TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

DISSERTATION

Efficient integrators for linear highly oscillatory ODEs based on asymptotic expansions

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines Doktors der technischen Wissenschaften unter der Leitung von

Univ.Prof. Dipl.-Ing. Dr.techn. Anton Arnold
E101
Institut für Analysis und Scientific Computing

eingereicht an der Technischen Universität Wien
bei der Fakultät für Mathematik und Geoinformation

von

Dipl.-Math. Jens Geier
Matrikelnummer: 0527997
Sommerergasse 15/5, 1130 Wien

Datum

Unterschrift

To Ines and Jost.

Ich schreibe nicht, euch zu gefallen;
Ihr sollt was lernen.

J.W.v. Goethe, Zahme Xenien I.

Acknowledgment

I want to thank all the people which have encouraged and supported me during my PhD. First of all I am grateful to my advisor Prof. Anton Arnold for his friendship, support, patience and the very instructive discussions on various topics within the last years. Thanks to all friends and the staff of the Institute for Analysis and Scientific Computing of the Vienna University of Technology, especially to Franz, Roberta, Lukas, Ilona, Jan, Maia, Dominik, Gabriela, Markus, Ewa, Peter and Ursula. Also special thanks to Prof. Claudia Negulescu from Institut de Mathématiques de Toulouse (Université Paul Sabatier), who agreed to examine this these. I also acknowledge the financial support of the Wissenschaftskolleg Differential Equations (supported by the Austrian Science Fund). Last but not least I thank my wife Ines who believed in me.

Deutsche Kurzfassung

Hoch oszillierende Phänomene treten in einer Vielzahl naturwissenschaftlicher Modelle aus den verschiedensten Bereichen auf, wie z. B. in der Elektrodynamik, in der Akustik, bei der Modellierung von Molekülen, beim Plasmatransport, bei der Computertomografie oder in der Quantenmechanik. Diese Arbeit konzentriert sich auf die numerische Lösung von hoch oszillierenden Problemen, die mittels Systemen linearer gewöhnlicher Differentialgleichungen (Dgl) beschrieben werden können.

Es sei $\varepsilon > 0$ eine Konstante und $V\colon \mathbb{R} \to \mathbb{R}$ eine glatte Funktion. Auf dem Intervall $[a, b] \subset \mathbb{R}$ betrachtet man das skalare Anfangswert-Problem (AWP)

$$\psi''(x) + \frac{1}{\varepsilon^2} V(x)\,\psi(x) \;=\; 0\,, \qquad (\psi(a), \psi'(a))^T \;=\; \psi_0 \in \mathbb{C}^2\,. \tag{1}$$

Die (skalare) Dgl (1) wird auch (1D) stationäre Schrödingergleichung (Sgl) genannt. Bei geeigneter Wahl des Parameters ε, der Funktion V und der Anfangsbedingungen ist das AWP ein einfaches quantenmechanisches Modell für ein eindimensionales Elektron im thermodynamischen Gleichgewicht (cf. [4]).

Ist $V > 0$ konstant, so ist $\psi(x) = c_1 \sin(\frac{\sqrt{V}}{\varepsilon} x) + c_2 \cos(\frac{\sqrt{V}}{\varepsilon} x)$ eine allgemeine Lösung von (1). Für $\varepsilon \ll 1$ ist ψ also eine hoch oszillierende Funktion, mit Schwingungsamplituden der Ordnung $\mathcal{O}(1)$. Dieser Charakter der Lösung bleibt auch für nicht konstante $V > 0$ erhalten. Je kleiner ε wird, desto stärker oszilliert ψ. Löst man das AWP (1) mit einem Standardverfahren, wie z. B. der klassischen Runge-Kutta-Methode, so benötigt man sehr fein auflösende Gitter (mit Ortsschrittweite $h < \varepsilon$), um verlässliche Resultate zu erzielen. Für ein einzelnes AWP dieser Art ist dies mit Sicherheit ein praktikabler Zugang. In vielen Anwendungen, wie z. B. bei der Modellierung von Halbleitern, ist man allerdings darauf angewiesen, sehr viele solcher Systeme zu lösen (siehe z. B. [4]). Dies führt zu einem immensen Rechenaufwand und folglich besteht Interesse an möglichst effizienten Lösungsmethoden für (1) und verwandte Systeme.

Das AWP (1) ist äquivalent zu einem System erster Ordnung der Gestalt

$$u'(x) \;=\; \frac{1}{\varepsilon} L(x) u(x) + B(x) u(x)\,, \qquad u(a) \;=\; u_0 \in \mathbb{C}^d\,, \tag{2}$$

mit einer reellen Diagonalmatrix $L(x)$ und einer (evtl. komplexen) Matrix $B(x)$ (siehe § 2.2). Wie in [4, 54, 27] beschrieben, lässt sich das AWP (2) bzw. (1) derart in ein System von Dgl erster Ordnung überführen, dass die dominanten Oszillationen mit Amplituden der Ordnung $\mathcal{O}(1)$ eliminiert werden. Das resultierende System aus [4] z. B. hat die Gestalt

$$z'(x) \;=\; \varepsilon A(x) z(x)\,, \qquad z(a) \;=\; z_0 \in \mathbb{C}^d\,. \tag{3}$$

Der Preis, den man für die positive Potenz von ε in (3) zu zahlen hat, sind hoch oszillierende Einträge in der Systemmatrix $A(x)$. Dennoch eignet sich das AWP (3) wesentlich besser für die Numerik als das äquivalente System (1), da die Amplituden der Schwingungen nun von der Ordnung $\mathcal{O}(\varepsilon)$ sind.

Im ersten Teil der Arbeit werden Ideen aus [4, 54] für die "analytische Vorbearbeitung" von (1) aufgegriffen und auf Systeme von linearen Dgl des Typs (2) ausgedehnt. In § 3.3 wird gezeigt, dass es unter bestimmten Voraussetzungen möglich ist, das AWP (2) auf die Gestalt ($n \in \mathbb{N}$)

$$z'(x) \;=\; \varepsilon^n A_n(x) z(x)\,, \qquad z(a) \;=\; z_0 \in \mathbb{C}^d \tag{4}$$

zu transformieren. Hierbei sind die Oszillationen der Einträge von $A_n(x)$ von der gleichen Art wie in der Systemmatrix $A(x)$ von (3).

Die Lösung der stationären Sgl (1) besitzt eine asymptotische Entwicklung. Diese wird oft als WKB-Entwicklung[1] bezeichnet. Da (1) in ein äquivalentes System der Gestalt (2) überführt werden kann, ist es naheliegend, auch hierfür eine solche Entwicklung der Lösung zu suchen. In § 3.5 wird gezeigt, dass eine WKB-artige asymptotische Entwicklung der Lösung von (2) existiert. Die hierbei abgeleiteten expliziten Formeln lassen darüber hinaus den Zusammenhang der WKB-Entwicklung mit dem Transformationsansatz erkennen. Es stellt sich heraus, dass die dominanten Oszillationen im Wesentlichen dadurch eliminiert werden, dass man die WKB-Lösung „heraus dividiert".

Nachdem das Ausgangsproblem analytisch „aufbereitet" ist, werden spezielle Einschrittverfahren entwickelt, um das äquivalente AWP zu lösen. Hier werden zunächst endlich viele Schritte der Picard-Iteration verwendet, die zur approximativen Lösung von AWP genutzt werden kann. Da allerdings die Matrix $A_n(x)$ von (4) stark oszillierende Einträge hat, benötigen man spezielle Quadraturformeln, um die entstehenden hoch oszillierenden Integrale geeignet zu approximieren. Es wird kurz auf einige Methoden zur Berechung solcher Integrale eingegangen und eine detaillierte Fehleranalyse des verwendeten Ansatzes durchgeführt. Obwohl die benutzte Quadratur schon in der Literatur diskutiert wird (siehe [60, 61]), sind die hergeleiteten Fehlerabschätzungen neu. Die genannten Arbeiten behandeln hauptsächlich das asymptotischen Verhalten bezüglich des kleinen Parameters ε, sodass Abschätzungen des Quadraturfehlers in Bezug auf die Länge des Integrationsintervalls fehlen. Diese Lücke wird in § 5 geschlossen.

Die hergeleiteten Einschrittverfahren sind aufgrund der speziell gewählten Diskretisierungstechniken asymptotisch korrekt. Das heißt im Grenzfall $\varepsilon \to 0$ geht der Konvergenzfehler der numerischen Methoden gegen Null. Dabei kann in manchen Fällen die asymptotische Ordnung des Fehlers $\mathcal{O}(\varepsilon^{2n+1})$ betragen. Sowohl die Konvergenzfehler als auch die Quadraturfehler werden anhand von numerischen Beispielen veranschaulicht. Zudem werden Diskretisierungen des Transformationsansatzes und eine Schrittweitensteuerung diskutiert.

Im letzten Teil der Arbeit (§ 10) geht es um Ansätze für die Diskretisierung von Einweg-Wellengleichungen (Ewgl), die im Zusammenhang mit der skalaren Helmholtz-Gleichung (Hgl) stehen. Unter bestimmten Voraussetzungen ist eine Lösung der Ewgl eine Lösung der Hgl. Die Ewgl ist eine Evolutionsgleichung, deren „ortsabhängiger" Teil die Wurzel eines Differentialoperators beinhaltet. Dies führt zu Problemen bei der numerischen Behandlung. Es werden einige dieser Probleme kurz skizziert und anschließend einen Funktionalkalkül für selbstadjungierte Operatoren bewiesen, wie er in [71] entwickelt ist. Dieser Ansatz scheint gut zur numerischen Berechnung von Funktionen selbstadjungierter Operatoren geeignet zu sein. Die Grundlage dieses letzten Teils der Arbeit ist im Wesentlichen ein ausgearbeitetes Vorlesungsmanuskript des Autors. Somit ist dieser Abschnitt hauptsächlich als Literaturarbeit einzustufen. Trotzdem finden sich auch eigene Resultate des Autors, wie etwa eine Variante des Darstellungssatzes von Riesz oder Formeln für die Berechnung der Wurzel eines selbstadjungierten Operators und eine (formale) Lösungsformel für die Ewgl.

[1] Benannt nach den Physikern Wentzel, Kramers und Brillouin.

Contents

List of Figures

Chapter 1

Introduction

Highly oscillatory phenomena occur in a multiplicity of scientific models of very different fields, e. g. in electrodynamics, acoustics, molecular modeling, plasma transport, computer tomography, quantum mechanics. In this work we focus on the numerical solution of problems, which can be described by systems of linear ordinary differential equations. One famous representative of the class of equations we have in mind is the one–dimensional (scalar) *stationary Schrödinger equation* (SE) (1.1).

Let $\varepsilon > 0$ be a real constant and let $V\colon \mathbb{R} \to \mathbb{R}$ be a smooth function. On the interval $[a, b] \subset \mathbb{R}$ we consider the initial value problem (IVP)

$$\psi''(x) + \frac{1}{\varepsilon^2} V(x)\,\psi(x) \;=\; 0\,, \qquad (\psi(a), \psi'(a))^T \;=\; \psi_0 \in \mathbb{C}^2\,. \tag{1.1}$$

For a suitable choice of the parameter ε, the function V, and the initial condition ψ_0 the IVP (1.1) is a simple quantum mechanical model for a single electron in a stationary or scattering model (cf. [4]).

Is $V > 0$ constant, then $\psi(x) = c_1 \sin(\frac{\sqrt{V}}{\varepsilon}x) + c_2 \cos(\frac{\sqrt{V}}{\varepsilon}x)$ is the general solution of the ordinary differential equation (ODE) (1.1). Hence, for $\varepsilon \ll 1$ the solution ψ is a highly oscillatory function with amplitude of order $\mathcal{O}(1)$ with respect to ε. This character of the solution is preserved also for non constant $V \geq V_0 > 0$, with $V_0 \in \mathbb{R}^+$. The local wave length λ of the solution ψ is approximately proportional to $\frac{2\pi\varepsilon}{\sqrt{V}}$. The smaller ε gets, the more oscillates ψ. For $V < 0$ the solution ψ shows a (very) different behavior compared to the oscillatory part. In this regime we observe an (ε–dependent) exponential growth and decay instead of oscillations. In quantum mechanical models this phenomena is known as tunneling. Here, ψ is the solution of a stiff ODE. Hence a numerical solver for an arbitrary function V must be able to deal with high oscillations and (rapid) exponential growth and decay. The transition from one (growth) behavior to another takes place in the neighborhoods of the zeros of V. Due to this, these points are called turning point. An "optimal" integrator also has to be quite accurate in this parts. In this work we shall only focus on the oscillatory regime. The connection to the stiff part, where the solution can be computed with already existing methods, is dedicated to future work.

If one solves the IVP (1.1) with a standard solver, like the *classical Runge–Kutta method*, then one has to use spatial grids which resolve the oscillations. Otherwise the results are not reliable. In Figure 1.1 we see what can happen

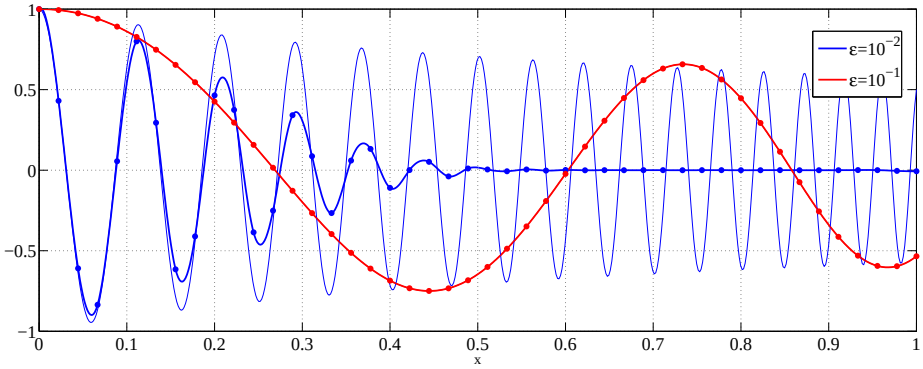

Figure 1.1: Exact solution (thin lines) of the IVP (1.1) for the function $V(x) = (x + \frac{1}{2})^2$ and the initial conditions $\psi(0) = 1, \psi'(0) = 0$ for $\varepsilon = 10^{-1}, 10^{-2}$. Furthermore the numerical approximation of the IVP with the classical Runge–Kutta method (dots) is plotted.

if one uses to few grid points. We plot the exact solution (thin lines) of the IVP (1.1) for the function $V(x) = (x + \frac{1}{2})^2$ and the initial conditions $\psi(0) = 1, \psi'(0) = 0$. Furthermore we see the numerical approximation of the IVP with the classical Runge–Kutta method (dots). For $\varepsilon = 10^{-1}$ there are enough points to resolve the oscillations and the (interpolated) result matches quite well with the exact solution. However this is not the case for $\varepsilon = 10^{-2}$. Here we use the same number of grid points as before. But this time there are to few abscissas to resolve the oscillations, which results in a totally wrong numerical solution.

To solve just a single IVP of type (1.1), even if it is highly oscillatory, is not a numerical challenge. One can of course use the standard solvers with a very fine grid. But in some applications, e. g. the modeling of semiconductor devices (cf. [4] for more details and references), one has to solve a large number of highly oscillatory systems in parallel or iteratively. This yields a tremendous numerical effort and hence one is interested in efficient solvers for (1.1) and related system.

It is well known that the solution of the stationary SE (1.1) has an *asymptotic expansion* (as $\varepsilon \to 0$), which is often called[1] WKB–expansion [32]. A basic idea to derive numerical solvers for (1.1) in the oscillatory regime (i. e. $V > 0$), is to use information of the asymptotic behavior of the solution as $\varepsilon \to 0$. In [57] the author suggests a *finite element method* (FEM), which uses elements built upon the WKB–approximation of ψ. Let $a = x_1 < x_2 < \cdots < x_N = b$ be the grid used for the FEM and let $h := \max\{x_{n+1} - x_n | n = 1, \ldots, N-1\}$. Then the (general, first order) WKB–approximation of ψ on the subinterval $[x_n, x_{n+1}]$ is given by ($c_{n,1}, c_{n,2} \in \mathbb{C}$)

$$\psi_n^{\mathrm{WKB}}(x) \quad := \quad \frac{1}{\sqrt[4]{V(x)}} \left(c_{n,1} e^{\frac{i}{\varepsilon}\varphi_n(x)} + c_{n,2} e^{-\frac{i}{\varepsilon}\varphi_n(x)} \right),$$

with $\varphi_n(x) := \int_{x_n}^{x} \sqrt{V(s)}\, ds$. It holds for all $x \in [x_n, x_{n+1}]$:

$$\psi_n^{\mathrm{WKB}}(x) \quad = \quad \psi_n^{\mathrm{WKB}}(x_n) w_n(x) + \psi_n^{\mathrm{WKB}}(x_{n+1}) v_n(x).$$

[1] Named after the physicists Wentzel, Kramers and Brillouin.

The functions w_n, v_n are defined as follows:

$$w_n(x) := \alpha_n(x)\sqrt[4]{\frac{V(x_n)}{V(x)}}, \qquad w_n(x) := \beta_n(x)\sqrt[4]{\frac{V(x_{n+1})}{V(x)}},$$
$$\alpha_n(x) := -\frac{\sin\frac{\varphi_{n+1}(x)}{\varepsilon}}{\sin\frac{\varphi_n(x_{n+1})}{\varepsilon}}, \qquad \beta_n(x) := \frac{\sin\frac{\varphi_n(x)}{\varepsilon}}{\sin\frac{\varphi_n(x_{n+1})}{\varepsilon}}.$$

"The functions α_n and β_n are the so–called WKB basis functions. They oscillate with a frequency close to that of the unknown wave function and actually permit solving the problem on coarser grids. In the limit $h \to 0$, these WKB basis functions reduce to usual linear interpolation functions" [57]. Due to construction of the WKB basis function we have to ensure, that $\frac{\varphi_n(x_{n+1})}{\varepsilon}$ does not get close to a multiple of π. This yields an (unnatural) artificial restriction for the grid. Despite the fact that the method works, this is the motivation for us to develop here a marching method for problem (1.1).

The IVP (1.1) can be transformed into an equivalent first order ODE system of type

$$u'(x) = \frac{1}{\varepsilon}L(x)u(x) + B(x)u(x), \qquad u(a) = u_0 \in \mathbb{C}^d, \tag{1.2}$$

with a real diagonal matrix $L(x) \in \mathbb{R}^{d\times d}$ and $B(x) \in \mathbb{C}^{d\times d}$ (cf. § 2.2). For the IVP (1.1) we have of course $d = 2$. As discussed in [4, 27, 39, 40, 54, 59] it is possible, with a further transformation, to eliminate the dominant oscillations with wave length $\sim \frac{1}{\varepsilon}$ and amplitude of order $\mathcal{O}(1)$ with respect to ε. Therefore one has to remove the negative powers of ε from the right–hand side of the ODE (1.2). For example, the resulting system from [4] has the form

$$z'(x) = \varepsilon A(x)z(x), \qquad z(a) = z_0 \in \mathbb{C}^d. \tag{1.3}$$

The price we have to pay for obtaining non negative power of ε on the right–hand side of (1.3) (also in the other approaches from literature) are highly oscillatory entries of the matrix valued function A. Nevertheless, the IVP (1.3) is much better suited for numerical treatment than the equivalent systems (1.1) or (1.2). The reason for this is that the solution z of (1.3) oscillates around the initial condition z_0 with amplitudes of order $\mathcal{O}(\varepsilon)$.

In the first part of this theses we shall seize the ideas from [4, 39, 40, 54] of the analytic preprocessing of problem (1.1) or (1.2). The article [4] only deals with the special case of the scalar IVP (1.1), but presents a transformation which results in a positive power of ε on the right–hand side of (1.3). On the other hand, in [39, 40, 54] the authors discuss a vector valued version of (1.1), but the used transformation approach only removes the negative powers of ε and yields a right–hand side that is $\mathcal{O}(1)$ with respect to ε. Thus the asymptotic behavior with respect to ε is not as accurate as in [4]. In § 3.3 we shall prove that it is possible (under certain assumptions) to transform the IVP (1.2) into

$$z'(x) = \varepsilon^n A_n(x)z(x), \qquad z(a) = z_0 \in \mathbb{C}^d, \tag{1.4}$$

with some $n \in \mathbb{N}$. Also $A_n(x)$ has highly oscillatory entries, which are of the same frequency as those in $A(x)$ of (1.3). We shall prove that our transformation

ansatz is an extension of the ansatz presented in [4, 54]. It combines and extends the ideas of both articles. The transformation ansatz in [27] is called *super–adiabatic transformation*. It is built on a similar strategy/philosophy as our approach and yields comparable results. In § 3.3.2 we shall compare it to our approach and point out the differences. The preprocessing discussed in [59] seems to be similar to our lowest order transformation.

Since (1.1) is equivalent to a system of type (1.2), it is quite natural to search also for an asymptotic expansion of the solution of (1.2). In § 3.5 we shall show that a WKB–type asymptotic expansion of the solution u exists. The derived, explicit formulas also reveal the connection between the WKB–expansion and the transformation approach from § 3.3. Let U_{WKB} be a WKB–type approximation of a fundamental system of solutions of (1.2). It turns out, that the dominant oscillations of the exact solution u from (1.2) are (essentially) removed by multiplying u with the (pointwise) matrix inverse of U_{WKB}.

Once the analytic preprocessing is finished, we shall start deriving specially designed one–step methods to solve the equivalent IVP of type (1.4). At first we make a finite number of *Picard iterations*, which yields an approximate (analytic) solution of the IVP. This is also the basis for the methods in [4, 27, 54]. Afterwards we have to find a suitable discretization of the derived expression. Since A_n from (1.4) is a matrix valued function with highly oscillatory entries, we need special quadratures to approximate the occurring highly oscillatory integrals of the form

$$I \quad := \quad \int_{\alpha}^{\beta} f(x) e^{-\frac{i}{\varepsilon}\varphi(x)}\, dx\,. \tag{1.5}$$

Here f, φ are smooth, real valued functions with $|\varphi'| \geq \delta > 0$. In [27, 54] integrals of type (1.5) are approximated by replacing f, φ by their Taylor approximations up to a certain order. Since this procedure does not take into account the asymptotic nature of I as $\varepsilon \to 0$, the method is not very efficient compared to most quadratures discussed in § 5. A more sophisticated method is used for the schemes in [4], the so called *shifted asymptotic method*. The presented quadrature is based on the *asymptotic method*, which can be found in [38, 37]. Essentially, it describes the asymptotic expansion of I as $\varepsilon \to 0$. We continue the work from [4] and derive an even improved version of the shifted asymptotic method. Nevertheless, it is still slightly less efficient than the method, we finally use for our one–step methods. Our quadrature, which we shall call *modified Filon–type method*, can be extracted from [60, 61]. It uses an interpolation approach for f, such that the benefit of the asymptotic method, i. e. the asymptotically correctness of the quadrature as $\varepsilon \to 0$, is gained. Since the focus of these articles is on the behavior of I (respectively the quadrature) as $\varepsilon \to 0$, there is no error analysis which takes the length of the integration interval into account. This gap shall be closed in § 5.

The combination of analytic preprocessing (WKB–type transformation) and sophisticated quadratures for highly oscillatory integrals yields a "zoo" of one–step methods, which are asymptotically correct as $\varepsilon \to 0$. I. e., even on a fixed spatial grid we obtain the right limit as $\varepsilon \to 0$. Here we can observe (for certain methods) a convergence error of order $\mathcal{O}(\varepsilon^{2n+1})$. In the literature we find this desirable feature for the methods discussed in [4]. The integrators constructed in [39, 54, 27] only show uniform (spatial) error estimates with respect to ε, but the error is $\mathcal{O}(1)$ as $\varepsilon \to 0$.

This thesis is organized as follows. In chapter 2 we start with two examples, which originate from quantum mechanical models of one dimensional semiconductors. We further discuss a procedure to transform (a certain class of) singularly perturbed second order IVP to first order systems of type (1.2). It is a slight modification of an approach from [54], which is also presented. The goal of the following chapter § 3 is the derivation of the analytic preprocessing of (1.2), see § 3.3. Beside this, in § 3.5 we also derive a WKB–type asymptotic approximation for the solution of our model problem from § 3.2. Furthermore we discuss the connection of the derived WKB–type transformation from § 3.3 with the approach from [4] and the super–adiabatic transformation from [27].

The basic idea for the numerical integration procedure of (1.2) is, at first, to apply the WKB–type transformation and afterwards use the specially designed integrators from § 6. Thus we discuss in chapter 4 a way to (numerically) approximate the analytic WKB–transformation. We shall not discuss it in its most general form. We rather fix one set of parameters, which are used for the numerical experiments in § 7, and derive a discretization for an equidistant and non–equidistant spatial grid. Additionally we give some remarks on the error inflicted by the discretized transformation.

If the eigenvalues of L from (1.2) approach each other but do not cross (avoided eigenvalue crossing), or if the norm of B gets very large compared to those of $\frac{1}{\varepsilon}L$, then the one–step methods may yield poor results. In this situation it is necessary to use a step size control strategy. One such algorithm shall be derive in § 4.4. It is motivated by a similar approach from [27].

Chapter 5 is devoted to the approximation of highly oscillatory integrals of type (1.5). We start with a brief review of some method in § 5.1. For more methods and references we refer to the review article [35]. Afterwards, in § 5.2 we specify the quadrature we shall use for the one–step methods and derive error estimates. Since the quadrature can be transformed such that is looks like the *Filon–type method* from [38, 37], we shall call it *modified Filon–type method*. Then in § 5.3 we review the shifted asymptotic method from [4] and derive an improved version of it. The performance of the modified Filon–type method and the shifted asymptotic methods are illustrated in § 5.4 by some numerical experiments.

In chapter 6 we derive and discuss the one–step method for equations of type (1.4). We shall give a detailed analysis of the local error and prove convergence of the schemes as the (maximum) spatial step size tends to zero. The theoretical results are compared with numerical experiments in chapter 7. Chapter 8 contains a collection of technical results from the author, which are used in different parts of the thesis.

The last chapter 10 has an exceptional position in the whole thesis. Here we (briefly) discuss ideas to discretize (special) one way wave equations (OWWE). This equations are connected with the (scalar) Helmholtz equation (HE). In certain situations the solution of the OWWE is also a solutions of the HE. Hence it is of interest to have numerical solvers for them. The OWWE is an evolution equation, whose "spatially dependent part" contains a square root of a partial differential operator. This causes some problems for the numerical treatment. We shall discuss some of the problems in § 10.2. In the following section § 10.3 we introduce a functional calculus for self–adjoint operators. It is a non standard approach from [71], which is based on the following idea. Let

f be a Schwartz function (rapidly decreasing function) and let $\widehat{f}$ be its *Fourier transform*. Then it holds for all $a \in \mathbb{R}$:

$$f(a) \;=\; \tfrac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \widehat{f}(\xi)\, u(\xi)\; dx\,, \qquad u(\xi) \;:=\; e^{i\xi a}\,,$$

The oscillatory part u of the integrand solves the IVP $u' \;=\; ia\, u$, $u(0) \;=\; 1$. Hence (formally) replacing a by a self–adjoint operator A and u by the unitary group U, with

$$U' \;=\; iA\, U\,, \qquad U(0) \;=\; \mathrm{Id}\,,$$

yields a formula for $f(A)$. The functional calculus may be well suited for the numerical computation of functions of self–adjoint operator. We give a (more or less) self–consistent proof. Therefore we collected and added a lot of results from literature, which are only mentioned in [71]. For the proof we also derive a version of Riesz' representation theorem, which we were not able to find in literature. The same holds for the main result from § 10.3.4. In § 10.4 we use the results from the previous section § 10.3 to (formally) derive explicit formulas for the square root of an self–adjoint operator and a solution of the OWWE. Furthermore we deduce from our (formal) solution formula a well known transformation, which connects the solution of the HE with the solution of a Schrödinger type equation.

Chapter 2

Highly oscillatory ODEs in application

In this chapter we introduce and briefly discuss some examples of highly oscillatory problems. Here the focus lies on the application to quantum mechanics. In § 2.1 we introduce the Kane–model and $k \cdot p$–model (two simple multi–band approaches for 1D semiconductors), along with their open boundary conditions that are needed in quantum transport applications. These boundary value problems are then transformed into equivalent initial value problems. The strategy of transforming the governing equation (2.10) for the two–band $k \cdot p$ model from § 2.1.2 to a first order system is generalized in § 2.2.

There are of course much more applications and problems. For example in [39, 40] the authors discuss integrators for singularly perturbed Schrödinger equations, where the time dependent Hamiltonian is modeled a by a finite-dimensional real symmetric matrix. This type of equations "arise as a computationally critical subproblem in mixed quantum-classical models of molecular dynamics [...], in the quantum–classical Liouville equation [...], or in the equations known as Ehrenfest or QCMD (quantum–classical molecular dynamics) model [...]" [40]. For more details we refer to the references cited in [39, 40].

In § 2.3 we derive a method to compute and store the Greens function for linear second order boundary value problems in an efficient way.

2.1 Two–band Schrödinger models

"For several novel semiconductor devices (like resonant interband tunneling diodes, see [55, 45]) single–band effective mass models become insufficient to simulate the quantum transport through such a device. Hence, it is getting ever more important to include realistic band structures in quantum transport models. In this section we shall discuss two independent, stationary two–band Schrödinger models (Kane–model and two–band $k \cdot p$–model) that are used for simulations of one dimensional semiconductor devices like a resonant tunnel diode. We assume that the considered semiconductor structure occupies the interval $[a, b]$ and is connected to reservoirs at $x = a$ and $x = b$. We also assume that both reservoirs are homogeneous and extend to $x = \pm\infty$. So all material (and energy) parameters are constant in each reservoir, which is hence

populated only by traveling plane waves.

In analogy to [4] we shall discuss in this paper only the numerically challenging oscillatory regime of traveling waves. The evanescent regime with tunneling is of course equally important for quantum transport, but the numerical treatment of those smooth wave functions is simpler (and need different tools), and will not be discussed here. However, WKB-based discretizations of the coupled oscillatory–evanescent situation are currently under investigation, and will be the topic of an upcoming work. In [46], transparent boundary conditions (TBCs) for the Kane–model and two–band $k \cdot p$–model were derived, as well as discrete TBCs for finite difference schemes. However, such classical schemes are numerically expensive in the oscillatory case. So it is the goal of this work to develop a more efficient alternative." [25]

2.1.1 The Two–Band Kane–Model

A simple multi–band approach is the two–band Kane-model (cf. [43, 46]). This is an inter-band model, describing the coupled electron transport in the conduction and the valence bands. Here the vector valued "wave function" $\psi(x) \in \mathbb{C}^2$ is a solution of the ODE

$$\mathbb{H}\psi \quad = \quad E\psi \tag{2.1}$$

with

$$\mathbb{H} \quad = \quad \begin{pmatrix} V(x) & -i\varepsilon p(x)\frac{d}{dx} \\ -i\varepsilon p(x)\frac{d}{dx} & V(x) - E_g(x) \end{pmatrix}.$$

We denote by $E > 0$ the prescribed (constant-in-x) injection energy of the electrons. Here, the potential V is the band edge of the conduction band, and $E_g > 0$ is the (x-dependent) band gap between the conduction and the valence band. The function $p > 0$ is related to the Kane-parameter. Furthermore, the real parameter $\varepsilon > 0$ is a small constant, which is often (depending on the model) proportional to the reduced Planck constant $\hbar$. Hence ε shall be assumed to be very small. The dispersion relation of this Kane model is discussed in detail in §3.1 of [46].

We want to model a finite semiconductor device on the bounded interval $[a, b]$. In order to have unique solutions (if we prescribe initial conditions) we assume:

Assumption 1. *The functions V, E_g, and p are continuous on $\mathbb{R}$ and continuously differentiable on $[a, b]$. Furthermore they are constant on the exterior domains $(-\infty, a]$ and $[b, \infty)$.*

Remark 2.1.1. *To ensure unique solvability of the IVP we could establish weaker assumptions, for example piecewise[1] Lipschitz continuity. Since we shall transform the derived IVP, such that it fits in a more general framework, we need more regularity.*

[1] Here, piecewise Lipschitz continuous means that there are finitely many (non trivial) pairwise disjoint intervals $I_1, \ldots, I_n$ such that $[a, b] \subset \bigcup_{j=1}^{n} I_j$ and the functions are globally Lipschitz continuous on each single interval $I_1, \ldots, I_n$.

We rewrite the ODE (2.1) in the more convenient form

$$\psi'(x) \;=\; \frac{i}{\varepsilon}\begin{pmatrix} 0 & \alpha(x) \\ \beta(x) & 0 \end{pmatrix}\psi(x)\,, \tag{2.2}$$

with

$$\alpha(x) \;:=\; \frac{E-V(x)+E_g(x)}{p(x)} \quad \text{and} \quad \beta(x) \;:=\; \frac{E-V(x)}{p(x)}\,.$$

We shall consider here a scattering model, subject to a given, incoming plane wave. Hence, we shall need transparent boundary conditions (TBCs) at both (artificial) boundary points a, b (as derived in §3.2 of [46]). It will be an inhomogeneous TBC at the influx boundary of the quantum structure, and a homogeneous TBC on the opposite side. Let us denote the system matrix of (2.2) by A. Hence the eigenvalues of $A(x)$ are given by

$$\begin{aligned}\lambda(x) \;&=\; \pm\frac{i}{\varepsilon}\sqrt{\alpha(x)\beta(x)} \\ &=\; \pm\frac{i}{\varepsilon p(x)}\sqrt{(E-V(x)+E_g(x))(E-V(x))} \;=:\; \pm\, i\,k(x)\,.\end{aligned}$$

If the given injection energy is larger than the conduction band edge, i. e. $E - V(x) > 0$ holds on the whole interval $[a,b]$ (and thus on the whole real line), the eigenvalues $\lambda = \pm ik$ are purely imaginary and hence we only have traveling waves[2]. Let $v_\pm(x)$ be (real valued) eigenvectors corresponding to the eigenvalues $\lambda(x) = \pm ik(x)$. Then the general solution ψ of (2.2) in the exterior domains reads

$$\psi_a(x) \;=\; c_a e^{ik(a)(x-a)}v_+(a) \;+\; d_a e^{-ik(a)(x-a)}v_-(a) \tag{2.3}$$

for $x \le a$ and

$$\psi_b(x) \;=\; c_b e^{ik(b)(x-b)}v_+(b) \;+\; d_b e^{-ik(b)(x-b)}v_-(b) \tag{2.4}$$

for $x \ge b$ with constants $c_a, c_b, d_a, d_b \in \mathbb{R}$. From the right exterior domain $[b,\infty)$ we now inject a left traveling electron wave with prescribed amplitude $d_b \neq 0$. Since it has to pass the semiconductor structure (in general) a part of the wave is reflected. Thus (we expect) $c_b \neq 0$ holds too. Once (a part of) the electron has passed the semiconductor and reached the constant regime $(-\infty, a]$, it will not be reflected there. Hence there is no right traveling part of the wave at $(-\infty, a]$, i. e. $c_a = 0$. This yields

$$\psi_a(x) \;=\; d_a e^{ik(a)(x-a)}\, v_-(a)\,.$$

We denote the solution in the interior domain $[a,b]$ simply by ψ. Due to Assumption 1 the (global) solution on the whole real line is continuously differentiable in certain open neighborhoods of the boundary points a, b. Hence we get the homogeneous boundary condition[3]

$$\begin{aligned}\psi(a) \;=\; \psi_a(a) \in \operatorname{span}(v_-(a)) \;&\Leftrightarrow\; \big(A(a) + ik(a)\,\mathrm{Id}\,\big)\psi(a) \;=\; 0 \\ &\Leftrightarrow\; \psi'(a) + ik(a)\psi(a) \;=\; 0\,.\end{aligned}$$

[2] One also gets purely imaginary eigenvalues if the energy is smaller than the valence band edge, i. e. $E - V(x) + E_g(x) < 0$. This energy corresponds to holes in the valence band, and the situation would be analogous to the case discussed here.

