TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

DIPLOMARBEIT

Numerische WKB-Methode für die stationäre Schrödingergleichung: Spektralmethode zur Phasenberechnung

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

unter Anleitung von
Univ-Prof. Dipl.-Ing. Dr. techn. Anton Arnold

durch
Bernhard Ujvari

Wien, 19. Oktober 2015
Datum

Unterschrift

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abstract

Die Arbeit behandelt ein numerisches WKB-Verfahren zur Näherung der hoch oszillatorischen Lösung einer stationären Schrödingergleichung. In der WKB-Näherung tritt ein Integral auf, das mit Hilfe von Spektralmethoden numerisch berechnet werden soll. Dabei wird untersucht, wie sich die Verwendung von Spektralmethoden, in der Arbeit die Clenshaw-Curtis Quadratur und eine baryzentrisch interpolierte Stammfunktion nach Entwicklung in eine Chebyshevreihe, auf den Gesamtfehler des Verfahrens im Vergleich zu gängigen Quadraturverfahren (wie Simpsonquadratur oder Trapezregel) auswirkt. Dazu wird in Kapitel 3 die Fehlerabschätzung des WKB-Verfahrens um eine Abschätzung des Fehlers erweitert, der bei der numerischen Berechnung des auftretenden Phasenintegrals entsteht.
In Kapitel 4 und 5 wird das WKB-Verfahren und die Integration mit Spektralmethoden anhand von zwei numerischen Beispielen behandelt.

Danksagungen

Ich möchte Prof. Arnold für die Unterstützung beim Schreiben dieser Arbeit danken, für die Zeit, um meine Fragen zu beantworten und die hilfreichen Hinweise. Weiters möchte Prof. Klein (Université de Bourgogne) für seine Anregungen zur baryzentrischen Interpolation und das Bereitstellen eines MATLAB-Codes zur Implementierung der Clenshaw-Curtis Stammfunktion danken.
Ich möchte auch Vanessa und Klaus dafür danken, dass sie die Arbeit Korrektur gelesen haben.
Zum Schluss möchte ich noch besonders meinen Eltern und allen anderen, die mich während meiner Studienzeit unterstützt haben danken.

Kapitel 1

Einleitung

1.1 Problemstellung und Ziel

Betrachtet man die, aus der Quantenmechanik kommende, Schrödingergleichung

$$\begin{cases} \varepsilon^2\varphi''(x) + a(x)\varphi(x) = 0, \quad x \in (0;1), & \text{(1.1a)} \\ \varphi(0) = \varphi_0, & \text{(1.1b)} \\ \varepsilon\varphi'(0) = \varphi_1, & \text{(1.1c)} \end{cases}$$

mit $0 < \varepsilon \ll 1$ und $a(x) \geq a_0 > 0$, für alle $x \in [0;1]$, einer glatten Funktion, ist die Lösung für kleines ε hoch oszillatorisch.
Die oszillierende Lösung von (1.1) soll möglichst effizient numerisch berechnet werden. Das bedeutet, dass das zur Berechnung verwendete numerische Verfahren für eine gute Approximation der Lösung weniger Gitterpunkte benötigen soll als $\varphi(x)$ Oszillationen hat.
Mit dem WKB-Verfahren, das in „WKB-based schemes for the oscillatory 1D Schrödinger equation in the semiclassical limit“ [ABN] präsentiert wird, lässt sich $\varphi(x)$ effizient berechnen. Bei der Berechnung der Lösung $\varphi(x)$ mit dem WKB-Verfahren, welches in Abschnitt 1.3 genauer erläutert wird, tritt ein Integral der Form

$$\phi(x) = \int_0^x \left(\sqrt{a(\tau)} - \varepsilon^2\beta(\tau)\right) \mathrm{d}\tau \tag{1.2}$$

mit

$$\beta := -\frac{1}{2a^{\frac{1}{4}}}\left(a^{-\frac{1}{4}}\right)'' \tag{1.3}$$

auf.
Ziel dieser Arbeit ist es, das Integral (1.2) mit Hilfe von Spektralmethoden

numerisch zu berechnen und zu untersuchen, welche Auswirkungen diese Methoden auf den Fehler im Gesamtverfahren im Vergleich zu herkömmlichen Quadraturverfahren (z.B. Simpsonverfahren oder Trapezregel) haben.
In [ABN, Theorem 3.1] wird eine Fehlerabschätzung für dieses WKB-Verfahren angegeben, jedoch wird der Quadraturfehler, der bei der numerischen Berechnung von (1.2) entsteht, nur teilweise in diesen Abschätzungen berücksichtigt. Daher ist ein weiteres Ziel dieser Arbeit, die Fehlerabschätzungen um den Quadraturfehler zu erweitern. Dies erfolgt in Kapitel 3, nachdem in Kapitel 2 Spektralmethoden und auf diesen basierende Quadraturverfahren präsentiert werden. Anschließend wird anhand von zwei Beispielen (in Kapitel 4 und 5) die Berechnung von (1.2) mit den Quadraturverfahren aus Kapitel 2 und der Simpsonquadratur verglichen und untersucht, wie sich die unterschiedlichen Methoden auf den Gesamtfehler des WKB-Verfahrens auswirken.

1.2 Die eindimensionale, stationäre Schrödingergleichung und WKB-Näherungen

In der Quantenmechanik wird die nach den Physikern Wentzel, Kramers und Brillouin benannte Methode verwendet, um eine asymptotische Lösung für kleines $\hbar$ der eindimensionalen, stationären Schrödingergleichung

$$\begin{aligned} \psi''(x) + k^2(x)\psi(x) &= 0 \\ \text{mit } k^2(x) &= \frac{2m}{\hbar^2}\left(E - V(x)\right) \end{aligned} \tag{1.4}$$

zu konstruieren. Dabei ist E die Energie und m die Masse des betrachteten Teilchens, dass sich in einem Potential $V(x)$ befindet. Für weitere Details zur Schrödingergleichung sei auf die entsprechende Literatur wie z.B. [De, Kapitel 4] verwiesen. Setzt man $a(x) = E - V(x)$ und $\varepsilon^2 = \frac{\hbar^2}{2m}$, erhält man aus (1.4) die in dieser Arbeit betrachtete Gleichung (1.1a).
Für eine WKB-Näherung wird der Ansatz

$$\psi(x) = c \cdot \exp\left(\frac{\imath}{\hbar}\phi(x)\right) \tag{1.5}$$

in (1.4) eingesetzt und anschließend $\phi(x)$ in Potenzen von $\imath\hbar$ entwickelt

$$\phi(x) \sim \sum_{p=0}^{\infty} (\imath\hbar)^p \phi_p(x). \tag{1.6}$$

Betrachtet man die einzelnen Potenzen von $\hbar$, und fordert, dass (1.4) für jede Potenz erfüllt wird, erhält man als Näherung zweiter Ordnung

$$\psi(x) \approx \frac{C}{\sqrt{k(x)}} \exp\left[\pm\frac{\imath}{\hbar}\int^x \left(\hbar k(\tau) - \frac{\hbar}{4}\frac{k''(\tau)}{k^2(\tau)} + \frac{3}{8}\hbar\frac{k'^2(\tau)}{k^3(\tau)}\right)\mathrm{d}\tau\right]. \quad (1.7)$$

Für die detaillierte Herleitung von (1.7) sei auf [No, Kapitel 7.4] verwiesen und noch erwähnt, dass für (1.7) bis zur zweiten Potenz von $\hbar$ entwickelt worden ist, jedoch diese und höhere $\hbar-$Potenzen in WKB-Näherungen nicht mehr berücksichtigt werden (vgl. dazu auch [Kr, Kapitel 11.1]).
Für (1.1) führt der WKB-Ansatz der Form

$$\varphi(x) \sim \exp\left(\frac{1}{\varepsilon}\sum_{p=0}^{\infty}\varepsilon^p\phi_p(x)\right) \quad (1.8)$$

mit Vergleich der Potenzen von ε bis zur zweiten Ordnung auf eine Näherung

$$\varphi(x) \approx \frac{C}{\sqrt[4]{a(x)}} \exp\left(\pm\frac{\imath}{\varepsilon}\phi(x)\right),\ \phi(x) = \int_0^x \left(\sqrt{a(x)} - \varepsilon^2\beta(\tau)\right)\mathrm{d}\tau. \quad (1.9)$$

1.3 Das WKB-Verfahren aus [ABN]

In [ABN] wird das Anfangswertproblem

$$\begin{cases} \varepsilon^2\varphi''(x) + a(x)\varphi(x) = 0, & 0 < x < 1, \\ \varphi(0) = \varphi_0 = 1, \\ \varepsilon\varphi'(0) = \varphi_1 = -\imath\sqrt{a(0)} \end{cases} \quad (1.10)$$

in ein Differentialgleichungssystem 1. Ordnung überführt, indem

$$U(x) = \begin{pmatrix} u_1(x) \\ u_2(x) \end{pmatrix} := \begin{pmatrix} a^{\frac{1}{4}}(x)\varphi(x) \\ \frac{\varepsilon\left(a^{\frac{1}{4}}(x)\varphi(x)\right)'}{\sqrt{a(x)}} \end{pmatrix} \quad (1.11)$$

gesetzt wird, und man erhält

$$\begin{cases} U'(x) = \left[\frac{1}{\varepsilon}A_0(x) + \varepsilon A_1(x)\right]U(x), & 0 < x < 1, \\ U(0) = U_I \end{cases} \quad (1.12)$$

mit

$$A_0(x) = \sqrt{a(x)}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \text{ und } A_1(x) = \begin{pmatrix} 0 & 0 \\ 2\beta(x) & 0 \end{pmatrix},$$

wobei $\beta(x)$ in (1.3) definiert ist. In (1.12) dominiert $\frac{1}{\varepsilon}A_0(x)$ und führt zu starken Oszillationen in der Lösung.
Um den dominanten Teil zu diagonalisieren, führt man die Transformation

$$Y(x) = PU(x)$$

mit den Matrizen

$$P = \frac{1}{\sqrt{2}} \begin{pmatrix} \imath & 1 \\ 1 & \imath \end{pmatrix} \tag{1.13}$$

$$P^{-1} = \frac{1}{\sqrt{2}} \begin{pmatrix} -\imath & 1 \\ 1 & -\imath \end{pmatrix} \tag{1.14}$$

durch. Damit erhält man aus (1.12) das System

$$\begin{cases} Y'(x) = \frac{1}{\varepsilon}D^{\varepsilon}(x)Y(x) + \varepsilon N(x)Y(x), & 0 < x < 1, \\ Y(0) = Y_I \end{cases} \tag{1.15}$$

mit den Matrizen

$$D^{\varepsilon}(x) = \begin{pmatrix} \sqrt{a(x)} - \varepsilon^2\beta(x) & 0 \\ 0 & -\sqrt{a(x)} + \varepsilon^2\beta(x) \end{pmatrix}$$

und

$$N(x) = \begin{pmatrix} 0 & \beta(x) \\ \beta(x) & 0 \end{pmatrix}.$$

Wegen der Matrix $\frac{\imath}{\varepsilon}D^{\varepsilon}(x)$ ist die Lösung auch hier wieder stark oszillierend. Um die Oszillationen zu eliminieren, bringt man mit

$$\Phi^{\varepsilon}(x) = \begin{pmatrix} \int_0^x \left(\sqrt{a(\tau)} - \varepsilon^2\beta(\tau)\right) \mathrm{d}\tau & 0 \\ 0 & -\int_0^x \left(\sqrt{a(\tau)} - \varepsilon^2\beta(\tau)\right) \mathrm{d}\tau \end{pmatrix} \tag{1.16}$$

das System (1.15) durch die Transformation

$$Z(x) = \begin{pmatrix} z_1(x) \\ z_2(x) \end{pmatrix} := \exp\left(-\frac{\imath}{\varepsilon}\Phi^{\varepsilon}(x)\right) Y(x)$$

auf die Form

$$\begin{cases} Z'(x) = \varepsilon N^{\varepsilon}(x)Z(x), & 0 < x < 1, \\ Z(0) = Z_I = Y_I. \end{cases} \tag{1.17}$$

Die Matrix N^{ε} hat die Form

$$N^{\varepsilon}(x) = \exp\left(-\frac{\imath}{\varepsilon}\Phi^{\varepsilon}(x)\right) N(x) \exp\left(\frac{\imath}{\varepsilon}\Phi^{\varepsilon}(x)\right) = \begin{pmatrix} 0 & \beta(x)e^{-\frac{2\imath}{\varepsilon}\phi(x)} \\ \beta(x)e^{\frac{2\imath}{\varepsilon}\phi(x)} & 0 \end{pmatrix},$$

wobei $\phi(x)$ die in (1.2) definierten und in (1.16) auftretenden Integrale sind. In [ABN, Propositon 2.2] wird die eindeutige Lösung von (1.17) mittels Picard-Iteration mit

$$Z(x) = Z_I + \sum_{p=1}^{\infty} \varepsilon^p M_p^\varepsilon(x;0) Z_I \tag{1.18}$$

dargestellt, wobei

$$\begin{aligned} M_p^\varepsilon(\eta;\xi) &= \int_\xi^\eta \int_\xi^{y_1} \dots \int_\xi^{y_{p-1}} N^\varepsilon(y_1) \dots N^\varepsilon(y_p) \mathrm{d}y_p \dots \mathrm{d}y_1 = \\ &= \int_\xi^\eta N^\varepsilon(y) M_{p-1}^\varepsilon(y,\xi) \mathrm{d}y, \ M_0^\varepsilon = \mathrm{Id}. \end{aligned}$$

Um (1.17) numerisch zu lösen, wird das Intervall $(0;1)$ in $T-1$ Teilintervalle der Länge $h = x_{t+1} - x_t$ mit $0 = x_1 < x_2 < \ldots < x_T = 1$ zerlegt. Ausgehend von (1.18) lässt sich $Z(x_{t+1})$ aus $Z(x_t)$ mit

$$Z(x_{t+1}) = Z(x_t) + \sum_{p=1}^{\infty} \varepsilon^p M_p^\varepsilon(x_{t+1};x_t) Z(x_t) \tag{1.19}$$

berechnen. In [ABN] wird die Entwicklung (1.19) bei $P = 1$ bzw. $P = 2$ abgebrochen, um Verfahren 1. bzw. 2. Ordnung in h zu erhalten. Für das Verfahren 1. Ordnung wird M_1^ε nach Potenzen von h und ε entwickelt und man erhält insgesamt ein Verfahren 1. Ordnung für (1.17) mit $Z_1 = Z_I$ und $t = 1, \ldots, T-1$ durch

$$Z_{t+1} = \left(I + A_t^1\right) Z_t. \tag{1.20}$$

Dabei bezeichne I die 2×2 Einheitsmatrix und $A_t^1 \in \mathbb{C}^{2\times 2}$ sei gegeben durch

$$A_t^1 := \tag{1.21}$$

$$\varepsilon^3 \beta_1(x_{t+1}) \begin{pmatrix} 0 & e^{-\frac{2\imath}{\varepsilon}\phi(x_t)} H_1\left(-\frac{2}{\varepsilon} S_t\right) \\ e^{\frac{2\imath}{\varepsilon}\phi(x_t)} H_1\left(\frac{2}{\varepsilon} S_t\right) & 0 \end{pmatrix}$$

$$-\imath\varepsilon^2 \begin{pmatrix} 0 & \beta_0(x_t) e^{-\frac{2\imath}{\varepsilon}\phi(x_t)} - \beta_0(x_{t+1}) e^{-\frac{2\imath}{\varepsilon}\phi(x_{t+1})} \\ \beta_0(x_{t+1}) e^{\frac{2\imath}{\varepsilon}\phi(x_{t+1})} - \beta_0(x_t) e^{\frac{2\imath}{\varepsilon}\phi(x_t)} & 0 \end{pmatrix},$$

mit

$$H_k(x) := e^{\imath x} - \sum_{p=0}^{k-1} \frac{(\imath x)^p}{p!}, \tag{1.22}$$

insbesondere

$$H_1(x) = e^{\imath x} - 1, \tag{1.23}$$

$$H_2(x) = e^{\imath x} - \imath x - 1, \tag{1.24}$$

$$\beta_0(x) := \frac{\beta}{2\phi'}(x),\ \beta_{k+1}(x) := \frac{1}{2\phi'(x)}\beta_k'(x) \tag{1.25}$$

und

$$S_t := \phi\left(x_{t+1}\right) - \phi\left(x_t\right) = \int_{x_t}^{x_{t+1}} \left(\sqrt{a(\tau)} - \varepsilon^2\beta(\tau)\right) \mathrm{d}\tau.$$

Entwickelt man zusätzlich noch M_2^ε nach h und ε, erhält man ein Verfahren 2. Ordnung, wieder mit $Z_1 = Z_I$ und $t = 1, \ldots, T-1$, durch

$$Z_{t+1} = \left(I + A_{mod,t}^1 + A_t^2\right) Z_t, \tag{1.26}$$

mit

$$A_{mod,t}^1 := \tag{1.27}$$

$$\begin{aligned}
&-\imath\varepsilon^2 \begin{pmatrix} 0 & \beta_0\left(x_t\right) e^{-\frac{2\imath}{\varepsilon}\phi(x_t)} - \beta_0\left(x_{t+1}\right) e^{-\frac{2\imath}{\varepsilon}\phi(x_{t+1})} \\ \beta_0\left(x_{t+1}\right) e^{\frac{2\imath}{\varepsilon}\phi(x_{t+1})} - \beta_0\left(x_t\right) e^{\frac{2\imath}{\varepsilon}\phi(x_t)} & 0 \end{pmatrix} \\
&+\varepsilon^3 \begin{pmatrix} 0 & \beta_1\left(x_{t+1}\right) e^{-\frac{2\imath}{\varepsilon}\phi(x_{t+1})} - \beta_1\left(x_t\right) e^{-\frac{2\imath}{\varepsilon}\phi(x_t)} \\ \beta_1\left(x_{t+1}\right) e^{\frac{2\imath}{\varepsilon}\phi(x_{t+1})} - \beta_1\left(x_t\right) e^{\frac{2\imath}{\varepsilon}\phi(x_t)} & 0 \end{pmatrix} \\
&+\imath\varepsilon^4\beta_2\left(x_{t+1}\right) \begin{pmatrix} 0 & -e^{-\frac{2\imath}{\varepsilon}\phi(x_t)} H_1\left(-\frac{2}{\varepsilon}S_t\right) \\ e^{\frac{2\imath}{\varepsilon}\phi(x_t)} H_1\left(\frac{2}{\varepsilon}S_t\right) & 0 \end{pmatrix} \\
&-\varepsilon^5\beta_3\left(x_{t+1}\right) \begin{pmatrix} 0 & e^{-\frac{2\imath}{\varepsilon}\phi(x_t)} H_2\left(-\frac{2}{\varepsilon}S_t\right) \\ e^{\frac{2\imath}{\varepsilon}\phi(x_t)} H_2\left(\frac{2}{\varepsilon}S_t\right) & 0 \end{pmatrix}
\end{aligned}$$

und

$$\begin{aligned}
A_t^2 := &-\varepsilon^3(x_{t+1} - x_t)\tfrac{\beta(x_{t+1})\beta_0(x_{t+1}) - \beta(x_t)\beta_0(x_t)}{2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \\
&-\varepsilon^4\beta_0\left(x_t\right)\beta_0\left(x_{t+1}\right) \begin{pmatrix} H_1\left(-\frac{2}{\varepsilon}S_t\right) & 0 \\ 0 & H_1\left(\frac{2}{\varepsilon}S_t\right) \end{pmatrix} \\
&+\imath\varepsilon^5\beta_1\left(x_{t+1}\right)\left[\beta_0\left(x_t\right) - \beta_0\left(x_{t+1}\right)\right] \begin{pmatrix} H_2\left(-\frac{2}{\varepsilon}S_t\right) & 0 \\ 0 & -H_2\left(\frac{2}{\varepsilon}S_t\right) \end{pmatrix}.
\end{aligned} \tag{1.28}$$

Um aus der mit (1.20) bzw. (1.26) berechneten Lösung von (1.17) eine Lösung für (1.12) zu erhalten, führt man eine Rücktransformation

$$U_t = P^{-1} e^{\frac{\imath}{\varepsilon}\Phi^\varepsilon(x_t)} Z_t \tag{1.29}$$

durch.
Die Hauptaussage von [ABN] ist folgende Fehlerabschätzung für (1.20), (1.26) und (1.29).

Satz 1.1 (Theorem 3.1 aus [ABN]). *Sei $a(x) \in C^{\infty}[0;1]$ eine beliebige, glatte und reellwertige Funktion, die $a(x) \geq a_0 > 0$ im Intervall $[0;1]$ erfülle, und $0 < \varepsilon < 1$ eine beliebige, aber feste, reelle Zahl, dann gilt für den Fehler im Verfahren 1. Ordnung (1.20), - wobei $Z(x_t)$ und $U(x_t)$ die exakten Lösungen bezeichnen und $\|\cdot\|$ eine beliebige Vektornorm in $\mathbb{C}^2$ sei -*

$$\|Z(x_t) - Z_t\| \leq C\varepsilon^2 \min(\varepsilon, h),\ 1 \leq t \leq T, \tag{1.30}$$

$$\|U(x_t) - U_t\| \leq C\frac{h^{\gamma}}{\varepsilon} + C\varepsilon^2 \min(\varepsilon, h),\ 1 \leq t \leq T. \tag{1.31}$$

Für den Fehler im Verfahren 2. Ordnung (1.26) gilt

$$\|Z(x_t) - Z_t\| \leq C\varepsilon^3 h^2,\ 1 \leq t \leq T, \tag{1.32}$$

$$\|U(x_t) - U_t\| \leq C\frac{h^{\gamma}}{\varepsilon} + C\varepsilon^3 h^2,\ 1 \leq t \leq T, \tag{1.33}$$

wobei C unabhängig von t, h und ε ist und γ die Ordnung der Quadratur für die numerische Berechnung von (1.16) ist.

Beweis. siehe [ABN, p. 1451f] □

Kapitel 2

Spektralmethoden

Spektralmethoden basieren auf der Idee, die gesuchte Lösungsfunktion $f(x)$ durch eine endliche Linearkombination von „Basisfunktionen“ $\phi_n(x)$,

$$f(x) \approx f_N(x) = \sum_{n=0}^{N} a_n \phi_n(x), \tag{2.1}$$

mit zu bestimmenden Koeffizienten a_n, anzunähern. Setzt man die Ansatzfunktion aus (2.1) in die Gleichung ein, deren Lösung gesucht ist, erhält man Bestimmungsgleichungen für die Koeffizienten a_n.
Zum Beispiel wird damit in [GO, example 1.1] die Lösung der eindimensionalen Wärmeleitungsgleichung

$$\begin{cases} \dfrac{\partial u(x,t)}{\partial t} = \dfrac{\partial^2 u(x,t)}{\partial x^2}, & 0 < x < \pi,\ t \geq 0, \quad (2.2a) \\ u(0,t) = u(\pi,0) = 0, & t > 0, \quad (2.2b) \\ u(x,0) = f(x), & 0 \leq x \leq \pi, \quad (2.2c) \end{cases}$$

genähert. Der Ansatz $u_N = \sum_{n=1}^{N} a_n(t) \sin nx$ eingesetzt in (2.2a) liefert

$$\sum_{n=1}^{N} \frac{\mathrm{d}a_n(t)}{\mathrm{d}t} \sin nx = -n^2 \sum_{n=1}^{N} a_n(t) \sin nx.$$

Daraus ergeben sich die Bestimmungsgleichungen für die Koeffizienten

$$\frac{\mathrm{d}a_n(t)}{\mathrm{d}t} = -n^2 a_n(t)$$

mit Anfangsbedingungen $a_n(0) = f_n$, $n = 1, \ldots, N$. Dabei sind f_n die Koeffizienten der Fourier-Sinus-Reihe der Anfangsbedingung (2.2c).

Eine weitere Möglichkeit, die Koeffizienten a_n zu bestimmen, ist zu fordern, dass, wenn die anzunähernde Funktion an Stellen x_j bekannt ist,

$$f_N(x_j) = \sum_{n=0}^{N} a_n \phi_n(x_j) = f(x_j),\ j = 0, \ldots, N \tag{2.3}$$

an diesen Kollokationspunkten x_j gilt. In der Literatur ([Bo1], [GHO], [GO]) werden diese Verfahren als Kollokations- oder Pseudospektralmethoden bezeichnet.
Für im Intervall $[-1; 1]$ definierte Funktionen werden hier als Kollokationspunkte die Punkte

$$x_j = \cos \frac{\pi j}{N},\ j = 0, \ldots, N, \tag{2.4}$$

und für die „Basisfunktionen“ $\phi_n(x) = T_n(x)$ die Chebyshevpolynome vom Grad n gewählt. Bei der Wahl dieser Basisfunktionen nennt man (2.1) eine Entwicklung der Funktion $f(x)$ in eine Chebyshevreihe. Die Punkte (2.4) werden im Weiteren als Chebyshevpunkte bezeichnet.

2.1 Chebyshevpolynome und Chebyshevreihen

Zuerst wird die Definition der Chebyshevpolynome und eine Zusammenfassung ihrer wichtigsten Eigenschaften angegeben, die, ebenso wie die Aussagen über die Entwicklung in Chebyshevreihen, aus [Ri] und [FP] entnommen sind.

Definition 2.1 (Chebyshevpolynome 1. Art)**.** Das Chebyshevpolynom vom Grad n oder auch $n-$tes Chebyshevpolynom ist definiert als

$$T_n(x) := \cos(n \arccos x) = \cos(n \cos^{-1} x),\ -1 \leq x \leq 1 \text{ und } n \in \mathbb{N}. \tag{2.5}$$

Bemerkung 2.2 (Chebyshevpolynome 2. Art)*.* Neben den in Definition 2.1 definierten Chebyshevpolynomen findet man in der Literatur auch noch die Polynome

$$U_{n-1} := \frac{\sin(n \arccos x)}{\sin(\arccos x)},\ -1 \leq x \leq 1 \text{ und } n \in \mathbb{N}, \tag{2.6}$$

die ebenfalls als Chebyshevpolynome bezeichnet werden. Im Weiteren bezieht sich die Bezeichnung Chebyshevpolynom immer auf die Chebyshevpolynome 1. Art.
Die Punkte (2.4) sind die Nullstellen der Chebyshevpolynome 2. Art und werden deshalb auch Chebyshevpunkte 2. Art genannt. Die Chebyshevpunkte 1. Art sind die Nullstellen der Chebyshevpolynome 1. Art.

Satz 2.3 (Eigenschaften von Chebyshevpolynomen)**.** *Für die in Definition 2.1 definierten Chebyshevpolynome gelten folgende Aussagen:*

1. $$T_n(\cos\theta) = \cos n\theta,\ \theta \in [0;\pi].$$

2. *Für $n \geq 1$ und $-1 \leq x \leq 1$ gilt die Rekursion*

$$\begin{aligned} &T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \\ &\text{mit } T_0(x) = 1 \text{ und } T_1(x) = x. \end{aligned} \tag{2.7}$$

3. *Für $n \geq 1$ ist der Koeffizient von x^n in $T_n(x)$ gleich 2^{n-1}.*

4. *Für alle x aus dem Intervall $[-1;1]$ gilt:*

$$|T_n(x)| \leq 1.$$

5. *Das Polynom $T_n(x)$ hat für $n \geq 1$ im Intervall $[-1;1]$ genau $n+1$ Extremstellen*

$$x_j^{(n)} = \cos\frac{\pi j}{n},\ j = 0,\ldots,n, \tag{2.8}$$

mit

$$T_n(x_j^{(n)}) = (-1)^j\,.$$

6. *Es gelten folgende Beziehungen für die unbestimmten Integrale von Chebyshevpolynomen*

$$\int T_n(y)\mathrm{d}y = \frac{1}{2}\left(\frac{T_{n+1}(x)}{n+1} - \frac{T_{n-1}(x)}{n-1}\right) + C,\ n \geq 2, \tag{2.9}$$

$$\int T_0(y)\mathrm{d}y = T_1(x) + C, \tag{2.10}$$

$$\int T_1(y)\mathrm{d}y = \frac{1}{4}\left(T_0(x) + T_2(x)\right) + C. \tag{2.11}$$

7. *Es gelten die Orthogonalitätsrelationen*

$$\int_{-1}^{1} T_k(x)T_m(x)\tfrac{\mathrm{d}x}{\sqrt{1-x^2}} = \begin{cases} 0, & k \neq m, \quad (2.12a) \\ \frac{\pi}{2}, & k = m \neq 0, \quad (2.12b) \\ \pi, & k = m = 0. \quad (2.12c) \end{cases}$$

Beweis. 1. Folgt aus der Variablentransformation $x = \cos\theta$.

2. Die expliziten Darstellungen von $T_0(x) = 1$ und $T_1(x) = x$ folgen unmittelbar aus (2.5). Für die Rekursion (2.7) verwendet man das Additionstheorem

$$\cos u + \cos v = 2\cos\left(\frac{u+v}{2}\right)\cos\left(\frac{u-v}{2}\right)$$

mit $u = (n+1)\theta$ und $v = (n-1)\theta$. Aus Punkt 1 erhält man mit $x = \cos\theta$

$$\begin{aligned} 2xT_x(x) - T_{n-1}(x) &= 2\cos\theta\cos n\theta - \cos(n-1)\theta \\ &= 2\cos\left(\tfrac{(n+1)\theta+(n-1)\theta}{2}\right)\cos\left(\tfrac{(n+1)\theta-(n-1)\theta}{2}\right) - \cos(n-1)\theta \\ &= \cos(n+1)\theta + \cos(n-1)\theta - \cos(n-1)\theta \\ &= T_{n+1}(x). \end{aligned}$$

3. Für $n = 1$ ist der führende Koeffizient $1 = 2^{1-1}$ nach Punkt 2. Gelte nun, dass $T_n(x)$ Leitkoeffizient 2^{n-1} habe, dann folgt mit (2.7), dass $T_{n+1}(x)$ Leitkoeffizient 2^n hat.

4. Folgt sofort aus Punkt 1.

5. Die Punkte $x_j^{(n)} = \cos\frac{\pi j}{n}$, $j = 0, \ldots, n$ sind alle verschieden und liegen im Intervall $[-1; 1]$. Da $\cos\pi j = (-1)^j$ für alle ganzen Zahlen j gilt, folgt aus Punkt 1

$$T_n(x_j^{(n)}) = \cos n\frac{\pi j}{n} = \cos\pi j = (-1)^j.$$

6. Verwendet man das Additionstheorem

$$\sin u\cos v = \frac{1}{2}\left(\sin(v+u) - \sin(v-u)\right)$$

mit $u = \theta$ und $v = n\theta$, erhält man mit Punkt 1

$$\begin{aligned} \int T_n(x)\mathrm{d}x &= -\int \cos n\theta\sin\theta\mathrm{d}\theta = -\frac{1}{2}\int\left(\sin(n+1)\theta - \sin(n-1)\theta\right)\mathrm{d}\theta \\ &= \frac{1}{2}\left(\frac{\cos(n+1)\theta}{n+1} - \frac{\cos(n-1)\theta}{n-1}\right) \\ &= \frac{1}{2}\left(\frac{T_{n+1}(x)}{n+1} - \frac{T_{n-1}(x)}{n-1}\right),\ n \geq 2. \end{aligned}$$

(2.10) folgt aus der expliziten Darstellung von $T_0(x)$ und $T_1(x)$ aus Punkt 2. Mit der Darstellung $T_2(x) = 2x^2 - 1$ erhält man
$\int T_1(x)\mathrm{d}x = \frac{x^2}{2} = \frac{1}{4}\left(T_0(x) + T_2(x)\right)$.

7. Unter Verwendung des Additionstheorems

$$\cos u \cos v = \frac{1}{2}\left(\cos(u+v) + \cos(u-v)\right)$$

erhält man mit von Null verschiedenen natürlichen Zahlen $k \neq m$, $u = k\theta$ und $v = m\theta$

$$\begin{aligned}\int_0^\pi \cos k\theta \cos m\theta \mathrm{d}\theta &= \frac{1}{2}\int_0^\pi \left(\cos(k+m)\theta + \cos(k-m)\theta\right)\mathrm{d}\theta \\ &= \frac{1}{2}\left(\frac{\sin(k+m)\theta}{k+m} + \frac{\sin(k-m)\theta}{k-m}\right)\Bigg|_0^\pi = 0.\end{aligned}$$

Für $k = m \neq 0$ ergibt sich

$$\int_0^\pi \cos^2 k\theta \mathrm{d}\theta = \frac{1}{2}\int_0^\pi (1+\cos 2k\theta)\,\mathrm{d}\theta = \frac{1}{2}\left(\theta + \frac{\sin 2k\theta}{2k}\right)\Bigg|_0^\pi = \frac{\pi}{2}$$

und für $k = m = 0$ erhält man

$$\int_0^\pi 1 \mathrm{d}\theta = \pi.$$

Unter Verwendung von Punkt 1 erhält man aus diesen Beziehungen

$$\int_{-1}^1 T_k(x)T_m(x)\tfrac{\mathrm{d}x}{\sqrt{1-x^2}} = \int_0^\pi \cos k\theta \cos m\theta \mathrm{d}\theta = \begin{cases} 0, & k \neq m, \\ \frac{\pi}{2}, & k = m \neq 0, \\ \pi, & k = m = 0. \end{cases}$$

□

Da Spektralmethoden auf der Entwicklung der Lösungsfunktion in eine Reihe beruhen, wird hier in Chebyshevreihen entwickelt. Das bedeutet, die Funktion in der Form

$$f(x) \sim {\sum_{k=0}^{\infty}}' b_k T_k(x) \tag{2.14}$$

mit zu bestimmenden Koeffizienten b_k darzustellen.

