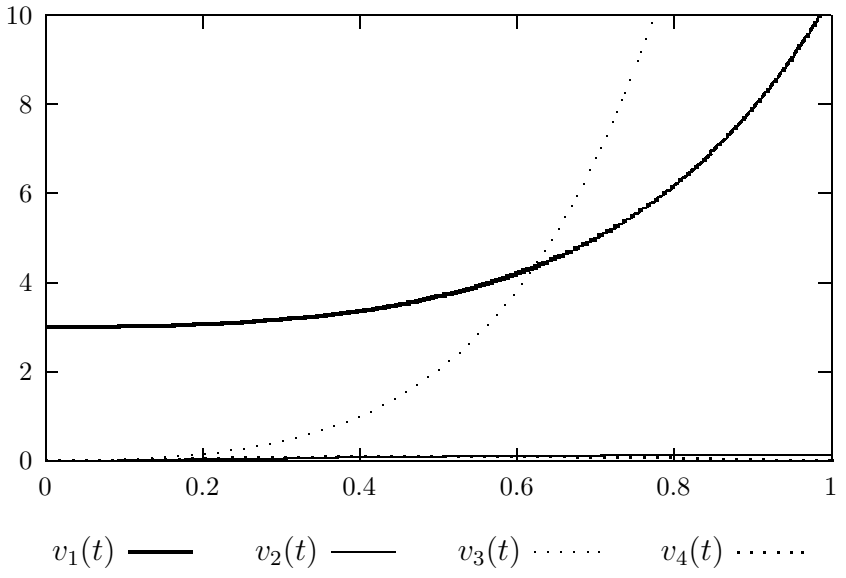


Abbildung 5.13: Graph der Lösung von Beispiel (5.16)

Die Entwicklung des Algorithmus ist auf die effiziente Lösung von singulären Randwertproblemen ausgerichtet. Es wurden jedoch auch reguläre Probleme herangezogen, um das Verhalten des Programms bei dieser Problemklasse beobachten zu können. Der Vollständigkeit halber werden nun die regulären Testbeispiele angeführt.

- Beispiel reg1

$$\begin{aligned} v'(t) = \begin{pmatrix} 0 & 1 \\ 4 & 0 \end{pmatrix} v(t) + -3 \begin{pmatrix} 0 \\ e^t \end{pmatrix}, \quad t \in [-1, 3], \\ \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} v(-1) + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} v(3) = \begin{pmatrix} e^{-1} \\ e^3 \end{pmatrix}, \end{aligned} \tag{5.17}$$

  mit exakter Lösung

$$v(t) = \begin{pmatrix} e^t \\ e^t \end{pmatrix}.$$

- Beispiel reg4

$$\begin{aligned} v'(t) = -\sin t, \quad t \in [0, 1], \\ v(0) = 1, \end{aligned} \tag{5.18}$$

  mit exakter Lösung

$$v(t) = \cos(t).$$

- Beispiel reg7

$$\begin{aligned} v'(t) = e^{-t} v(t), \quad t \in [0, 5], \\ v(0) = 1, \end{aligned} \tag{5.19}$$

  mit exakter Lösung

$$v(t) = e^{1-e^{-t}}.$$

- Beispiel reg8

$$\begin{aligned} v'(t) = \begin{pmatrix} \psi(t) & 1 \\ 2\psi(t) & -\psi(t) \end{pmatrix} v(t) + \begin{pmatrix} (1 - \psi(t))e^t \\ 2e^t \end{pmatrix}, \quad t \in [0, 1], \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} v(0) = \begin{pmatrix} 1 + e \\ 2 + 2e \end{pmatrix}, \end{aligned} \tag{5.20}$$

mit exakter Lösung

$$v(t) = \begin{pmatrix} e^t \\ 2e^t \end{pmatrix}.$$

Es gilt

$$\psi(t) = 20 \sin t + 20t \cos t.$$

- Beispiel reg9

$$\begin{gathered} v'(t) = \begin{pmatrix} -5 & 0 \\ 0 & 5 \end{pmatrix} v(t), \quad t \in [0,1], \\ \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} v(0) + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} v(1) = \begin{pmatrix} 3 \\ e^5 \end{pmatrix}, \end{gathered} \tag{5.21}$$

mit exakter Lösung

$$v(t) = \begin{pmatrix} 3e^{-5t} \\ e^{5t} \end{pmatrix}.$$

Kapitel 6

Numerische Resultate

6.1 Numerische Bestimmung der Konvergenzordnungen

Das Basisverfahren der Kollokation hat für reguläre Randwertprobleme die in (2.20) beschriebenen Konvergenzordnungen. Die folgenden Ergebnisse zeigen, dass dieses Verhalten auch für singuläre Randwertprobleme der Form (1.1) gilt.
Zunächst wird die Kollokation auf dem Initialgitter Δ^m ausgeführt. Dann wird der Fehler $\|\eta_{\Delta^m}\|_{\Delta^m}$ ermittelt. Das Gitter Δ wird anschließend kohärent verfeinert, indem jedes Teilintervall halbiert wird. Das entstehende Gitter wird mit $\tilde{\Delta}^m$ bezeichnet. Nach der Ermittlung der Basislösung auf diesem Gitter $\tilde{\Delta}^m$ kann das Gleichungssystem

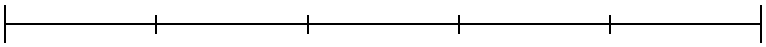
$$\begin{aligned} \|\eta_{\Delta^m}\|_{\Delta^m} &= c_e\, \mathbf{h}^{m_e}, \\ \|\eta_{\tilde{\Delta}^m}\|_{\tilde{\Delta}^m} &= c_e\, (\frac{\mathbf{h}}{2})^{m_e}, \end{aligned}$$

gelöst werden. Dabei gibt m_e die empirische Konvergenzordnung an, und c_e die zugehörige Fehlerkonstante. Es gilt

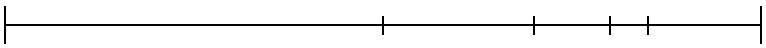
$$m_e = \log \frac{\|\eta_{\Delta^m}\|_{\Delta^m}}{\|\eta_{\tilde{\Delta}^m}\|_{\tilde{\Delta}^m}} / \log 2, \quad c_e = \frac{\|\eta_{\Delta^m}\|_{\Delta^m}}{\mathbf{h}^{m_e}}.$$

Diese Werte werden für jedes weitere durch sukzessives Halbieren der Intervalle entstandene Gitter berechnet.
Da diese Testläufe auch für nicht äquidistante Anfangsgitter durchgeführt wurden, ist vor den tabellarisch angeführten Ergebnissen das erste Gitter graphisch dargestellt.

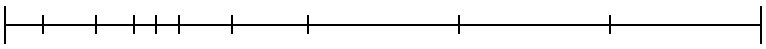
Beispiel (5.8) mit verschiedenen Startgittern (m=4):



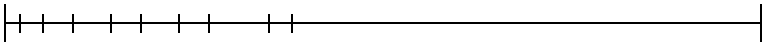
h	$\|\eta_{\Delta^m}\|_{\Delta^m}$	m_e	c_e
2.00E-001	1.23E+000	0.00E+000	0.00E+000
1.00E-001	2.62E-001	2.23E+000	4.47E+001
5.00E-002	4.77E-003	5.78E+000	1.58E+005
2.50E-002	1.87E-004	4.67E+000	5.68E+003
1.25E-002	1.02E-005	4.20E+000	1.01E+003
6.25E-003	6.21E-007	4.04E+000	4.88E+002
3.13E-003	3.85E-008	4.01E+000	4.34E+002
1.56E-003	2.40E-009	4.00E+000	4.14E+002



h	$\|\eta_{\Delta^m}\|_{\Delta^m}$	m_e	c_e
5.00E-001	8.41E+000	0.00E+000	0.00E+000
2.50E-001	2.19E+000	1.94E+000	3.23E+001
1.25E-001	3.07E-001	2.83E+000	1.11E+002
6.25E-002	2.34E-002	3.71E+000	6.95E+002
3.12E-002	5.52E-004	5.40E+000	7.53E+004
1.56E-002	2.60E-005	4.41E+000	2.37E+003
7.81E-003	1.53E-006	4.09E+000	6.38E+002
3.91E-003	9.39E-008	4.02E+000	4.58E+002
1.95E-003	5.85E-009	4.00E+000	4.13E+002

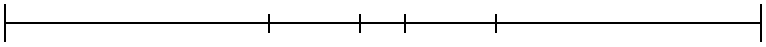


h	$\|\eta_{\Delta^m}\|_{\Delta^m}$	m_e	c_e
2.00E-001	2.87E-002	0.00E+000	0.00E+000
1.00E-001	2.94E-003	3.29E+000	5.67E+000
5.00E-002	1.56E-004	4.23E+000	5.04E+001
2.50E-002	7.50E-006	4.38E+000	7.84E+001
1.25E-002	3.73E-007	4.33E+000	6.50E+001
6.25E-003	2.21E-008	4.08E+000	2.14E+001
3.13E-003	1.35E-009	4.03E+000	1.67E+001
1.56E-003	8.44E-011	4.01E+000	1.46E+001

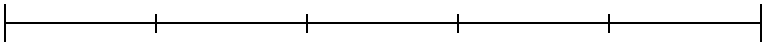


h	$\|\eta_{\Delta^m}\|_{\Delta^m}$	m_e	c_e
6.20E-001	5.37E-001	0.00E+000	0.00E+000
3.10E-001	1.95E-001	1.46E+000	1.08E+000
1.55E-001	2.61E-002	2.90E+000	5.78E+000
7.75E-002	1.54E-003	4.09E+000	5.36E+001
3.88E-002	6.76E-005	4.50E+000	1.55E+002
1.94E-002	2.96E-006	4.51E+000	1.59E+002
9.69E-003	1.59E-007	4.22E+000	4.96E+001
4.84E-003	9.61E-009	4.05E+000	2.30E+001

Beispiel (5.10) mit zwei verschiedenen Startgittern (m=4):



h	$\|\eta_{\Delta^m}\|_{\Delta^m}$	m_e	c_e
3.50E-001	1.86E+000	0.00E+000	0.00E+000
1.75E-001	5.88E-001	1.66E+000	1.07E+001
8.75E-002	3.69E-002	3.99E+000	6.19E+002
4.38E-002	1.54E-003	4.58E+000	2.60E+003
2.19E-002	7.00E-005	4.46E+000	1.77E+003
1.09E-002	2.87E-006	4.61E+000	3.13E+003
5.47E-003	1.38E-007	4.38E+000	1.13E+003
2.73E-003	7.90E-009	4.12E+000	2.94E+002



h	$\|\eta_{\Delta^m}\|_{\Delta^m}$	m_e	c_e
2.00E-001	6.41E+000	0.00E+000	0.00E+000
1.00E-001	1.35E+000	2.25E+000	2.40E+002
5.00E-002	2.51E-002	5.74E+000	7.41E+005
2.50E-002	1.74E-003	3.86E+000	2.62E+003
1.25E-002	7.19E-005	4.59E+000	3.98E+004
6.25E-003	3.30E-006	4.44E+000	2.06E+004
3.13E-003	1.85E-007	4.15E+000	4.75E+003
1.56E-003	1.12E-008	4.04E+000	2.49E+003
7.81E-004	6.97E-010	4.01E+000	2.04E+003

6.2 Startgitterberechnungen

Wir fassen nun die numerischen Ergebnisse zu den Auswertungen auf den Gittern $\bar{\Delta}_{1,j}$ zusammen. Dabei wird demonstriert, dass bei der Auswertung auf dem Startgitter $\bar{\Delta}_{1,1}$ gemäß Tabelle 4.4 für Beispiele mit einer glatten Lösung meistens die Toleranz sofort erfüllt wird. Hingegen sind bei Beispielen mit unglatter Lösungsstruktur Basisgitterverfeinerungen und anschließende Gleichverteilung notwendig.
Die Tabellen beschreiben, wie oft für ein Beispiel an einer bestimmten Stelle des Algorithmus die Phase 1 beendet ist. Dabei wurden vorerst nur die (m,aTOL)-Paare behandelt, die in der Tabelle 4.4 hervorgehoben sind, also automatisch berechnet wurden. Insgesamt handelt es sich um 12 Testläufe. Die Zahlen geben an, wie oft die folgenden Fälle eintreten[1]:

- $\bar{\Delta}_{1,1}^m = \bar{\Delta}_1^m, \quad TOL_q < 1$

  Schon auf dem ersten Gitter wird die Toleranzanforderung erfüllt, sodass der Algorithmus beendet wird.

- $\bar{\Delta}_{1,1}^m = \bar{\Delta}_1^m, \quad \|\varepsilon_{\bar{\Delta}_1^m}\|_{\bar{\Delta}_1^m} < \|p_{\bar{\Delta}_1^m}\|_{\bar{\Delta}_1^m}$

  Nach der Auswertung auf dem ersten Gitter wird der Fehler gleichverteilt.

- $\bar{\Delta}_{1,2}^m = \bar{\Delta}_1^m, \quad TOL_q < 1$

  Das nach (4.15) verfeinerte Gitter erfüllt die Toleranzanforderung, sodass der Algorithmus beendet wird.

- $\bar{\Delta}_{1,2}^m = \bar{\Delta}_1^m, \quad \|\varepsilon_{\bar{\Delta}_{1,2}^m}\|_{\bar{\Delta}_{1,2}^m} < \|p_{\bar{\Delta}_{1,2}^m}\|_{\bar{\Delta}_{1,2}^m}$

  Nach der Auswertung auf dem nach (4.15) verfeinerten Gitter wird der Fehler gleichverteilt.

- $\bar{\Delta}_{1,j}^m = \bar{\Delta}_1^m, \quad TOL_q < 1$

  Die Auswertung auf dem durch Einfügen von $\lceil \frac{N(\bar{\Delta}_{1,i-1})}{2} \rceil$ Intervallen entstandenen Gitter erfüllt die Toleranzanforderung, sodass der Algorithmus beendet wird.

- $\bar{\Delta}_{1,j}^m = \bar{\Delta}_1^m, \quad \|\varepsilon_{\bar{\Delta}_{1,j}^m}\|_{\bar{\Delta}_{1,j}^m} < \|p_{\bar{\Delta}_{1,j}^m}\|_{\bar{\Delta}_{1,j}^m}$

  Nach der Auswertung auf dem durch Einfügen von $\lceil \frac{N(\bar{\Delta}_{1,i-1})}{2} \rceil$ Intervallen entstandenen Gitter wird der Fehler gleichverteilt.

[1] Diese Fälle sind genau die weiterführenden Verzweigungen im Struktogramm 4.7.

Beispiel (5.21):

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$
9	3

Tabelle 6.1: Basisgitterberechnungen für das Beispiel (5.21) mit der automatischen Berechnung von m

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$
31	17

Tabelle 6.2: Basisgitterberechnungen für das Beispiel (5.21)

Da sich die Fälle ausschliessen, ergibt sich als Summe die Anzahl der Testläufe. Die zweite Tabelle für das jeweilige Beispiel zeigt die Zusammenfassung für alle (m,aTOL)-Paare, die in der Tabelle 4.4 zu sehen sind. Die auftretenden Ergebnisse sind wie vorher zu verstehen, insgesamt ergeben sich hier 48 Testläufe für ein Beispiel.

Beispiel (5.17):

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$
12	0

Tabelle 6.3: Basisgitterberechnungen für das Beispiel (5.17) mit der automatischen Berechnung von m

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$
37	11

Tabelle 6.4: Basisgitterberechnungen für das Beispiel (5.17)

Beispiel (5.18):

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$
12	0

Tabelle 6.5: Basisgitterberechnungen für das Beispiel (5.18) mit der automatischen Berechnung von m

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$
42	6

Tabelle 6.6: Basisgitterberechnungen für das Beispiel (5.18)

Beispiel (5.19):

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$
12	0

Tabelle 6.7: Basisgitterberechnungen für das Beispiel (5.19) mit der automatischen Berechnung von m

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$
37	11

Tabelle 6.8: Basisgitterberechnungen für das Beispiel (5.19)

Beispiel (5.20):

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$
12	0

Tabelle 6.9: Basisgitterberechnungen für das Beispiel (5.20) mit der automatischen Berechnung von m

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$
42	6

Tabelle 6.10: Basisgitterberechnungen für das Beispiel (5.20)

Beispiel (5.12):

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$
0	12

Tabelle 6.11: Basisgitterberechnungen für das Beispiel (5.12) mit der automatischen Berechnung von m

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$
7	41

Tabelle 6.12: Basisgitterberechnungen für das Beispiel (5.12)

Beispiel (5.13):

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$
12	0

Tabelle 6.13: Basisgitterberechnungen für das Beispiel (5.13) mit der automatischen Berechnung von m

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$
39	9

Tabelle 6.14: Basisgitterberechnungen für das Beispiel (5.13)

Beispiel (5.14):

$\bar{\Delta}_{1,1}^m = \bar{\Delta}_1^m$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}_1^m}\|_{\bar{\Delta}_1^m} < \|p_{\bar{\Delta}_1^m}\|_{\bar{\Delta}_1^m}$
12	0

Tabelle 6.15: Basisgitterberechnungen für das Beispiel (5.14) mit der automatischen Berechnung von m

$\bar{\Delta}_{1,1}^m = \bar{\Delta}_1^m$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}_1^m}\|_{\bar{\Delta}_1^m} < \|p_{\bar{\Delta}_1}\|_{\bar{\Delta}_1}$
39	9

Tabelle 6.16: Basisgitterberechnungen für das Beispiel (5.14)

Beispiel (5.15):

$\bar{\Delta}_{1,1}^m = \bar{\Delta}_1^m$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}_1^m}\|_{\bar{\Delta}_1^m} < \|p_{\bar{\Delta}_1^m}\|_{\bar{\Delta}_1^m}$
12	0

Tabelle 6.17: Basisgitterberechnungen für das Beispiel (5.15) mit der automatischen Berechnung von m

$\bar{\Delta}_{1,1}^m = \bar{\Delta}_1^m$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}_1^m}\|_{\bar{\Delta}_1^m} < \|p_{\bar{\Delta}_1^m}\|_{\bar{\Delta}_1^m}$
37	11

Tabelle 6.18: Basisgitterberechnungen für das Beispiel (5.15)

Beispiel (5.16):

$\bar{\Delta}_{1,1}^m = \bar{\Delta}_1^m$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}_1^m}\|_{\bar{\Delta}_1^m} < \|p_{\bar{\Delta}_1^m}\|_{\bar{\Delta}_1^m}$
9	3

Tabelle 6.19: Basisgitterberechnungen für das Beispiel (5.16) mit der automatischen Berechnung von m

$\bar{\Delta}_{1,1}^m = \bar{\Delta}_1^m$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}_1^m}\|_{\bar{\Delta}_1^m} < \|p_{\bar{\Delta}_1^m}\|_{\bar{\Delta}_1^m}$
30	18