[3] $\operatorname{span}(v)$ denotes the vector space spanned by v.

At the right boundary we combine the first derivative ψ' with ψ to eliminate a the reflection constant c_b. From (2.4) we obtain the inhomogeneous right TBC

$$\begin{aligned}\psi'(b) - i\,k(b)\psi(b) &= ik(b)\big(c_b v_+(b) - d_b v_-(b)\big)\\ &\quad - ik(b)\big(c_b v_+(b) + d_b v_-(b)\big)\\ &= -2i\,k(b)\,d_b\,v_-(b)\,.\end{aligned}$$

Thus our BVP with TBC for a left traveling plane wave coming from the exterior domain $[b,\infty)$ with prescribed "amplitude" $d_b \neq 0$ reads

$$\begin{aligned}\psi'(x) - A(x)\psi(x) &= 0\,, \qquad x\in[a,b]\,, &(2.5)\\ \psi'(a) + ik(a)\,\mathrm{Id}\,\psi(a) &= 0 &(2.6)\\ \psi'(b) - ik(b)\,\mathrm{Id}\,\psi(b) &= -2ik(b)d_b v_-(b)\,. &(2.7)\end{aligned}$$

In §3.3 of [46] existence and uniqueness of a solution ψ of (2.5)–(2.7) is discussed. Thus we state without a proof

Lemma 2.1.2. *The BVP* (2.5)–(2.7) *has a unique solution in* $\big(W^{2,\infty}(a,b)\big)^2$.

Remark 2.1.3. *Since the function* ψ *satisfies the ODE* (2.5) *even on the boundary, we can use it to reformulate the boundary conditions* (2.6), (2.7)*: It holds*

$$\begin{aligned}\big(A(a) + ik(a)\,\mathrm{Id}\,\big)\psi(a) &= 0\,,\\ \big(A(b) - ik(b)\,\mathrm{Id}\,\big)\psi(b) &= -2ik(b)d_b v_-(b)\,.\end{aligned}$$

We shall now reformulate the BVP (2.5)–(2.7) as an IVP, which is easier to solve numerically (particularly for our highly oscillatory regime). To this end we first solve (using the left boundary condition) the IVP

$$\begin{aligned}\psi_-'(x) &= A(x)\psi_-(x)\,, \qquad x\in[a,b]\,, &(2.8)\\ \psi_-(a) &= v_-(a)\in\mathbb{C}^2\,.\end{aligned}$$

Due to Assumption 1, any IVP of ODE (2.5) is uniquely solvable. Since ψ has to fulfill (2.6) (which is equivalent to $\psi(a)\in\mathrm{span}(v_-(a))$), there exists a constant $c_-\in\mathbb{R}$ such that $\psi(a) = c_- v_-(a)$. Hence we get $\psi = c_-\psi_-$. To determine the unknown factor c_- we use the remaining boundary condition (2.6). From Remark 2.1.3 we get

$$c_-\,\big(A(b) - ik(b)\,\mathrm{Id}\,\big)\psi_-(b) = -2ik(b)d_b v_-(b)\,.$$

The inner product of this equation with $v_-(b)$ yields

$$c_- = \frac{-2ik(b)d_b\|v_-(b)\|^2}{v_-(b)^T\big(A(b) - ik(b)\,\mathrm{Id}\,\big)\psi_-(b)} = \frac{-2ik(b)d_b\|v_-(b)\|^2}{v_-(b)^T\big(\psi_-'(b) - ik(b)\psi_-(b)\big)}\,.$$

Analogously we get for a right traveling plane wave in the left exterior domain $(-\infty,a]$ with prescribed "amplitude" $c_a\neq 0$:

$$\begin{aligned}\psi'(x) - A(x)\psi(x) &= 0\\ \psi'(a) + ik(a)\,\mathrm{Id}\,\psi(a) &= 2ik(a)c_a v_+(a)\\ \psi'(b) - ik(b)\,\mathrm{Id}\,\psi(b) &= 0\,.\end{aligned}$$

It holds $\psi = c_+\psi_+$, where ψ_+ solves

$$\psi_+'(x) \;=\; A(x)\psi_+(x)\,, \qquad \psi_+(b) \;=\; v_+(b) \in \mathbb{C}^2\,.$$

and

$$c_+ \;=\; \frac{2ik(a)c_a\|v_+(a)\|^2}{v_+(a)^T\big(A(a)+ik(a)\,\mathrm{Id}\,\big)\psi_+(a)} \;=\; \frac{2ik(a)c_a\|v_+(a)\|^2}{v_+(a)^T\big(\psi_+'(a)+ik(a)\psi_+(a)\big)}\,.$$

Recall from (2.2) that the system matrix $A(x)$ is proportional to $\frac{1}{\varepsilon}$ and off-diagonal. We now aim to transform out the dominant oscillations in the IVP (2.8). Therefore we have to diagonalize the matrix function A. A simple calculation yields

$$v_+(x) \;=\; \begin{pmatrix} \sqrt{\alpha(x)} \\ \sqrt{\beta(x)} \end{pmatrix} \quad\text{and}\quad v_-(x) \;=\; \begin{pmatrix} -\sqrt{\alpha(x)} \\ \sqrt{\beta(x)} \end{pmatrix}.$$

Thus we have $A(x) = \frac{i}{\varepsilon}Q(x)^{-1}L(x)Q(x)$ with

$$Q(x)^{-1} \;=\; \begin{pmatrix} \sqrt{\alpha(x)} & -\sqrt{\alpha(x)} \\ \sqrt{\beta(x)} & \sqrt{\beta(x)} \end{pmatrix} \quad\text{and}\quad L(x) \;=\; \begin{pmatrix} \varepsilon k(x) & 0 \\ 0 & -\varepsilon k(x) \end{pmatrix}.$$

Note that the matrix valued functions Q and L are actually independent of ε. Since it holds

$$Qv_- \;=\; \frac{1}{2\sqrt{\alpha\beta}}\begin{pmatrix} \sqrt{\beta} & \sqrt{\alpha} \\ -\sqrt{\beta} & \sqrt{\alpha} \end{pmatrix}\begin{pmatrix} -\sqrt{\alpha} \\ \sqrt{\beta} \end{pmatrix} \;=\; \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

the ansatz $u := Q\psi_-$ yields the IVP

$$\begin{aligned} u'(x) \;&=\; \frac{i}{\varepsilon}L(x)u(x) \,+\, B(x)u(x)\,, \qquad x \in [a,b]\,, \\ u(a) \;&=\; \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \end{aligned} \tag{2.9}$$

with $B := Q'Q^{-1}$. The same transformation also works for the other case of a right traveling plane wave with prescribed amplitude c_a in $(-\infty, a]$. We only have to replace the initial condition by $u(b) = (1,0)^T$.

2.1.2 The two–band $k\cdot p$–model

In this section we discuss a slightly more involved inter-band model for the coupled quantum transport in the conduction and the valence bands (cf. [5, 45]). A different inter-band $k\cdot p$–model is analyzed in §4 of [46]. And for extended multi–band $k\cdot p$–models (including the intra-band coupling of heavy and light holes within the valence band) we refer to [42], §6 of [45], and §5 of [46]. In our two–band model, the "wave function" $\psi = (\psi_1, \psi_2)^T \in \mathbb{C}^2$ solves a 2×2 Schrödinger boundary value problem

$$\mathbb{H}(x)\psi(x) \;=\; E\,\psi(x)\,, \qquad x \in (a,b) \tag{2.10}$$

$$\psi'(a) - K_a(E)\psi(a) \;=\; 0 \tag{2.11}$$

$$\psi'(b) - K_b(E)\psi(b) \;=\; r \in \mathbb{C}^2\,, \tag{2.12}$$

with the Hamiltonian

$$\mathbb{H} \;:=\; -\varepsilon^2 \frac{\partial^2}{\partial x^2} + \varepsilon\, P(x)\frac{\partial}{\partial x} + \begin{pmatrix} V(x) & 0 \\ 0 & V(x) - E_g(x) \end{pmatrix}$$

and

$$P(x) \;:=\; \begin{pmatrix} 0 & ip(x) \\ ip(x) & 0 \end{pmatrix}.$$

The real parameter $\varepsilon > 0$ is a small constant, which is often (depending on the model) proportional to the reduced Planck constant $\hbar$. By E we denote the given injection energy of the particles. The potential $V(x)$ is the band edge of the conduction band, and $E_g(x) > 0$ is the energy gap between the conduction and the (light hole) valence bands. Further, $p(x) > 0$ is the coupling coefficient (related to the Kane-parameter) between the two bands. We remark that (2.10)-(2.12) is a scattering problem with given E, and *not* an eigenvalue problem. As for the Kane model in § 2.1.1 Assumption 1 should hold.

The matrices K_a, $K_b \in \mathbb{C}^{2\times 2}$ and the vector r in (2.11), (2.12) constitute the TBCs for the two–band $k \cdot p$–model. (2.12) models the injection of plane waves at $x = b$. Their derivation follows the same strategy as for the Kane model in § 2.1.1. But for the more involved details we refer to [5].

The self-consistent potential V is the solution of the following Poisson problem:

$$\begin{aligned} V''(x) &= n(x)\,, \qquad x \in (a,b)\,, \\ V(a) &= V_1 > 0\,, \\ V(b) &= 0\,. \end{aligned}$$

The charge density n is defined by

$$n(x) \;=\; \int_0^\infty f(E)\,|\psi(x,E)|^2\,dE\,,$$

for a prescribed function f that models the semiconductor statistics. Well-posedness of this nonlinear BVP is established in Th. 2.2 of [5].

If one is interested in a numerical approximation of $n(x)$ or of the current density

$$j(x) \;=\; \int_0^\infty f(E)\,\big(-\operatorname{Im}\langle\psi',\psi\rangle + 2p\operatorname{Re}(\overline{\psi}_1\psi_2)\big)(x,E)\,dE\,,$$

one has to use an iterative scheme, like the Gummel method, to solve the nonlinear problem. In each step one has to calculate a suitable approximation for the charge density n, and hence one has to solve (2.10) for a large number of (discrete) energies. Since $0 < \varepsilon \ll 1$ is a small constant the wave function ψ is highly oscillatory for $E > \|V\|_\infty$. In order to speed up the calculations it is very useful to have a numerical scheme that produces an accurate approximation for ψ, without having to resolve all oscillations of the wave function.

It is often more convenient to solve, instead of a BVP, an equivalent initial value problem. As we will see in a moment, it is possible to determine the solution ψ of the BVP (2.10)–(2.12) from just one matrix valued IVP-solution.

Due to Assumption 1, the functions p, V and E_g are Lipschitz continuous on $[a,b]$. Hence the IVP [4] for a matrix-valued wave function $\Phi(x) \in \mathbb{C}^{2\times 2}$,

$$\begin{aligned} \mathbb{H}(x)\Phi(x) &= E\Phi(x)\,, \qquad x \in (a,b), && (2.13)\\ \Phi(a) &= \mathrm{Id}, && (2.14)\\ \Phi'(a) &= K_a\,, && (2.15) \end{aligned}$$

has a unique solution. Further, it holds for every vector valued solution ϕ of

$$\begin{aligned} \mathbb{H}(x)\phi(x) &= E\phi(x)\,, \qquad x \in (a,b), && (2.16)\\ \phi'(a) - K_a\phi(a) &= 0\,, \end{aligned}$$

that $\phi(x) = \Phi(x)\phi(a)$. Since the solution ψ of the BVP (2.10)–(2.12) solves (2.16), we can write $\psi(x) = \Phi(x)\psi(a)$. Hence we get from the remaining right boundary condition (2.12)

$$\big(\Phi'(b) - K_b\Phi(b)\big)\psi(a) \;=\; r\,.$$

Since the BVP (2.10)–(2.12) is well-posed (cf. [45], Prop. 2.1), the above equation has a unique solution which yields $\psi(a)$ and consequently ψ.

As we have seen, the solution ψ of the BVP (2.10)–(2.12) is (uniquely) determined by the solution Φ of the IVP (2.13)–(2.15). Thus in the sequel we shall derive and discuss an efficient numerical method to solve the IVP for the (vector valued) equation (2.16). Since each column of Φ is a solution of (2.13) we further restrict ourself to vector valued solutions, in order to simplify notation. All results derived below also hold for matrix valued solutions.

The ODE of the IVP (2.13) Equation (2.16) for $\phi \in \mathbb{C}^2$ takes the form

$$\varepsilon^2\phi'' - \varepsilon P(x)\phi' + A(x)\phi \;=\; 0\,, \qquad (2.17)$$

with $A(x) := \mathrm{diag}(E - V(x), E - V(x) + E_g(x))$. For $E > \|V\|_\infty$ the matrix $A(x)$ is positive definite for all $x \in [a,b]$. Now we want to rewrite (2.17) as a first order IVP, with the same form as (2.9). To this end we set

$$v_1 := A^{\frac12}\phi\,, \qquad v_2 := \varepsilon\phi'\,,$$

which yields for $v(x) = (v_1(x),\, v_2(x))^T \in \mathbb{C}^4$:

$$\begin{aligned} v' &= \frac{1}{\varepsilon}\begin{pmatrix} 0 & A^{\frac12} \\ -A^{\frac12} & P \end{pmatrix} v + \begin{pmatrix} {A^{\frac12}}' A^{-\frac12} & 0 \\ 0 & 0 \end{pmatrix} v\,, && (2.18)\\ v(a) &= \begin{pmatrix} A^{\frac12}(a) \\ \varepsilon K_a \end{pmatrix}. \end{aligned}$$

Let us denote by $\tilde{L}$ the first matrix of (2.18). Since $P(x)$ is skew–symmetric for all $x \in [a,b]$, the same holds for $\tilde{L}$. Hence there exists a matrix function $Q\colon [a,b] \to \mathbb{C}^{4\times 4}$, such that for all $x \in [a,b]$ it holds (cf. § 2.1.2)

$$\tilde{L}(x) \;=\; iQ^*(x)L(x)Q(x)\,,$$

[4] Simply rewrite the IVP as a first order IVP to prove existence and uniqueness of Φ.

with $L(x)$ real and diagonal. Finally we set

$$u(x) \;:=\; Q(x)\,v(x) \in \mathbb{C}^4\,,$$

which yields

$$u' = \frac{i}{\varepsilon} Lu + Bu\,, \qquad u(a) = u_0\,, \tag{2.19}$$

with

$$B(x) \;=\; Q'(x)Q^*(x) + Q(x) \begin{pmatrix} A^{\frac{1}{2}\prime}(x)A^{-\frac{1}{2}}(x) & 0 \\ 0 & 0 \end{pmatrix} Q^*(x)\,.$$

Of course, the above transformation procedure is not limited to the special case (2.10). One can apply it to any ODE of type (2.17) with $P(x)$ skew–symmetric and $A(x)$ positive definite (see § 2.2).

The matrix Q for the two–band $k \cdot p$–model

For the two band $k \cdot p$–model we can explicitly compute the transformation Q and the eigenvalues of $\tilde{L}$ (cf. (2.18)). The matrix $\tilde{L}$ is given by

$$\tilde{L} \;=\; \begin{pmatrix} 0 & 0 & \sqrt{E_1} & 0 \\ 0 & 0 & 0 & \sqrt{E_2} \\ -\sqrt{E_1} & 0 & 0 & ip \\ 0 & -\sqrt{E_2} & ip & 0 \end{pmatrix},$$

where we set $E_1 = E - V$ and $E_2 = E - V + E_g$. We use *Maple 14* to derive the characteristic polynomial χ and get

$$\chi(\tilde{\lambda}) \;=\; \tilde{\lambda}^4 \,+\, (p^2 + E_1 + E_2)\tilde{\lambda}^2 \,+\, E_1E_2\,.$$

Hence the eigenvalues are

$$\tilde{\lambda} \;=\; \pm\frac{i}{\sqrt{2}}\sqrt{p^2 + E_1 + E_2 \pm \sqrt{(p^2 + E_1 + E_2)^2 - 4E_1E_2}}\,.$$

Again with *Maple 14* we compute a corresponding eigenvector v_λ for the eigenvalue λ, which is

$$v_{\tilde{\lambda}} \;=\; \left(-\frac{i\sqrt{E_1}pE_2}{\tilde{\lambda}^2(\tilde{\lambda}^2 + p^2 + E_1)}\,,\; \frac{\sqrt{E_2}}{\tilde{\lambda}}\,,\; -\frac{ipE_2}{\tilde{\lambda}(\tilde{\lambda}^2 + p^2 + E_1)}\,,\; 1 \right)^T.$$

Since $\tilde{\lambda}$ is a root of χ we get

$$\tilde{\lambda}^2(\tilde{\lambda}^2 + p^2 + E_1) \;=\; -E_2(\tilde{\lambda}^2 + E_1)\,,$$

which yields

$$v_{\tilde{\lambda}} \;=\; \left(\frac{ip\sqrt{E_1}}{\tilde{\lambda}^2 + E_1}\,,\; \frac{\sqrt{E_2}}{\tilde{\lambda}}\,,\; \frac{ip\tilde{\lambda}}{\tilde{\lambda}^2 + E_1}\,,\; 1 \right)^T.$$

Let $i\lambda_1, \ldots, i\lambda_4$ be the four pairwise different eigenvalues of $\tilde{L}$, e. g.

$$\begin{aligned}
\lambda_1 &:= \frac{1}{\sqrt{2}}\sqrt{p^2 + E_1 + E_2 + \sqrt{(p^2 + E_1 + E_2)^2 - 4E_1E_2}} \\
\lambda_2 &:= \frac{1}{\sqrt{2}}\sqrt{p^2 + E_1 + E_2 - \sqrt{(p^2 + E_1 + E_2)^2 - 4E_1E_2}} \\
\lambda_3 &:= -\frac{1}{\sqrt{2}}\sqrt{p^2 + E_1 + E_2 + \sqrt{(p^2 + E_1 + E_2)^2 - 4E_1E_2}} \\
\lambda_4 &:= -\frac{1}{\sqrt{2}}\sqrt{p^2 + E_1 + E_2 - \sqrt{(p^2 + E_1 + E_2)^2 - 4E_1E_2}}
\end{aligned}$$

and let $v_1, \ldots, v_4$ be the corresponding eigenvectors, i. e.$v_j = v_{i\lambda_j}$. Then it holds

$$\tilde{L} \;=\; i\,Q^*LQ\,,$$

with $L = \operatorname{diag}(\lambda_1, \ldots, \lambda_4)$ and

$$Q \;=\; \left(\frac{v_1}{\|v_1\|}, \ldots, \frac{v_4}{\|v_4\|}\right)^*.$$

2.2 Singularly perturbed second order ODE

Let $\varepsilon_0 > 0$ and let $[a, b] \subset \mathbb{R}$ be a non–trivial bounded interval with $x_0 \in [a, b]$. Further let $\Omega := [a, b] \times (0, \varepsilon_0)$ be the domain of the matrix valued functions $A, P\colon \Omega \to \mathbb{C}^{d\times d}$ and the vector valued function $g\colon \Omega \to \mathbb{C}^d$. We assume that $P(x, \varepsilon)$ is skew symmetric and $A(x, \varepsilon)$ is positive definite for all $(x, \varepsilon) \in \Omega$. Additionally we assume that A is uniformly coercive on its domain Ω, i. e. there exists a constant $c_A > 0$ such that it holds for all $(x, \varepsilon) \in \Omega$:

$$v^*A(x, \varepsilon)v \;\geq\; c_A v^*v \quad \text{for all} \quad v \in \mathbb{C}^d\,.$$

The reformulated Kane– and two–band $k \cdot p$–model from § 2.1.1 and § 2.1.2 respectively are the motivation for our Model Problem 1 from § 3.2. As announced in the end of § 2.1.2 the transformation procedure used for the two–band $k \cdot p$–model is not limited to that special case. In this section we extend the ansatz to the larger class of singularly perturbed (vector valued) second order IVP of the form

$$\begin{aligned}
\varepsilon^2\psi'' - \varepsilon P\psi' + A\psi &= g\,, \\
\psi(x_0) &= \psi_0 \in \mathbb{C}^d\,, \\
\psi'(x_0) &= \psi_1 \in \mathbb{C}^d\,.
\end{aligned} \tag{2.20}$$

Here we assume that for fixed $\varepsilon \in (0, \varepsilon_0)$ the matrix and vector valued functions P, A, f are smooth (with respect to x) and all their x–derivatives are uniformly bounded with respect to ε. For the applications we have in mind (e. g. the two–band $k \cdot p$ model from § 2.1.2), ε is a very small positive constant. The following transformation is constructed such that the $\frac{1}{\varepsilon}$ part of the resulting system matrix is skew–symmetric. This term determines the dominant oscillations, of the solution of the transformed and 'original' problem (2.20) respectively, with frequency $\sim \varepsilon$ and amplitudes of order $\mathcal{O}(1)$ as $\varepsilon \to 0$. For the case $P = 0$ we

present a modified version of a reformulation strategy, which is already discussed in [54].

Since $A(x,\varepsilon)$ is positive definite the Cholesky decomposition exists, cf. [67, 68]. I. e. there exists a unique upper triangular matrix[5] $\mathbb{L}(x,\varepsilon)$ with positive diagonal elements, such that

$$A(x,\varepsilon) \;=\; \mathbb{L}(x,\varepsilon)^*\,\mathbb{L}(x,\varepsilon)\,.$$

The entries of $\mathbb{L}(x,\varepsilon)$ are constructed row by row as follows (cf. [67]):

- for $i=j$ set

$$\mathbb{L}_{ii} \;=\; \left(A_{ii} \,-\, \sum_{k=1}^{i-1} |\mathbb{L}_{ki}|^2\right)^{\frac{1}{2}}$$

- then successively compute for $i<j$

$$\mathbb{L}_{ij} \;=\; \frac{1}{\mathbb{L}_{ii}}\left(A_{ij} \,-\, \sum_{k=1}^{i-1} \overline{\mathbb{L}_{ki}}\,\mathbb{L}_{kj}\right).$$

Due to construction of $\mathbb{L}(x,\varepsilon)$, we deduce that the matrix valued function $\mathbb{L}$ has the same smoothness properties as A. Further $\mathbb{L}(x,\varepsilon)$ is regular for all $(x,\varepsilon)\in\Omega$. Hence we can make the ansatz:

$$\tilde{u}_1 \;:=\; \mathbb{L}\psi\,, \qquad \tilde{u}_1 \;:=\; \varepsilon\psi'\,, \tag{2.21}$$

which yields

$$\tilde{u}' \;=\; \frac{1}{\varepsilon}\begin{pmatrix} 0 & \mathbb{L} \\ -\mathbb{L}^* & P \end{pmatrix}\tilde{u} \,+\, \begin{pmatrix} \mathbb{L}'\mathbb{L}^{-1} & 0 \\ 0 & 0 \end{pmatrix}\tilde{u} \,+\, \frac{1}{\varepsilon}\begin{pmatrix} 0 \\ g \end{pmatrix}, \tag{2.22}$$

$$\tilde{u}(x_0) \;=\; \tilde{u}_0 \;:=\; \begin{pmatrix} \mathbb{L}\psi_0 \\ \varepsilon\psi_1 \end{pmatrix}. \tag{2.23}$$

Since $P(x,\varepsilon)$ is skew symmetric, the same holds for the first matrix of (2.22), which we denote by $\tilde{L}$. Hence there exists a unitary matrix $Q(x,\varepsilon)$ and a real diagonal matrix $L(x,\varepsilon)$ such that

$$\tilde{L}(x,\varepsilon) \;=\; Q(x,\varepsilon)^*\, iL(x,\varepsilon)\, Q(x,\varepsilon)\,.$$

If (for fixed ε) A is differentiable this also holds for the matrix valued function $\tilde{L}$. Unfortunately this does not have to hold for the matrix valued function Q, as the following example illustrates.

Example 2.2.1. *Let $x\in(-1,1)$ and let $M\colon(-1,1)\to\mathbb{R}^{2\times 2}$ defined by*

$$M(x) \;:=\; \begin{pmatrix} x+2 & x \\ x & x+4 \end{pmatrix}.$$

[5] Here we changed the classical notation for the Cholesky decomposition $A=\mathbb{L}\mathbb{L}^*$ in order to simplify the notation of the transformed equation.

The matrix valued function M is obviously differentiable, with even analytic components. Furthermore $M(x)$ is symmetric and hence diagonalizable with eigenvalues

$$\lambda_1(x) \;=\; x \,+\, 3 \,-\, \sqrt{1+x^2}\,, \qquad \lambda_2(x) \;=\; x \,+\, 3 \,+\, \sqrt{1+x^2}$$

and corresponding eigenvectors[6]

$$v_1(x) \;=\; \begin{pmatrix} \frac{x}{1-\sqrt{1+x^2}} \\ 1 \end{pmatrix}, \qquad v_2(x) \;=\; \begin{pmatrix} \frac{x}{1+\sqrt{1+x^2}} \\ 1 \end{pmatrix}.$$

To get a diagonalization of the matrix $M(x) = Q(x)^*\Lambda(x)Q(x)$ with a unitary transformation $Q(x)$ we simply have to set

$$Q(x) \;:=\; \left(\frac{v_1(x)}{\|v_1(x)\|}, \frac{v_2(x)}{\|v_2(x)\|} \right)^*.$$

Unfortunately the first component of $v_1(x)$ has a pole at $x = 0$. With the rule of de l'Hospital (cf. [23]) *one gets*

$$\lim_{x \searrow 0} v_{1,1}(x) \;=\; -\infty\,, \qquad \lim_{x \nearrow 0} v_{1,1}(x) \;=\; \infty\,.$$

Hence the first component of $\frac{v_1}{\|v_1\|}$ has a jump at $x = 0$ and is consequently not continuous:

$$\lim_{x \searrow 0} \frac{v_{1,1}(x)}{\sqrt{v_1(x)^2+1}} \;=\; -1\,, \qquad \lim_{x \nearrow 0} \frac{v_{1,1}(x)}{\sqrt{v_1(x)^2+1}} \;=\; 1\,.$$

In this example the problem for Q can be fixed by piecewise definition. With v_1 also $-v_1$ is an eigenvector with respect to the eigenvalue λ_1. Hence we can set

$$Q(x) \;:=\; \left(\operatorname{sgn}(v_1(x)) \frac{v_1(x)}{\|v_1(x)\|}, \frac{v_2(x)}{\|v_2(x)\|} \right)^*.$$

This matrix valued function is continuously differentiable.

As we have seen in the above Example 2.2.1, where we start with analytic components and get a discontinuous transformation Q, the smoothness of $\tilde{L}$ does not automatically guarantee smoothness of Q. Hence, in order to continue with a further transformation we need

Assumption 2. *The matrix valued function Q has the same smoothness properties as A.*

Thus, now we can make the ansatz $u := Q\,\tilde{u}$ and the IVP (2.22)–(2.23) transforms into

$$u' \;=\; \frac{i}{\varepsilon} L u \,+\, B u \,+\, \frac{1}{\varepsilon} f\,, \quad u(x_0) \;=\; Q(x_0)\tilde{u}_0\,,$$

with

$$B \;:=\; Q' Q^{-1} \,+\, Q \begin{pmatrix} \mathbb{L}'\mathbb{L}^{-1} & 0 \\ 0 & 0 \end{pmatrix} Q^{-1}, \quad \text{and} \quad f \;:=\; Q \begin{pmatrix} 0 \\ g \end{pmatrix}.$$

The negative ε–order of the inhomogeneity is no drawback of the method. It can be replaced in a constructive way (with an additional transformation), such that the new inhomogeneity is of positive order with respect to ε. For the details we refer to § 3.4.

[6] Computed with *Maple 14*.

2.2.1 A special case discussed by Lorenz et al. [54]

In this section we consider the special of (2.20), where we set $P = 0$. Since A is positive definite, there exists a unique positive definite square root of A. Hence we can replace the first transformation (2.21) by

$$\tilde{u}_1 \;:=\; A^{\frac{1}{2}}\Psi\,, \qquad \tilde{u}_1 \;:=\; \varepsilon\Psi'\,, \tag{2.24}$$

which yields

$$\tilde{u}' \;=\; \frac{1}{\varepsilon}\begin{pmatrix} 0 & A^{\frac{1}{2}} \\ -A^{\frac{1}{2}} & 0 \end{pmatrix}\tilde{u} + \begin{pmatrix} (A^{\frac{1}{2}})'A^{-\frac{1}{2}} & 0 \\ 0 & 0 \end{pmatrix}\tilde{u}\,. \tag{2.25}$$

To compute[7] $A^{\frac{1}{2}}$ one can use a diagonalization of A. Since A is real and symmetric we can write $A = U^*\Lambda U$ with U unitary and Λ diagonal, e. g.

$$\Lambda(x) \;=\; \operatorname{diag}\left(\lambda_1(x)\operatorname{Id}_{\mu_1}, \ldots, \lambda_\sigma(x)\operatorname{Id}_{\mu_\sigma}\right).$$

We denote the vector of geometrical multiplicities $(\mu_1, \ldots, \mu_\sigma)$ by μ. Since it holds

$$\begin{pmatrix} i & 1 \\ 1 & i \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} -i & 1 \\ 1 & -i \end{pmatrix} \;=\; \begin{pmatrix} 2i & 0 \\ 0 & -2i \end{pmatrix}$$

the matrix

$$Q(x) \;:=\; \frac{1}{\sqrt{2}}\begin{pmatrix} i & 1 \\ 1 & i \end{pmatrix} \otimes U(x) \tag{2.26}$$

diagonalize the first matrix of (2.25), which we denote $\tilde{L}$. Further it holds

$$\begin{pmatrix} i & 1 \\ 1 & i \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} -i & 1 \\ 1 & -i \end{pmatrix} \;=\; \begin{pmatrix} 1 & i \\ -i & 1 \end{pmatrix}$$

and hence the variable transformation $u = Q\tilde{u}$ yields

$$u' \;=\; \frac{i}{\varepsilon}\begin{pmatrix} \Lambda^{\frac{1}{2}} & 0 \\ 0 & -\Lambda^{\frac{1}{2}} \end{pmatrix}u + Bu + \frac{1}{\varepsilon}Q\begin{pmatrix} 0 \\ g \end{pmatrix},$$

with

$$B \;=\; \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes (U'U^*) + \frac{1}{2}\begin{pmatrix} 1 & i \\ -i & 1 \end{pmatrix} \otimes \left(U(A^{\frac{1}{2}})'A^{-\frac{1}{2}}U^*\right). \tag{2.27}$$

An advantage of this approach is that one can use the diagonalization of A to diagonalize the matrix $\tilde{L}$, i. e. $\tilde{L} = iQ^*LQ$. The vector ν of geometrical multiplicities for the eigenvalues of L is given by $\nu = (\mu_1, \ldots, \mu_\sigma, \mu_1, \ldots, \mu_\sigma)$ and consequently $s = 2\sigma$.

Up to know (beside the adaption of notation) we followed the discussion from [54]. In the article the authors consider the case where the multiplicities of the eigenvalues are equal to one. Since we allow larger eigenspaces, we have

[7] It is also possible to compute $A^{\frac{1}{2}}$ without knowing the diagonalization of A. For example the *Matlab* function `sqrtm` is based on an algorithm which uses the *Schur* form of A. See [11] for more details.

to invest some additional work to adapt the approach from the article to our slightly more general setting. This is done in the following paragraph which results in Remark 2.2.3.

The ν–diagonal part of B plays an important role for the WKB–type transformation in § 3.3 (cf. Remark 3.3.2). It determines the 0^{th} order of the transformation with respect to ε. The following consideration shows that we can choose the matrix valued function U, such that the ν–diagonal part of B is a simple diagonal matrix. For a moment let us go back to the diagonalization of A and let W be a unitary matrix, which commutes with Λ, i. e. $\operatorname{diag}_\mu(W) = W$. Hence we can write

$$A \;=\; U^*\Lambda U \;=\; U^*W^*\Lambda W U\,.$$

Thus we can also use WU instead of U to construct Q (see (2.26)). If we do so, we have to deal with the matrix

$$(WU)'(WU)^* \;=\; W'W^* + WU'U^*W^*$$

to construct the matrix B. Since W commutes with Λ it holds

$$\operatorname{diag}_\mu\big((WU)'(WU)^*\big) \;=\; W'W^* + W\operatorname{diag}_\mu\big(U'U^*\big)W^*\,.$$

Furthermore we observe that

$$0 \;=\; \mathrm{Id}' \;=\; (U^*)'U + U^*U' \;\Leftrightarrow\; U^*U' \;=\; -(U^*)'U \;=\; -(U^*U')^*$$

which means that U^*U' is skew symmetric. Hence $\operatorname{diag}_\mu\big(U'U^*\big)$ is skew symmetric too.

Remark 2.2.2. *Let $M\colon I \to \mathbb{C}^{d\times d}$ be a skew symmetric matrix and let W be the unique solution of the IVP*

$$W'(x) \;=\; W(x)M(x)\,, \qquad W(x_0) \;=\; Id\,.$$

Then $W(x)$ is unitary for all $x \in I$.