Bemerkung 2.4. Im Weiteren werden die abkürzenden Notationen

$$\begin{aligned}{\sum_{j=0}^{N}}' a_j &:= \frac{a_0}{2} + a_1 + \ldots + a_{N-1} + a_N \text{ bzw.} \\ {\sum_{j=0}^{\infty}}' a_j &:= \frac{a_0}{2} + a_1 + \ldots \text{ und} \\ {\sum_{j=0}^{N}}'' a_j &:= \frac{a_0}{2} + a_1 + \ldots + a_{N-1} + \frac{a_N}{2}\end{aligned}$$

verwendet.

Wie der nächste Satz zeigt, lassen sich Polynome in eine endliche Chebyshevreihe entwickeln.

Satz 2.5. *Jedes Polynom*

$$p(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n$$

kann in der Form

$$p(x) = \frac{b_0}{2} + b_1 T_1(x) + b_2 T_2(x) + \ldots + b_n T_n(x) \tag{2.15}$$

$$\text{mit } b_k = \frac{2}{\pi} \int_{-1}^{1} p(x) T_k(x) \frac{\mathrm{d}x}{\sqrt{1-x^2}},\ k = 0, \ldots, n \tag{2.16}$$

dargestellt werden. Es gilt weiters, dass $b_n = 2^{-(n-1)} a_n$.

Beweis. Dass $b_n = 2^{-(n-1)} a_n$ ist, folgt aus Satz 2.3 Punkt 3. Für die Darstellung (2.16) multipliziert man (2.15) mit $T_m(x) \frac{1}{\sqrt{1-x^2}}$ und integriert anschließend von -1 bis 1,

$$\int_{-1}^{1} p(x) T_m(x) \frac{\mathrm{d}x}{\sqrt{1-x^2}} = \sum_{k=0}^{n}{}' b_k \int_{-1}^{1} T_k(x) T_m(x) \frac{\mathrm{d}x}{\sqrt{1-x^2}}.$$

Verwendet man die Orthogonalitätsrelationen (2.12) erhält man

$$b_k = \frac{2}{\pi} \int_{-1}^{1} p(x) T_k(x) \frac{\mathrm{d}x}{\sqrt{1-x^2}}.$$

□

Nach [Ri, Chapter 3] können auf $[-1; 1]$ integrierbare Funktionen in eine Chebyshevreihe in der Form

$$f(x) \sim \sum_{k=0}^{\infty}{}' b_k T_k(x) = \lim_{N \to \infty} \sum_{k=0}^{N}{}' b_k T_k(x) \tag{2.17}$$

mit Koeffizienten

$$b_k = \frac{2}{\pi} \int_{-1}^{1} f(x) T_k(x) \frac{\mathrm{d}x}{\sqrt{1-x^2}} \tag{2.18}$$

entwickelt werden. Bricht man die Entwicklung von (2.17) nach N Gliedern ab, erhält man eine Näherung $f_N(x)$, die die Funktion im Sinne der kleinsten Quadrate

$$\int_{-1}^{1} \left(f(x) - f_N((x))\right)^2 \frac{\mathrm{d}x}{\sqrt{1-x^2}} \to \min$$

approximiert (siehe dazu [FP, section 2.13]).
Hat $f(x)$ Unstetigkeiten, tritt bei $f_N(x)$ an diesen Stellen das Gibbs-Phänomen auf, ein Überschwingen des Funktionswerts an der Stelle der Unstetigkeit. Somit kann (2.17) nicht für alle Funktionen gleichmäßig konvergieren.
Damit die Reihe (2.17) gleichmäßig gegen die Funktion $f(x)$ konvergiert, müssen zusätzliche Forderungen gestellt werden. Im Entwicklungssatz, [HH, Abschnitt 4.4.8], wird $f \in C^2[-1;1]$ gefordert, dann konvergiert (2.17) gleichmäßig, und für die Koeffizienten (2.18) gilt die Abschätzung

$$|b_k| \leq \frac{C}{k^2},$$

wobei C nur von f abhängt. In [Bo1, Theorem 4] wird die Abschätzung für die Fourier-Koeffizienten einer Funktion, die

$$f(\pi) = f(-\pi),\ f'(\pi) = f'(-\pi), \ldots, f^{(n-2)}(\pi) = f^{(n-2)}(-\pi)$$

erfüllt und deren $n-$te Ableitung integrierbar ist, erweitert. Da die Chebyshevreihe einer Funktion $f(x)$ die Fourier-Cosinus-Reihe von $f(\cos\theta)$ ist, erhält man damit eine Abschätzung

$$|b_k| \leq \frac{C}{k^n}$$

und somit sehr rasche Konvergenz (siehe [Bo1, Section 2.11]).
In [Ri, Theorem 3.4] ist mit

$$\lim_{N\to\infty} \omega\left(\frac{1}{N}\right) \log N = 0 \tag{2.19}$$

eine Bedingung für stetige Funktionen angegeben, damit (2.17) gleichmäßig konvergiert. Dabei ist

$$\omega(f;[a;b];\delta) = \omega(\delta) = \sup_{\substack{x,y\in[a;b]\\ |x-y|\leq\delta}} |f(x) - f(y)|$$

das Stetigkeitsmodul. Diese Bedingung (2.19) wird von allen hölderstetigen Funktionen mit beliebigem Hölder-Exponenten $0 < \alpha \leq 1$ erfüllt (siehe [Tr2, p. 75] und [Ri, p. 135]).

Da die Koeffizienten in (2.18) im Allgemeinen umständlich zu berechnen sind, erhält man durch die Substitution $x = \cos\theta$ den Ausdruck

$$b_k = \frac{2}{\pi}\int_0^\pi \cos k\theta f(\cos\theta)\mathrm{d}\theta.$$

Wendet man darauf die Trapezregel mit $h = \frac{\pi}{N}$ und den Stützstellen $\theta_r = \frac{r\pi}{N}$, $r = 0, \ldots, N$ an, erhält man

$$\begin{aligned} b_k \approx a_k &= \frac{2}{\pi}\frac{\pi}{N}\left(\frac{1}{2}f(1) + \frac{1}{2}\cos k\pi f(\cos k\pi) + \sum_{r=1}^{N-1}\cos k\theta_r f(\cos\theta_r)\right) \\ &= \frac{2}{N}{\sum_{r=0}^{N}}'' \cos k\theta_r f(\cos\theta_r) \\ &= \frac{2}{N}{\sum_{r=0}^{N}}'' T_k(x_r) f(x_r). \end{aligned}$$

Insgesamt erhält man die Näherung

$$\begin{cases} f_N(x) = {\sum_{k=0}^{N}}'' a_k T_k(x), \\ a_k = \frac{2}{N}{\sum_{r=0}^{N}}'' f(x_r) T_k(x_r), \quad x_r = \cos\frac{r\pi}{N}. \end{cases} \tag{2.20}$$

Die Näherung (2.20) ist ein Polynom vom Grad N, das an den Stellen x_r mit der Funktion $f(x)$ übereinstimmt. Der Abschneidefehler, der durch die Verwendung von (2.20) statt $f(x)$ entsteht, ist in [FP, Section 4.4] mit

$$\left| f(x) - {\sum_{k=0}^{N}}'' a_k T_k(x) \right| \leq 2 \sum_{k=N+1}^{\infty} |b_k| \tag{2.21}$$

angegeben. Da für die Konvergenz von (2.17)

$$\lim_{k\to\infty} b_k = 0$$

notwendig ist, erhält man mit (2.21) aus der Konvergenz von (2.17) die Konvergenz von (2.20) für $N \to \infty$.
Unter den Voraussetzungen, dass $f(z)$ auf dem Abschluss eines Gebietes $G \subseteq \mathbb{C}$ analytisch ist und die paarweise verschiedenen Punkte z_j, $j = 0, \ldots, n$, innerhalb der geschlossenen Kurve $\mathcal{C}$ in G liegen, ist in [Da, Theorem 3.6.1] mit

$$f(x) - p_n(x) = \frac{1}{2\pi\imath}\int_{\mathcal{C}} \frac{(x-z_0)(x-z_1)\ldots(x-z_n)f(z)\mathrm{d}z}{(z-z_0)(z-z_1)\ldots(z-z_n)(z-x)} \tag{2.22}$$

der Fehler angegeben, wenn man f durch p_n an den Stellen z_j interpoliert. Für den Fehler durch (2.20) gilt der folgende

Satz 2.6. *Es sei $f(z)$ auf dem Abschluss des Gebietes G analytisch und liegen die Punkte $x_r = \cos\frac{\pi r}{N}$ innerhalb der geschlossenen Kurve $\mathcal{C}$ in G. Dann gilt für den Fehler*

$$e_N(x) := f(x) - f_N(x), \tag{2.23}$$

$$e_N(x) = \frac{T_{N+1}(x) - T_{N-1}(x)}{2\pi\imath} \int_{\mathcal{C}} \frac{f(z)\mathrm{d}z}{(z-x)(T_{N+1}(z) - T_{N-1}(z))}. \tag{2.24}$$

Lemma 2.7. *Die Stellen $x_r = \cos\frac{\pi r}{N}$ sind die Nullstellen von $T_{N+1}(x) - T_{N-1}(x)$.*

Beweis. Aus Satz 2.3 Punkt 1 und dem Additionstheorem

$$\cos(u \pm v) = \cos u \cos v \mp \sin u \sin v$$

mit $u = N\theta$ und $v = \theta$ folgt

$$\begin{aligned} T_{N+1}(x) - T_{N-1}(x) &= \cos(N+1)\theta - \cos(N-1)\theta \\ &= \cos(N\theta + \theta) - \cos(N\theta - \theta) \\ &= \cos N\theta\cos\theta - \sin N\theta\sin\theta - \cos N\theta\cos\theta - \sin N\theta\sin\theta \\ &= -2\sin N\theta\sin\theta. \end{aligned}$$

Da $x = \cos\theta$, folgt mit $\theta_r = \frac{\pi r}{N}$ die Behauptung. □

Beweis von Satz 2.6. Die Darstellung (2.24) folgt aus (2.22) mit Hilfe von Lemma 2.7. □

Bemerkung 2.8. Polynome vom Grad N lassen sich auch durch (2.20) in eine Chebyshevreihe entwickeln. Dabei gilt dann für den Koeffizienten $\frac{a_N}{2} = b_N$ nach Satz 2.5.

Funktionen $f(x)$, die auf einem Intervall $[a; b]$ definiert und integrierbar sind, lassen sich durch die Transformation

$$x = \frac{1}{2}\left(b + a + t(b-a)\right), t \in [-1; 1] \tag{2.25}$$

in eine Funktion $F(t)$ überführen. Die Funktion $F(t)$ lässt sich in eine Chebyshevreihe entwickeln.

Beispiel 2.9. Sei $f(x) = x + \frac{1}{2}$, $x \in [0; 1]$. Die Transformation (2.25) ergibt

$$f(x) = F(t) = \frac{t}{2} + 1.$$

Nach Satz 2.5 hat $F(t)$ eine endliche Chebysheventwicklung mit Koeffezienten $b_0 = a_0 = 2$ und $b_1 = \frac{1}{2}$ bzw. nach Bemerkung 2.8 $a_1 = 1$. Die Darstellung $F(t) = T_0(t) + \frac{1}{2}T_1(t)$ folgt auch sofort aus Satz 2.3 Punkt 2.

2.2 Clenshaw-Curtis Quadratur

Die Quadratur von Clenshaw und Curtis [CC] basiert darauf, dass Funktionen, die auf $[a;b]$ integrierbar sind, durch (2.25) in das Intervall $[-1;1]$ transformiert werden und durch (2.20) in eine Chebyshevreihe entwickelt werden. Dabei werden die zu integrierende Funktion $f(x)$ und ihre Stammfunktion $\int_a^x f(\tau)\mathrm{d}\tau$ in eine Chebyshevreihe entwickelt und die Koeffizienten der Stammfunktion aus den Koeffizienten des Integranden berechnet.
Mit (2.25) gilt

$$\begin{aligned}\frac{2}{b-a}\int_a^x f(\tau)\mathrm{d}\tau &= \int_{-1}^t F(u)\mathrm{d}u = \int_{-1}^t \left(\frac{a_0}{2} + a_1T_1(u) + a_2T_2(u) + \ldots\right)\mathrm{d}u \\ &= \frac{b_0}{2} + b_1T_1(t) + b_2T_2(t) + \ldots\ . \end{aligned} \tag{2.26}$$

Verwendet man Satz 2.3 Punkt 6, lassen sich die Koeffizienten b_n der Entwicklung der Stammfunktion aus den Koeffizienten a_n bestimmen. Aus (2.9), (2.10) und (2.11) erhält man durch gliedweise Integration eine Stammfunktion

$$\begin{aligned}&\int \left(\frac{a_0}{2} + a_1T_1(t) + a_2T_2(t) + \ldots + a_{N-1}T_{N-1}(t) + \frac{a_N}{2}T_N(t)\right)\mathrm{d}t \\ =\ & \frac{a_0}{2}T_1(t) + \frac{a_1}{4}\left(T_0(t) + T_2(t)\right) + \frac{a_2}{2}\left(\frac{T_3(t)}{3} - T_1(t)\right) + \ldots + \\ & + \frac{a_{N-1}}{2}\left(\frac{T_N(t)}{N} - \frac{T_{N-2}(t)}{N-2}\right) + \frac{a_N}{2}\left(\frac{T_{N+1}(t)}{N+1} - \frac{T_{N-1}(t)}{N-1}\right) \\ =\ & \frac{a_0}{2}T_1(t) + \frac{a_1}{4}T_0(t) + \frac{a_1}{2}T_2(t) + \frac{a_2}{6}T_3(t) - \frac{a_2}{2}T_1(t) + \ldots + \\ & + \frac{a_{N-1}}{2N}T_N(t) - \frac{a_{N-1}}{2N-2}T_{N-2}(t) + \frac{a_N}{2N+2}T_{N+1}(t) - \frac{a_N}{2N-2}T_{N-1}(t) \\ =\ & \frac{b_0}{2} + b_1T_1(t) + b_2T_2(t) + b_3T_3(t) + \ldots + b_NT_N(t) + b_{N+1}T_{N+1}(t)\end{aligned} \tag{2.27}$$

und Koeffizientenvergleich liefert

$$b_1 = \frac{a_0 - a_2}{2}, \ldots, b_k = \frac{a_{k-1} - a_{k+1}}{2k}, \ldots, b_N = \frac{a_{N-1}}{2N}, b_{N+1} = \frac{a_N}{2N+2}. \tag{2.28}$$

Die Konstante b_0 wird so bestimmt, dass die Stammfunktion an der Stelle $t = -1$ den Wert 0 annimmt. Aus Satz 2.3 Punkt 5 folgt $T_n(-1) = (-1)^n$. Dies eingesetzt in (2.27) liefert

$$0 = \frac{b_0}{2} - b_1 + b_2 - b_3 + b_4 \pm \ldots + (-1)^N b_N + (-1)^{N+1}b_{N+1},$$

somit ergibt sich

$$b_0 = 2b_1 - 2b_2 + 2b_3 - 2b_4 \pm \ldots + (-2)^N b_{N+1}. \tag{2.29}$$

Um das bestimmte Integral $\int_{-1}^{1} F(t)\mathrm{d}t$ zu berechnen, verwendet man, dass $T_n(1) = 1$ und man erhält somit

$$\begin{aligned}\int_{-1}^{1} F(t)\mathrm{d}t &\approx (b_1 + b_2 + b_3 + \ldots + b_{N+1}) - \left(-b_1 + b_2 - b_3 \pm \ldots + (-1)^{N+1} b_{N+1}\right) \\ &= 2\left(b_1 + b_3 + b_5 + \ldots + b_{2K+1}\right), \end{aligned} \tag{2.30}$$

wobei $2K + 1 \leq N + 1 \leq 2K + 2$.
Um aus (2.27) das ursprüngliche bestimmte oder unbestimmte Integral zu erhalten, muss mit $\frac{b-a}{2}$ multipliziert werden,

$$\int_a^x f(\tau)\mathrm{d}\tau \approx \frac{b-a}{2}\left(\frac{b_0}{2} + b_1 T_1(t) + \ldots + b_N T_N(t) + b_{N+1} T_{N+1}(t)\right). \tag{2.31}$$

Beispiel 2.10 (Fortsetzung von Beispiel 2.9)**.** Es sollen $\int_0^x \left(\tau + \frac{1}{2}\right)\mathrm{d}\tau$ und $\int_0^1 \left(x + \frac{1}{2}\right)\mathrm{d}x$ berechnet werden. Nach Beispiel 2.9 hat die Entwicklung von $f(x)$ die Koeffizienten $a_0 = 2$ und $a_1 = 1$. Mit (2.28) und (2.29) hat die Stammfunktion die Koeffizienten $b_0 = 2$ und $b_1 = 1$. Die genäherte Stammfunktion hat somit die Form

$$\int_0^x \left(\tau + \frac{1}{2}\right)\mathrm{d}\tau \approx \frac{1}{2}\left(1 + T_1(t)\right).$$

Das bestimmte Integral hat den Wert

$$\int_0^1 \left(x + \frac{1}{2}\right)\mathrm{d}x = \frac{1}{2} \cdot 2 \cdot 1 = 1.$$

Dieser Wert stimmt mit dem exakten Wert überein.
Eine bessere Näherung der Stammfunktion ergibt eine Entwicklung mit (2.20) bis zum Grad 2. Die Koeffizienten sind dann $a_0 = 2$, $a_1 = \frac{1}{2}$ und $a_2 = 0$. Diese Entwicklung ändert nichts an der Darstellung der Funktion als endliche Chebyshevreihe, jedoch ergeben sich jetzt mit (2.28) und (2.29) andere Werte für die Koeffizienten der Stammfunktion, und zwar $b_1 = 1$, $b_2 = \frac{1}{8}$ und $b_0 = \frac{7}{4}$. Somit erhält man eine Stammfunktion

$$\int_0^x \left(\tau + \frac{1}{2}\right)\mathrm{d}\tau = \frac{1}{2}\left(\frac{7}{8} + T_1(t) + \frac{1}{8}T_2(t)\right).$$

Verwendet man die explizite Darstellung der Chebyshevpolynome aus Satz 2.3 Punkt 2, erhält man

$$\int_0^x \left(\tau + \frac{1}{2}\right) \mathrm{d}\tau = \frac{1}{2}\left(\frac{7}{8} + t + \frac{1}{8}\left(2t^2 - 1\right)\right) = \frac{t^2}{8} + \frac{t^2}{2} + \frac{3}{8}.$$

Transformiert man mit $t = 2x - 1$ zurück in das Intervall $[0; 1]$, ergibt sich

$$\int_0^x \left(\tau + \frac{1}{2}\right) \mathrm{d}\tau = \frac{x^2}{2} + \frac{x}{2}.$$

Dies ist genau die Stammfunktion von $f(x)$, die $F(0) = 0$ erfüllt. Eine Entwicklung mit $N \geq 3$ ist nicht weiter sinnvoll, da alle weiteren Koeffizienten den Wert 0 haben. Jedoch erkennt man, dass die Stammfunktion dieser Funktion für $N \geq 2$ exakt berechnet werden kann.

Dass der Wert des bestimmten Integrals aus Beispiel 2.10 mit dem exakten Wert übereinstimmt, liegt an der Konstruktion der Clenshaw-Curtis Quadratur. Da nach Satz 2.5 jedes Polynom als Linearkombination von Chebyshevpolynomen dargestellt werden kann, folgt sofort

Satz 2.11. *Sei p_n ein Polynom vom Grad n, dann stimmt der mit der Clenshaw-Curtis Quadratur berechnete Wert mit dem exakten Wert von*

$$\int_{-1}^{1} p_n(x)\mathrm{d}x$$

überein.

Da in den Kapiteln 4 und 5 Funktionen betrachtet werden, bei denen der Integrand in (1.2),

$$\phi(x) = \int_0^x \left(\sqrt{a(\tau)} - \varepsilon^2\beta(\tau)\right) \mathrm{d}\tau,$$

analytisch ist, wird eine Abschätzung des Fehlers angegeben, der bei der Berechnung von (1.2) mit der Clenshaw-Curtis Quadratur auftritt.

2.2.1 Fehlerabschätzung für die Clenshaw-Curtis Quadratur für analytische Funktionen

Die folgende Abschätzung des Quadraturfehlers für analytische Funktionen stammt von Chawla [Ch]. Auch der Beweis der Abschätzung wurde daraus

entnommen.
Für die Abschätzung benötigt man zuerst die Abbildung

$$\xi = \rho e^{\imath\varphi} \mapsto z = \frac{1}{2}\left(\xi + \xi^{-1}\right),\ 0 \leq \varphi \leq 2\pi, \tag{2.32}$$

die Kreise mit Radien ρ bzw. $\frac{1}{\rho}$ auf Ellipsen $\mathcal{E}_\rho$ mit Brennpunkten in ± 1 und Summe der Halbachsen ρ abbildet. Das Intervall $[-1;1]$ ist dabei genau das Bild des Einheitskreises (siehe [Ri, p.143]).
Weiters bezeichne E_N den Fehler, der durch die Clenshaw-Curtis Quadratur entsteht, wobei mit (2.23)

$$E_N = \int_{-1}^{1} e_N(x)\mathrm{d}x \tag{2.33}$$

gilt. Mit dieser Notation lässt sich eine Abschätzung für E_N formulieren.

Satz 2.12 (Theorem 1 aus [Ch])**.** *Sei $f(z)$ analytisch auf dem Abschluss der Ellipse $\mathcal{E}_\rho$ mit Brennpunkten in ± 1 und Summe der Halbachsen $\rho > 1$, dann gilt für gerades N*

$$|E_N| \leq \left(\frac{16N^2}{4N^2-1}\right)\frac{M(\rho)}{(\rho^2-1)\left(\rho^N - \rho^{-N}\right)} \tag{2.34}$$

mit $M(\rho) = \max_{z\in\mathcal{E}_\rho} |f(z)|$.

Für den Beweis von Satz 2.12 benötigt man noch ein Lemma und vereinfachende Notation. Mit (2.24) lässt sich (2.33) weiterschreiben als

$$E_N = \frac{1}{\pi\imath}\int_{\mathcal{C}} \frac{\left(Q_{N+1}(z) - Q_{N-1}(z)\right) f(z)\mathrm{d}z}{T_{N+1}(z) - T_{N-1}(z)} \tag{2.35}$$

mit

$$Q_N(z) = \frac{1}{2}\int_{-1}^{1} \frac{T_N(x)\mathrm{d}x}{z-x} \tag{2.36}$$

und einer geschlossen Kurve $\mathcal{C}$ innerhalb derer die Punkte $x_j = \cos\frac{j\pi}{N}$, $0 \leq j \leq N$, liegen.

Lemma 2.13. *Sei $z \in \mathcal{E}_\rho$, mit $\rho > 1$, dann lässt sich $Q_N(z)$ wie folgt darstellen,*

$$Q_N(z) = \xi^{-N-1} \sum_{k=-\left[\frac{N}{2}\right]}^{\infty} \frac{\sigma_{N,N+2k+1}}{\xi^{2k}} \tag{2.37}$$

mit

$$\sigma_{N,N+2k+1} = \frac{2(N+2k+1)}{(2N+2k+1)(2k+1)}.$$

Beweis. Setzt man in (2.36) $x = \cos\theta$ und $z = \frac{1}{2}\left(\xi + \xi^{-1}\right)$, erhält man

$$Q_N(z) = \frac{1}{\xi}\int_0^\pi \frac{\cos N\theta \sin\theta \mathrm{d}\theta}{1 - 2\xi^{-1}\cos\theta + \xi^{-2}}. \tag{2.38}$$

Betrachtet man

$$\Im\left(\sum_{m=0}^{\infty} \frac{\cos m\theta + \imath \sin m\theta}{\xi^m}\right),$$

erhält man

$$\sum_{m=1}^{\infty} \frac{\sin m\theta}{\xi^m} = \frac{\sin\theta}{\xi\left(1 - 2\xi^{-1}\cos\theta + \xi^{-2}\right)}, \tag{2.39}$$

wenn $|\xi| \geq \rho > 1$. Setzt man (2.39) in (2.38) ein, ergibt sich

$$Q_N(z) = \sum_{m=1}^{\infty} \frac{\sigma_{N,m}}{\xi^m}$$

mit

$$\sigma_{N,m} = \int_0^\pi \cos N\theta \sin m\theta \mathrm{d}\theta.$$

Setzt man $v = N\theta$ und $u = m\theta$, erhält man aus dem Additionstheorem

$$\cos v \sin = \frac{1}{2}\left(\sin(v+u) - \sin(v-u)\right),$$

$$\begin{aligned}
\sigma_{N,m} &= \int_0^\pi \left(\sin(N+m)\theta - \sin(N-m)\theta\right) \mathrm{d}\theta \\
&= \frac{1}{2}\left[\frac{\cos(N-m)\theta}{N-m} - \frac{\cos(N+m)\theta}{N+m}\right]\Bigg|_0^\pi \\
&= \frac{1}{2}\left[\frac{(N+m)(-1)^{N-m} - (N-m)(-1)^{N+m} - 2m}{N^2 - m^2}\right].
\end{aligned}$$

Da $N-m$ genau dann gerade ist, wenn $N+m$ gerade ist, und $N-m$ genau dann ungerade ist, wenn $N+m$ ungerade ist, ergibt sich

$$\sigma_{N,m} = \begin{cases} \frac{2m}{m^2-N^2}, & m-N \text{ ist ungerade,} \\ 0, & m+N \text{ ist gerade.} \end{cases}$$

Setzt man $m - N = 2k+1$, $N, m = 1, 2, 3, \ldots$, folgt $k \geq -\left[\frac{N}{2}\right]$ und damit (2.37). □

Korollar 2.14. *Für $z \in \mathcal{E}_\rho$, mit $\rho > 1$, gilt*

$$Q_{N+1}(z) = \xi^{-N} \sum_{k=-\left[\frac{N-1}{2}\right]}^{\infty} \frac{\sigma_{N+1,N+2k}}{\xi^{2k}} \tag{2.40}$$

mit

$$\sigma_{N+1,N+2k} = \frac{2(N+2k)}{(2N+2k+1)(2k-1)}$$

und

$$\sigma_{N+1,N+2k} \leq \frac{2N}{2N+1}.$$

Beweis. Setze $N = N + 1$ in (2.37). □

Korollar 2.15. *Für $z \in \mathcal{E}_\rho$, mit $\rho > 1$, gilt*

$$Q_{N+1}(z) - Q_{N-1}(z) = \xi^{-N} \sum_{k=-\left[\frac{N-1}{2}\right]}^{\infty} \frac{\lambda_{N,k}}{\xi^{2k}} \tag{2.41}$$

mit

$$\lambda_{N,k} = \frac{8N(N+2k)}{(4(N+k)^2-1)\,(4k^2-1)}$$

und

$$\lambda_{N,k} \leq \frac{8N^2}{4N^2-1}. \tag{2.42}$$

Beweis. Ersetzt man in (2.37) N durch $N-1$ und bildet die Differenz mit (2.40), ergibt sich

$$Q_{N+1}(z) - Q_{N-1}(z) = \xi^{-N} \sum_{k=-\left[\frac{N-1}{2}\right]}^{\infty} \frac{\lambda_{N,k}}{\xi^{2k}}$$

mit

$$\begin{aligned} \lambda_{N,k} &= \sigma_{N+1,N+2k} - \sigma_{N-1,N+2k} \\ &= \frac{8N(N+2k)}{(4(N+k)^2-1)\,(4k^2-1)} \end{aligned}$$

und

$$\lambda_{N,k} \leq |\lambda_{N,0}| = \frac{8N^2}{4N^2-1}.$$

□

Beweis von Satz 2.12. Wählt man in (2.35) $\mathcal{C} = \mathcal{E}_\rho$, erhält man

$$|E_N| \leq \frac{1}{\pi} \int_{\mathcal{E}_\rho} \frac{|Q_{N+1}(z) - Q_{N-1}(z)|\,|f(z)|\,|\mathrm{d}z|}{|T_{N+1}(z) - T_{N-1}(z)|}. \tag{2.43}$$

Für gerades N erhält man aus (2.41) und (2.42)

$$\begin{aligned}
|Q_{N+1}(z) - Q_{N-1}(z)| &\leq \rho^{-N} \left(\frac{8N^2}{4N^2-1}\right) \sum_{k=-\frac{N}{2}+1}^{\infty} \rho^{-2k} \\
&= \frac{8N^2}{4N^2-1} \rho^{-N} \left(\sum_{k=-\frac{N}{2}+1}^{-1} \rho^{-2k} + \sum_{k=0}^{\infty} \rho^{-2k} \right) \\
&= \frac{8N^2}{4N^2-1} \left(\rho^{-2} + \rho^{-4} + \ldots + \rho^{2-N} + \frac{\rho^{-N}}{1-\rho^{-2}} \right) \\
&= \frac{8N^2}{4N^2-1} \left(\rho^{-2} \frac{1-(\rho^{-2})^{\frac{N}{2}-1}}{1-\rho^{-2}} + \frac{\rho^{-N}}{1-\rho^{-2}} \right) \\
&= \frac{8N^2}{4N^2-1} \frac{1}{\rho^2-1}.
\end{aligned}$$

Nach [Ri, exercise 2.4.11] und [Ri, exercise 2.4.12] gilt

$$\max_{z \in \mathcal{E}_\rho} |T_N(z)| = \frac{\rho^N + \rho^{-N}}{2}.$$

Damit folgt

$$|T_{N+1}(z) - T_{N-1}(z)| \geq |T_{N+1}(z)| - |T_{N-1}(z)| = \frac{\rho^{N+1} + \rho^{-N-1}}{2} - \frac{\rho^{N-1} + \rho^{-N+1}}{2}.$$

Mit

$$|\mathrm{d}z| = \frac{|\xi^2 - 1|}{|2\xi^2|}\,|\mathrm{d}\xi|$$

und $|\xi| = \rho$ ergibt sich insgesamt

$$\frac{|\mathrm{d}z|}{|T_{N+1}(z) - T_{N+1}(z)|} \leq \frac{|\mathrm{d}\xi|}{\rho^{N+1} - \rho^{-N+1}} = \frac{|\mathrm{d}\xi|}{\rho(\rho^N - \rho^{-N})}.$$

Setzt man dies alles mit $|f(z)| \leq M(\rho)$ in (2.43) zusammen und integriert, erhält man (2.34). □

Bemerkung 2.16. 1. Schon in [Ch] selbst wird angemerkt, dass (2.34) für ρ nahe bei 1 schlecht ist.

2. In (2.34) fällt der Fehler exponentiell, $|E_N| \leq \mathcal{O}(\rho^{-N})$. Dabei handelt es sich um die sogenannte „spectral accuracy “, wie in [Tr1, Chapter 4] beschrieben.

3. Die Voraussetzung „für gerades N“ ist mehr beweistechnischer Natur. Die Abschätzung hält auch bei ungeradem N.

4. In [CC, Section 4] ist schon eine Abschätzung

$$|E_N| \leq 2 \max \{a_{N-4}, a_{N-2}, a_N\}$$

angegeben. Nach [Ri, Theorem 3.8] gilt für die Koeffizienten der Chebysheventwicklung von f, $|a_N| \leq 2M(\rho)\rho^{-N}$, wenn $f(z)$ analytisch innerhalb und auf der Ellipse $\mathcal{E}_\rho$ ist. Wendet man dies auf die Abschätzung aus [CC] an, erhält man

$$|E_N| \leq \frac{4M(\rho)}{\rho^N}. \tag{2.44}$$

Auch hier erkennt man, dass der Fehler exponentiell fällt.

5. Für Funktionen mit $k-1$ absolut stetigen Ableitungen und beschränkter $k-$ter Ableitung findet man in [Tr2, Theorem 5.1] eine Abschätzung für den Fehler durch die Clenshaw-Curtis Quadratur.

6. Für Funktionen, die nicht analytisch sind, aber für die $f \in C^\infty [-1; 1]$ gilt, verhält sich der Fehler durch die Clenshaw-Curtis Quadratur ähnlich wie der Fehler analytischer Funktionen. Siehe dazu das Beispiel $f(x) = e^{-x^{-2}}$ in [Tr1, Output 30b] bzw. [Tr2, Fig. 2].