Tabelle 6.20: Basisgitterberechnungen für das Beispiel (5.16)

Beispiel (5.2):

$\bar{\Delta}_{1,1}^m = \bar{\Delta}_1^m$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}_1^m}\|_{\bar{\Delta}_1^m} < \|p_{\bar{\Delta}_1^m}\|_{\bar{\Delta}_1^m}$
2	10

Tabelle 6.21: Basisgitterberechnungen für das Beispiel (5.2) mit der automatischen Berechnung von m

$\bar{\Delta}_{1,1}^m = \bar{\Delta}_1^m$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}_1^m}\|_{\bar{\Delta}_1^m} < \|p_{\bar{\Delta}_1^m}\|_{\bar{\Delta}_1^m}$
9	39

Tabelle 6.22: Basisgitterberechnungen für das Beispiel (5.2)

Beispiel (5.3):

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$		$\bar{\Delta}^m_{1,2} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$	$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_{1,2}}\|_{\bar{\Delta}^m_{1,2}} < \|p_{\bar{\Delta}^m_{1,2}}\|_{\bar{\Delta}^m_{1,2}}$
0	8	0	4

Tabelle 6.23: Basisgitterberechnungen für das Beispiel (5.3) mit der automatischen Berechnung von m

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$		$\bar{\Delta}^m_{1,2} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$	$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_{1,2}}\|_{\bar{\Delta}^m_{1,2}} < \|p_{\bar{\Delta}^m_{1,2}}\|_{\bar{\Delta}^m_{1,2}}$
0	38	3	7

Tabelle 6.24: Basisgitterberechnungen für das Beispiel (5.3)

Beispiel (5.4):

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$		$\bar{\Delta}^m_{1,2} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$	$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_{1,2}}\|_{\bar{\Delta}^m_{1,2}} < \|p_{\bar{\Delta}^m_{1,2}}\|_{\bar{\Delta}^m_{1,2}}$
0	5	0	5

$\bar{\Delta}^m_{1,4} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_{1,4}}\|_{\bar{\Delta}^m_{1,4}} < \|p_{\bar{\Delta}^m_{1,4}}\|_{\bar{\Delta}^m_{1,4}}$
0	2

Tabelle 6.25: Basisgitterberechnungen für das Beispiel (5.4) mit der automatischen Berechnung von m

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$		$\bar{\Delta}^m_{1,2} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$	$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_{1,2}}\|_{\bar{\Delta}^m_{1,2}} < \|p_{\bar{\Delta}^m_{1,2}}\|_{\bar{\Delta}^m_{1,2}}$
0	35	0	9

$\bar{\Delta}^m_{1,4} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_{1,4}}\|_{\bar{\Delta}^m_{1,4}} < \|p_{\bar{\Delta}^m_{1,4}}\|_{\bar{\Delta}^m_{1,4}}$
0	4

Tabelle 6.26: Basisgitterberechnungen für das Beispiel (5.4)

Beispiel (5.11):

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$
12	0

Tabelle 6.27: Basisgitterberechnungen für das Beispiel (5.11) mit der automatischen Berechnung von m

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$
35	13

Tabelle 6.28: Basisgitterberechnungen für das Beispiel (5.11)

Beispiel (5.8):

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$		$\bar{\Delta}^m_{1,2} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$	$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_{1,2}}\|_{\bar{\Delta}^m_{1,2}} < \|p_{\bar{\Delta}^m_{1,2}}\|_{\bar{\Delta}^m_{1,2}}$
0	9	2	1

Tabelle 6.29: Basisgitterberechnungen für das Beispiel (5.8) bei automatischer Berechnung von m

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$		$\bar{\Delta}^m_{1,2} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$	$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_{1,2}}\|_{\bar{\Delta}^m_{1,2}} < \|p_{\bar{\Delta}^m_{1,2}}\|_{\bar{\Delta}^m_{1,2}}$
0	33	13	2

Tabelle 6.30: Basisgitterberechnungen für das Beispiel (5.8)

Beispiel (5.7):

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$		$\bar{\Delta}^m_{1,2} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$	$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_{1,2}}\|_{\bar{\Delta}^m_{1,2}} < \|p_{\bar{\Delta}^m_{1,2}}\|_{\bar{\Delta}^m_{1,2}}$
0	2	3	4

$\bar{\Delta}^m_{1,3} = \bar{\Delta}^m_1$		$\bar{\Delta}^m_{1,4} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_{1,3}}\|_{\bar{\Delta}^m_{1,3}} < \|p_{\bar{\Delta}^m_{1,3}}\|_{\bar{\Delta}^m_{1,3}}$	$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_{1,4}}\|_{\bar{\Delta}^m_{1,4}} < \|p_{\bar{\Delta}^m_{1,4}}\|_{\bar{\Delta}^m_{1,4}}$
0	1	1	0

$\bar{\Delta}^m_{1,5} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_{1,5}}\|_{\bar{\Delta}^m_{1,5}} < \|p_{\bar{\Delta}^m_{1,5}}\|_{\bar{\Delta}^m_{1,5}}$
1	0

Tabelle 6.31: Basisgitterberechnungen für das Beispiel (5.7) bei automatischer Berechnung von m

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$		$\bar{\Delta}^m_{1,2} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$	$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_{1,2}}\|_{\bar{\Delta}^m_{1,2}} < \|p_{\bar{\Delta}^m_{1,2}}\|_{\bar{\Delta}^m_{1,2}}$
0	22	10	8

$\bar{\Delta}^m_{1,3} = \bar{\Delta}^m_1$		$\bar{\Delta}^m_{1,4} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_{1,3}}\|_{\bar{\Delta}^m_{1,3}} < \|p_{\bar{\Delta}^m_{1,3}}\|_{\bar{\Delta}^m_{1,3}}$	$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_{1,4}}\|_{\bar{\Delta}^m_{1,4}} < \|p_{\bar{\Delta}^m_{1,4}}\|_{\bar{\Delta}^m_{1,4}}$
4	1	3	0

$\bar{\Delta}^m_{1,5} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_{1,5}}\|_{\bar{\Delta}^m_{1,5}} < \|p_{\bar{\Delta}^m_{1,5}}\|_{\bar{\Delta}^m_{1,5}}$
1	0

Tabelle 6.32: Basisgitterberechnungen für das Beispiel (5.7)

Beispiel (5.9):

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$		$\bar{\Delta}^m_{1,2} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$	$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_{1,2}}\|_{\bar{\Delta}^m_{1,2}} < \|p_{\bar{\Delta}^m_{1,2}}\|_{\bar{\Delta}^m_{1,2}}$
0	11	1	0

Tabelle 6.33: Basisgitterberechnungen für das Beispiel (5.9) bei automatischer Berechnung von m

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$		$\bar{\Delta}^m_{1,2} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$	$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_{1,2}}\|_{\bar{\Delta}^m_{1,2}} < \|p_{\bar{\Delta}^m_{1,2}}\|_{\bar{\Delta}^m_{1,2}}$
5	42	1	0

Tabelle 6.34: Basisgitterberechnungen für das Beispiel (5.9)

Beispiel (5.10):

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$		$\bar{\Delta}^m_{1,2} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$	$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_{1,2}}\|_{\bar{\Delta}^m_{1,2}} < \|p_{\bar{\Delta}^m_{1,2}}\|_{\bar{\Delta}^m_{1,2}}$
0	4	3	4

$\bar{\Delta}^m_{1,4} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_{1,4}}\|_{\bar{\Delta}^m_{1,4}} < \|p_{\bar{\Delta}^m_{1,4}}\|_{\bar{\Delta}^m_{1,4}}$
1	0

Tabelle 6.35: Basisgitterberechnungen für das Beispiel (5.10) bei automatischer Berechnung von m

$\bar{\Delta}^m_{1,1} = \bar{\Delta}^m_1$		$\bar{\Delta}^m_{1,2} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1} < \|p_{\bar{\Delta}^m_1}\|_{\bar{\Delta}^m_1}$	$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_{1,2}}\|_{\bar{\Delta}^m_{1,2}} < \|p_{\bar{\Delta}^m_{1,2}}\|_{\bar{\Delta}^m_{1,2}}$
0	26	15	6

$\bar{\Delta}^m_{1,4} = \bar{\Delta}^m_1$	
$TOL_q < 1$	$\|\varepsilon_{\bar{\Delta}^m_{1,4}}\|_{\bar{\Delta}^m_{1,4}} < \|p_{\bar{\Delta}^m_{1,4}}\|_{\bar{\Delta}^m_{1,4}}$
1	0

Tabelle 6.36: Basisgitterberechnungen für das Beispiel (5.10)

6.3 m-TOL - Tabelle

Die m-TOL-Tabelle dient zur automatischen Ermittlung von m bei gegebener Toleranz. Die hier angeführten numerischen Resultate wurden wie folgt gewonnen.
Für alle Testläufe wird aTOL=rTOL gesetzt. Anschließend werden alle m-TOL-Paare aus der Tabelle 4.4 der Reihe nach untersucht. Für jedes m-TOL-Paar wird der Algorithmus mit folgenden Standardparametern ausgeführt:

`RECOVER_PEAKS`=1, `SMOOTHING_FACTOR`=5%.

Ist die Toleranz unterschritten werden die Berechnungen beendet. Gemessen wird dabei die Gesamtlaufzeit in Sekunden[2]. Diese Zeit wird in den folgenden Tabellen mit *time* bezeichnet. Zu beachten ist, dass während der Zeitmessung keine zeitaufwändigen Routinen (z.B. Ausgabe der Lösung) ausgeführt werden, die dem Algorithmus nicht direkt zugeordnet sind.
Wegen der sehr langen Rechenzeiten werden die Ergebnisse für die Paare

$$\begin{aligned} m = 2, &\quad aTOL < 1E-4, \\ m = 4, &\quad aTOL < 1E-7, \\ m = 6, &\quad aTOL < 1E-10, \end{aligned}$$

nicht berücksichtigt.
Die numerischen Ergebnisse werden in den folgenden Tabellen zusammengefasst. Dabei ist in der rechten Spalte die kürzeste Zeit

$$\min_{m\in\{2,4,6,8\}} time$$

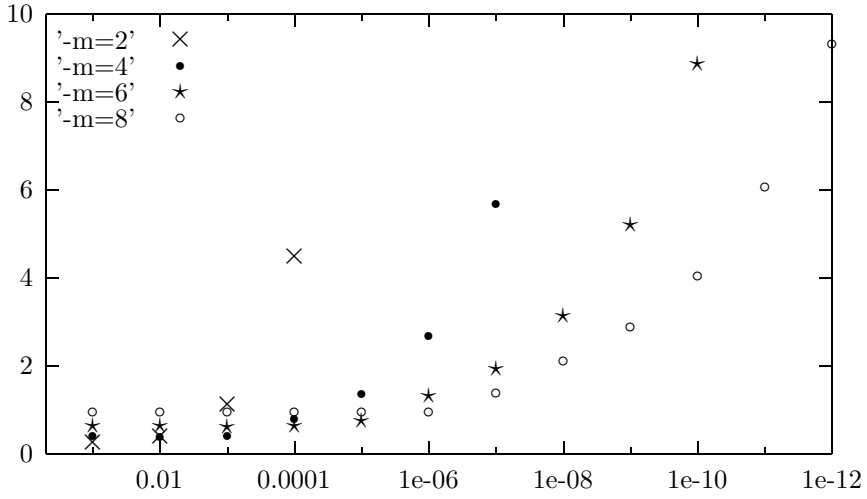
für die jeweilige Toleranz wiederholt. Außerdem ist dieser Wert in der jeweiligen Zeile gesondert hervorgehoben, um das zugehörige m besser erkennen zu können. Die Abbildungen unter den Tabellen geben Überblick über die beste Wahl von m für die vorgegebene Toleranz.
Insgesamt zeigt sich, dass für gröbere Toleranzen eine kleinere Ordnung m sich bewährt, bei sehr strengen Toleranzen jedoch die Wahl von $m = 8$ immer die Beste ist. Insgesamt kann die Wahl von m bei gegebener Toleranz gemäß folgender Tabelle getroffen werden:

aTOL	m
$\in [1E\text{-}1, 1E\text{-}2]$	2
$\in (1E\text{-}2, 1E\text{-}4]$	4
$\in (1E\text{-}4, 1E\text{-}6]$	6
$\in (1E\text{-}6, 1E\text{-}12]$	8

[2]In MATLAB kann dieser Wert durch das Kommando `CPUTIME` ermittelt werden.

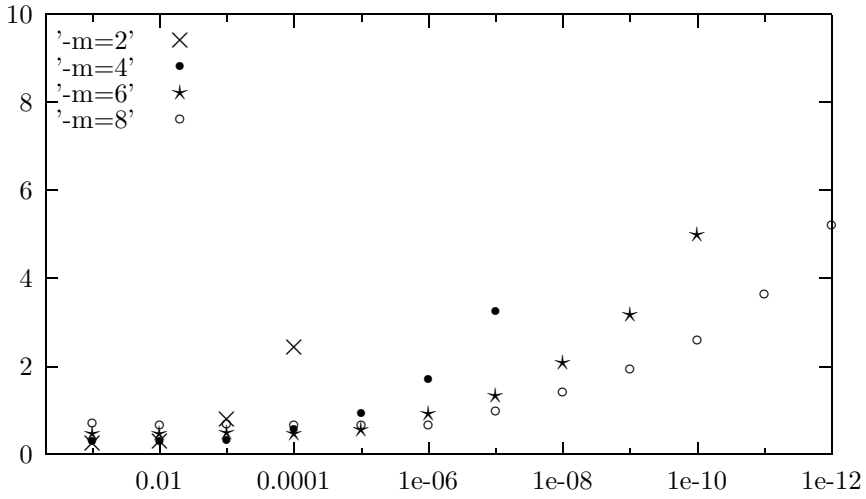
Beispiel (5.17):

TOL	m=2	m=4	m=6	m=8	min *time*
1.00E-01	**0.26**	0.39	0.62	0.95	0.26
1.00E-02	0.39	**0.38**	0.62	0.94	0.38
1.00E-03	1.14	**0.39**	0.61	0.94	0.39
1.00E-04	4.48	0.79	**0.62**	0.94	0.62
1.00E-05		1.35	**0.76**	0.95	0.76
1.00E-06		2.66	1.33	**0.94**	0.94
1.00E-07		5.68	1.93	**1.37**	1.37
1.00E-08			3.15	**2.09**	2.09
1.00E-09			5.2	**2.88**	2.88
1.00E-10			8.88	**4.05**	4.05
1.00E-11				**6.06**	6.06
1.00E-12				**9.32**	9.32



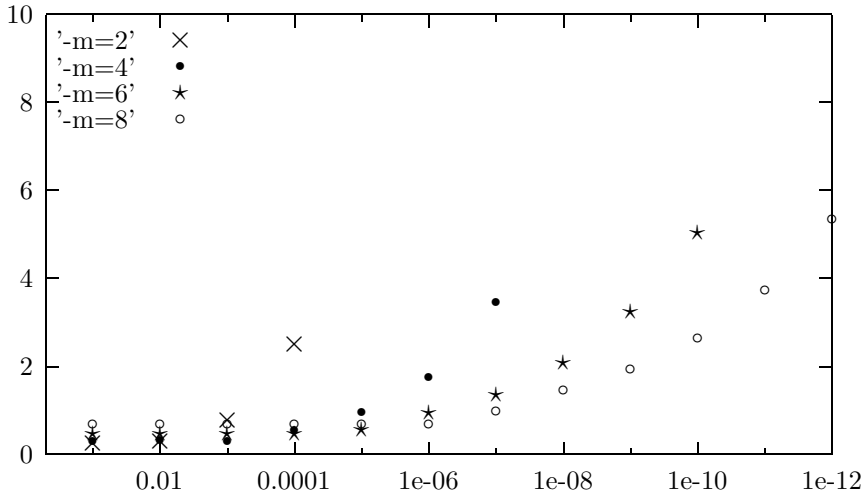
Beispiel (5.18):

TOL	m=2	m=4	m=6	m=8	min *time*
1.00E-01	**0.25**	0.3	0.47	0.71	0.25
1.00E-02	**0.3**	0.3	0.47	0.67	0.3
1.00E-03	0.8	**0.33**	0.49	0.7	0.33
1.00E-04	2.45	0.57	**0.47**	0.67	0.47
1.00E-05		0.95	**0.55**	0.67	0.55
1.00E-06		1.72	0.92	**0.67**	0.67
1.00E-07		3.26	1.32	**0.98**	0.98
1.00E-08			2.09	**1.43**	1.43
1.00E-09			3.17	**1.93**	1.93
1.00E-10			5.01	**2.59**	2.59
1.00E-11				**3.66**	3.66
1.00E-12				**5.23**	5.23



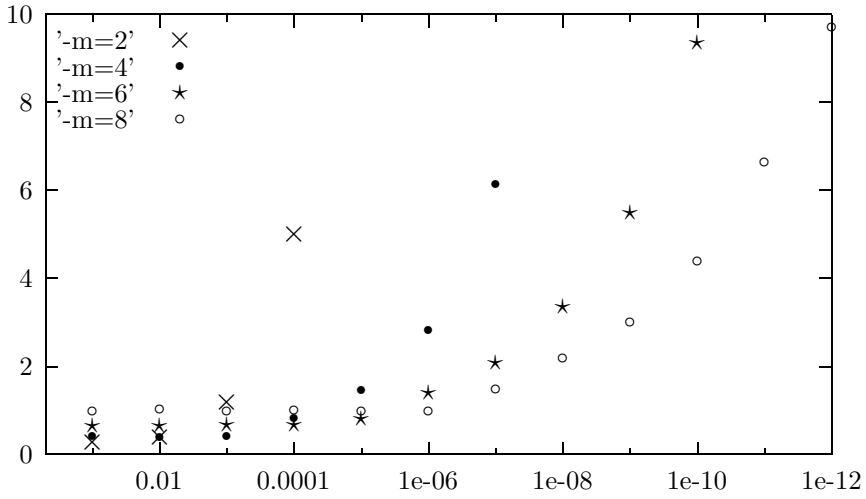
Beispiel (5.19):

TOL	m=2	m=4	m=6	m=8	min *time*
1.00E-01	**0.25**	0.3	0.48	0.69	0.25
1.00E-02	**0.3**	0.34	0.47	0.68	0.3
1.00E-03	0.78	**0.29**	0.47	0.68	0.29
1.00E-04	2.51	0.56	**0.48**	0.68	0.48
1.00E-05		0.96	**0.57**	0.68	0.57
1.00E-06		1.76	0.94	**0.7**	0.7
1.00E-07		3.47	1.36	**0.99**	0.99
1.00E-08			2.08	**1.45**	1.45
1.00E-09			3.25	**1.95**	1.95
1.00E-10			5.03	**2.63**	2.63
1.00E-11				**3.72**	3.72
1.00E-12				**5.34**	5.34