Proof. It holds

$$\begin{aligned}(WW^*)' \;&=\; W'W^* + W(W^*)' \;=\; WMW^* + W(WM)^*\\ &=\; WMW^* + WM^*W^* \;=\; WMW^* - WMW^* \;=\; 0\,.\end{aligned}$$

Hence WW^* is constant and due to the initial condition it is $WW^* = \mathrm{Id}$. □

Due to Remark 2.2.2 we can choose W to be the solution of the IVP

$$W' \;=\; -W\operatorname{diag}_\mu(U'U^*)\,, \qquad W(x_0) \;=\; \mathrm{Id}\,, \tag{2.28}$$

which yields

$$\operatorname{diag}_\mu\big((WU)'(WU)^*\big) \;=\; 0\,.$$

Remark 2.2.3. *Thus, without restriction of generality, we can choose the matrix valued function U such that*

$$\operatorname{diag}_\mu(U'U^*) \;=\; 0\,. \tag{2.29}$$

Now we come back to the matrix B. Due to (2.29) the μ–diagonal part of B is determined by the second matrix of (2.27). It follows

$$\begin{aligned} U(A^{\frac{1}{2}})'A^{-\frac{1}{2}}U^* &= U(U^*\Lambda^{\frac{1}{2}}U)'U^*\Lambda^{-\frac{1}{2}} \\ &= U(U^*)' + (\Lambda^{\frac{1}{2}})'\Lambda^{-\frac{1}{2}} + \Lambda^{\frac{1}{2}}U'U^*\Lambda^{-\frac{1}{2}} \\ &= -U'U^* + \tfrac{1}{2}\Lambda'\Lambda^{-1} + \Lambda^{\frac{1}{2}}U'U^*\Lambda^{-\frac{1}{2}} \end{aligned} \tag{2.30}$$

and we get

$$\operatorname{diag}_\nu(B) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes \left(\tfrac{1}{4}\Lambda'\,\Lambda^{-1}\right).$$

Hence the IVP for the quantity $\mathcal{T}$ from Remark 3.3.2 reads

$$\begin{aligned} \mathcal{T}' &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes \left(\tfrac{1}{4}\Lambda'\,\Lambda^{-1}\right)\mathcal{T}, \\ \mathcal{T}(x_0) &= \mathrm{Id}, \end{aligned}$$

and has the unique solution

$$\mathcal{T}(x) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes \left(\Lambda^{\frac{1}{4}}(x)\Lambda^{-\frac{1}{4}}(x_0)\right). \tag{2.31}$$

If all geometric multiplicities of the eigenvalues of A are equal to one, then $\operatorname{diag}_\nu(B) = \operatorname{diag}(B)$. Since A is real, we can always choose U to be a real, orthogonal matrix. As we have seen before $U'U^*$ is skew symmetric and consequently the diagonal of it must be zero. Thus, in this case we (even) do not have to solve an IVP (cf. discussion above that lead to (2.28)) to get a correct transformation U. Hence the accuracy of the 0^{th}–order transformation is (only) determined by the accuracy of the diagonalization of A.

2.3 Linear second order BVPs

As we have seen in § 2.1.2, we can construct the solution of the two–band $k \cdot p$–model, which is a BVP, from the solution of an appropriate IVP. This procedure is not limited to this special example. In this section we present a method to construct the solution of a special class of BVPs from corresponding IVPs. This is an extension of the procedure described in [63], p. 111 ff. for the scalar case.

Let $K_1^a, K_2^a, K_1^b, K_2^b \in \mathbb{C}^{n\times n}$ and let $A, P\colon [a,b] \to \mathbb{C}^{n\times n}$ and $f\colon [a,b] \to \mathbb{C}^n$. We consider the following vector valued *Robin*–, mixed or third type boundary value problem (BVP):

$$y_{xx}(x) + P(x)\,y_x(x) + A(x)\,y(x) = f(x), \qquad x \in (a,b) \tag{2.32}$$

$$K_1^a\,y(a) + K_2^a\,y_x(a) = r^a \in \mathbb{C}^n, \tag{2.33}$$

$$K_1^b\,y(b) + K_2^b\,y_x(b) = r^b \in \mathbb{C}^n. \tag{2.34}$$

In this section we do not discuss solvability conditions for the BVP (2.32)–(2.34). Thus we need

Assumption 3. *The system* (2.32)–(2.34) *has a unique solution and we assume that A, P, f and the boundary data are such that the subsequent BVPs and IVPs in* § 2.3.1 *are uniquely solvable.*

In this case it is quite obvious that we can write

$$y(x) \;=\; y^B(x) + y^S(x)\,, \qquad x \in [a,b]\,,$$

where y^B, y^S are the unique solutions of

$$\begin{aligned} y^B_{xx}(x) + P(x)\,y_x(x) + A(x)\,y^B(x) &= 0\,, \qquad x \in (a,b) \qquad (2.35)\\ K_1^a\,y^B(a) + K_2^a\,y^B_x(a) &= r^a\,,\\ K_1^b\,y^B(b) + K_2^b\,y^B_x(b) &= r^b\,. \end{aligned}$$

and

$$\begin{aligned} y^S_{xx}(x) + P(x)\,y_x(x) + A(x)\,y^S(x) &= f(x)\,, \qquad x \in (a,b) \qquad (2.36)\\ K_1^a\,y^S(a) + K_2^a\,y^S_x(a) &= 0\,,\\ K_1^b\,y^S(b) + K_2^b\,y^S_x(b) &= 0\,. \end{aligned}$$

The solution of (2.36) can be constructed as follows. Let $G\colon [a,b]\times[a,b] \to \mathbb{C}^{n\times n}$ be continuous such that for every $\xi \in (a,b)$ it holds (in the classical sense)

$$\begin{aligned} G_{xx}(x,\xi) + P(x)G_x(x,\xi) + A(x)\,G(x,\xi) &= 0\,, \quad x \in (a,b)\backslash\{\xi\} \qquad (2.37)\\ K_1^a\,G(a,\xi) + K_2^a\,G_x(a,\xi) &= 0 \qquad (2.38)\\ K_1^b\,G(b,\xi) + K_2^b\,G_x(b,\xi) &= 0 \qquad (2.39)\\ G_x(\xi,\xi_-) - G_x(\xi,\xi_+) &= \mathrm{Id} \in \mathbb{C}^{n\times n}\,. \qquad (2.40) \end{aligned}$$

Then (if G exists) y^S is pointwise given by

$$y^S(x) \;:=\; \int_a^b G(x,\xi) f(\xi)\,d\xi\,.$$

The map G is the Greens function of the BVP with homogeneous boundary conditions.

Formal proof. Let A, P and f be continuous. Hence $G(\cdot,\xi)$ is a classical solution of (2.37) on the intervals (a,ξ) and (ξ,b). This yields

$$\begin{aligned} \frac{d}{dx}y^S(x) &= \frac{d}{dx}\Big(\int_a^x G(x,\xi)f(\xi)\,d\xi + \int_x^b G(x,\xi)f(\xi)\,d\xi\Big)\\ &= \big(G(x,x_-) - G(x,x_+)\big)\,f(x) + \int_a^b G_x(x,\xi)\,f(\xi)\,d\xi \end{aligned}$$

Since G is continuous the first summand is zero which yields

$$\begin{aligned} \frac{d^2}{dx^2}y^S(x) &= \frac{d}{dx}\Big(\int_a^x G_x(x,\xi)f(\xi)\,d\xi + \int_x^b G_x(x,\xi)f(\xi)\,d\xi\Big)\\ &= \big(G_x(x,x_-) - G_x(x,x_+)\big)\,f(x) + \int_a^b G_{xx}(x,\xi)\,f(\xi)\,d\xi\,. \end{aligned}$$

Hence y^S is two times continuously differentiable and we get

$$\begin{aligned} &y^S_{xx}(x) + P(x)y^S_x(x) + A(x)\,y^S(x)\\ &\quad= \big(G_x(x,x_-) - G_x(x,x_+)\big)\,f(x)\\ &\qquad + \int_a^b \big(G_{xx}(x,\xi) + P(x)G_x(x,\xi) + A(x)G(x,\xi)\big)\,f(\xi)\,d\xi\\ &\quad= f(x)\,. \end{aligned}$$

Since

$$y_x^S(x) \quad = \quad \int_a^b G_x(x,\xi)\, f(\xi)\, d\xi$$

we immediately see that y^S fulfill the homogeneous boundary conditions (2.38) and (2.39) . □

Another possible decomposition of y is as follows: Let y^S be the unique solution of the IVP

$$\begin{aligned} y_{xx}^S(x) + P(x)\, y_x^S(x) + A(x)\, y^S(x) &= f(x)\,, \qquad x \in (a,b) \\ y^S(a) &= 0\,, \\ y_x^S(a) &= 0 \end{aligned}$$

and let y^B be the solution of

$$\begin{aligned} y_{xx}^B(x) + P(x)\, y_x^B(x) + A(x)\, y^B(x) &= 0\,, \qquad x \in (a,b) \\ K_1^a\, y^B(a) + K_2^a\, y_x^B(a) &= r^a\,, \\ K_1^b\, y^B(b) + K_2^b\, y_x^B(b) &= r^b - K_1^b\, y^S(b) - K_2^b\, y_x^S(b)\,. \end{aligned}$$

Obviously $y = y^S + y^B$ solves the inhomogeneous differential equation (2.32) on (a,b) and fulfills the boundary conditions (2.33), (2.34).

In the following sections § 2.3.1 and § 2.3.2 we show how to compute y^B and the Greens function G from the solutions of suitable IVPs.

2.3.1 Deriving y^B from IVPs

Let us define the vectors $p := y^B(a)$ and $q := y_x^B(a)$. Then the boundary condition of (2.35) at $x = a$, i. e.

$$K_1^a p + K_2^a q \quad = \quad r^a \in \mathbb{C}^n\,, \tag{2.41}$$

can be rewritten as

$$M^a \begin{pmatrix} p \\ q \end{pmatrix} \quad := \quad \left(\begin{array}{c|c} K_1^a & K_2^a \end{array} \right) \begin{pmatrix} p \\ q \end{pmatrix} \quad = \quad r^a\,, \tag{2.42}$$

Let us forget for moment that p and q are connected with the solution y of the BVP (2.35). Since $M^a \in \mathbb{C}^{n\times 2n}$, it has a non trivial kernel. Let us assume that the vectors

$$\begin{pmatrix} p^1 \\ q^1 \end{pmatrix}, \ \cdots, \ \begin{pmatrix} p^m \\ q^m \end{pmatrix}$$

are a basis of the kernel of M^a, which we can compute (e. g.) with *Gauss elimination*. Furthermore let

$$\begin{pmatrix} \tilde{p} \\ \tilde{q} \end{pmatrix}$$

be an inhomogeneous solution of (2.42), which can be computed together with the basis of the kernel. Hence the general solution of (2.42) is given by

$$\begin{pmatrix} p \\ q \end{pmatrix} \quad = \quad \begin{pmatrix} \tilde{p} \\ \tilde{q} \end{pmatrix} + \left(\begin{array}{c|c|c} p^1 & \cdots & p^m \\ q^1 & \cdots & q^m \end{array} \right) c\,,$$

with $c \in \mathbb{C}^m$. Since $(y^B(a), y_x^B(a))^T$ is a solution of (2.41) there exists a unique vector $c^B \in \mathbb{C}^m$ such that

$$\begin{pmatrix} y^B(a) \\ y_x^B(a) \end{pmatrix} = \begin{pmatrix} \tilde{p} \\ \tilde{q} \end{pmatrix} + \left(\begin{array}{c|c|c} p^1 & \cdots & p^m \\ q^1 & \cdots & q^m \end{array}\right) c^B .$$

A priori we do not know the data of y^B at $x = a$, but we can compute the vectors on the right–hand side. Let $\tilde{y}, y^1, \ldots, y^m$ be the unique solutions of the following initial value problems:

$$\begin{aligned} \tilde{y}_{xx}(x) + P(x)\,\tilde{y}_x(x) + A(x)\,\tilde{y}(x) &= 0\,, \qquad x \in (a,b) \\ \tilde{y}(a) &= \tilde{p}\,, \\ \tilde{y}_x(a) &= \tilde{q} \end{aligned}$$

and for $j = 1, ..., m$

$$\begin{aligned} y_{xx}^j(x) + P(x)\,y_x^j(x) + A(x)\,y^j(x) &= 0\,, \qquad x \in (a,b) \\ y^j(a) &= p^j\,, \\ y_x^j(a) &= q^j\,. \end{aligned}$$

We define

$$Y(x) := \left(\; y^1(x) \;\middle|\; \ldots \;\middle|\; y^m(x) \; \right) .$$

By Assumption 3 y^B is the unique solution of (2.35) and hence

$$y^B = \tilde{y} + Y c^B\,,$$

with the not yet known vector c^B. Since y^B also has to fulfill the boundary condition at $x = b$ we get

$$K_1^b \left(\tilde{y}(b) + Y(b)\, c_j^B \right) + K_2^b \left(\tilde{y}_x(b) + Y_x(b)\, c_j^B \right) = r^b\,.$$

This is a linear equation for c^B which has a unique solution since y^B is the unique solution of (2.35).

2.3.2 Deriving the Greens function from IVPs

With the same ideas as in § 2.3.1 we shall construct the Greens function G, which is the solution of the BVP (2.37)–(2.40). Therefor we have to solve a suitable set of IVP. Let

$$\begin{pmatrix} p_a^1 \\ q_a^1 \end{pmatrix}, \cdots, \begin{pmatrix} p_a^{m_a} \\ q_a^{m_a} \end{pmatrix}, \qquad \begin{pmatrix} p_b^1 \\ q_b^1 \end{pmatrix}, \cdots, \begin{pmatrix} p_b^{m_b} \\ q_b^{m_b} \end{pmatrix} \in \mathbb{C}^{2n}$$

be corresponding kernel bases of the matrices

$$\left(\; K_1^a \;\middle|\; K_2^a \; \right) \in \mathbb{C}^{n \times 2n} \quad \text{and} \quad \left(\; K_1^b \;\middle|\; K_2^b \; \right) \mathbb{C}^{n \times 2n}\,.$$

Furthermore let $y^{a,1}, \ldots, y^{a,m_a}$ and $y^{b,1}, \ldots, y^{b,m_b}$ be the (corresponding) unique solutions of the IVPs

$$\begin{aligned} y_{xx}^{a,j}(x) + P(x)\,y_x^{a,j}(x) + A(x)\,y^{a,j}(x) &= 0\,, \qquad x \in (a,b) \\ y^{a,j}(a) &= p_a^j\,, \\ y_x^{a,j}(a) &= q_a^j \end{aligned}$$

and

$$\begin{aligned} y_{xx}^{b,j}(x) + P(x)\, y_x^{b,j}(x) + A(x)\, y^{b,j}(x) &= 0\,, \qquad x \in (a,b) \\ y^{b,j}(a) &= p_b^j\,, \\ y_x^{b,j}(a) &= q_b^j\,. \end{aligned}$$

We define for all $x \in [a,b]$

$$\begin{aligned} Y^a(x) &:= \left(\; y^{a,1}(x) \;\middle|\; \dots \;\middle|\; y^{a,m_a}(x) \; \right) \in \mathbb{C}^{n \times m_a}\,, \\ Y^b(x) &:= \left(\; y^{b,1}(x) \;\middle|\; \dots \;\middle|\; y^{b,m_b}(x) \; \right) \in \mathbb{C}^{n \times m_b}\,. \end{aligned}$$

Since $G(\cdot,\xi)$ solves the homogeneous boundary conditions at $x = a$ we deduce from the unique solvability that there exists a matrix[8] $C^a(\xi) \in \mathbb{C}^{m_a \times n}$ such that

$$G(x,\xi)|_{[a,\xi)} = Y^a(x)\, C^a(\xi)\,.$$

Analog we get from the right boundary a matrix $C^b(\xi) \in \mathbb{C}^{m_b \times n}$ such that

$$G(x,\xi)|_{(\xi,b]} = Y^b(x)\, C^b(\xi)\,.$$

Since G is assumed to be continuous on $[a,b] \times [a,b]$ and fulfills the jump condition (2.40) we get the following linear system

$$\left(\begin{array}{c|c} Y^a(\xi) & -Y^b(\xi) \\ \hline Y_x^a(\xi) & -Y_x^b(\xi) \end{array} \right) \left(\begin{array}{c} C^a \\ \hline C^b \end{array} \right) = \left(\begin{array}{c} 0 \\ \hline \mathrm{Id} \end{array} \right). \tag{2.43}$$

Hence the matrix valued functions Y^a, Y^b contain the whole information one needs to construct G.

Remark 2.3.1. *In application it can be of interest to store the values of the Greens function at certain grid points $x_1, \dots, x_l$. Since the Greens function has two arguments and maps to $\mathbb{R}^{n \times n}$ or $\mathbb{C}^{n \times n}$, one has to store $l^2 n^2$ scalar complex or real numbers. Of course, the storage cost may be (approximately) reduced by a factor one half, if the Greens function is symmetric in its arguments, but it stays quadratic in l.*

On the other hand side, as we have seen above, we can construct the Greens function from the matrix functions Y^a, Y^b. Hence the amount of storage we need to construct the Greens function for l grid points is given by $l(m_a + m_b)n$ scalar complex or real numbers. This is only linear in the crucial variable l instead of quadratic as for the storage of the whole matrix. Here we trade storage for computational speed.

[8] Each column of $G(\cdot, x')$ solves the vector valued IVP and hence is a linear combination of the kernel bases vectors.

Chapter 3

Analytic preprocessing: WKB–type transformations

The WKB method, named after the physicists Gregor Wentzel, Hendrik Anthony Kramers [47], and Leon Brillouin, is an approach to determine the asymptotic behavior (as $\varepsilon \to 0$) of the scalar stationary Schrödinger equation

$$\psi''(x) + \frac{1}{\varepsilon^2} a(x)\psi(x) \quad = \quad 0\,. \tag{3.1}$$

For $a(x) \neq 0$ one gets the first–term approximation of the general solution ψ from (3.1) (cf. [32] p.162f)

$$\psi(x) \quad \sim \quad a(x)^{-\frac{1}{4}} \left(c_1 e^{\frac{i}{\varepsilon}\int_{x_0}^{x} \sqrt{a(s)}\, ds} + c_2 e^{-\frac{i}{\varepsilon}\int_{x_0}^{x} \sqrt{a(s)}\, ds} \right). \tag{3.2}$$

Thus, if $a(x) > 0$, the solution is oscillatory and the local wave length tends to zero as $\varepsilon \to 0$. Hence ψ gets more and more oscillatory. The high oscillations are due to the $\mathcal{O}(\varepsilon^{-2})$ term in the ODE (3.1). Hence our aim is to find a reformulation of the ODE, such that these (or equivalent terms) are eliminated in the gained ODE. In the sequel we are only interested in the highly oscillatory case, i. e. $a(x) > 0$.

Equation (3.1) can be transformed, such that it fits into the setting of § 3.2 (cf. § 2.2). Thus we expect to find an asymptotic expansion of our Model Problem 1, which is the vector valued analogon (or at least very closely related) to the classical WKB method. For vector valued systems of form (3.1) (replace $a(x)$ by a positive matrix $A(x)$), we shall find an expansion similar to (3.2) (cf. Remark 3.5.8). This work is done in § 3.5. The transformation derived in § 3.3 could also be established with the WKB–type expansion from § 3.5 (see Remark 3.5.12). This is the reason why we call it *WKB–type transformation*. It transforms a linear first order ODE with a system matrix of $\mathcal{O}(\varepsilon^{-1})$ to a linear first order ODE with a system matrix of $\mathcal{O}(\varepsilon^{n})$ for some $n \in \mathbb{N}$. The price we have to pay for this are highly oscillatory entries in the gained system matrix. Due to this we shall need specially designed quadratures to derive our efficient marching methods in 6.

In § 3.4 we briefly describe an approach to transform an inhomogeneous ODE to a more convenient form in the spirit of § 3.3. It is based on the WKB–type

transformation. Since dealing with matrix equations needs a lot of notation, we assign the first section § 3.1 to this topic. There we also prove some technical results, which we shall use in the later upcoming discussions. Finally we give in § 3.6 a (very) brief introduction to asymptotic approximations, as done in [32].

3.1 Notation and technical results

We use this section to introduce our notations and shall prove some technical results.

The natural numbers including zero are denoted by $\mathbb{N}_0$, i. e. $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. By $\mathbb{R}^+$ we denote the strict positive real numbers, i. e. $\mathbb{R}^+ = \{x \in \mathbb{R} | \, x > 0\}$. Furthermore $\mathbb{R}_0^+ = \mathbb{R}^+ \cup \{0\}$ denotes the nonnegative real numbers. The real and complex numbers excluded 0 are denoted by $\mathbb{R}^*$ and $\mathbb{C}^*$ respectively.

By $\|\cdot\|$ we denote the Euclidean norm on $\mathbb{C}^d$ and the subordinated matrix norm on $\mathbb{C}^{d\times d}$ respectively. Let $I \subset \mathbb{R}$ be a closed non empty bounded interval. For continuous maps $M\colon I \to \mathbb{C}^d$ or $M\colon I \to \mathbb{C}^{d\times d}$ we define

$$\|M\|_\infty \quad := \quad \sup_{x\in I} \|M(x)\| \,. \tag{3.3}$$

The space $C^j(I, \mathbb{C}^{d\times d})$ of j-times continuously differentiable matrix valued functions $M\colon I \to \mathbb{C}^{d\times d}$ is also denoted by $C^j(I)$. Further we set

$$\|M\|_{C^k(I)} \quad := \quad \max_{j=0,\dots,k} \|M^{(j)}\|_\infty \,.$$

Lemma 3.1.1. *Let $I \subset \mathbb{R}$ and let $M\colon I \to \mathbb{C}^{d\times d}$ be regular for all $x \in I$. If M is differentiable at $x \in I$, then M^{-1} is differentiable at x too and the following holds:*

$$\big(M^{-1}(x)\big)' M(x) = -M^{-1}(x)M'(x) \tag{3.4}$$

Proof. It is $\det M(\xi) \neq 0$ for all $\xi \in I$ and hence $(\det M(\xi))^{-1}$ is differentiable in x. By Cramer's rule (cf. [7]) we get the differentiability of M^{-1} at x. Hence we can apply the product rule on $M^{-1}M$ which yields

$$0 \;=\; \mathrm{Id}' \;=\; \big(M^{-1}M\big)'(x) \;=\; M^{-1}(x)'M(x) \,+\, M^{-1}(x)M'(x) \,.$$

□

Corollary 3.1.2. *Let the matrix valued function $M\colon I \to \mathbb{C}^{d\times d}$ be regular for all $x \in I$. If $M \in C^r(I, \mathbb{C}^{d\times d})$ the same holds for M^{-1}. Furthermore*

$$\big(M^{-1}\big)' \quad = \quad -M^{-1}M'M^{-1} \,. \tag{3.5}$$

Proof. From equation (3.4) we immediately get $\big(M^{-1}\big)' = -M^{-1}M'M^{-1}$. Using this formula and Lemma 3.1.1 we inductively see that M^{-1} is n-times continuously differentiable. □

For the upcoming computations it is important that the involved quantities are bounded independent of the small parameter ε.

Definition 3.1.3. *Let $\varepsilon_0 > 0$ and $r \in \mathbb{N}$. The function $M\colon I \times (0, \varepsilon_0) \to \mathbb{C}^{d\times d}$ is called* C^n–bounded independently of ε *if and only if*

(i) $\forall\, \varepsilon \in (0,\varepsilon_0)\colon\ M(\cdot,\varepsilon) \in C^n(I)$,

(ii) $\exists\, c>0\ \forall\, \varepsilon \in (0,\varepsilon_0)\ \forall\, j \in \{0,\dots,n\}\colon\ \|M^{(j)}(\cdot,\varepsilon)\|_\infty \ \leq\ c.$

Remark 3.1.4. *The second condition of Definition 3.1.3 can also be written in the shorter form*

$$\exists\, c>0\ \forall\, \varepsilon \in (0,\varepsilon_0)\colon\ \|M(\cdot,\varepsilon)\|_{C^n(I)} \ \leq\ c\,.$$

Next we specify the notation for matrices. By Id_d we denote the identity matrix acting on $\mathbb{C}^d$. For $M \in \mathbb{C}^{m\times n}$ we denote by $M^* \in \mathbb{C}^{n\times m}$ its adjoint and by M^T its transposed matrix. The (i,j)–th component of M is labeled by M_{ij}, unless noted otherwise.

The *Kronecker product* (or tensor product) of $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{p\times q}$ is denoted by $A\otimes B$ and is defined as the block matrix

$$A\otimes B \ :=\ \begin{pmatrix} A_{11}B & \dots & A_{1n}B \\ \vdots & \ddots & \vdots \\ A_{m1}B & \dots & A_{mn}B \end{pmatrix} \in \mathbb{C}^{mp\times nq}\,. \tag{3.6}$$

For our applications the most important properties are the linearity in both components of $A\otimes B$ and the multiplication rule

$$(A\otimes B)(C\otimes D) \ =\ AC\otimes BD$$

for $A,C \in \mathbb{C}^{m\times n}$ and $B,D \in \mathbb{C}^{p\times q}$. For more details and a proof of the above mixed–product property we refer to [34].

The *Hadamard product* of $A,B \in \mathbb{C}^{m\times n}$ is denoted by $A\odot B$ and is defined by the entry–wise multiplication, i. e.

$$A\odot B \ :=\ \begin{pmatrix} A_{11}B_{11} & \dots & A_{1n}B_{1n} \\ \vdots & \ddots & \vdots \\ A_{m1}B_{m1} & \dots & A_{mn}B_{mn} \end{pmatrix} \in \mathbb{C}^{m\times n}\,. \tag{3.7}$$

We refer to [34] for more details. Obviously, the ones matrix $\mathbb{1}^{m\times n}$ defined by

$$\mathbb{1}^{m\times n}_{ij} \ =\ 1\,, \quad \text{for} \quad 1\leq i\leq m, 1\leq j\leq n \tag{3.8}$$

is the neutral element of the $\odot$–product on $\mathbb{C}^{m\times n}$. If $n=m$, we shortly write $\mathbb{1}_n$ instead of $\mathbb{1}^{n\times n}$. Furthermore we set

$$A^{\odot 0} \ :=\ \mathbb{1}^{m\times n}$$

and inductively define the $\odot$–powers of A by

$$A^{\odot n+1} \ =\ A\odot A^{\odot n}\,.$$

Let $d,s\in\mathbb{N}$, such that $s\leq d$ and let $\nu\in\mathbb{N}^s$ with $\sum_{j=1}^s \nu_j = d$. For arbitrary matrices $M_j \in \mathbb{C}^{\nu_j\times\nu_j}$, $j=1,\dots,s$, we denote by $\mathrm{diag}(M_1,\dots,M_s)$ the following block diagonal matrix:

$$\mathrm{diag}(M_1,\dots,M_s) \ :=\ \begin{pmatrix} M_1 & & 0 \\ & \ddots & \\ 0 & & M_s \end{pmatrix} \in \mathbb{C}^{d\times d}\,.$$

Conversely, for given $M \in \mathbb{C}^{d\times d}$ we denote by $\mathrm{diag}_\nu(M)$ the following block diagonal matrix:

$$\mathrm{diag}_\nu(M) \;:=\; \mathrm{diag}(\mathbb{1}_{\nu_1},\ldots,\mathbb{1}_{\nu_s})\odot M\,. \tag{3.9}$$

Furthermore we define

$$M^{\mathrm{dia}_\nu} \;:=\; \mathrm{diag}_\nu(M)\,,\qquad M^{\mathrm{off}_\nu} \;:=\; M - M^{\mathrm{dia}_\nu}\,. \tag{3.10}$$

Example. Let $\nu = (2,1,3)$ and hence $d = 6$ and let

$$M \;=\; \begin{pmatrix} 35 & 1 & 6 & 26 & 19 & 24\\ 3 & 32 & 7 & 21 & 23 & 25\\ 31 & 9 & 2 & 22 & 27 & 20\\ 8 & 28 & 33 & 17 & 10 & 15\\ 30 & 5 & 34 & 12 & 14 & 16\\ 4 & 36 & 29 & 13 & 18 & 11 \end{pmatrix}.$$

Then

$$\begin{aligned} M^{\mathrm{dia}_\nu} \;&=\; \mathrm{diag}_\nu(M) \;=\; \mathrm{diag}(\mathbb{1}_2,\mathbb{1}_1,\mathbb{1}_3)\odot M\\ &=\; \begin{pmatrix} 1 & 1 & & & & \\ 1 & 1 & & & & \\ & & 1 & & & \\ & & & 1 & 1 & 1\\ & & & 1 & 1 & 1\\ & & & 1 & 1 & 1 \end{pmatrix} \odot \begin{pmatrix} 35 & 1 & 6 & 26 & 19 & 24\\ 3 & 32 & 7 & 21 & 23 & 25\\ 31 & 9 & 2 & 22 & 27 & 20\\ 8 & 28 & 33 & 17 & 10 & 15\\ 30 & 5 & 34 & 12 & 14 & 16\\ 4 & 36 & 29 & 13 & 18 & 11 \end{pmatrix}\\ &=\; \begin{pmatrix} 35 & 1 & 0 & 0 & 0 & 0\\ 3 & 32 & 0 & 0 & 0 & 0\\ 0 & 0 & 2 & 0 & 0 & 0\\ 0 & 0 & 0 & 17 & 10 & 15\\ 0 & 0 & 0 & 12 & 14 & 16\\ 0 & 0 & 0 & 13 & 18 & 11 \end{pmatrix} \end{aligned}$$

and consequently

$$M^{\mathrm{off}_\nu} \;=\; M - M^{\mathrm{dia}_\nu} \;=\; \begin{pmatrix} 0 & 0 & 6 & 26 & 19 & 24\\ 0 & 0 & 7 & 21 & 23 & 25\\ 31 & 9 & 0 & 22 & 27 & 20\\ 8 & 28 & 33 & 0 & 0 & 0\\ 30 & 5 & 34 & 0 & 0 & 0\\ 4 & 36 & 29 & 0 & 0 & 0 \end{pmatrix}.$$

□

Sometimes it is simpler to estimate each coefficient of a matrix M instead of the norm $\|\cdot\|$. Therefore we define for $M \in \mathbb{C}^{d\times d}$

$$\|M\|_{\sup} \;=\; \sup_{i,j\in\{1,\ldots,d\}} |M_{ij}|\,. \tag{3.11}$$

It is easy to check that $\|\cdot\|_{\sup}$ is a norm on $\mathbb{C}^{d\times d}$. It holds for all $A, B \in \mathbb{C}^{d\times d}$

$$\|A \odot B\|_{\sup} \quad \leq \quad \|A\|_{\sup} \|B\|_{\sup} . \tag{3.12}$$

In the upcoming computations we have to deal with matrices that have a special block structure. Since computations with block matrices are often very similar to that of ordinary matrices we define

Definition 3.1.5. *A matrix $M \in \mathbb{C}^{d\times d}$ is called* ν–block diagonal *if and only if*

$$M \quad = \quad \mathrm{diag}_\nu(M) .$$

A segmentation of $X = (X_{ij})_{1\leq i,j\leq s} \in \mathbb{C}^{d\times d}$ into block matrices $X_{ij} \in \mathbb{C}^{\nu_i\times\nu_j}$ for all $i, j \in \{1, \ldots, s\}$ is called ν–segmentation of X.

In the following Lemma 3.1.6 we prove that the (ordinary) matrix multiplication carries over to ν–segmented matrices.

Lemma 3.1.6. *Let the matrices $A, B \in \mathbb{C}^{d\times d}$ and $C := AB$. Furthermore let $(A_{ij}), (B_{ij}), (C_{ij})$ be ν–segmentations of A, B, and C respectively. Then*

$$C_{ij} \quad = \quad \sum_{r=1}^{s} A_{ir} B_{rj}$$

hold for all $1 \leq i, j \leq s$.

Proof. In the proof we shall denote the matrix components by lower–case letters, e. g. a_{ij} is the ij^{th} entry of A. For $1 \leq m \leq s$ we define

$$\Sigma(m) \quad := \quad \sum_{n=1}^{m-1} \nu_i .$$

Let $1 \leq i, j \leq s$ and let $1 \leq k \leq \nu_i$, $1 \leq l \leq \nu_j$. Then for any matrix $X \in \mathbb{C}^{d\times d}$ with ν–segmentation (X_{ij}) we get

$$(X_{ij})_{kl} \quad = \quad x_{\Sigma(i)+k,\, \Sigma(j)+l} \in \mathbb{C} .$$

This yields

$$\begin{aligned}
\Big(\sum_{n=1}^{s} A_{in} B_{nj}\Big)_{kl} \quad &= \quad \sum_{n=1}^{s} \big(A_{in} B_{nj}\big)_{kl} \quad = \quad \sum_{n=1}^{s} \sum_{m=1}^{\nu_n} \big(A_{in}\big)_{km} \big(B_{nj}\big)_{ml} \\
&= \quad \sum_{n=1}^{s} \sum_{m=1}^{\nu_n} a_{\Sigma(i)+k,\, \sigma(n)+m}\, b_{\Sigma(n)+m,\, \Sigma(j)+l} \\
&= \quad \sum_{r=1}^{d} a_{\Sigma(i)+k,\, r}\, b_{r,\, \Sigma(j)+l} \quad = \quad c_{\Sigma(i)+k,\, \Sigma(j)+l} \\
&= \quad \big(C_{ij}\big)_{kl} .
\end{aligned}$$

□

Lemma 3.1.7. *Let $A, B \in \mathbb{C}^{d\times d}$. Then*

$$\operatorname{diag}_\nu(AB) \;=\; \operatorname{diag}_\nu(A)\,\operatorname{diag}_\nu(B) \,+\, \operatorname{diag}_\nu(A^{\operatorname{off}_\nu} B^{\operatorname{off}_\nu})$$

and

$$\operatorname{diag}_\nu(A^{\operatorname{off}_\nu} B^{\operatorname{off}_\nu}) \;=\; \operatorname{diag}_\nu(A\, B^{\operatorname{off}_\nu}) \;=\; \operatorname{diag}_\nu(A^{\operatorname{off}_\nu} B)\,.$$

Proof. Let $A = (A_{ij})_{1\le i,j\le s}$, $B = (B_{ij})_{1\le i,j\le s}$, $(C_{ij})_{1\le i,j\le s}$ be ν–segmentations of A, B, and $C := AB$. By Lemma 3.1.6 it holds

$$C_{ii} \;=\; \sum_{j=1}^{s} A_{is}B_{si} \;=\; A_{ii}B_{ii} \,+\, \sum_{j\neq i} A_{is}B_{si}\,.$$

This yields the first claim. Let $X \in \mathbb{C}^{n\times n}$ and let $(X_{ij}^{\operatorname{off}_\nu})$ be a ν–segmentation of $X^{\operatorname{off}_\nu}$. Hence $X_{jj}^{\operatorname{off}_\nu} = 0$ for $j = 1, \ldots, s$ and thus the second identity holds. □

3.1.1 The Sylvester equation

Some crucial points in the upcoming computations are a solvability condition and existence of regular solutions $X \in \mathbb{C}^{d\times d}$ of the *Sylvester equation* (cf. [13])

$$A\,X \,-\, X\,B \;=\; C \tag{3.13}$$

for given matrices $A, B, C \in \mathbb{C}^{d\times d}$. To be more precise we have to deal with the two following special cases

$$A\,X \,-\, X\,A \;=\; C \quad\text{and}\quad A\,X \,-\, X\,B \;=\; 0\,,$$

where the matrices A, B are diagonalizable. As mentioned in [34] it is often very useful to reformulate the equation by independent similarity transformations. Let $S, T \in \mathbb{C}^{d\times d}$ be regular. Then equation (3.13) is equivalent to

$$(SAS^{-1})\,SXT \,-\, SXT\,(T^{-1}BT) \;=\; SCT$$

which may be written as

$$A'X' \,-\, X'B' \;=\; C'\,.$$

Hence we can assume without loss of generality that

$$A = \operatorname{diag}(a_1, \ldots, a_d) \quad\text{and}\quad B = \operatorname{diag}(b_1, \ldots, b_d) \tag{3.14}$$

are diagonal matrices. From [34, Theorem 4.4.6] we deduce that

> (3.13) has a unique solution if and only if $\sigma(A) \cap \sigma(B) = \emptyset$, where $\sigma(A), \sigma(B)$ denotes the spectra of A and B respectively.