Beispiel 2.17 (aus [Ch])**.** Sei $f(z) = \frac{1}{z+4}$ analytisch innerhalb und auf der Ellipse mit großer Halbachse $a < 4$. Somit erhält man ein maximales $\rho_{\max} = 4 + \sqrt{15} \approx 7.87$, sodass $f(z)$ auf $\mathcal{E}_{\rho_{\max}}$ nicht mehr analytisch ist, da die Funktion bei $z = -4$ einen Pol erster Ordnung hat. Wählt man $\rho < \rho_{\max}$, ist die Funktion auch auf $\mathcal{E}_\rho$ analytisch. Wie in [Ch] wird auch hier $\rho = 7$ gewählt. Mit $\max_{z \in \mathcal{E}_7} |f(z)| = M(7) = \frac{7}{3}$ erhält man aus (2.34) die Abschätzung

$$|E_N| \leq \frac{7N^2}{9\,(4N^2 - 1)\,(7^N - 7^{-N})} \tag{2.45}$$

für die Berechnung von

$$\int_{-1}^{1} \frac{1}{x+4} \mathrm{d}x = \ln \frac{5}{3}.$$

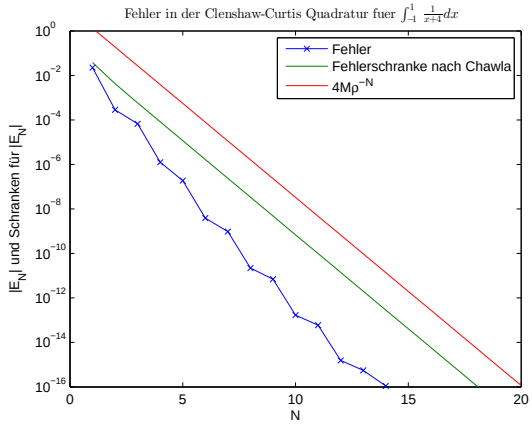

Abbildung 2.1: Beispiel zur Fehlerabschätzung in der Clenshaw-Curtis Quadratur

In Abb. 2.1 sind der wahre Fehler (blau), die Fehlerschranke aus (2.45) (grün) und die Abschätzung (2.44) (rot) gezeigt. Man sieht, dass (2.34) auch für ungerades N funktioniert und dass es eine bessere Schranke für den Fehler ist als (2.44). Beide haben jedoch $|E_N| \leq \mathcal{O}(\rho^{-N})$ gemeinsam.

2.3 Ein weiteres auf Spektralmethoden basierendes Quadraturverfahren

In [Tr1, Chapter 12] wird neben der Clenshaw-Curtis Quadratur noch ein weiteres Quadraturverfahren vorgestellt, das hier ebenfalls kurz präsentiert wird.
In [GHO, Section 4] wird eine Matrix $D_N \in \mathbb{R}^{(N+1)\times(N+1)}$ hergeleitet, mit der man die Ableitung einer Funktion F an Chebyshevpunkten berechnen kann. Diese Matrix hat die Einträge

$$(D_N)_{00} = \frac{2N^2+1}{6}, \quad (D_N)_{NN} = -\frac{2N^2+1}{6},$$

$$(D_N)_{jj} = \frac{-x_j}{2\left(1-x_j^2\right)}, \quad j = 1, \ldots, N-1,$$

$$(D_N)_{ij} = \frac{c_i}{c_j}\frac{(-1)^{i+j}}{(x_i - x_j)}, \quad i \neq j, \quad i, j = 0, \ldots, N,$$

mit $c_0 = c_N = 2$ und $c_j = 1$, $j = 1, \ldots, N-1$.
Mit D_N und $F_N = (F(x_0), F(x_1), \ldots, F(x_N))^T \in \mathbb{R}^{N+1}$, dem Vektor mit den Funktionswerten der Stammfunktion an den Chebyshevpunkten x_j, $j = 0, \ldots, N$, gilt dann

$$f_N := D_N \cdot F_N, \tag{2.46}$$

wobei die Einträge von f_N die Näherung der Ableitung von F an den Stellen x_j sind. Mit Hilfe dieser Differentationsmatrix D_N kann das Integral über eine

Funktion $f(x)$, die an Chebyshevpunkten gegeben ist, berechnet werden.
Ist nun eine Funktion $f(x)$ auf dem Intervall $[-1;1]$ gegeben, von der eine Stammfunktion $F(x)$, die $F(-1) = 0$ erfüllt, gesucht ist, lässt sich diese Aufgabe als Anfangswertproblem

$$\begin{cases} f(x) = F'(x), & -1 < x \leq 1, \quad (2.47a) \\ F(-1) = 0, & \quad (2.47b) \end{cases}$$

betrachten. Analytisch ist die Lösung von (2.47) gegeben mit

$$F(x) = \int_{-1}^{x} f(\tau)\,\mathrm{d}\tau$$

und somit ergibt sich

$$\int_{-1}^{1} f(\tau)\,\mathrm{d}\tau = F(1). \tag{2.48}$$

Dieser Wert $F(1)$ soll numerisch berechnet werden.
Um die Anfangsbedingung (2.47b) sicherzustellen, streicht man in D_N die letzte Zeile und die letzte Spalte und erhält eine Matrix $\widetilde{D}_N \in \mathbb{R}^{N\times N}$. Ebenso streicht man in dem Vektor $f_N = (f(x_0), f(x_1), \ldots, f(x_N))^T \in \mathbb{R}^{N+1}$, der die gegebenen Funktionswerte an den Chebyshevpunkten enthält, den letzten Eintrag und man erhält $\widetilde{f}_N \in \mathbb{R}^N$. Der Vektor $\widetilde{F}_N \in \mathbb{R}^N$, der die gesuchten Werte für die angenäherte Stammfunktion an den Chebyshevpunkten $x_j = \cos\frac{\pi j}{N}$, $j = 0, \ldots, N-1$,[1] enthält, ist die Lösung des aus (2.46) resultierenden Gleichungssystems

$$\widetilde{f}_N = \widetilde{D}_N \cdot \widetilde{F}_N. \tag{2.49}$$

Benötigt man nur den Wert des bestimmten Integrals (2.48), also $F(x_0)$ mit $x_0 = 1$, der dem ersten Eintrag des Vektors $\widetilde{F}_N$ entspricht, muss (2.49) nicht vollständig gelöst werden, sondern es reicht, $\widetilde{f}_N$ mit $w_N^T := \widetilde{D}_N^{-1}(1,:)$, der ersten Zeile von $\widetilde{D}_N^{-1}$, zu multiplizieren. Die Einträge des Vektors w_N^T sind Quadraturgewichte, die für bestimmte Wahl von N denen von anderen Verfahren entsprechen (siehe dazu Bemerkung 2.18).
Eine Nachfrage bei Trefethen hat ergeben, dass für dieses Verfahren keine Fehlerabschätzungen bekannt sind. Jedoch wird in [Tr1, Chapter 12] darauf hingewiesen, dass Polynome vom Grad $N-1$ exakt integriert werden.

Bemerkung 2.18. Wie schon erwähnt, erhält man für bestimmte Wahl von N bekannte Quadraturverfahren oder es ergeben sich Ähnlichkeiten zu diesen.

[1] Da $x_N = -1$ ist, ist dieser Wert durch (2.47b) vorgegeben und muss nicht berechnet werden.

1. Für $N = 1$ erhält man die Rechteckregel $\int_{-1}^{1} f(x)\mathrm{d}x \approx 2f(1)$.

2. Für $N = 2$ ergibt sich die Mittelpunktsregel $\int_{-1}^{1} f(x)\mathrm{d}x \approx 2f(0)$.

3. Wählt man $N = 4$, erhält man ein Verfahren, das dieselbe Fehlerordnung wie das Simpsonverfahren hat. Die Gewichte an den Punkten $x_0 = 1$, $x_1 \approx 0.7071$, $x_2 = 0$ und $x_3 \approx -0.7071$ sind $w_4^T = \left[0; \frac{2}{3}; \frac{2}{3}; \frac{2}{3}\right]$, somit erhält man $\int_{-1}^{1} f(x)\mathrm{d}x \approx \frac{2}{3}\left(f(0.7071) + f(0) + f(-0.7071)\right)$.

Beispiel 2.19. Der Fehler des Quadraturverfahrens soll anhand der Funktion $f(x) = \frac{1}{x+4}$ mit $\int_{-1}^{1} f(x)\mathrm{d}x = \ln\frac{5}{3}$ aus Beispiel 2.17 illustriert werden.

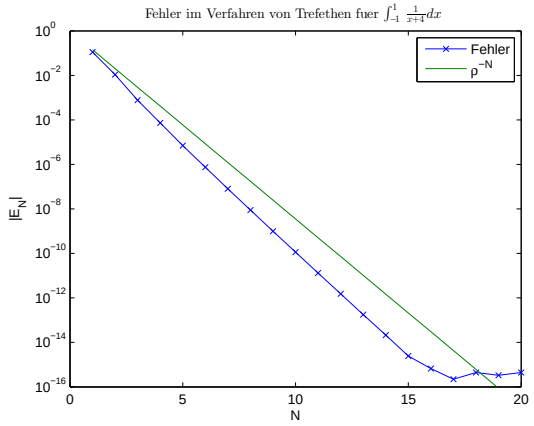

Abbildung 2.2: Beispiel zum Fehler im Verfahren aus [Tr1]

Neben dem Fehler (blau) ist in Abb. 2.2 auch $\rho^{-N} = 7^{-N}$ (grün) zum Vergleich abgebildet. Man erkennt, dass der Verfahrensfehler ebenso wie bei der Clenshaw-Curtis Quadratur exponentiell fällt. In [Tr1, Output 30a] ist Ähnliches für die Funktion $f(x) = \frac{1}{1+x^2}$ gezeigt, die innerhalb und auf der Ellipse mit $\rho < 1 + \sqrt{2}$ analytisch ist.

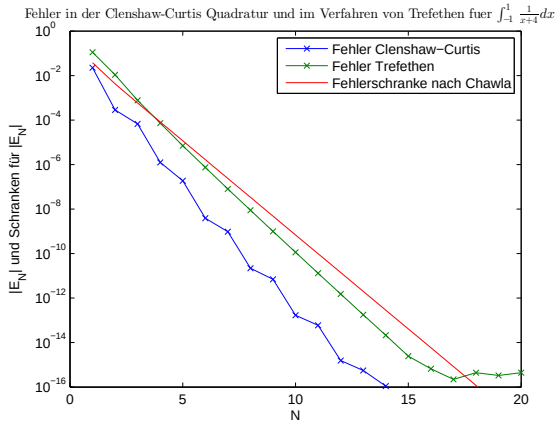

Abbildung 2.3: Fehler im Verfahren aus [Tr1] und Clenshaw-Curtis Quadratur

In Abb. 2.3 sind zum Vergleich der Fehler der Clenshaw-Curtis Quadratur (blau), der Fehler des Verfahrens von Trefethen (grün) und die Fehlerabschätzung (2.45) (rot) abgebildet. Für das Verfahren von Trefethen ist (2.34) für kleine Werte von N keine Abschätzung, ebenso ist der Fehler der Clenshaw-Curtis Quadratur kleiner.

Bemerkung 2.20. Das Verfahren lässt sich durch eine lineare Transformation wie (2.25) auch auf in einem Intervall $[a; b]$ integrierbare Funktionen anwenden.

2.4 Baryzentrische Interpolation

Die Baryzentrische Interpolation ist eine Erweiterung der Lagrange-Interpolation [BK, BT] bzw. der Hermite-Interpolation [SV]. Da die Stammfunktion mit der Clenshaw-Curtis Quadratur nur an den Stützstellen berechnet wird und keine Ableitungen an den Stützstellen verwendet werden, wird hier nur die baryzentrische Lagrange-Interpolation betrachtet.
Ausgehend von der Lagrange-Interpolationsformel

$$p_n(x) = \sum_{j=0}^{n} f(x_j) l_j(x) \tag{2.50}$$

mit den Lagrangepolynomen

$$l_j(x) = \prod_{\substack{k=0 \\ k \neq j}}^{n} \frac{x - x_k}{x_j - x_k} \tag{2.51}$$

nennt man

$$w_j := \frac{1}{\prod_{k \neq j}(x_j - x_k)},\ j = 0, \ldots, n \tag{2.52}$$

die baryzentrischen Gewichte und mit

$$l(x) := \prod_{k=0}^{n} (x - x_k)$$

erhält man die erste baryzentrische Form

$$p_n(x) = l(x) \sum_{j=0}^{n} \frac{w_j}{x - x_j} f(x_j). \tag{2.53}$$

Betrachtet man die Funktion $f(x) \equiv 1$, erhält man für die erste Form

$$1 = l(x) \sum_{j=0}^{n} \frac{w_j}{x - x_j}. \tag{2.54}$$

Dividiert man (2.53) durch (2.54), erhält man die Formel für das baryzentrische Interpolationspolynom

$$p_n(x) = \frac{\sum_{j=0}^{n} \frac{w_j}{x-x_j} f(x_j)}{\sum_{j=0}^{n} \frac{w_j}{x-x_j}}. \tag{2.55}$$

In [BK, BT] sind die Vorteile der baryzentrischen Interpolation gegenüber herkömmlicher Lagrange-Interpolation oder Newton-Interpolation mittels dividierter Differenzen angeführt. Die Auswertung von $p_n(x)$ hat einen Aufwand von $\mathcal{O}(n)$, sobald die Gewichte berechnet sind. Der Aufwand für die Berechnung der von $f(x_j)$ unabhängigen Gewichte ist $\mathcal{O}(n^2)$. Neue Datenpunkte $(x_{n+1}, f(x_{n+1}))$ können ebenfalls mit einem Aufwand von $\mathcal{O}(n)$ hinzugefügt werden.
Ein weiterer Vorteil ist, dass für die Interpolation an den Chebyshevpunkten (2.4), wie sie für die Clenshaw-Curtis Quadratur gewählt werden, die Gewichte durch

$$w_j = (-1)^j \, \delta_j, \;\; \delta_j = \begin{cases} \frac{1}{2}, & j = 0 \text{ oder } j = n, \\ 1, & \text{sonst}, \end{cases} \tag{2.56}$$

gegeben sind. Würde man ein allgemeines Intervall $[a; b]$ betrachten, müsste man die Gewichte (2.56) mit $2^n(b-a)^{-n}$ multiplizieren. Da in (2.55) die Gewichte in Zähler und Nenner auftreten, hat dieser Faktor jedoch keine Auswirkung.
Die baryzentrische Interpolation zeichnet sich dadurch aus, dass sie sehr stabil ist, auch wenn x nahe an einer Stützstelle x_j liegt. Falls $x = x_j$ ist und in (2.55) eine Division durch 0 auftritt, ersetzt man den unbestimmten Ausdruck durch den gegebenen Wert $f(x_j)$ und somit ist an diesen Stellen $p_n(x_j) = f(x_j)$.
Für allgemeine, mit (2.52) berechnete Gewichte wird in [Hi, Theorem 4.1] gezeigt, dass (2.55) stabil ist. Für die Gewichte (2.56) findet man in [Ma, Theorem 1] und [Ma, Theorem 2] schärfere Abschätzungen für den Fehler bei der Auswertung von (2.55) in Gleitkommaarithmetik.
Für glatte, auf $[-1; 1]$ definierte Funktionen, die innerhalb und auf der Ellipse $\mathcal{E}_\rho$ (vgl. (2.32)) analytisch sind, ist in [BT, (6.1)] eine Abschätzung des Interpolationsfehlers mit

$$\max_{x \in [-1;1]} |f(x) - p_n(x)| \leq C \rho^{-n} \tag{2.57}$$

mit Konstante $C > 0$ und Summe der Halbachsen $\rho > 1$ angegeben.

Kapitel 3

Berücksichtigung des Quadraturfehlers im WKB-Verfahren

In der Fehlerabschätzung aus Satz 1.1 für das Anfangswertproblem (1.17),

$$\begin{cases} Z'(x) = \varepsilon N^{\varepsilon}(x) Z(x), \quad 0 < x < 1, \\ Z(0) = Z_I, \end{cases}$$

wird der Quadraturfehler, der bei der Berechnung der Phase entsteht, nicht berücksichtigt. Im Folgenden sollen nun die Abschätzung (1.30),

$$\|Z(x_t) - Z_t\| \leq C\varepsilon^2 \min(\varepsilon, h), \ 1 \leq t \leq T,$$

des Verfahrens 1. Ordnung und die Abschätzung (1.32),

$$\|Z(x_t) - Z_t\| \leq C\varepsilon^3 h^2, \ 1 \leq t \leq T,$$

des Verfahrens 2. Ordnung um den Quadraturfehler, der bei der Berechnung von (1.2),

$$\phi(x) = \int_0^x \left(\sqrt{a(\tau)} - \varepsilon^2 \beta(\tau) \right) \mathrm{d}\tau,$$

entsteht, erweitert werden.
Dazu betrachtet man die Verfahren 1. Ordnung aus (1.20),

$$Z_{t+1} = \left(I + A_t^1 \right) Z_t, \ t = 1, \ldots, T-1,$$

und 2. Ordnung aus (1.26),

$$Z_{t+1} = \left(I + A_{mod,t}^1 + A_t^2 \right) Z_t, \ t = 1, \ldots, T-1,$$

mit Matrizen $\widetilde{A}_t^1$, $\widetilde{A}_{mod,t}^1$ und $\widetilde{A}_t^2$ anstelle der Matrizen A_t^1, in (1.21) definiert, $A_{mod,t}^1$, in (1.27) definiert, und A_t^2, in (1.28) definiert. In diesen Matrizen wird die exakte Phase $\phi(x)$ durch eine numerische Näherung $\widetilde{\phi}(x)$ ersetzt.
Abhängig von der Art der Berechnung von $\widetilde{\phi}\,(x)$ entsteht ein Fehler

$$E := \max_{1\leq t\leq T} \left|\phi\,(x_t) - \widetilde{\phi}\,(x_t)\right| \tag{3.1}$$

an den Gitterpunkten x_t, auf denen die Lösung von (1.17) berechnet werden soll.

3.1 Möglichkeiten zur Phasenberechnung

3.1.1 Mit baryzentrischer Interpolation

Hier wird, im Gegensatz zu gängigen numerischen Quadraturverfahren wie man sie z. B. in [HH, Kapitel 7] oder [Pl, Kapitel 6] findet, kein bestimmtes Integral der Form $\int_a^b f(x)\mathrm{d}x$ berechnet, sondern man nähert mit (2.27) die gesuchte Stammfunktion. Dazu wird im ersten Schritt der Integrand mit (2.20) in eine Chebyshevreihe entwickelt und die Koeffizienten der Reihe zur Näherung der Stammfunktion an den Chebyshevpunkten verwendet.
Im zweiten Schritt werden mit baryzentrischer Interpolation die gesuchten Stammfunktionswerte an den Gitterpunkten ausgehenden von den genäherten Werten an den Chebyshevpunkten berechnet.
Da (1.17) auf dem Intervall $[0;1]$ gelöst wird, die Chebyshevpolynome jedoch auf dem Intervall $[-1;1]$ definiert sind, führt eine lineare Transformation wie (2.25) darauf, dass eine Stammfunktion auf einem beliebigen Intervall $[a;b]$ mit (2.31),

$$\int_a^x f(\tau)\mathrm{d}\tau \approx F_N(x) = \frac{b-a}{2}\left(\frac{b_0}{2} + b_1T_1(u) + \ldots + b_NT_N(u)\right),$$

berechnet werden kann, wobei $u \in [-1;1]$ durch (2.25) bestimmt ist.
Die genäherte Stammfunktion $F_N(x)$ wird mit (2.31) an den Chebyshevpunkten $\widetilde{x}_j$, $0 \leq j \leq N$, berechnet und muss nicht mit der exakten Stammfunktion an den Chebyshevpunkten übereinstimmen. Für den Fehler, der durch die Näherung der Stammfunktion mit (2.31) entsteht, sind trotz intensiver Literaturrecherche, keine Abschätzungen gefunden worden.
Im nächsten Schritt wird $F_N(x)$ mit der baryzentrischen Interpolationsformel (2.55),

$$F_N(x_t) = \frac{\sum_{j=0}^N \frac{w_j}{x_t-\widetilde{x}_j}F_N(\widetilde{x}_j)}{\sum_{j=0}^N \frac{w_j}{x_t-\widetilde{x}_j}},$$

ausgehend von den approximierten Funktionswerten an den Chebyshevpunkten $\widetilde{x}_j$, $0 \leq j \leq N$, an den Gitterpunkten x_t, $1 \leq t \leq T$, berechnet.

Ist die Funktion $G(x)$, die interpoliert werden soll, bekannt und ist $G_N(x)$ die mit (2.55) baryzentrisch interpolierte Näherung von $G(x)$, ist in Abschnitt 2.4 mit (2.57) eine Abschätzung des Interpolationsfehlers

$$\max_{x\in[-1;1]} |G(x) - G_N(x)| \leq C\rho^{-N}$$

angegeben, wobei $\rho > 1$ die Summe der Halbachsen der Ellipse $\mathcal{E}_\rho$ (vgl. (2.32)) ist, innerhalb und auf der die Funktion $G(x)$ noch analytisch ist. Durch (2.25) kann die Abschätzung (2.57) auf ein Intervall $[a; b]$ transformiert werden,

$$\max_{x\in[a;b]} |G(x) - G_N(x)| \leq C\rho^{-N}. \tag{3.2}$$

Die Ellipse $\mathcal{E}_\rho$ ist dann jene, innerhalb und auf der die Transformation von $G(x)$ in das Intervall $[-1; 1]$ noch analytisch ist.
Wird $G(x)$ an Gitterpunkten x_t, $1 \leq t \leq T$, baryzentrisch interpoliert, wird (3.2) zu

$$\max_{1\leq t\leq T} |G(x_t) - G_N(x_t)| \leq C\rho^{-N}.$$

Somit sind die Abschätzungen (2.57) und (3.2) unabhängig von der Wahl der Gitterpunkte und daher auch von der Schrittweite $h = x_{t+1} - x_t$.

Da für die Konvergenz des approximierten Integranden $f_N(x)$ gegen $f(x)$ für die Koeffizienten seiner Entwicklung $\lim_{k\to\infty} a_k = 0$ gelten muss, reicht es in der Praxis nur so weit zu entwickeln, bis seine Koeffizienten vernachlässigbar klein im Vergleich zur Maschinengenauigkeit sind. Die Koeffizienten b_k der approximierten Stammfunktion sind nach (2.28) gegeben mit $b_k = \frac{a_{k-1}-a_{k+1}}{2k}$ und bilden somit ebenfalls eine Nullfolge.

Beispiel 3.1. Gesucht ist eine genäherte Stammfunktion von $f(x) = \frac{-3}{(2x+1)^3}$ im Intervall $[0; 1]$, die $F(0) = 0$ erfüllen soll. Die Stammfunktion von $f(x)$, die diese Forderung erfüllt, ist $F(x) = \frac{3}{4(2x+1)^2} - \frac{3}{4}$.
Entwickelt man $f(x)$ mit (2.20) und (2.25) in eine Chebyshevreihe bis zum Grad $N = 34$, ist $a_{34} = 2.7756 \cdot 10^{-17}$, also für eine Implementierung in IEEE double precision, wie sie in MATLAB verwendet wird, vernachlässigbar klein.

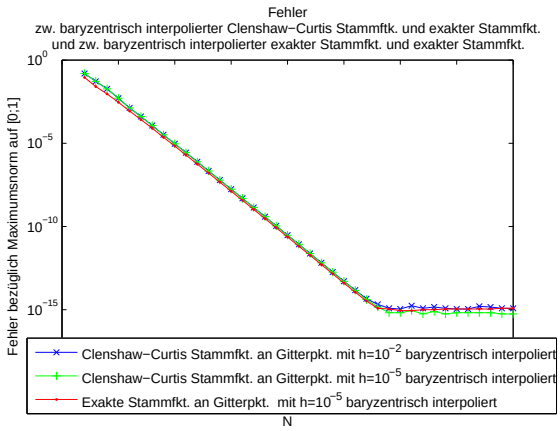

Abbildung 3.1: Fehler einer baryzentrisch interpolierten Stammfunktion in Abhängigkeit vom Grad N der Chebyshe-ventwicklung

In Abb. 3.1 sind mehrere Fehlerkurven zum Vergleich abgebildet. Es ist $\max_{1\leq t\leq T}|F(x_t) - F_N(x_t)|$ auf einem groben Gitter, $h = 10^{-2}$ bzw. $T = 101$ (blau), und einem feinen Gitter, $h = 10^{-5}$ bzw. $T = 100001$ (grün), in Abhängigkeit von N gezeigt. $F_N(x_t)$ bezeichnet dabei die mit (2.31) genäherte und anschließend mit (2.55) berechnet Stammfunktion und $F(x_t)$ die exakte Stammfunktion an den Gitterpunkten. Die rote Kurve ist (3.2) mit der, an den Gitterpunkten baryzentrisch interpolierten, exakten Stammfunktion und der exakten Stammfunktion an den Gitterpunkten mit $h = 10^{-5}$. Die blaue und die grüne Fehlerkurve verlaufen annähernd deckungsgleich und die Rote parallel dazu. Man erkennt, dass der Fehler, wenn die exakte Stammfunktion baryzentrisch interpoliert wird, wie in (3.2) ersichtlich, exponentiell abnimmt.

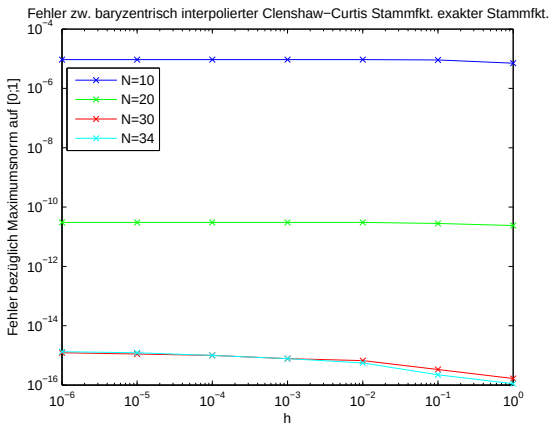

Abbildung 3.2: Fehler einer baryzentrisch interpolierten Stammfunktion in Abhängigkeit von h

Entscheidend ist, dass sich der Fehler bei der Auswertung der genäherten Stammfunktion $F_N(x)$ wie (3.2) verhält und somit diese Abschätzung für den Quadraturfehler verwendet werden kann. Weiters erkennt man,dass der Fehler unabhängig von h ist.
In Abb. 3.2 ist der Fehler $\max_{1\leq t\leq T}|F(x_t) - F_N(x_t)|$ für verschiedene Werte von N (10 (blau), 20 (grün), 30 (rot) und 34 (hellblau)) in Abhängigkeit von h bzw. T dargestellt und man erkennt, dass (3.2) von der Schrittweite h bzw. der Anzahl der Gitterpunkte T unabhängig ist.

Beispiel 3.1 hat gezeigt, dass der Fehler (3.1), der bei dieser Art der Phasenberechnung entsteht, unabhängig von der Schrittweite $h = x_{t+1} - x_t$ ist, die bei der Lösung von (1.17) gewählt wird. Weiters ergibt sich aus diesem Beispiel, dass (3.1) durch (3.2) nach oben mit $C\rho^{-N}$ beschränkt ist und somit nur vom Grad N der Entwicklung des Integranden in Chebyshevpolynomen abhängt. Deshalb kann (3.1) beliebig klein werden, wenn N nur groß genug gewählt wird.

Bemerkung 3.2. Anstelle von (2.31) kann mit der Lösung des linearen Gleichungssystems (2.49) aus Abschnitt 2.3,

$$\widetilde{f}_N = \widetilde{D}_N \cdot \widetilde{F}_N,$$

eine approximierte Stammfunktion an Chebyshevpunkten berechnet werden. Diese Stammfunktion kann ebenfalls baryzentrisch interpoliert werden, um die gesuchten Funktionswerte an den Gitterpunkten x_t, $1 \leq t \leq T$, zu bestimmen. Da jedoch bei der Berechnung der Stammfunktion $\widetilde{F}_N$ nur die Funktionswerte an N Punkten anstelle von $N+1$ Punkten wie, bei der Clenshaw-Curtis Quadratur, verwendet werden, ist die Berechnung der Stammfunktion über die Clenshaw-Curtis Quadratur für kleine Werte von N besser. Entwickelt man aber den Integranden soweit, bis die Koeffizienten a_k vernachlässigbar klein sind, besteht praktisch kein Unterschied mehr zwischen den beiden Verfahren.

Beispiel 3.3 (Fortsetzung von Beispiel 3.1)**.**

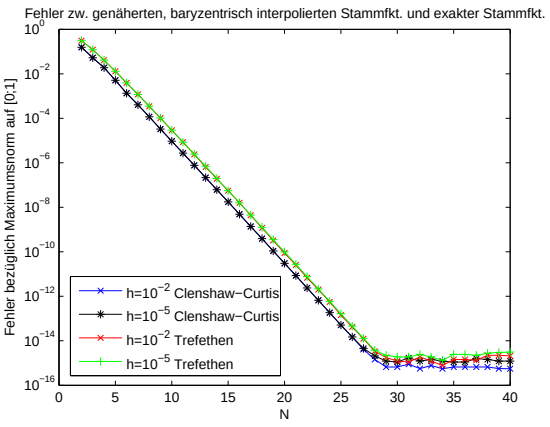

Abbildung 3.3: Fehler einer baryzentrisch interpolierten Stammfunktion mit dem Verfahren aus Abschnitt 2.3

Die Stammfunktion von $f(x) = \frac{-3}{(2x+1)^3}$ wird jetzt mit (2.49) berechnet und anschließend auf den Gitterpunkten baryzentrisch interpoliert.
In Abb. 3.3 ist der Fehler zwischen der mit (2.49) berechneten und anschließend baryzentrisch interpolierten Stammfunktion und der exakten Lösung abgebildet. Die Interpolation erfolgte auf einem groben Gitter, $h = 10^{-2}$ (rot), und einem feinen Gitter, $h = 10^{-5}$ (grün). Man erkennt, dass auch hier der Fehler unabhängig von der Schrittweite h ist. Zum Vergleich dazu sind die

beiden Fehlerkurven aus Abb. 3.1 ebenfalls eingezeichnet (blau und schwarz). Man sieht deutlich die Verschiebung der Fehlerkurven um eins nach links, da bei der Berechnung mit (2.31) ein Punkt mehr verwendet wird als bei (2.49). Ebenso ist ersichtlich, dass, wenn N groß genug gewählt wird, praktisch kein Unterschied mehr zwischen den beiden Verfahren besteht.

Da, wie aus Bemerkung 3.2 und Beispiel 3.3 hervorgeht, für großes N in der numerischen Implementierung kein Unterschied zwischen der baryzentrischen Interpolation von (2.31) und der interpolierten Lösung von (2.49) besteht, wird in den Kapiteln 4 und 5 nur die baryzentrische Interpolation von (2.31) verwendet, um das Phasenintegral (1.2),

$$\phi(x) = \int_0^x \left(\sqrt{a(\tau)} - \varepsilon^2 \beta(\tau) \right) \mathrm{d}\tau,$$

an den Gitterpunkten x_t, $1 \leq t \leq T$, zu berechnen.