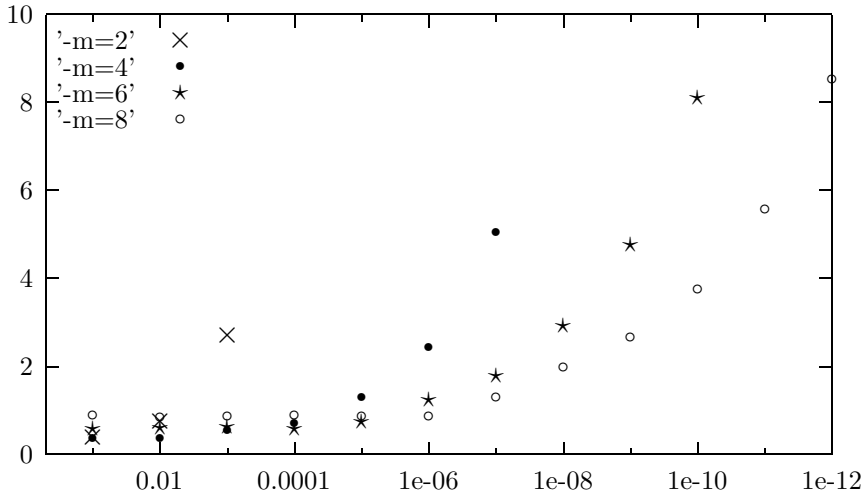
Beispiel (5.20):

TOL	m=2	m=4	m=6	m=8	min *time*
1.00E-01	**0.27**	0.41	0.66	0.99	0.27
1.00E-02	**0.4**	0.4	0.66	1.02	0.4
1.00E-03	1.2	**0.42**	0.68	0.98	0.42
1.00E-04	5.01	0.81	**0.67**	1.01	0.67
1.00E-05		1.47	**0.8**	0.99	0.8
1.00E-06		2.83	1.4	**0.98**	0.98
1.00E-07		6.15	2.09	**1.49**	1.49
1.00E-08			3.36	**2.2**	2.2
1.00E-09			5.5	**3.02**	3.02
1.00E-10			9.37	**4.4**	4.4
1.00E-11				**6.65**	6.65
1.00E-12				**9.73**	9.73



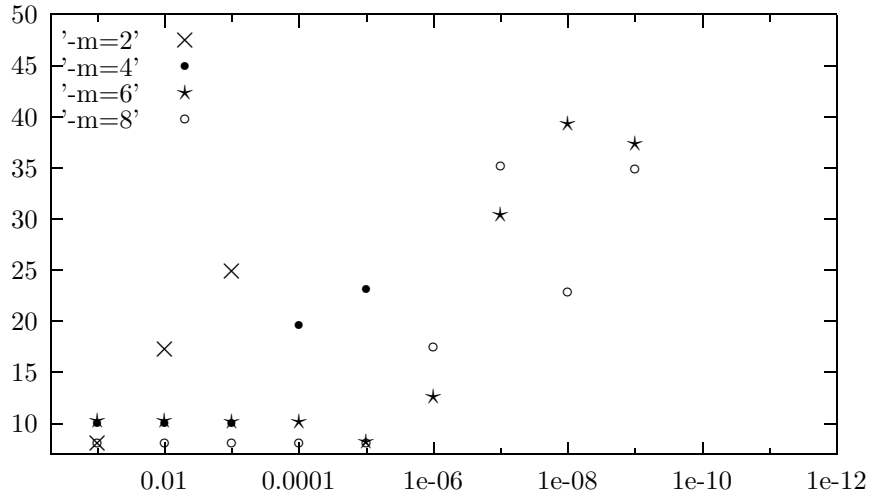
Beispiel (5.21):

TOL	m=2	m=4	m=6	m=8	min *time*
1.00E-01	0.39	**0.37**	0.58	0.89	0.37
1.00E-02	0.75	**0.36**	0.59	0.86	0.36
1.00E-03	2.71	**0.55**	0.63	0.87	0.55
1.00E-04	12.44	0.71	**0.58**	0.89	0.58
1.00E-05		1.29	**0.74**	0.87	0.74
1.00E-06		2.45	1.23	**0.87**	0.87
1.00E-07		5.07	1.79	**1.3**	1.3
1.00E-08			2.92	**1.98**	1.98
1.00E-09			4.76	**2.66**	2.66
1.00E-10			8.12	**3.76**	3.76
1.00E-11				**5.58**	5.58
1.00E-12				**8.55**	8.55



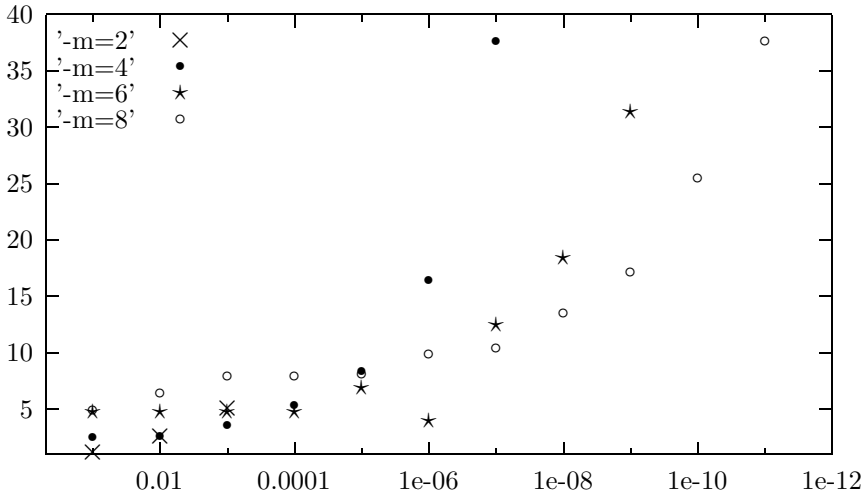
Beispiel (5.7):

TOL	m=2	m=4	m=6	m=8	min *time*
1.00E-01	**8.07**	10.05	10.25	8.07	8.07
1.00E-02	17.25	10.08	10.23	**8.11**	8.11
1.00E-03	24.99	10.08	10.22	**8.09**	8.09
1.00E-04	55.05	19.66	10.2	**8.13**	8.13
1.00E-05		23.18	8.25	**8.06**	8.06
1.00E-06		61.58	**12.66**	17.45	12.66
1.00E-07		85.04	**30.4**	35.23	30.4
1.00E-08			39.31	**22.84**	22.84
1.00E-09			37.39	**34.89**	34.89
1.00E-10			94.1	**54.68**	54.68
1.00E-11				**68.42**	68.42
1.00E-12				**208.72**	208.72



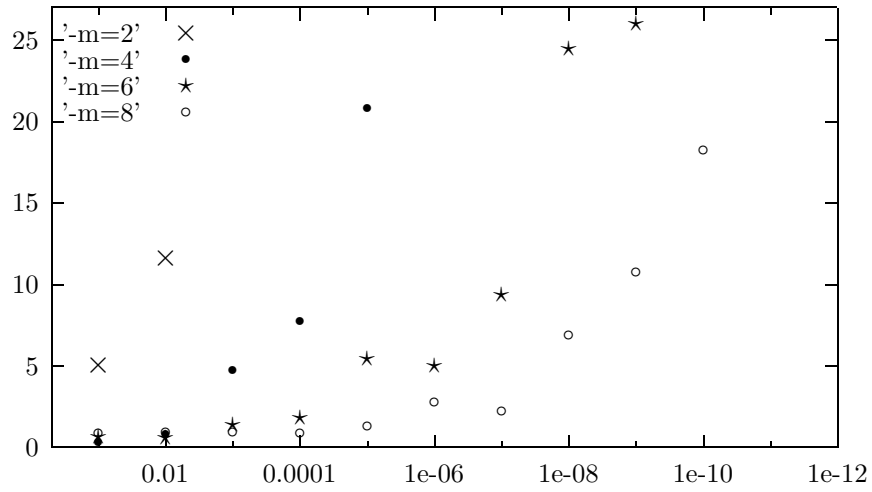
Beispiel (5.8):

TOL	m=2	m=4	m=6	m=8	min *time*
1.00E-01	**1.25**	2.56	4.75	4.9	1.25
1.00E-02	2.58	**2.57**	4.76	6.38	2.57
1.00E-03	5.07	**3.63**	4.76	7.99	3.63
1.00E-04	43.39	5.36	**4.77**	7.97	4.77
1.00E-05	0	8.38	**6.92**	8.1	6.92
1.00E-06	0	16.46	**4.01**	9.85	4.01
1.00E-07	0	37.65	12.45	**10.41**	10.41
1.00E-08	0	0	18.44	**13.5**	13.5
1.00E-09	0	0	31.41	**17.18**	17.18
1.00E-10	0	0	57.5	**25.5**	25.5
1.00E-11	0	0	0	**37.69**	37.69
1.00E-12	0	0	0	**57.9**	57.9



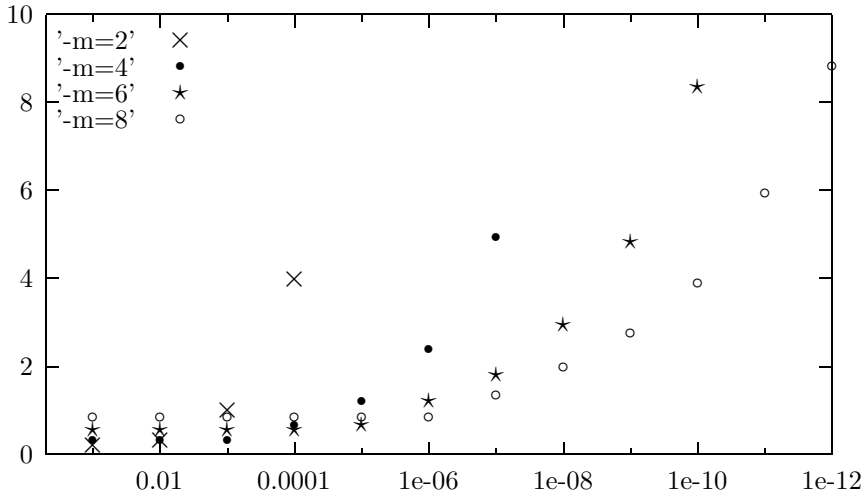
Beispiel (5.12):

TOL	m=2	m=4	m=6	m=8	min *time*
1.00E-01	5.05	**0.36**	0.61	0.92	0.36
1.00E-02	11.62	0.78	**0.6**	0.93	0.6
1.00E-03	73.6	4.72	1.37	**0.95**	0.95
1.00E-04	557.1	7.75	1.84	**0.92**	0.92
1.00E-05		20.82	5.47	**1.31**	1.31
1.00E-06		42.03	5.04	**2.8**	2.8
1.00E-07		204.44	9.41	**2.24**	2.24
1.00E-08			24.48	**6.93**	6.93
1.00E-09			26	**10.77**	10.77
1.00E-10			54.3	**18.22**	18.22
1.00E-11				**41.71**	41.71
1.00E-12				**56.95**	56.95



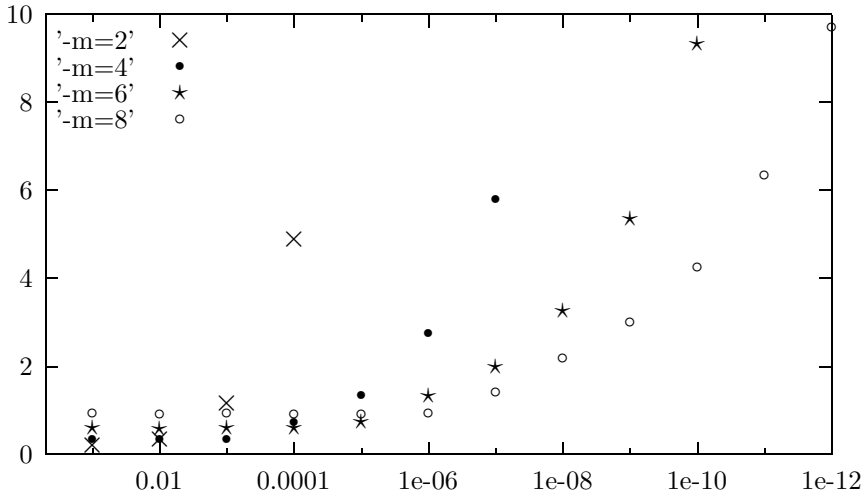
Beispiel (5.13):

TOL	m=2	m=4	m=6	m=8	min *time*
1.00E-01	**0.2**	0.34	0.56	0.86	0.2
1.00E-02	**0.32**	0.32	0.55	0.85	0.32
1.00E-03	1.01	**0.32**	0.55	0.86	0.32
1.00E-04	3.99	0.67	**0.55**	0.85	0.55
1.00E-05	0	1.21	**0.68**	0.86	0.68
1.00E-06	0	2.39	1.22	**0.85**	0.85
1.00E-07	0	4.95	1.82	**1.34**	1.34
1.00E-08	0	0	2.94	**2**	2
1.00E-09	0	0	4.83	**2.75**	2.75
1.00E-10	0	0	8.36	**3.9**	3.9
1.00E-11	0	0	0	**5.94**	5.94
1.00E-12	0	0	0	**8.84**	8.84



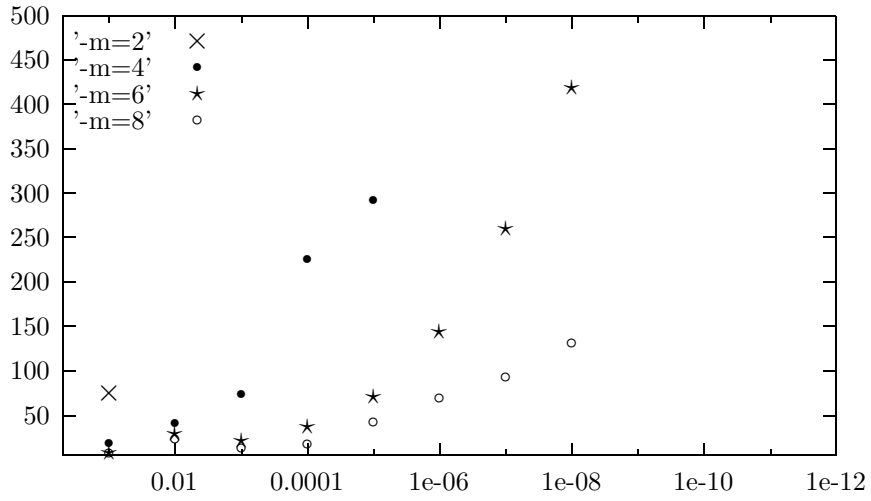
Beispiel (5.14):

TOL	m=2	m=4	m=6	m=8	min *time*
1.00E-01	**0.21**	0.35	0.6	0.93	0.21
1.00E-02	**0.35**	0.35	0.58	0.92	0.35
1.00E-03	1.16	**0.35**	0.61	0.93	0.35
1.00E-04	4.9	0.74	**0.61**	0.92	0.61
1.00E-05		1.36	**0.74**	0.92	0.74
1.00E-06		2.75	1.34	**0.93**	0.93
1.00E-07		5.81	2	**1.41**	1.41
1.00E-08			3.28	**2.19**	2.19
1.00E-09			5.37	**3**	3
1.00E-10			9.35	**4.26**	4.26
1.00E-11				**6.35**	6.35
1.00E-12				**9.73**	9.73



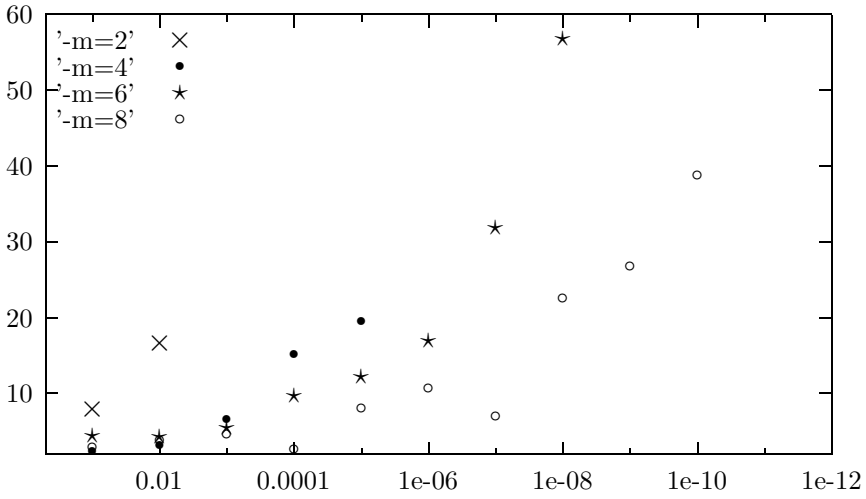
Beispiel (5.4):

TOL	m=2	m=4	m=6	m=8	min *time*
1.00E-01	75.01	18.46	**8.01**	6.91	8.01
1.00E-02	723.08	41.34	29.13	**22.93**	22.93
1.00E-03	0	73.76	21.19	**12.72**	12.72
1.00E-04	0	226	37.77	**18.27**	18.27
1.00E-05	0	292.51	70.98	**42.08**	42.08
1.00E-06	0	1456.39	144.29	**68.86**	68.86
1.00E-07	0	0	260.2	**93.23**	93.23
1.00E-08	0	0	419.61	**131.03**	131.03
1.00E-09	0	0	1110.41	**943.97**	943.97
1.00E-10	0	0	0	**1577.41**	1577.41



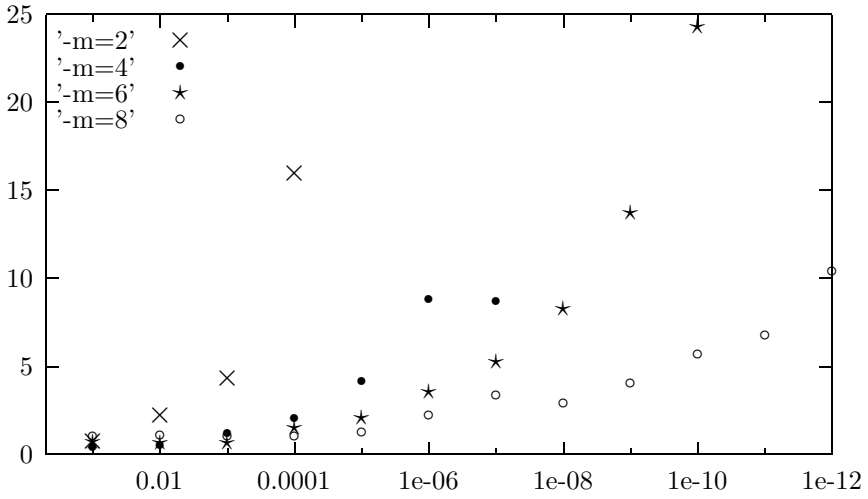
Beispiel (5.3):