Thus for $C = 0$ a necessary condition for the existence of a regular solution is that A and B must have at least one common eigenvalue. Additionally it follows that A, B, C cannot be arbitrarily chosen. In literature usually the unique solvability of (3.13) is discussed. But the existence of regular solutions of the homogeneous Sylvester equation (3.13) is not covered by the consulted literature. Hence we shall prove a necessary and sufficient condition for it in Lemma 3.1.8.

In the proof of Lemma 3.1.8 we use the following notation: Let $M \in \mathbb{C}^{d\times d}$ be a quadratic matrix and let $\lambda \in \mathbb{C}$ be an eigenvalue of M. Then we denote by $\mu(\lambda, M)$ the geometrical multiplicity of λ with respect to M.

Lemma 3.1.8. *Let the matrices $A, B \in \mathbb{C}^{d\times d}$ be diagonal. Then the following two statements are equivalent:*

(i) *The matrix equation $AX - XB = 0$ has a regular solution*

(ii) *There exists a permutation matrix[1] P such that $A = PBP^*$.*

Proof. The implication (ii) $\Rightarrow$ (i) is almost trivial since the permutation matrix P is a regular solution of $AX - XB = 0$.

To prove (i) $\Rightarrow$ (ii) it is enough to show that $\sigma(A) = \sigma(B)$ and for all $\lambda \in \sigma(A)$ it is

$$\mu(\lambda, A) \quad = \quad \mu(\lambda, B)\,.$$

Since A, B are diagonal, the algebraic multiplicity coincides with geometrical multiplicity of the eigenvalue λ.

We define for a given matrix $M \in \mathbb{C}^{d\times d}$ the vector

$$\vec{M} \quad := \quad (m_{11}, \ldots, m_{d,1}, m_{12}, \ldots, m_{d,2}, \ldots, m_{d1}, \ldots, m_{d,d})^T \in \mathbb{C}^{d^2}\,.$$

Hence the equation $AX - XB = 0$ is equivalent to (cf. [34])

$$[\mathrm{Id}_d \otimes A \; - \; B^T \otimes \mathrm{Id}_d]\vec{X} \quad = \quad 0\,.$$

Thus the j–th column X_j of the regular solution X satisfies

$$(A - b_j\,\mathrm{Id}_d)X_j \quad = \quad 0\,.$$

Since X is regular, $X_j \neq 0$ holds and hence b_j is an eigenvalue of A and consequently $\sigma(B) \subset \sigma(A)$.

Let $\lambda \in \sigma(B)$ and assume $\mu := \mu(\lambda, B) > \mu(\lambda, A)$. Then there exist indices $j_1, \ldots, j_\mu$ such that

$$(A - \lambda\,\mathrm{Id}_d)X_{j_s} \;=\; 0\,, \quad \text{for} \quad s = 1, \ldots, \mu\,.$$

Since $X_{j_1}, \ldots, X_{j_\mu}$ are eigenvectors of A, they have to be linearly dependent, which is a contradiction to the regularity of X. Hence it is $\mu(\lambda, B) \leq \mu(\lambda, A)$.

From the characteristic polynomial we get

$$d \;=\; \sum_{\lambda\in\sigma(B)} \mu(\lambda, B) \;\leq\; \sum_{\lambda\in\sigma(B)} \mu(\lambda, A) \;\leq\; \sum_{\lambda\in\sigma(A)} \mu(\lambda, A) \;=\; d\,.$$

Hence it has to be $\mu(\lambda, B) = \mu(\lambda, A)$ and $\sigma(B) = \sigma(A)$. □

Lemma 3.1.9. *Let $A = \mathrm{diag}(\lambda_1\,\mathrm{Id}_{\nu_1}, \ldots, \lambda_s\,\mathrm{Id}_{\nu_s}) \in \mathbb{C}^{d\times d}$ with pairwise distinct λ_j and let $\nu = (\nu_1, \ldots, \nu_s)^T$ with $\sum_j \nu_j = d$. The matrix equation*

$$AX \; - \; XA \quad = \quad C \tag{3.15}$$

[1] From [33]: A quadratic matrix P is a permutation matrix if exactly one entry in each row and column is equal to 1, and all other entries are 0. P is regular and $P^{-1} = P^*$. In the context of linear equations with coefficient matrix A the coordinate transformation $A \to PAP^*$ corresponds to a renumbering of the variables.

has a solution if and only if $\mathrm{diag}_\nu(C) = 0$. *In this case the general solution is given by*

$$X \;=\; D_A^- \odot C \,+\, M\,,$$

with arbitrary ν*–block diagonal matrix* $M \in \mathbb{C}^{d\times d}$. *The matrix* D_A^- *is defined by*

$$(D_A^-)_{ij} \;:=\; \left\{ \begin{array}{c} 0 \\ (A_{ii}-A_{jj})^{-1}, \end{array} \right. \begin{array}{c} A_{ii}-A_{jj}=0\,, \\ \textit{else}\,. \end{array}$$

Proof. Let $X = (X_{ij})_{1\le i,j\le s}$, $C = (C_{ij})_{1\le i,j\le s}$ be ν–segmentations of X and C. We compute for $i, j \in \{1, \ldots, s\}$

$$C_{ij} \;=\; (AX - XA)_{ij} \;=\; \lambda_i \,\mathrm{Id}_{\nu_i}\, X_{ij} \,-\, X_{ij}\lambda_j \,\mathrm{Id}_{\nu_j} \;=\; (\lambda_i - \lambda_j)X_{ij}\,.$$

Thus $\mathrm{diag}_\nu(C) = \mathrm{diag}(C_{11}, \ldots, C_{ss}) = 0$ is a necessary condition for the solvability of (3.15). Since $\lambda_1, \ldots, \lambda_s$ are pairwise distinct we have for $i \neq j$

$$X_{ij} \;=\; \frac{1}{\lambda_i - \lambda_j}\, C_{ij} \;=\; \frac{1}{\lambda_i - \lambda_j} \mathbb{1}^{\nu_i\times\nu_j} \odot C_{ij}$$

Hence, if $\mathrm{diag}_\nu(C) = 0$, then

$$X_p \;:=\; D_A^- \odot C$$

is a particular solution of (3.15) where D_A^- is defined via the ν–segmentation

$$(D_A^-)_{ij} \;:=\; \left\{ \begin{array}{c} (\lambda_i - \lambda_j)^{-1},\ i \neq j \\ 0 \quad ,\ \text{else} \end{array} \right\} \mathbb{1}^{\nu_i\times\nu_j}\,.$$

Analog to D_A^- we define D_A via

$$(D_A)_{ij} \;:=\; (\lambda_i - \lambda_j)\, \mathbb{1}^{\nu_i\times\nu_j}\,.$$

Hence (3.15) is given by $D_A \odot X = C$. Since this is an inhomogeneous linear equation in X all solutions are given by

$$X \;=\; X_p \,+\, M$$

with $D_A \odot M \;=\; 0$, which is equivalent to M is ν–block diagonal. □

Remark 3.1.10. *Sylvester's matrix equation* (3.15) *shows up in the proof of Proposition 3.3.1. Due to its importance for the proof we shall collect some results associated with the quantities of Lemma 3.1.9. Let* $A \in \mathbb{C}^{d\times d}$ *be a diagonal matrix.*

(i) *From Lemma 3.1.9 we deduce (set* $C = 0$*)*

$$[A, M] \;=\; 0 \;\;\Leftrightarrow\;\; M \textit{ is } \nu\textit{–block diagonal.}$$

(ii) *From the proof of Lemma 3.1.9: for arbitrary* $M \in \mathbb{C}^{d\times d}$ *it holds*

$$[A, M] \;=\; D_A \odot M\,.$$

The matrix D_A *is defined by the* ν*–segmentation*

$$(D_A)_{ij} \;:=\; (\lambda_i - \lambda_j)\, \mathbb{1}^{\nu_i\times\nu_j}\,. \tag{3.16}$$

(iii) *By definition of* D_A *and* D_A^- *it holds*

$$D_A \odot D_A^- \quad = \quad \mathbb{1}_d - \mathrm{diag}(\mathbb{1}_{\nu_1}, \ldots, \mathbb{1}_{\nu_s})\,.$$

(iv) *Let* $B = (B_{ij}), M = (M_{ij}) \in \mathbb{C}^{d\times d}$ *be* ν*–segmentations of* B, M*. Furthermore let* M *be* ν*–block diagonal. Then* $(i \neq j)$

$$\begin{aligned}\big(D_A^- \odot (BM)\big)_{ij} &= (\lambda_i - \lambda_j)^{-1}\, \mathbb{1}^{\nu_i\times\nu_j} \odot (B_{ij}M_{jj}) \\ &= \big((\lambda_i - \lambda_j)^{-1} B_{ij}\big) M_{jj} \;=\; \big((D_A^-)_{ij} \odot B_{ij}\big) M_{jj}\end{aligned}$$

and hence we get

$$D_A^- \odot (BM) \quad = \quad (D_A^- \odot B)M\,.$$

3.1.2 The matrix E_Φ^ε

A further notation concerned with the (entry–wise) matrix product $\odot$ arises from the similarity transformation used in § 3.3 (see (3.25)–(3.24)).

Let $\nu \in \mathbb{N}^s$ with $\sum_{j=1}^s \nu_j = d$ and let $\Phi = \mathrm{diag}(\lambda_1 \,\mathrm{Id}_{\nu_1}, \ldots, \lambda_s \,\mathrm{Id}_{\nu_s}) \in \mathbb{C}^{d\times d}$ with pairwise distinct eigenvalues λ_j. For $\varepsilon > 0$ we define the matrix E_Φ^ε componentwise by

$$(E_\Phi^\varepsilon)_{rs} \quad := \quad e^{-\frac{i}{\varepsilon}(\Phi_{rr} - \Phi_{ss})}\,. \tag{3.17}$$

It is easy to check that for any $B \in \mathbb{C}^{d\times d}$ it holds

$$\exp\big(-\tfrac{i}{\varepsilon}\Phi\big)\; B\; \exp\big(\tfrac{i}{\varepsilon}\Phi\big) \quad = \quad E_\Phi^\varepsilon \odot B\,. \tag{3.18}$$

The matrix E_Φ^ε can also be generated by the formula

$$E_\Phi^\varepsilon \quad = \quad \exp_\odot\big(-\tfrac{i}{\varepsilon} D_\Phi\big)\,,$$

where D_Φ is given by (3.16). Here $\odot$ indicates to apply the exponential function componentwise, in contrast to $\exp(-\frac{i}{\varepsilon}\Phi)$ where we use the matrix exponential function.

The following properties of the oscillatory matrix E_Φ^ε are used to derive the finite difference methods in chapter 6.

(i) Let $\Phi(x)$ be a smooth x–dependent diagonal matrix. It holds

$$\frac{d}{dx} E_\Phi^\varepsilon(x) \quad = \quad -\tfrac{i}{\varepsilon}\; D_{\Phi'}(x) \odot E_\Phi^\varepsilon(x)\,.$$

(ii) Let $A, B, \Phi \in \mathbb{C}^{d\times d}$ be arbitrary quadratic matrices. Furthermore we assume that Φ is diagonal. It follows with $E := \exp(\frac{i}{\varepsilon}\Phi)$ and (3.18) that

$$\begin{aligned}(E_\Phi^\varepsilon \odot A)(E_\Phi^\varepsilon \odot B) &= E^{-1}AEE^{-1}BE \;=\; E^{-1}ABE \\ &= E_\Phi^\varepsilon \odot (AB)\,.\end{aligned}$$

(iii) Let Φ be a real valued diagonal matrix. Hence $\exp\big(-\frac{i}{\varepsilon}\Phi\big)$ is unitary and it holds for all $B \in \mathbb{C}^{d\times d}$

$$\|B\| \quad = \quad \|\exp\big(-\tfrac{i}{\varepsilon}\Phi\big)\; B\; \exp\big(\tfrac{i}{\varepsilon}\Phi\big)\| \;=\; \|E_\Phi^\varepsilon \odot B\|\,. \tag{3.19}$$

(iv) Let $\lambda_1, \dots, \lambda_s \in C^1(I, \mathbb{C})$ such hat $\lambda_1(x), \dots, \lambda_s(x)$ are pairwise distinct for all $x \in I$. Furthermore let

$$\Phi(x) \quad := \quad \int_{x_0}^{x} \operatorname{diag}(\lambda_1(\xi)\operatorname{Id}_{\nu_1}, \dots, \lambda_s(\xi)\operatorname{Id}_{\nu_s})\, d\xi$$

and let $M\colon I \times (0, \varepsilon_0) \to \mathbb{C}^{d\times d}$ be C^1–bounded independently of ε such that $\operatorname{diag}_\nu M(x,\varepsilon) = 0$ for all $x \in I \times (0,\varepsilon_0)$. It holds:

$$\begin{aligned}\int_{x_0}^{x} E_\Phi^\varepsilon(s) \odot M(s,\varepsilon)\, ds \quad &\overset{\text{(i)}}{=} \quad i\varepsilon \int_{x_0}^{x} \big(E_\Phi^\varepsilon(s)\big)' \odot \big(D_{\Phi'}^-(s) \odot M(s,\varepsilon)\big)\, ds \\ &= \quad i\varepsilon E_\Phi^\varepsilon(s) \odot D_{\Phi'}^-(s) \odot M(s,\varepsilon)\big|_{s=x_0}^{x} \\ &\qquad -i\varepsilon \int_{x_0}^{x} E_\Phi^\varepsilon(s) \odot \big(D_{\Phi'}^-(s) \odot M(s,\varepsilon)\big)'\, ds\,.\end{aligned}$$

Hence there exists a constant $c \geq 0$ such that for all $\varepsilon > 0$ and all $x \in I$

$$\left\| \int_{x_0}^{x} E_\Phi^\varepsilon(s) \odot M(s,\varepsilon)\, ds \right\| \quad \leq \quad c\,\varepsilon\,. \tag{3.20}$$

3.2 Formulation of the problem

The IVPs (2.9) and (2.19) derived in § 2.1.1 and § 2.1.2 respectively have the same structure. Since this are only two examples of a much larger class of problems which can be transformed to equations of similar form, we shall continue our discussion for a more general problem (of the form (2.19)).

Model Problem 1. *Let $\varepsilon_0 > 0$ and let $[a,b] \subset \mathbb{R}$ be a bounded non–trivial interval. We define $\Omega := [a,b] \times (0,\varepsilon_0)$. Further let $\nu_1, \dots, \nu_s \in \mathbb{N}$ with $\sum_{j=1}^{s} \nu_j = d$ and let $l_1, \dots, l_s\colon \Omega \to \mathbb{R}$. We set*

$$L(x,\varepsilon) \quad = \quad \operatorname{diag}(l_1 \operatorname{Id}_{\nu_1}, \dots, l_s \operatorname{Id}_{\nu_s})(x,\varepsilon) \in \mathbb{R}^{d\times d} \subset \mathbb{C}^{d\times d}$$

where $\operatorname{Id}_{\nu_j}$ denotes the identity matrix of $\mathbb{C}^{\nu_j \times \nu_j}$. For $x_0 \in [a,b]$ and (fixed) $\varepsilon \in (0,\varepsilon_0)$ we shall consider the initial value problem for $u(x,\varepsilon) \in \mathbb{C}^d$:

$$u'(x,\varepsilon) \quad = \quad \frac{i}{\varepsilon} L(x,\varepsilon)\, u(x,\varepsilon) \,+\, B(x,\varepsilon)\, u(x,\varepsilon)\,, \quad x \in [a,b]\,, \tag{3.21}$$

$$u(x_0,\varepsilon) \quad = \quad u_0(\varepsilon) \in \mathbb{C}^d\,, \tag{3.22}$$

and make the following assumptions:

(A1) *For every fixed $\varepsilon \in (0,\varepsilon_0)$ the matrix valued functions $L\colon \Omega \to \mathbb{R}^{d\times d}$ and $B\colon \Omega \to \mathbb{C}^{d\times d}$ are smooth (in the spatial variable x) and B, L and all their x–derivatives are uniformly bounded on $[a,b]$ with respect to ε.*

(A2) *There exists a positive constant $\delta > 0$, such that for all $(x,\varepsilon) \in \Omega$ and all admissible indices $i \neq j$ it holds*

$$|l_i(x,\varepsilon) - l_j(x,\varepsilon)| \quad \geq \quad \delta\,.$$

(A3) *The map $u_0\colon (0,\varepsilon_0) \to \mathbb{C}^d$ is bounded.*

Remark 3.2.1. *Assumptions* (A2) *exclude the case of crossing eigenvalues. This case will be the content of future work.*

Since L is diagonal and real valued, the solution u of the IVP (3.21)–(3.22) is (highly) oscillatory. Its norm is bounded by a constant independently of ε. In order to prove this we introduce a smoother, "adiabatic" variable η, which coincides with the "η" from [40, 54, 27]. This can be interpreted as the lowest order WKB–type transformation in the context of § 3.3.

Lemma 3.2.2. *Let u be the unique solution of the IVP* (3.21)–(3.22) *and let*

$$E_\varepsilon(x) \;:=\; \exp\left(\frac{i}{\varepsilon}\int_{x_0}^{x} L(s,\varepsilon)\,ds\right) \in \mathbb{C}^{d\times d}\,.$$

Then the new quantity $\eta(x) := E_\varepsilon^*(x)\,u(x)$ solves the IVP

$$\eta' \;=\; (E_\varepsilon^* B E_\varepsilon)\,\eta\,, \qquad \eta(x_0,\varepsilon) \;=\; \eta_0(\varepsilon) \;:=\; u_0(\varepsilon)\,.$$

There exists a constant $c \geq 0$, such that it holds for all $(x,\varepsilon) \in \Omega$:

$$\|u(x,\varepsilon)\| \;=\; \|\eta(x,\varepsilon)\| \;\leq\; c\,.$$

Proof. Since $L(x,\varepsilon)$ is diagonal and real valued $E_\varepsilon(x)$ is unitary. Differentiation of the ansatz $\eta = E_\varepsilon^* u$ yields the IVP. By Corollary 6.1.5 it holds

$$\|\eta(x,\varepsilon)\| \;\leq\; e^{|x-x_0|\|E_\varepsilon^* B(\cdot,\varepsilon)E_\varepsilon\|_\infty}\,\|\eta_0(\varepsilon)\|\,.$$

Since E_ε is unitary it follows

$$\|u(x,\varepsilon)\| \;=\; \|\eta(x,\varepsilon)\| \;\leq\; e^{|x-x_0|\|B(\cdot,\varepsilon)\|_\infty}\,\|u_0(\varepsilon)\|\,.$$

Due to Assumption (A1) of our Model Problem 1 the matrix valued function $B(\cdot,\varepsilon)$ is bounded independently of ε. By (A3) the same holds for $u_0(\varepsilon)$. □

3.3 Reformulation of the initial value problem

If $\varepsilon \ll 1$, then the solution u of the Model Problem 1 (p. 34) is highly oscillatory with (local) wavelength $\sim \varepsilon$. Hence standard integrators need a very fine grid (with step sizes $h < \varepsilon$) in order to produce reliable results. The goal of this thesis is to derive a marching method which does not have this restriction, i. e. which can use (in the best case) ε independent grids. The first step to achieve this goal is an analytical preprocessing of the initial value problem (3.21)–(3.22), which is discussed in the sequel.

Our transformation ansatz (3.25) is mainly inspired by [54] and [4]. In the first article we find an analytic preprocessing for a vector valued second order initial value problem (IVP), which is almost identical to our zeroth order transformation ($n = 0$). The procedure discussed in the second article for the special case of a scalar second order equation shows that more sophisticated transformations (compared to [54]) are possible, which yield system matrices of order $\mathcal{O}(\varepsilon)$. The combination of the results in both articles has been our motivation to search for a generalization of the ansatz from [4] for the more general setting in [54].

A result comparable to our approach presented in this thesis can be found in [27]. There a product ansatz for the transformation T_ε (see (3.25)) is used which is called *super–adiabatic transformation* (cf. § 3.3.2). Another ansatz, which seems to have a similar structure as our transformation (3.25), is (independently from this work) discussed in [17]. A major difference of our method compared to the mentioned articles and books is the incorporation of the case of multiple eigenvalues of L with constant multiplicity on the interval I.

The following transformation ansatz is designed to get rid of the ε^{-1}-term in (3.21). This is done such that the gained system matrix is of order $\mathcal{O}(\varepsilon^\alpha)$ for some $\alpha > 0$. Let the assumptions of our Model Problem 1 from § 3.2 hold and let $\Omega := I \times (0, \varepsilon_0)$. Furthermore let u be the (unique) solution of the IVP (3.21)–(3.22), i. e.

$$u'(x) \;=\; \frac{i}{\varepsilon} L\, u \,+\, B\, u\,, \qquad u(x_0) \;=\; u_0\,, \tag{3.23}$$

with $L, B\colon \Omega \to \mathbb{C}^{d\times d}$ and

$$L(x,\varepsilon) \;=\; \operatorname{diag}\left(l_1(x,\varepsilon)\operatorname{Id}_{\nu_1}, \ldots, l_s(x,\varepsilon)\operatorname{Id}_{\nu_s}\right) \in \mathbb{R}^{d\times d}.$$

The matrix valued functions L, B are C^m–bounded independently of ε. Let the matrix valued functions $T_0, \ldots, T_n\colon \Omega \to \mathbb{C}^{d\times d}$ be (at least) C^1–bounded independently of ε, such that

$$T_\varepsilon(x) \;:=\; \sum_{j=0}^{n} \varepsilon^j\, T_j(x,\varepsilon) \tag{3.24}$$

is regular for all $(x,\varepsilon) \in \Omega$. In order to eliminate the dominant oscillations with frequency $\sim \frac{1}{\varepsilon}$ and amplitude $\mathcal{O}(1)$ as $\varepsilon \to 0$ we make the following transformation ansatz:

$$y(x) \;:=\; E_\varepsilon^{-1}(x) T_\varepsilon^{-1}(x)\, u(x)\,, \tag{3.25}$$

where we set

$$E_\varepsilon(x) \;:=\; \exp\left(\frac{i}{\varepsilon}\int_{x_0}^{x} L(s,\varepsilon)\, ds\right). \tag{3.26}$$

Since $L(x,\varepsilon)$ is real for all $(x,\varepsilon) \in \Omega$ the matrix $E_\varepsilon(x)$ is unitary, i. e.

$$E_\varepsilon^{-1}(x) \;=\; E_\varepsilon^{*}(x) \;=\; \exp\left(-\frac{i}{\varepsilon}\int_{x_0}^{x} L(s,\varepsilon)\, ds\right).$$

The following Proposition 3.3.1 states that we can determine the matrix valued functions $T_0, \ldots, T_n$ such that the new variable y is the (unique) solution of an IVP, whose system matrix is bounded by a constant times ε^n. The matrix T_ε^{-1} is implicitly defined by its inverse in order to point out the connection between the WKB–type approximation of u as discussed in § 3.5 and the transformation here.

Proposition 3.3.1 (WKB–type transformation)**.** *Let $L, B\colon \Omega \to \mathbb{C}^{d\times d}$ be C^r–bounded independently of ε and let $r \geq n \in \mathbb{N}$. Then there exists an $\varepsilon_0 \geq \varepsilon_1 > 0$ and matrix valued functions $T_0, \ldots, T_n\colon \tilde{\Omega} := I \times (0, \varepsilon_1) \to \mathbb{C}^{d\times d}$ such that T_j*

is C^{r-j+1}–bounded independently of ε for $j = 0, \dots, n$. Furthermore the matrix $T_\varepsilon(x)$ is regular for all $(x, \varepsilon) \in \tilde{\Omega}$ and the new variable y satisfies the IVP (for all $0 < \varepsilon < \varepsilon_1$)

$$y' \;=\; \varepsilon^n E_\varepsilon^* S_n E_\varepsilon \, y \,, \qquad y(x_0) = y_0 \,. \tag{3.27}$$

The function $S_n \colon \tilde{\Omega} \to \mathbb{C}^{d\times d}$ is C^{r-n}–bounded independently of ε.

Proof. The proof consists of three parts:

(i) formal derivation of conditional equations for $T_0, \dots, T_n$

(ii) solving the conditional equations yields an explicit recurrence relation and regularity for $T_0, \dots, T_n$ on Ω

(iii) restriction to $(0, \varepsilon_1) \subset (0, \varepsilon_0)$ yields regularity of T_ε^{-1} and justifies (i)

(i): Formal differentiation of the above ansatz (3.25) yields

$$y' \;=\; E_\varepsilon^* \left[\tfrac{i}{\varepsilon}(T_\varepsilon^{-1} L \,-\, L T_\varepsilon^{-1}) \,+\, T_\varepsilon^{-1\prime} \,+\, T_\varepsilon^{-1} B \right] (E_\varepsilon^* T_\varepsilon^{-1})^{-1} y \,. \tag{3.28}$$

Since T_ε^{-1} is implicitly defined by its inverse it is not easy to derive conditional equations for the matrices $T_0, \dots, T_n$ from the differential equation (3.28). Hence we reformulate the right–hand side such that the terms between the squared brackets only contains T_ε. We can write

$$T_\varepsilon^{-1} L - L T_\varepsilon^{-1} \;=\; T_\varepsilon^{-1} [L, T_\varepsilon] T_\varepsilon^{-1} \,,$$

which yields with Lemma 3.1.1 (and E_ε^* being unitary)

$$y' \;=\; E_\varepsilon^* T_\varepsilon^{-1} \left(\tfrac{i}{\varepsilon}[L, T_\varepsilon] \,-\, T_\varepsilon' \,+\, B T_\varepsilon \right) E_\varepsilon y$$

and hence we get ($T_\varepsilon = \sum \varepsilon^j T_j$)

$$\begin{aligned} y' = E_\varepsilon^* T_\varepsilon \big(\tfrac{i}{\varepsilon}[L, T_0] + i[L, T_1] + \,\dots\, + \varepsilon^{n-1} i[L, T_n] \\ +\quad B T_0 + \,\dots\, + \varepsilon^{n-1} B T_{n-1} + \varepsilon^n B T_n \\ -\quad T_0' - \,\dots\, - \varepsilon^{n-1} \quad T_{n-1}' - \varepsilon^n \quad T_n' \big) E_\varepsilon y \,. \end{aligned} \tag{3.29}$$

Now the idea is to choose $T_0, \dots, T_n$ such that the coefficients of $\varepsilon^{-1}, \dots, \varepsilon^{n-1}$ in equation (3.29) vanish, which leads to the following system of equations

$$i[L, T_j] \,+\, B T_{j-1} \,-\, T_{j-1}' \;=\; 0 \,, \qquad j = 0, \dots, n \tag{3.30}$$

where we set $T_{-1} := 0$. Provided there exists a solution of the above system, we get

$$y' \;=\; \varepsilon^n \, E_\varepsilon^* S_n E_\varepsilon \, y \quad \text{with} \quad S_n = T_\varepsilon^{-1} (B T_n - T_n') \,. \tag{3.31}$$

(ii): Lemma 3.1.9 yields that (3.30) has a solution for $j = 0, \dots, n$ if and only if

$$\operatorname{diag}_\nu (T_{j-1}' \,-\, B T_{j-1}) \;=\; 0 \,, \tag{3.32}$$

which is equivalent to

$$\left(T_{j-1}^{\mathrm{dia}_\nu} \right)' \;=\; \operatorname{diag}_\nu (B)\, T_{j-1}^{\mathrm{dia}_\nu} \,+\, \operatorname{diag}_\nu \left(B^{\mathrm{off}_\nu} T_{j-1}^{\mathrm{off}_\nu} \right) . \tag{3.33}$$

Here we use notation (3.10) and Lemma 3.1.7 to expand $\mathrm{diag}_\nu\left(BT_{j-1}\right)$. Let $\mathcal{T}$ be the unique solution of the homogeneous linear IVP (on the interval I)

$$\mathcal{T}' \;=\; \mathrm{diag}_\nu(B)\,\mathcal{T}\,, \qquad \mathcal{T}(x_0) \;=\; \mathrm{Id}_d\,. \tag{3.34}$$

Hence by variation of constants (3.33) is equivalent to

$$\begin{aligned} &T_{j-1}^{\mathrm{dia}_\nu}(x) \\ &\quad=\; \mathcal{T}(x)\Big(T_{j-1}^{\mathrm{dia}_\nu}(x_0) \,+\, \int_{x_0}^{x} \mathcal{T}(\xi)^{-1}\,\mathrm{diag}_\nu\left(B^{\mathrm{off}_\nu} T_{j-1}^{\mathrm{off}_\nu}\right)(\xi)\,d\xi\Big) \end{aligned} \tag{3.35}$$

Again by Lemma 3.1.9 we see that (3.30) uniquely determines the matrix $T_j^{\mathrm{off}_\nu}$, while $T_j^{\mathrm{dia}_\nu}$ is arbitrary. Hence for $j = 0, \ldots, n$ we find that the order equation (3.30) and the solvability condition (3.32) are equivalent to the following system

$$\begin{aligned} T_j^{\mathrm{off}_\nu} &= i\,D_L^- \odot \left(BT_{j-1} - T_{j-1}'\right) \qquad (3.36)\\ T_j^{\mathrm{dia}_\nu}(x) &= \mathcal{T}(x)\Big(T_j^{\mathrm{dia}_\nu}(x_0) \,+\, \int_{x_0}^{x} \mathcal{T}(\xi)^{-1}\,\mathrm{diag}_\nu\left(B^{\mathrm{off}_\nu} T_j^{\mathrm{off}_\nu}\right)(\xi)\,d\xi\Big)\,. \end{aligned}$$

Since $T_{-1} = 0$ fulfills the solvability condition (3.32) we derived an explicit recurrence relation for the matrices $T_j^{\mathrm{off}_\nu}, T_j^{\mathrm{dia}_\nu}$. This yields the existence of $T_0, \ldots, T_n$. Due to $T_{-1} = 0$ we also find

$$T_0(x) \;=\; T_0^{\mathrm{dia}_\nu}(x) \;=\; \mathcal{T}(x) T_0^{\mathrm{dia}_\nu}(x_0)\,.$$

We additionally deduce from the above construction that $\mathcal{T}$ is C^{r+1}–bounded independently of ε. Since we want $T_0, \ldots, T_n$ to be bounded independently of ε we have to choose bounded integration constants $T_0^{\mathrm{dia}_\nu}(x_0, \varepsilon), \ldots, T_n^{\mathrm{dia}_\nu}(x_0, \varepsilon)$. Hence we deduce by induction from (3.36) for $j = 0, \ldots, n$:

$$T_j \;=\; T_j^{\mathrm{dia}_\nu} + T_j^{\mathrm{off}_\nu} \text{ is } C^{r+1-j}\text{–bounded independently of } \varepsilon\,.$$

It follows from the definition of T_ε (cf. (3.24)) that T_ε is C^{r-n+1}–bounded independently of ε.
(iii): Since $T_\varepsilon(x)$ has to be regular for $\varepsilon \in (0, \varepsilon_1)$, for some not yet determined $\varepsilon_1 > 0$, we choose a regular initial condition $T_0^{\mathrm{dia}_\nu}(x_0, \varepsilon)$ such that $\|T_0^{\mathrm{dia}_\nu}(x_0, \varepsilon)^{-1}\|$ is bounded independently of ε. Hence $T_0(x, \varepsilon)$ is regular on I and the norm of its inverse is bounded by a constant $c_0 > 0$ independently of ε. Since all T_j are at least C^{r-n}–bounded independently of ε there exists a constant $c > 0$ such that it holds for all $(x, \varepsilon) \in I \times (0, \varepsilon_1)$

$$\|T_j(x, \varepsilon) T_0^{-1}(x, \varepsilon)\| \;\leq\; c\,,$$

which yields for all $y \in \mathbb{C}^d$ (with the lower triangle inequality)

$$\begin{aligned} \|T_\varepsilon(x) y\| &= \left\| \Big(\mathrm{Id} + \varepsilon \sum_{j=1}^{n} \varepsilon^{j-1} T_j(x, \varepsilon) T_0^{-1}(x, \varepsilon) \Big) T_0(x, \varepsilon) y \right\| \\ &\geq \left(1 - \frac{\varepsilon}{1-\varepsilon} c\right) \|T_0(x, \varepsilon) y\|\,. \end{aligned}$$

For $0 \leq \varepsilon \leq \varepsilon_1 := \min(\varepsilon_0, \frac{1-\tau}{1+c})$ with $\tau \in (0, 1]$ the right–hand side is positive for all $y \neq 0$ and hence $T_\varepsilon(x)$ is injective and consequently regular (cf. [19]). Since

$T_0, \dots, T_n$ are continuously differentiable and T_ε is regular on I for $\varepsilon \in (0, \varepsilon_1)$, the above formal derivation of the system (3.30) is a posteriori justified.