3.1.2 Mit aufsummierter Clenshaw-Curtis Quadratur

Eine weitere Möglichkeit, die Phase an den Gitterpunkten x_t, $1 \leq t \leq T$, zu bestimmen, ist (1.2) zwischen zwei Gitterpunkten x_t und x_{t+1} zu berechnen und anschließend diese Werte zu summieren. In [Pl, Kapitel 6.5] finden man z. B. die aufsummierte Trapezregel oder das aufsummierte Simpsonverfahren und die globalen Fehlerordnungen h^γ dieser Verfahren.
Ebenso können das Quadraturverfahren aus Abschnitt (2.3) oder die Clenshaw-Curtis Quadratur (2.30) zur Berechnung eines bestimmten Integrals aufsummiert werden.
Dazu wird der Integrand

$$\phi'(x) = \sqrt{a\,(x)} - \varepsilon^2 \beta\,(x)$$

vom Intervall $[x_t; x_{t+1}]$ mit (2.25) linear in das Intervall $[-1; 1]$ transformiert, dort in eine Chebyshevreihe entwickelt und integriert und der Wert des Integrals wird wieder zurücktransformiert. Bezeichne $\psi_t\,(u)\,,\ u \in [-1; 1]\,,$ die durch (2.25) transformierte Funktion $\phi'\,(x)\,,\ x \in [x_t; x_{t+1}]\,,$ und I_t das durch die Clenshaw-Curtis Quadratur genäherte Integral, dann folgt mit (2.26) bzw. (2.31),

$$I_t \approx \int_{-1}^{1} \psi_t\,(u)\,\mathrm{d}u = \frac{2}{h} \int_{x_t}^{x_{t+1}} \phi'\,(\tau)\,\mathrm{d}\tau. \tag{3.3}$$

Der lokale Fehler E_t für das Intervall $[x_t; x_{t+1}]$ ist gegeben mit

$$\begin{aligned}\left|\int_{x_t}^{x_{t+1}} \phi'(\tau)\,\mathrm{d}\tau - \frac{h}{2}I_t\right| &= \left|\frac{h}{2}\int_{-1}^{1}\psi_t(u)\,\mathrm{d}u - \frac{h}{2}I_t\right| \\ &= \frac{h}{2}\left|\int_{-1}^{1}\psi_t(u)\,\mathrm{d}u - I_t\right| \\ &= \frac{h}{2}E_t.\end{aligned}$$

Nach Satz 2.12 ist

$$E_t \leq \frac{16N^2}{4N^2-1}\frac{M(\rho_t)}{(\rho_t^2-1)\left(\rho_t^N-\rho_t^{-N}\right)},$$

wobei $\rho_t > 1$ die Summe der Halbachsen der Ellipse $\mathcal{E}_{\rho_t}$ mit Brennpunkten in ± 1 ist, auf der die Funktion $\psi_t(u)$ analytisch ist, $M(\rho_t)$ das Maximum von $\psi_t(u)$ auf $\mathcal{E}_{\rho_t}$ ist und N der Grad des Chebyshevpolynoms ist, in das $\psi_t(u)$ entwickelt wird.
Auf Grund der Linearität des Integrals gilt mit (3.3)

$$\phi(x_t) = \int_0^{x_t}\phi'(\tau)\,\mathrm{d}\tau \approx \widetilde{\phi}(x_t) = \frac{h}{2}\sum_{i=1}^{t-1} I_i \text{ mit } \widetilde{\phi}(0) = 0,$$

und somit für den globalen Fehler, der bei der Berechnung von $\phi(1)$ entsteht,

$$\begin{aligned}\left|\int_0^1 \phi'(\tau)\,\mathrm{d}\tau - \frac{h}{2}\sum_{t=1}^{T-1} I_t\right| &= \left|\sum_{t=1}^{T-1}\int_{x_t}^{x_{t+1}}\phi'(\tau)\,\mathrm{d}\tau - \frac{h}{2}\sum_{t=1}^{T-1} I_t\right| \\ &= \left|\sum_{t=1}^{T-1}\frac{h}{2}\int_{-1}^{1}\psi_t(u)\,\mathrm{d}u - \frac{h}{2}\sum_{t=1}^{T-1} I_t\right| \\ &= \left|\frac{h}{2}\sum_{t=1}^{T-1}\left(\int_{-1}^{1}\psi_t(u)\,\mathrm{d}u - I_t\right)\right| \\ &\leq \frac{h}{2}\sum_{t=1}^{T-1}\left|\int_{-1}^{1}\psi_t(u)\,\mathrm{d}u - I_t\right| = \frac{h}{2}\sum_{t=1}^{T-1}E_t \\ &\leq \frac{h}{2}\sum_{t=1}^{T-1}E = \frac{h(T-1)}{2}E = \frac{E}{2},\end{aligned}$$

mit $E := \max_{1\leq t\leq T-1} E_t = \max_{2\leq t\leq T}\left|\phi(x_t) - \widetilde{\phi}(x_t)\right|$.
Im Gegensatz zur Berechnung mit baryzentrischer Interpolation ist hier der

Fehler E sowohl von N als auch der Schrittweite h abhängig, da hier die Intervalle $[x_t; x_{t+1}]$ mit (2.25) linear in das Intervall $[-1; 1]$ transformiert werden und somit ρ_t von h abhängt. Auf die genaue $h-$Abhängigkeit von ρ_t hat die Form der Funktion $f(x)$ Einfluss. Es gilt jedoch bei dieser Art der Phasenberechnung nicht nur bei festem h $\lim_{N\to\infty} E = 0$, sondern auch bei festem N $\lim_{h\to 0} E = 0$. Das nächste Beispiel soll dies veranschaulichen.

Beispiel 3.4. Wie in den Beispielen 3.1 und 3.3 wird auch hier die Stammfunktion von $f(x) = \frac{-3}{(2x+1)^3}$ mit $F(0) = 0$ an den Gitterpunkten x_t,
$1 \le t \le T$, berechnet. Dabei wird mit der Clenshaw-Curtis Quadratur das bestimmte Integral im Intervall $[x_t; x_{t+1}]$ berechnet.
Die Transformation (2.25) ist hier

$$x = \frac{1}{2}\left(x_{t+1} + x_t + u\left(x_{t+1} - x_t\right)\right) \text{ mit } u \in [-1; 1]. \tag{3.4}$$

Mit (3.4) wird die im Intervall $[x_t; x_{t+1}]$ betrachtete Funktion $f(x)$ zu

$$\psi_t(u) = \frac{-3}{(x_{t+1} + x_t + u(x_{t+1} - x_t) + 1)^3}$$

transformiert, die sich mit $x_{t+1} = th$ und $x_t = (t-1)h$, zu

$$\begin{aligned} \psi_t(u) &= \frac{-3}{((2t-1)h + uh + 1)^3} \\ &= \frac{-3}{h^3(2t - 1 + u + \frac{1}{h})^3} \end{aligned}$$

umschreiben lässt. Die Funktion $\psi_t(u)$ ist innerhalb und auf der Ellipse $\mathcal{E}_{\rho_t}$ mit $\rho_t < 2t - 1 + \frac{1}{h} + \sqrt{(2t - 1 + \frac{1}{h})^2 - 1}$ analytisch. Da für $a > 1$ die Beziehung $\sqrt{a} - 1 < \sqrt{a-1}$ gilt, folgt $\sqrt{(2t - 1 + \frac{1}{h})^2} - 1 < \sqrt{(2t - 1 + \frac{1}{h})^2 - 1}$.
Somit kann $\rho_t = 4t - 3 + \frac{2}{h}$ gewählt werden.
Da nach Satz 2.12 der Fehler der Quadratur von der Größenordnung $\mathcal{O}(\rho^{-N})$ ist, verwendet man für eine globale Abschätzung

$$\rho = \min_{1\le t\le T-1} \rho_t = 1 + \frac{2}{h}. \tag{3.5}$$

Somit erhält man bei festem h einen Fehler der Ordnung $\mathcal{O}\left((1 + \frac{2}{h})^{-N}\right)$, der wenn h klein genug ist, zu $\mathcal{O}\left(h^N\right)$ wird. Wählt man N fest, erhält man aus der Fehlerabschätzung (2.34),

$$E \le \left(\frac{16N^2}{4N^2 - 1}\right) \frac{M(\rho)}{(\rho^2 - 1)(\rho^N - \rho^{-N})},$$

eine Fehlerordnung $\mathcal{O}(\rho^{-N-2})$, die mit (3.5) einer Ordnung in h von $\mathcal{O}(h^{N+2})$ entspricht.

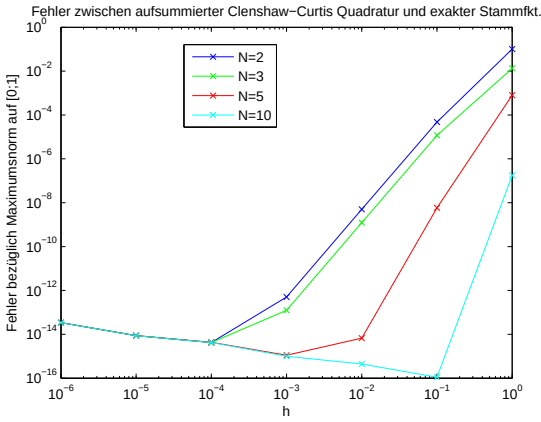

Abbildung 3.4: Fehler einer aufsummierten Clenshaw-Curtis Quadratur

In Abb. 3.4 sind die Fehlerkurven in Abhängigkeit der Schrittweite h für verschiedene, steigende Werte von N (2 (blau), 3 (grün), 5 (rot) und 10 (hellblau)) gezeigt. Man erkennt, dass der Fehler bei festem N mit h gegen Null geht. Ebenso geht der Fehler gegen Null, wenn bei festem h N gegen Unendlich geht. Bei sehr kleinem h steigt der Fehler auf Grund von numerischen Effekten wie Rundungs- und Auslöschungsfehlern wieder leicht an.

3.2 Erweiterung der Fehlerabschätzung um den Quadraturfehler bei der Phasenberechnung

Mit einer genäherten Phase $\widetilde{\phi}(x)$, die (3.1) erfüllt, betrachtet man den Fehler nach t Schritten zwischen Z_{t+1} und $\widetilde{Z}_{t+1}$,

$$\left\| Z_{t+1} - \widetilde{Z}_{t+1} \right\|, \tag{3.6}$$

der von (3.1) abhängt. Ob die Schrittweite h Einfluss auf (3.1) und somit auf (3.6) hat, hängt von der Methode der Phasenberechnung ab und wird in den Abschätzungen vorläufig nicht berücksichtigt.
Für den zusätzlichen Fehler, der in einem Schritt gemacht wird, werden auf Grund der Struktur der Matrizen A_t^1, $A_{mod,t}^1$ und A_t^2 die Elemente $\left(A_t^1 - \widetilde{A}_t^1\right)_{2,1}$, $\left(A_{mod,t}^1 - \widetilde{A}_{mod,t}^1\right)_{2,1}$ bzw. $\left(A_t^2 - \widetilde{A}_t^2\right)_{2,2}$ betrachtet und abgeschätzt.
In den Matrizen treten die Funktionen $H_1\left(\frac{2}{\varepsilon}S_t\right)$ und $H_2\left(\frac{2}{\varepsilon}S_t\right)$ auf, die in (1.23),

$$H_1(x) = e^{\imath x} - 1,$$

und (1.24),

$$H_2(x) = e^{\imath x} - 1 - \imath x,$$

definiert worden sind. Mit $S_t = \phi(x_{t+1}) - \phi(x_t)$ folgt

$$H_1\left(\frac{2}{\varepsilon}S_t\right) = e^{\frac{2\imath}{\varepsilon}(\phi(x_{t+1})-\phi(x_t))} - 1$$

und

$$H_2\left(\frac{2}{\varepsilon}S_t\right) = e^{\frac{2\imath}{\varepsilon}(\phi(x_{t+1})-\phi(x_t))} - 1 - \frac{2\imath}{\varepsilon}\left(\phi(x_{t+1}) - \phi(x_t)\right).$$

Diese Ausdrücke werden im Folgenden für $H_1\left(\frac{2}{\varepsilon}S_t\right)$ und $H_2\left(\frac{2}{\varepsilon}S_t\right)$ verwendet. Die weiteren auftretenden Funktionen β_k, die in (1.25) definiert sind, sind auch für die Matrizen $\widetilde{A}_t^1$, $\widetilde{A}_{mod,t}^1$ und $\widetilde{A}_t^2$ mit denen $\widetilde{Z}_t$ berechnet wird exakt, da darin nur

$$\phi'(x) = \sqrt{a(x)} - \varepsilon^2\beta(x)$$

und höhere Ableitungen davon auftreten. Diese lassen sich bei gegebener Funktion $a(x)$ exakt berechnen.
Für die folgenden Fehlerabschätzungen sei weiters vorausgesetzt, dass die Ableitungen von $\widetilde{\phi}$ folgende Bedingungen für alle x aus dem Intervall $[0;1]$ und alle $0 < \varepsilon \leq \varepsilon_0$ erfüllen,

$$\begin{cases} \left\|\phi' - \widetilde{\phi}'\right\|_{L^\infty(0;1)} \leq E', & \text{(3.7a)} \\ 0 < \left|\widetilde{\phi}'(x)\right| \leq C_1, & \text{(3.7b)} \\ 0 < \left|\widetilde{\phi}''(x)\right| \leq C_2. & \text{(3.7c)} \end{cases}$$

Da $\phi'(x)$ exakt gegeben ist und $a(x) \geq a_0 > 0$ vorausgesetzt wird, sind (3.7b) für $\phi'(x)$ und (3.7c) für $\phi''(x)$ erfüllt.

3.2.1 Quadraturfehler im Verfahren 1. Ordnung

Satz 3.5. *Es sei $a(x) \in C^\infty[0;1]$ eine beliebige, glatte und reellwertige Funktion, die $a(x) \geq a_0 > 0$ im Intervall $[0;1]$ erfülle, und $0 < \varepsilon < 1$ eine beliebige, aber feste reelle Zahl und sei weiters (3.7) erfüllt. Dann gilt für den Fehler im Verfahren 1. Ordnung (1.20) unter der Berücksichtigung des Fehlers (3.1) bei der Phasenberechnung*

$$\left\|Z(x_t) - \widetilde{Z}_t\right\| \leq C_1\varepsilon^2 \min(\varepsilon, h) + C_2\varepsilon \min\left(\frac{\varepsilon}{h}, \frac{E}{h}, (E+E'), 1\right), \ 1 \leq t \leq T. \tag{3.8}$$

Dabei bezeichne $Z(x_t)$ die exakte Lösung, $\widetilde{Z}_t$ die numerische Lösung mit fehlerbehafteter Phase $\widetilde{\phi}(x_t)$, E den Fehler in der Phasenberechnung (3.1) und $\|\cdot\|$ eine beliebige Vektornorm in $\mathbb{C}^2$.

Satz 3.6. *Es sei (3.7) erfüllt. Dann gilt für den Fehler (3.6) im Verfahren 1. Ordnung*

$$\left\| Z_t - \widetilde{Z}_t \right\| \leq \varepsilon C \min\left(\frac{\varepsilon}{h}, \frac{E}{h}, (E+E'), 1 \right), \ 1 \leq t \leq T. \tag{3.9}$$

Beweis. Für den zusätzlichen Fehler in einem Schritt betrachtet man

$$\left(A_t^1 - \widetilde{A}_t^1 \right)_{2,1} =$$

$$\begin{aligned} = \ & \varepsilon^3 \beta_1(x_{t+1}) e^{\frac{2\imath}{\varepsilon}\phi(x_t)} \left(e^{\frac{2\imath}{\varepsilon}(\phi(x_{t+1})-\phi(x_t))} - 1 \right) - \imath\varepsilon^2 \beta_0(x_{t+1}) e^{\frac{2\imath}{\varepsilon}\phi(x_{t+1})} + \imath\varepsilon^2 \beta_0(x_t) e^{\frac{2\imath}{\varepsilon}\phi(x_t)} \\ & -\varepsilon^3 \beta_1(x_{t+1}) e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_t)} \left(e^{\frac{2\imath}{\varepsilon}\left(\widetilde{\phi}(x_{t+1})-\widetilde{\phi}(x_t)\right)} - 1 \right) + \imath\varepsilon^2 \beta_0(x_{t+1}) e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_{t+1})} - \imath\varepsilon^2 \beta_0(x_t) e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_t)} \end{aligned}$$

$$\begin{aligned} = \ & \varepsilon^3 \beta_1(x_{t+1}) \left(e^{\frac{2\imath}{\varepsilon}\phi(x_{t+1})} - e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_{t+1})} \right) + \varepsilon^3 \beta_1(x_{t+1}) \left(e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_t)} - e^{\frac{2\imath}{\varepsilon}\phi(x_t)} \right) \\ & +\imath\varepsilon^2 \beta_0(x_{t+1}) \left(e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_{t+1})} - e^{\frac{2\imath}{\varepsilon}\phi(x_{t+1})} \right) + \imath\varepsilon^2 \beta_0(x_t) \left(e^{\frac{2\imath}{\varepsilon}\phi(x_t)} - e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_t)} \right). \end{aligned}$$

Bildet man den Betrag, erhält man mit der Dreiecksungleichung

$$\left| \left(A_t^1 - \widetilde{A}_t^1 \right) \right| \leq$$

$$\leq \varepsilon^3 \left| \beta_1(x_{t+1}) \right| \left| e^{\frac{2\imath}{\varepsilon}\phi(x_{t+1})} - e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_{t+1})} + e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_t)} - e^{\frac{2\imath}{\varepsilon}\phi(x_t)} \right| \tag{3.10}$$

$$+\varepsilon^2 \left| \beta_0(x_{t+1}) \left(e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_{t+1})} - e^{\frac{2\imath}{\varepsilon}\phi(x_{t+1})} \right) - \beta_0(x_t) \left(e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_t)} - e^{\frac{2\imath}{\varepsilon}\phi(x_t)} \right) \right| \tag{3.11}$$

Zuerst wird (3.11) weiter abgeschätzt.
Da $\left| e^{\frac{2\imath}{\varepsilon}\eta} \right| = 1$ gilt und β_0 nach [ABN, (2.13)] im Intervall $[0;1]$ beschränkt ist, folgt mit der Dreiecksungleichung

$$(3.11) \leq 4\varepsilon^2 \left\| \beta_0 \right\|_{L^\infty(0;1)}. \tag{3.12}$$

Der Mittelwertsatz für vektorwertige Funktionen, [Heu2, Satz 167.4]:

> Die Funktion $\vec{f} : G \subset \mathbb{R}^p \to \mathbb{R}^q$ (G offen) sei stetig differenzierbar, und $\vec{x}_0$, $\vec{x}_0 + \vec{h}$ seien Punkte, die mitsamt ihrer Verbindungsstrecke S in G liegen. Dann gilt
>
> $$\vec{f}(\vec{x}_0 + \vec{h}) - \vec{f}(\vec{x}_0) = \left(\int_0^1 \vec{f}'(\vec{x}_0 + t\vec{h}) \mathrm{d}t \right) \vec{h}.$$

Ist $|||\cdot|||$ irgendeine mit den Normen von $\mathbb{R}^p$ und $\mathbb{R}^q$ verträgliche Matrixnorm auf $\mathfrak{M}(q,p)$ [die Menge aller (q,p)-Matrizen, Anm.] so haben wir die Abschätzung

$$\left\|\vec{f}(\vec{x}_0+\vec{h})-\vec{f}(\vec{x}_0)\right\| \leq M\left\|\vec{h}\right\| \text{ mit } M := \max_{\vec{x}\in S}|||\vec{f}'(\vec{x})|||\,.$$

angewandt auf die Funktion

$$\vec{f}(\phi) = e^{\frac{2\imath}{\varepsilon}\phi} = \begin{pmatrix} \cos\frac{2\phi}{\varepsilon} \\ \sin\frac{2\phi}{\varepsilon} \end{pmatrix} \tag{3.13}$$

mit

$$\vec{f}'(\phi) = \frac{2\imath}{\varepsilon}e^{\frac{2\imath}{\varepsilon}\phi} = \begin{pmatrix} -\frac{2}{\varepsilon}\sin\frac{2\phi}{\varepsilon} \\ \frac{2}{\varepsilon}\cos\frac{2\phi}{\varepsilon} \end{pmatrix}$$

ergibt mit $x_0=\phi$ und $h=\widetilde{\phi}-\phi$

$$e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}} - e^{\frac{2\imath}{\varepsilon}\phi} = \underbrace{\int_0^1 \imath e^{\frac{2\imath}{\varepsilon}\left(\phi+s\left(\widetilde{\phi}-\phi\right)\right)}\mathrm{d}s}_{=:A}\frac{2}{\varepsilon}\left(\widetilde{\phi}-\phi\right) \tag{3.14}$$

und

$$\left|e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}} - e^{\frac{2\imath}{\varepsilon}\phi}\right| \leq \frac{2}{\varepsilon}\left|\widetilde{\phi}-\phi\right|$$

da $|A|\leq 1$.
Damit erhält man

$$(3.11) = \varepsilon^2\left|\beta_0\left(x_{t+1}\right)A_1\frac{2}{\varepsilon}\left(\widetilde{\phi}\left(x_{t+1}\right)-\phi\left(x_{t+1}\right)\right) - \beta_0\left(x_t\right)A_2\frac{2}{\varepsilon}\left(\widetilde{\phi}\left(x_t\right)-\phi\left(x_t\right)\right)\right|$$

und daraus mit $\left|\widetilde{\phi}\left(x_t\right)-\phi\left(x_t\right)\right| \leq E$ und der Dreiecksungleichung eine weitere Abschätzung,

$$(3.11) \leq 4\varepsilon\left\|\beta_0\right\|_{L^\infty(0;1)}E. \tag{3.15}$$

Formt man nach der Anwendung des Mittelwertsatzes noch weiter um, erhält man

$$\begin{aligned}
(3.11) \;=\; & \varepsilon^2\Big|\left(\beta_0\left(x_{t+1}\right)-\beta_0\left(x_t\right)\right)A_1\frac{2}{\varepsilon}\left(\widetilde{\phi}\left(x_{t+1}\right)-\phi\left(x_{t+1}\right)\right) && (3.16)\\
& +\beta_0\left(x_t\right)A_1\frac{2}{\varepsilon}\left(\widetilde{\phi}\left(x_{t+1}\right)-\widetilde{\phi}\left(x_t\right)-\phi\left(x_{t+1}\right)+\phi\left(x_t\right)\right) && (3.17)\\
& +\beta_0\left(x_t\right)\left(A_1-A_2\right)\frac{2}{\varepsilon}\left(\widetilde{\phi}\left(x_t\right)-\phi\left(x_t\right)\right)\Big|\,. && (3.18)
\end{aligned}$$

Für die weiteren Abschätzungen verwendet man den Mittelwertsatz der Differentialrechnung mit Stellen η, $\widetilde{\eta}$ und $\overline{\eta}$, alle aus dem Inneren des Intervalls $[x_t; x_{t+1}]$ der Länge h. Damit ergibt sich

$$(3.16) \quad = \quad \beta_0'(\overline{\eta}) A_1 \frac{2h}{\varepsilon} \left(\widetilde{\phi}\left(x_{t+1}\right) - \phi\left(x_{t+1}\right) \right), \tag{3.19}$$

$$(3.17) \quad = \quad \beta_0\left(x_t\right) A_1 \frac{2h}{\varepsilon} \left(\widetilde{\phi}'(\widetilde{\eta}) - \phi'(\eta) \right). \tag{3.20}$$

Da nach [ABN, (2.13)] β_0' im Intervall $[0;1]$ beschränkt ist, erhält man

$$|(3.19)| \leq \|\beta_0'\|_{L^\infty(0;1)} \frac{2Eh}{\varepsilon}. \tag{3.21}$$

Betrachtet man mit dem Mittelwert der Differentialrechnung an einer Stelle ζ den Term in der Klammer aus (3.20), erhält man

$$\widetilde{\phi}'(\widetilde{\eta}) - \phi'(\eta) = \widetilde{\phi}'(\widetilde{\eta}) - \phi'(\widetilde{\eta}) + \phi'(\widetilde{\eta}) - \phi'(\eta) = \widetilde{\phi}'(\widetilde{\eta}) - \phi'(\widetilde{\eta}) + \phi''(\zeta)(\eta - \widetilde{\eta}).$$

Da $\left|\widetilde{\phi}'(\widetilde{\eta}) - \phi'(\widetilde{\eta})\right|$ nach (3.7a) mit E' beschränkt ist, $\phi''(\zeta)$ ebenfalls beschränkt ist und $|\eta - \widetilde{\eta}| < h$ ist, erhält man mit Dreiecksungleichung

$$|(3.20)| \leq \|\beta_0\|_{L^\infty(0;1)} \frac{2h}{\varepsilon} \left(E' + Ch\right). \tag{3.22}$$

Für (3.18) betrachtet man

$$A_1 = \int_0^1 \imath e^{\frac{2\imath}{\varepsilon}\left(\phi(x_{t+1}) + s\left(\widetilde{\phi}(x_{t+1}) - \phi(x_{t+1})\right)\right)} \mathrm{d}s$$

und

$$A_2 = \int_0^1 \imath e^{\frac{2\imath}{\varepsilon}\left(\phi(x_t) + s\left(\widetilde{\phi}(x_t) - \phi(x_t)\right)\right)} \mathrm{d}s,$$

die sich aus dem Mittelwertsatz für vektorwertige Funktionen ergeben.
Wendet man auf die beiden Integranden von A_1 und A_2 wieder den Mittelwertsatz für vektorwertige Funktionen an, erhält man unter Verwendung des Mittelwertsatzes der Differentialrechnung

$$\begin{gathered}
\imath e^{\frac{2\imath}{\varepsilon}\left(\phi(x_{t+1}) + s\left(\widetilde{\phi}(x_{t+1}) - \phi(x_{t+1})\right)\right)} - \imath e^{\frac{2\imath}{\varepsilon}\left(\phi(x_t) + s\left(\widetilde{\phi}(x_t) - \phi(x_t)\right)\right)} = \\
a(s)\tfrac{2}{\varepsilon} \left(\phi\left(x_{t+1}\right) - \phi\left(x_t\right) + s\left(\widetilde{\phi}\left(x_{t+1}\right) - \widetilde{\phi}\left(x_t\right) - \phi\left(x_{t+1}\right) + \phi\left(x_t\right) \right) \right) = \\
a(s)\tfrac{2}{\varepsilon} \left(\phi'(\eta)h + s\left(\widetilde{\phi}'(\widetilde{\eta})h - \phi'(\eta)h \right) \right)
\end{gathered}$$

wobei $a(s)$ wieder über das Integral aus dem Mittelwertsatz gegeben ist und $|a(s)| \leq 1$ gilt. Da alle Ableitungen nach (3.7b) beschränkt sind und $0 \leq s \leq 1$ ist, erhält man damit

$$|A_1 - A_2| \leq 2 \min\left(1, \frac{Ch}{\varepsilon}\right). \tag{3.23}$$

Aus (3.23) erhält man

$$|(3.18)| \leq \|\beta_0\|_{L^\infty(0;1)} \frac{4E}{\varepsilon} \min\left(1, \frac{Ch}{\varepsilon}\right). \tag{3.24}$$

Fasst man (3.21), (3.22) und (3.24) zusammen, erhält man

$$\begin{aligned} (3.11) \quad \leq \quad & \varepsilon \Big(2 \|\beta_0'\|_{L^\infty(0;1)} Eh + 2 \|\beta_0\|_{L^\infty(0;1)} h(E' + Ch) \\ & + 4E \|\beta_0\|_{L^\infty(0;1)} \min\left(1, \tfrac{Ch}{\varepsilon}\right) \Big) \end{aligned} \tag{3.25}$$

Formt man (3.11) noch vor der Anwendung des Mittelwertsatzes für vektorwertige Funktionen um und wendet erst danach den Mittelwertsatz für vektorwertige Funktionen und den Mittelwertsatz der Differentialrechnung an, erhält man

$$\begin{aligned} (3.11) \quad = \quad & \varepsilon^2 \Big| \beta_0(x_{t+1}) \left(e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_{t+1})} - e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_t)} \right) - \beta_0(x_t) \left(e^{\frac{2\imath}{\varepsilon}\phi(x_{t+1})} - e^{\frac{2\imath}{\varepsilon}\phi(x_t)} \right) \\ & + (\beta_0(x_{t+1}) - \beta_0(x_t)) \left(e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_t)} - e^{\frac{2\imath}{\varepsilon}\phi(x_{t+1})} \right) \Big| \\ = \quad & \varepsilon^2 \Big| \beta_0(x_{t+1}) A_3 \frac{2}{\varepsilon} \left(\widetilde{\phi}(x_{t+1}) - \widetilde{\phi}(x_t) \right) - \beta_0(x_t) A_4 \frac{2}{\varepsilon} (\phi(x_{t+1}) - \phi(x_t)) \\ & + (\beta_0(x_{t+1}) - \beta_0(x_t)) \left(e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_t)} - e^{\frac{2\imath}{\varepsilon}\phi(x_{t+1})} \right) \Big| \\ = \quad & \varepsilon^2 \left| \beta_0(x_{t+1}) A_3 \frac{2}{\varepsilon} \widetilde{\phi}'(\widetilde{\eta}) h - \beta_0(x_t) A_4 \frac{2}{\varepsilon} \phi'(\eta) h + \beta_0'(\overline{\eta}) h \left(e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_t)} - e^{\frac{2\imath}{\varepsilon}\phi(x_{t+1})} \right) \right|. \end{aligned}$$

Da alle Ableitungen beschränkt sind, erhält man daraus mit Dreiecksungleichung

$$(3.11) \leq \varepsilon h \left(4 \|\beta_0\|_{L^\infty(0;1)} C + 2\varepsilon \|\beta_0'\|_{L^\infty(0;1)} \right). \tag{3.26}$$

Fasst man (3.12), (3.15), (3.25) und (3.26) zusammen und vernachlässigt dabei $2 \|\beta_0\|_{L^\infty(0;1)} Ch^2 + 4E \|\beta_0\|_{L^\infty(0;1)} \min\left(1, \frac{Ch}{\varepsilon}\right)$ aus (3.25) und $2\varepsilon \|\beta_0'\|_{L^\infty(0;1)}$ aus (3.26), da dieser Terme im Vergleich zu den anderen in den Abschätzungen auftretenden Termen klein sind, erhält man

$$(3.11) \leq \varepsilon C \min(\varepsilon, E, h(E + E'), h). \tag{3.27}$$

Für (3.10) erhält man sofort, da β_1 nach [ABN, p. 1446] im Intervall $[0;1]$ beschränkt ist, mit der Dreiecksungleichung

$$(3.10) \leq 4\varepsilon^3 \left\|\beta_1\right\|_{L^\infty(0;1)}. \tag{3.28}$$

Wendet man den Mittelwertsatz für vektorwertige Funktionen an, erhält man

$$\begin{aligned}(3.10) &= \varepsilon^3 \left|\beta_1\left(x_{t+1}\right)\right| \left|\frac{2}{\varepsilon}A_1\left(\phi\left(x_{t+1}\right)-\widetilde{\phi}\left(x_{t+1}\right)\right)-\frac{2}{\varepsilon}A_2\left(\phi\left(x_t\right)-\widetilde{\phi}\left(x_t\right)\right)\right| \\ &\leq 4\varepsilon^2 \left\|\beta_1\right\|_{L^\infty(0;1)} E. \end{aligned} \tag{3.29}$$

Formt man (3.10) um und wendet erst anschließend den Mittelwertsatz für vektorwertige Funktionen und den Mittelwertsatz der Differentialrechnung an, erhält man mit (3.7b)

$$\begin{aligned}(3.10) &= \varepsilon^3 \left|\beta_1\left(x_{t+1}\right)\right| \left|\frac{2}{\varepsilon}A_3\left(\phi\left(x_{t+1}\right)-\phi\left(x_t\right)\right)-\frac{2}{\varepsilon}A_4\left(\widetilde{\phi}\left(x_{t+1}\right)-\widetilde{\phi}\left(x_t\right)\right)\right| \\ &= \varepsilon^3 \left|\beta_1\left(x_{t+1}\right)\right| \frac{2}{\varepsilon}\left|A_3\phi'(\eta)h - A_4\widetilde{\phi}'(\widetilde{\eta})h\right| \\ &\leq 4C\varepsilon^2 \left\|\beta_1\right\|_{L^\infty(0;1)} h. \end{aligned} \tag{3.30}$$

Fasst man (3.28), (3.29) und (3.30) zusammen, erhält man

$$(3.10) \leq \varepsilon^2 C \min\left(\varepsilon, E, h\right). \tag{3.31}$$

Mit (3.27) und (3.31) ergibt sich

$$\left|\left(A_t^1 - \widetilde{A}_t^1\right)\right| \leq \varepsilon C \min\left(\varepsilon, E, h(E+E'), h\right) + \varepsilon^2 C \min\left(\varepsilon, E, h\right). \tag{3.32}$$

Da nach Voraussetzung $\varepsilon < 1$ ist, ist $\varepsilon^2 < \varepsilon$. Daher ist (3.32) von der Größenordnung $\mathcal{O}\left(\varepsilon \min\left(\varepsilon, E, h(E+E'), h\right)\right)$ für $\varepsilon \to 0$.
Führt man den ersten Schritt des Verfahrens aus, erhält man daher, da Z_1 die vorgegebene Anfangsbedingung ist,

$$\begin{aligned}\left\|Z_2 - \widetilde{Z}_2\right\| &= \left\|\left(I + A_1^1\right)Z_1 - \left(I + \widetilde{A}_1^1\right)Z_1\right\| \\ &= \left\|\left(A_1^1 - \widetilde{A}_1^1\right)Z_1\right\|. \end{aligned}$$

Daraus ergibt sich

$$\left\|Z_2 - \widetilde{Z}_2\right\| \leq \varepsilon C \min\left(\varepsilon, E, h(E+E'), h\right).$$

Um, von $Z_1 = Z_I$ ausgehend, Z_T und $\widetilde{Z}_T$ zu berechnen, benötigt man insgesamt $(T-1) = \frac{1}{h}$ Schritte.
Somit erhält man für den globalen Fehler

$$\left\| Z_t - \widetilde{Z}_t \right\| \leq \varepsilon C \min\left(\frac{\varepsilon}{h}, \frac{E}{h}, (E+E'), 1 \right). \tag{3.33}$$

□

Beweis von Satz 3.5. Aus der Dreiecksungleichung, (1.30) aus Satz 1.1 und (3.9) aus Satz 3.6 folgt

$$\begin{aligned} \left\| Z(x_t) - \widetilde{Z}_t \right\| &= \left\| Z(x_t) - Z_t + Z_t - \widetilde{Z}_t \right\| \\ &\leq \| Z(x_t) - Z_t \| + \left\| Z_t - \widetilde{Z}_t \right\| \\ &\leq C_1 \varepsilon^2 \min(\varepsilon, h) + \varepsilon C_2 \min\left(\frac{\varepsilon}{h}, \frac{E}{h}, (E+E'), 1 \right). \end{aligned}$$

□

In [ABN, Section 3.3] wird neben dem Verfahren 1. Ordnung zusätzlich ein modifiziertes Verfahren 1. Ordnung

$$Z_{t+1} = \left(I + \overline{A}_t^1 \right) Z_t, \ 1 \leq t \leq T-1,$$

mit Anfangsbedingung $Z_1 = Z_I$, präsentiert. Dabei ist

$$\overline{A}_t^1 := -\imath \varepsilon^2 \beta_0(x_{t+1}) \begin{pmatrix} 0 & e^{\frac{-2\imath}{\epsilon}\phi(x_t)} - e^{\frac{-2\imath}{\varepsilon}\phi(x_{t+1})} \\ e^{\frac{2\imath}{\varepsilon}\phi(x_{t+1})} - e^{\frac{2\imath}{\varepsilon}\phi(x_{t+1})} & 0 \end{pmatrix}.$$

Dieses Verfahren soll hier nur kurz erwähnt werden und die Fehlerabschätzung wieder um den Quadraturfehler erweitert werden. In den weiteren Kapiteln wird dieses Verfahren nicht mehr verwendet, da der Fehler im Vergleich zum ersten Teil von (3.8) schlechter ist.
Für die Abschätzung des Fehlers dieses Verfahrens gelten dieselben Voraussetzungen wie in Satz 1.1.