TOL	m=2	m=4	m=6	m=8	min *time*
1.00E-01	8.01	**2.42**	4.43	2.9	2.42
1.00E-02	16.67	**3.27**	4.4	3.82	3.27
1.00E-03	109.71	6.67	5.52	**4.71**	4.71
1.00E-04	1166.17	15.21	9.76	**2.67**	2.67
1.00E-05		19.56	12.31	**8.07**	8.07
1.00E-06		61.71	16.93	**10.78**	10.78
1.00E-07		164.14	31.89	**7.04**	7.04
1.00E-08			56.88	**22.57**	22.57
1.00E-09			77.12	**26.84**	26.84
1.00E-10			150.38	**38.79**	38.79
1.00E-11				**325.69**	325.69
1.00E-12				**243.81**	243.81



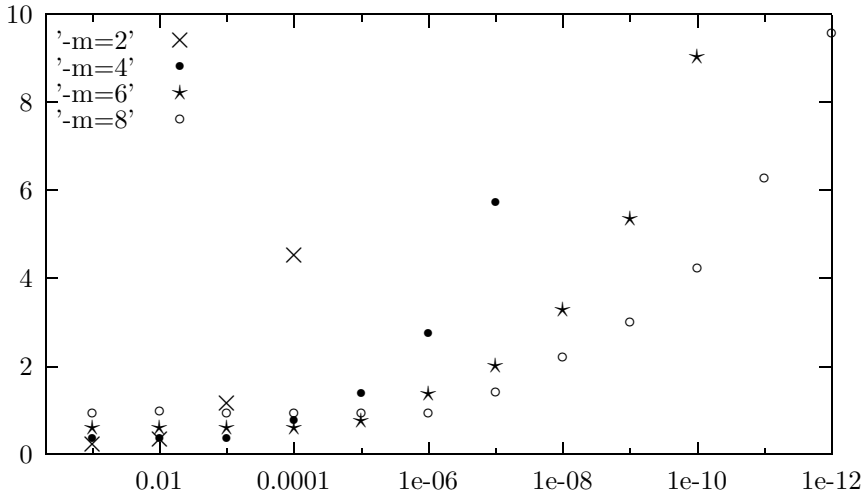
Beispiel (5.2):

TOL	m=2	m=4	m=6	m=8	min *time*
1.00E-01	0.75	**0.43**	0.71	1.03	0.43
1.00E-02	2.22	**0.54**	0.69	1.08	0.54
1.00E-03	4.33	1.2	**0.69**	1.07	0.69
1.00E-04	15.96	2.09	1.51	**1.06**	1.06
1.00E-05		4.18	2.11	**1.26**	1.26
1.00E-06		8.82	3.58	**2.26**	2.26
1.00E-07		8.69	5.28	**3.35**	3.35
1.00E-08			8.26	**2.92**	2.92
1.00E-09			13.71	**4.05**	4.05
1.00E-10			24.36	**5.67**	5.67
1.00E-11				**6.8**	6.8
1.00E-12				**10.45**	10.45



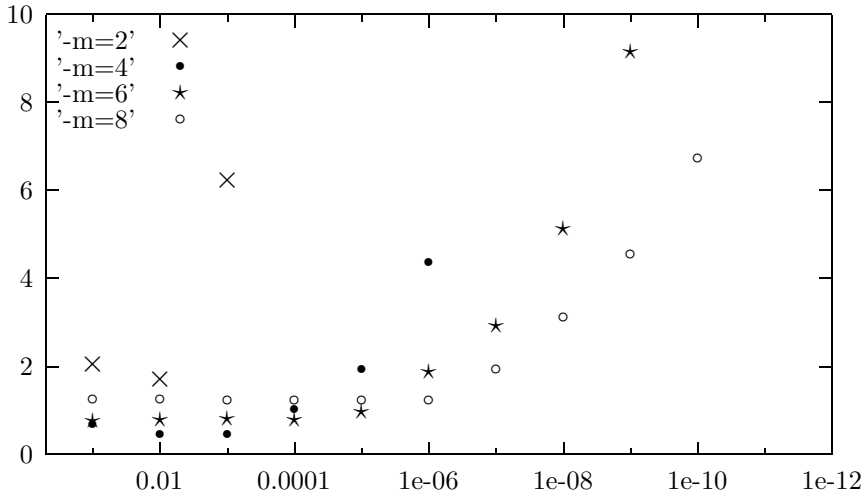
Beispiel (5.15):

TOL	m=2	m=4	m=6	m=8	min *time*
1.00E-01	**0.24**	0.37	0.61	0.94	0.24
1.00E-02	**0.35**	0.36	0.61	0.97	0.35
1.00E-03	1.16	**0.36**	0.61	0.93	0.36
1.00E-04	4.52	0.77	**0.61**	0.94	0.61
1.00E-05		1.39	**0.75**	0.93	0.75
1.00E-06		2.77	1.38	**0.95**	0.95
1.00E-07		5.74	2.01	**1.42**	1.42
1.00E-08			3.29	**2.21**	2.21
1.00E-09			5.35	**3.01**	3.01
1.00E-10			9.04	**4.24**	4.24
1.00E-11				**6.28**	6.28
1.00E-12				**9.58**	9.58



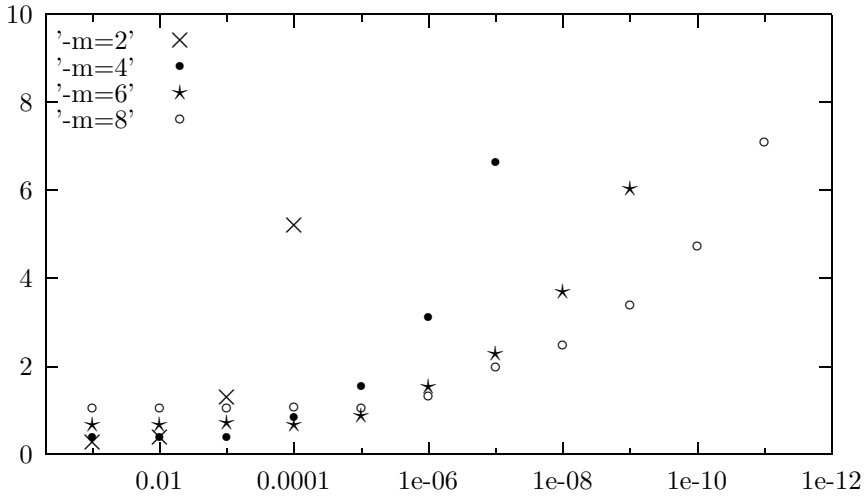
Beispiel (5.16):

TOL	m=2	m=4	m=6	m=8	min *time*
1.00E-01	2.05	**0.68**	0.77	1.25	0.68
1.00E-02	1.7	**0.45**	0.78	1.25	0.45
1.00E-03	6.24	**0.45**	0.8	1.23	0.45
1.00E-04	35.96	1.02	**0.78**	1.23	0.78
1.00E-05		1.95	**0.98**	1.23	0.98
1.00E-06		4.38	1.88	**1.23**	1.23
1.00E-07		10.38	2.93	**1.94**	1.94
1.00E-08			5.12	**3.13**	3.13
1.00E-09			9.17	**4.55**	4.55
1.00E-10			17.46	**6.74**	6.74
1.00E-11				**11.26**	11.26
1.00E-12				**340.36**	340.36



Beispiel (5.11):

TOL	m=2	m=4	m=6	m=8	min *time*
1.00E-01	**0.27**	0.4	0.69	1.05	0.27
1.00E-02	**0.4**	0.4	0.68	1.05	0.4
1.00E-03	1.29	**0.4**	0.71	1.05	0.4
1.00E-04	5.21	0.86	**0.68**	1.07	0.68
1.00E-05		1.56	**0.87**	1.05	0.87
1.00E-06		3.11	1.54	**1.33**	1.33
1.00E-07		6.64	2.29	**2**	2
1.00E-08			3.7	**2.48**	2.48
1.00E-09			6.04	**3.4**	3.4
1.00E-10			10.36	**4.74**	4.74
1.00E-11				**7.11**	7.11
1.00E-12				**10.94**	10.94



6.4 Ermittlung des Glättungsfaktors

Die folgenden Ausführungen motivieren die Wahl des Glättungsfaktors. Der relative Glättungsfaktor `SMOOTHING_FACTOR` wird dabei im Intervall

$$\texttt{SMOOTHING\_FACTOR} \in [0, 50],$$

wobei die Angaben in Prozent zu verstehen sind, variiert. Die Formel für die Anzahl der Nachbarn, über die geglättet wird lautet:

$$s_a := \left\lfloor (N(\bar{\Delta}_1^m) + 1) \cdot \frac{\texttt{SMOOTHING\_FACTOR}}{100} \right\rfloor .$$

Damit der Wert $s_a = 0$ möglich ist, wurde s_a nicht gemäß (4.18) berechnet, wo $s_a \geq 2$ gilt. In der vorliegenden Testreihe wurden die singulären Modelle (5.2), (5.3), (5.8), (5.7), (5.12), (5.16) und das reguläre Problem (5.21) behandelt. Für jeden Wert des Parameters `SMOOTHING_FACTOR`$= 0, 1, 2, \ldots, 50$ wird der Algorithmus sofort nach der Auswertung auf dem Gitter $\bar{\Delta}_2^m$ gestoppt. Die folgenden Auswertungsdaten werden in einer Tabelle zusammengefasst:

- s_{min} ... jener Glättungsfaktor `SMOOTHING_FACTOR`, bei dem $\|\varepsilon_{\bar{\Delta}_2^m}\|_{\bar{\Delta}_2^m}$ am kleinsten ist,
- ε_{min} ... geschätzter Fehler $\|\varepsilon_{\bar{\Delta}_2^m}\|_{\bar{\Delta}_2^m}$ bei s_{min},
- $N(\bar{\Delta}_2)_{s_{min}} + 1$ ... Anzahl der Gitterpunkte des Gitters $\bar{\Delta}_2$ bei s_{min},
- ε_5 ... geschätzter Fehler $\|\varepsilon_{\bar{\Delta}_2^m}\|_{\bar{\Delta}_2^m}$ bei `SMOOTHING_FACTOR` $= 5\%$,
- $N(\bar{\Delta}_2)_{s_5} + 1$ ... Anzahl der Gitterpunkte des Gitters $\bar{\Delta}_2$ bei `SMOOTHING_FACTOR` $= 5\%$.

Da die Berechnung von $N(\bar{\Delta}_2)$ nach (4.26) von I abhängig ist, ist $N(\bar{\Delta}_2)$ auch von `SMOOTHING_FACTOR` abhängig. Daher ist $N(\bar{\Delta}_2)$ sowohl für s_5 als auch für s_{min} angegeben. Für jede Auswertung ist `RECOVER_PEAKS`=1 gesetzt. Die zugehörigen Abbildungen zeigen den Verlauf von $\|\varepsilon_{\bar{\Delta}_2^m}\|_{\bar{\Delta}_2^m}$ in Abhängigkeit von `SMOOTHING_FACTOR`. Die in manchen Abbildungen auftretenden "Plateaus" können folgendermaßen entstehen:

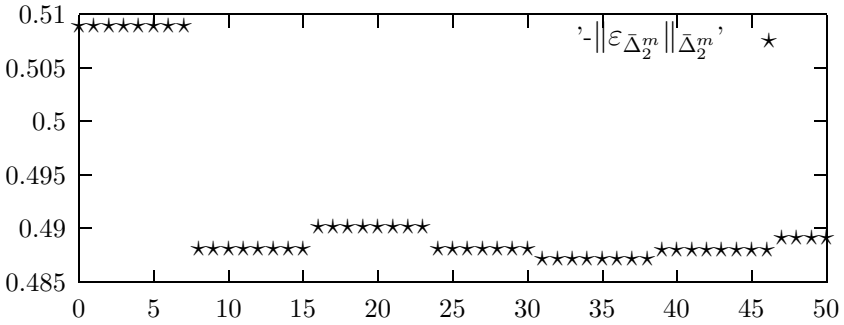
- Für kleine $N(\bar{\Delta}_1)$ ändert sich die Anzahl der Nachbarn s_a, über die gemittelt wird, nicht, auch dann wenn `SMOOTHING_FACTOR` in einem gewissen Bereich variiert. $\|\varepsilon_{\bar{\Delta}_2^m}\|_{\bar{\Delta}_2^m}$ bleibt dann auch konstant.
- Änderung von `SMOOTHING_FACTOR` beeinflusst die Intervallzahl $N(\bar{\Delta}_2)$. In den Abbildungen sind dann "Sprungstellen" zu sehen.

Zunächst sieht man, dass Glättung im Vergleich zu $s_a = 0$ den Fehler fast immer auf ein niedrigeres Niveau bringt. Die numerischen Resultate zeigen, dass häufig `SMOOTHING_FACTOR` $\approx 5\%$ optimale Resultate liefert. Beachtet man die Größenordnung[3] auf der senkrechten Achse, so sieht man, dass die Variation in $\|\varepsilon_{\bar{\Delta}_2^m}\|_{\bar{\Delta}_2^m}$ für verschiedene Werte von `SMOOTHING_FACTOR` sehr klein ist. Wir haben deshalb als Standardwert `SMOOTHING_FACTOR`=5% festgelegt. Damit soll einerseits der Vorteil der Glättung genutzt werden und andererseits die lokale Information möglichst vollständig erhalten bleiben.

Beispiel (5.16):

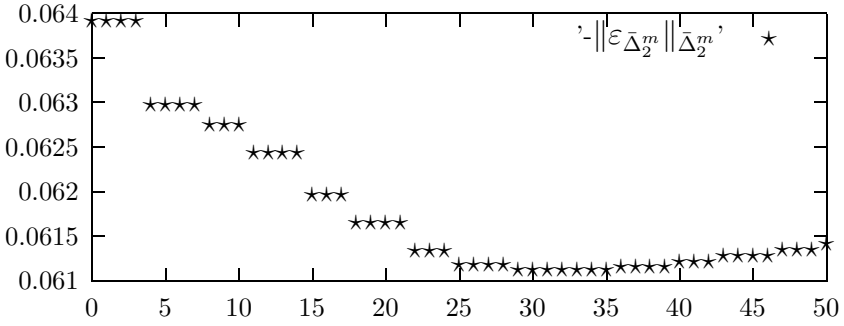
m=2, aTOL=rTOL=1E−1

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
31	4.87E-001	6	5.09E-001	6



m=2, aTOL=rTOL=1E−2

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
29	6.11E-002	15	6.30E-002	15

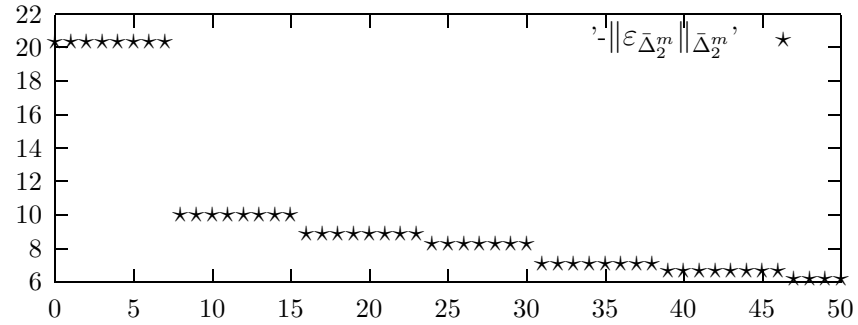


[3]Wir benützen jetzt in den Grafiken lineare Maßstäbe.

Beispiel (5.12):

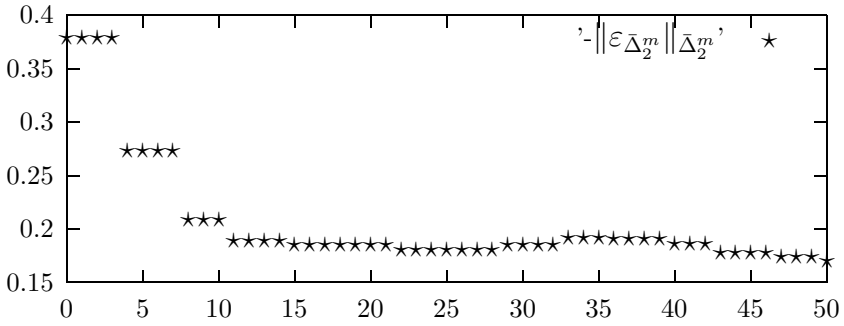
m=2, aTOL=rTOL=1E−1

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
47	6.20E+000	21	2.04E+001	18



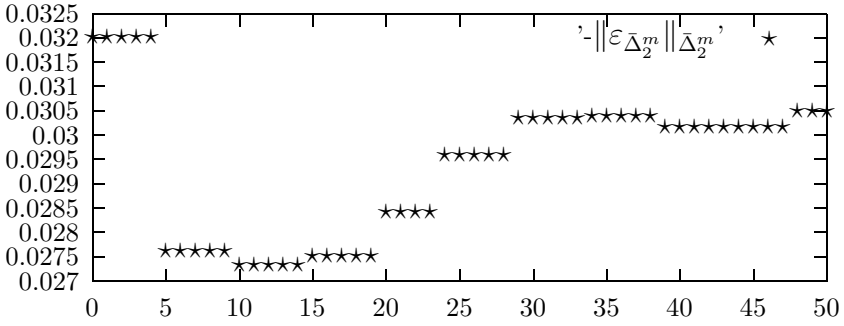
m=2, aTOL=rTOL=1E−2

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
50	1.70E-001	135	2.73E-001	121



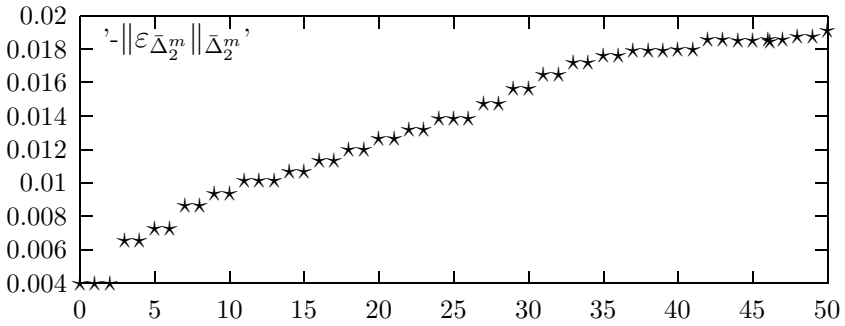
m=4, aTOL=rTOL=1E−3

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
10	2.73E-002	16	2.76E-002	16



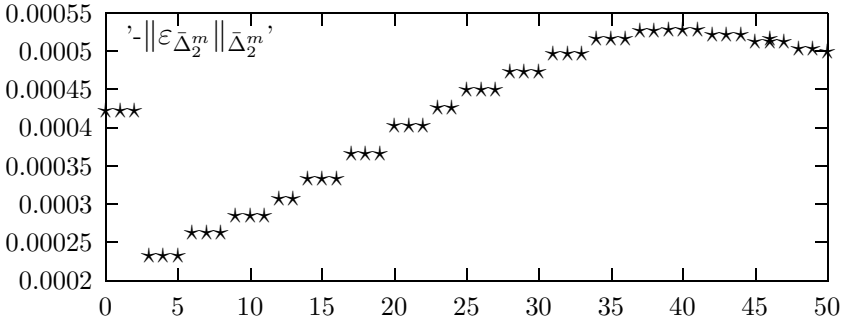
m=4, aTOL=rTOL=1E−4

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
0	4.02E-003	24	7.24E-003	24