It remains to show that S_n is C^{r-n}–bounded independently of ε. Since this trivially holds for T_n', T_n, B we simply have to show that the same is true for T_ε^{-1}. From Corollary 3.1.2 we get $T_\varepsilon^{-1} \in C^{r-n}(I)$. Furthermore

$$\|T_\varepsilon^{-1}\| \;=\; \sup_{y\neq 0} \frac{\|y\|}{\|T_\varepsilon y\|} \;\leq\; \sup_{y\neq 0} \frac{1-\varepsilon}{1-(1+c)\varepsilon} \frac{\|y\|}{\|T_0 y\|} \;\leq\; \frac{\|T_0^{-1}(x,\varepsilon)\|}{\tau} \;\leq\; \frac{c_0}{\tau}\,.$$

Since T_ε is C^{r-n+1} bounded independently of ε, we inductively deduce with equation (3.5) from Corollary 3.1.2 that $\|(T_\varepsilon^{-1})^{(j)}\|$, for $j \in \{0, \dots, r-n\}$, is bounded independently of ε. □

In the proof of Proposition 3.3.1 we derive an explicit recurrence relation for the matrices T_j. For the numerical approximation we choose (except for $j = 0$) all integration constants equal to zero (i. e. $T_j^{\mathrm{dia}_\nu}(x_0) = 0$). For this special case we summarize the computation procedure in

Remark 3.3.2. *Let T_0 be the unique solution of the IVP*

$$T_0' \;=\; \mathrm{diag}_\nu(B)\, T_0\,, \qquad T_0(x_0) \;=\; \mathrm{Id}\,. \tag{3.37}$$

Furthermore define the matrix valued functions $T_1, \dots, T_n \colon \tilde{\Omega} \to \mathbb{C}^{d\times d}$ by the explicit recurrence relation

$$T_j^{\mathrm{off}_\nu} \;=\; i\, D_L^- \odot \big(B T_{j-1} - T_{j-1}'\big)\,, \tag{3.38}$$

$$T_j^{\mathrm{dia}_\nu}(x) \;=\; T_0(x) \int_{x_0}^{x} T_0(\xi)^{-1}\, \mathrm{diag}_\nu(B^{\mathrm{off}\,\nu} T_j^{\mathrm{off}\,\nu})(\xi)\, d\xi\,. \tag{3.39}$$

By Lemma 3.1.7, we can replace $\mathrm{diag}_\nu(B^{\mathrm{off}\,\nu} T_j^{\mathrm{off}\,\nu})$ *by* $\mathrm{diag}_\nu(B\, T_j^{\mathrm{off}\,\nu})$ *in the integrand of* (3.39)*. Additionally let*

$$\Phi(x) \;=\; \int_{x_0}^{x} L(\xi)\, d\xi\,, \qquad T_\varepsilon(x) \;=\; \sum_{j=0}^{n} \varepsilon^j T_j(x)\,.$$

Then the ansatz (3.25) $y = E_\varepsilon^* T_\varepsilon^{-1}\, u$ *yields*

$$\begin{aligned} y' \;&=\; \varepsilon^n\, (E_\varepsilon^* S_n E_\varepsilon)\, y \;=\; \varepsilon^n\, (E_\Phi^\varepsilon \odot S_n)\, y \\ y(x_0) \;&=\; T_\varepsilon^{-1}(x_0) u_0 \end{aligned} \tag{3.40}$$

with E_Φ^ε from § 3.1.2*. The matrix valued function S_n is given by*

$$S_n \;=\; T_\varepsilon^{-1}\big(B T_n - T_n'\big)\,. \tag{3.41}$$

It is not necessary to compute the ν–diagonal part of T_n with the above relation. Here one is free to choose it, such that the whole problem gets simpler.

The IVP (3.21) can be reformulated such that the system matrix of the equivalent IVP (3.27) is of order $\mathcal{O}(\varepsilon^n)$. We can express the unique solution y of (3.27) by the limit of the Picard iteration (cf. Lemma 6.1.4). This shall be

discussed in more detail in § 6.1. Let us briefly point out the idea. Integration of the linear first order IVP (here A is a wild–card character for $E_\varepsilon^* S_n E_\varepsilon$)

$$y' \;=\; \varepsilon^n A(x) y\,, \qquad y(x_0) = y_0 \tag{3.42}$$

yields the integral equation

$$y(x) \;=\; y_0 \,+\, \varepsilon^n \int_{x_0}^{x} A(s) y(s)\, ds\,. \tag{3.43}$$

Now we can replace[2] $y(s)$ in the integrand by the integral equation (3.43) and hence

$$y(x) \;=\; y_0 \,+\, \varepsilon^n \int_{x_0}^{x} A(s) y_0 \,+\, \varepsilon^{2n} \int_{x_0}^{x} \int_{x_0}^{s} A(r) y(r)\, dr\,.$$

We continue with this procedure and get an (infinite) sum of y independent, multiple integrals. We shall call them *iterated integrals* and they shall be denoted by $\mathcal{I}_{x_0}^j$ (cf. Remark 6.1.2). It holds

$$\mathcal{I}_{x_0}^0 \;=\; \mathrm{Id} \quad \text{and} \quad \mathcal{I}_{x_0}^{j+1}(x) \;=\; \int_{x_0}^{x} A(x)\, \mathcal{I}_{x_0}^j(s)\, ds \quad \text{for all } j \in \mathbb{N}.$$

Proposition 3.3.3. *Let the assumptions of Proposition 3.3.1 hold. Then the IVP* (3.27) *admits a solution* $y \in C^{r-n+1}(I, \mathbb{C}^d)$ *with the expansion*

$$y(x) \;=\; \sum_{j=0}^{\infty} \varepsilon^{jn} \mathcal{I}_{x_0}^j(x)\, y_0\,, \tag{3.44}$$

where the iterated integrals $\mathcal{I}_{x_0}^j$ *are given by Definition 6.1.1 with* $M = E_\varepsilon^* S_n E_\varepsilon$. *Moreover we have for all* $x \in I$

$$\|y(x) - y_0\| \;\le\; c\,\varepsilon^n\,, \quad \|y^{(j)}(x)\| \;\le\; c\,\varepsilon^{n-j+1}\,,$$

for $j = 1, \ldots, r-n+1$ *with a constant* c *independently of* ε.

Proof. From Proposition 3.3.1 we get that S_n is C^{r-n}–bounded independently of ε. Thus we deduce from ODE (3.27) that the solution $y \in C^{r-n+1}(I, \mathbb{C}^d)$. Since E_ε is unitary

$$\|(E_\varepsilon^* S_n E_\varepsilon)(x)\| = \|S_n(x)\|$$

(cf. § 3.1.2) and hence we can apply Lemma 6.1.4 which yields the series representation of y. Therefor we use that $\mathcal{I}_{x_0}$ is linear in M, in order to write ε^{jn} in front of the integrals.

In order to prove the first estimate we use the series representation (3.44). It holds that

$$y(x) - y_0 \;=\; \varepsilon^n \mathcal{I}_\xi^1 \Big(\sum_{j=0}^{\infty} \varepsilon^{jn} \mathcal{I}_{x_0}^j(x)\, y_0 \Big) \;=\; \varepsilon^n (\mathcal{I}_\xi^1 y)(x)\,.$$

[2] We can also use the fundamental theorem of calculus and the IVP (3.42).

By Lemma 6.1.3 we immediately get

$$\|y(x) - y_0\| \quad \leq \quad \varepsilon^n |x - x_0| \|S_n\|_\infty \|y\|_\infty \,.$$

From Corollary 6.1.5 we deduce that $\|y\|_\infty$ is finite and uniformly bounded in ε. Due to $\|E'_\varepsilon\| = \mathcal{O}(\frac{1}{\varepsilon})$ the other estimates inductively follow from the first one in connection with ODE (3.27). □

The estimates in Proposition 3.3.3 are a direct consequence of the series representation of y and the ε–order of the system matrix of (3.27). Since the series summands are highly oscillatory integrals, at least on the ν-off diagonal blocks, it is possible to improve the first estimate of Proposition 3.3.3. This is done by an additional (final) transformation. The following Corollary 3.3.4 is an adaption of [4, Proposition 2.2].

Corollary 3.3.4 (Arnold, BenAbdallah, Negulescu [4])**.** *Let the assumptions of Proposition 3.3.3 hold and let $R\colon I \to \mathbb{C}^{d\times d}$ be the unique solution of the IVP*

$$R' \;=\; \varepsilon^n \operatorname{diag}_\nu(S_n)\, R\,, \qquad R(x_0) \;=\; \mathrm{Id}\,. \tag{3.45}$$

If y is the unique solution of (3.27) then $z := R^{-1}y$ solves the IVP

$$z' \;=\; \varepsilon^n \left(E_\varepsilon^* \mathcal{S}_n E_\varepsilon\right) z\,, \qquad z(x_0) = z_0\,, \tag{3.46}$$

with $\mathcal{S}_n := R^{-1} S_n^{\mathrm{off}_\nu} R$. Furthermore z admits the improved estimates

$$\|z(x) - z_0\| \leq c\,\varepsilon^{n+1}\,, \quad \|z^{(j)}(x)\| \leq c\,\varepsilon^{n-j+1}\,,$$

for $j = 1, \ldots, r-n+1$ with an ε independent constant $0 < c < \infty$.

Proof. Since $R(x_0) = \mathrm{Id}$, it holds $R = \operatorname{diag}_\nu(R)$, which yields $[E_\varepsilon, R] = 0$. Hence differentiation of the ansatz for z yields the IVP (3.46). Furthermore from $S_n \in C^{r-n}(I)$ we deduce $R \in C^{r-n+1}(I)$. This yields $z \in C^{r-n+1}(I)$. Let $\varepsilon^n M$ be the system matrix of (3.46), i. e.

$$M \;:=\; E_\varepsilon^* R^{-1} S_n^{\mathrm{off}_\nu} R E_\varepsilon \;=\; E_\Phi^\varepsilon \odot \left(R^{-1} S_n^{\mathrm{off}_\nu} R\right).$$

As in the proof of Proposition 3.3.3 we use the series representation to write down the following identity:

$$z(x) - z_0 \;=\; \varepsilon^n \sum_{j=0}^{\infty} \varepsilon^{jn} \big(\mathcal{I}_{x_0}^j \mathcal{I}_{x_0}^1\big)(x) z_0\,.$$

It holds $\operatorname{diag}_\nu(M) = 0$ and hence we get from (3.20)

$$\|\mathcal{I}_{x_0}^1(x)\| \;=\; \left\| \varepsilon^n \int_{x_0}^x E_\Phi^\varepsilon \odot (R S_n R^{-1})(\xi)\, d\xi \right\| \;=\; \mathcal{O}(\varepsilon^{n+1})\,.$$

With Lemma 6.1.3 it follows

$$\|z(x) - z_0\| \quad \leq \quad \varepsilon^n e^{\varepsilon^n \|M\|_\infty |x-x_0|} \|\mathcal{I}_{x_0}^1(x)\| \;\leq\; c\,\varepsilon^{n+1}\,.$$

The other estimates directly follow from the differential equation for z. □

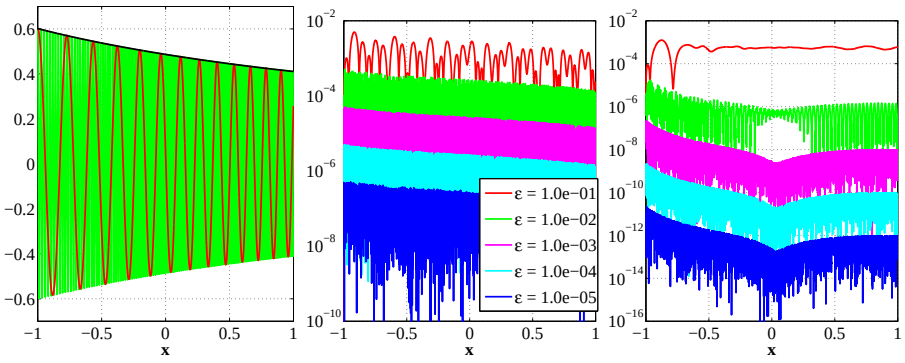

Figure 3.1: The left picture shows the real part of the first component of u (i. e. $\mathrm{Re}(u_1)$) for the numerical problem from § 7.1 for five values of ε $(10^{-1}, \ldots, 10^{-5})$. In the middle, the absolute value of $\mathrm{Re}(\eta_1 - \eta_1^*)$ is plotted with semilogarithmic axis. The function $\mathrm{Re}(\eta_1^*)$ is the black line in the upper part of the left picture. The right picture shows a semilogarithmic plot of the absolute value of $\mathrm{Re}(z_1(x) - z_1(x_0))$. The legend from the η plot in the middle is valid for all three plots.

Remark 3.3.5. *Since S_n is at least C^0–bounded independently of ε we get from* (3.45) *and the Gronwall Lemma 8.4.3, that R is C^0–bounded independently of ε too. Using once again the ODE, we further find that R is at least C^1–bounded independently of ε.*

In Figure 3.1 we illustrate the effect of the discussed transformations for the vector valued example from § 7.1, which is also used to illustrate the performance of the one–step methods derived in § 6. For our WKB–type transformation that finally yields the variable z we set $n = 1$. We plot the real part of the first component of u (left) (cf. (3.23)) and the absolute values of $\mathrm{Re}(\eta_1 - \eta_1^*)$ (middle) (cf. Lemma 3.2.2) and $\mathrm{Re}(z_1(x) - z_1(x_0))$ (right) (cf. Corollary 3.3.4). For the last two quantities we choose a semilogarithmic representation. We see that u is highly oscillatory. The variable η is oscillatory too (even with a higher frequency than u), but the amplitude of $\eta - \eta^*$ decreases with decreasing ε. Here η^* is a smooth function, which is the solution of the *adiabatic limit equation* as discussed in [54, §2.5]. As derived in the article, we see oscillations of $\mathcal{O}(\varepsilon)$ around η^*. Since we choose $n = 1$ we expect, due to the estimates of Corollary 3.3.4, that z oscillates around its initial condition with amplitudes of $\mathcal{O}(\varepsilon^2)$. And indeed this can be observed in the right picture of Figure 3.1.

3.3.1 Application to a scalar second order IVP discussed by Arnold et al. [4]

In [4] the authors discuss an efficient numerical method for the integration of the following linear scalar second order IVP (which is a special case of the IVPs discussed in § 2.2):

$$\begin{aligned} \varepsilon^2 \varphi''(x) + a(x)\varphi(x) &= 0\,, \\ \varphi(x_0) &= \varphi_0 \in \mathbb{C}\,, \\ \varepsilon \varphi'(x_0) &= \varphi_1 \in \mathbb{C}\,. \end{aligned} \tag{3.47}$$

They use a single transformation to get an equivalent first order system. The gained system matrix splits into a diagonal matrix of order $\mathcal{O}(\varepsilon^{-1})$ and a remainder of order $\mathcal{O}(\varepsilon)$. In order to remove the (diagonal) $\mathcal{O}(\varepsilon^{-1})$ term they use a final transformation, where the variable is multiplied with a highly oscillatory matrix. This matrix is defined as E_ε from (3.26).

In order to see the connection between our approach and the method discussed in [4], we shall apply our WKB–type transformation to the IVP (3.47). First we have to rewrite it as a first order system. To get a comparable result we set $n = 1$ with respect to Remark 3.3.2. Since no first derivative is present in (3.47), both approaches from § 2.2 (the ansatz with Cholesky decomposition and the transformation from [54]) are equal. Additionally we do not have to diagonalize the matrix valued function $a(x)$, i. e. $U = 1$ which yields $U' = 0$. Hence the new quantity

$$u(x) \;:=\; \frac{1}{\sqrt{2}} \begin{pmatrix} i & 1 \\ 1 & i \end{pmatrix} \begin{pmatrix} a^{\frac{1}{2}}(x)\varphi(x) \\ \varepsilon\varphi'(x) \end{pmatrix}$$

solves the IVP

$$\begin{aligned} u' &= \frac{i}{\varepsilon} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} a^{\frac{1}{2}}\, u + \begin{pmatrix} 1 & i \\ -i & 1 \end{pmatrix} \frac{(a^{\frac{1}{2}})' a^{-\frac{1}{2}}}{2}\, u\,, \\ u(x_0) &= u_0\,. \end{aligned}$$

Thus we have

$$L(x) \;=\; \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} a^{\frac{1}{2}}(x) \quad \text{and} \quad B(x) \;=\; \begin{pmatrix} 1 & i \\ -i & 1 \end{pmatrix} \frac{a'(x)}{4a(x)}\,.$$

From (2.31) of § 2.2.1 we know that the quantity $\mathcal{T} = T_0$ is given by

$$T_0(x) \;=\; \mathcal{T}(x) \;=\; \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \frac{a^{\frac{1}{4}}(x)}{a^{\frac{1}{4}}(x_0)}\,.$$

To simplify the computations we set

$$t_0(x) \;:=\; \frac{a^{\frac{1}{4}}(x)}{a^{\frac{1}{4}}(x_0)}\,. \tag{3.48}$$

Furthermore

$$D_L(x) \;=\; \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} 2a^{\frac{1}{2}}(x)\,.$$

This yields (cf. Remark 3.3.2)

$$\begin{aligned} T_1^{\mathrm{off}_\nu}(x) &= iD_L^-(x) \odot \big(B(x)T_0(x) - T_0(x)'\big) \\ &= \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \frac{i}{2a^{\frac{1}{2}}(x)} \odot \begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix} \frac{a'(x)}{4a(x)} t_0(x) \\ &= -t_0(x) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \frac{a'(x)}{8a^{\frac{3}{2}}(x)}\,. \end{aligned}$$

Since we do not consider T_2 and higher order coefficients we are free to choose arbitrary diagonal elements of T_1. The following choice is made, such that the

determinant of T_ε is equal to t_0^2. This significantly simplifies the formula for the inverse matrix T_ε^{-1}. We set

$$T_1(x) \;:=\; -t_0(x)t_1(x)\begin{pmatrix} -i & 1 \\ 1 & i \end{pmatrix} \quad \text{with} \quad t_1(x) \;:=\; \frac{a'(x)}{8a^{\frac{3}{2}}(x)}\,.$$

It follows

$$T_\varepsilon \;=\; t_0\begin{pmatrix} 1+i\varepsilon t_1 & -\varepsilon t_1 \\ -\varepsilon t_1 & 1-i\varepsilon t_1 \end{pmatrix},$$

which yields ($\det T_\varepsilon = t_0^2$)

$$T_\varepsilon^{-1} \;=\; \frac{1}{t_0}\begin{pmatrix} 1-i\varepsilon t_1 & \varepsilon t_1 \\ \varepsilon t_1 & 1+i\varepsilon t_1 \end{pmatrix} \;=\; \frac{1}{t_0}\,\mathrm{Id} + \frac{\varepsilon t_1}{t_0}\begin{pmatrix} -i & 1 \\ 1 & i \end{pmatrix}.$$

Remark 3.3.6. *Due to our special choice of the diagonal elements of T_1, the matrix T_ε is always regular independently of ε. Hence the restriction in Proposition 3.3.1 to a smaller ε–interval $(0,\varepsilon_1)$ can be neglected.*

Now we can compute the matrix valued function S_1 from Remark 3.3.2:

$$S_1(x) \;=\; T_\varepsilon^{-1}(x)\big(B(x)T_1(x) - T_1'(x)\big)\,.$$

In order to make the computations more traceable we start with the expression in the brackets:

$$\begin{aligned} &BT_1 - T_1' \\ &= -t_0\begin{pmatrix} 1 & i \\ -i & 1 \end{pmatrix}\begin{pmatrix} -i & 1 \\ 1 & i \end{pmatrix}\frac{(a')^2}{32a^{\frac{5}{2}}} + \begin{pmatrix} -i & 1 \\ 1 & i \end{pmatrix}\left(t_0\frac{a'}{8a^{\frac{3}{2}}}\right)' \\ &= \begin{pmatrix} -i & 1 \\ 1 & i \end{pmatrix}\left(t_0'\frac{a'}{8a^{\frac{3}{2}}} + t_0\frac{a''a^{\frac{3}{2}} - \frac{3}{2}(a')^2a^{\frac{1}{2}}}{8a^3}\right). \end{aligned}$$

From Remark 3.3.2 or from the definition of t_0 in (3.48) we deduce

$$t_0' \;=\; \frac{1}{4}\frac{a'}{a}t_0$$

which yields

$$BT_1 - T_1' \;=\; t_0\begin{pmatrix} -i & 1 \\ 1 & i \end{pmatrix}\frac{4a''a - 5(a')^2}{32a^{\frac{5}{2}}}\,.$$

Let us denote the last scalar factor of the above equation by β, i. e.

$$\beta \;:=\; \frac{4a''a - 5(a')^2}{32a^{\frac{5}{2}}} \;=\; -\frac{1}{2}\,a^{-\frac{1}{4}}\big(a^{-\frac{1}{4}}\big)''\,.$$

Since

$$\begin{pmatrix} -i & 1 \\ 1 & i \end{pmatrix}\begin{pmatrix} -i & 1 \\ 1 & i \end{pmatrix} \;=\; \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

i. e. the matrix is nilpotent, we immediately compute

$$S_1(x) \;=\; T_\varepsilon^{-1}(x)\big(B(x)T_1(x) - T_1'(x)\big) \;=\; \beta(x)\begin{pmatrix} -i & 1 \\ 1 & i \end{pmatrix}.$$

Furthermore we get (see Remark 3.3.2)

$$\Phi(x) \;=\; \int_{x_0}^{x} a^{\frac{1}{2}}(t)\,dt \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Hence the transformation ($E_\varepsilon(x) = \exp(\frac{i}{\varepsilon}\Phi(x))$)

$$y(x) \;=\; E_\varepsilon^*(x) T_\varepsilon^{-1}(x) u(x)$$

yields the IVP

$$\begin{aligned} y' &= \varepsilon\, E_\varepsilon^*(x)\,\beta(x) \begin{pmatrix} -i & 1 \\ 1 & i \end{pmatrix} E_\varepsilon(x)\, y\,, \\ y(x_0) &= y_0\,. \end{aligned}$$

Now we can compute the (last) transformation from Corollary 3.3.4. The variable R solves the IVP

$$R' \;=\; \varepsilon\,\beta(x) \begin{pmatrix} -i & 0 \\ 0 & i \end{pmatrix} R\,, \qquad R(x_0) \;=\; \mathrm{Id}\,,$$

and thus

$$R(x) \;=\; \exp\left(\varepsilon \int_{x_0}^{x} \beta(t)\,dt \begin{pmatrix} -i & 0 \\ 0 & i \end{pmatrix}\right).$$

Hence, $z = R^{-1}y$ is the unique solution of

$$\begin{aligned} z' &= \varepsilon\, R(x)^{-1} E_\varepsilon^*(x)\,\beta(x) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} E_\varepsilon(x) R(x)\, z\,, \\ z(x_0) &= z_0 \;:=\; y_0\,. \end{aligned}$$

By construction it holds

$$E_\varepsilon(x)R(x) \;=\; \exp\left(\frac{i}{\varepsilon}\int_{x_0}^{x} a^{\frac{1}{2}}(t) - \varepsilon^2\beta(t)\,dt \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\right).$$

We set $\phi^\varepsilon(x) := \int_{x_0}^{x} a^{\frac{1}{2}}(t) - \varepsilon^2\beta(t)\,dt$. This yields

$$\begin{aligned} z'(x) &= \varepsilon\,\beta(x) \begin{pmatrix} 0 & e^{-\frac{2i}{\varepsilon}\phi^\varepsilon(x)} \\ e^{\frac{2i}{\varepsilon}\phi^\varepsilon(x)} & 0 \end{pmatrix} z(x)\,, \qquad (3.49) \\ z(x_0) &= z_0\,. \end{aligned}$$

The ODE (3.49) is exactly the reformulation of the problem which is discussed in [4]. To find out how our approach is related to that discussed in the article, let us have a look on our transformation that connects y and φ, φ':

$$\begin{aligned} y &= R^{-1}E_\varepsilon^* T_\varepsilon^{-1} \frac{1}{\sqrt{2}} \begin{pmatrix} i & 1 \\ 1 & i \end{pmatrix} \begin{pmatrix} a^{\frac{1}{2}}\varphi \\ \varepsilon\varphi' \end{pmatrix} \\ &= R^{-1}E_\varepsilon^* \left(\frac{1}{t_0}\,\mathrm{Id} + \frac{\varepsilon t_1}{t_0} \begin{pmatrix} -i & 1 \\ 1 & i \end{pmatrix}\right) \frac{1}{\sqrt{2}} \begin{pmatrix} i & 1 \\ 1 & i \end{pmatrix} \begin{pmatrix} a^{\frac{1}{2}}\varphi \\ \varepsilon\varphi' \end{pmatrix} \\ &= R^{-1}E_\varepsilon^* \frac{1}{\sqrt{2}\,t_0} \begin{pmatrix} i & 1 \\ 1 & i \end{pmatrix} \left(\mathrm{Id} + \frac{\varepsilon t_1}{2} \begin{pmatrix} -i & 1 \\ 1 & -i \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 2i & 0 \end{pmatrix}\right) \begin{pmatrix} a^{\frac{1}{2}}\varphi \\ \varepsilon\varphi' \end{pmatrix} \\ &= R^{-1}E_\varepsilon^* \frac{1}{\sqrt{2}\,t_0} \begin{pmatrix} i & 1 \\ 1 & i \end{pmatrix} \begin{pmatrix} a^{\frac{1}{2}}\varphi \\ \varepsilon(\varphi' + 2t_1 a^{\frac{1}{2}}\varphi) \end{pmatrix}. \end{aligned}$$

Since it holds

$$\varphi' + 2t_1 a^{\frac{1}{2}}\varphi \;=\; \varphi' + \frac{a'}{4a}\varphi \;=\; \frac{(a^{\frac{1}{4}}\varphi)'}{a^{1/4}}$$

we find

$$z(x) \;=\; \exp\left(-\frac{i}{\varepsilon}\phi^{\varepsilon}(x)\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}\right)\frac{a^{\frac{1}{4}}(x_0)}{\sqrt{2}}\begin{pmatrix}i & 1\\ 1 & i\end{pmatrix}\begin{pmatrix}a^{\frac{1}{4}}\varphi\\ \varepsilon\frac{(a^{\frac{1}{4}}\varphi)'}{a^{1/2}}\end{pmatrix}.$$

Up to the constant factor $a^{\frac{1}{4}}(x_0)$ this is exactly the transformation established in [4]. Hence our WKB–type approach is a generalization of the transformation discussed in the article to the vector valued case.

In order to reproduce the ansatz from [4] with our WKB–type transformation, we make a special choice of the diagonal values of T_1. The result is a transformation matrix T_ε^{-1}, which has a simple structure, is easy to compute, and regular for all $\varepsilon \in \mathbb{C}$. However, in Proposition 3.3.1 (which holds for the (general) vector valued case) ε is restricted to the interval $(0, \varepsilon_1)$ in order to guarantee regularity of T_ε. Hence, naturally the question arises, if it is also possible in the (general) vector valued case to choose the diagonal, such that the matrix T_ε is regular independently of ε? We shall discuss this in sequel for

$$T_\varepsilon \;=\; T_0 + \varepsilon T_1 \;=\; (\mathrm{Id} + \varepsilon T_1 T_0^{-1})T_0$$

where the diagonal of T_1 is arbitrary. Motivated by the above discussion (for the problem from [4]) we shall make the ansatz

$$(\mathrm{Id} - \varepsilon M)^{-1} \;=\; \sum_{j=0}^{n} \varepsilon^j M_j\,,$$

where M is a wild card for $-T_1 T_0^{-1}$ with an arbitrary diagonal part. Multiplication with $(\mathrm{Id} - \varepsilon M)$ yields

$$\begin{aligned}\mathrm{Id} \;&=\; \Big(\sum_{j=0}^{n} \varepsilon^j M_j\Big)(\mathrm{Id} - \varepsilon M)\\ &=\; M_0 + \sum_{j=1}^{n} \varepsilon^j (M_j - M_{j-1}M) + \varepsilon^{n+1} M_n M\,. \qquad (3.50)\end{aligned}$$

The limit $\varepsilon \to 0$ yields $M_0 = \mathrm{Id}$. Since all the remaining coefficients of the ε powers have to be zero, we inductively deduce

$$M_j \;=\; M^j\,, \quad j = 1, \ldots, n \quad \text{and} \quad M^{n+1} \;=\; 0\,. \qquad (3.51)$$

This means M is nilpotent, which cause the Neumann series to terminate after a finite number of summands. We record the essence of the above calculations in the following

Lemma 3.3.7. *Let $M \in \mathbb{C}^{d\times d}$ be a quadratic matrix. It is equivalent*

(i) *For all $\varepsilon \in \mathbb{C}$ the matrix $\mathrm{Id} - \varepsilon M$ is regular,*

(ii) *M is nilpotent.*

Proof. (ii)⇒(i): If M is nilpotent, then the matrix $\sum_{j=0}^{d} \varepsilon^j M^j$ is the inverse of $\mathrm{Id} - \varepsilon M$, as shown above.
(i)⇒(ii): For $\varepsilon \neq 0$ we write

$$\mathrm{Id} - \varepsilon M \quad = \quad \varepsilon \left(\tfrac{1}{\varepsilon} \mathrm{Id} - M \right).$$

Since $\mathrm{Id} - \varepsilon M$ is regular, $\frac{1}{\varepsilon}$ cannot be an eigenvalue of M. Hence 0 is the only eigenvalue of M. Thus the characteristic polynomial of M is $\chi_M(\lambda) = \lambda^d$. This yields $M^d = 0$, which means M is nilpotent. □

Now one can argue that we only have $\varepsilon \in (0, \varepsilon_0) \subset \mathbb{R}$. But also in this situation it is true that the inverse of the matrix $(\mathrm{Id} - \varepsilon M)$ is a polynomial in ε, if and only if M is nilpotent.

Proposition 3.3.8. *Let $\varepsilon_0 > 0$. For $M \in \mathbb{C}^{d \times d}$ it is equivalent:*

(i) *For all $\varepsilon \in (0, \varepsilon_0)$ the matrix $(\mathrm{Id} - \varepsilon M)$ has an inverse of the form*

$$(\mathrm{Id} - \varepsilon M)^{-1} \quad = \quad \sum_{j=0}^{n} \varepsilon^j B_j \,,$$

with ε–independent matrices $B_0, \ldots, B_n \in \mathbb{C}^{d \times d}$ and a fixed $n \in \mathbb{N}$.

(ii) *There exists a constant $c \in \mathbb{C}$, such that for all $\varepsilon \in (0, \varepsilon_0)$ it holds*

$$\det(\mathrm{Id} - \varepsilon M) \quad = \quad c \,.$$

(iii) *There exists a constant $c \in \mathbb{C}$ and $d+1$ pairwise distinct complex numbers $\zeta_1, \ldots, \zeta_{d+1} \in \mathbb{C}$, such that*

$$\det(\mathrm{Id} - \zeta_j M) \;=\; c \,, \qquad j = 1, \ldots, d+1 \,.$$

(iv) *$\det(\mathrm{Id} - \varepsilon M) = 1$ for all $\varepsilon \in \mathbb{C}$.*

(v) *$(\mathrm{Id} - \varepsilon M)$ is regular for all $\varepsilon \in \mathbb{C}$.*

(vi) *M is nilpotent.*

Proof. Obviously it holds (iv) ⇒ (ii) ⇒ (iii). To prove (iii) ⇒ (iv) we remark that $\det(\mathrm{Id} - \varepsilon M) = p(\varepsilon)$ is a polynomial in ε of degree d. Hence the polynomial $p(\varepsilon) - c$ has degree d too, but $d+1$ pairwise distinct roots. Thus (due to the Fundamental Theorem of Algebra, cf. [64]) it has to be zero. Hence $p(\varepsilon) = c$ holds for all $\varepsilon \in \mathbb{C}$. Since $p(0) = \det(\mathrm{Id}) = 1$ we get $c = 1$. By Lemma 3.3.7 and (3.50) (with the discussion that leads to (3.51)) we get (iv) ⇒ (v) ⇔ (vi) ⇔ (i). Finally we prove (v) ⇒ (iv). Since $(\mathrm{Id} - \varepsilon M)$ is regular, the polynomial $p(\varepsilon) = \det(\mathrm{Id} - \varepsilon M)$ has no roots in $\mathbb{C}$. By the Fundamental Theorem of Algebra it has to be constant. □

Remark 3.3.9. *Hence the matrix T_ε is regular independently of ε, iff $T_1 T_0^{-1}$ is nilpotent. The cases in which the diagonal of T_1 can be chosen such that this holds are still not characterized. But for special situations Proposition 3.3.8 can be an additional criteria for the determination of* $\mathrm{diag}(T_1)$.

3.3.2 Comparison with the Super–Adiabatic Transformation by Hairer et al. [27]

In [27] the authors (briefly) discuss a transformation ansatz similar to (3.25) for a special case of our setting, as we shall see in a moment. In the textbook the authors start with the first order system ($v(x) \in \mathbb{C}^d$)

$$v'(x) \quad = \quad \frac{1}{\varepsilon} A(x)\, v(x)\,,$$

where $A(x) \in \mathbb{C}^{d\times d}$ is skew–hermitian for all $x \in [a,b]$. Furthermore they assume that A and its derivatives are bounded independently of ε. Since $A(x)$ is skew–hermitian, there exists a unitary matrix $Q(x)$ and a real valued diagonal matrix $L(x)$, such that $A(x) = iQ(x)^* L(x) Q(x)$. Additionally it is assumed that $A(x)$ has d distinct eigenvalues for all $x \in [a,b]$. Hence the transformation ansatz $u = Qv$ yields (provided that Q is differentiable)

$$u'(x) \quad = \quad \frac{i}{\varepsilon} L(x)u(x) \,+\, Q'(x)Q(x)^* u(x)\,.$$

This is a special case of (3.21) from § 3.2, with $B := Q'Q^*$. By Lemma 3.1.1 we get that $Q'Q^*$ is skew–hermitian.

For all $x \in [a,b]$ let $\Phi_1(x), \ldots, \Phi_n(x) \in \mathbb{C}^{d\times d}$ be real diagonal matrices and $X_1(x), \ldots, X_n(x) \in \mathbb{C}^{d\times d}$ skew–hermitian. They shall be determined in the sequel (at least Φ_1, X_1 to point out the procedure) in order to reproduce the transformation from [27]. We define the matrix valued function

$$\widehat{T}_\varepsilon \quad := \quad \exp(\varepsilon^n X_n) \exp(i\varepsilon^{n-1}\Phi_n) \ldots \exp(\varepsilon^1 X_1) \exp(i\varepsilon^0 \Phi_1)\,.$$

Since each factor is a unitary matrix, so is $\widehat{T}_\varepsilon$. The *super–adiabatic transformation* (SAT) ansatz now reads

$$\widehat{y} \quad = \quad E_\varepsilon^* \widehat{T}_\varepsilon^*\, u\,. \tag{3.52}$$

Here the matrix valued function E_ε is given by (3.26). We slightly changed the notation with respect to [27]. The matrix $X_j(x)$ in the textbook is equivalent to $-X_j(x)$ here. In [27] we find the following relations between X_j, Φ_j that have to be fulfilled for the SAT:

$$-i[L, X_j] \,+\, i\Phi_j' \;=\; W_{j-1}\,, \qquad j = 1, \ldots, n\,. \tag{3.53}$$

The matrix valued function W_0 is equal to $B = Q'Q^*$. For $j \geq 1$ the variables W_j are not specified in the textbook. It is only mentioned that W_j can be built up from the variables up to the index $j-1$. Hence the linear system (3.53) can recursively be solved. If $\{\phi_j, X_j,\, j = 1, \ldots, n\}$ is a solution of the linear system, then the system matrix of the gained equivalent IVP for $\widehat{y}$ is of order $\mathcal{O}(\varepsilon^n)$. In the textbook this fact is briefly described by the relation $\widehat{y}' = \mathcal{O}(\varepsilon^n)$.

For $n = 1$ we shall derive the matrices Φ_1 and X_1 from the ansatz (3.52), instead of using (3.53). This (hopefully) yields a better understanding how (3.53) can be derived and additionally we get the IVP for the transformed quantity $\widehat{y}$. Afterwards we shall compare the $\widehat{y}$–IVP and SAT variables to our WKB–type approach.