Satz 3.7 (Theorem 3.5 aus [ABN])**.** *Für den Fehler des modifizierten Verfahrens 1. Ordnung gilt*

$$\| Z(x_t) - Z_t \|_2 \leq C\varepsilon \min(\varepsilon, h), \ 1 \leq t \leq T,$$

und

$$\| U(x_t) - U_t \|_2 \leq C \frac{h^\gamma}{\varepsilon} + C\varepsilon \min(\varepsilon, h), \ 1 \leq t \leq T.$$

Beweis. siehe [ABN, p. 1455] □

Satz 3.8. *Unter den Voraussetzungen von Satz 3.5 gilt für das modifizierte Verfahren 1. Ordnung unter Berücksichtigung des Quadraturfehlers*

$$\left\| Z\left(x_t\right) - \widetilde{Z}_t \right\|_2 \leq C_1 \varepsilon \min\left(\varepsilon, h\right) + \varepsilon C_2 \min\left(\frac{\varepsilon}{h}, \frac{E}{h}, 1\right).$$

Satz 3.9. *Für den Fehler (3.6) im modifizierten Verfahren 1. Ordnung gilt*

$$\left\| Z_t - \widetilde{Z}_t \right\|_2 \leq \varepsilon C \min\left(\frac{\varepsilon}{h}, \frac{E}{h}, 1\right), \ 1 \leq t \leq T.$$

Beweis. Hier geht man vor wie im Beweis von Satz 3.6 und betrachtet

$$\begin{aligned}
\left(\overline{A}_t^1 - \widetilde{\overline{A}}_t^1\right)_{2,1} &= -\imath\varepsilon^2 \beta_0\left(x_{t+1}\right)\left(e^{\frac{2\imath}{\varepsilon}\phi(x_{t+1})} - e^{\frac{2\imath}{\varepsilon}\phi(x_t)}\right) \\
&\quad + \imath\varepsilon^2 \beta_0\left(x_{t+1}\right)\left(e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_{t+1})} - e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_t)}\right) \\
&= \imath\varepsilon^2 \beta_0\left(x_{t+1}\right)\left(e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_{t+1})} - e^{\frac{2\imath}{\varepsilon}\phi(x_{t+1})} + e^{\frac{2\imath}{\varepsilon}\phi(x_t)} - e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_t)}\right).
\end{aligned}$$

Diese Differenz ist von ähnlicher Struktur wie (3.10) und man erhält auf gleiche Weise

$$\left| \left(\overline{A}_t^1 - \widetilde{\overline{A}}_t^1\right)_{2,1} \right| \leq \varepsilon C \min\left(\varepsilon, E, h\right),$$

und damit für den ersten Schritt des modifizierten Verfahrens 1. Ordnung

$$\begin{aligned}
\left\| Z_2 - \widetilde{Z}_2 \right\|_2 &= \left\| \left(I + \overline{A}_1^1\right) Z_1 - \left(I + \widetilde{\overline{A}}_1^1\right) Z_1 \right\|_2 \\
&= \left\| \left(\overline{A}_1^1 - \widetilde{\overline{A}}_1^1\right) Z_1 \right\|_2 \\
&\leq \varepsilon C \min\left(\varepsilon, E, h\right).
\end{aligned}$$

Führt man die Iteration fort, erhält man insgesamt eine Abschätzung

$$\left\| Z_t - \widetilde{Z}_t \right\|_2 \leq \varepsilon C \min\left(\frac{\varepsilon}{h}, \frac{E}{h}, 1\right).$$

□

Beweis von Satz 3.8.

$$\begin{aligned}\left\|Z(x_t) - \widetilde{Z}_t\right\|_2 &= \left\|Z(x_t) - Z_t + Z_t - \widetilde{Z}_t\right\|_2 \\ &\leq \|Z(x_t) - Z_t\|_2 + \left\|Z_t - \widetilde{Z}_t\right\|_2 \\ &\leq C_1 \varepsilon \min(\varepsilon, h) + \varepsilon C_2 \min\left(\frac{\varepsilon}{h}, \frac{E}{h}, 1\right).\end{aligned}$$

□

Bemerkung 3.10.

1. Die Abschätzung (3.9) zeigt, dass der zusätzliche Fehler durch die Näherung des Phasenintegrals (1.2) beschränkt bleibt, wenn h gegen Null geht. Dies ist bei der Berechnung mit der baryzentrisch interpolierten Stammfunktion wie in Abschnitt 3.1.1 wichtig, da hier der Fehler von der Schrittweite h unabhängig ist.

2. Der Fehler E' aus Bedingung (3.7a) kann, wenn $\widetilde{\phi}'$ aus ϕ mit Hilfe der Differentationsmatrix aus Abschnitt 2.3 (vgl. (2.46)) an den Chebyshevpunkten berechnet wird, mit [Tr1, Theorem 6] abgeschätzt werden und ist dann wieder von der Form

   $$\left|\phi'(x) - \widetilde{\phi}'(x)\right| \leq C\rho^{-N}.$$

   Dabei ist ϕ wieder analytisch auf und innerhalb der Ellipse mit Summe der Halbachsen $\rho > 1$.

3. Man kann die Bedingung (3.7a) vermeiden, wenn man in (3.20) den Term $\left|\widetilde{\phi}'(\widetilde{\eta}) - \phi'(\eta)\right|$ mit $2\max\left(\left|\widetilde{\phi}'(\widetilde{\eta})\right|, |\phi'(\eta)|\right) = K$ abschätzt. Dabei wird nur die Beschränktheit der Ableitungen (3.7b) vorausgesetzt, die auch in den anderen Abschätzungen verwendet wird.

4. Wird $\widetilde{\phi}(x)$ schrittweise, so wie in Abschnitt 3.1.2, berechnet, ist ρ und somit (3.1) von der Schrittweite h abhängig und es gilt

   $$E \leq C\rho(h)^{-N-2} \to 0 \text{ für } h \to 0.$$

5. In den Sätzen in diesem Abschnitt wurde vorausgesetzt, dass die Funktionen $\beta(x)$ und $\beta_k(x)$ aus $a(x)$ exakt berechnet werden können. Sollte es nicht möglich sein, diese exakt zu berechnen, kann man die auftretenden Ableitungen mit Hilfe der Differentationsmatrix aus Abschnitt 2.3 an Chebyshevpunkten näherungsweise berechnen und diese Werte dann mit Hilfe der baryzentrischen Interpolation an den Gitterpunkten

x_t, $1 \leq t \leq T$, nähern. Dabei entsteht ein weiterer zusätzlicher Fehler, der ebenfalls in den Abschätzungen berücksichtigt werden kann. In dieser Arbeit wurde nur der Fehler bei der Näherung des Phasenintegrals (1.2) berücksichtigt, da in den Kapiteln 4 und 5 die Funktionen $\beta(x)$ und $\beta_k(x)$ exakt berechnet worden sind.

3.2.2 Quadraturfehler im Verfahren 2. Ordnung

Das Verfahren 2. Ordnung setzt sich aus 2 Matrizen, $A^1_{mod,t}$ und A^2_t, zusammen, die auf Grund ihrer Struktur im folgenden Beweis der Abschätzung getrennt betrachtet werden.

Satz 3.11. *Es sei $a(x) \in C^\infty[0;1]$ eine beliebige, glatte und reellwertige Funktion, die $a(x) \geq a_0 > 0$ im Intervall $[0;1]$ erfülle, und $0 < \varepsilon < 1$ eine beliebige, aber feste reelle Zahl und sei weiters (3.7) erfüllt. Dann gilt für den Fehler im Verfahren 2. Ordnung (1.26) unter der Berücksichtigung des Quadraturfehlers*

$$\begin{aligned}\left\|Z(x_t) - \widetilde{Z}_t\right\| \;&\leq\; C_1\varepsilon^3 h^2 + \varepsilon C_2 \min\left(\frac{\varepsilon}{h}, \frac{E}{h}, (E+E'), 1\right) \\ &\quad + \varepsilon^3 C_3 \min\left(\frac{\varepsilon}{h}, \frac{E}{h}, 1\right),\ 1 \leq t \leq T. \end{aligned} \tag{3.34}$$

Dabei bezeichne $Z(x_t)$ die exakte Lösung, $\widetilde{Z}_t$ die numerische Lösung mit fehlerbehafteter Phase $\widetilde{\phi}(x_t)$, E den Fehler in der Phasenberechnung (3.1) und $\|\cdot\|$ eine beliebige Vektornorm in $\mathbb{C}^2$.

Satz 3.12. *Es sei (3.7) erfüllt. Für den Fehler (3.6) im Verfahren 2. Ordnung gilt*

$$\left\|Z_t - \widetilde{Z}_t\right\| \leq \varepsilon C_1 \min\left(\frac{\varepsilon}{h}, \frac{E}{h}, (E+E'), 1\right) + \varepsilon^3 C_2 \min\left(\frac{\varepsilon}{h}, \frac{E}{h}, 1\right),\ 1 \leq t \leq T. \tag{3.35}$$

Beweis.

$$\left|\left(A^1_{mod,t} - \widetilde{A}^1_{mod,t}\right)_{2,1}\right| =$$

$$\begin{aligned} = \;& \Big|-\imath\varepsilon^2\left(\beta_0(x_{t+1})\, e^{\frac{2\imath}{\varepsilon}\phi(x_{t+1})} - \beta_0(x_t)\, e^{\frac{2\imath}{\varepsilon}\phi(x_t)}\right) \\ & +\varepsilon^3\left(\beta_1(x_{t+1})\, e^{\frac{2\imath}{\varepsilon}\phi(x_{t+1})} - \beta_1(x_t)\, e^{\frac{2\imath}{\varepsilon}\phi(x_t)}\right) \\ & +\imath\varepsilon^4\beta_2(x_{t+1})\, e^{\frac{2\imath}{\varepsilon}\phi(x_t)}\left(e^{\frac{2\imath}{\varepsilon}(\phi(x_{t+1})-\phi(x_t))} - 1\right) \\ & -\varepsilon^5\beta_3(x_{t+1})\, e^{\frac{2\imath}{\varepsilon}\phi(x_t)}\left(e^{\frac{2\imath}{\varepsilon}(\phi(x_{t+1})-\phi(x_t))} - 1 - \frac{2\imath}{\varepsilon}\left(\phi(x_{t+1}) - \phi(x_t)\right)\right) \end{aligned}$$

$$
\begin{aligned}
&+\imath\varepsilon^2\left(\beta_0\left(x_{t+1}\right)e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_{t+1})}-\beta_0\left(x_t\right)e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_t)}\right)\\
&-\varepsilon^3\left(\beta_1\left(x_{t+1}\right)e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_{t+1})}-\beta_1\left(x_t\right)e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_t)}\right)\\
&-\imath\varepsilon^4\beta_2\left(x_{t+1}\right)e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_t)}\left(e^{\frac{2\imath}{\varepsilon}\left(\widetilde{\phi}(x_{t+1})-\widetilde{\phi}(x_t)\right)}-1\right)\\
&+\varepsilon^5\beta_3\left(x_{t+1}\right)e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_t)}\left(e^{\frac{2\imath}{\varepsilon}\left(\widetilde{\phi}(x_{t+1})-\widetilde{\phi}(x_t)\right)}-1-\frac{2\imath}{\varepsilon}\left(\widetilde{\phi}\left(x_{t+1}\right)-\widetilde{\phi}\left(x_t\right)\right)\right)\Bigg|
\end{aligned}
$$

$$
\begin{aligned}
\leq\ & \varepsilon^2\left|\beta_0\left(x_{t+1}\right)\left(e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_{t+1})}-e^{\frac{2\imath}{\varepsilon}\phi(x_{t+1})}\right)-\beta_0\left(x_t\right)\left(e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_t)}-e^{\frac{2\imath}{\varepsilon}\phi(x_t)}\right)\right| && (3.36)\\
&+\varepsilon^3\left|\beta_1\left(x_{t+1}\right)\left(e^{\frac{2\imath}{\varepsilon}\phi(x_{t+1})}-e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_{t+1})}\right)-\beta_1\left(x_t\right)\left(e^{\frac{2\imath}{\varepsilon}\phi(x_t)}-e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_t)}\right)\right| && (3.37)\\
&+\varepsilon^4\left|\beta_2\left(x_{t+1}\right)\left(e^{\frac{2\imath}{\varepsilon}\phi(x_{t+1})}-e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_{t+1})}+e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_t)}-e^{\frac{2\imath}{\varepsilon}\phi(x_t)}\right)\right| && (3.38)\\
&+\varepsilon^5\left|\beta_3\left(x_{t+1}\right)\left(e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_{t+1})}-e^{\frac{2\imath}{\varepsilon}\phi(x_{t+1})}+e^{\frac{2\imath}{\varepsilon}\phi(x_t)}-e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_t)}\right)\right| && (3.39)\\
&+2\varepsilon^4\left|\beta_3\left(x_{t+1}\right)\left(e^{\frac{2\imath}{\varepsilon}\phi(x_t)}\left(\phi\left(x_{t+1}\right)-\phi\left(x_t\right)\right)-e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_t)}\left(\widetilde{\phi}\left(x_{t+1}\right)-\widetilde{\phi}\left(x_t\right)\right)\right)\right|. && (3.40)
\end{aligned}
$$

Die Terme (3.36) und (3.37) sind von der Struktur ähnlich zu (3.11) und man erhält durch analoges Vorgehen

$$
\begin{aligned}
(3.36) &\leq \varepsilon C\min\left(\varepsilon,E,h(E+E'),h\right), && (3.41)\\
(3.37) &\leq \varepsilon^2 C\min\left(\varepsilon,E,h(E+E'),h\right). && (3.42)
\end{aligned}
$$

Die Terme (3.38) und (3.39) sind von der Struktur ähnlich zu (3.10) und man erhält auf analoge Weise

$$
\begin{aligned}
(3.38) &\leq \varepsilon^3 C\min\left(\varepsilon,E,h\right), && (3.43)\\
(3.39) &\leq \varepsilon^4 C\min\left(\varepsilon,E,h\right). && (3.44)
\end{aligned}
$$

Für (3.40) erhält man mit dem Mittelwertsatz der Differentialrechnung

$$
\begin{aligned}
(3.40) &= 2\varepsilon^4\left|\beta_3\left(x_{t+1}\right)\left(e^{\frac{2\imath}{\varepsilon}\phi(x_t)}\phi'(\eta)h-e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}(x_t)}\widetilde{\phi}'(\widetilde{\eta})h\right)\right|\\
&\leq 4\varepsilon^4 h\left\|\beta_3\right\|_{L^\infty(0;1)}\max\left(\left|\phi'(\eta)\right|,\left|\widetilde{\phi}'(\widetilde{\eta})\right|\right). && (3.45)
\end{aligned}
$$

Wendet man die Abschätzung aus dem Mittelwertsatz für vektorwertige Funktionen auf die Funktion

$$
\vec{f}(x,y)=e^{\frac{2\imath}{\varepsilon}x}(y-x)=\begin{pmatrix}(y-x)\cos\frac{2x}{\varepsilon}\\(y-x)\sin\frac{2x}{\varepsilon}\end{pmatrix}
$$

mit Jacobimatrix

$$\vec{f}'(x,y) = \begin{pmatrix} -\frac{2}{\varepsilon}(y-x)\sin\frac{2x}{\varepsilon} - \cos\frac{2x}{\varepsilon} & \cos\frac{2x}{\varepsilon} \\ \frac{2}{\varepsilon}(y-x)\cos\frac{2x}{\varepsilon} - \sin\frac{2x}{\varepsilon} & \sin\frac{2x}{\varepsilon} \end{pmatrix}$$

an, erhält man

$$\begin{aligned} (3.40) &\leq 2\varepsilon^4 M \left\|\beta_3\right\|_{L^\infty(0;1)} \sqrt{\left(\phi(x_t) - \widetilde{\phi}(x_t)\right)^2 + \left(\phi(x_{t+1}) - \widetilde{\phi}(x_{t+1})\right)^2} \\ &\leq 2\sqrt{2}\varepsilon^4 M \left\|\beta_3\right\|_{L^\infty(0;1)} E, \end{aligned} \tag{3.46}$$

mit

$$M = \max_{s\in[0;1]} \left|\left|\left| \vec{f}' \left(\widetilde{\phi}(x_t) + s(\phi(x_t) - \widetilde{\phi}(x_t)), \widetilde{\phi}(x_{t+1}) + s(\phi(x_{t+1}) - \widetilde{\phi}(x_{t+1})) \right) \right|\right|\right|.$$

Dabei bezeichne $|||\cdot|||$ eine mit $\|\cdot\|_2 = |\cdot|$ verträgliche Matrixnorm auf $\mathbb{R}^{2\times 2}$, z. B. die Spektralnorm.
Fasst man (3.45) und (3.46) zusammen, ergibt sich

$$(3.40) \leq \varepsilon^4 C \min(h, E). \tag{3.47}$$

Insgesamt hat man mit (3.41), (3.42), (3.43), (3.44) und (3.47)

$$\begin{aligned} \left| \left(A^1_{mod,t} - \widetilde{A}^1_{mod,t} \right)_{2,1} \right| \leq\ & \varepsilon C \min(\varepsilon, E, h(E+E'), h) \\ & + \varepsilon^2 C \min(\varepsilon, E, h(E+E'), h) \\ & + \varepsilon^3 C \min(\varepsilon, E, h) \\ & + \varepsilon^4 C \min(\varepsilon, E, h) \\ & + \varepsilon^4 C \min(h, E). \end{aligned} \tag{3.48}$$

Da $\varepsilon < 1$ vorausgesetzt wird, ist (3.48) für $\varepsilon \to 0$ von der Größenordnung $\mathcal{O}(\varepsilon \min(\varepsilon, E, h(E+E'), h))$ und somit ergibt sich für diesen Teil

$$\left| \left(A^1_{mod,t} - \widetilde{A}^1_{mod,t} \right)_{2,1} \right| \leq \varepsilon C \min(\varepsilon, E, h(E+E'), h). \tag{3.49}$$

Für den zweiten Teil betrachtet man

$$\left| \left(A^2_t - \widetilde{A}^2_t \right)_{2,2} \right| =$$

$$\begin{aligned} = \Bigg| & \imath\varepsilon^3 \left((x_{t+1}) - (x_t) \right) \frac{\beta(x_{t+1})\beta_0(x_{t+1}) + \beta(x_t)\beta_0(x_t)}{2} \\ & - \varepsilon^4 \beta_0(x_t)\beta_0(x_{t+1}) \left(e^{\frac{2\imath}{\varepsilon}(\phi(x_{t+1}) - \phi(x_t))} - 1 \right) \\ & - \imath\varepsilon^5 \beta_1(x_{t+1}) \left(\beta_0(x_t) - \beta_0(x_{t+1}) \right) \left(e^{\frac{2\imath}{\varepsilon}(\phi(x_{t+1}) - \phi(x_t))} - 1 - \frac{2\imath}{\varepsilon}\left(\phi(x_{t+1}) - \phi(x_t)\right) \right) \end{aligned}$$

$$
\begin{aligned}
&-\imath\varepsilon^3\left((x_{t+1})-(x_t)\right)\frac{\beta\left(x_{t+1}\right)\beta_0\left(x_{t+1}\right)+\beta\left(x_t\right)\beta_0\left(x_t\right)}{2}\\
&+\varepsilon^4\beta_0\left(x_t\right)\beta_0\left(x_{t+1}\right)\left(e^{\frac{2\imath}{\varepsilon}\left(\widetilde{\phi}(x_{t+1})-\widetilde{\phi}(x_t)\right)}-1\right)\\
&+\imath\varepsilon^5\beta_1\left(x_{t+1}\right)\left(\beta_0\left(x_t\right)-\beta_0\left(x_{t+1}\right)\right)\left(e^{\frac{2\imath}{\varepsilon}\left(\widetilde{\phi}(x_{t+1})-\widetilde{\phi}(x_t)\right)}-1-\frac{2\imath}{\varepsilon}\left(\widetilde{\phi}\left(x_{t+1}\right)-\widetilde{\phi}\left(x_t\right)\right)\right)\Bigg|
\end{aligned}
$$

$$
\begin{aligned}
\leq\ & \varepsilon^4\left|\beta_0\left(x_t\right)\beta_0\left(x_{t+1}\right)\left(e^{\frac{2\imath}{\varepsilon}\left(\widetilde{\phi}(x_{t+1})-\widetilde{\phi}(x_t)\right)}-e^{\frac{2\imath}{\varepsilon}\left(\phi(x_{t+1})-\phi(x_t)\right)}\right)\right| && (3.50)\\
&+\varepsilon^5\left|\beta_1\left(x_{t+1}\right)\left(\beta_0\left(x_t\right)-\beta_0\left(x_{t+1}\right)\right)\left(e^{\frac{2\imath}{\varepsilon}\left(\widetilde{\phi}(x_{t+1})-\widetilde{\phi}(x_t)\right)}-e^{\frac{2\imath}{\varepsilon}\left(\phi(x_{t+1})-\phi(x_t)\right)}\right)\right| && (3.51)\\
&+2\varepsilon^4\left|\beta_1\left(x_{t+1}\right)\left(\widetilde{\phi}\left(x_{t+1}\right)-\phi\left(x_{t+1}\right)+\phi\left(x_t\right)-\widetilde{\phi}\left(x_t\right)\right)\right| && (3.52)
\end{aligned}
$$

Mit der Dreiecksungleichung erhält man für (3.50) die Abschätzung

$$
(3.50)\leq 2\varepsilon^4\left\|\beta_0\right\|^2_{L^\infty(0;1)}. \tag{3.53}
$$

Wendet man die Abschätzung aus dem Mittelwertsatz für vektorwertige Funktionen auf

$$
\vec{f}(x,y)=e^{\frac{2\imath}{\varepsilon}(x-y)}=\begin{pmatrix}\cos\frac{2}{\varepsilon}(x-y)\\ \sin\frac{2}{\varepsilon}(x-y)\end{pmatrix} \tag{3.54}
$$

mit Jacobimatrix

$$
\vec{f}(x,y)=\frac{2}{\varepsilon}\begin{pmatrix}-\sin\frac{2}{\varepsilon}(x-y) & \sin\frac{2}{\varepsilon}(x-y)\\ \cos\frac{2}{\varepsilon}(x-y) & -\cos\frac{2}{\varepsilon}(x-y)\end{pmatrix} \tag{3.55}
$$

an, erhält man

$$
(3.50)\leq 4\varepsilon^3\left\|\beta_0\right\|^2_{L^\infty(0;1)}E, \tag{3.56}
$$

da die Jacobimatrix (3.55) Spektralnorm $\frac{2\sqrt{2}}{\varepsilon}$ hat.
Wendet man zuerst den Mittelwertsatz der Differentialrechnung im Exponenten an und erst anschließend den Mittelwertsatz für vektorwertige Funktionen mit der in (3.13) definierten Funktion, erhält man

$$
\begin{aligned}
(3.50) &= \varepsilon^4\left|\beta_0\left(x_t\right)\beta_0\left(x_{t+1}\right)\left(e^{\frac{2\imath}{\varepsilon}\widetilde{\phi}'(\widetilde{\eta})h}-e^{\frac{2\imath}{\varepsilon}\phi'(\eta)h}\right)\right|\\
&\leq 2\varepsilon^3\left\|\beta_0\right\|^2_{L^\infty(0;1)}h\left|\widetilde{\phi}'(\widetilde{\eta})-\phi'(\eta)\right|\\
&\leq 4\varepsilon^3\left\|\beta_0\right\|^2_{L^\infty(0;1)}h\max\left(\left|\widetilde{\phi}'(\widetilde{\eta})\right|,\left|\phi'(\eta)\right|\right) && (3.57)
\end{aligned}
$$

Aus (3.53), (3.56) und (3.57) erhält man

$$(3.50) \leq \varepsilon^3 C \min(\varepsilon, E, h). \tag{3.58}$$

Für (3.51) erhält man mit der Dreiecksungleichung

$$(3.51) \leq 4\varepsilon^5 \|\beta_0\|_{L^\infty(0;1)} \|\beta_1\|_{L^\infty(0;)}. \tag{3.59}$$

Wendet man wieder den Mittelwertsatz für vektorwertige Funktionen mit der in (3.54) definierten Funktion und den Mittelwertsatz der Differentialrechnung an, erhält man

$$(3.51) \leq 4\varepsilon^4 \|\beta_1\|_{L^\infty(0;1)} \|\beta_0'\|_{L^\infty(0;1)} hE \tag{3.60}$$

Fasst man (3.59) und (3.60) zusammen, ergibt sich

$$(3.51) \leq \varepsilon^4 C \min(\varepsilon, hE). \tag{3.61}$$

Für (3.52) erhält man mit der Dreiecksungleichung und Bedingung (3.7b)

$$(3.52) \leq 8\varepsilon^4 \|\beta_1\|_{L^\infty(0;1)} \max\left(|\phi(x_t)|, |\phi(x_{t+1})|, \left|\widetilde{\phi}(x_t)\right|, \left|\widetilde{\phi}(x_{t+1})\right|\right). \tag{3.62}$$

Unter Verwendung von $\left|\phi - \widetilde{\phi}\right| \leq E$ erhält man

$$(3.52) \leq 8\varepsilon^4 \|\beta_1\|_{L^\infty(0;1)} E. \tag{3.63}$$

Mit dem Mittelwertsatz der Differentialrechnung und Bedingung (3.7b) erhält man

$$\begin{aligned} (3.52) &= 2\varepsilon^4 \left|\beta_1(x_{t+1}) h \left(\widetilde{\phi}'(\widetilde{\eta}) - \phi'(\eta)\right)\right| \\ &\leq 2\varepsilon^4 \|\beta_1\|_{L^\infty(0;1)} h \max\left(\left|\widetilde{\phi}'(\widetilde{\eta})\right|, |\phi'(\eta)|\right). \end{aligned} \tag{3.64}$$

Somit erhält man die Abschätzung

$$(3.52) \leq \varepsilon^4 C \min(1, E, h). \tag{3.65}$$

Aus (3.58), (3.61) und (3.65) erhält man die Abschätzung

$$\begin{aligned} \left|\left(A_t^2 - \widetilde{A}_t^2\right)_{2,2}\right| \leq\ & \varepsilon^3 C \min(\varepsilon, E, h) \\ & + \varepsilon^4 C \min(\varepsilon, hE) \\ & + \varepsilon^4 C \min(1, E, h). \end{aligned} \tag{3.66}$$

Aus (3.66) erhält man, da $\varepsilon < 1$ ist,

$$\left| \left(A_t^2 - \widetilde{A}_t^2 \right)_{2,2} \right| \leq \varepsilon^3 C \min\left(\varepsilon, E, h\right). \tag{3.67}$$

Somit ergibt sich aus (3.49) und (3.67) für den ersten Schritt des Verfahrens

$$\begin{aligned}
\left\| Z_2 - \widetilde{Z}_2 \right\| &= \left\| \left(I + A_{mod,1}^1 + A_1^2 \right) Z_1 - \left(I + \widetilde{A}_{mod,1}^1 + \widetilde{A}_1^2 \right) Z_1 \right\| \\
&= \left\| I Z_1 + A_{mod,1}^1 Z_1 + A_1^2 Z_1 - I Z_1 - \widetilde{A}_{mod,1}^1 Z_1 - \widetilde{A}_1^2 Z_1 \right\| \\
&= \left\| \left(A_{mod,1}^1 - \widetilde{A}_{mod,1}^1 \right) Z_1 + \left(A_1^2 - \widetilde{A}_1^2 \right) Z_1 \right\| \\
&\leq \left\| \left(A_{mod,1}^1 - \widetilde{A}_{mod,1}^1 \right) Z_1 \right\| + \left\| \left(A_1^2 - \widetilde{A}_1^2 \right) Z_1 \right\| \\
&\leq \varepsilon C_1 \min\left(\varepsilon, E, h(E+E'), h\right) + \varepsilon^3 C_2 \min\left(\varepsilon, E, h\right).
\end{aligned}$$

Da $(T-1) = \frac{1}{h}$ Schritte notwendig sind, um Z_T und $\widetilde{Z}_T$ zu berechnen, erhält man für den globalen Fehler

$$\left\| Z_t - \widetilde{Z}_t \right\| \leq \varepsilon C_1 \min\left(\frac{\varepsilon}{h}, \frac{E}{h}, (E+E'), 1 \right) + \varepsilon^3 C_2 \min\left(\frac{\varepsilon}{h}, \frac{E}{h}, 1 \right). \tag{3.68}$$

□

Beweis von Satz 3.11. Aus der Dreiecksungleichung, (1.32) aus Satz 1.1 und (3.35) aus Satz 3.11 folgt

$$\begin{aligned}
\left\| Z\left(x_t\right) - \widetilde{Z}_t \right\| &= \left\| Z\left(x_t\right) - Z_t + Z_t - \widetilde{Z}_t \right\| \\
&\leq \left\| Z\left(x_t\right) - Z_t \right\| + \left\| Z_t - \widetilde{Z}_t \right\| \\
&\leq C_1 \varepsilon^3 h^2 + \varepsilon C_2 \min\left(\frac{\varepsilon}{h}, \frac{E}{h}, (E+E'), 1 \right) + \varepsilon^3 C_3 \min\left(\frac{\varepsilon}{h}, \frac{E}{h}, 1 \right).
\end{aligned}$$

□

Bemerkung 3.13. Bemerkung 3.10 lässt sich auch auf diesen Abschnitt mit den Abschätzungen für das Verfahren 2. Ordnung übertragen.

3.2.3 Quadraturfehler in der Berechnung von U

Aus Z kann U durch (1.29),

$$U_t = P^{-1} e^{\frac{t}{\varepsilon} \Phi^{\varepsilon}(x_t)} Z_t,$$

berechnet werden. Da die Matrizen P^{-1}, (1.14,)

$$P^{-1} = \frac{1}{\sqrt{2}} \begin{pmatrix} -\imath & 1 \\ 1 & -\imath \end{pmatrix},$$

und $\Phi^{\varepsilon}(x_t)$, (1.16),

$$\Phi^{\varepsilon}(x) = \begin{pmatrix} \int_0^x \left(\sqrt{a(\tau)} - \varepsilon^2 \beta(\tau) \right) \mathrm{d}\tau & 0 \\ 0 & -\int_0^x \left(\sqrt{a(\tau)} - \varepsilon^2 \beta(\tau) \right) \mathrm{d}\tau \end{pmatrix},$$

unitär sind, ist das Produkt $P^{-1} e^{\frac{\imath}{\varepsilon}\Phi^{\varepsilon}(x_t)}$ ebenfalls unitär. Da für eine unitäre Matrix $A \in \mathbb{C}^{2\times 2}$

$$\|Ax\|_2 = \|x\|_2 \ \forall x \in \mathbb{C}^2$$

gilt, folgt

$$\|U_t\|_2 = \left\| P^{-1} e^{\frac{\imath}{\varepsilon}\Phi^{\varepsilon}(x_t)} Z_t \right\|_2 = \|Z_t\|_2 .$$

Betrachtet man jedoch

$$\begin{aligned} \left\| U(x_t) - \widetilde{U}_t \right\|_2 &= \left\| P^{-1} e^{\frac{\imath}{\varepsilon}\Phi^{\varepsilon}(x_t)} Z(x_t) - P^{-1} e^{\frac{\imath}{\varepsilon}\widetilde{\Phi}^{\varepsilon}(x_t)} \widetilde{Z}_t \right\|_2 \\ &= \left\| P^{-1} \left(e^{\frac{\imath}{\varepsilon}\Phi^{\varepsilon}(x_t)} Z(x_t) - e^{\frac{\imath}{\varepsilon}\widetilde{\Phi}^{\varepsilon}(x_t)} \widetilde{Z}_t \right) \right\|_2 \\ &= \left\| e^{\frac{\imath}{\varepsilon}\Phi^{\varepsilon}(x_t)} Z(x_t) - e^{\frac{\imath}{\varepsilon}\widetilde{\Phi}^{\varepsilon}(x_t)} \widetilde{Z}_t \right\|_2 , \end{aligned}$$

muss der Quadraturfehler bei der Berechnung von $\widetilde{\Phi}^{\varepsilon}(x_t)$ noch berücksichtigt werden.
In Satz 1.1 wird dies soweit berücksichtigt, dass

$$\|U(x_t) - U_t\|$$

mit (1.31),

$$\|U(x_t) - U_t\| \leq C \frac{h^{\gamma}}{\varepsilon} + C\varepsilon^2 \min(\varepsilon, h),$$

im Verfahren 1. Ordnung und mit (1.33),

$$\|U(x_t) - U_t\| \leq C \frac{h^{\gamma}}{\varepsilon} + C\varepsilon^3 h^2,$$

im Verfahren 2. Ordnung, jeweils für $1 \leq t \leq T$, abgeschätzt wird. Dabei wurde jedoch vorausgesetzt, dass das Phasenintegral (1.2),

$$\phi(x) = \int_0^x \left(\sqrt{a(\tau)} - \varepsilon^2 \beta(\tau) \right) \mathrm{d}\tau,$$

in der Berechnung von Z exakt berechnet worden war. Diese Abschätzungen können erweitert werden, indem man auch bei der Berechnung von Z den Quadraturfehler berücksichtigt.