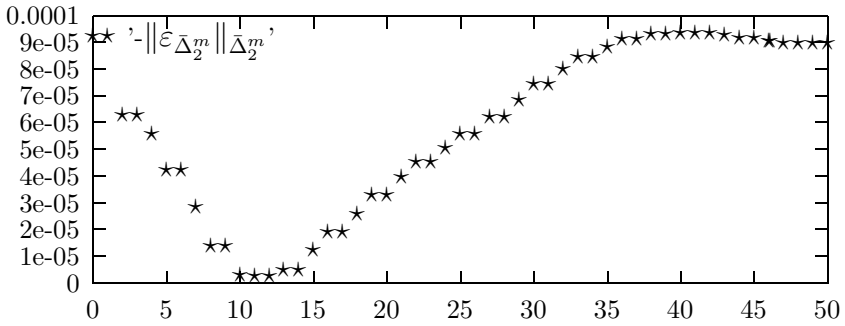
m=6, aTOL=rTOL=1E−5

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
3	2.32E-004	11	2.32E-004	11



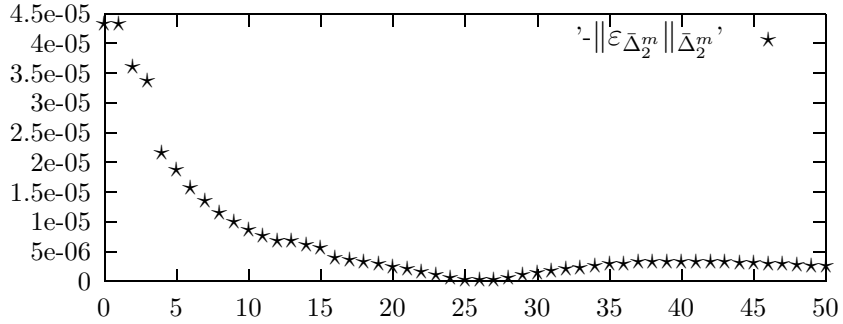
m=6, aTOL=rTOL=1E−6

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
11	2.62E-006	18	4.27E-005	17



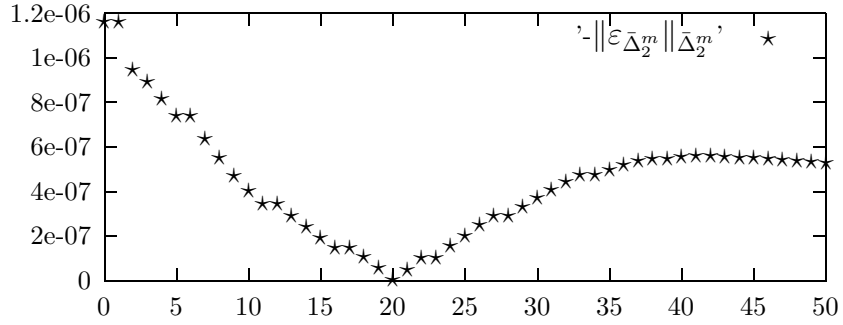
m=6, aTOL=rTOL=1E−7

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
25	2.00E-007	29	1.87E-005	28



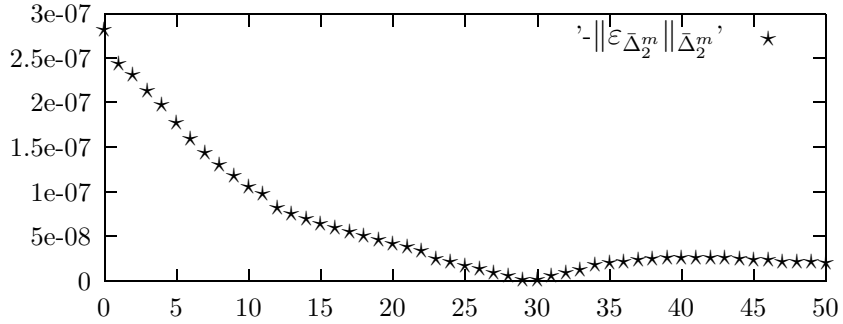
m=8, aTOL=rTOL=1E−8

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
20	4.87E-009	14	7.41E-007	14



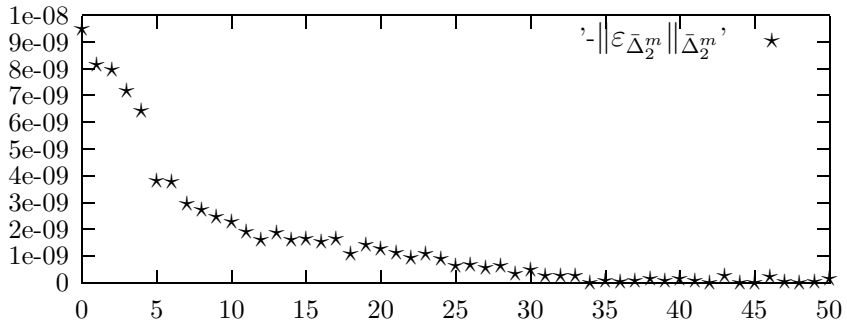
m=8, aTOL=rTOL=1E−9

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
29	1.32E-009	20	1.77E-007	20



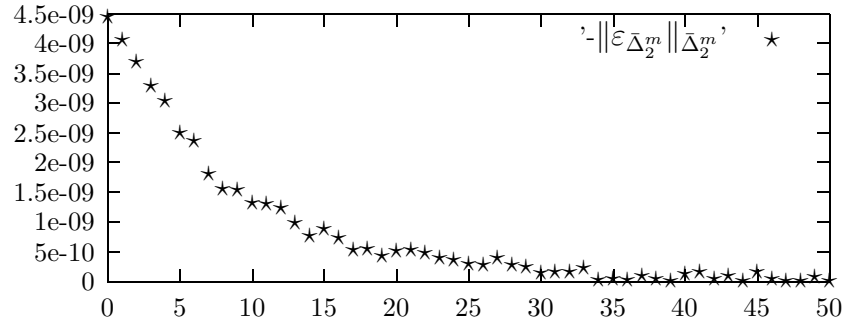
m=8, aTOL=rTOL=1E−10

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
34	9.27E-012	30	3.82E-009	30



m=8, aTOL=rTOL=1E−11

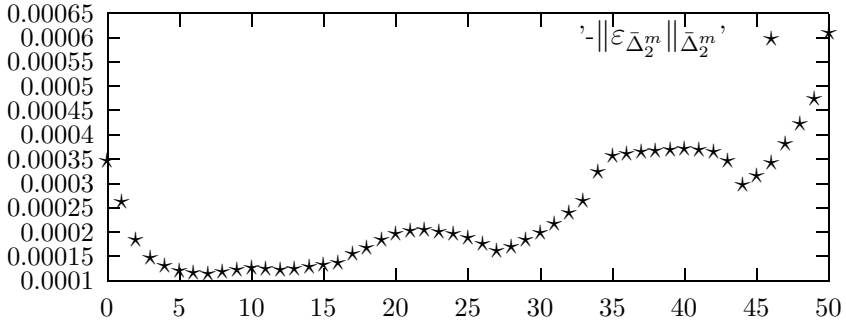
s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
48	6.88E-012	33	2.51E-009	32



Beispiel (5.8):

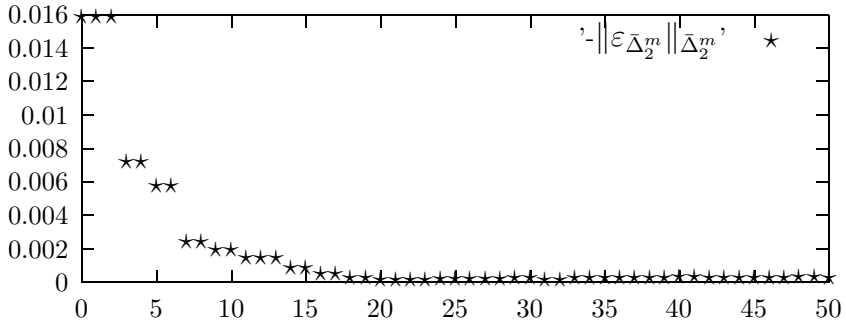
m=4, aTOL=rTOL=1E−3

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
7	1.14E-004	25	1.21E-004	25



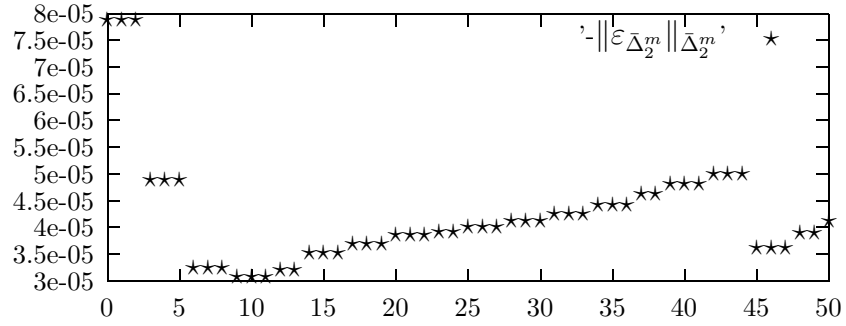
m=4, aTOL=rTOL=1E−4

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
20	1.74E-004	23	5.79E-003	22



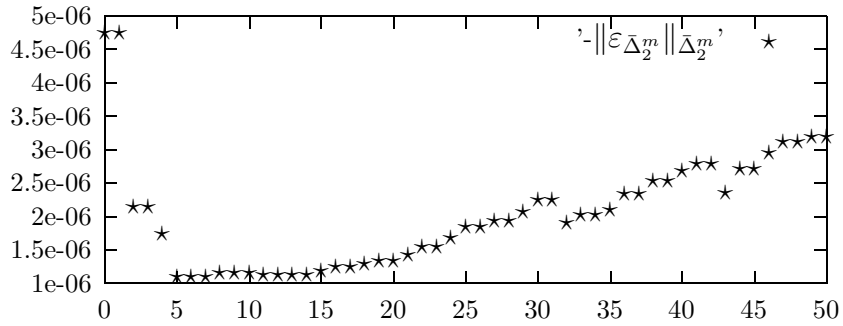
m=6, aTOL=rTOL=1E−5

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
9	3.08E-005	20	4.90E-005	20



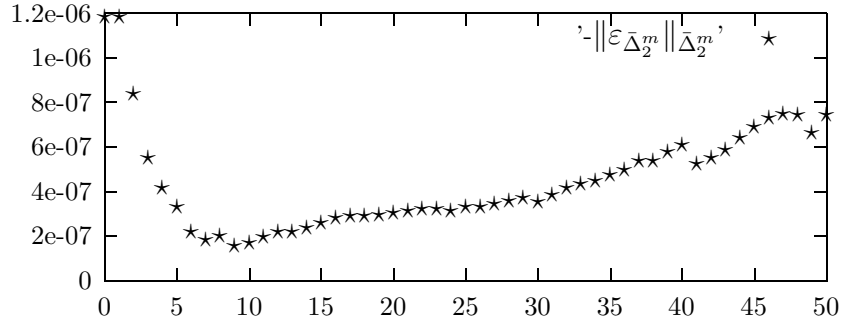
m=6, aTOL=rTOL=1E−6

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
5	1.09E-006	24	1.09E-006	24



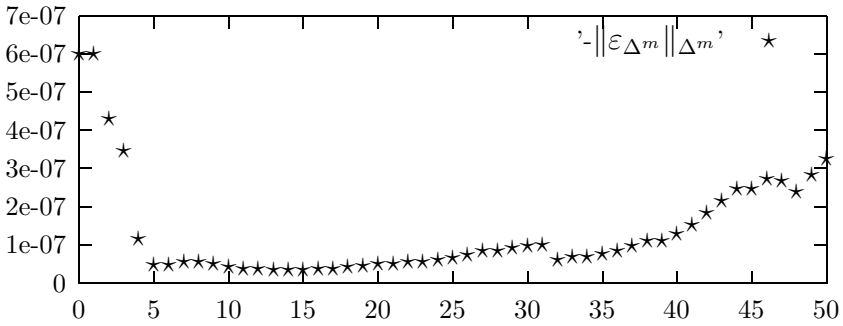
m=6, aTOL=rTOL=1E−7

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
9	1.54E-007	32	3.31E-007	31



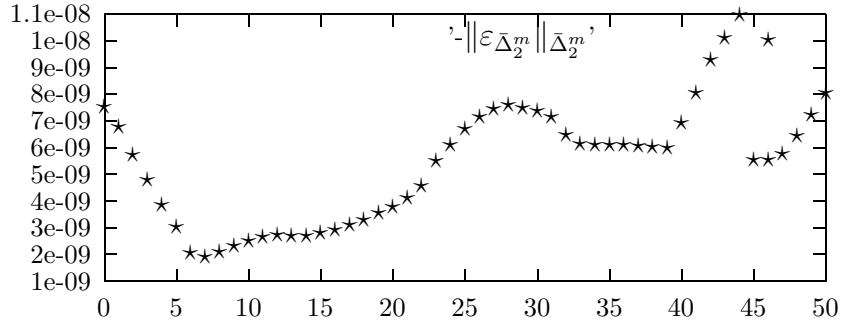
m=8, aTOL=rTOL=1E−8

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
14	3.57E-008	22	4.80E-008	22



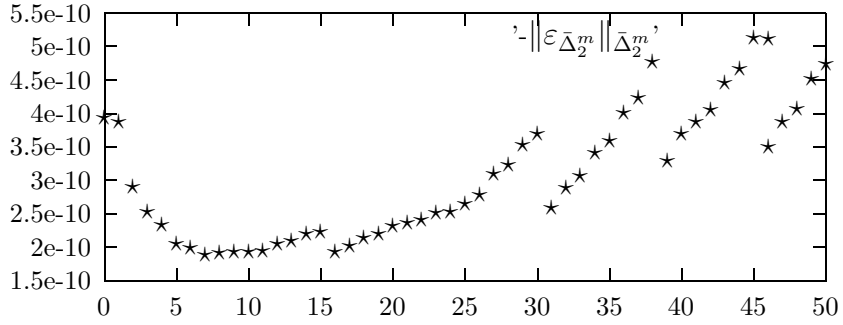
m=8, aTOL=rTOL=1E−9

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
7	1.90E-009	27	3.06E-009	26



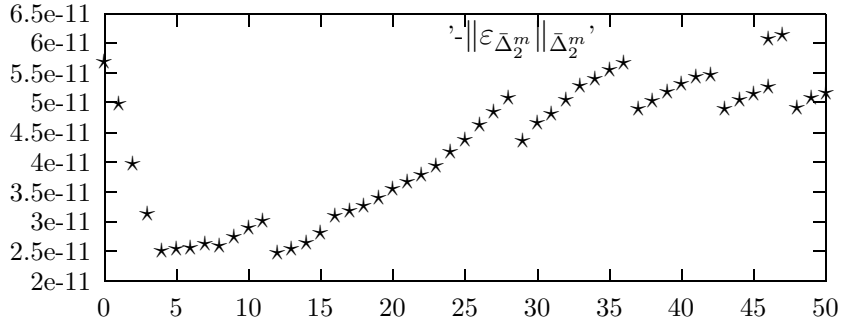
m=8, aTOL=rTOL=1E−10

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
7	1.88E-010	35	2.06E-010	35



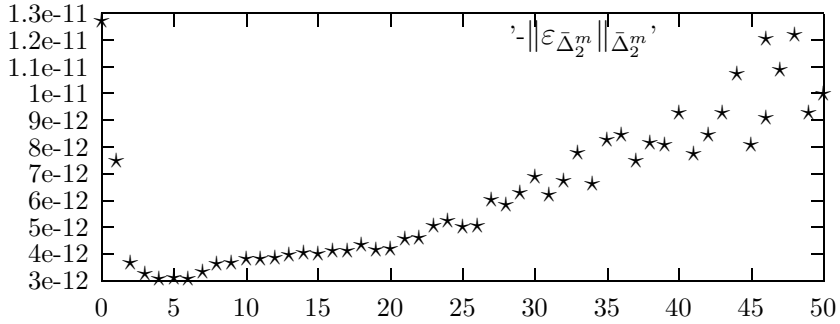
m=8, aTOL=rTOL=1E−11

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
12	2.49E-011	44	2.53E-011	45



m=8, aTOL=rTOL=1E−12

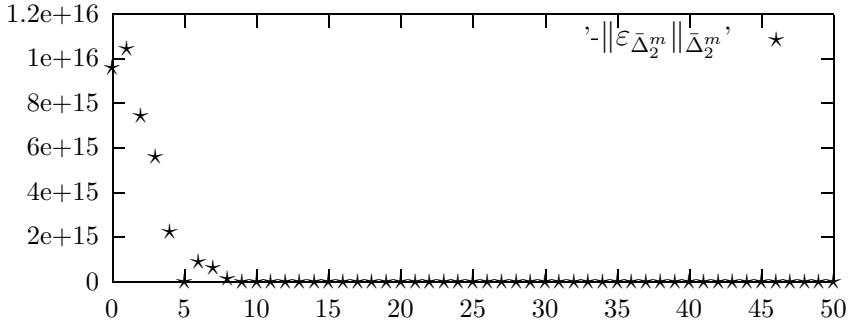
s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
4	3.06E-012	59	3.12E-012	59



Beispiel (5.7):

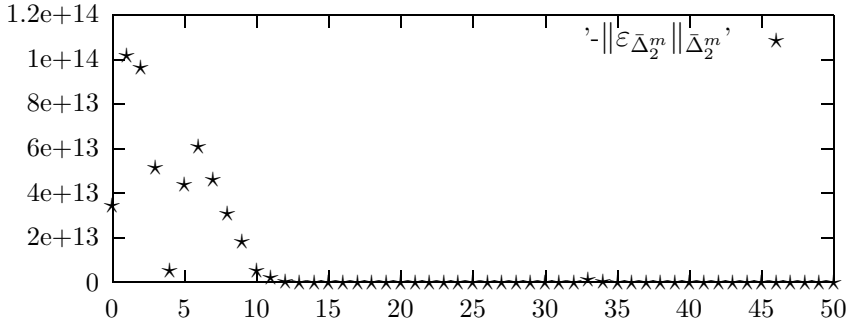
m=2, aTOL=rTOL=1E−2

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
25	4.29E-003	104	2.10E+013	103