If one passes to $n \geq 2$, the equation that determines X_1, Φ_1 (i. e. (3.53) for $j = 1$) remains unchanged. Hence the results from the subsequent computations are valid also for higher order SAT. The following discussion is an elaborated version of the computations from the textbook. Let us denote the (unitary-) matrix valued functions $\exp(i\Phi_1)$ and $\exp(\varepsilon X_1(x))$ by $\widehat{T}_0$ and $\widehat{T}_1$ respectively. Hence the ansatz (3.52) reads for $n = 1$:

$$\widehat{y}(x) \quad = \quad E_\varepsilon(x)^* \, \widehat{T}_0(x)^* \, \widehat{T}_1(x)^* \, u(x) \, . \tag{3.54}$$

Furthermore we set

$$F_1(x) \;:=\; \sum_{j=1}^{\infty} \varepsilon^{j-1} \frac{X_1(x)^j}{j!} \quad \text{and} \quad F_2(x) \;:=\; \sum_{j=2}^{\infty} \varepsilon^{j-2} \frac{X_1(x)^j}{j!} \, .$$

For all $x \in [a, b]$ the matrices $F_1(x), F_2(x)$ are $\mathcal{O}(1)$ as $\varepsilon \to 0$. Since $E_\varepsilon(x)$ and $\widehat{T}_0(x)$ are diagonal matrices they commutate. Thus, differentiating equation (3.54) yields (with Lemma 3.1.1)

$$\begin{aligned} \widehat{y}' \;&=\; E_\varepsilon^* \widehat{T}_0^* \big(-\tfrac{i}{\varepsilon} L \widehat{T}_1^* \,-\, i\Phi_1' \widehat{T}_1^* \,+\, (\widehat{T}_1^*)' \,+\, \widehat{T}_1^* \big(\tfrac{i}{\varepsilon} L \,+\, B \big) \big) \, u \\ &=\; E_\varepsilon^* \widehat{T}_0^* \widehat{T}_1^* \big(-\tfrac{i}{\varepsilon} \widehat{T}_1 L \widehat{T}_1^* \,-\, i\widehat{T}_1 \Phi_1' \widehat{T}_1^* \,+\, \widehat{T}_1 (\widehat{T}_1^*)' \,+\, \tfrac{i}{\varepsilon} L \,+\, B \big) \widehat{T}_1 \widehat{T}_0 E_\varepsilon \, \widehat{y} \\ &=\; E_\varepsilon^* \widehat{T}_0^* \widehat{T}_1^* \big(\tfrac{i}{\varepsilon} \big[L, \widehat{T}_1 \big] \,-\, i\widehat{T}_1 \Phi_1' \,-\, \widehat{T}_1' \,+\, B\widehat{T}_1 \big) \widehat{T}_0 E_\varepsilon \, \widehat{y} \, . \end{aligned}$$

The matrix valued function between the brackets (in the last line) is denoted by $\widehat{S}_1$. Now we use the identities $\widehat{T}_1 = \mathrm{Id} + \varepsilon F_1$ and $\widehat{T}_1 = \mathrm{Id} + \varepsilon X_1 + \varepsilon^2 F_2$ to separate the lowest order terms with respect to ε of $\widehat{S}_1$. We get

$$\begin{aligned} \widehat{S}_1 \;&=\; \tfrac{i}{\varepsilon} \big[L, \mathrm{Id} + \varepsilon X_1 + \varepsilon^2 F_2 \big] \,-\, i \big(\mathrm{Id} + \varepsilon F_1 \big) \Phi_1' \\ &\qquad -\varepsilon F_1' \,+\, B \big(\mathrm{Id} + \varepsilon F_1 \big) \\ &=\; i \big[L, X_1 \big] \,-\, i\Phi_1' \,+\, B \\ &\qquad + \varepsilon \big(BF_1 - iF_1\Phi_1' - F_1' + i[L, F_2] \big) \, . \end{aligned}$$

Hence $\widehat{S}_1$ is of order $\mathcal{O}(\varepsilon)$, if and only if

$$\big[L, X_1 \big] \quad = \quad iB \,+\, \Phi_1' \, . \tag{3.55}$$

By Lemma 3.1.9, there exists a solution of (3.55), if and only if $\mathrm{diag}(iB + \Phi_1') = 0$ and thus $\Phi_1' = -i \, \mathrm{diag}(B)$. Furthermore we deduce that

$$X_1 \quad = \quad D_L^- \odot \big(iB + \Phi_1 \big) \quad = \quad i \, D_L^- \odot B$$

is a (partial) solution. Since B is skew–hermitian, X_1 is skew–hermitian too. We choose the integration constant for Φ_1, such that $\Phi_1(x_0) = 0$, which yields

$$\widehat{T}_0(x) \quad = \quad \exp \left(\int_{x_0}^{x} \mathrm{diag} \big(B(s) \big) \, ds \right) .$$

Hence $\widehat{T}_0$ solves the IVP

$$T_0' \;=\; \mathrm{diag}_\nu(B) \, T_0 \, , \qquad T_0(x_0) \;=\; \mathrm{Id} \, .$$

which coincides with (3.37) from Remark 3.3.2 ($\nu = (1, \ldots, 1)^T \in \mathbb{C}^d$). Thus our quantity T_0 from Remark 3.3.2 and $\widehat{T}_0$ from the SAT are equal. Since B is skew–hermitian, its diagonal has to be purely imaginary. Hence $\widehat{T}_0$ is unitary as assumed. We also find that T_1 from Remark 3.3.2 and X_1 are equal. But the ODE system matrices for the variables y (WKB–type transformation) and $\widehat{y}$ (SAT) differ. For $\widehat{y}$ we get the quite lengthy term

$$\varepsilon\, E_\varepsilon^* \widehat{T}_\varepsilon^* \big(BF_1 - iF_1\Phi_1' - F_1' + i[L, F_2]\big)\widehat{T}_0 E_\varepsilon\,,$$

while the system matrix for y reads (cf. Remark 3.3.2)

$$\varepsilon\, E_\varepsilon^* T_\varepsilon^{-1}\big(BT_1 - T_1'\big)E_\varepsilon\,.$$

The structure of the equations that determine the variables X_j, Φ_j of the SAT (cf. (3.53)) and the equations determining the variables T_j of our WKB–type approach (cf. (3.30) in the proof of Proposition 3.3.1) are very similar. From this point of view the computational effort is the same. But for higher orders, the right–hand side of (3.53) (i. e. W_j) gets more and more involved, while for our approach it is of the same (simple) type for all stages (cf. Remark 3.3.2).

The advantage of the SAT compared to our transformation is that $\widehat{T}_\varepsilon$ is unitary. Hence (as long as $X_1, \ldots, X_n$ are skew hermitian) one does not have to solve a linear system in order to compute the system matrix for $\widehat{y}$. Furthermore, errors on the $\widehat{y}$ level are not enhanced when transforming back to u. However, this advantage is lost, if one extends the SAT to our more general setting from § 3.2. Here $B(x)$ does not have to be skew hermitian. The generalization of the SAT to this case is straight forward. One has to replace * by $^{-1}$ in the previous discussion. Since we have not used that $\widehat{T}_0, \widehat{T}_1$ are unitary, the equations that determine Φ_1 and X_1 remain unchanged. Thus, if B is not skew–hermitian, X_1 is not skew–hermitian either and hence $\widehat{T}_1$ is not unitary. Furthermore (for ε small), the matrix inverse of our transformation matrix $T_\varepsilon(x)$ is given by a von Neumann series. To be more precise it holds

$$\begin{aligned} T_\varepsilon(x)^{-1} &= \left(T_0(x)\left(\mathrm{Id} + \varepsilon T_0^{-1}\sum_{j=1}^{n}\varepsilon^{j-1}T_j\right)\right)^{-1} \\ &= \left(\sum_{k=0}^{\infty}(-\varepsilon)^k\left(T_0^{-1}\sum_{j=1}^{n}\varepsilon^{j-1}T_j\right)^k\right)T_0(x)^{-1}\,. \end{aligned}$$

Hence errors of our WKB–type transformation variable y are only moderately amplified when transforming back to u.

The exclusive usage of unitary matrices for the SAT has the (very little) drawback that one has to compute the matrix exponential $\exp\big(\varepsilon X_1(x)\big)$ (which is the variable $\widehat{T}_1(x)$ in our notation). If one simply truncates the series, the result is not unitary. In this case one implicitly uses an ansatz similar to our approach. Thus, one could have directly started with our transformation. Other methods, like the *Matlab* function `expm`, have to solve a linear matrix equation (cf. [30]). Thus also in this case the computational advantage of the SAT (compared to our approach) is significantly reduced. Moreover one has to find suitable approximations for $F_1(x)$ and $F_2(x)$. One can either use a truncation of the series or the formulas

$$F_1 = \frac{1}{\varepsilon}\big(\widehat{T}_1 - \mathrm{Id}\,\big)\,, \qquad F_2 = \frac{1}{\varepsilon^2}\big(\widehat{T}_1 - \mathrm{Id} - \varepsilon X_1\big)\,.$$

The disadvantage of the formulas is that the error made while deriving $\widehat{T}_1$ is enhanced with a factor ε^{-2}.

From the computational point of view our approach is much simpler for the more general problem from § 3.2, yielding a comparable result to the SAT. Even in the special case of skew–hermitian B (which is considered in [27]) there seems to be no significant draw back of our ansatz compared to the SAT.

3.4 The inhomogeneous case

Let $m \in \mathbb{Z}$ and let $\Omega := I \times (0, \varepsilon_0)$, with $\varepsilon_0 > 0$ and $I \subset \mathbb{R}$ a bounded open (non trivial) interval. In this section we shall consider the inhomogeneous equation

$$u' \;=\; \tfrac{i}{\varepsilon} L u \,+\, B u \,+\, \varepsilon^m f\,.$$

with the assumptions of Proposition 3.3.1 for the matrix valued functions L, B. Additionally it has to hold

Assumption 4. *The functions $L\colon \Omega \to \mathbb{C}^{d\times d}$ and $f\colon \Omega \to \mathbb{C}^d$ are C^{r-m}–bounded independently of ε.*

Assumption 5. *There exists a constant $c_l > 0$ independently of ε, such that for all $(x,\varepsilon) \in \Omega = I \times (0,\varepsilon_0)$ it holds*

$$|l_j(x,\varepsilon)| \;\geq\; c_l\,, \qquad \text{for } j = 1,\ldots,s\,.$$

A consequence of Assumption 5 is

Lemma 3.4.1. *The matrix valued function L is regular for all $(x,\varepsilon) \in \Omega$. Furthermore, L^{-1} is $C^{r'}$–bounded independently of ε, with $r' = \max(r, r-m)$.*

Proof. Since the eigenvalues of L are bounded away from zero, L is obviously regular for all $(x,\varepsilon) \in \Omega$ and it holds $\|L^{-1}\| \leq \frac{1}{c_l}$. From Corollary 3.1.2 we get for every fixed $\varepsilon \in (0,\varepsilon_0)$, that $L^{-1}(\cdot,\varepsilon) \in C^r$. Furthermore it holds

$$\|(L^{-1})'\| \;=\; \|-L^{-1}L'L^{-1}\| \;\leq\; \tfrac{1}{c_l^2}\,\|L'\|\,.$$

Since L is C^r–bounded independently of ε, we get $(L^{-1})'$ is bounded independently of ε. By induction (using (3.5)) it follows that $(L^{-1})^{(j)}$ is bounded independently of ε for $j = 1,\ldots,r$. □

Now we can prove the main result of this section.

Proposition 3.4.2. *Let the assumptions of Proposition 3.3.1 hold and let $T_0,\ldots,T_n,S_n$ be the matrix valued function from the Proposition. Additionally let Assumption 4, 5 hold. Then there are vector valued functions $g_{m+1},\ldots,g_n$, such that the new variable y defined by* (cf. (3.25)–(3.24))

$$y \;:=\; E_\varepsilon^* T_\varepsilon^{-1}(u-g) \quad \textit{with} \quad g \;:=\; \sum_{j=m+1}^{n} \varepsilon^j g_j$$

solves the inhomogeneous ODE

$$y' \;=\; \varepsilon^n\, E_\varepsilon^* S_n E_\varepsilon\, y \,+\, \varepsilon^n E_\varepsilon^* \widehat{f}\,. \tag{3.56}$$

The function $\widehat{f}$ is C^{r-m}–bounded independently of ε.

Proof. Let $E_\varepsilon, T_\varepsilon$ be given by Remark 3.3.2 in § 3.3. It holds

$$\big((E_\varepsilon^* T_\varepsilon^{-1})' + E_\varepsilon^* T_\varepsilon^{-1}(\tfrac{i}{\varepsilon}L + B)\big)(E_\varepsilon^* T_\varepsilon^{-1})^{-1} \quad = \quad \varepsilon^n E_\varepsilon^* S_n E_\varepsilon$$

and the ansatz $y = E_\varepsilon^* T_\varepsilon^{-1}(u - g)$ yields

$$\begin{aligned} y' &= (E_\varepsilon^* T_\varepsilon^{-1})'(u-g) + (E_\varepsilon^* T_\varepsilon^{-1})\big((\tfrac{i}{\varepsilon}L + B)u + \varepsilon^m f - g'\big) \\ &= \varepsilon^n E_\varepsilon^* S_n E_\varepsilon\, y + (E_\varepsilon^* T_\varepsilon^{-1})(\varepsilon^m f + (\tfrac{i}{\varepsilon}L + B)g - g')\,. \end{aligned}$$

To increase the ε–order of the inhomogeneity we make the ansatz[3]

$$g(x) \quad = \quad \sum_{j=m+1}^{n} \varepsilon^j g_j(x)\,,$$

which yields with $g_{n+1} := 0$

$$\varepsilon^m f + (\tfrac{i}{\varepsilon}L + B)g - g' \quad = \quad \varepsilon^m\big(f + iLg_{m+1}\big) + \sum_{j=m+1}^{n} \varepsilon^j\big(Bg_j - g_j' + iLg_{j+1}\big)\,.$$

We set $g_{m+1} = iL^{-1}f$ and for $j = m+2, \ldots, n$

$$g_j \quad = \quad iL^{-1}(Bg_{j-1} - g_{j-1}')\,.$$

This yields

$$y' \quad = \quad \varepsilon^n E_\varepsilon^* S_n E_\varepsilon\, y + \varepsilon^n E_\varepsilon^* T_\varepsilon^{-1}(g_n' - Bg_n)\,.$$

By Lemma 3.4.1 L^{-1} is $C^{r'}$ bounded independently of ε. Hence g_{m+1} is C^{r-m_-} bounded independently of ε. This yields (by induction) that g_j is C^{r-j+1_-} bounded independently of ε for $j = 1, \ldots, n$. Hence the (non oscillatory) vector valued function

$$\widehat{f} \quad := \quad T_\varepsilon^{-1}(g_n' - Bg_n)$$

of the inhomogeneity in ODE (3.56) is C^{r-n}–bounded independently of ε. □

Remark 3.4.3. *From the proof of Proposition 3.4.2 we get that the functions g_j are constructed as follows: We set $g_{m+1} := iL^{-1}f$ and for $j = m+2, \ldots, n$*

$$g_j \quad = \quad iL^{-1}(Bg_{j-1} - g_{j-1}')\,.$$

The function $\widehat{f}$ from (3.56) *is given by $\widehat{f} = T_\varepsilon^{-1}(g_n' - Bg_n)$.*

3.5 WKB approximation

The *WKB–Method* or *Phase Integral Method* is a technique which became popular with the rise of the quantum mechanics. It was used by Wentzel, Kramers and Brillouin in the 1920s to find approximate solutions of the Schrödinger equation (cf. [32]). The basic idea is the assumption that the fast variation of

[3] For $m \geq n$ we have an empty sum and hence $g = 0$.

the solution of a linear singular perturbed ODE[4], like the stationary one dimensional Schrödinger equation in the oscillatory regime, is of exponential nature. For more than a century the WKB–Method is used in different fields of physics like quantum or solid mechanics to approximate solutions of singular perturbed linear ODE. Hence it is not surprising that one finds two different definitions of it in literature. One definition (cf. [6, 48]) for the scalar problem

$$\varepsilon^2\psi''(x) + V(x)\psi(x) \quad = \quad 0 \tag{3.57}$$

is given by an expansion of the form

$$\psi(x) \quad \sim \quad \exp\left(\tfrac{i}{\varepsilon}\sum_{j=0}^{\infty}\varepsilon^j\phi_j(x)\right). \tag{3.58}$$

The (formal) ansatz $\psi = e^{\frac{i}{\varepsilon}\int\phi dx}$ leads to the Riccati equation

$$i\varepsilon\phi' \quad = \quad \phi^2 - a\,.$$

Hence the above given expansion for ψ corresponds to an asymptotic approximation of the nonlinear first order ODE.

Another approach (cf. [32]) is to find a phase function ϕ and an asymptotic expansion[5] of ψ in the sense of § 3.6, i. e.

$$\psi(x) \quad \sim \quad \left(\sum_{j=0}^{\infty}\varepsilon^j\, c_j(x)\right)\exp\left(\tfrac{i}{\varepsilon}\phi(x)\right). \tag{3.59}$$

We call the second ansatz (3.59) WKB–Method and the first one (3.58) physical WKB–Method, due to the appearance of this ansatz in almost all physical textbooks dealing with this topic.

In §2.2 we present a technique to transform the ODE (3.57) to a first order system of the form

$$u'(x) \quad = \quad \frac{i}{\varepsilon}L(x)u(x) \,+\, B(x)u(x)\,.$$

Hence, also u has a WKB approximation and it is quite natural to ask if for more general matrix valued functions L, B an asymptotic approximation of u exists. There is of course a positive answer. In the sequel we shall derive a WKB–type approximation for a fundamental system of solutions U of ODE (3.21) on the bounded interval $I = [a, b]$. I. e. the (square) matrix valued function U solves

$$U'(x) \quad = \quad \frac{i}{\varepsilon}L(x)U(x) \,+\, B(x)U(x) \tag{3.60}$$

and is regular for all $x \in I$.

As in the previous sections, the matrix L is diagonal with

$$L \quad = \quad \mathrm{diag}(l_1\,\mathrm{Id}_{\nu_1}, \ldots, l_s\,\mathrm{Id}_{\nu_s})$$

[4]At least the highest derivative is multiplied by a small parameter, which significantly changes the behavior of the ODE if set to zero.

[5]To be more precise this means $\psi e^{-\frac{i}{\varepsilon}\phi} \sim \sum_j \varepsilon^j c_j$

with pairwise distinct $l_j(x)$ for every $x \in I$. Furthermore the matrix functions L, B are assumed to be smooth and we set $\nu = (\nu_1, \ldots, \nu_s)^T \in \mathbb{N}^s$. Since two fundamental systems of solutions are equal up to the multiplication with a regular constant matrix from the right–hand side we focus, without restriction of generality, on a solution with $U(x_0) = \mathrm{Id}$ for some $x_0 \in I$.

Since the WKB–ansatz (3.59) can directly be generalized to the vector case and structurally yields the same equations (for the second order ODE) to solve as in the scalar case its our method of choice. The physical WKB–ansatz yields more problems, since generally $\exp(A)' \neq A' \exp(A)$ for a matrix valued function $A \in C^1(I, \mathbb{C}^{\nu\times\nu})$. How it can be applied or modified for the vector case is not yet clear, but it seems to be connected to the so called super–adiabatic transformations briefly discussed in [27] and § 3.3.2.

The basic strategy to find an approximation for U is based on the variation of constants principle. Assume we have given two matrix valued functions U_{wkb} and S_ε such that it holds for all $x \in I$:

$$U'_{\mathrm{wkb}} \;=\; \frac{i}{\varepsilon} L U_{\mathrm{wkb}} \,+\, B U_{\mathrm{wkb}} \,+\, S_\varepsilon\,, \qquad U_{\mathrm{wkb}}(x_0) \;=\; U(x_0)\,,$$

for some $x_0 \in I$. Hence $U_{\mathrm{wkb}} - U$ solves the same inhomogeneous IVP as U_{wkb}, but with trivial initial data. Since U is a fundamental system of solutions of the homogeneous equation, we get by variation of constants (cf. Lemma 8.4.1)

$$U_{\mathrm{wkb}}(x) - U(x) \;\;=\;\; U(x) \int_{x_0}^{x} U^{-1}(\xi) S_\varepsilon(\xi)\, d\xi\,.$$

With Lemma 3.5.2 we deduce from the previous equation

$$\|U_{\mathrm{wkb}} - U\| \;\;\leq\;\; c\,\|S_\varepsilon\|\,, \tag{3.61}$$

with a constant c independently of ε. If $\|S_\varepsilon\| \ll 1$, e. g. $\|S_\varepsilon\| = \mathcal{O}(\varepsilon^\alpha)$ with $\alpha \geq 1$, then U_{wkb} is a good approximation for U.

Remark 3.5.1. *Due to the estimate* (3.61) *we shall use the following strategy to determine a suitable approximation* U_{wkb}*.*

(i) *make a suitable ansatz for* U_{wkb} *(motivated by* (3.59)*)*

(ii) *insert it into the homogeneous ODE* (3.60)

(iii) *determine the free parameter from the ansatz function such that the remainder* S_ε *is getting small*

In the next Lemma 3.5.2 we collect some properties of U in order to get an idea what are natural assumptions for the desired WKB approximation U_{wkb}.

Lemma 3.5.2. *Let* U *be a fundamental system of solutions of the IVP* (3.21)*. Then it holds:*

(i) U *is regular on* I *and* U, U^{-1} *are continuously differentiable.*

(ii) $\exists\; c > 0$ *independently of* ε*, such that:* $\|U\|, \|U^{-1}\| \leq c\,.$

Proof. (i) U is regular on I (cf. [2]) and hence we can apply Corollary 3.1.2.
(ii) To derive a bound of U we introduce a new quantity $Y := E^*U$ with

$$E = \exp\left(\tfrac{i}{\varepsilon}\int_{x_0}^{x} L(\xi)\, d\xi\right).$$

Differentiation yields

$$\begin{aligned} Y' &= -\tfrac{i}{\varepsilon}LY + E^*\big(\tfrac{i}{\varepsilon}L + B\big)EY \\ &= E^*BE\,Y\,. \end{aligned}$$

Integration of the differential equation yields

$$\|Y(x)\| \;\leq\; \|Y_0\| + \left|\int_{x_0}^{x} \|E^*(\xi)\|\|B(\xi)\|\|E(\xi)\|\, d\xi\right|.$$

Since L is real, E is unitary and hence $\|E^*\| = \|E\| = 1$. By a Gronwall argument (cf. Lemma 8.4.3) we get

$$\|Y(x)\| \;\leq\; \|Y_0\|\, e^{\left|\int_{x_0}^{x} \|B(s)\|\, ds\right|}.$$

The smooth matrix B is bounded independently of ε. Since I is a bounded interval, there exists a constant $c > 0$ independently of ε, such that for all $x \in I$

$$\|U(x)\| \;\leq\; \|Q^{-1}(x)\|\|E(x)\|\|Y(x)\| \;\leq\; c\,.$$

To show the existence of an ε independent bound for $\|U(x)^{-1}\|$ we use the fact that U^{-1} is differentiable. Equation (3.5) yields

$$\big(U^{-1}\big)' \;=\; -U^{-1}\big(\tfrac{i}{\varepsilon}L + B\big)\,.$$

Hence a similar calculation as done for U yields a bound for U^{-1}. □

In the following Definition 3.5.3 we specify our WKB ansatz. Since we want to approximate a fundamental system of solutions of ODE (3.60), we shall demand similar properties for the ansatz function as listed in Lemma 3.5.2.

Definition 3.5.3. *Let $n \in \mathbb{N}$ and let $C_0, \ldots, C_n, \Phi\colon I \to \mathbb{C}^{d\times d}$ independently of ε with $\Phi(x) \in \mathbb{R}^{d\times d}$ diagonal for all $x \in I$. Then the matrix function*

$$W(x) \;:=\; \left(\sum_{j=0}^{n} \varepsilon^j\, C_j(x)\right) E_\varepsilon(x) \quad \textit{with} \tag{3.62}$$

$$E_\varepsilon(x) \;:=\; \exp\big(\tfrac{i}{\varepsilon}\Phi(x)\big) \tag{3.63}$$

is called a WKB ansatz function of degree n *for the ODE* (3.60)*, if and only if*

(i) *all quantities are continuously differentiable on I,*

(ii) *W is regular on I,*

(iii) $\exists\, c > 0\ \exists\, \varepsilon_0 > 0\ \forall\, \varepsilon \in (0, \varepsilon_0)\colon\ \|W^{-1}\| \leq c.$

From Corollary 3.1.2 we know that W^{-1} is differentiable and hence we do not have to claim it additionally. Due to (i) and since Φ is a real diagonal matrix, we easily see that $\|W\|$ is bounded independently of $\varepsilon \in (0, \varepsilon_0)$ for some $\varepsilon_0 > 0$. Since W is assumed to be regular, the same holds for $\sum \varepsilon^j C_j$. It is well known that a sufficient condition for the sum being regular is that C_0 is regular and ε small. By the following Lemma 3.5.4 this is also necessary for a WKB ansatz function.

Lemma 3.5.4. *Let W be a WKB ansatz function of degree n. Then the following conditions are equivalent*

(i) $\exists\, c > 0\ \exists\, \varepsilon_0 > 0\ \forall\, \varepsilon \in (0, \varepsilon_0)\colon\ \|W^{-1}\| \le c$.

(ii) *$C_0(x)$ is regular for all $x \in I$.*

Proof. (ii) $\Rightarrow$ (i): Let C_0 be regular on I and $0 \neq v \in \mathbb{C}^d$. For any regular matrix $M \in \mathbb{C}^{d\times d}$ it holds

$$\|v\| = \|M^{-1}Mv\| \quad \Rightarrow \quad \|Mv\| \ge \frac{1}{\|M^{-1}\|}\|v\|\,, \tag{3.64}$$

which yields

$$\|W(x)v\| \quad \ge \quad \left(\frac{1}{\|(C_0(x))^{-1}\|} - \varepsilon \sum_{j=1}^{n} \varepsilon^{j-1}\|C_j(x)\|\right)\|E_\varepsilon(x)v\|\,.$$

Since Φ is real, E_ε is unitary and hence $\|E_\varepsilon(x)v\| = \|v\|$. Since all quantities are continuous on the compact interval I, there are constants $\varepsilon_0, \tau > 0$, such that the right–hand side is strictly positive for all $\varepsilon \in (0, \varepsilon_0)$, i. e.

$$\|W(x)v\| \quad \ge \quad \tau\|v\|\,.$$

Hence $W(x)$ is injective and consequently regular. Furthermore we compute

$$\|W^{-1}(x)\| \quad = \quad \sup_{v\neq 0} \frac{\|v\|}{\|W(x)v\|} \quad \le \quad \frac{1}{\tau}\,.$$

The boundedness of $\|W(x)\|$ is clear due to the definition of W.

(i) $\Rightarrow$ (ii) is proven by contradiction. Assume there exists an $x \in I$, such that $C_0(x)$ is not regular and let $v \in \mathbb{C}^d$. Since $W(x)$ is regular, we can find for any prescribed v a vector u, such that $W(x)u = v$. It follows with $C_0(x) = TJT^{-1}$ (Jordan normal form)

$$W(x)u \;=\; v \quad \Leftrightarrow \quad T^{-1}v \;=\; Jw + \varepsilon Bw\,,$$

with

$$B \;=\; \sum_{j=1}^{n} \varepsilon^{j-1} T^{-1} C_j T \quad \text{and} \quad w = T^{-1}u\,.$$

Since $C_0(x)$ is assumed not to be regular, we can assume without restriction that the last row of J is identically zero. Let $\varepsilon \in (0, \varepsilon_0)$ and choose v, such that

$T^{-1}v = (0,\dots,0,1)^T \in \mathbb{C}^d$. Since $C_0(x)$ does not depend on ε, the same holds for the vector v. Hence we get

$$\begin{aligned} 1 &= \varepsilon \left| \sum_{j=0}^{d} B_{dj} w_j \right| \leq \varepsilon \, \|B\|_\infty \|w\|_\infty \\ &\leq \varepsilon \, c \, \|T\| \|T^{-1}\| \left(\sum_{j=0}^{N} \varepsilon_0^{j-1} \|C_j\| \right) \|T^{-1} W^{-1} v\| \leq c \, \varepsilon \, \|W^{-1}\| \,. \end{aligned}$$

Here we denote by $\|\cdot\|_\infty$ the ∞–norm on $\mathbb{C}^d$, which is at once replaced by the euclidean norm due to the equivalence of norms on $\mathbb{C}^d$ (cf. [68]). Since c is an ε–independent constant, $\|W^{-1}(x)\|$ cannot be bounded as $\varepsilon \to 0$, which is a contradiction. □

If L, B are sufficiently smooth, the following Lemma 3.5.5 guarantees the existence of a WKB ansatz function, such that (3.60) is solved up to a remainder of order $\mathcal{O}(\varepsilon^n)$. But without prescribing initial conditions. Our strategy to construct an approximation U_{wkb} for an IVP is as follows: From the constructive proof of Lemma 3.5.5 we extract a special set of WKB ansatz functions $\mathbb{W}_0, \dots, \mathbb{W}_n$ of orders $0, \dots, n$ (cf. Corollary 3.5.7). Afterwards we prove in Lemma 3.5.9 that there exists a unique linear combination W which approximates the IVP up to a remainder of $\mathcal{O}(\varepsilon^n)$. The function W is in general not a WKB ansatz function.

Lemma 3.5.5. *Let $m \geq n$ and $L, B \in C^m(I, \mathbb{C}^{d\times d})$. Then there exists an $\varepsilon_0 > 0$ and a WKB ansatz function W of degree n, such that*

$$W' - \tfrac{i}{\varepsilon} L W - B W \;=\; \varepsilon^n \, S \, E_\varepsilon \,.$$

The matrix function S is given by

$$S \;=\; C_n' - B C_n$$

and hence is independently of ε. Furthermore there exists a permutation matrix P, such that it holds for all $x \in I$:

$$\Phi'(x) \;=\; P^* L(x) P \,. \tag{3.65}$$

The matrix coefficient functions are smooth. To be more precise it holds for $j = 0, \dots, n$: $C_j \in C^{m-j+1}(I, \mathbb{C}^{d\times d})$, which yield $S \in C^{m-n}$.

Proof. We formally compute:

$$W' - \tfrac{i}{\varepsilon} L W - B W \;=\; \sum_{j=0}^{n} \big(C_j' + \tfrac{i}{\varepsilon} C_j \Phi' - \tfrac{i}{\varepsilon} L C_j - B C_j \big) \varepsilon^j \, E_\varepsilon \,. \tag{3.66}$$

Now the idea is to determine $C_0, \dots, C_n$, such that the coefficient matrix in front of the factor ε^j is zero up to ε^{n-1}. This yields for $j \in \{0, \dots, n\}$

$$i \, C_j \Phi' \,-\, i \, L C_j \,+\, C_{j-1}' \,-\, B C_{j-1} \;=\; 0 \,, \tag{3.67}$$

where we set $C_{-1} := 0$. For $j = 0$ we get

$$C_0 \Phi' - L C_0 \;=\; 0 \,.$$

Due to Lemma 3.5.4 C_0 has to be regular, which yields with Lemma 3.1.8 that

$$\Phi'(x) = P^* L(x) P\,, \tag{3.68}$$

with an arbitrary permutation matrix P which has to be constant; otherwise we would get a discontinuous jump on the diagonal of Φ'.

Multiplying (3.67) from the right–hand side with P^* yields for $j = 0, \ldots, n$

$$i[L, \widehat{C}_j] \,+\, B\,\widehat{C}_{j-1} \,-\, \widehat{C}'_{j-1} \;\;=\;\; 0\,, \tag{3.69}$$

where we set $\widehat{C}_j = C_j P^*$. A comparison of equation (3.69) with (3.30) from the proof of Lemma 3.3.1 yields that both systems of equations are equal, i. e. $\widehat{C}_j = T_j$ for $j = 0, \ldots, n$. Hence for $j = 0, \ldots, n$

$$\widehat{C}_j^{\mathrm{off}_\nu} \;\;=\;\; i\,D_L^- \odot \big(B\widehat{C}_{j-1} - \widehat{C}'_{j-1}\big)\,, \tag{3.70}$$

$$\widehat{C}_j^{\mathrm{dia}_\nu}(x) \;\;=\;\; \mathcal{T}(x)\bigg(\widehat{C}_j^{\mathrm{dia}_\nu}(x_0) \,+\, \int_{x_0}^{x} \mathcal{T}(\xi)^{-1}\,\mathrm{diag}_\nu(B\,\widehat{C}_j^{\mathrm{off}})(\xi)\;d\xi\bigg), \tag{3.71}$$

where $\mathcal{T}$ is the unique solution of the IVP

$$\mathcal{T}' \;\;=\;\; \mathrm{diag}_\nu(B)\,\mathcal{T}\,, \qquad \mathcal{T}(x_0) \;\;=\;\; \mathrm{Id}\;.$$

Since $\widehat{C}_{-1} = 0$, we have $\widehat{C}_0(x) = \mathcal{T}(x)\widehat{C}_0^{\mathrm{dia}_\nu}(x_0)$ and hence (due to Lemma 3.5.4) $C_0^{\mathrm{dia}_\nu}(x_0)$ has to be a regular matrix. Thus we have an explicit recurrence relation for $\widehat{C}_0, \ldots, \widehat{C}_n$. And by construction it is $\mathcal{T} \in C^{m+1}(I)$ and hence $\widehat{C}_j \in C^{m-j+1}(I)$. Thus there exists an $\varepsilon_0 > 0$, such that $\sum \varepsilon^j \widehat{C}_j$ is regular on I for all $\varepsilon \in (0, \varepsilon_0)$. For the rest of the proof let $\varepsilon \in (0, \varepsilon_0)$. Since all quantities are continuously differentiable, the formal derivation of (3.67) is justified.

Going back to equation (3.66) we get

$$W' \,-\, \tfrac{i}{\varepsilon} LW \,-\, BW \;\;=\;\; \varepsilon^n \,\big(C'_n - BC_n\big)E_\varepsilon$$

and hence

$$S \;\;=\;\; C'_n - BC_n\,.$$

The matrix function S is (obviously) ε–independent and $S \in C^{m-n}$, which completes the proof. □

In the above proof we derived an explicit recurrence relation for the matrix functions $C_0, \ldots, C_n$. As we have seen the matrix valued functions $C_0, \ldots, C_n$, as well as the permutation matrix P are not unique. In order to characterize all WKB ansatz functions and derive an approximation for U we shall choose a special set $\mathcal{C}_0, \ldots, \mathcal{C}_m$ of matrix coefficient functions.