Satz 3.14. *Es sei $a(x) \in C^\infty[0;1]$ eine beliebige, glatte und reellwertige Funktion, die $a(x) \geq a_0 > 0$ im Intervall $[0;1]$ erfülle, und $0 < \varepsilon < 1$ eine beliebige, aber feste reelle Zahl und sei weiters (3.7) erfüllt. Dann gilt für den Fehler in U im Verfahren 1. Ordnung für $1 \leq t \leq T$*

$$\left\| U(x_t) - \widetilde{U}_t \right\|_2 \leq C_1 \frac{E}{\varepsilon} + C_2 \varepsilon^2 \min(\varepsilon, h) + C_3 \varepsilon \min\left(\frac{\varepsilon}{h}, \frac{E}{h}, (E + E'), 1 \right), \tag{3.69}$$

und im Verfahren 2. Ordnung für $1 \leq t \leq T$

$$\begin{aligned} \left\| U(x_t) - \widetilde{U}_t \right\|_2 \quad \leq \quad & C_1 \frac{E}{\varepsilon} + C_2 \varepsilon^3 h^2 + \varepsilon C_3 \min\left(\frac{\varepsilon}{h}, \frac{E}{h}, (E + E'), 1 \right) \\ & + \varepsilon^3 C_4 \min\left(\frac{\varepsilon}{h}, \frac{E}{h}, 1 \right). \end{aligned} \tag{3.70}$$

Dabei bezeichne E wieder den Fehler (3.1), der bei der Berechnung der Phase (1.2) entsteht.

Beweis. Verwendet man die Dreiecksungleichung, erhält man

$$\begin{aligned} \left\| U(x_t) - \widetilde{U}_t \right\|_2 &= \left\| U(x_t) - U_t + U_t - \widetilde{U}_t \right\|_2 \\ &\leq \left\| U(x_t) - U_t \right\|_2 + \left\| U_t - \widetilde{U}_t \right\|_2 \\ &= \left\| U(x_t) - U_t \right\|_2 + \left\| P^{-1} e^{\frac{\imath}{\varepsilon}\widetilde{\Phi}^\varepsilon(x_t)} Z_t - P^{-1} e^{\frac{\imath}{\varepsilon}\widetilde{\Phi}^\varepsilon(x_t)} \widetilde{Z}_t \right\|_2 \\ &= \left\| U(x_t) - U_t \right\|_2 + \left\| P^{-1} e^{\frac{\imath}{\varepsilon}\widetilde{\Phi}(x_t)} \left(Z_t - \widetilde{Z}_t \right) \right\|_2 \\ &= \left\| U(x_t) - U_t \right\|_2 + \left\| Z_t - \widetilde{Z}_t \right\|_2. \end{aligned}$$

Mit (1.31) und (3.9) aus Satz 3.6 folgt die Abschätzung für das Verfahren 1. Ordnung. Mit (1.33) und (3.35) aus Satz 3.12 folgt die Abschätzung für das Verfahren 2. Ordnung. Dabei wurde jeweils der Ausdruck h^γ durch einen allgemeinen Fehler E aus (3.1) ersetzt. □

Bemerkung 3.15. Verwendet man statt der 2−Norm eine andere Vektornorm, gilt die Gleichheit der Normen von U und Z nicht mehr. Satz 3.14 lässt sich auch mit einer beliebigen Norm formulieren, man verwendet dann im Beweis statt der Eigenschaft unitärer Matrizen, dass die entsprechende Norm der Transformationmatrizen konstant ist.

Kapitel 4

Numerisches Beispiel mit $a(x) = (x + \frac{1}{2})^2$

In diesem Kapitel wird

$$a(x) = \left(x + \frac{1}{2}\right)^2 \tag{4.1}$$

gewählt, um zu zeigen, welche Auswirkungen die Verwendung einer auf Spektralmethoden basierenden Quadratur hat. Die Funktion erfüllt die Voraussetzungen $a(x) \in C^\infty [0;1]$ und $a(x) \geq a_0 > 0$ für alle x aus $[0;1]$.
Die Funktionen $\beta(x)$, $\beta_1(x)$, $\beta_2(x)$ und $\beta_3(x)$, die in den WKB-Verfahren verwendet werden, ergeben sich aus (1.3) und (1.25) und haben hier die Form

$$\beta(x) = \frac{-3}{(2x+1)^3}, \tag{4.2}$$

$$\beta_0(x) = \frac{-3}{(2x+1)^4 + 6\varepsilon^2}, \tag{4.3}$$

$$\beta_1(x) = \frac{24\,(2x+1)^6}{\left((2x+1)^4 + 6\varepsilon^2\right)^3}, \tag{4.4}$$

$$\beta_2(x) = \frac{-288\,(2x+1)^8\left((2x+1)^4 - 6\varepsilon^2\right)}{\left((2x+1)^4 + 6\varepsilon^2\right)^5}, \tag{4.5}$$

und

$$\beta_3(x) = \frac{4608\,(2x+1)^{10}\left((2x+1)^8 - 18\varepsilon^2\,(2x+1)^4 + 36\varepsilon^4\right)}{\left((2x+1)^4 + 6\varepsilon^2\right)^7}. \tag{4.6}$$

Setzt man in (1.11) $x = 0$ und verwendet die Anfangsbedingung von (1.10), erhält man die Anfangsbedingung für (1.12)

$$U_I = U(0) = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \varepsilon\sqrt{2} - \frac{\imath}{\sqrt{2}} \end{pmatrix}.$$

Mit der Matrix P aus (1.13) erhält man daraus die Anfangsbedingung für (1.17)

$$Z_I = Z(0) = \begin{pmatrix} \varepsilon \\ 1 + \imath\varepsilon \end{pmatrix}.$$

Das Phasenintegral (1.2) lässt sich für diese Wahl von $a(x)$ exakt berechnen,

$$\begin{aligned} \phi(x) &= \int_0^x \left(\sqrt{a(\tau)} - \varepsilon^2\beta(\tau)\right) \mathrm{d}\tau \\ &= \int_0^x \left(\tau + \frac{1}{2} + \frac{3\varepsilon^2}{(2\tau+1)^3}\right) \mathrm{d}\tau \\ &= \frac{\tau^2}{2} + \frac{\tau}{2} - \frac{3\varepsilon^2}{4(2\tau+1)^2}\Bigg|_0^x \\ &= \frac{x^2}{2} + \frac{x}{2} - \frac{3\varepsilon^2}{4(2x+1)^2} + \frac{3\varepsilon^2}{4}. \end{aligned} \tag{4.7}$$

Mit (4.7) wurde eine Referenzlösung Z_{ref} für die beiden Verfahren auf einem Gitter der Schrittweite $h = 10^{-7}$ berechnet, mit der die Lösungen Z_{num}, in der die Funktion $\phi(x)$ mit einer Quadratur berechnet worden ist, verglichen werden. Dazu wird die L^2-Norm auf $[0;1]$ mit

$$\begin{aligned} \|Z_{ref} - Z_{num}\|_{L^2[0;1]} &= \sqrt{\sum_{t=1}^{T} \|Z_{ref}(x_t) - Z_{num}(x_t)\|_2^2 \, h} \\ &= \sqrt{\sum_{t=1}^{T} \left((Z_{ref}(x_t) - Z_{num}(x_t))_1^2 + (Z_{ref}(x_t) - Z_{num}(x_t))_2^2\right) h} \end{aligned} \tag{4.8}$$

berechnet und auch der Fehler in U wird mit dieser Norm gemessen.

4.1 Quadraturfehler

4.1.1 mit dem Verfahren aus Abschnitt 2.3

Verwendet man zur Berechnung von (1.2) die Quadratur aus Abschnitt 2.3, lässt sich im Gegensatz zur Clenshaw-Curtis Quadratur keine a priori Fehlerabschätzung, wie in Satz 2.12, angeben, sondern es können nur a posteriori

Fehlerschranken mit (4.7) bestimmt werden.
Dazu betrachtet man (3.1),

$$E = \left\| \phi(x) - \widetilde{\phi}(x) \right\|_{\infty} = \max_{1 \le t \le T} \left| \phi(x_t) - \widetilde{\phi}(x_t) \right|,$$

wobei der Parameter ε und die Schrittweite $h = x_{t+1} - x_t$ fest bleiben, um eine Fehlerabschätzung der Form

$$E_N \le C \cdot 10^{\kappa N} = C \cdot e^{\lambda N}$$

zu erhalten. Diese Konstanten findet man in Tab. A.1. Diese Fehlerabschätzung gilt ab $N \ge 2$, da $\phi'(x)$ als Differenz eines Polynoms, $\sqrt{a(x)} = x + \frac{1}{2}$, und einer analytischen Funktion, $\varepsilon^2 \beta(x) = \frac{-3\varepsilon^2}{(2x+1)^3}$, geschrieben werden kann und das Verfahren Polynome vom Grad $N-1$ exakt integriert. Somit liefert der Polynomanteil, abgesehen von Rundungsfehlern im Bereich von 10^{-16}, keine Beiträge zum Quadraturfehler. Dies erkennt man auch daran, dass die Exponenten κ bzw. λ für festes h unabhängig von ε sind, einzig die Konstante C fällt auf Grund des Faktors ε^2 mit dem Faktor 10^{-2}, wenn ε um einen Faktor 10 abnimmt.
Wählt man den Grad N des interpolierenden Chebyshevpolynoms und ε fest, lässt sich eine Fehlerabschätzung

$$E_h \le C \cdot h^{\gamma}$$

in Abhängigkeit der Schrittweite h angeben. Die Konstanten C und γ sind für verschiedene Werte von N in Tab. A.2 angegeben.
Die Konvergenzordnung γ entspricht den Aussagen über das Verfahren, die in Abschnitt 2.3 in Bemerkung 2.18 gemacht worden sind. Für $N = 2$ ergibt sich $\gamma \approx 2$, wie man es für die Mittelpunktsregel erwartet und für $N = 4$ ist, wie bei der Simpsonregel, $\gamma \approx 4$. Für $N = 5$ ist nur mehr für $\varepsilon = 10^{-1}$ $\gamma \approx 5$ und für $N = 10$ trifft $\gamma \approx N$ nicht mehr zu, da hier die maximale Genauigkeit schon sehr schnell erreicht ist und daher die Konstanten nicht mehr sinnvoll berechnet werden können.

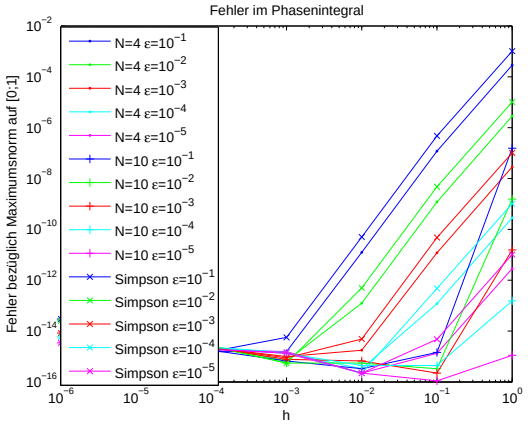

Abbildung 4.1: Fehler im Phasenintegral mit dem Verfahren aus Abschnitt 2.3 und Simpsonregel

In. Abb. 4.1 sind die Fehler für das Simpsonverfahren, $N = 4$ und $N = 10$ abgebildet. Man erkennt deutlich, dass das Konvergenzverhalten für $N = 4$ und das des Simpsonverfahrens gleich ist. Weiters sieht man, dass für $N = 10$ und $\varepsilon = 10^{-1}$ bzw. $\varepsilon = 10^{-5}$ der Fehler im ersten Schritt um h^γ mit γ aus Tab. A.2 fällt.

4.1.2 mit der Clenshaw-Curtis Quadratur

In Satz 2.12 wird eine Abschätzung des Fehlers in Abhängigkeit von ρ, der Summe der beiden Halbachsen der Ellipse $\mathcal{E}_\rho$ (vgl. (2.32)), auf der die Funktion analytisch ist, angegeben. Daher ist es das Ziel dieses Abschnittes, einen Zusammenhang zwischen der Schrittweite h und ρ herzustellen. Ähnlich wie die Funktion aus Beispiel 2.9 wird dazu der analytische Teil des Integranden vom Intervall $[x_t; x_{t+1}]$ mit (2.25) in des Intervall $[-1; 1]$ transformiert und dort in eine Chebyshevreihe entwickelt. Der Polynomanteil wird nach Satz 2.11 für $N \geq 1$ exakt integriert und deshalb im Quadraturfehler nicht weiter berücksichtigt.
Da es sich bei der Funktion $\beta(x)$ genau um die Funktion aus Beispiel 3.4 handelt, kann der Zusammenhang zwischen ρ und h aus (3.5),

$$\rho = 1 + \frac{2}{h},$$

übernommen werden. Somit lässt sich der Fehler bei der Berechnung des Phasenintegrals (1.2),

$$\phi(x) = \int_0^x \left(\tau + \frac{1}{2} + \frac{3\varepsilon^2}{(2\tau+1)^3}\right) \mathrm{d}\tau,$$

auf den Fehler der Berechnung von

$$\widetilde{\phi}_2(x) \approx \int_0^x \frac{-3}{(2\tau+1)^3} \mathrm{d}\tau$$

multipliziert mit ε^2 einschränken.
Analog zu den Fehlerkonstanten aus Abschnitt 4.1.1 lassen sich auch für die aufsummierte Clenshaw-Curtis Quadratur a posteriori Abschätzungen

$$E_N \leq C \cdot 10^{\kappa N} = C \cdot e^{\lambda N}$$

bei festem h und ε bzw.

$$E_h \leq C \cdot h^{\gamma}$$

bei festem N und ε mit Hilfe von (4.7) angeben. Die Werte der Konstanten von E_N findet man für Schrittweiten $h = 10^0$ und $h = 10^{-1}$ in Tab. A.3. Für kleinere Schrittweiten lassen sich die Konstanten nicht mehr sinnvoll berechnen, da hier E_N auf Größenordnung der Maschinengenauigkeit gefallen ist. Die Konstanten für E_h sind in Tab. A.4 angegeben. Man erkennt die in Bsp. 3.4 erwähnte Fehlerordnung $\mathcal{O}(h^{N+2})$ und der Fehler kann mit

$$E \leq C \cdot h^{N+2} \tag{4.9}$$

abgeschätz werden. Da jedoch, wenn ε kleiner und zusätzlich N größer wird, der Fehler sehr schnell im Bereich der Maschinengenauigkeit ist, können die Konstanten nur mehr eingeschränkt berechnet werden und die berechnete Fehlerordnung stimmt nicht mehr mit der theoretischen Fehlerordnung überein.

4.1.3 mit baryzentrischer Interpolation

Der erste Teil des Phasenintegrals (1.2),

$$\phi(x) = \int_0^x \left(\tau + \frac{1}{2} + \frac{3\varepsilon^2}{(2\tau+1)^3}\right) \mathrm{d}\tau,$$

ist schon in Beispiel 2.10 berechnet worden und dort ist schon gezeigt worden, dass er für $N \geq 2$ exakt berechnet werden kann. Somit liefert auch hier der Polynomanteil des Integranden keinen Beitrag zum Fehler. Der Fehler ergibt sich durch die Berechnung von

$$\widetilde{\phi}_2(x) \approx \int_0^x \frac{-3}{(2\tau+1)^3} \mathrm{d}\tau.$$

Diese Funktion ist schon in Beispiel 3.1 betrachtet worden und jetzt soll zusätzlich eine Abschätzung des Fehlers angegeben werden.
Mit (3.2) erhält man eine, wie in Abschnitt 3.1.1 gezeigt, von h unabhängige Fehlerabschätzung, die durch die in das Intervall $[-1; -1]$ transformierte Funktion bestimmt ist. Da die mit (2.25) erhaltene Transformierte von $\beta(x)$,

$$\beta_T(u) = \frac{-3}{(u+2)^3},$$

bei $u = -2$ einen Pol dritter Ordnung hat, hat die Stammfunktion an dieser Stelle einen Pol zweiter Ordnung. Daher kann man die Summe der Halbachsen

$$\rho < 2 + \sqrt{3}$$

bestimmen, sodass die Funktion innerhalb und auf der Ellipse $\mathcal{E}_\rho$ analytisch ist. In den Beispielen aus [BT, Section 6] wird für die Abschätzungen des Interpolationsfehlers der Wert für ρ verwendet, bei dem der Pol auf $\mathcal{E}_\rho$ liegt. Das ergibt für diese Art der Phasenberechnung einen Fehler

$$E \leq C \left(2 + \sqrt{3}\right)^{-N} \approx C \cdot 3.73^{-N}. \tag{4.10}$$

Bestimmt man ρ und C experimentell, erhält man $\rho \approx 3.5$ und $C \approx 3$.

4.2 Verfahren 1. Ordnung

Mit den Abschätzungen aus Abschnitt 4.1 lassen sich die Abschätzungen (3.8) aus Satz 3.5 für Z und (3.69) aus Satz 3.14 für U für das konkrete Beispiel mit

$$\phi(x) = \int_0^x \left(\tau + \frac{1}{2} + \frac{3\varepsilon^2}{(\tau+1)^3}\right) \mathrm{d}\tau \tag{4.11}$$

präzisieren. Zur Vereinfachung wird das Integral in zwei Teile,

$$\widetilde{\phi}_1(x) \approx \int_0^x \left(\tau + \frac{1}{2}\right) \mathrm{d}\tau \tag{4.12}$$

und

$$\widetilde{\phi}_2(x) \approx \int_0^x \frac{-3}{(2\tau+1)^3} \mathrm{d}\tau, \tag{4.13}$$

geteilt, die getrennt voneinander berechnet werden.

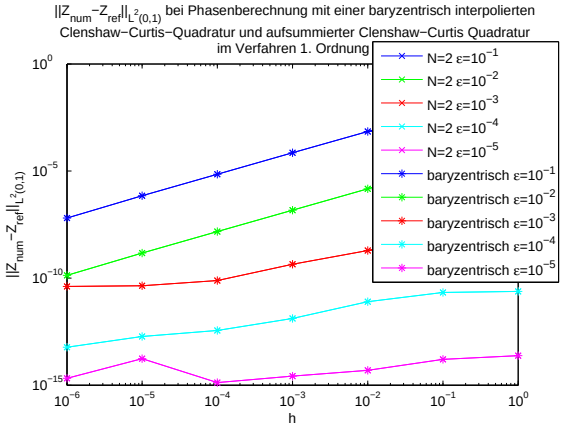

Abbildung 4.2: Fehler bei der Berechnung von Z mit dem WKB-Verfahren 1. Ordnung

In Abb. 4.2 ist der Fehler in Z mit der in (4.8) definierten Norm abgebildet. Für die baryzentrische Interpolation der Phase wurde der Integrand von (4.12) bis $N = 2$ und der Integrand von (4.13) bis $N = 34$ entwickelt (siehe dazu Beispiel 3.1). Man erkennt, dass hier die Art der Phasenberechnung keinen Einfluss auf den Fehler im Verfahren hat, da nur (4.13) einen Beitrag liefert und dieser noch mit ε^2 multipliziert wird.

Für diese Wahl der Funktion $a(x)$ lässt sich mit Satz 3.5 eine Fehlerabschätzung formulieren.

Satz 4.1. *Es sei* $a(x) = \left(x + \frac{1}{2}\right)^2$ *und* $0 < \varepsilon < 1$ *eine beliebige, aber feste reelle Zahl. Dann gilt für den Fehler im Verfahren 1. Ordnung (1.20),*

1.

$$\left\| Z\left(x_t\right) - \widetilde{Z}_t \right\| \leq C_1 \varepsilon^2 \min\left(\varepsilon, h\right) + C_2 \varepsilon^3 h^{N+2}, \ 1 \leq t \leq T, \tag{4.14}$$

wenn (4.11) mit einer aufsummierten Clenshaw-Curtis Quadratur, die bis zur Ordnung N entwickelt worden ist, berechnet wird;

2.

$$\left\| Z\left(x_t\right) - \widetilde{Z}_t \right\| \leq C_1 \varepsilon^2 \min\left(\varepsilon, h\right) + C_2 \varepsilon^3 \left(2 + \sqrt{3}\right)^{-N}, \ 1 \leq t \leq T, \tag{4.15}$$

wenn (4.11) mit einer baryzentrisch interpolierten Clenshaw-Curtis Stammfunktion, die bis zur Ordnung N entwickelt worden ist, berechnet wird.

Beweis. 1. Für die Berechnung mit aufsummierter Clenshaw-Curtis Quadratur entnimmt man Tab. A.4, dass die größte Konstante, um den Faktor ε^2 bereinigt, $C = 0.45$ ist. Somit ist, wenn man E noch mit ε^2 multipliziert, da nur der Fehler zwischen $\phi_2(x)$ und $\widetilde{\phi}_2(x)$ einen Beitrag liefert,

$$\min\left(\frac{\varepsilon}{h}, C\varepsilon^2 h^{N+1}, C\varepsilon^2 h^{N+2}, 1\right) \leq C\varepsilon^2 h^{N+2}$$

und man erhält (4.14) aus (3.8).

2. Für die Berechnung mit baryzentrisch interpolierter Stammfunktion ist $E = \left(2+\sqrt{3}\right)^{-N} \approx 3 \cdot 3.73^{-N}$ nach (4.10) aus Abschnitt 4.1.3, multipliziert mit ε^2 ergibt

$$\min\left(\frac{\varepsilon}{h}, \frac{3\varepsilon^2 3.73^{-N}}{h}, 3\varepsilon^2 3.73^{-N}, 1\right) \leq 3\varepsilon^2 3.73^{-N}$$

und man erhält (4.15) aus (3.8).

Nach Bemerkung 3.10 Punkt 3 kann auf (3.7a) verzichtet werden und somit ist hier der Fehler E' nicht berücksichtigt worden. □

Bemerkung 4.2. Wie schon in Bemerkung 3.10 Punkt 1 erwähnt, muss der Fehler, der durch die Berechnung von (4.11) entsteht, nicht gegen Null gehen. Entwickelt man bei der Berechnung mit baryzentrischer Interpolation weit genug, sodass die Koeffizienten der Chebyshevreihe im Bereich der Maschinengenauigkeit liegen, für das konkrete Beispiel bis $N = 34$, ist

$$E \approx 3 \cdot 3.73^{-34} \approx 10^{-19}.$$

Somit ist der zusätzliche Fehler im Vergleich zu $\varepsilon^2 \min(\varepsilon, h)$ vernachlässigbar klein.

Ist N nicht groß genug gewählt, gibt es eine gewisse Schrittweite h_0, sodass für Schrittweiten $h \leq h_0$

$$\varepsilon^2 \min(\varepsilon, h) \leq \varepsilon^3 \left(2+\sqrt{3}\right)^{-N}$$

gilt und der Fehler somit konstant bleibt.

Berechnet man aus Z mit (1.29) U, wird der zusätzliche Fehlerterm in der Fehlerabschätzung von U je nach gewähltem Verfahren zu

$$\frac{E}{\varepsilon} = \varepsilon C h^{N+2},$$

wenn eine aufsummierte Clenshaw-Curtis Quadratur verwendet wird und

$$\frac{E}{\varepsilon} = \varepsilon C \left(2+\sqrt{3}\right)^{-N},$$

wenn die genäherte Stammfunktion baryzentrisch interpoliert wird. Damit erhält man

Satz 4.3. *Unter den Voraussetzungen von Satz 4.1 gilt für den Fehler in U im Verfahren 1. Ordnung für* $1 \leq t \leq T$,

1.

$$\left\| U\left(x_t\right) - \widetilde{U}_t \right\| \leq C_1 \varepsilon h^{N+2} + C_2 \varepsilon^2 \min\left(\varepsilon, h\right) + C_3 \varepsilon^3 h^{N+2}, \tag{4.16}$$

wenn (4.11) mit einer aufsummierten Clenshaw-Curtis Quadratur, die bis zur Ordnung N entwickelt worden ist, berechnet wird;

2.

$$\left\| U\left(x_t\right) - \widetilde{U}_t \right\| \leq C_1 \varepsilon \left(2+\sqrt{3}\right)^{-N} + C_2 \varepsilon^2 \min\left(\varepsilon, h\right) + C_3 \varepsilon^3 \left(2+\sqrt{3}\right)^{-N}, \tag{4.17}$$

wenn (4.11) mit einer baryzentrisch interpolierten Clenshaw-Curtis Stammfunktion, die bis zur Ordnung N entwickelt worden ist, berechnet wird.

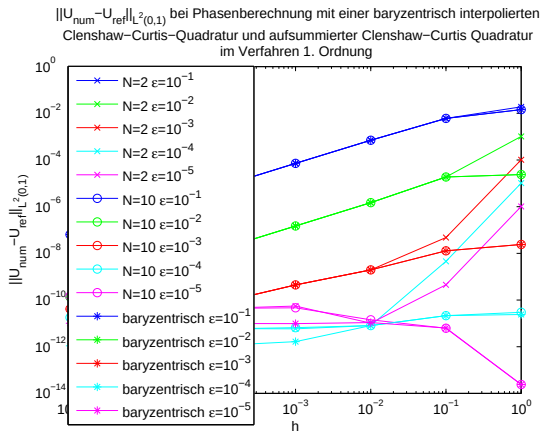

Abbildung 4.3: Fehler bei der Berechnung von U mit dem WKB-Verfahren 1. Ordnung

In Abb. 4.3 ist der Fehler in U für eine baryzentrisch interpolierte Phase mit $N = 34$ und zwei mit summierter Clenshaw-Curtis Quadratur ($N = 2$ und $N = 10$) berechneten Phasen abgebildet. Man sieht, dass bei der summierten Clenshaw-Curtis Quadratur bei kleinen Schrittweiten die Wahl von N Auswirkung auf den Fehler hat. Einzig wenn $\varepsilon = 10^{-1}$ gewählt wird, hat die Wahl von N fast keinen Einfluss. Ob die Phase mit einer summierten Clenshaw-Curtis Quadratur mit $N = 10$ oder mit baryzentrischer Interpolation mit $N = 34$ berechnet wird, hat nahezu keinen Einfluss auf den Fehler. Durch numerische Effekte wie Rundungs- und Auslöschungsfehler steigt bei kleinem ε der Fehler wieder an, da hier $\varepsilon^2 \min\left(\varepsilon, h\right)$ sehr schnell im Bereich der Maschinengenauigkeit ist.

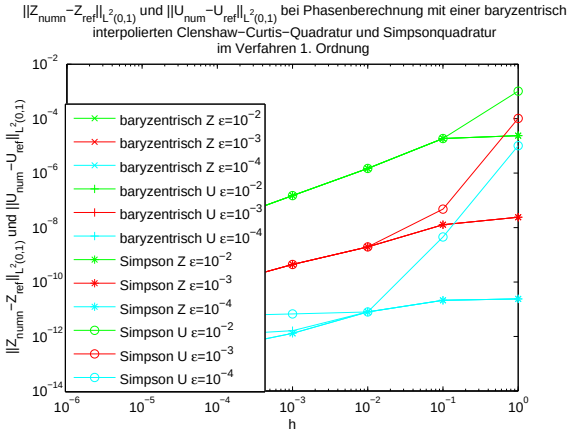

Abbildung 4.4: Vergleich von Spektralmethoden mit Simpsonverfahren im Verfahren 1. Ordnung

Um die Verbesserung des Fehlers bei der Verwendung von Spektralmethoden im Gegensatz zu anderen Quadraturverfahren wie z. B. einer Simpsonquadratur deutlich zu machen, ist in Abb. 4.4 sowohl der Fehler bei der Berechnung von Z als auch von U abgebildet.
Wie schon bei der Berechnung von Z erwähnt, hat die Wahl des Quadraturverfahrens keinen Einfluss auf den Fehler in Z. Der Fehler in der Berechnung von U wird jedoch bei der Verwendung eines auf Spektralmethoden basierenden Quadraturverfahrens verbessert, wenn bei großer Schrittweite h der Parameter ε klein ist. In Abb. 4.3 ist ersichtlich,dass für $\varepsilon = 10^{-1}$ die aufsummierte Clenshaw-Curtis Quadratur, die mit $N = 2$ von derselben Fehlerordnung wie das Simpsonverfahren ist, denselben Fehler wie die Variante mit baryzentrisch interpolierter Phase liefert. Wird ε kleiner, verbessert die Verwendung von Spektralmethoden den Verfahrensfehler bei der Berechnung von U.

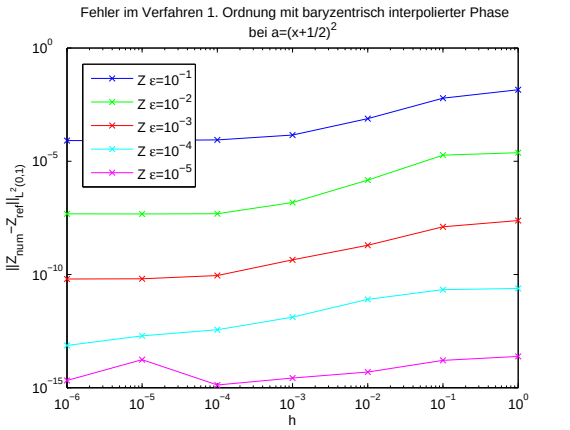

Abbildung 4.5: Beispiel für konstanten Fehler

Abb. 4.5 zeigt, das der Fehler konstant bleibt, wenn N bei der baryzentrischen Interpolation zu klein gewählt wird, sodass die Koeffizienten der Reihenentwicklung des Integranden im Vergleich zur Maschinengenauigkeit nicht vernachlässigbar klein werden.
Im konkreten Fall wurde $N = 2$ gewählt, sodass nur $\phi_1(x)$ (4.12) exakt berechnet wird. Für $\varepsilon = 10^{-1}$ ergibt das für den Fehlerterm der Phasenberechnung $3\varepsilon^3 3.73^{-2} \approx 2.16 \cdot 10^{-4}$, für $\varepsilon = 10^{-2}$ den Wert $3\varepsilon^3 3.73^{-2} \approx 2.16 \cdot 10^{-7}$ und für $\varepsilon = 10^{-3}$ erhält man $3\varepsilon^3 3.73^{-2} \approx 2.16 \cdot 10^{-10}$. Diese Werte sind auch in Abb. 4.5 ersichtlich. Die in Bemerkung 4.2 erwähn-

te Schrittweite h_0 liegt für $\varepsilon = 10^{-1}$ bei $h_0 = 10^{-3}$. Bei $\varepsilon = 10^{-5}$ ist nicht mehr zu erkennen, dass der Fehler konstant ist, da hier der zusätzliche Fehler, ebenso wie $\varepsilon^2 \min(\varepsilon, h)$ schon im Bereich der Maschinengenauigkeit ist.

4.3 Verfahren 2. Ordnung

Analog zu den Fehlerabschätzungen für das WKB-Verfahren 1. Ordnung lassen sich mit (4.9) und (4.10) Abschätzungen für das WKB-Verfahren angeben, wenn $a(x) = \left(x + \frac{1}{2}\right)^2$ gewählt wird. Man erhält

Satz 4.4. *Es sei $a(x) = \left(x + \frac{1}{2}\right)^2$ und $0 < \varepsilon < 1$ eine beliebige, aber feste reelle Zahl. Dann gilt für den Fehler im Verfahren 2. Ordnung (1.26) für $1 \le t \le T$,*

1.
$$\left\| Z(x_t) - \widetilde{Z}_t \right\| \le C_1 \varepsilon^3 h^2 + C_2 \varepsilon^3 h^{N+2} + C_3 \varepsilon^5 h^{N+1}, \tag{4.18}$$
wenn (4.11) mit einer aufsummierten Clenshaw-Curtis Quadratur, die bis zur Ordnung N entwickelt worden ist, berechnet wird;

2.
$$\begin{aligned} \left\| Z(x_t) - \widetilde{Z}_t \right\| \;\le\; & C_1 \varepsilon^3 h^2 + C_2 \varepsilon^3 \left(2 + \sqrt{3}\right)^{-N} \\ & + C_3 \varepsilon^3 \min\left(\tfrac{\varepsilon}{h}, \tfrac{3\varepsilon^2 \left(2+\sqrt{3}\right)^{-N}}{h}, 1 \right), \end{aligned} \tag{4.19}$$
wenn (4.11) mit einer baryzentrisch interpolierten Clenshaw-Curtis Stammfunktion, die bis zur Ordnung N entwickelt worden ist, berechnet wird.

Beweis. Der erste Teil des zusätzlichen Fehlers entspricht für beide Arten der Phasenberechnung dem zusätzlichen Fehler des Verfahrens 1. Ordnung (Satz 4.1). Für den Fehler, der durch A_t^2 entsteht, ergibt sich

1. bei aufsummierter Clenshaw-Curtis Quadratur mit (4.9) und den Konstanten aus Tab. A.4
$$\min\left(\frac{\varepsilon}{h}, \frac{\varepsilon^2 E}{h}, 1 \right) \le \varepsilon^2 h^{N+1}$$
und man erhält damit (4.18);

2. bei baryzentrischer Interpolation (4.19) mit (4.10).