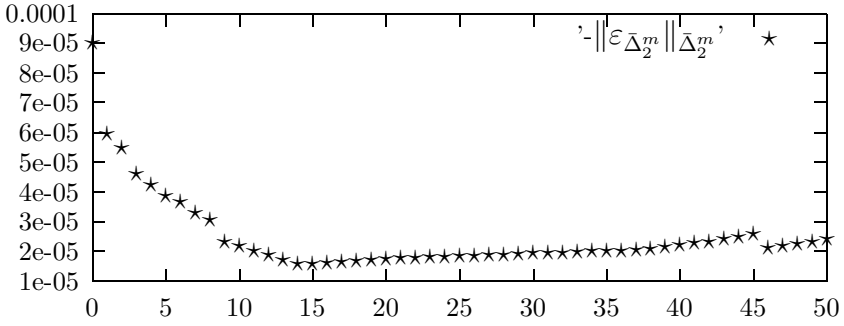
m=4, aTOL=rTOL=1E−4

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
32	1.73E-005	62	4.41E+013	62



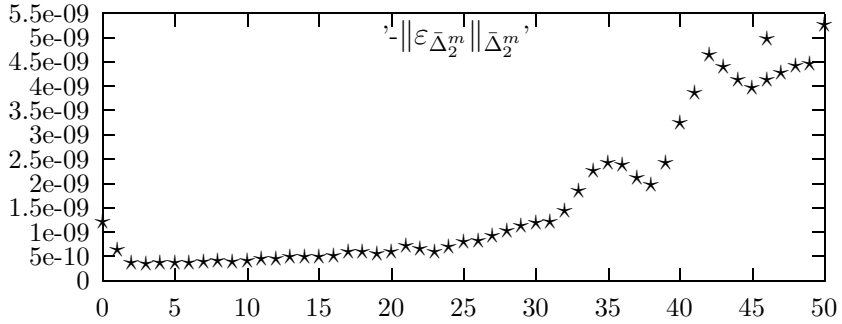
m=6, aTOL=rTOL=1E−5

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
15	1.58E-005	31	3.87E-005	30



m=6, aTOL=rTOL=1E−7

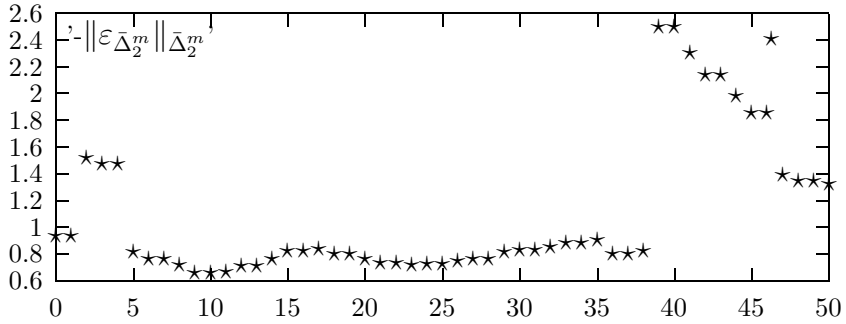
s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
3	3.54E-010	82	3.82E-010	82



Beispiel (5.3):

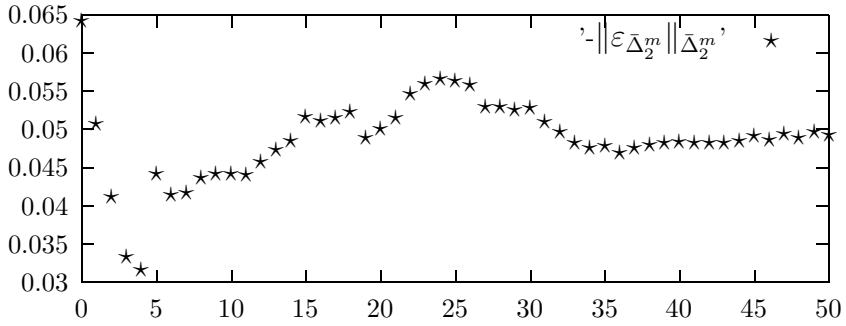
m=2, aTOL=rTOL=1E−1

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
9	6.58E-001	26	8.15E-001	26



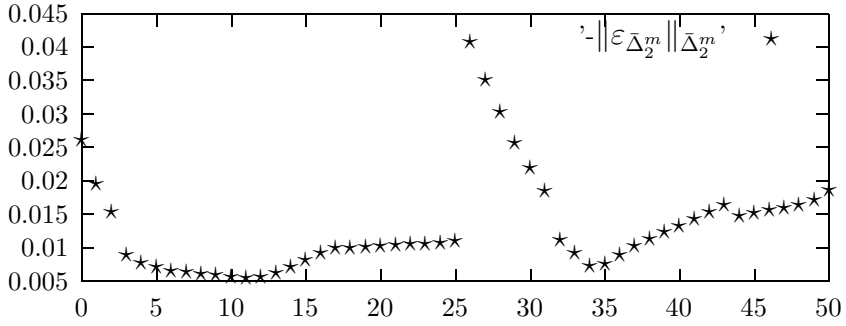
m=2, aTOL=rTOL=1E−2

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
4	3.16E-002	119	4.43E-002	120



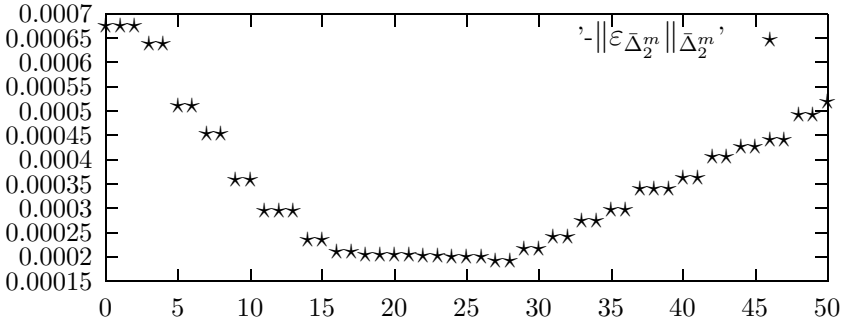
m=4, aTOL=rTOL=1E−3

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
11	5.55E-003	28	7.11E-003	28



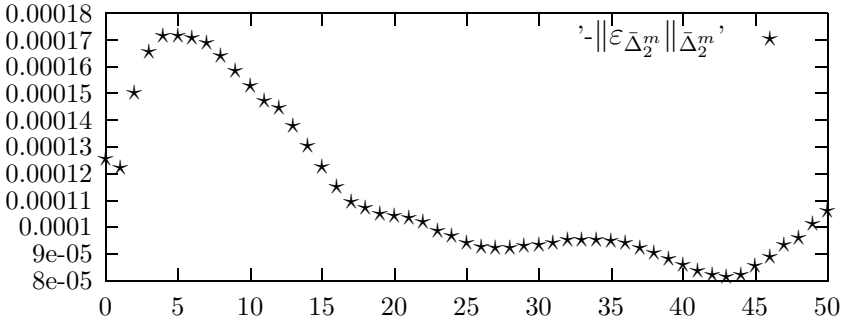
m=4, aTOL=rTOL=1E−4

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
27	1.93E-004	68	5.11E-004	65



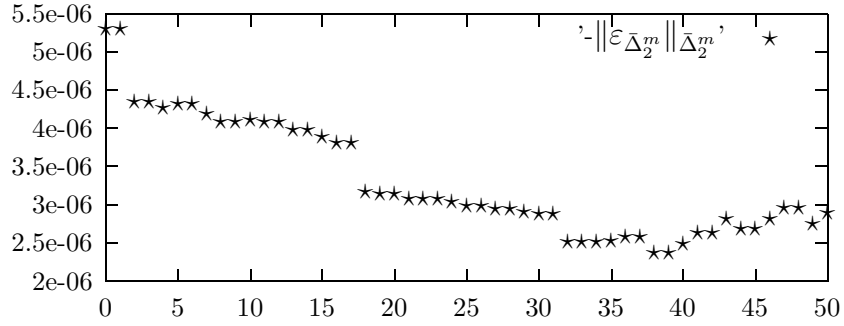
m=6, aTOL=rTOL=1E−5

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
43	8.17E-005	27	1.72E-004	27



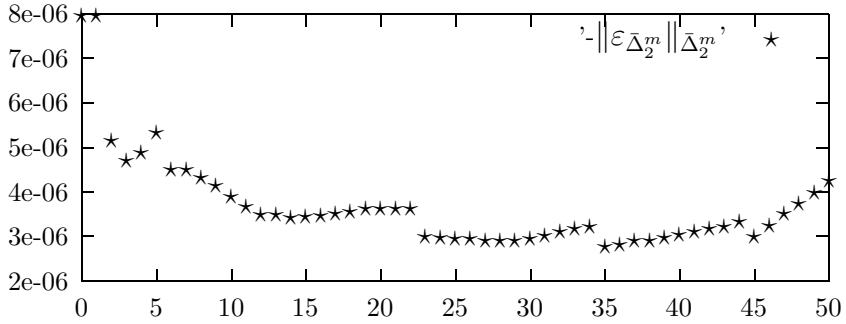
m=6, aTOL=rTOL=1E−6

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
38	2.37E-006	46	4.32E-006	43



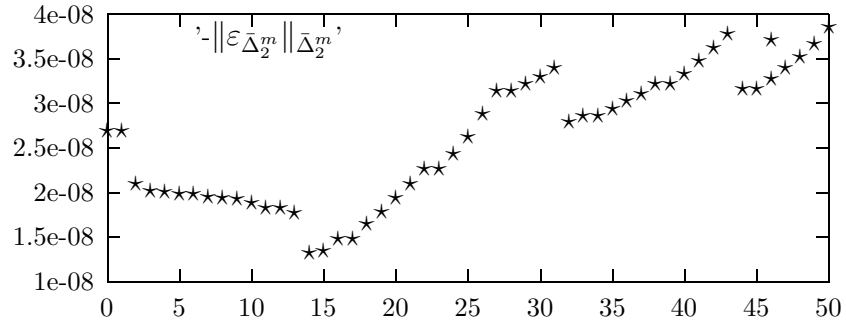
m=6, aTOL=rTOL=1E−7

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
35	2.77E-006	41	5.35E-006	38



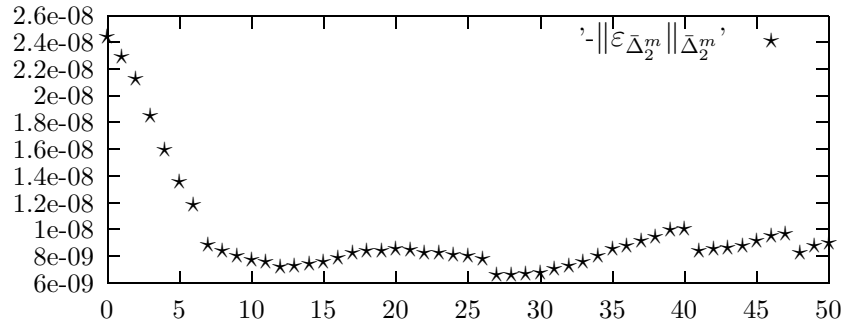
m=8, aTOL=rTOL=1E−8

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
14	1.33E-008	36	1.99E-008	35



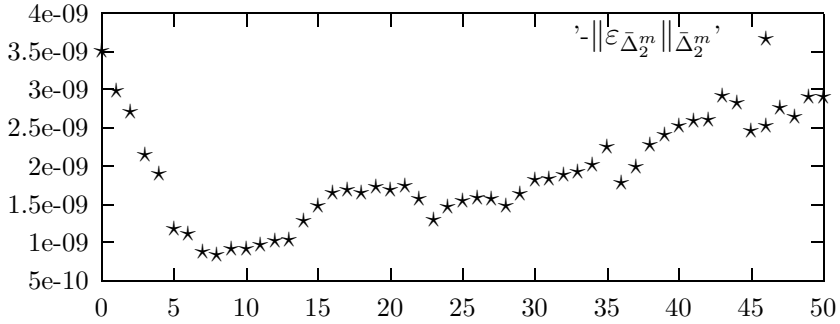
m=8, aTOL=rTOL=1E−9

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
28	6.61E-009	41	1.36E-008	39



m=8, aTOL=rTOL=1E−10

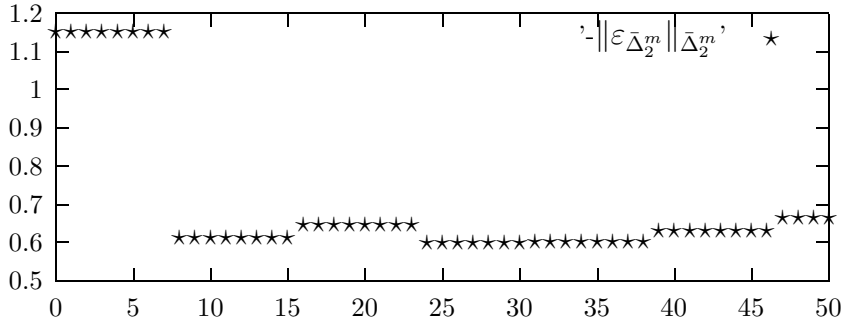
s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
8	8.46E-010	49	1.19E-009	49



Beispiel (5.21):

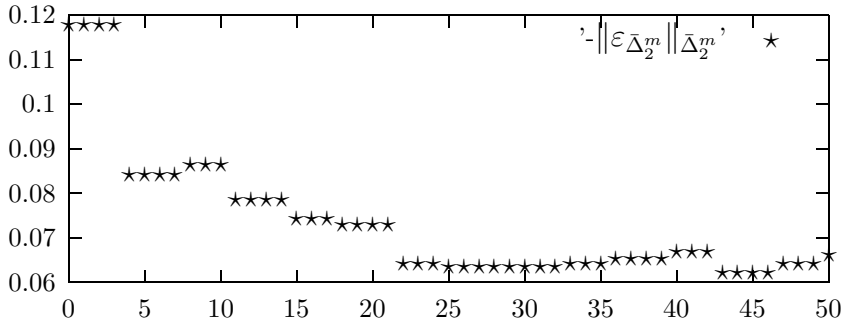
m=2, aTOL=rTOL=1E−1

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
24	5.99E-001	8	1.15E+000	7



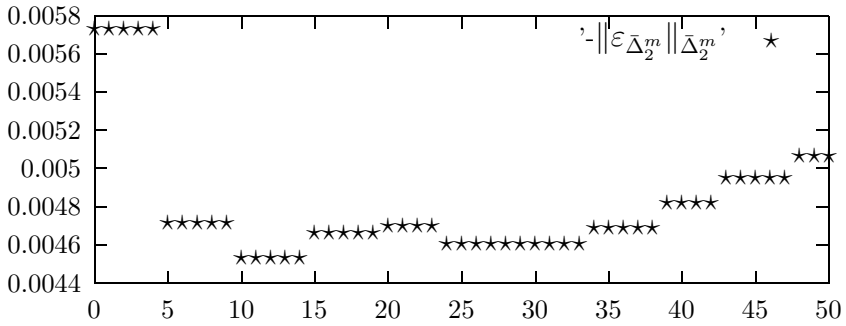
m=2, aTOL=rTOL=1E−2

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
43	6.22E-002	22	8.41E-002	20



m=4, aTOL=rTOL=1E−3

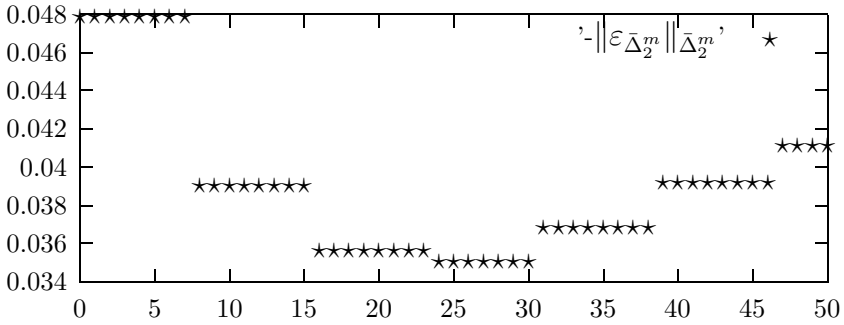
s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
10	4.53E-003	7	4.72E-003	7



Beispiel (5.2):

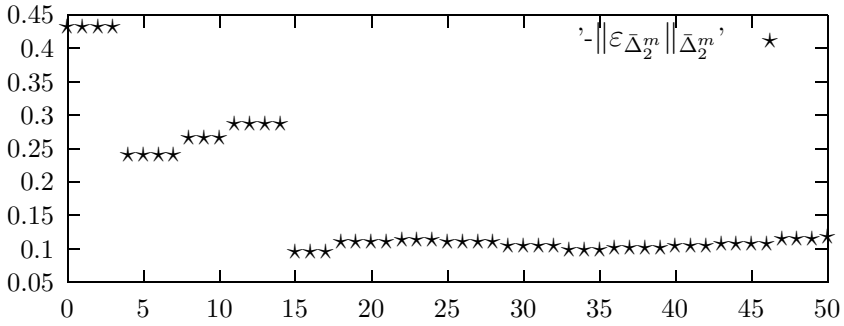
m=2, aTOL=rTOL=1E−1

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
24	3.50E-002	18	4.79E-002	17



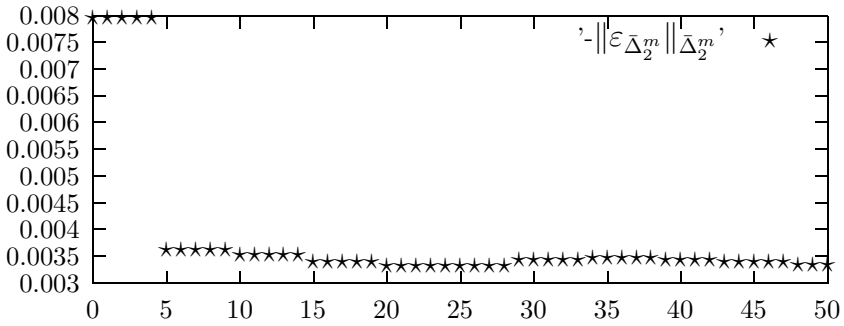
m=2, aTOL=rTOL=1E−2

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
15	9.63E-002	14	2.42E-001	13