Definition 3.5.6. *Since L was already assumed to be diagonal, we shall set $P = \mathrm{Id}$ and let $\mathcal{T}$ be the unique solution of the IVP*

$$\mathcal{T}' \;\;=\;\; \mathrm{diag}_\nu(B)\,\mathcal{T}\,, \qquad \mathcal{T}(x_0) \;\;=\;\; \mathrm{Id}\,, \tag{3.72}$$

and let

$$\Phi(x) \;\;:=\;\; \int_{x_0}^{x} L(\xi)\;d\xi\,. \tag{3.73}$$

Furthermore we set $\mathcal{C}_0 := \mathcal{T}$ *and define* $\mathcal{C}_1, \dots, \mathcal{C}_n$ *by the following recurrence*

$$\mathcal{C}_j^{\mathrm{off}_\nu} = i\, D_L^- \odot \left(B\mathcal{C}_{j-1} - \mathcal{C}_{j-1}'\right) \tag{3.74}$$

$$\mathcal{C}_j^{\mathrm{dia}_\nu}(x) = \mathcal{T}(x) \int_{x_0}^{x} \mathcal{T}(\xi)^{-1}\, \mathrm{diag}_\nu(B\,\mathcal{C}_j^{off})(\xi)\, d\xi\,. \tag{3.75}$$

With these definitions we get from the proof of Lemma 3.5.5

Corollary 3.5.7. *Let* $\Phi, \mathcal{T}$ *and* $\mathcal{C}_0, \dots, \mathcal{C}_m$ *be given by equations* (3.72)–(3.75). *Then the special WKB ansatz functions*

$$\mathbb{W}_k(x) := \sum_{j=0}^{k} \varepsilon^j \mathcal{C}_j(x)\, E_\varepsilon(x)\,, \qquad k = 0, \dots, m$$

satisfy

$$\mathbb{W}_k' - \tfrac{i}{\varepsilon} L\mathbb{W}_k - B\mathbb{W}_k = \varepsilon^k S_k E_\varepsilon\,, \tag{3.76}$$

with

$$S_k = \mathcal{C}_k' - B\mathcal{C}_k\,.$$

Remark 3.5.8. *If* (3.60) *originates from a second order ODE as discussed in* § 2.2.1*, then* $\mathcal{T}$ *from Definition 3.5.6 is given by*

$$\mathcal{T}(x) = L(x)^{\frac{1}{4}} L(x_0)^{-\frac{1}{4}}\,.$$

This factor corresponds to the non oscillatory "amplitude" in the first–term WKB approximation (3.2).

As stated in the beginning of this section the basic idea to derive an approximation U_{wkb} of U is to find an approximate solution for the ODE (3.60) with suitable initial conditions. Since the ODE is linear we can use a linear combination of the special WKB ansatz functions to construct U_{wkb}. Lemma 3.5.9 is even a stronger result. Every approximate solution of ODE (3.60) (with remainder of order $\mathcal{O}(\varepsilon^{n'})$) can be uniquely approximated by a linear combination of $\mathbb{W}_0, \dots, \mathbb{W}_m$, up to a remainder of order $\mathcal{O}(\varepsilon^{\min(m,n')})$.

Lemma 3.5.9. *Let the matrix valued functions* $L, B \in C^m(I, \mathbb{C}^{d\times d})$ *and let* $x_0 \in I$, $\varepsilon_0 > 0$ *and* $n' \in \mathbb{N}$. *Further let* $V\colon I \times (0, \varepsilon_0) \to \mathbb{C}^{d\times d}$ *be continuously differentiable in the first variable for every fixed* $\varepsilon \in (0, \varepsilon_0)$ *and let*

$$V(x_0, \varepsilon) = \sum_{j=0}^{n'} \varepsilon^j V_j\, \exp\left(\tfrac{i}{\varepsilon}\Phi_0\right) + \mathcal{O}(\varepsilon^{n+1})\,,$$

with matrices $V_0, \dots, V_{n'} \in \mathbb{C}^{d\times d}$ *and* $\Phi_0 \in \mathbb{R}^{d\times d}$ *diagonal. Furthermore we assume that there exists a constant* $c' > 0$, *such that for all* $(x, \varepsilon) \in I \times (0, \varepsilon_0)$

$$\|V'(x) - \tfrac{i}{\varepsilon} L(x)V(x) - B(x)V(x)\| \leq c'\, \varepsilon^{n'}\,.$$

Then there exists unique matrices $X_0, \dots, X_n$ *and a constant* $c > 0$ *independently of* ε *such that*

$$\|V - W\| \leq c\, \varepsilon^n$$

with $n := \min(m, n')$ *and*

$$W(x) \;:=\; \sum_{k=0}^{n} \varepsilon^{n-k} \mathbb{W}_k(x)\, X_k\, \exp\left(\tfrac{i}{\varepsilon}\Phi_0\right). \tag{3.77}$$

Proof. We start with determining the matrices $X_0, \ldots, X_n$. The approximation W has to coincide with V at x_0 up to order $\mathcal{O}(\varepsilon^n)$ which leads to

$$\sum_{j=0}^{n} \varepsilon^j V_j \;=\; \sum_{k=0}^{n} \varepsilon^{n-k} \mathbb{W}_k(x_0)\, X_k \;=\; \sum_{j=0}^{n} \varepsilon^j \Big(\sum_{s=0}^{j} \mathcal{C}_{j-s}(x_0) X_{n-s} \Big).$$

Since this has to be true for all $\varepsilon \in (0, \varepsilon_0)$, we get the following linear system where we write $\mathcal{C}_j$ instead of $\mathcal{C}_j(x_0)$:

$$\begin{pmatrix} \mathcal{C}_0 & & \\ \vdots & \ddots & \\ \mathcal{C}_n & \ldots & \mathcal{C}_0 \end{pmatrix} \begin{pmatrix} X_n \\ \vdots \\ X_0 \end{pmatrix} \;=\; \begin{pmatrix} V_0 \\ \vdots \\ V_n \end{pmatrix}.$$

By definition $\mathcal{C}_0(x_0) = \mathrm{Id}$ and hence there exists unique solutions $X_0, \ldots, X_n$. Let R_V be the remainder of ODE (3.60) with respect to V. Due to the assumptions it is $\|R_V\| \le c' \varepsilon^{n'}$. This yields

$$(W - V)' \,-\, \tfrac{i}{\varepsilon} L(W - V) \,-\, B(W - V) \;=\; \varepsilon^n \sum_{k=0}^{n} S_k E_\varepsilon X_k \,+\, R_V$$

and we get with variation of constants

$$\|V - W\| \;\le\; c\, \varepsilon^n \left\| \sum_{k=0}^{n} S_k E_\varepsilon X_k \right\| \,+\, c' \varepsilon^{n'}.$$

To finish the proof we remark that $n \le n'$. □

A direct consequence of Lemma 3.5.9 is that any WKB ansatz function can uniquely be represented by a "linear combination" of $\mathbb{W}_0, \ldots, \mathbb{W}_m$. For our fundamental system of solutions U we derive

Corollary 3.5.10. *Let U be a fundamental system of solutions of* (3.60) *such that $U(x_0) = \mathrm{Id}$. Then there exists unique matrices $X_0, \ldots, X_m$ such that*

$$\|U - U_{\mathrm{wkb}}\| \;\le\; c\, \varepsilon^m,$$

with a constant c independently of ε and U_{wkb} given by (3.77).

Vice versa, there exists for any WKB ansatz function W a fundamental system of solutions U such that the difference between W and U is of the same order as the residuum of W with respect to the ODE (3.60).

Corollary 3.5.11. *For any WKB ansatz function W of degree $n \le m$ with*

$$W' \,-\, \tfrac{i}{\varepsilon} LW \,-\, BW \;=\; \varepsilon^r S E_\varepsilon,$$

there exists a fundamental system of solutions U to (3.60)*, such that*

$$\|U - W\| \le c\, \varepsilon^n,$$

with a constant $c > 0$ independently of ε.

Proof. The fundamental system of solutions is determined by the following condition:

$$U(x_0) \;=\; \sum_{j=0}^{n} \varepsilon^j C_j(x_0)\,.$$

By Definition 3.5.3, W is regular on I which yields that $U(x_0)$ is regular too. □

Finally we discuss the connection between the WKB approximation W and the derived transformation T_ε in § 3.3. In the proof of Lemma 3.5.5 we already observed that the matrix functions T_j and $\mathcal{C}_j$ solve the same equations.

Remark 3.5.12. *Let u be the unique solution of the IVP* (3.23) *and let the matrix valued function $W \in C^1(I)$ be regular for all $x \in I$ and satisfy*

$$W' \;=\; \tfrac{i}{\varepsilon} L W \,+\, B W \,+\, \tilde{S}$$

with a matrix valued function S. Then the new quantity

$$y \;:=\; W^{-1}\, u$$

is the unique solution of the IVP

$$y' \;=\; -(W^{-1}\tilde{S})\, y\,, \qquad y(x_0) \;=\; W^{-1}(x_0) u_0\,. \tag{3.78}$$

Is W a WKB ansatz function, then W^{-1} is bounded independently of ε and hence the norm of the system matrix of ODE (3.78) *is of order $\mathcal{O}(\|\tilde{S}\|)$.*

Since the matrix valued functions $T_0, \ldots, T_n$ from Remark 3.3.2 coincide with $\mathcal{C}_0, \ldots, \mathcal{C}_n$ from Definition 3.5.6, we get for the special choice $W = \mathbb{W}_n$ the IVP (3.40) *from* § 3.3*, i. e. we have*

$$y' \;=\; \varepsilon^n E_\varepsilon^* S_n E_\varepsilon\, y\,, \qquad y(x_0) = y_0\,.$$

Thus in § 3.3 *we constructed $\mathbb{W}_n = T_\varepsilon E_\varepsilon$.*

Proof. Differentiation yields with $(W^{-1})' = -W^{-1}W'W^{-1}$

$$\begin{aligned} y' \;&=\; -W^{-1}W'W^{-1}u \,+\, W^{-1}\big(\tfrac{i}{\varepsilon}A + B\big)u \\ &=\; -W^{-1}\big[(\tfrac{i}{\varepsilon}A + B)W \,+\, S\big]W^{-1}u \,+\, W^{-1}\big(\tfrac{i}{\varepsilon}A + B\big)u \\ &=\; -\,(W^{-1}S)\, y\,. \end{aligned}$$

Since $T_0, \ldots, T_n$ from Remark 3.5.12 coincide with $\mathcal{C}_0, \ldots, \mathcal{C}_n$ it is $T_\varepsilon E_\varepsilon = \mathbb{W}_n$ and hence we get from Lemma 3.5.5

$$\begin{aligned} -\mathbb{W}_n^{-1}\tilde{S} \;&=\; -(T_\varepsilon E_\varepsilon)^{-1}\varepsilon^n(\mathcal{C}_n' \,-\, B\mathcal{C}_n)E_\varepsilon \\ &=\; \varepsilon^n\, E_\varepsilon^* T_\varepsilon^{-1}(BT_n - T_n')E_\varepsilon \;=\; \varepsilon^n\, E_\varepsilon^* S_n E_\varepsilon\,. \end{aligned}$$

□

3.6 Asymptotic expansions

In § 3.5 and in the numerical part about the oscillatory integrals we come close to the field of asymptotic analysis. Due to this we follow the textbook of Holmes [32] and use this section to give a brief introduction of the basic definitions and give some simple examples to illustrate the concept.

To warm–up we start with the repetition of the order symbols. Therefor we define the term neighborhood. In the sequel a neighborhood of $x \in \mathbb{R}$ always denotes an open subset of $\mathbb{R}$ which contains x. Since we want to define the order symbols also for $\pm\infty$, we have to define neighborhoods for them. A neighborhood of ∞ is an open set $U \subset \mathbb{R}$, such that U contains an interval (a, ∞) with some $a \in \mathbb{R}$. Analogously we define neighborhoods of $-\infty$.

In the sequel $I \subset \mathbb{R}$ denotes a non empty open interval and $\overline{I}$ its closure with respect to the euclidean topology.

Definition 3.6.1 (Order Symbols)**.** *Let $\varepsilon_0 \in \overline{I}$ and $f, \phi \colon I \to \mathbb{C}$.*

(i) *We write $f = \mathcal{O}(\phi)$ as $\varepsilon \to \varepsilon_0$, if and only if*

$$\exists\, c > 0 \;\; \exists\, U \subset \mathbb{R} \text{ neighborhood of } \varepsilon_0 \;\; \forall\, \varepsilon \in I \cap U \colon \qquad |f(\varepsilon)| \leq c\, |\phi(\varepsilon)| \; .$$

(ii) *We write $f = o(\phi)$ as $\varepsilon \to \varepsilon_0$, if and only if*

$$\forall\, c > 0 \;\; \exists\, U \subset \mathbb{R} \text{ neighborhood of } \varepsilon_0 \;\; \forall\, \varepsilon \in I \cap U \colon \qquad |f(\varepsilon)| \leq c\, |\phi(\varepsilon)| \; .$$

Since ε_0 can be a boundary point of I the above definition includes the one–sided convergence of ε to ε_0.

The following Lemma 3.6.2 gives a sufficient criteria for f being of order $\mathcal{O}(\phi)$ or $o(\phi)$, which can be more useful then Definition 3.6.1.

Lemma 3.6.2. *Let $f, \phi \colon I \to \mathbb{C}$ be functions, such that*

$$\lambda \;\; := \;\; \lim_{\varepsilon \to \varepsilon_0} \frac{|f(\varepsilon)|}{|\phi(\varepsilon)|} \;\; \in \;\; \mathbb{R} \cup \{\infty\}$$

exists. It holds

(i) *If $\lambda < \infty$, then $f = \mathcal{O}(\phi)$.*

(ii) *$f = o(\phi)$, if and only if $\lambda = 0$.*

Proof. Let $\lambda < \infty$. Hence it holds:

$$\forall\, c > 0 \;\; \exists\, U \subset \mathbb{R} \text{ neighborhood of } \varepsilon_0 \;\; \forall \varepsilon \in I \cap U \colon \qquad \left| \tfrac{|f(\varepsilon)|}{|\phi(\varepsilon)|} - \lambda \right| \leq c \, .$$

From the inequality we deduce, with the lower triangle inequality, that

$$|f(\varepsilon)| \leq (\lambda + c)\, |\phi(\varepsilon)| \; .$$

Hence the existence of the finite limit λ is a sufficient condition for f to be of order $\mathcal{O}(\phi)$ as $\varepsilon \to \varepsilon_0$. Is $\lambda = 0$ we additionally deduce from the above calculation that f is of order $o(\phi)$ as $\varepsilon \to \varepsilon_0$. □

Remark 3.6.3. *In Chapter § 5 and § 6 we derive error estimates, e. g. of a quadrature rules for highly oscillatory integrals in § 5.2 or the local error of one step integrators in § 6.6, that depend on two small parameters ε and h. Hence we shall give a precise definition of the order symbol $\mathcal{O}$ in the presence of two variables. Here we restrict ourself to the special situations that appear in this thesis.*

Let B be a vector space with norm $\|\cdot\|$ and let $\alpha, \beta \in \mathbb{R}$. Furthermore let $\varepsilon_*, h_* > 0$ and let the function $f\colon (0, \varepsilon_*) \times (0, h_*) \to B$. Analogue to Definition 3.6.1 we write

$$f(\varepsilon, h) \;=\; \mathcal{O}(\varepsilon^\alpha h^\beta) \quad \textit{as} \quad (\varepsilon, h) \to 0\,,$$

if and only if there exists a constant $c > 0$, such that

$$\forall\ (\varepsilon, h) \in (0, \varepsilon_*) \times (0, h_*)\colon \qquad \|f(\varepsilon, h)\| \;\leq\; c\,\varepsilon^\alpha h^\beta\,.$$

Since ε and h are small positive numbers, we skip the adjunct $(\varepsilon, h) \to 0$ and (often) only write $f(\varepsilon, h) = \mathcal{O}(\varepsilon^\alpha h^\beta)$.

Our next goal is to characterize the behavior of a function as $\varepsilon \to \varepsilon_0$. The suitable term for our applications is the so called *asymptotic expansion*, which is defined in Definition 3.6.6. In order to give a precise definition of it, we shall first define an *asymptotic approximation* (Definition 3.6.4) and an *asymptotic sequence* (Definition 3.6.5).

Definition 3.6.4. *Let $f, \phi\colon I \to \mathbb{C}$. The function ϕ is an* asymptotic approximation *to f as $\varepsilon \to \varepsilon_0$, if and only if $f - \phi = o(\phi)$ as $\varepsilon \to \varepsilon_0$. In this case we write $f \sim \phi$.*

The following example from [32] illustrates the meaning of asymptotic approximations. Let $\varepsilon_0 = 0$ and $f(\varepsilon) = \sin(\varepsilon)$. From the power series of $\sin(\varepsilon)$ at $\varepsilon = 0$ we get

$$f(\varepsilon) \;=\; \varepsilon - \tfrac{1}{6}\varepsilon^3 + \mathcal{O}(\varepsilon^5)\,.$$

It is easy to check that $f \sim \phi_j$ holds for

$$\phi_1(\varepsilon) = \varepsilon\,, \quad \phi_2(\varepsilon) = \varepsilon + 1000\varepsilon^2\,, \quad \phi_3(\varepsilon) = \varepsilon - \tfrac{1}{6}\varepsilon^3\,.$$

Hence we do not have a unique asymptotic approximation, since all of the above given functions serve as an approximation of f as $\varepsilon \to 0$. Obviously, ϕ_3 is a better approximation of f than ϕ_2 for $|\varepsilon|$ small. To take also the comparative accuracy into account we proceed with

Definition 3.6.5. *A sequence of functions $\{\phi_j\colon I \to \mathbb{C}\}_{j\in\mathbb{N}}$ is called an* asymptotic sequence *as $\varepsilon \to \varepsilon_0$, if and only if for all $n \in \mathbb{N}$ it holds that $\phi_n = o(\phi_m)$ as $\varepsilon \to \varepsilon_0$ for all $m < n$.*

A simple example of an asymptotic sequence is $\phi_j(\varepsilon) = (\varepsilon - \varepsilon_0)^{\gamma_j}$, where $\{\gamma_j\}_{j\in\mathbb{N}}$ is a (real) strictly monotone increasing sequence. Since $\gamma_n - \gamma_m > 0$ for $m < n$,

$$\frac{\phi_n(\varepsilon)}{\phi_m(\varepsilon)} \;=\; (\varepsilon - \varepsilon_0)^{(\gamma_n - \gamma_m)} \overset{\varepsilon\to\varepsilon_0}{\longrightarrow} 0$$

and hence $\phi_n = o(\phi_m)$.

Definition 3.6.6. *We say that $f\colon I \to \mathbb{C}$ has an* asymptotic expansion *to n terms at ε_0, if and only if there exists an asymptotic sequence $\{\phi_j\}_{j\in\mathbb{N}}$ as $\varepsilon \to \varepsilon_0$ and (constant) complex coefficients $\{a_k\}_{k\in\mathbb{N}}$, such that for all $m \leq n$*

$$f \;=\; \sum_{j=1}^{m} a_j\phi_j \;+\; o(\phi_m) \qquad \text{as } \varepsilon \to \varepsilon_0\,. \tag{3.79}$$

In this case we write $f \sim \sum_{j=1}^{n} a_j\phi_j$.

We deduce from (3.79) and Lemma 3.6.2 that for all $m \leq n$

$$\frac{f(\varepsilon) - \sum_{j=1}^{m-1} a_j\phi_j(\varepsilon)}{\phi_m(\varepsilon)} - a_m \quad \stackrel{\varepsilon\to\varepsilon_0}{\longrightarrow} \quad 0\,.$$

Hence a_m and consequently the asymptotic expansion is unique for a prescribed asymptotic sequence.

One way to derive an asymptotic expansion for real–valued functions is to use Taylor approximation. Let $\varepsilon_0 \in I$ and $f\colon I \to \mathbb{R}$ be in $C^{n+1}(I)$. Due to Taylor's theorem

$$f(\varepsilon) \;=\; \sum_{j=0}^{n} \tfrac{f^{(j)}(\varepsilon_0)}{j!}(\varepsilon-\varepsilon_0)^j \;+\; R_{n+1}(\varepsilon)\,,$$

with the remainder $R_{n+1}(\varepsilon)$ being of order $o((\varepsilon-\varepsilon_0)^n)$. Since we have proven that $\phi_j(\varepsilon) = (\varepsilon-\varepsilon_0)^j$ is an asymptotic sequence, the Taylor polynomial is an asymptotic expansion of f as $\varepsilon \to \varepsilon_0$. Hence, for our example we immediately see that

$$f(\varepsilon) \;=\; \sin(\varepsilon) \sim \sum_{j=0}^{n} a_j\phi_j(\varepsilon)$$

for all $n \in \mathbb{N}$ and $a_j = \frac{(-1)^j}{(2j+1)!}$.

Nevertheless a function can have multiple asymptotic expansions. It is only unique for a prescribed asymptotic sequence. For example take the Taylor polynomial and add an exponentially small function to ϕ_j, e. g.

$$\tilde{\phi}_j(\varepsilon) \;=\; (\varepsilon-\varepsilon_0)^n \;+\; c_j e^{-\frac{1}{(\varepsilon-\varepsilon_0)^2}}\,.$$

We know that for all $j \in \mathbb{N}$ $\frac{e^{-\frac{1}{(\varepsilon-\varepsilon_0)^2}}}{(\varepsilon-\varepsilon_0)^j} \stackrel{\varepsilon\to\varepsilon_0}{\longrightarrow} 0$ and hence the set $\{\tilde{\phi}_j\}$ is an asymptotic sequence and

$$f \;\sim\; \sum_{j=0}^{n} \tfrac{f^{(j)}(\varepsilon_0)}{j!}\,\tilde{\phi}_j\,.$$

Since $c_j \in \mathbb{R}$ can be arbitrary numbers the asymptotic series

$$\sum_{j=0}^{\infty} \tfrac{f^{(j)}(\varepsilon_0)}{j!}\,\tilde{\phi}_j$$

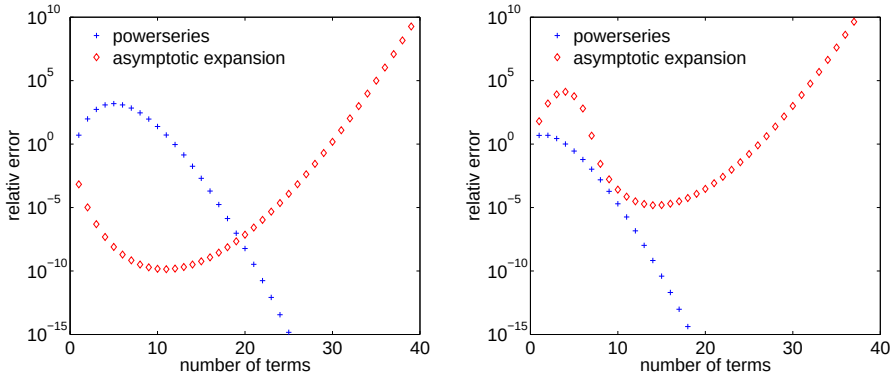

Figure 3.2: The left plot shows the relative error of the asymptotic expansion (3.81) (red diamond) compared to the power series (3.80) (blue cross) of f for $\varepsilon = 0.1$ and $\nu = 0$. The values are given as functions of the number of terms used. In the the second figure the order of J_ν is increased to $\nu = 14$.

generally does not converge for $\varepsilon > \varepsilon_0$, even if the function f can be represented by its Taylor series. The reason why an asymptotic expansion does not necessarily converge as $n \to \infty$ lies in the fact that an asymptotic expansion only makes a statement about the behavior as $\varepsilon \to \varepsilon_0$.

Thus, for a problem whose solution has an asymptotic expansion the series itself is (in general) not a reliable way to find an appropriate approximation for a fixed $\varepsilon > \varepsilon_0$. The derived expansion might be divergent and an increasing number of terms could lead to increasing errors. Nevertheless it can be a powerful tool to simplify a given problem, as done in § 3.3.

The following illustrative example about accuracy versus convergence of an asymptotic expansion is an extended example from [32]. Let

$$J_\nu(z) \quad = \quad \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(\nu+k)!}\left(\frac{z}{2}\right)^{\nu+2k} \tag{3.80}$$

be the *Bessel function* of first kind ν–th order. In [51] it is shown (for integer ν) that $f(\varepsilon) := J_\nu(\frac{1}{\varepsilon})$ has an asymptotic expansion as $\varepsilon \to 0$ of the form

$$f \quad \sim \quad \sqrt{\tfrac{2\varepsilon}{\pi}}\left[\alpha(\varepsilon)\cos(\tfrac{1}{\varepsilon} - \tfrac{\pi}{4} - \tfrac{\pi}{2}\nu) - \beta(\varepsilon)\sin(\tfrac{1}{\varepsilon} - \tfrac{\pi}{4} - \tfrac{\pi}{2}\nu)\right], \tag{3.81}$$

with

$$\begin{aligned}
\alpha \quad &= \quad \sum_{k=0}^{n} \frac{(-1)^k}{(2k)!}\left(\frac{\prod_{l=1}^{2k}(4\nu^2-(2l-1)^2)}{8^{2k}}\right)\varepsilon^{2k} \; + \; \mathcal{O}(\varepsilon^{2n+2}), \\
\beta \quad &= \quad \sum_{k=0}^{n} \left(\frac{(-1)^{k+1}}{(2k+1)!}\,\frac{\prod_{l=1}^{2k+1}(4\nu^2-(2l-1)^2)}{8^{2k+1}}\right)\varepsilon^{2k+1} \; + \; \mathcal{O}(\varepsilon^{2n+3}).
\end{aligned}$$

It is not hard to check, at least for $\nu = 0$, that the above given expansions for α and β are divergent for all nonzero ε. Nevertheless, as we can see in the right plot of Figure 3.6 for $f(\varepsilon) = J_0(\frac{1}{\varepsilon})$, an asymptotic expansion can yield a sufficient good approximation for a small number of terms taken into account,

but it does not have to. The left plot shows the relative error of the asymptotic expansion (3.81) (red diamond) compared to the power series (3.80) (blue cross) of $f(\varepsilon)$ for $\varepsilon = 0.1$ and $\nu = 0$. The values are given as functions of the number of terms used. In the plot on the right–hand side the order of J_ν is increased to $\nu = 14$. As we expect from theory, the power series always yields very accurate results for a large number of terms taken into account. However the asymptotic expansions are divergent and hence the approximation errors starts to grow monotonously for increasing number of terms.

Thus for the numerical treatment of a problem, e. g. solving an ODE, the asymptotic structure of the desired solution can be a powerful tool. But it should always be supported by additional techniques that guarantee that the approximation error can be decreased below a certain bound. An example for this procedure is discussed in § 5.2, where the approximation of highly oscillatory integrals of the form $\int_a^b f(x) e^{\frac{i}{\varepsilon}\phi(x)} dx$ is discussed.

Chapter 4

Computing the WKB–type transformation of § 3.3

The ODE integrators (one–step methods) from § 6.4 are designed to (efficiently) approximate the solution y or z of the IVP (6.2) or (6.16) respectively. For the variable z the IVP reads

$$z' \;=\; \varepsilon^n \; E_\varepsilon^* \mathcal{S}_n E_\varepsilon \; z \,, \qquad z(x_0) = z_0 \,. \tag{4.1}$$

In general, the problems we want to solve do not have this nice form. Usually on starts (possibly after a further preprocessing, e. g. as described in § 2.2) with the quantity u which solves

$$u' \;=\; \frac{i}{\varepsilon} L u \,+\, B u \,, \qquad u(x_0) \;=\; u_0 \,. \tag{4.2}$$

Thus, if we want to derive the solution of the IVP (4.2) (using the one–step methods from § 6.4), we firstly have to compute the WKB–type transformation (3.25)–(3.24) from § 3.3 (see also Proposition 3.3.1). In this chapter we discuss a discretization approach for it. It is incorporated in the fully discretized schemes used in § 7.2.

In § 3.3 the variable n denotes the polynomial degree of the transformation matrix $T_\varepsilon(x)$ with respect to ε. It is also the exponent of ε in the transformed IVP (4.1). Since in this chapter n is reserved to mark quantities at the (numerical) grid points and subintervals, we use ϑ_1 to denote the order of T_ε instead. This is a consistent notation with respect to § 6.6.1

As in [54] we shall only use values at the grid points to compute the variables. Only for some quantities close to the boundary of the integration interval I (e. g. $\Phi(x_1)$) we compute additional values, in order to guarantee a certain accuracy with respect to h.

In § 4.1 we shall derive approximation strategies on equidistant grids for $\vartheta_1 = 1, 2$. The diagram shown in Figure 4.1 sketches the variables we have to compute (for $\vartheta_1 = 2$) and there interdependence. A variable at the beginning of an arrow appears in the formula used to compute the variable the arrow points to. For example (only) T_0 shows up in the formula for T_1 (beside the given matrix valued functions L, B, which are not included in the diagram). From the diagram we get the order in which the variables have to be approximated. We

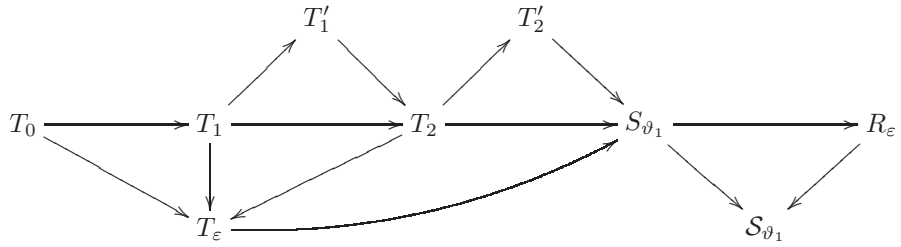

Figure 4.1: The diagram shows the interdependence of the variables, which show up in the transformation (3.26), (3.24) from § 3.3. A variable at the beginning of an arrow appears in the formula used to compute the variable the arrow points to. For example we need T_0 to compute T_1.

start with the discretization of T_0 (p.69f). Afterwards we discuss the strategy for T_1, S_1 (and T_2, S_2) (p.70ff), followed by a section which deals with the remaining variables $R_\varepsilon, \mathcal{S}_{\vartheta_1}$ (p.73f). All the discretizations are designed for equidistant grids. How they can be modified for non equidistant discretizations is discussed in § 4.2. In § 4.3 we discuss the error which originates from the WKB–type transformation or rather its numerical approximations. Furthermore we identify the crucial part of the transformation, which shall be the matrix valued phase function $\Phi = \int L\, dx$. In § 4.4 we sketch an idea to construct a step size control algorithm for the computation of the WKB–type transformation, which does not use a local error estimator. It is based on ideas from [27, VIII.2]. We also construct an algorithm based on the mentioned approach for the WKB–type transformation from § 3.3.

4.1 Equidistant grid

In the sequel let $h > 0$ and $x_0 \in \mathbb{R}$. For $j \in \mathbb{Z}$ we define $x_j = x_0 + jh$. Independent of T_0, T_1 and the related quantities (see Figure 4.1) we additionally have to compute the (matrix valued) phase function Φ. We start our discussion with its discretization.

Computation of Φ

We start with the approximation of the matrix valued function

$$\Phi(x) \quad = \quad \int_{x_0}^{x} L(\xi)\, d\xi\,.$$

Therefor we use the well known Simpson rule (cf. [28, 29, 68]). In the sequel we denote the numerical approximation of a quantity at the grid point x_n by the quantities name with subscript n, i. e. Φ_n is our approximation of $\Phi(x_n)$.

(i) Set $\Phi_0 = 0$.

(ii) Compute $L_0 = L(x_0)$, $L_{\frac{1}{2}} = L\big(\frac{x_0+x_1}{2}\big)$, $L_1 = L(x_1)$ and set

$$\Phi_1 \quad = \quad \frac{h}{6}(L_0 + 4L_{\frac{1}{2}} + L_1)\,.$$

(iii) `for` $n = 2 : N$
compute $L_n = L(x_n)$ and

$$\Phi_n \;=\; \Phi_{n-2} \,+\, \frac{h}{3}\big(L_{n-2} + 4L_{n-1} + L_n\big)\,.$$

`end`

This yields a local quadrature error of $\mathcal{O}(h^4)$. The Simpson rule can be obtained by integrating the differential equation $\Phi' = L$ with the classical Runge–Kutta (RK)[1] method (cf. [63, 68]). Its update routine from $y_n \mapsto y_{n+1}$ for the (non-linear) ODE $y'(x) = f(x, y(x))$ reads (with $h_n = x_{n+1} - x_n$):

$$\begin{aligned}
k_1 &:= f(x_n,\, y_n)\,,\\
k_2 &:= f(x_n + \tfrac{1}{2}h_n,\, y_n + \tfrac{1}{2}h_n k_1)\,,\\
k_3 &:= f(x_n + \tfrac{1}{2}h_n,\, y_n + \tfrac{1}{2}h_n k_2)\,,\\
k_4 &:= f(x_n + \;\; h_n,\, y_n + \;\; h_n k_3)\,,\\
\\
y_{n+1} &= y_n \,+\, \tfrac{h_n}{6}(k_1 + 2k_2 + 2k_3 + k_4)\,.
\end{aligned}$$

Hence we shall use the RK method also for the upcoming IVP we have to solve.

Computation of T_0

Since we implicitly use the classical RK method for Φ, there is no reason[2] why we should not use it to solve the IVP (3.37) for T_0, i. e.

$$T_0' \;=\; \mathrm{diag}_\nu(B)\, T_0\,, \qquad T_0(x_0) \;=\; \mathrm{Id}\,.$$

We only want to use values at the given grid. Since the RK method needs an evaluation of the flow function (right–hand side of the ODE) at an intermediate point, we cannot directly use the integration method. Instead we apply the RK method with the step size $2h$ on the sub–grids $x_0, x_2, x_4, \ldots$ and $x_1, x_3, \ldots$. Thus we solve (alternating) two (original) Runge–Kutta problems. Only for the first step (i. e. for the computation of $T_{0,1}$) we compute an additional value for $B_{\frac{1}{2}}$ and apply the formula from the textbooks [63, 68]. For the remaining computations we have to replace h by $2h$.

(i) Set $T_{0,0} = \mathrm{Id}$.

(ii) Compute $B_0 = B(x_0)$, $B_{\frac{1}{2}} = B\big(\frac{x_0+x_1}{2}\big)$, $B_1 = B(x_1)$ and set

$$\begin{aligned}
K_1 &= \mathrm{diag}_\nu(B_0)\, T_{0,0}\,,\\
K_2 &= \mathrm{diag}_\nu(B_{\frac{1}{2}})(T_{0,0} + \tfrac{h}{2}K_1)\,,\\
K_3 &= \mathrm{diag}_\nu(B_{\frac{1}{2}})(T_{0,0} + \tfrac{h}{2}K_2)\,,\\
K_4 &= \mathrm{diag}_\nu(B_1)(T_{0,0} + hK_3)\,,\\
T_{0,1} &= T_{0,0} \,+\, \frac{h}{6}(K_1 + 2K_2 + 2K_3 + K_4)\,.
\end{aligned}$$

[1] The abbreviation RK always denotes the classical Runge–Kutta scheme and do not mean the whole class of integrators.