□

Bemerkung 4.5. Wenn N groß genug gewählt wird, ist

$$\min\left(\frac{\varepsilon}{h}, \frac{3\varepsilon^2\left(2+\sqrt{3}\right)^{-N}}{h}, 1\right) = \frac{3\varepsilon^2\left(2+\sqrt{3}\right)^{-N}}{h}.$$

Ist $N = 34$, wie in den folgenden Plots, hat man für den zusätzlichen Fehlerterm

$$\min\left(\frac{\varepsilon}{h}, \frac{3\varepsilon^2\left(2+\sqrt{3}\right)^{-34}}{h}, 1\right) \approx \frac{\varepsilon^2 \cdot 10^{-19}}{h}.$$

Ist $\varepsilon = 10^{-1}$ und $h = 10^{-6}$, ist dieser Wert immer noch im Bereich der Maschinengenauigkeit und somit vernachlässigbar klein.

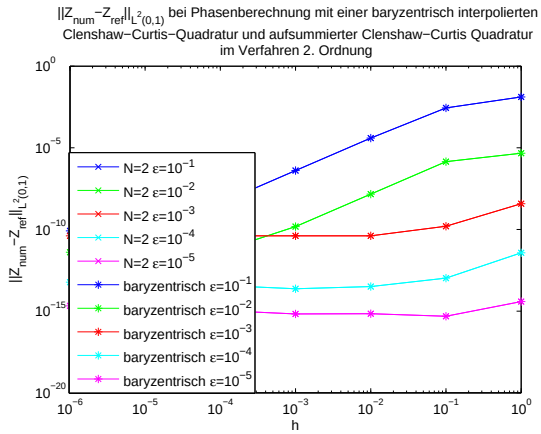

Abbildung 4.6: Fehler bei der Berechnung von Z mit dem WKB-Verfahren 2. Ordnung

Dass, sowie im Verfahren 1. Ordnung, auch im Verfahren 2. Ordnung die Wahl der Quadratur keinen Einfluss auf den Verfahrensfehler bei der Berechnung von Z hat, ist in Abb. 4.6 ersichtlich. Auch hier wurde für die baryzentrische Interpolation (4.12) bis $N = 2$ und (4.13) bis $N = 34$ entwickelt. Wählt man $\varepsilon = 10^{-1}$ oder $\varepsilon = 10^{-2}$, erkennt man die Fehlerordnung h^2 des Verfahrens. Wird ε kleiner gewählt, haben numerische Effekte Einfluss auf den Fehler, da die Grenze der Genauigkeit in MATLAB durch den Faktor ε^3 sehr schnell erreicht ist.

Auch im Verfahren 2. Ordnung lässt sich eine Fehlerabschätzung für den Fehler in U für diese Wahl der Funktion $a(x)$ formulieren. Wenn man in dem zusätzlichen Fehlerterm $\frac{E}{\varepsilon}$ den Quadraturfehler durch (4.9) bzw. (4.10) ersetzt und wieder berücksichtigt, dass nur (4.13) einen Beitrag liefert und daher wieder mit ε^2 multipliziert werden muss, erhält man

Satz 4.6. *Unter den Voraussetzungen von Satz 4.4 gilt für den Fehler in U im Verfahren 2. Ordnung für $1 \le t \le T$,*

1.

$$\left\| U(x_t) - \widetilde{U}_t \right\| \le C_1\varepsilon h^{N+2} + C_2\varepsilon^3 h^2 + C_3\varepsilon^3 h^{N+2} + C_4\varepsilon^5 h^{N+1}, \quad (4.20)$$

wenn (4.11) mit einer aufsummierten Clenshaw-Curtis Quadratur, die bis zur Ordnung N entwickelt worden ist, berechnet wird;

2.

$$\left\|U(x_t) - \widetilde{U}_t\right\| \leq C_1\varepsilon\left(2+\sqrt{3}\right)^{-N} + C_3\varepsilon^3\left(2+\sqrt{3}\right)^{-N} + C_2\varepsilon^3 h^2 + C_4\varepsilon^3 \min\left(\frac{\varepsilon}{h}, \frac{3\varepsilon^2\left(2+\sqrt{3}\right)^{-N}}{h}, 1\right), \tag{4.21}$$

wenn (4.11) mit einer baryzentrisch interpolierten Clenshaw-Curtis Stammfunktion, die bis zur Ordnung N entwickelt worden ist, berechnet wird.

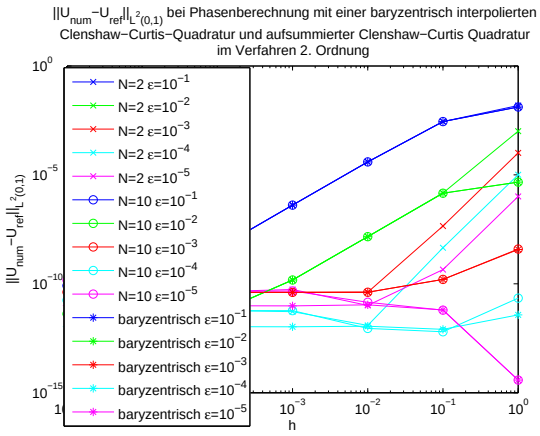

Abbildung 4.7: Fehler bei der Berechnung von U mit dem WKB-Verfahren 2. Ordnung

Abb. 4.7 zeigt den Fehler in U, wenn die Phase (4.11) mit einer aufsummierten Clenshaw-Curtis Quadratur mit $N = 2$ und $N = 10$ und mit einer baryzentrischen Interpolation, mit einer Entwicklung von (4.12) bis $N = 2$ und (4.13) bis $N = 34$, berechnet wird. Für kleine Werte des Parameters ε ist ein deutlicher Unterschied zwischen den Varianten der Phasenberechnung erkennbar. Besonders bei $\varepsilon = 10^{-3}$ und $\varepsilon = 10^{-4}$ zeigt sich, dass die Methode mit baryzentrischer Interpolation, die fast keinen Unterschied zur summierten Variante mit $N = 10$ aufweist, der summierten Quadratur mit $N = 2$ überlegen ist.

Vergleicht man die Phasenberechnung mit baryzentrischer Interpolation mit der Phasenberechnung mit dem Simpsonverfahren, erkennt man in Abb. 4.8, dass bei der baryzentrischen Interpolation im Vergleich zum Simpsonverfahren der zusätzliche Fehlerterm in der Berechnung von U keine Auswirkung auf den Gesamtfehler hat, da $\left(2+\sqrt{3}\right)^{-34} \approx 3.5 \cdot 10^{-20}$ ist und somit unterhalb der Genauigkeit von MATLAB liegt.

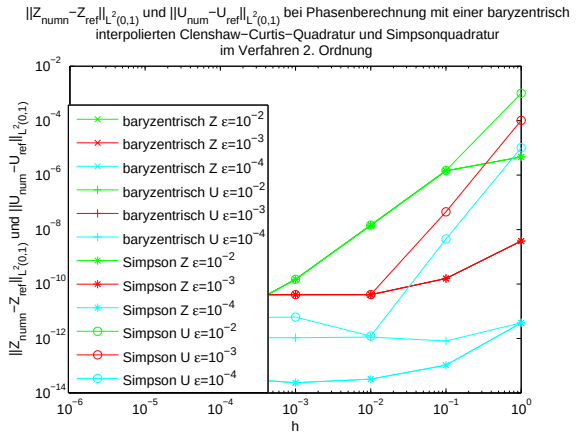

Abbildung 4.8: Vergleich von Spektralmethoden mit Simpsonverfahren im Verfahren 2. Ordnung

Dies ist besonders deutlich bei $\varepsilon = 10^{-3}$ zu erkennen. Hier sieht man den zusätzlichen Fehlerterm bei der Berechnung von U, wenn das Simpsonverfahren verwendet worden ist. Bei der Verwendung der baryzentrischen Interpolation haben Z und U denselben Fehler. Wird ε zu klein gewählt, führen Rundungs- und Auslöschungsfehler dazu, dass der Gesamtfehler auf Grund der begrenzten Genauigkeit wieder ansteigt, wie hier bei $\varepsilon = 10^{-4}$ zu erkennen ist.

Kapitel 5

Numerisches Beispiel mit $a(x) = e^{-x^2}$

In diesem Kapitel wird

$$a(x) = e^{-x^2} \tag{5.1}$$

gewählt. Diese Funktion erfüllt die Voraussetzungen $a(x) \in C^\infty[0;1]$ und $a(x) \geq a_0 > 0$ für alle x aus $[0;1]$.
Die Funktionen $\beta(x)$, $\beta_1(x)$, $\beta_2(x)$ und $\beta_3(x)$, die verwendet werden und in (1.3) und (1.25) definiert sind, haben hier die Form

$$\beta(x) = \frac{-\left(2+x^2\right) e^{\frac{x^2}{2}}}{8}, \tag{5.2}$$

$$\beta_0(x) = \frac{-\left(2+x^2\right)}{2\left(8e^{-x^2}+2\varepsilon^2+\varepsilon^2x^2\right)}, \tag{5.3}$$

$$\beta_1(x) = \frac{-32x\left(3+x^2\right) e^{-\frac{3x^2}{2}}}{\left(8e^{-x^2}+2\varepsilon^2+\varepsilon^2x^2\right)^3}, \tag{5.4}$$

$$\begin{aligned}\beta_2(x) = {} & \frac{384e^{-2x^2}\left(-8e^{-x^2}+9\varepsilon^2x^2-2\varepsilon^2-32x^2e^{-x^2}\right)}{\left(8e^{-x^2}+2\varepsilon^2+\varepsilon^2x^2\right)^4} \\ & + \frac{384e^{-2x^2}\left(6x^4\varepsilon^2-8x^4e^{-x^2}+x^6\varepsilon^2\right)}{\left(8e^{-x^2}+2\varepsilon^2+\varepsilon^2x^2\right)^5}\end{aligned} \tag{5.5}$$

sowie

$$\begin{aligned}\beta_3(x) &= \frac{-6144xe^{-\frac{5x^2}{2}}\left(x^8\varepsilon^4 - 24e^{-x^2}\varepsilon^2 + 27\varepsilon^4x^4 + 22\varepsilon^4x^2\right)}{\left(8e^{-x^2} + 2\varepsilon^2 + \varepsilon^2x^2\right)^7} \\ &\quad - \frac{6144xe^{-\frac{5x^2}{2}}\left(320x^2e^{-2x^2} + 9x^6\varepsilon^4 + 64x^4e^{-2x^2}\right)}{\left(8e^{-x^2} + 2\varepsilon^2 + \varepsilon^2x^2\right)^7} \\ &\quad - \frac{6144xe^{-\frac{5x^2}{2}}\left(-18\varepsilon^4 - 312e^{-x^2}\varepsilon^2x^2 - 24x^6e^{-x^2}\varepsilon^2\right)}{\left(8e^{-x^2} + 2\varepsilon^2 + \varepsilon^2x^2\right)^7} \\ &\quad - \frac{6144xe^{-\frac{5x^2}{2}}\left(-168x^4e^{-x^2}\varepsilon^2 + 192e^{-2x^2}\right)}{\left(8e^{-x^2} + 2\varepsilon^2 + \varepsilon^2x^2\right)^7}. \end{aligned} \tag{5.6}$$

Die Anfangsbedingungen für diese Wahl von $a(x)$ sind für (1.12)

$$U_I = U(0) = \begin{pmatrix} 1 \\ -\imath \end{pmatrix}$$

und

$$Z_I = Z(0) = \begin{pmatrix} 0 \\ \sqrt{2} \end{pmatrix}$$

für (1.17).
Das Phasenintegral (1.2) hat hier die Form

$$\phi(x) = \int_0^x \left(e^{-\frac{\tau^2}{2}} + \frac{\varepsilon^2\left(2+\tau^2\right)e^{\frac{\tau^2}{2}}}{8}\right)\mathrm{d}\tau \tag{5.7}$$

und besitzt keine in geschlossener Form darstellbare Stammfunktion.
Die Referenzlösung Z_{ref}, mit der die numerischen Lösungen Z_{num} verglichen werden, ist auf einem Gitter mit Schrittweite $h = 10^{-7}$ berechnet worden. Die Phase in der Referenzlösung Z_{ref} ist mit (2.31) berechnet worden. Dabei ist der erste Teil des Integrals

$$\widetilde{\phi}_1(x) \approx \int_0^x e^{-\frac{\tau^2}{2}}\mathrm{d}\tau \tag{5.8}$$

bis $N = 17$ und der zweite Teil

$$\widetilde{\phi}_2(x) \approx \int_0^x \frac{\left(2+\tau^2\right)e^{\frac{\tau^2}{2}}}{8}\mathrm{d}\tau \tag{5.9}$$

bis $N = 18$ entwickelt worden. Anschließend ist die auf diese Weise erhaltene Näherung der Stammfunktion auf den Gitterpunkten baryzentrisch interpoliert worden, um die Werte $\phi\left(x_t\right)$ zu erhalten.

5.1 Quadraturfehler

5.1.1 mit dem Verfahren aus Abschnitt 2.3

Wie schon in Abschnitt 2.3 erwähnt, gibt es für diese Art der Quadratur keine Fehlerabschätzungen. Da (5.7) auch keine Stammfunktion in geschlossener Form hat, wird hier zur Berechnung der Konstanten einer a posteriori Abschätzung

$$E_h \leq C \cdot h^{\gamma}$$

die genäherte Phase mit der Phase der Referenzlösung verglichen. Die Konstanten sind für verschiedene Werte von N in Tab. A.5 aufgelistet. Man erkennt an dem Wert der Konstanten C, dass $\int_0^x \sqrt{a(\tau)}\mathrm{d}\tau$ im Gegensatz zu Kapitel 4 nicht mehr exakt integriert wird, da der Wert von ε keine Auswirkungen hat. Die Ordnung γ entspricht wieder dem Wert N, wobei bei großem N die maximale Genauigkeit sehr schnell erreicht ist und die experimentell bestimmte Konstante γ nicht mehr mit N übereinstimmt.

5.1.2 mit der Clenshaw-Curtis Quadratur

Berechnet man das Phasenintegral (5.7) in einem Intervall $[x_t; x_{t+1}]$, werden die Integranden von (5.8) und (5.9) mit (2.25) in das Intervall $[-1; 1]$ transformiert und man erhält

$$\sqrt{a_T(u)} = e^{-\frac{1}{8}((2t-1)h+uh)^2}$$

und

$$\beta_T(u) = -\frac{(8 + (2t-1)\,h + uh)}{32} e^{\frac{1}{2}((2t-1)h+uh)^2}.$$

Bei diesen beiden Funktionen handelt es sich um ganze Funktionen. Für ganze Funktionen sind in der Literatur keine Fehlerabschätzungen gefunden worden. Vergleicht man jedoch das Konvergenzverhalten der Clenshaw-Curtis Quadratur von auf $\mathcal{E}_\rho$ analytischen Funktionen und ganzen Funktionen, erkennt man, dass sie sich ähnlich verhalten. Siehe dazu auch die Beispiele $f(x) = e^x$ und $f(x) = e^{-x^2}$ in [Tr2, Fig. 2].
Dies legt die Vermutung nahe, dass sich der Fehler bei der Summation der Teilintegrale ebenfalls mit

$$E \leq C \cdot h^{N+2} \tag{5.10}$$

angeben lässt.
Abb. 5.1 mit $\varepsilon = 10^{-5}$ in (5.7) und die Werte in Tab. A.6 bestätigen diese Vermutung zumindest für $N = 2$, $N = 4$ und $N = 6$. Für $N = 3$ und $N = 5$ ergibt sich nur h^{N+1}. Für $N = 10$ ist der Fehler schon im Bereich der

Maschinengenauigkeit und somit können die Konstanten nicht mehr sinnvoll berechnet werden.

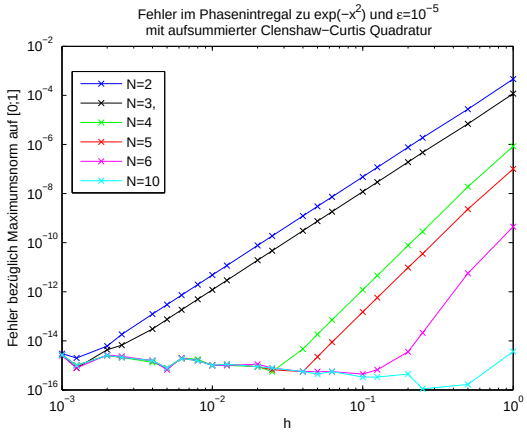

Abbildung 5.1: Fehler im Phasenintegral mit aufsummierter Clenshaw-Curtis Quadratur

Daraus lässt sich schließen, dass (5.10) nur für gerade Werte von N eine Abschätzung des Quadraturfehlers ist. Ist N ungerade, hat man nur mehr $E \leq C \cdot h^{N+1}$ als Abschätzung. Weiters erkennt man an der Konstanten C aus Tab. A.6, dass die Wahl des Parameters ε keinen Einfluss auf den Fehler bei der Berechnung des Phasenintegrals (5.7) hat.

5.1.3 mit baryzentrischer Interpolation

Wird das Phasenintegral (5.7) im Intervall $[0;1]$ mit (2.31) an den Chebyshevpunkten $\widetilde{x}_j$, $0 \leq j \leq N$ berechnet und anschließend mit (2.55) an den Gitterpunkten x_t, $1 \leq t \leq T$ berechnet, kann der Fehler mit (3.2) abgeschätzt werden, da das Problem der baryzentrischen Interpolation ganzer Funktionen in [WTG] behandelt wird und dort ρ so gewählt wird, dass

$$\rho^N = \texttt{eps}^{-1} \tag{5.11}$$

ist, wobei N der Grad der Entwicklung in die Chebyshevreihe ist, sodass die Funktion bis zur Maschinengenauigkeit `eps` interpoliert werden kann.
Die Entwicklung des Integranden von (5.8) in eine Chebyshevreihe (2.20) hat ab $N \geq 17$ vernachlässigbar kleine Koeffizienten, somit sind auch die Koeffizienten in der Clenshaw-Curtis Stammfunktion (2.31) vernachlässigbar klein. Der Integrand kann auch für beliebige Gitterpunkte durch baryzentrische Interpolation der Funktionswerte an den zu dieser Reihe gehörigen Chebyshevpunkten bis auf Maschinengenauigkeit approximiert werden. Die Koeffizienten der Entwicklung von (5.9) sind ab $N \geq 18$ vernachlässigbar klein. Wählt man $N = 18$ für beide Teile von (5.7), ergibt das einen Wert $\rho \approx 7.4$, der in der Abschätzung (3.2) verwendet werden kann.

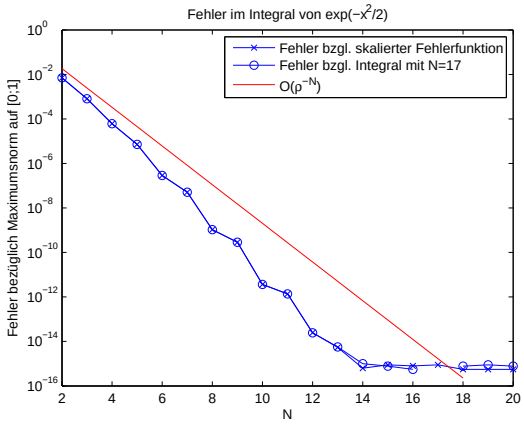

Abbildung 5.2: Fehler in der baryzentrisch interpolierten Stammfunktion von $e^{-\frac{x^2}{2}}$

Durch Skalierung der in MATLAB implementierten Fehlerfunktion `erf(x)`[1] erhält man (5.8). In Abb. 5.2 ist der Fehler der baryzentrisch interpolierten Stammfunktion von (5.8) bezüglich $\sqrt{\frac{\pi}{2}}\texttt{erf}\left(\frac{\texttt{x}}{\sqrt{2}}\right)$ (blau x) und bezüglich einer Lösung mit $N = 17$ (blau o) gezeigt. Weiters ist zum Vergleich ρ^{-N} mit $\rho = 7.4$ abgebildet (rot). Bei der Berechnung von (5.7) haben der Parameter ε und die Schrittweite h wieder keinen Einfluss auf den Fehler, der entsteht, und somit erhält man für diese Art der Phasenberechnung

$$E \leq C \cdot 7.4^{-N} \tag{5.12}$$

als Abschätzung für den Quadraturfehler.

5.2 Verfahren 1. Ordnung

Mit den Abschätzungen für den Quadraturfehler aus Abschnitt 5.1 lassen sich die Fehlerabschätzungen für die numerischen Lösungen an den Gitterpunkten x_t, $1 \leq t \leq T$, von (1.17) bzw. (1.12) aus Satz 3.5 und Satz 3.14 für die konkrete Wahl der Funktion $a(x) = e^{-x^2}$ angeben.
Setzt man die Abschätzungen (5.10) bzw. (5.12) in (3.8) ein, erhält man

Satz 5.1. *Es sei $a(x) = e^{-x^2}$ und $0 < \varepsilon < 1$ eine beliebige, aber feste reelle Zahl. Dann gilt für den Fehler im Verfahren 1. Ordnung (1.20) für $1 \leq t \leq T$,*

1. $$\left\| Z(x_t) - \widetilde{Z}_t \right\| \leq C_1 \varepsilon^2 \min(\varepsilon, h) + C_2 \varepsilon h^{N+2}, \tag{5.13}$$

   wenn N gerade ist und (5.7) mit einer aufsummierten Clenshaw-Curtis Quadratur, die bis zur Ordnung N entwickelt worden ist, berechnet wird;

2. $$\left\| Z(x_t) - \widetilde{Z}_t \right\| \leq C_1 \varepsilon^2 \min(\varepsilon, h) + C_2 \varepsilon \cdot 7.4^{-N}, \tag{5.14}$$

[1] $\texttt{erf(x)} := \frac{2}{\sqrt{\pi}} \int_0^x e^{-\tau^2} \mathrm{d}\tau$

wenn (5.7) mit einer baryzentrisch interpolierten Clenshaw-Curtis Stammfunktion, die bis zur Ordnung N entwickelt worden ist, berechnet wird.

Beweis. Analog zu Beweis von 4.1 erhält man mit Tab. A.6 die Abschätzung (5.13). Da 7.4^{-N} immer kleiner ist als die anderen im Minimum auftretenden Terme erhält man (5.14). □

Bemerkung 5.2. Verwendet man in der aufsummierten Clenshaw-Curtis Quadratur ungerades N für die Entwicklung, verschlechtert sich die Fehlerabschätzung und der Exponent von h ist nur mehr $N+1$ statt $N+2$.

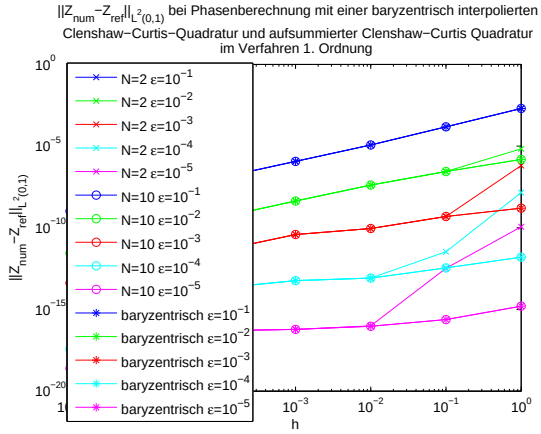

Abbildung 5.3: Fehler bei der Berechnung von Z mit dem WKB-Verfahren 1. Ordnung

Berechnet man (5.7) mit einer aufsummierten Clenshaw-Curtis Quadratur, sieht man in Abb. 5.3, dass die Wahl von N fast keinen Einfluss auf den Gesamtfehler bei kleinen Schrittweiten hat, da der Fehler des Verfahrens $\varepsilon^2 \min(\varepsilon, h)$ gegenüber dem Quadraturfehler dominiert. Für große Schrittweiten ist bei kleiner werdendem ε eine Verbesserung des Fehlers mit wachsendem N erkennbar. Ein Unterschied zwischen der Berechnung mit einer summierten Clenshaw-Curtis Quadratur mit $N = 10$ und einer baryzentrisch interpolierten Stammfunktion, bei der der Integrand bis $N = 18$ entwickelt worden ist, ist nicht erkennbar.
Berechnet man aus den Werten von Z mit (1.29) die Werte von U, lässt sich auch hier eine Abschätzung des Fehlers angeben. Aus (3.69) erhält man mit (5.10) bzw. (5.12)

Satz 5.3. *Unter den Voraussetzungen von Satz 5.1 gilt für den Fehler in U im Verfahren 1. Ordnung für $1 \leq t \leq T$,*

1.

$$\left\| U(x_t) - \widetilde{U}_t \right\| \leq C_1 \frac{h^{N+2}}{\varepsilon} + C_2 \varepsilon^2 \min(\varepsilon, h) + C_3 \varepsilon h^{N+2}, \qquad (5.15)$$

wenn N gerade ist und (5.7) mit einer aufsummierten Clenshaw-Curtis Quadratur, die bis zur Ordnung N entwickelt worden ist, berechnet wird;

2.

$$\left\| U\left(x_t\right) - \widetilde{U}_t \right\| \leq C_1 \frac{7.4^{-N}}{\varepsilon} + C_2 \varepsilon^2 \min\left(\varepsilon, h\right) + C_3 \varepsilon \cdot 7.4^{-N}, \qquad (5.16)$$

wenn (5.7) mit einer baryzentrisch interpolierten Clenshaw-Curtis Stammfunktion, die bis zur Ordnung N entwickelt worden ist, berechnet wird.

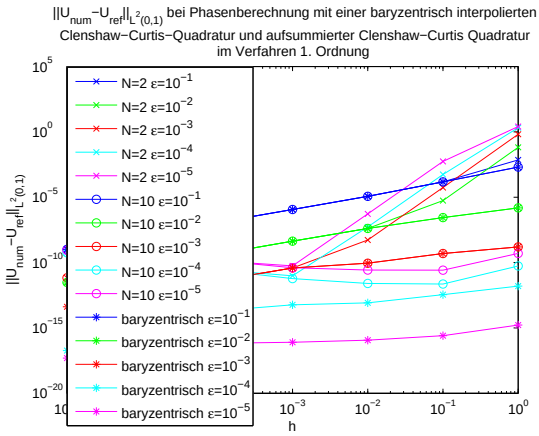

Abbildung 5.4: Fehler bei der Berechnung von U mit dem WKB-Verfahren 1. Ordnung

Berechnet man U mit einer aufsummierten Clenshaw-Curtis Quadratur mit kleinem N, z. B. wie in Abb. 5.4 mit $N = 2$, erkennt man den zusätzlichen Fehlerterm $\frac{h^4}{\varepsilon}$. Hier kehrt sich dadurch im Vergleich zu Abb. 5.3 die Reihenfolge der Fehlerkurven und der kleinste Parameterwert liefert den größten Fehler. Wird $N = 10$ gewählt, unterscheidet sich der Fehler für großes ε kaum von dem Fehler der Berechnung mit baryzentrischer Interpolation mit $N = 18$. Bei der Berechnung mit baryzentrischer Interpolation mit N, das groß genug gewählt ist, wie hier $N = 18$, hat der zusätzliche Fehlerterm $\frac{E}{\varepsilon}$ für große Werte von ε keine Auswirkung, da hier $7.4^{-18} \approx 2.2 \cdot 10^{-16} \ll \varepsilon$ ist. Wählt man $\varepsilon = 10^{-5}$ bleibt der Fehler bei der baryzentrischen Interpolation, wie schon bei Z im Bereich der Maschinengenauigkeit.
Bei der Berechnung der Phase mit der aufsummierten Clenshaw-Curtis Quadratur mit $N = 10$ erkennt man, dass bei $\varepsilon = 10^{-4}$ und $\varepsilon = 10^{-5}$ der Fehler auf Grund von Rundungsfehlern wieder zu wachsen beginnt, da hier der Fehler (5.13) sehr schnell im Bereich der Maschinengenauigkeit ist.
Um die Verwendung von Spektralmethoden zur Berechnung von (5.7) mit der Berechnung mittels Simpsonquadratur zu vergleichen, sind in Abb. 5.5 sowohl die Fehler bei der Berechnung von Z als auch die von der Berechnung von U für verschiedene Werte von ε eingezeichnet. Dabei wurde auf die Darstellung der Kurven für $\varepsilon = 10^{-1}$ und $\varepsilon = 10^{-5}$ verzichtet, da wie aus den beiden Abbildungen davor (Abb. 5.3 und Abb. 5.4) ersichtlich ist, alle Verfahren für $\varepsilon = 10^{-1}$ einen annähernd gleichen Fehler liefern. Für $\varepsilon = 10^{-5}$ liefert die Verwendung von baryzentrischer Interpolation mit $N = 18$ jeweils

Fehler im Bereich der Maschinengenauigkeit und ist somit wesentlich besser als das Simpsonverfahren.

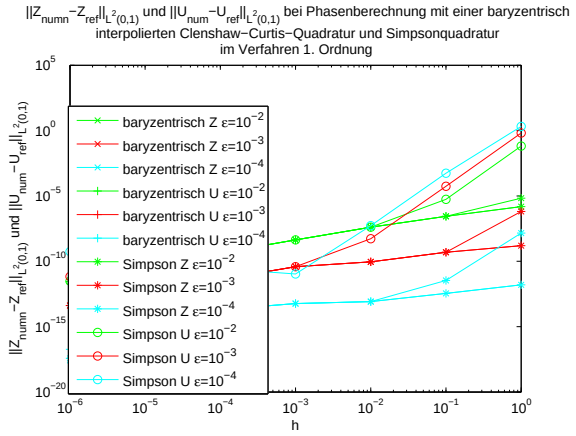

Abbildung 5.5: Vergleich von Spektralmethoden mit Simpsonverfahren im Verfahren 1. Ordnung

Für die in Abb. 5.5 gewählten Werte von ε erkennt man, dass sich die Fehler für Z und U bei der Berechnung mit baryzentrischer Interpolation kaum unterscheiden, während bei der Berechnung mit der Simpsonquadratur der zusätzliche Fehlerterm, der durch die Transformation (1.29) entsteht, erkennbar ist. Auch die Fehlerkurven für Z der beiden Verfahren unterscheiden sich für große Schrittweiten und die Verwendung von baryzentrischer Interpolation liefert den geringeren Fehler.

5.3 Verfahren 2. Ordnung

Berechnet man die Lösung von (1.17) mit dem Verfahren 2. Ordnung (1.26), lassen sich mit (5.10) und (5.12) Fehlerabschätzungen angeben, wenn $a(x) = e^{-x^2}$ gewählt wird. Aus Satz 3.11 erhält man als Spezialfall analog zu den bisherigen Sätzen

Satz 5.4. *Es sei $a(x) = e^{-x^2}$ und $0 < \varepsilon < 1$ eine beliebige, aber feste reelle Zahl. Dann gilt für den Fehler im Verfahren 2. Ordnung (1.26) für $1 \leq t \leq T$,*

1. $$\left\| Z\left(x_t\right) - \widetilde{Z}_t \right\| \leq C_1 \varepsilon^3 h^2 + C_2 \varepsilon h^{N+2} + C_3 \varepsilon^3 h^{N+1}, \tag{5.17}$$

   wenn N gerade ist und (5.7) mit einer aufsummierten Clenshaw-Curtis Quadratur, die bis zur Ordnung N entwickelt worden ist, berechnet wird;

2. $$\left\| Z\left(x_t\right) - \widetilde{Z}_t \right\| \leq C_1 \varepsilon^3 h^2 + C_2 \varepsilon \cdot 7.4^{-N} + C_3 \varepsilon^3 \min\left(\frac{\varepsilon}{h}, \frac{7.4^{-N}}{h}, 1\right), \tag{5.18}$$

wenn (5.7) mit einer baryzentrisch interpolierten Clenshaw-Curtis Stammfunktion, die bis zur Ordnung N entwickelt worden ist, berechnet wird.

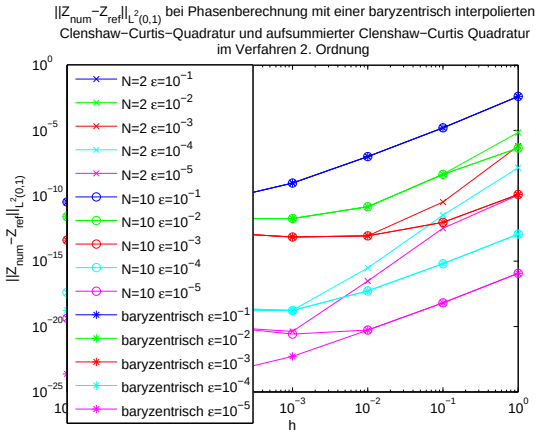

Abbildung 5.6: Fehler bei der Berechnung von Z mit dem WKB-Verfahren 2. Ordnung

In Abb. 5.6 sind Fehlerkurven gezeigt, wie sich die Art der Phasenberechnung auf den Gesamtfehler auswirkt. Es wurden dafür aufsummierte Clenshaw-Curtis Quadraturen mit $N = 2$ und $N = 10$ verwendet und weiters eine bis $N = 18$ entwickelte baryzentrisch interpolierte Stammfunktion. Für große Werte von ε hat die Art der Phasenberechnung nahezu keinen Einfluss auf den Fehler des Gesamtverfahrens, da hier der Term $\varepsilon^3 h^2$ des Verfahrensfehlers größer ist als der Quadraturfehler. Hier erkennt man gut, dass das Verfahren 2. Ordnung ist.
Wird ε kleiner gewählt, ergeben sich Unterschiede zwischen der Berechnung von (5.7) mit einer aufsummierten Clenshaw-Curtis Quadratur mit $N = 2$ und $N = 10$. Die Variante mit $N = 10$ unterscheidet sich im Gesamtfehler kaum von dem Verfahren, in dem baryzentrische Interpolation verwendet worden ist. Bei diesen beiden Methoden ist der Quadraturfehler kleiner als der Verfahrensfehler $\varepsilon^3 h^2$ und an den Fehlerkurven erkennt man wieder den Fehler des Verfahrens. Bei der aufsummierten Clenshaw-Curtis Quadratur mit $N = 2$ ist der Verfahrensfehler $\varepsilon^3 h^2$ kleiner als der Quadraturfehler h^4 und man erkennt den Quadraturfehler in der Fehlerkurve für $\varepsilon = 10^{-4}$ und $\varepsilon = 10^{-5}$.
Die Abschätzungen für den Fehler bei der Berechnung von U aus Z mittels (1.29) lassen sich ebenfalls angeben.