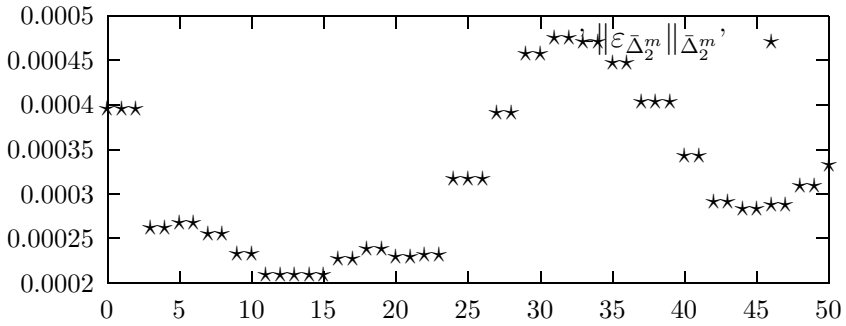
m=4, aTOL=rTOL=1E−3

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
24	3.33E-003	7	3.62E-003	7



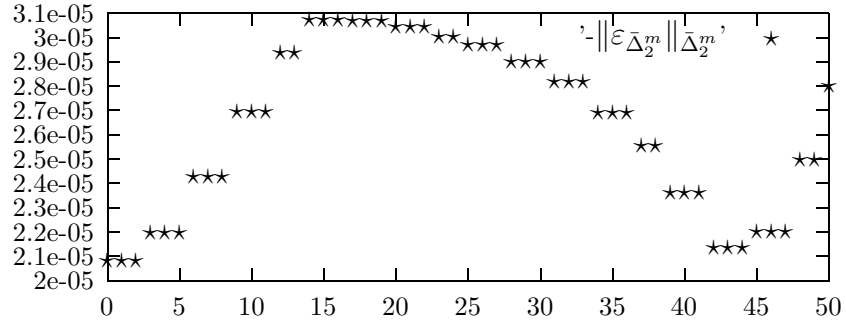
m=4, aTOL=rTOL=1E−4

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
11	2.10E-004	10	2.68E-004	10



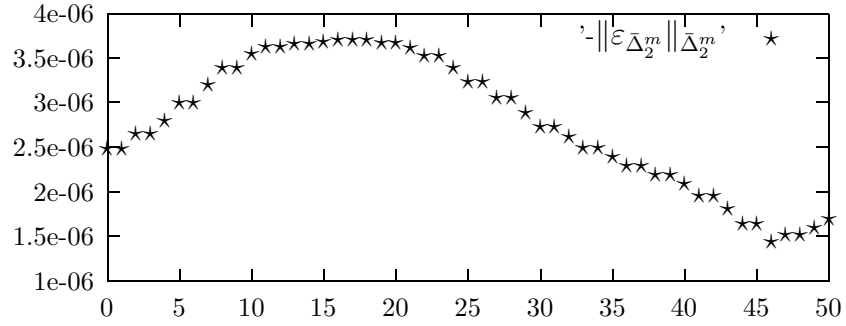
m=6, aTOL=rTOL=1E−5

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
0	2.08E-005	7	2.20E-005	7



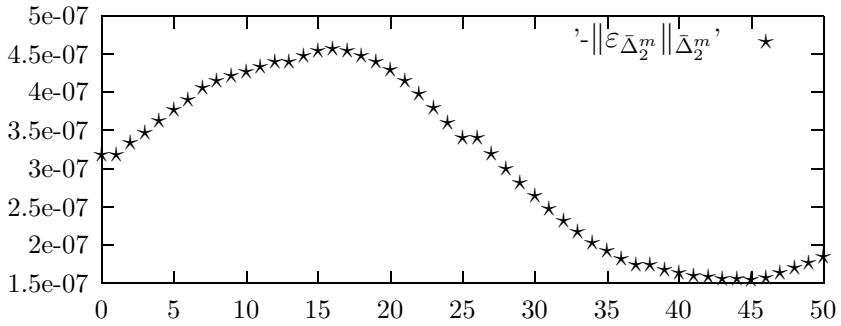
m=6, aTOL=rTOL=1E−6

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
46	1.45E-006	10	3.00E-006	10



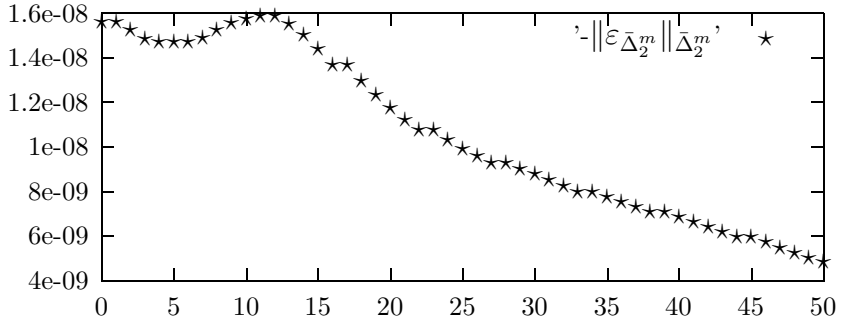
m=6, aTOL=rTOL=1E−7

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
45	1.55E-007	14	3.77E-007	14



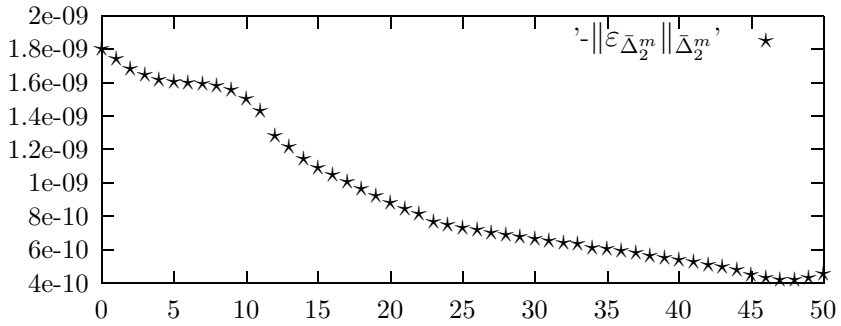
m=8, aTOL=rTOL=1E−8

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
50	4.83E-009	10	1.47E-008	10



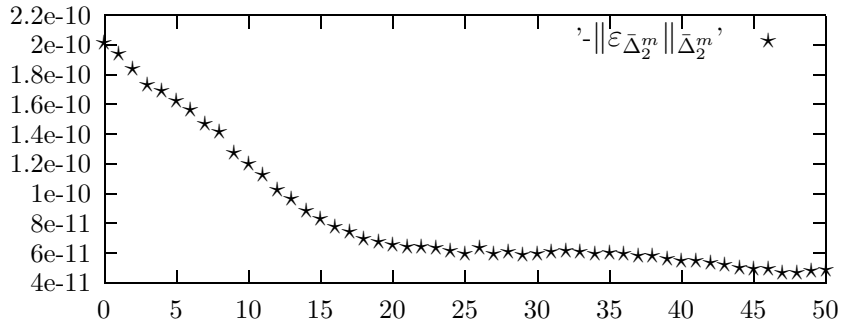
m=8, aTOL=rTOL=1E−9

s_{min}	ε_{min}	$N(\bar{\Delta}_2)_{s_{min}}+1$	ε_5	$N(\bar{\Delta}_2)_{s_5}+1$
48	4.19E-010	13	1.60E-009	13



m=8, aTOL=rTOL=1E−10

s_{min}	ε_{min}	$N(\Delta_2)_{s_{min}}+1$	ε_5	$N(\Delta_2)_{s_5}+1$
47	4.67E-011	17	1.63E-010	17



6.5 Ermittlung der $\frac{\mathbf{h}}{h_{min}}$ Schranke

Dieser Abschnitt ist der Wahl der Schranke für den Quotienten $\frac{\mathbf{h}}{h_{min}}$ im Gitter $\bar{\Delta}_2$ gewidmet. Große Werte für $\frac{\mathbf{h}}{h_{min}}$ sind bei jenen Beispielen zu erwarten, wo in gewissen kleinen Bereichen im Lösungsverlauf steile Flanken auftreten. Anhand der Beispielreihe 100 ((5.7), (5.8) und (5.10)) motivieren wir die Wahl der Schranke $\frac{\mathbf{h}}{h_{min}} \leq 10$.

- Beispiel (5.7)

  Wir untersuchen zuerst das Beispiel (5.7) mit den Eingabedaten aTOL=rTOL=1E−7, m=6. Die Abbildung 6.1 zeigt die Auswertung auf dem Basisgitter $\bar{\Delta}_1^6$.

  In den Abbildungen 6.2 und 6.3 folgen Darstellungen der Gitter $\Delta_{2,1}^6$ und $\Delta_{2,2}^6$, um die Größenordnung der auftretenden Verhältnisse $\frac{\mathbf{h}}{h_{min}}$ zu illustrieren.

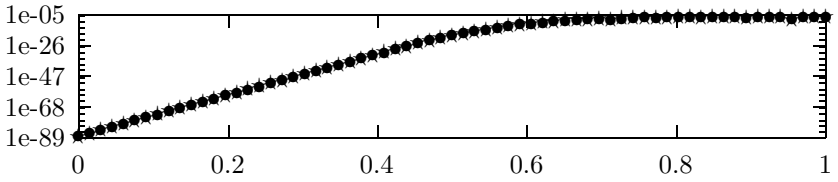
  Die Abbildung 6.4 zeigt schließlich das stabilisierte Gitter $\bar{\Delta}_2^6$ mit $\frac{\mathbf{h}}{h_{min}} \leq 10$. Die Auswertung auf diesem Gitter erfüllt sofort die Toleranzanforderungen, wobei $Q(\bar{\Delta}_2^6) = 466$ gilt.

  In der Abbildung 6.5 wird zum Vergleich die Auswertung auf einem Gitter $\bar{\Delta}_2^6$ mit $\frac{\mathbf{h}}{h_{min}} \leq 20$ angeführt. Man sieht, dass die Toleranz hier nicht sofort erreicht wurde und weitere Gitterverfeinerungen notwendig sind[4].

[4]Für dieses Gitter gilt nur $Q(\bar{\Delta}_3^6) \approx 24.5$.

Auswertung auf dem Gitter $\bar{\Delta}^6_{1,2} = \bar{\Delta}^6_1$:

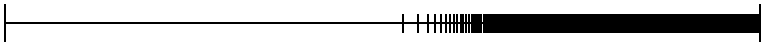
$N(\bar{\Delta}_1)+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
6.70E+001	1.52E-002	1.52E-002	1.00E+000



$\|\eta_{\bar{\Delta}^6_1}\|_{\bar{\Delta}^6_1}$	$\|\varepsilon_{\bar{\Delta}^6_1}\|_{\bar{\Delta}^6_1}$	TOL_q	TOL_g
1.10E-006	6.63E-007	1.09E+000	4.22E-007

Abbildung 6.1: Auswertung auf dem Basisgitter, Beispiel (5.7), aTOL=rTOL=1E−7, m=6

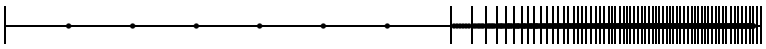
Das Gitter $\Delta^6_{2,1}$:



$N(\Delta_{2,1})+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
6.70E+001	3.99E-003	5.90E-001	1.48E+002

Abbildung 6.2: Das Gitter $\Delta^6_{2,1}$, Beispiel (5.7), aTOL=rTOL=1E−7, m=6

Das Gitter $\Delta^6_{2,2}$:

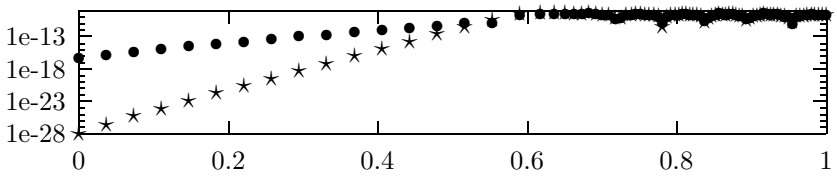


$N(\Delta_{2,2})+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
6.70E+001	3.99E-003	5.90E-001	1.48E+002

Abbildung 6.3: Das Gitter $\Delta^6_{2,2}$, Beispiel (5.7), aTOL=rTOL=1E−7, m=6

Auswertung auf dem Gitter $\bar{\Delta}^6_2$:

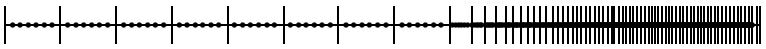
$N(\bar{\Delta}_2)+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
8.20E+001	3.99E-003	3.69E-002	9.24E+000



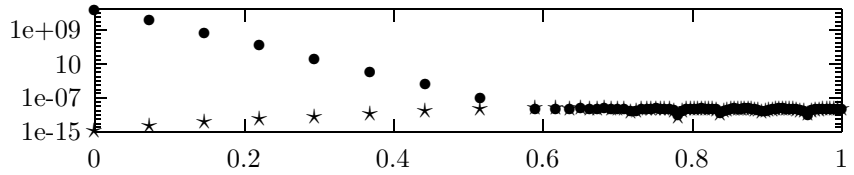
$\|\eta_{\bar{\Delta}^6_2}\|_{\bar{\Delta}^6_2}$	$\|\varepsilon_{\bar{\Delta}^6_2}\|_{\bar{\Delta}^6_2}$	TOL_q	TOL_g
5.01E-010	3.82E-010	3.70E-003	1.00E-007

Abbildung 6.4: Das Gitter $\bar{\Delta}^6_2$ mit $\frac{\mathbf{h}}{h_{min}} \leq 10$, Beispiel (5.7), aTOL=rTOL=1E−7, m=6

Auswertung auf dem Gitter $\bar{\Delta}_2^6$:



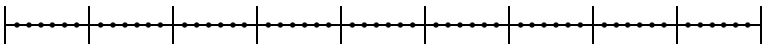
$N(\bar{\Delta}_2)+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
7.40E+001	3.99E-003	7.37E-002	1.85E+001



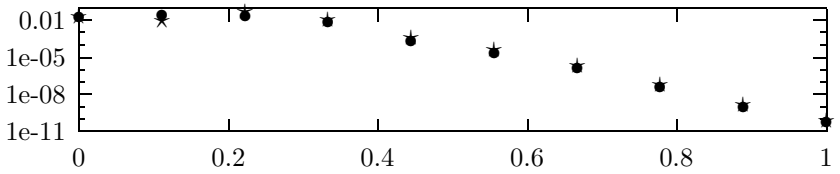
$\|\eta_{\bar{\Delta}_2^6}\|_{\bar{\Delta}_2^6}$	$\|\varepsilon_{\bar{\Delta}_2^6}\|_{\bar{\Delta}_2^6}$	TOL_q	TOL_g
5.06E-010	6.91E+013	6.91E+020	1.00E-007

Abbildung 6.5: Das Gitter $\bar{\Delta}_2^6$ mit $\frac{\mathbf{h}}{h_{min}} \leq 20$, Beispiel (5.7), aTOL=rTOL=1E−7, m=6

Auswertung auf dem Gitter $\bar{\Delta}_1^6$:



$N(\bar{\Delta}_1)+1$	h_{min}	**h**	$\frac{\mathbf{h}}{h_{min}}$
1.00E+001	1.11E-001	1.11E-001	1.00E+000



$\|\eta_{\bar{\Delta}_1^6}\|_{\bar{\Delta}_1^6}$	$\|\varepsilon_{\bar{\Delta}_1^6}\|_{\bar{\Delta}_1^6}$	TOL_q	TOL_g
5.32E-002	2.69E-002	2.11E+004	1.02E-006

Abbildung 6.6: Auswertung auf dem Basisgitter, Beispiel (5.8), aTOL=rTOL=1E−6, m=6

- Beispiel (5.8)

  Wir testen nun das Beispiel (5.8) mit den Eingabedaten aTOL=rTOL=1E−6, m=6. Abbildung 6.6 zeigt die Auswertung auf dem Basisgitter $\bar{\Delta}_1^6$. In den Abbildungen 6.7 und 6.8 folgen wieder die Darstellungen der Gitter $\Delta_{2,1}^6$ und $\Delta_{2,2}^6$. Auf dem Gitter $\Delta_{2,2}^6$ ergibt sich $\frac{\mathbf{h}}{h_{min}} = 13.9$.

  Die Lösung, die auf dem Gitter $\bar{\Delta}_2^6$ mit $\frac{\mathbf{h}}{h_{min}} \leq 10$ ermittelt wurde erfüllt die Toleranzforderung.

  Man sieht wieder, dass die Auswertung auf dem nicht stabilisierten Gitter $\Delta_{2,2}^6 =: \bar{\Delta}_{2,2}^6$ nicht erfolgreich ist, vgl. Abbildung 6.10. Hier wären weitere Gitterverfeinerungen notwendig.

Das Gitter $\Delta^6_{2,1}$:

$N(\Delta_{2,1})+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
2.30E+001	2.03E-002	2.83E-001	1.39E+001

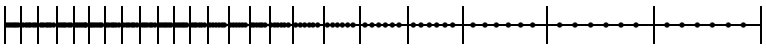
Abbildung 6.7: Das Gitter $\Delta^6_{2,1}$, Beispiel (5.8), aTOL=rTOL=1E−6, m=6

Das Gitter $\Delta^6_{2,2}$:

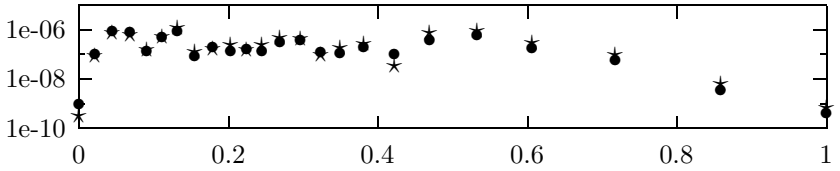
$N(\Delta_{2,2})+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
2.30E+001	2.03E-002	2.83E-001	1.39E+001

Abbildung 6.8: Das Gitter $\Delta^6_{2,2}$, Beispiel (5.8), aTOL=rTOL=1E−6, m=6

Auswertung auf dem Gitter $\bar{\Delta}_2^6$:



$N(\Delta_2)+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
2.40E+001	2.03E-002	1.41E-001	6.96E+000

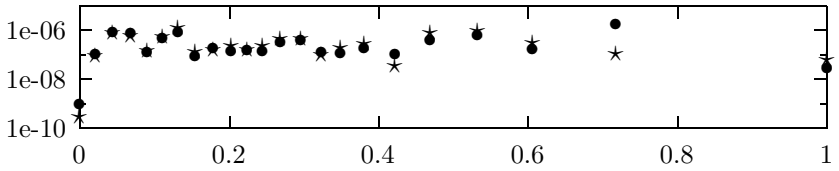


$\|\eta_{\bar{\Delta}_2^6}\|_{\bar{\Delta}_2^6}$	$\|\varepsilon_{\bar{\Delta}_2^6}\|_{\bar{\Delta}_2^6}$	TOL_q	TOL_g
1.40E-006	1.09E-006	8.31E-001	1.00E-006

Abbildung 6.9: Das Gitter $\bar{\Delta}_2^6$ mit $\frac{\mathbf{h}}{h_{min}} \leq 10$, Beispiel (5.8), aTOL=rTOL=1E−6, m=6

Auswertung auf dem Gitter $\Delta^6_{2,2} =: \bar{\Delta}^6_2$:

$N(\bar{\Delta}_2)+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
2.30E+001	2.03E-002	2.83E-001	1.39E+001



$\|\eta_{\bar{\Delta}^6_2}\|_{\bar{\Delta}^6_2}$	$\|\varepsilon_{\bar{\Delta}^6_2}\|_{\bar{\Delta}^6_2}$	TOL_q	TOL_g
1.40E-006	1.80E-006	1.80E+000	1.00E-006

Abbildung 6.10: Auswertung auf dem Gitter $\bar{\Delta}^6_{2,2}$, Beispiel (5.8), aTOL=rTOL=1E−6, m=6

- Beispiel (5.10)

  Hier betrachten wir das Beispiel (5.10) mit den Eingabedaten aTOL=rTOL=1E−5, m=6.