Satz 5.5. *Unter den Voraussetzungen von Satz 5.4 gilt für den Fehler in U im Verfahren 2. Ordnung für $1 \leq t \leq T$,*

1.

$$\left\| U(x_t) - \widetilde{U}_t \right\| \leq C_1 \frac{h^{N+2}}{\varepsilon} + C_2 \varepsilon^3 h^2 + C_3 \varepsilon h^{N+2} + C_4 \varepsilon^3 h^{N+1}, \quad (5.19)$$

wenn N gerade ist und (5.7) mit einer aufsummierten Clenshaw-Curtis Quadratur, die bis zur Ordnung N entwickelt worden ist, berechnet wird;

2.

$$\left\|U\left(x_t\right)-\widetilde{U}_t\right\| \leq C_1\frac{7.4^{-N}}{\varepsilon}+C_2\varepsilon^3h^2+C_3\varepsilon\cdot 7.4^{-N}+C_4\varepsilon^3\min\left(\frac{\varepsilon}{h},\frac{7.4^{-N}}{h},1\right), \tag{5.20}$$

wenn (5.7) mit einer baryzentrisch interpolierten Clenshaw-Curtis Stammfunktion, die bis zur Ordnung N entwickelt worden ist, berechnet wird.

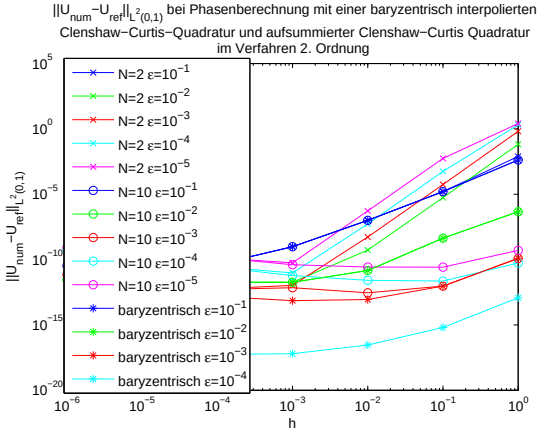

Abbildung 5.7: Fehler bei der Berechnung von U mit dem WKB-Verfahren 2. Ordnung

Die Auswirkung der Art der Phasenberechnung im Fehler von U ist in Abb. 5.7 gezeigt. Dabei ist (5.7) mit aufsummierten Clenshaw-Curtis Quadraturen mit $N = 2$ und $N = 10$ und mit einer baryzentrisch interpolierten Stammfunktion, die bis $N = 18$ entwickelt worden ist, berechnet worden. Wird $\varepsilon = 10^{-1}$ gewählt, hat die Art der Phasenberechnung fast keinen Einfluss auf den Fehler, da hier der Verfahrensfehler gegenüber dem Quadraturfehler dominiert. Für kleinere Werte von ε ist bei der Phasenberechnung mit der aufsummierten Clenshaw-Curtis Quadratur mit $N = 2$ wieder der zusätzliche Term $\frac{E}{\varepsilon}$, der durch die Transformation (1.29) entsteht, erkennbar, da der kleinste $\varepsilon-$Wert den größten Fehler liefert. Dieser Term dominiert den Gesamtfehler. Dies erkennt man auch daran, dass der Gesamtfehler mit $h^{N+2} = h^4$ fällt und nicht mit h^2 wie der Verfahrensfehler.
Dass der Fehlerterm durch die Transformation von Z auf U dominiert, tritt bei der Verwendung einer aufsummierten Clenshaw-Curtis Quadratur mit $N = 10$ erst bei kleinen Werten von ε auf. Bei der Berechnung der Phase mit baryzentrisch interpolierter Stammfunktion tritt dieser Effekt nicht auf.
Der Fehler für $\varepsilon = 10^{-5}$ bei der baryzentrischen Interpolation ist in Abb. 5.7 nicht abgebildet, da schon bei der Berechnung von Z der Fehler im Bereich der Maschinengenauigkeit ist. Man erkennt bei $\varepsilon = 10^{-3}$ und $\varepsilon = 10^{-4}$, dass

der Fehler sehr schnell im Bereich der Maschinengenauigkeit ist. Bei der Berechnung mit aufsummierter Quadratur erkennt man wieder ein Ansteigen des Fehlers durch Rundungsfehler. Vergleicht man die Verwendung von Spektralmethoden mit anderen Quadraturverfahren, erkennt man wieder, dass die Spektralmethoden die besseren Ergebnisse liefern.

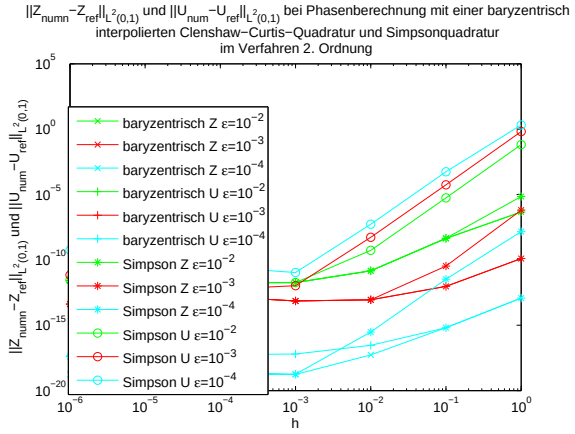

Abbildung 5.8: Vergleich von Spektralmethoden mit Simpsonverfahren im Verfahren 2. Ordnung

In Abb. 5.8 sind die Gesamtfehler in der Berechnung von Z und U abgebildet, wenn für die Phasenberechnung ein Simpsonverfahren und eine baryzentrisch interpolierte Clenshaw-Curtis Stammfunktion mit $N = 18$ verwendet werden. Der zusätzliche Fehlerterm bei der Berechnung von U ist auch bei der Verwendung der Simpsonquadratur erkennbar und der kleinste Wert von ε liefert den größten Fehler. Im Gegensatz dazu ist bei $\varepsilon = 10^{-2}$ und $\varepsilon = 10^{-3}$ fast kein Unterschied zwischen den Fehlern in Z und U erkennbar, wenn die Phase mit baryzentrisch interpolierter Clenshaw-Curtis Stammfunktion berechnet wird. Bei $\varepsilon = 10^4$ ist ein Unterschied zwischen den Fehlerkurven von Z und U erkennbar, wenn die Schrittweite kleiner wird, da hier wieder die Maschinengenauigkeit erreicht ist. Für kleine Werte von ε ist auch eine Verbesserung des Fehlers bei der Berechnung von Z erkennbar, wenn anstelle des Simpsonverfahrens eine baryzentrisch interpolierte Stammfunktion verwendet wird.

Kapitel 6

Conclusio

Die beiden Beispiele aus den vorigen Kapiteln haben gezeigt, dass die Berechnung des Phasenintegrals (1.2) mit Spektralmethoden den Fehler gegenüber herkömmlichen Quadraturverfahren der Ordnung h^γ deutlich verbessern.
Ganz allgemein haben diese Quadraturverfahren Vorteile gegenüber klassischen Verfahren.
Will man ein Integral der Form $\int_a^b f(x)\mathrm{d}x$ berechnen, eignen sich auf Spektralmethoden basierende Quadraturverfahren ausgezeichnet, wenn das Intervall $[a;b]$ nicht in Teilintervalle $[x_t;x_{t+1}]$ der Länge h unterteilt wird. Dann ist das Verfahren aus Abschnitt 2.3 oder die Clenshaw-Curtis Quadratur effizienter als Verfahren, bei denen für eine gewünschte Genauigkeit eine Unterteilung von $[a;b]$ in Teilintervalle notwendig ist. Im direkten Vergleich ist dabei die Clenshaw-Curtis Quadratur die bessere Wahl als das Verfahren aus Abschnitt 2.3. Man könnte natürlich dieses Integral $\int_a^b f(x)\mathrm{d}x$ auch mit einer Gauß-Quadratur berechnen. Die Unterschiede zwischen Gauß-Quadratur und Clenshaw-Curtis Quadratur werden ausführlich in [Tr2] behandelt. Dabei zeigt sich, dass sich die Fehler von Gauß-Quadratur und Clenshaw-Curtis Quadratur für C^∞- und C^k-Funktionen kaum unterscheiden (vgl. [Tr2, Fig. 2] bzw. [Tr2, Theorem 4.5] und [Tr2, Theorem 5.1]).
Muss das Intervall $[a;b]$ in Teilintervalle $[x_t;x_{t+1}]$ der Länge h geteilt werden, um z. B. so wie in den WKB-Verfahren die Phase an diesen Stellen x_t zu berechnen, liefert die Clenshaw-Curtis Quadratur oder das Verfahren aus Abschnitt 2.3 immer noch bessere Ergebnisse als andere Quadraturverfahren wie die Simpsonquadratur oder die Trapezregel. Die beste Variante, um Werte einer Stammfunktion an diesen Punkten x_t zu ermitteln, ist die Berechnung der Stammfunktion mit (2.31) und anschließender Berechnung der gesuchten Funktionswerte an den Stellen x_t mit baryzentrischer Interpolation. Die Methode der baryzentrischen Interpolation hat gegenüber der Clenshaw-Curtis Quadratur den Vorteil, dass die Stammfunktion nur einmal

berechnet werden muss und anschließend für verschiedene Wahl von x_t interpoliert werden kann. Im Gegensatz dazu müssen bei jeder Änderung der Stellen x_t mit einer Clenshaw-Curtis Quadratur oder auch einem Simpsonverfahren die gesuchten Werte neu berechnet werden. Dies führt, wenn die Anzahl an Stellen x_t groß ist, zu einem erheblichen Mehraufwand.
In der konkreten Anwendung in dieser Arbeit führte die Verwendung einer baryzentrisch interpolierten Stammfunktion von (1.2) zu einer Verbesserung des Gesamtfehlers im Vergleich zu einer Clenshaw-Curtis Quadratur oder einem Simpsonverfahren. Bei der Berechnung von Z ist diese Verbesserung nicht so deutlich zu erkennen wie in der Berechnung von U. Besonders das Beispiel aus Kapitel 5 zeigt diese Verbesserung in eindrucksvoller Weise. Hier hat der zusätzliche Fehlerterm, der durch die Transformation (1.29) von Z auf U entsteht, fast keinen Einfluss, da die Funktionswerte der Phase bis auf Maschinengenauigkeit approximiert werden können.
Lässt sich der erste Teil des Phasenintegrals exakt berechnen, so wie dies in Kapitel 4 der Fall war, ist der Unterschied im Fehler zwischen den einzelnen Arten der Phasenberechnung vernachlässigbar, da dadurch der Faktor ε^2, mit dem das Integral von β multipliziert wird, den Fehler noch zusätzlich verkleinert. Auch der zusätzliche Fehlerterm in U wirkt sich nicht so stark aus, da die Division durch ε gekürzt wird. In diesem Beispiel ist trotzdem die Verwendung der baryzentrisch interpolierten Stammfunktion von (1.2) vorteilhaft, da diese, vor allem bei kleinen Schrittweiten, effizienter berechnet werden kann.
Abschließend kann man festhalten, dass auf Spektralmethoden basierende Quadraturverfahren gegenüber einer Trapezregel oder einer Simpsonquadratur die bessere Wahl sind.

Anhang A

Fehlertabellen

A.1 Quadraturfehler für das Beispiel mit $a(x) = \left(x + \frac{1}{2}\right)^2$

A.1.1 mit dem Verfahren aus Abschnitt 2.3

Fehlerkonstanten für $E_N \leq C \cdot e^{\lambda N} = C \cdot 10^{\kappa N}$.

	C	λ	κ
$\varepsilon = 10^{-1}$			
$h = 10^{0}$	$8.3 \cdot 10^{-2}$	-1.2915	-0.5609
$h = 10^{-1}$	$5.91 \cdot 10^{-2}$	-3.3527	-1.4560
$h = 10^{-2}$	$6.4912 \cdot 10^{-2}$	-5.6906	-2.4714
$h = 10^{-3}$	$6.5553 \cdot 10^{-2}$	-7.9979	-3.4735
$h = 10^{-4}$	$2.8526 \cdot 10^{-2}$	-9.8845	-4.2928
$\varepsilon = 10^{-2}$			
$h = 10^{0}$	$5.2743 \cdot 10^{-4}$	-1.2689	-0.5510
$h = 10^{-1}$	$5.9132 \cdot 10^{-4}$	-3.3527	-1.4560
$h = 10^{-2}$	$6.4912 \cdot 10^{-4}$	-5.6906	-2.4714
$h = 10^{-3}$	$6.3990 \cdot 10^{-4}$	-7.9859	-3.4682
$\varepsilon = 10^{-3}$			
$h = 10^{0}$	$4.769 \cdot 10^{-6}$	-1.2615	-0.5479
$h = 10^{-1}$	$5.9132 \cdot 10^{-6}$	-3.3527	-1.4560
$h = 10^{-2}$	$6.4974 \cdot 10^{-6}$	-5.6910	-2.4714

Fortsetzung

	C	λ	κ
$\varepsilon = 10^{-4}$			
$h = 10^{0}$	$4.3 \cdot 10^{-8}$	-1.2535	-0.5444
$h = 10^{-1}$	$5.9122 \cdot 10^{-8}$	-3.3526	-1.4560
$h = 10^{-2}$	$5.7299 \cdot 10^{-8}$	-5.6279	-2.4442
$\varepsilon = 10^{-5}$			
$h = 10^{0}$	$4.5674 \cdot 10^{-10}$	-1.2647	-0.5493
$h = 10^{-1}$	$5.817 \cdot 10^{-10}$	-3.3446	-1.4525

Tabelle A.1: A posteriori Konstanten für $N \geq 2$ zu Abschnitt 4.1.1

Fehlerkonstanten für $E_h \leq C \cdot h^{\gamma}$.

	C	γ		C	γ
$N = 2$			$N = 4$		
$\varepsilon = 10^{-1}$	$7.5339 \cdot 10^{-3}$	2.0018	$\varepsilon = 10^{-1}$	$1.1459 \cdot 10^{-3}$	3.9815
$\varepsilon = 10^{-2}$	$7.4505 \cdot 10^{-5}$	2.0008	$\varepsilon = 10^{-2}$	$1.1447 \cdot 10^{-5}$	3.9811
$\varepsilon = 10^{-3}$	$7.4067 \cdot 10^{-7}$	2.0000	$\varepsilon = 10^{-3}$	$8.0492 \cdot 10^{-8}$	3.8281
$\varepsilon = 10^{-4}$	$8.3155 \cdot 10^{-9}$	2.0250	$\varepsilon = 10^{-4}$	$2.8547 \cdot 10^{-10}$	3.3767
$\varepsilon = 10^{-5}$	$6.9406 \cdot 10^{-11}$	1.9817	$\varepsilon = 10^{-5}$	$2.8548 \cdot 10^{-12}$	3.3310
$N = 5$			$N = 10$		
$\varepsilon = 10^{-1}$	$5.0591 \cdot 10^{-4}$	4.9781	$\varepsilon = 10^{-1}$	$1.5490 \cdot 10^{-7}$	8.0307
$\varepsilon = 10^{-2}$	$3.1872 \cdot 10^{-6}$	4.7775	$\varepsilon = 10^{-2}$	$1.5490 \cdot 10^{-9}$	6.6675
$\varepsilon = 10^{-3}$	$8.0280 \cdot 10^{-9}$	4.1787	$\varepsilon = 10^{-3}$	$1.5490 \cdot 10^{-11}$	4.8436
$\varepsilon = 10^{-4}$	$8.0280 \cdot 10^{-11}$	4.1780	$\varepsilon = 10^{-4}$	$1.5477 \cdot 10^{-13}$	2.5422
$\varepsilon = 10^{-5}$	$8.0291 \cdot 10^{-13}$	3.5582			

Tabelle A.2: A posteriori Konstanten für festes h zu Abschnitt 4.1.1

A.1.2 mit Clenshaw-Curtis Quadratur

Fehlerkonstanten für $E_N \leq C \cdot e^{\lambda N} = C \cdot 10^{\kappa N}$.

	C	λ	κ		C	λ	κ
$\varepsilon = 10^{-1}$				$\varepsilon = 10^{-2}$			
$h = 10^{0}$	0.028	-1.4	-0.6	$h = 10^{0}$	$2.8 \cdot 10^{-4}$	-1.4	-0.6

Fortsetzung

	C	λ	κ		C	λ	κ
$\varepsilon = 10^{-1}$				$\varepsilon = 10^{-2}$			
$h = 10^{-1}$	0.01	-3.8	-1.6	$h = 10^{-1}$	10^{-4}	-3.8	-1.6
$\varepsilon = 10^{-3}$				$\varepsilon = 10^{-4}$			
$h = 10^{0}$	$2.8 \cdot 10^{-6}$	-1.4	-0.6	$h = 10^{0}$	$2.8 \cdot 10^{-8}$	-1.4	-0.6
$h = 10^{-1}$	10^{-6}	-3.8	-1.6	$h = 10^{-1}$	$1.4 \cdot 10^{-9}$	-3.1	-1.4
$\varepsilon = 10^{-5}$							
$h = 10^{0}$	$2.6 \cdot 10^{-10}$	-1.4	-0.6				

Tabelle A.3: A posteriori Konstanten für $N \geq 1$ zu Abschnitt 4.1.2

Fehlerkonstanten für $E_h \leq C \cdot h^{\gamma}$.

	C	γ		C	γ
$N = 2$			$N = 4$		
$\varepsilon = 10^{-1}$	0.0045	3.9795	$\varepsilon = 10^{-1}$	$3.7 \cdot 10^{-4}$	5.9117
$\varepsilon = 10^{-2}$	$4.5 \cdot 10^{-5}$	3.9794	$\varepsilon = 10^{-2}$	$2.47 \cdot 10^{-7}$	4.7362
$\varepsilon = 10^{-3}$	$4.6 \cdot 10^{-7}$	3.9888	$\varepsilon = 10^{-3}$	$2.47 \cdot 10^{-9}$	4.7362
$\varepsilon = 10^{-4}$	$1 \cdot 10^{-9}$	3.331	$\varepsilon = 10^{-4}$	$2.47 \cdot 10^{-11}$	4.5687
$\varepsilon = 10^{-5}$	$1 \cdot 10^{-11}$	3.3291	$\varepsilon = 10^{-5}$	$2.47 \cdot 10^{-13}$	3.3467
$N = 5$			$N = 10$		
$\varepsilon = 10^{-1}$	10^{-5}	5.2457	$\varepsilon = 10^{-1}$	$1.7 \cdot 10^{-9}$	6.9
$\varepsilon = 10^{-2}$	$8 \cdot 10^{-8}$	5.1349	$\varepsilon = 10^{-2}$	$1.7 \cdot 10^{-11}$	4.8
$\varepsilon = 10^{-3}$	$8 \cdot 10^{-10}$	5.1415	$\varepsilon = 10^{-3}$	$1.7 \cdot 10^{-13}$	3.2003
$\varepsilon = 10^{-4}$	$8 \cdot 10^{-12}$	4.5565			
$\varepsilon = 10^{-5}$	$8 \cdot 10^{-14}$	2.8573			

Tabelle A.4: A posteriori Konstanten für festes h zu Abschnitt 4.1.2

A.2 Quadraturfehler für das Beispiel mit $a(x) = e^{-x^2}$

A.2.1 mit dem Verfahren aus Abschnitt 2.3

Fehlerkonstanten für $E_h \leq C \cdot h^{\gamma}$.

	C	γ		C	γ
$N = 2$			$N = 4$		
$\varepsilon = 10^{-1}$	0.0264	2.0279	$\varepsilon = 10^{-1}$	$1.1884 \cdot 10^{-4}$	3.9880
$\varepsilon = 10^{-2}$	0.0269	2.0264	$\varepsilon = 10^{-2}$	$1.1449 \cdot 10^{-4}$	3.9813
$\varepsilon = 10^{-3}$	0.0269	2.0264	$\varepsilon = 10^{-3}$	$1.1445 \cdot 10^{-4}$	3.9812
$\varepsilon = 10^{-4}$	0.0269	2.0264	$\varepsilon = 10^{-4}$	$1.1445 \cdot 10^{-4}$	3.9812
$\varepsilon = 10^{-5}$	0.0269	2.0264	$\varepsilon = 10^{-5}$	$1.1445 \cdot 10^{-4}$	3.9812
$N = 5$			$N = 6$		
$\varepsilon = 10^{-1}$	$1.6417 \cdot 10^{-5}$	5.0938	$\varepsilon = 10^{-1}$	$4.2169 \cdot 10^{-7}$	5.8083
$\varepsilon = 10^{-2}$	$1.6784 \cdot 10^{-5}$	5.0874	$\varepsilon = 10^{-2}$	$4.6663 \cdot 10^{-7}$	5.8412
$\varepsilon = 10^{-3}$	$1.6788 \cdot 10^{-5}$	5.0873	$\varepsilon = 10^{-3}$	$4.6708 \cdot 10^{-7}$	5.8416
$\varepsilon = 10^{-4}$	$1.6788 \cdot 10^{-5}$	5.0873	$\varepsilon = 10^{-3}$	$4.6709 \cdot 10^{-7}$	5.8415
$\varepsilon = 10^{-5}$	$1.6788 \cdot 10^{-5}$	5.0873	$\varepsilon = 10^{-3}$	$4.6709 \cdot 10^{-7}$	5.8416

Tabelle A.5: A posteriori Konstanten für festes h zu Abschnitt 5.1.1

A.2.2 mit Clenshaw-Curtis Quadratur

Fehlerkonstanten für $E_h \leq C \cdot h^\gamma$.

	C	γ		C	γ
$N = 2$			$N = 3$		
$\varepsilon = 10^{-1}$	$4.7916 \cdot 10^{-4}$	3.9914	$\varepsilon = 10^{-1}$	$1.2214 \cdot 10^{-4}$	3.9997
$\varepsilon = 10^{-2}$	$4.6216 \cdot 10^{-4}$	3.9852	$\varepsilon = 10^{-2}$	$1.1814 \cdot 10^{-4}$	3.9947
$\varepsilon = 10^{-3}$	$4.6199 \cdot 10^{-4}$	3.9851	$\varepsilon = 10^{-3}$	$1.1809 \cdot 10^{-4}$	3.9946
$\varepsilon = 10^{-4}$	$4.6199 \cdot 10^{-4}$	3.9851	$\varepsilon = 10^{-4}$	$1.1809 \cdot 10^{-4}$	3.9946
$\varepsilon = 10^{-5}$	$4.6199 \cdot 10^{-4}$	3.9851	$\varepsilon = 10^{-5}$	$1.1809 \cdot 10^{-4}$	3.9946
$N = 4$			$N = 5$		
$\varepsilon = 10^{-1}$	$7.6258 \cdot 10^{-7}$	5.8154	$\varepsilon = 10^{-1}$	$9.0184 \cdot 10^{-8}$	5.7921
$\varepsilon = 10^{-2}$	$8.3921 \cdot 10^{-7}$	5.8460	$\varepsilon = 10^{-2}$	$1.0105 \cdot 10^{-7}$	5.8304
$\varepsilon = 10^{-3}$	$8.3997 \cdot 10^{-7}$	5.8463	$\varepsilon = 10^{-3}$	$1.0116 \cdot 10^{-7}$	5.8309
$\varepsilon = 10^{-4}$	$8.3998 \cdot 10^{-7}$	5.8463	$\varepsilon = 10^{-4}$	$1.0116 \cdot 10^{-7}$	5.8306
$\varepsilon = 10^{-5}$	$8.3998 \cdot 10^{-7}$	5.8463	$\varepsilon = 10^{-5}$	$1.0116 \cdot 10^{-7}$	5.8309
$N = 6$			$N = 10$		
$\varepsilon = 10^{-1}$	$6.0661 \cdot 10^{-10}$	8.0603	$\varepsilon = 10^{-1}$	$2.1094 \cdot 10^{-15}$	0.8016
$\varepsilon = 10^{-2}$	$4.4826 \cdot 10^{-10}$	8.0593	$\varepsilon = 10^{-2}$	$3.7748 \cdot 10^{-15}$	1.2304
$\varepsilon = 10^{-3}$	$4.4668 \cdot 10^{-10}$	8.0627	$\varepsilon = 10^{-3}$	$3.7748 \cdot 10^{-15}$	1.0544

Fortsetzung

	C	γ		C	γ
$\varepsilon = 10^{-4}$	$4.4666 \cdot 10^{-10}$	8.0627	$\varepsilon = 10^{-4}$	$3.7748 \cdot 10^{-15}$	1.0544
$\varepsilon = 10^{-5}$	$4.4666 \cdot 10^{-10}$	8.0590	$\varepsilon = 10^{-5}$	$3.6637 \cdot 10^{-15}$	1.2175

Tabelle A.6: A posteriori Konstanten für festes h zu Abschnitt 5.1.2

Literaturverzeichnis

[ABN] Arnold, A., Ben Abdallah, N. und Negulescu, C. *WKB-based schemes for the oscillatory 1D Schrödinger equaquation in the semiclassical limit,* SIAM J. Numer. Anal. 49, Nr. 4 (2011), pp. 1436-1460.

[Ba] Basu, N. K. *Error Estimates for a Chebyshev Quadrature Method,* Math. Comp. Vol 24, 112 (1970), pp. 863-867.

[BK] Berrut, J.-P. und Klein, G. *Recent advances in linear barycentric rational interpolation,* Journal of Computational and Applied Mathematics, Vol. 259 (2014), pp. 95-107.

[BR] Brezinksi, C. und Redivo-Zaglia, M. *Padé-type rational and barycentric interpolation,* Numerische Mathematik 125, 1 (2013), pp. 89-113.

[BT] Berrut, J.-P. und Trefethen, L. N. *Barycentric Lagrange Interpolation,* SIAM Review, Vol. 46, 3 (2004), pp. 501-517.

[Bo1] Boyd, J. P. *Chebyshev and Fourier Spectral Methods,* 2d. edition, Dover Publishers, 2000. online erhältlich http://www-personal.umich.edu/∼jpboyd/BOOK_Spectral2000.html, abgerufen am 12.12.2013.

[Bo2] Boyd, J. P. *The Rate of Convergence of Fourier Coefficients for Entire Functions of Infinite Order with Application to the Weidemann-Cloot Sinh-Mapping for Pseudospectral Computations on an Infinite Intervall,* Journal of Computational Physics, 111, 2 (1994), pp. 360-372.

[Ch] Chawla, M. M. *Error Estimates for the Clenshaw-Curtis Quadrature,* Math. Comp. Vol 22, 103 (1968), pp. 651-656.

[CC] Clenshaw, C. W. und Curtis, A. R. *A method for numerical integration on an automatic computer,* Numerische Mathematik 2, 1 (1960), pp. 197-205.

[Da] Davis, P. J. *Interpolation And Approximation,* Blaisdell Publishin Company, Waltham Toronto London, 1963.

[DR] Davis, P. J. und Rabinowitz, P. *Numerical Integration,* Blaisdell Publishing Company, Waltham Toronto London, 1967.

[De] Demtröder, W. *Experimentalphysik 3 Atome, Moleküle und Festkörper,* 4. Auflage, Springer Verlag, Berlin Heidelberg, 2009.

[Fi] Filippi, S. *Angenäherte Tschebyscheff-Approximation - eine Modifikation des Verfahrens von Clenshaw und Curtis,* Numerische Mathematik 6, 1 (1964), pp. 320-328.

[FK] Frauendiener, J. und Klein, Ch. *Hyperelliptic Theta-Functions and Spectral Methods: KdV and KP Solutions,* Letters in Mathematical Physics 76 (2006), pp. 249-267.

[FP] Fox, L. und Parker, I. B. *Chebyshev Polynomials in Numerical Analysis,* Oxford University Press, London, 1968.

[Ge] Geier, J. *Efficient integrators for linear highly oscillatory ODEs based on asymptotic expansions,* Dissertation, TU Wien, 2011.

[GHO] Gottlieb, D., Hussaini, M.Y. und Orszag, S. A. *Introduction: Theory and Applications of Spectral Methods,* in Voigt, R. G., Gottlieb, D. und Hussaini, M. Y., eds., Spectral Methods for Partial Differential Equations, pp. 1-54, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1984.

[GO] Gottlieb, D. und Orszag, S. A. *Numerical analysis of spectral methods: Theory and Applications,* CBMS Regional Conference Series in Applied Mathematics 26, 5. Auflage, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1989.

[GRB] Gillette, A., Rand, A. und Bajaj, Ch. *Error estimates for generalized barycentric interpolation,* Advances in Computational Mathematics, Vol. 37, 3 (2012), pp. 417-439.

[GST] Gil, A., Segura, J. und Temme, N. M. *Numerical Methods for Special Functions,* Societey for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007.

[HH] Hämmerlin, G. und Hoffmann, K.-H. *Numerische Mathematik,* Springer-Verlag, Berlin Heidelberg, 1989.

[Hel1] Held, K. *Quantentheorie I,* Vorlesungsskript, TU Wien, 2013.

[Hel2] Held, K. *Quantentheorie II,* Vorlesungsskript, TU Wien, 2012.

[Heu1] Heuser, H. *Lehrbuch der Analysis Teil 1,* 15. Auflage, Teubner, Stuttgart Leipzig Wiesbaden, 2003.

[Heu2] Heuser, H. *Lehrbuch der Analysis Teil 2,* 12. Auflage, Teubner, Stuttgart Leipzig Wiesbaden, 2002.

[Hi] Higham, N. J. *The numerical stabilty of barycentric Lagrange interpolation,* IMA Journal of Numerical Analysis, Vol. 24, 4 (2004), pp. 547-556.

[HJ1] Horn, R. A. und Johnson, C. R. *Matrix analysis,* Cambridge University Press, Cambridge, 1999.

[HJ2] Horn, R. A. und Johnson, C. R. *Topics in matrix analysis,* Cambridge University Press, Cambridge, 1999.

[Kr] Kreuzer, M. *Quantum Theory,* Vorlesungsskript, TU Wien, 2009.

[Ma] Mascarenhas, W. F. *The stability of barycentric interpolation at the Chebyshev points of the second kind,* Numerische Mathematik 128, 2 (2014), pp. 265-300.

[No] Nolting, W. *Grundkurs Theoretische Physik 5/2 Quantenmechanik - Methoden und Anwendung,* 7. Auflage, Springer Verlag, Heidelberg Dodrecht London New York, 2012.

[Pl] Plato, R. *Numerische Mathematik kompakt Grundlagen für Studium und Praxis,* 4. Auflage, Vieweg+Teubner, Wiesbaden, 2010.

[Pr] Praetorius, D. *Numerische Mathematik,* Vorlesungsskript, TU Wien, 2005.

[QuV] Quarteroni, V. und Valli, A. *Numerical Approximation of Partial Differential Equations,* 2. Auflage, Springer Verlag, Berlin Heidelberg, 1997.

[Ri] Rivlin, T. J. *The Chebyshev Polynomials,* John Wiley & Sons, New York London Sydney Toronto, 1974.

[SV] Sadiq, B. und Viswanath, D. *Baryzentric Hermite Interpolation,* SIAM Journal on Scientific Computing, Vol. 35, 3 (2013), pp. 1254-1270.

[Tr1] Trefethen, L. N. *Spectral Methods in MATLAB, vol. 10 of Software, Environments, and Tools,* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000.

[Tr2] Trefethen, L. N. *Is Gauss Quadrature Better than Clenshaw-Curtis?* SIAM Review, Vol. 50, 1 (2008), pp. 67-87.

[ÜK] Überhuber, Ch. und Katzenbeisser St. *Einführung in MATLAB,* TU Wien, 2002.

[WT] Weideman, J. A. C. und Trefethen, L. N. *The kink phenomenon in Fejér and Clenshaw-Curtis quadrature,* Numerische Mathematik 107, 4 (2007), pp. 707-727.

[WTG] Webb, M., Trefethen, L. N. und Gonnet, P. *Stability of Barycentric Interpolation Formulas for Extrapolation,* SIAM Journal on Scientific Computing, Vol. 34, 6 (2012), pp. A3009-A3015.

[Zh] Zhan, X. *Matrix Theory,* AMS, Providence, RI, 2013.