  Abbildung 6.11 zeigt wieder die Auswertung auf dem Basisgitter $\bar{\Delta}_1^6$, die als Grundlage zur Fehlergleichverteilung dient.

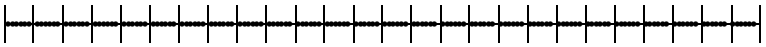
  In den Abbildungen 6.12 und 6.13 folgen die Darstellungen der Gitter $\Delta_{2,1}^6$ und $\Delta_{2,2}^6$.

  Für dieses Beispiel reicht das einmalige Halbieren des längsten Teilintervalls, um die geforderte Beschränkung für $\frac{\mathbf{h}}{h_{min}}$ zu erzielen, vgl. Abbildung 6.14. Die Auswertung auf dem Gitter $\bar{\Delta}_2^6$ erfüllt jedoch noch nicht die Toleranzanforderung.

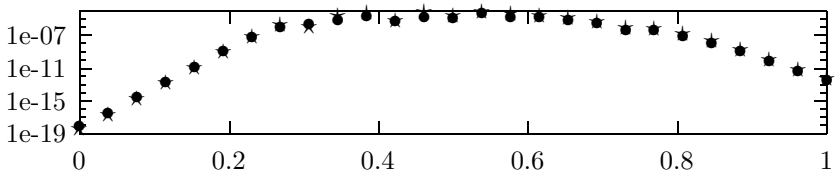
  Die Wahl von $\frac{\mathbf{h}}{h_{min}} \leq 7$ würde in diesem Fall bewirken, dass die Toleranzanforderung sofort nach der Gleichverteilung erfüllt ist, siehe Abbildung 6.15.

Natürlich kann man nicht hoffen, dass man eine universelle Schranke für $\frac{\mathbf{h}}{h_{min}}$ finden kann, sodass die unmittelbar nach der Gleichverteilung errechnete Lösung die geforderte Güte aufweist. Als Standardwert wählen wir $\frac{\mathbf{h}}{h_{min}} \leq 10$.

Auswertung auf dem Gitter $\Delta^6_{1,2} = \bar{\Delta}^6_1$:



$N(\bar{\Delta}_1)+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
2.70E+001	3.85E-002	3.85E-002	1.00E+000



$\|\eta_{\bar{\Delta}^6_1}\|_{\bar{\Delta}^6_1}$	$\|\varepsilon_{\bar{\Delta}^6_1}\|_{\bar{\Delta}^6_1}$	TOL_q	TOL_g
9.06E-005	6.98E-005	2.18E+000	2.09E-005

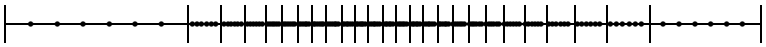
Abbildung 6.11: Auswertung auf dem Basisgitter, Beispiel (5.10), aTOL=rTOL=1E−5, m=6

Das Gitter $\Delta^6_{2,1}$:

$N(\Delta_{2,1})+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
2.70E+001	1.71E-002	2.41E-001	1.41E+001

Abbildung 6.12: Das Gitter $\Delta^6_{2,1}$, Beispiel (5.10), aTOL=rTOL=1E−5, m=6

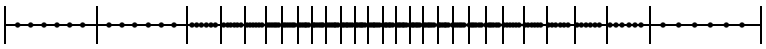
Das Gitter $\Delta^6_{2,2}$:



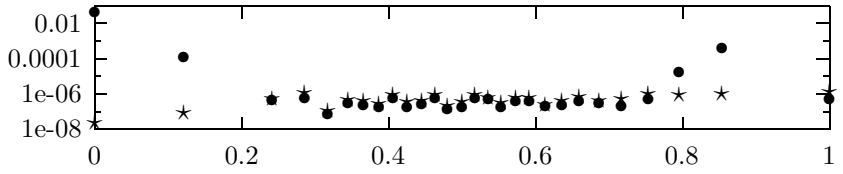
$N(\Delta_{2,2})+1$	h_{min}	**h**	$\frac{\mathbf{h}}{h_{min}}$
2.70E+001	1.71E-002	2.41E-001	1.41E+001

Abbildung 6.13: Das Gitter $\Delta^6_{2,2}$, Beispiel (5.10), aTOL=rTOL=1E−5, m=6

Auswertung auf dem Gitter $\bar{\Delta}^6_2$:



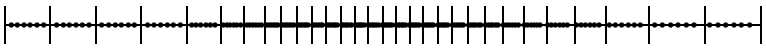
$N(\bar{\Delta}_2)+1$	h_{min}	**h**	$\frac{\mathbf{h}}{h_{min}}$
2.80E+001	1.71E-002	1.47E-001	8.57E+000



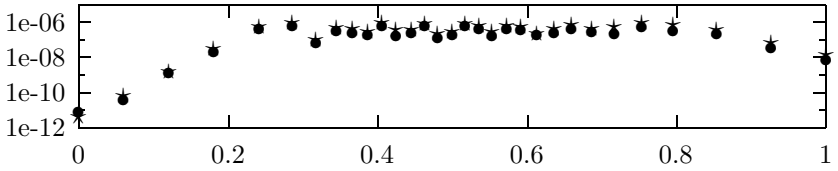
$\|\eta_{\bar{\Delta}^6_2}\|_{\bar{\Delta}^6_2}$	$\|\varepsilon_{\bar{\Delta}^6_2}\|_{\bar{\Delta}^6_2}$	TOL_q	TOL_g
1.27E-006	4.40E-002	4.40E+003	1.00E-005

Abbildung 6.14: Das Gitter $\bar{\Delta}^6_2$ mit $\frac{\mathbf{h}}{h_{min}} \leq 10$, Beispiel (5.10), aTOL=rTOL=1E−5, m=6

Auswertung auf dem Gitter $\bar{\Delta}_2^6$:



$N(\Delta_2)+1$	h_{min}	$\mathbf{h}$	$\frac{\mathbf{h}}{h_{min}}$
3.10E+001	1.71E-002	7.33E-002	4.29E+000



$\|\eta_{\bar{\Delta}_2^6}\|_{\bar{\Delta}_2^6}$	$\|\varepsilon_{\bar{\Delta}_2^6}\|_{\bar{\Delta}_2^6}$	TOL_q	TOL_g
1.08E-006	6.38E-007	6.01E-002	1.00E-005

Abbildung 6.15: Das Gitter $\bar{\Delta}_2^6$ mit $\frac{\mathbf{h}}{h_{min}} \leq 7$, Beispiel (5.10), aTOL=rTOL=1E−5, m=6

6.6 Qualität der Gittersteuerung

Abschließend soll demonstriert werden, dass unsere Gittersteuerungsstrategie für fast alle Beispiele einen Qualitätsgewinn bringt. Dazu wird das Qualitätsmaß $Q(\Delta)$ nach (4.6) bestimmt. Da in der Praxis der exakte Fehler nicht vorliegt, arbeiten wir hier mit der Fehlerschätzung. Mit Δ bezeichnen wir das letzte Gitter auf dem die Näherungslösung die Toleranz erfüllt. Das Gitter Δ_a ist das äquidistante Vergleichsgitter mit gleicher Intervallanzahl. Es gilt

$$\bar{Q}(\Delta) = \frac{\|\varepsilon_{\Delta_a}\|_{\Delta_a}}{\|\varepsilon_{\Delta}\|_{\Delta}}, \quad N(\Delta_a) = N(\Delta).$$

Das oben beschriebene Qualitätsmaß wird in den folgenden Tabellen dokumentiert. Dabei wurden nur jene Beispiele berücksichtigt, wo Gittersteuerung auftritt. Alle Testläufe wurden mit

$$\texttt{SMOOTHIG\_FACTOR} = 5\%, \quad \texttt{RECOVER\_PEAKS} = 1$$

durchgeführt.

Beispiel (5.16)

m	aTOL=rTOL	$\bar{Q}(\Delta)$
2	1.00E−01	1.12E+00
2	1.00E−02	1.08E+00
2	1.00E−03	1.05E+00

Beispiel (5.9)

m	aTOL=rTOL	$\bar{Q}(\Delta)$
4	1.00E−03	4.66E−01
6	1.00E−06	4.09E+00
8	1.00E−09	3.66E+00
8	1.00E−11	5.33E+00

Beispiel (5.10)

m	aTOL=rTOL	$Q(\Delta)$
4	1.00E−04	1.72E+01
8	1.00E−09	3.47E+01
8	1.00E−10	4.11E+01
8	1.00E−11	3.90E+01

Beispiel (5.11)

m	aTOL=rTOL	$\bar{Q}(\Delta)$
8	1.00E−06	2.16E+00
8	1.00E−07	3.10E+00

Beispiel (5.21)

m	aTOL=rTOL	$Q(\Delta)$
2	1.00E−01	1.45E+00
2	1.00E−02	1.31E+00
2	1.00E−03	1.14E+00
4	1.00E−03	1.56E+00

Beispiel (5.8)

m	aTOL=rTOL	$\bar{Q}(\Delta)$
2	1.00E−03	6.66E+00
4	1.00E−03	1.88E+01
4	1.00E−04	1.33E+01
4	1.00E−05	1.81E+01
4	1.00E−06	1.78E+01
4	1.00E−07	1.84E+01
6	1.00E−05	4.60E+00
6	1.00E−07	3.46E+01
6	1.00E−08	5.23E+01
6	1.00E−09	7.75E+01
6	1.00E−10	5.35E+01
8	1.00E−06	2.84E+02
8	1.00E−07	1.14E+02
8	1.00E−08	3.26E+01
8	1.00E−09	2.31E+02
8	1.00E−10	4.14E+02
8	1.00E−12	1.78E+01

Beispiel (5.7)

m	aTOL=rTOL	$Q(\Delta)$
2	1.00E−01	1.00E+00
2	1.00E−02	1.19E+01
2	1.00E−03	1.32E+01
4	1.00E−04	6.56E+01
4	1.00E−05	6.74E+01
4	1.00E−06	6.22E+01
4	1.00E−07	3.38E+01
6	1.00E−07	4.66E+02
6	1.00E−09	1.49E+02
6	1.00E−10	1.22E+02
8	1.00E−06	3.57E+01
8	1.00E−07	4.98E+01
8	1.00E−11	2.29E+00

Beispiel (5.2)

m	aTOL=rTOL	$Q(\Delta)$
2	1.00E−01	1.60E+00
2	1.00E−02	4.90E+00
2	1.00E−03	4.41E+00
4	1.00E−02	1.30E+00
4	1.00E−03	4.52E+00
4	1.00E−04	6.58E+00
4	1.00E−05	4.58E+00
4	1.00E−06	3.08E+00
4	1.00E−07	8.79E+00
6	1.00E−04	2.52E+00
6	1.00E−05	3.60E+00
6	1.00E−06	2.16E+00
6	1.00E−07	2.99E+00
6	1.00E−08	5.83E+00
6	1.00E−09	5.46E+00
6	1.00E−10	4.58E+00
8	1.00E−05	1.70E+00
8	1.00E−06	6.98E+00
8	1.00E−07	7.82E+00
8	1.00E−08	4.02E+00
8	1.00E−09	3.59E+00
8	1.00E−10	3.48E+00

Beispiel (5.3)

m	aTOL=rTOL	$Q(\Delta)$
2	1.00E−01	7.81E+00
2	1.00E−02	8.74E+00
2	1.00E−03	6.01E+00
4	1.00E−02	3.20E+01
4	1.00E−03	2.99E+01
4	1.00E−04	1.02E+01
4	1.00E−05	2.79E+01
4	1.00E−06	2.85E+01
4	1.00E−07	2.42E+01
6	1.00E−03	1.78E+01
6	1.00E−04	2.03E+01
6	1.00E−05	2.71E+01
6	1.00E−06	2.75E+01
6	1.00E−07	4.28E+01
6	1.00E−08	3.69E+01
6	1.00E−09	4.77E+01
6	1.00E−10	5.89E+01
8	1.00E−01	5.16E+00
8	1.00E−02	4.18E+00
8	1.00E−03	3.21E+00
8	1.00E−04	1.64E+01
8	1.00E−05	4.96E+00
8	1.00E−06	1.04E+01
8	1.00E−07	4.57E+01
8	1.00E−08	5.99E+01
8	1.00E−09	4.68E+01
8	1.00E−10	4.10E+01
8	1.00E−12	1.20E+00

Literaturverzeichnis

[1] U. Ascher, J. Christiansen, and R.D. Russel. *A Collocation Solver for Mixed Order Systems of Boundary Value Problems.* Math. Comp., 33 (1978), pp. 659–679.

[2] U. Ascher, J. Christiansen, and R.D. Russel. *Collocation Software for Boundary Value ODEs.* ACM Transactions on Mathematical Software, 7/2 (1981), pp. 209–222.

[3] 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.

[4] W. Auzinger, O. Koch, P. Kofler, and E. Weinmüller. *The Application of Shooting to Singular Boundary Value Problems.* Techn. Rep. Nr. 126/99, Inst. for Appl. Math. and Numer. Anal., Vienna Univ. of Technology, Austria, 1999. Avalaible at `http://fsmat.htu.tuwien.ac.at/~othmar/research.html`.

[5] W. Auzinger, O. Koch, and E. Weinmüller. *Efficient Collocation Schemes for Singular Boundary Value Problems.* Submitted to Numer. Algorithms.

[6] E. Badralexe and A.J. Freeman. *Eigenvalue Equation for a General Periodic Potential and its Multipole Expansion Solution.* Phys. Rev. B, 37/3 (1988), pp. 1067–1084.

[7] L. Bauer, E.L. Reiss, and H.B. Keller. *Axisymmetric Buckling of Hollow Spheres and Hemispheres.* Comm. Pure Appl. Math., 23 (1970), pp. 529–568.

[8] C. de Boor and B. Swartz. *Collocation at Gaussian Points.* SIAM J. Numer. Anal., 10 (1973), pp. 582–606.

[9] T.W. Carr and T. Erneux. *Understanding the Bifurcation to Traveling Waves in a Class-B Laser Using a Degenerate Ginzburg-Landau Equation.* Phys. Rev. A, 50 (1994), pp. 4219–4227.

[10] C.Y. Chan and Y.C. Hon. *A Constructive Solution for a Generalized Thomas-Fermi Theory of Ionized Atoms.* Quart. Appl. Math., 45 (1987), pp. 591–599.

[11] M. Drmota, R. Scheidl, H. Troger, and E. Weinmüller. *On the Imperfection Sensitivity of Complete Spherical Shells.* Comp. Mech., 2 (1987), pp. 63–74.

[12] R. Fazio. *A Novel Approach to the Numerical Solution of Boundary Value Problems on Infinite Intervals.* SIAM J. Numer. Anal., 33 (1996), pp. 1473–1483.

[13] R. Frank. *Schätzungen des globalen Diskretisierungsfehlers bei Runge-Kutta-Methoden.* ISNM, 27 (1975), pp. 45–70.

[14] R. Frank. *The Method of Iterated Defect Correction and Its Application to Two-Point Boundary Value Problems, Part I.* Numer. Math., 25 (1976), pp. 409–419.

[15] R. Frank and C. Überhuber. *Iterated Defect Correction for Differential Equations, Part I: Theoretical Results.* Computing, 20 (1978), pp. 207–228.

[16] M. Gräff and E. Weinmüller. *Schätzungen des lokalen Diskretisierungsfehlers bei singulären Anfangswertproblemen.* Techn. Rep. Nr. 66/86, Inst. for Appl. Math. and Numer. Anal., Vienna Univ. of Technolgy, Austria, 1986.

[17] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I.* Springer-Verlag, Berlin-Heidelberg-New York, 1987.

[18] F.R. de Hoog and R. Weiss. *Difference Methods for Boundary Value Problems with a Singularity of the First Kind.* SIAM J. Numer. Anal., 13 (1976), pp. 775–813.

[19] F.R. de Hoog and R. Weiss. *The Application of Linear Multi-Step Methods to Singular Initial Value Problems.* Math. Comp., 32 (1977), pp. 676–690.

[20] F.R. de Hoog and R. Weiss. *The Application of Runge-Kutta Schemes to Singular Initial Value Problems.* Math. Comp., 44 (1985), pp. 93–103.

[21] O. Koch, P. Kofler, and E. Weinmüller. *Analysis of Singular Initial and Terminal Value Problems*. Techn. Rep. Nr. 125/99, Inst. for Appl. Math. and Numer. Anal., Vienna Univ. of Technology, Austria, 1999. Avalaible at `http://fsmat.htu.tuwien.ac.at/~othmar/research.html`.

[22] O. Koch and E. Weinmüller. *The Convergence of Shooting Methods for Singular Boundary Value Problems*. To appear in Math. Comp.

[23] O. Koch and E. Weinmüller. *Iterated Defect Correction for the Solution of Singular Initial Value Problems*. To appear in SIAM J. Numer. Anal.

[24] P. Kofler. *Theorie und numerische Lösung singulärer Anfangswertprobleme gewöhnlicher Differentialgleichungen mit der Singularität erster Art*. Ph. D. Thesis, Inst. for Appl. Math. and Numer. Anal., Vienna Univ. of Technology, Austria, 1998.

[25] X. Liu. *A Note on the Sturmian Theorem for Singular Boundary Value Problems*. J. Math. Anal. Appl., 237 (1999), pp. 393–403.

[26] R. März and E. Weinmüller. *Solvability of Boundary Value Problems for Systems of Singular Differential-Algebraic Equations*. SIAM J. Math. Anal., 24 (1993), pp. 200–215.

[27] G. Moore. *Computation and Parametrization of Periodic and Connecting Orbits*. IMA J. Numer. Anal., 15 (1995), pp. 245–263.

[28] S.V. Parter, M.L. Stein, and P.R. Stein. *On the Multiplicity of Solutions of a Differential Equation Arising in Chemical Reactor Theory*. Techn. Rep. Nr. 194, Dept. Computer Sciences, Univ. of Wisconsin, 1973.

[29] W. Ruess and E. Weinmüller. *Beschleunigte Algorithmen zur effizienten numerischen Lösung von singulären Randwertaufgaben*. Techn. Rep. Nr. 127/99, Inst. for Appl. Math. and Numer. Anal., Vienna Univ. of Technology, 1999.

[30] H. J. Stetter. *Analysis of Discretization Methods for Ordinary Differential Equations*. Springer-Verlag, Berlin-Heidelberg-New York, 1973.

[31] H. J. Stetter. *The Defect Correction Principle and Discretization Methods*. Numer. Math., 29 (1978), pp. 425–443.

[32] P.E. Zadunaisky. *On the Estimation of Errors Propagated in the Numerical Integration of ODEs*. Numer. Math., 27 (1976), pp. 21–39.