[2] If one is interested in the construction of a symmetric solver for u, one has to be more careful and may use a symmetric solver for T_0 too. Since this is not our aim an explicit method is enough.

(iii) `for` $n = 2 : N$
compute $B_n = B(x_n)$ and

$$\begin{aligned} K_1 &= \operatorname{diag}_\nu(B_{n-2})T_{0,n-2}\,, \\ K_2 &= \operatorname{diag}_\nu(B_{n-1})(T_{0,n-2} + hK_1)\,, \\ K_3 &= \operatorname{diag}_\nu(B_{n-1})(T_{0,n-2} + hK_2)\,, \\ K_4 &= \operatorname{diag}_\nu(B_n)(T_{0,n-2} + 2hK_3)\,, \\ T_{0,n} &= T_{0,n-2} + \frac{h}{3}(K_1 + 2K_2 + 2K_3 + K_4)\,. \end{aligned}$$

`end`

As for Φ we get a local error of $\mathcal{O}(h^4)$.

Computation of T_1 and S_1 and optional T_2, S_2

Next we compute the matrix valued functions T_1, S_1, and T_2, S_2 respectively.

(i) By (3.38) from Remark 3.3.2 it holds

$$T_1^{\mathrm{off}_\nu} = i\,D_L^- \odot (BT_0 - T_0')\,.$$

Since $T_0' = \operatorname{diag}_\nu(B)T_0$, we do not have to approximate T_0'. We replace it by using the ODE. This yields

$$T_{1,n}^{\mathrm{off}_\nu} = i\,D_{L_n}^- \odot ((B_n - \operatorname{diag}_\nu(B_n))T_{0,n}) = i\,D_{L_n}^- \odot (B_nT_{0,n})\,.$$

For the last equality we used $\operatorname{diag}_\nu(D_L^-) = 0$ and $\operatorname{diag}_\nu(T_0) = T_0$. Since B_n is already computed to derive $T_{0,n}$, we can incorporate the computation of $T_{1,n}^{\mathrm{off}_\nu}$ in the loop for T_0. Obviously, the approximation error for $T_{1,n}^{\mathrm{off}_\nu}$ is of the same order as for T_0.

(ii) If $\vartheta_1 = 1$, i. e. we do not have to compute T_2, we set $T_1 = T_1^{\mathrm{off}_\nu}$ and can skip (iii). From Remark 3.3.2 we deduce

$$S_1 = T_\varepsilon^{-1}\big(BT_1 - T_1'\big)\,.$$

Now we could directly use the finite differences from § 8.1 to approximate $(T_1^{\mathrm{off}_\nu})'(x_n)$. In this case the accuracy of the finite difference approximation is limited (in the worst case) by the data error (of T_0) divided by h. Thus, if we want to guarantee the same spatial convergence behavior for S_1 as for T_0 and T_1, we have to proceed in a different way. However, if T_0 is exactly given, e. g. for the problem in § 2.2.1, we can directly use the finite differences, since in this case T_1 is more or less exact.

To compute the (exact) derivative of $T_1^{\mathrm{off}_\nu}$ we differentiate (3.38) with respect to x. This yields[3]

$$\begin{aligned} (T_1^{\mathrm{off}_\nu})' &= \big(iD_L^- \odot (BT_0)\big)' \\ &= i(D_L^-)' \odot (BT_0) + iD_L^- \odot (B'T_0 + BT_0') \\ &= iD_L^- \odot \big(iD_{L'} \odot T_1^{\mathrm{off}_\nu} + (B' + B\operatorname{diag}_\nu(B))T_0\big)\,. \end{aligned}$$

[3]By Lemma 3.1.9 $(D_L^-)_{ij} = (L_{ii} - L_{jj})^{-1}$. Hence the first derivative with respect to x is given by $(D_L^-)'_{ij} = -(L_{ii} - L_{jj})_{ij}^{-2}(L'_{ii} - L'_{jj})$, which yields $(D_L^-)' = -(D_L^-)^{\odot 2} \odot D_{L'}$.

Here we use ODE (3.37) to replace the derivative of T_0 in the second line. Since $\operatorname{diag}_\nu D_{L'} = 0$ we can also replace $T_1^{\operatorname{off}_\nu}$ by T_1 in the last line.

Let us denote numerically computed derivatives with respect to the spatial variable x by †, i. e. $f^\dagger(x)$ is a numerical approximation of $f'(x)$. If the derivative is approximated at a grid point x_n, we simply write $f_n^\dagger$ instead of $f^\dagger(x_n)$. With this notation we get from the previous calculation

$$\big(T_{1,n}^{\operatorname{off}_\nu}\big)^\dagger \;=\; iD_{L_n}^- \odot \big((B_n^\dagger + B_n \operatorname{diag}_\nu(B_n))T_{0,n} \,+\, iD_{L_n^\dagger} \odot T_{1,n}\big)\,.$$

Thus it remains to derive suitable approximations for L' and B'. Since it holds

$$T_{0,n} = T_0(x_n) + \mathcal{O}(h^4)\,, \qquad T_{1,n} = T_1(x_n) + \mathcal{O}(h^4)\,,$$

we shall use the finite differences from Definition 8.1.5 of § 8.1, which yield an approximation error of $\mathcal{O}(h^4)$.

(iii) Otherwise, if $\vartheta_1 = 2$ and hence we have to compute T_2, we firstly have to determine the ν–diagonal part of T_1. By (3.39) from Remark 3.3.2 it holds

$$T_1^{\operatorname{dia}_\nu}(x) \;=\; T_0(x) \int_{x_0}^{x} T_0(\xi)^{-1} \operatorname{diag}_\nu(BT_1^{\operatorname{off}_\nu})(\xi)\, d\xi\,.$$

As before the Simpson rule is our method of choice to approximate the integral.

(a) In order to apply the Simpson rule on the interval $[x_0, x_1]$ we have to evaluate the integrand at $x_{\frac12} = x_0 + \frac{h}{2}$. Since $B_{\frac12}$ is already computed to approximate $T_{0,1}$, it remains to derive a suitable approximation for $T_1^{\operatorname{off}_\nu}(x_{\frac12})$ and $T_0(x_{\frac12})$ respectively. Therefor we consider two approaches.

(1) We make one RK step to derive $T_{0,\frac12}$. Since B_0 and $B_{\frac12}$ are already computed, it remains to determine $B_{\frac14} = B\big(x_0 + \frac{h}{4}\big)$. Then we compute

$$\begin{aligned}
K_1 &= \operatorname{diag}_\nu(B_0)\, T_{0,0}\,,\\
K_2 &= \operatorname{diag}_\nu(B_{\frac14})(T_{0,0} + \tfrac{h}{4}K_1)\,,\\
K_3 &= \operatorname{diag}_\nu(B_{\frac14})(T_{0,0} + \tfrac{h}{4}K_2)\,,\\
K_4 &= \operatorname{diag}_\nu(B_{\frac12})(T_{0,0} + \tfrac{h}{2}K_3)\,,\\
T_{0,\frac12} &= T_{0,0} \,+\, \frac{h}{12}(K_1 + 2K_2 + 2K_3 + K_4)\,.
\end{aligned}$$

(2) Another idea is to use the interpolation approach from § 8.3. This is a bit less accurate, but it does not need an additional function evaluation. Since we use an equidistant grid we deduce from Lemma 8.3.1 ($\theta_l = \theta_r = \frac12$)

$$\begin{aligned}
T_{0,\frac12} \;:=\; & \tfrac14\big[2\,\mathrm{Id} + \tfrac12 h_n \operatorname{diag}_\nu(B_0)\big]\, T_{0,0}\\
& + \tfrac14\big[2\,\mathrm{Id} - \tfrac12 h_n \operatorname{diag}_\nu(B_1)\big]\, T_{0,1}\,.
\end{aligned}$$

Here the approximation error of $T_{0,\frac12}$ with respect to $T_0(x_{\frac12})$ is of order $\mathcal{O}(h^4)$.

(b) Now we set $T_{1,0}^{\mathrm{dia}_\nu} = I_{T_{1,0}} = 0$ and

$$T_{1,\frac{1}{2}}^{\mathrm{off}_\nu} \;=\; i\, D^-_{L_{\frac{1}{2}}} \odot (B_{\frac{1}{2}} T_{0,\frac{1}{2}})\,.$$

Next we compute the first Simpson approximation

$$I_{T_{1,1}} \;=\; \frac{h}{6}\,\mathrm{diag}_\nu\left(T_{0,0}^{-1} B_0 T_{1,0}^{\mathrm{off}_\nu} \,+\, 4\, T_{0,\frac{1}{2}}^{-1} B_{\frac{1}{2}} T_{1,\frac{1}{2}}^{\mathrm{off}_\nu} \,+\, T_{0,1}^{-1} B_1 T_{1,1}^{\mathrm{off}_\nu}\right).$$

and set $T_{1,1}^{\mathrm{dia}_\nu} = T_{0,1}\, I_{T_{1,1}}$.

(c) `for` $n = 2 : N$

$$\begin{aligned} I_{T_{1,n}} \;&=\; I_{T_1,n-2} \,+\, \frac{h}{3}\,\mathrm{diag}_\nu\Big(\; T_{0,n-2}^{-1} B_{n-2} T_{1,n-2}^{\mathrm{off}_\nu} \\ &\qquad +\, 4\, T_{0,n-1}^{-1} B_{n-1} T_{1,n-1}^{\mathrm{off}_\nu} \\ &\qquad +\; T_{0,n}^{-1} B_n T_{1,n}^{\mathrm{off}_\nu}\Big)\,, \\ T_{1,n}^{\mathrm{dia}_\nu} \;&=\; T_{0,n}\, I_{T_{1,n}}\,. \end{aligned}$$

`end`

The data we use for the Simpson rule have an error of $\mathcal{O}(h^4)$. Thus in each step we add an error term (with respect to the Simpson rule with unperturbed data) of order $\mathcal{O}(h^5)$. Summing up these defects yields an accuracy of $I_{T_{1,n}}$ of $\mathcal{O}(h^4)$. Hence the perturbation of the data is small enough, such that it does not influence the asymptotic behavior of the Simpson rule.

(d) Once the ν–diagonal part of T_1 is computed, we us Remark 3.3.2 to derive T_2. Since we do not compute T_3, we can set the ν–diagonal part of T_2 equal to zero. This yields

$$T_{2,n} \;=\; i\, D^-_{L_n} \odot \big(B_n T_{1,n} - T_1'(x_n)\big)\,.$$

Since the ν–diagonal part of $D^-_{L_n}$ is zero, we do not have to compute the derivative of the ν–diagonal elements of T_1. This yields

$$T_{2,n} \;=\; i\, D^-_{L_n} \odot \big(B_n T_{1,n} - (T_1^{\mathrm{off}_\nu})'(x_n)\big)\,.$$

We use the approximation procedure from (ii) to compute $(T_1^{\mathrm{off}_\nu})'$. This yields an approximation error for $T_{2,n}$ of $\mathcal{O}(h^4)$.

Now we have all quantities to compute S_2. By Remark 3.3.2 it holds

$$S_2 \;=\; T_\varepsilon^{-1}\big(BT_2 \,-\, T_2'\big)\,.$$

For simplicity we use the finite difference schemes from § 8.1 to determine T_2'. This yields (at most) a local error of $\mathcal{O}(\varepsilon^0 h^3)$ for S_2.

Computation of R and $\mathcal{S}_{\vartheta_1}$

The last transformation step is to remove the ν–diagonal part of S_{ϑ_1}. This is done as in Corollary 3.3.4. Therefor we have to find the solution of

$$R' \;=\; \varepsilon^{\vartheta_1}\,\mathrm{diag}_\nu(S_{\vartheta_1})\,R\,, \qquad R(x_0) \;=\; \mathrm{Id}$$

and set $z := R^{-1}y$ which yields

$$z' \;=\; \varepsilon^{\vartheta_1}\left(E_\varepsilon^* \mathcal{S}_{\vartheta_1} E_\varepsilon\right) z\,, \qquad z(x_0) = y(x_0) = T_\varepsilon^{-1}(x_0)\,u(x_0)\,,$$

with $\mathcal{S}_{\vartheta_1} := R^{-1} S_{\vartheta_1}^{\mathrm{off}_\nu} R$.

Notation. *In the sequel we simply write $S, \mathcal{S}$ instead of S_{ϑ_1} and $\mathcal{S}_{\vartheta_1}$ respectively. Thus if we write S_j, we mean $S_{\vartheta_1}(x_j)$ (or an appropriate approximation) and not the matrix valued function $S_{\vartheta_1=j}$.*

Again we apply the Runge–Kutta method to derive an approximation for R. Since the system matrix of the R–ODE is not oscillatory and of order $\mathcal{O}(\varepsilon^{\vartheta_1})$, we expect the RK method to be very accurate for the first steps. Hence, this time we do not compute an additional value $S_{\frac{1}{2}}$ for the first integration step. Instead we use the interpolation approach from § 8.3 to approximate R_1. In detail this means:

(i) Since $R(x_0) = R_0 = \mathrm{Id}$ we immediately get $\mathcal{S}_0 = S_0 - \mathrm{diag}_\nu(S_0)$.

(ii) Compute an approximation of $R(x_2)$ with the RK method, i. e.

$$\begin{aligned}
K_1 &= \varepsilon^{\vartheta_1}\,\mathrm{diag}_\nu(S_0)R_0\,,\\
K_2 &= \varepsilon^{\vartheta_1}\,\mathrm{diag}_\nu(S_1)(R_0 + hK_1)\,,\\
K_3 &= \varepsilon^{\vartheta_1}\,\mathrm{diag}_\nu(S_1)(R_0 + hK_2)\,,\\
K_4 &= \varepsilon^{\vartheta_1}\,\mathrm{diag}_\nu(S_2)(R_0 + 2hK_3)\,,\\
R_2 &= R_0 + \frac{h}{3}(K_1 + 2K_2 + 2K_3 + K_4)
\end{aligned}$$

and set $\mathcal{S}_2 = R_2^{-1}(S_2 - \mathrm{diag}_\nu(S_2))R_2$.

(iii) With Lemma 8.3.1 ($\theta_l = \theta_r = \frac{1}{2}$, $h_n = 2h$) we get

$$\begin{aligned}
R_1 \;=\;& \tfrac{1}{4}\left[2\,\mathrm{Id} + h\,\varepsilon^{\vartheta_1}\,\mathrm{diag}_\nu(S_0)\right] R_0\\
&+ \tfrac{1}{4}\left[2\,\mathrm{Id} - h\,\varepsilon^{\vartheta_1}\,\mathrm{diag}_\nu(S_2)\right] R_2\,.
\end{aligned}$$

Than we set $\mathcal{S}_1 = R_1^{-1}(S_1 - \mathrm{diag}_\nu(S_1))R_1$.

(iv) `for` $n = 3 : N$

$$\begin{aligned}
K_1 &= \varepsilon^{\vartheta_1}\,\mathrm{diag}_\nu(S_{n-2})R_{n-2}\,,\\
K_2 &= \varepsilon^{\vartheta_1}\,\mathrm{diag}_\nu(S_{n-1})(R_{n-2} + hK_1)\,,\\
K_3 &= \varepsilon^{\vartheta_1}\,\mathrm{diag}_\nu(S_{n-1})(R_{n-2} + hK_2)\,,\\
K_4 &= \varepsilon^{\vartheta_1}\,\mathrm{diag}_\nu(S_n)(R_{n-2} + 2hK_3)\,,\\
R_n &= R_{n-2} + \frac{h}{3}(K_1 + 2K_2 + 2K_3 + K_4)\,,\\
\\
\mathcal{S}_n &= R_n^{-1}(S_n - \mathrm{diag}_\nu(S_n))R_n
\end{aligned}$$

`end`

The data we use to approximate R_n are perturbed. The error of S_n is of order $\mathcal{O}(h^4)$ for $\vartheta_1 = 1$ and (at least) $\mathcal{O}(h^3)$ for $\vartheta_1 = 2$. Hence in each RK step we get, beside the unavoidable defect from the RK method, an additionally defect of order $\mathcal{O}(\varepsilon^1 h^5)$ or $\mathcal{O}(\varepsilon^2 h^4)$ respectively. Summing up these errors yields an accuracy of R_n with respect to $R(x_n)$ of $\mathcal{O}(\varepsilon h^4)$ or $\mathcal{O}(\varepsilon^2 h^3)$ respectively. Hence the approximation error of $\mathcal{S}_n$ is dominated by the accuracy of S_n.

4.2 Non–equidistant grids

If we only want to use values at a given arbitrary grid (not necessary equidistant), we cannot directly apply the Simpson or Runge–Kutta method as in § 4.1. But we also do not want to derive an entire new scheme to approximate the transformation. We rather want to make slight modifications of the ideas presented in § 4.1. To solve the (inhomogeneous) linear IVPs which show up to compute the transformation, e. g.

$$\begin{aligned} \Phi' &= L\,, & \Phi(x_0) &= 0\,,\\ T_0' &= \operatorname{diag}_\nu(B)T_0\,, & T_0(x_0) &= \mathrm{Id}\,,\\ R' &= \varepsilon^1 \operatorname{diag}_\nu(S_1)R\,, & R(x_0) &= \mathrm{Id}\,, \end{aligned}$$

we still want to use the RK method. Unfortunately the needed intermediate points are (in general) not at the grid. Hence we have to approximate them using the available data at the given nodes. What is the accuracy we should demand of this approximations?

To answer this question let us consider the IVP (on $[a, b]$)

$$y'(x) = A(x)y(x) + f(x)\,, \qquad y(x_0) = y_0\,. \tag{4.3}$$

We assume that the matrix valued function A and the vector (or matrix) valued function f are $C^{(4)}([a, b])$. This yields $y \in C^{(5)}([a, b])$. Furthermore let Y be the unique fundamental system of solutions that solve

$$Y'(x) = A(x)Y(x)\,, \qquad y(x_0) = \mathrm{Id}\,.$$

From literature (cf. [68]) we know that the convergence error of the RK method is $\mathcal{O}(h^4)$, where h is the maximum step size of the used prescribed spatial grid $a = x_{n_a} < \cdots < x_{n_b} = b$. In each step one has to evaluate A and f at the intermediate point

$$x_{n+\frac{1}{2}} := \frac{x_{n+1} - x_n}{2}\,.$$

Now assume that we only have perturbed values $A_{n+\frac{1}{2}}, f_{n+\frac{1}{2}}$ at $x_{n+\frac{1}{2}}$, such that (c independent of the grid)

$$\|A_{n+\frac{1}{2}} - A(x_{n+\frac{1}{2}})\| \le c\,h_n^4\,, \qquad \|f_{n+\frac{1}{2}} - f(x_{n+\frac{1}{2}})\| \le c\,h_n^4\,. \tag{4.4}$$

As usual we set $h_n := x_{n+1} - x_n$. Applying the RK method with the perturbed intermediate data $A_{n+\frac{1}{2}}$ and $f_{n+\frac{1}{2}}$ yields a sequence $\{\widehat{y}_n\}$. The update routine

reads (cf. [68] p.438)

$$\begin{aligned} K_1 &= A_n\widehat{y}_n + f_n\,, \\ K_2 &= A_{n+\frac{1}{2}}\,(\widehat{y}_n + \tfrac{1}{2}h_nK_1) \,+\, f_{n+\frac{1}{2}}\,, \\ K_3 &= A_{n+\frac{1}{2}}\,(\widehat{y}_n + \tfrac{1}{2}h_nK_2) \,+\, f_{n+\frac{1}{2}}\,, \\ K_4 &= A_{n+1}\,(\widehat{y}_n + hK_3) \,+\, f_{n+1}\,, \\ \widehat{y}_{n+1} &= \widehat{y}_n \,+\, \frac{h_n}{6}(K_1 + 2K_2 + 2K_3 + K_4)\,. \end{aligned}$$

This can be written in the form

$$\widehat{y}_{n+1} \;=\; \big(\mathrm{Id} + h_n\,\mathbb{K}_n\big)\widehat{y}_n \,+\, h_n\,F_n \tag{4.5}$$

with suitable $\mathbb{K}_n$ and F_n. By assumption (4.4) and since the matrix valued function A is uniformly bounded, there exists a constant $c_K \geq 0$, such that for all n it holds $\|\mathbb{K}_n\| \leq c_K$.

Now we construct an initial value problem, which is "well" approximated by $\{\widehat{y}_n\}$ and whose exact solution $\widehat{y}$ stays close to the desired solution y of the IVP (4.3). Therefor we define the functions ΔA and Δf piecewise on each subinterval $[x_n, x_{n+1}]$ by

$$\begin{aligned} \Delta A(x) &:= \frac{16(x-x_n)^2(x_{n+1}-x)^2}{h_n^4}\,(A_{n+\frac{1}{2}} - A(x_{n+\frac{1}{2}}))\,, \\ \Delta f(x) &:= \frac{16(x-x_n)^2(x_{n+1}-x)^2}{h_n^4}\,(f_{n+\frac{1}{2}} - f(x_{n+\frac{1}{2}}))\,. \end{aligned}$$

It holds

$$\begin{aligned} \Delta A(x_n) \;=\; \Delta A(x_{n+1}) \;=\; \Delta A'(x_n) \;=\; \Delta A'(x_{n+1}) \;&=\; 0\,, \\ \Delta f(x_n) \;=\; \Delta f(x_{n+1}) \;=\; \Delta f'(x_n) \;=\; \Delta f'(x_{n+1}) \;&=\; 0\,, \end{aligned}$$

and

$$\begin{aligned} \Delta A(x_{n+\frac{1}{2}}) &= A_{n+\frac{1}{2}} - A(x_{n+\frac{1}{2}})\,, \\ \Delta f(x_{n+\frac{1}{2}}) &= f_{n+\frac{1}{2}} - f(x_{n+\frac{1}{2}})\,. \end{aligned}$$

Hence the functions

$$\widehat{A}(x) \;:=\; A(x) \,+\, \Delta A(x) \quad\text{and}\quad \widehat{f}(x) \;:=\; f(x) \,+\, \Delta f(x)$$

are $C^1([a,b])$ and coincide with A, f at the grid. Furthermore they coincide with $A_{n+\frac{1}{2}}$ and $f_{n+\frac{1}{2}}$ at the intermediate points. Thus, if one uses the RK scheme to solve the IVP (on $[a,b]$)

$$\widehat{y}'(x) \;=\; \widehat{A}(x)\widehat{y}(x) \,+\, \widehat{f}(x)\,, \qquad \widehat{y}(x_0) = y_0\,, \tag{4.6}$$

one gets the sequence $\{\widehat{y}_n\}$. Now let us make one step with the RK method, starting at x_n with the exact solution $\widehat{y}(x_n)$ and denote the result by $\tilde{y}_{n+1}$. Since $\widehat{A}$ and $\widehat{f}$ are only $C^1([a,b])$ we cannot directly benefit from the convergence results of the RK method. But for smooth data the RK scheme has convergence

order $\mathcal{O}(h^4)$. Hence the local error (and locally $\widehat{A}$ and $\widehat{f}$ are $C^{(4)}([x_n, x_{n+1}])$) is of order $\mathcal{O}(h_n^5)$, which yields

$$\|\widehat{y}(x_{n+1}) - \tilde{y}_{n+1}\| \quad \leq \quad c_n\, h_n^5\,.$$

The constant depends on the forth derivative of the functions $\widehat{A}$ and $\widehat{f}$ on the interval $[x_n, x_{n+1}]$. By assumption (4.4) we get (and analog for f)

$$\begin{aligned}\sup_{x\in[x_n,x_{n+1}]} \|\widehat{A}^{(4)}(x)\| &= \sup_{x\in[x_n,x_{n+1}]} \left\| A^{(4)}(x) + \frac{16\cdot 4!}{h_n^4}\big(A_{n+\frac{1}{2}} - A(x_{n+\frac{1}{2}})\big)\right\| \\ &\leq \sup_{x\in[x_n,x_{n+1}]} \|A^{(4)}(x)\| + \tilde{c}\,.\end{aligned}$$

The constant $\tilde{c}$ is independent of A, f, n and the grid. Hence there exists a constant c_*, such that for all admissible n it holds $c_n \leq c_*$. Thus we get the same local error for the RK scheme as we would get for globally smooth functions A, f. Furthermore, the sequence $\{\widehat{y}(x_n)\}$ solves the inhomogeneous difference equation (cf. (4.5))

$$\widehat{y}(x_{n+1}) \quad = \quad (\mathrm{Id} + h_n\mathbb{K}_n)\widehat{y}(x_n) + h_n F_n + R_n\,,$$

with $\|R_n\| \leq c h_n^5$. Hence $\Delta_n := \widehat{y}(x_n) - \widehat{y}_n$ is a solution of the inhomogeneous difference equation

$$\Delta_{n+1} \;=\; (\mathrm{Id} + h_n K_n)\Delta_n + R_n\,, \qquad \Delta_0 \;=\; 0\,,$$

which yields

$$\|\Delta_{n+1}\| \;\leq\; \|1 + h_n K_n\|\|\Delta_n\| + \|R_n\| \;\leq\; (1 + h_n c_K)\|\Delta_n\| + c h_n^5\,.$$

By induction it follows:

$$\|\Delta_n\| \;\leq\; \prod_{j=0}^{n-1}(1 + h_j c_K)\,\|\Delta_0\| + \sum_{j=0}^{n-1}\prod_{l=j+1}^{n-1}(1 + h_l c_K)\, c h_j^5\,.$$

Since $h_n \leq h$ and $\prod_{l=j+1}^{n-1}(1 + h_l c_K) \leq e^{c_K(x_n - x_j)}$ it holds

$$\|\Delta_n\| \;\leq\; c_K h^4 e^{c_K(b-a)} \sum_{j=0}^{n-1} h_j \;\leq\; c_K h^4 e^{c_K(b-a)}(b-a)\,.$$

Hence $\|\widehat{y}(x_n) - \widehat{y}_n\| \leq c\, h^4$, with a grid and n independent constant c. Furthermore $u := \widehat{y} - y$ solves the inhomogeneous IVP

$$u'(x) \;=\; A(x)u(x) + \Delta A(x)\widehat{y}(x) + \Delta f(x)\,, \qquad u(x_0) = 0\,.$$

Using Variation of constants

$$(\widehat{y} - y)(x) \;=\; u(x) \;=\; Y(x)\int_{x_0}^{x} Y(t)^{-1}\big(\Delta A\widehat{y} + \Delta f\big)(t)\, dt\,.$$

This yields

$$\begin{aligned}\|y - \widehat{y}\|_\infty &\leq \|Y\|_\infty\|Y^{-1}\|_\infty|a-b|(\|\Delta A\|_\infty\|\widehat{y}\|_\infty + \|\Delta f\|_\infty) \\ &\leq c\, h^4\,\|Y\|_\infty\|Y^{-1}\|_\infty|a-b|(1 + \|\widehat{y}\|)\,.\end{aligned}$$

Thus the difference of the exact solution y and the solution of the modified problem is of order $\mathcal{O}(h^4)$.

Now let us consider a sequence of grids $\{X_m\}$, which admits (4.4). Let h_m be the maximum step size of the grid X_m. For every single grid we construct functions $\widehat{A}_m$, $\widehat{f}_m$, which yields a solution $\widehat{y}_m$ of (4.6). Further we get a sequence $\{\widehat{y}_{m,n}\}_n$ of a corresponding solution of the RK method. If $h_m \to 0$ the functions $\widehat{A}_m, \widehat{f}_m$ uniformly converge to A and f. Hence it is possible to derive an upper bound for $\|\widehat{y}_m\|_\infty$, which is independent of the currently used grid (but of course may depend on the family of grids). For this purpose one can use a Gronwall argument (cf. Lemma 8.4.3). Thus we get

$$\|y - \widehat{y}_m\|_\infty \quad \leq \quad c\, h_m^4\,,$$

with $c \geq 0$ independent of m. This yields

$$\|y(x_n) - \widehat{y}_{m,n}\| \quad \leq \quad \|y(x_n) - \widehat{y}_m(x_n)\| \,+\, \|\widehat{y}_m(x_n) - \widehat{y}_{m,n}\| \quad = \quad \mathcal{O}(h_m^4)\,.$$

Remark 4.2.1. *If we use perturbed (intermediate) data with an error of $\mathcal{O}(h^4)$ we preserve the asymptotic nature of the RK method.*

The missing intermediate value shall be approximated by the finite difference approach from § 8.1. In this case one has to solve a 4×4 linear system in each step. An alternative approach is to use polynomial interpolation instead. Here one can use Neville's algorithm (cf. [68]). But using computer algebra programs like *Maple14* (or carrying out the tedious computations by hand) one can a priori solve the linear system. Hence the approximation procedure via the finite difference ansatz requires only the evaluation of four scalar algebraic expressions, four scalar–matrix multiplications and three matrix summations. This is much less effort compared to Neville's algorithm[4].

We want to approximate a given functions f at the point $x_* = \frac{1}{2}(x_n + x_{n+2})$. If $x_n < x_* \leq x_{n+1}$ we use the abscissas $\{x_{n-1}, x_n, x_{n+1}, x_{n+2}\}$. Otherwise we take $\{x_n, x_{n+1}, x_{n+2}, x_{n+3}\}$. In both cases let $\eta_1, \ldots, \eta_4$ be the relative coordinates with respect to x_*, i. e.

$$x_{n_j} = x_* \,+\, h_n \eta_j \quad \text{with} \quad h_n = x_{n+2} - x_n\,.$$

This yields the ansatz

$$f_* \quad = \quad \sum_{j=1}^{4} v_j f(x_* + h_n \eta_j)\,.$$

By Lemma 8.1.2 and Remark 8.3 we get that

$$\|f(x_*) - f_*\| \quad \leq \quad c\, h_n^4\,,$$

if and only if v solves the linear system

$$\begin{pmatrix} 1 & \ldots & 1 \\ \eta_1^1 & \ldots & \eta_4^1 \\ \vdots & & \vdots \\ \eta_1^3 & \ldots & \eta_4^3 \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_4 \end{pmatrix} \quad = \quad \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}\,.$$

[4] One needs twelve scalar–matrix multiplications and six matrix summations

The unique solution of this linear system of equations is given by

$$v_i \;=\; -\frac{\eta_j\eta_k\eta_l}{(\eta_i-\eta_j)(\eta_i-\eta_k)(\eta_i-\eta_l)}\,,$$

with $i,j,k,l\in\{1,2,3,4\}$ pairwise distinct.

Remark 4.2.2. *The coefficient v_i can also be obtained by evaluating the corresponding Lagrange polynomial at x_*, i. e.*

$$\frac{(x_*-x_j)(x_*-x_k)(x_*-x_l)}{(x_i-x_j)(x_i-x_k)(x_i-x_l)} \;=\; \frac{(-\eta_j)(-\eta_k)(-\eta_l)}{(\eta_i-\eta_j)(\eta_i-\eta_k)(\eta_i-\eta_l)} \;=\; v_i\,.$$

Hence the finite difference ansatz is (in this case) equivalent to polynomial interpolation.

4.3 Crucial part of the transformation error

In this section we shall discuss the numerical approximation of the IVP (3.21) from the introduction of § 3. For this purpose let us mark the numerically derived quantities with $\widehat{\ }$, i. e. if f is a given analytical quantity its approximation coming from the algorithm is denoted by $\widehat{f}$.

Let $\vartheta_1\in\mathbb{N}$. In order to reformulate the IVP, such that it fits into the setting of § 6.2, we use the WKB–type transformation (3.25) from § 3.3. Let us briefly summarize the approximation procedure for E_ε, T_ε and R as discussed § 4.1.

(i) Fix a grid $a=x_{n_a}<x_{n_a+1}<\cdots<x_{n_b}=b$ and let

$$h \;:=\; \max_{n\in\{n_a,\ldots,n_b-1\}}(x_n-x_{n+1})$$

be the maximum step size.

(ii) If Φ is not analytically given, derive an approximation $\widehat{\Phi}$ with an ordinary quadrature, like Trapezoid or Simpson rule. This yields the local estimate

$$\|\Phi(x_n)-\widehat{\Phi}_n\| \;\le\; c\,h^{\gamma_\Phi}\,.$$

(iii) Solve the IVP (3.37)

$$T_0' \;=\; \mathrm{diag}_\nu(B)\,T_0\,,\qquad T_0(x_0) \;=\; \mathrm{Id}$$

on the fixed grid with an ordinary ODE solver. This yields a numerical approximation $\widehat{T}_0$ with the local error estimate

$$\|T_0(x_n)-\widehat{T}_{0,n}\| \;\le\; c\,h^{\gamma_0}\,.$$

The exponent $\gamma_0>0$ is prescribed by the integrator used to solve the IVP. For some problems, e. g. the second order equations from § 2.2, the IVP for T_0 is analytically solvable and hence there is no approximation error.

(iv) For $j = 1, \dots, \vartheta_1$ use Remark 3.3.2 to compute an approximation of T_j, i. e. one has to approximate

$$\begin{aligned} \widehat{T}_{j,n}^{\mathrm{off}_\nu} &= i\, D_L^-(x_n) \odot \big(B(x_n)\widehat{T}_{j,n} - \widehat{T}'_{j,n}\big) \\ \widehat{T}_{j,n}^{\mathrm{dia}_\nu} &= \widehat{T}_{0,n} \int_{x_0}^{x_n} \mathrm{diag}_\nu(BT_j^{\mathrm{off}\,\nu})(\xi)\, d\xi\,. \end{aligned}$$

For the ν–diagonal part use a suitable quadrature with order γ_j. By induction we see that the local error is always at most of order $\mathcal{O}(h^{\gamma_0})$. This yields

$$\|T_j(x_n) - \widehat{T}_{j,n}\| \;\le\; c\,h^{\gamma_0} + c\,h^{\gamma_j}\,.$$

If T_0 is exactly given, i. e. $\widehat{T}_0 = T_0$, it also holds $\widehat{T}_{1,n}^{\mathrm{off}_\nu} = T_{1,n}^{\mathrm{off}_\nu}$. Hence for $\vartheta_1 = 1$ there is no approximation error for T_ε, since the diagonal part $T_{1,n}^{\mathrm{dia}_\nu}$ has no influence of the asymptotic order and thus can be arbitrarily chosen.

(v) Compute $\widehat{S}_{\vartheta_1}$ with Remark 3.3.2.

(vi) Solve the IVP

$$\tilde{R}' \;=\; \varepsilon^{\vartheta_1}\, \mathrm{diag}_\nu(\widehat{S}_{\vartheta_1})\, \tilde{R}\,, \qquad \tilde{R}(x_0) \;=\; \mathrm{Id}$$

on the grid with an ordinary ODE solver.

(vii) Compute $\widehat{S} = \widehat{\mathcal{S}}_{\vartheta_1}$ from Corollary 3.3.4.

Even if the outstanding computations would be exact, we make some errors in the transformation that leads to (3.46) (see Corollary 3.3.4). Let us collect them. The quantities L, B of the original IVP for u (see § 3.2) are assumed to be given. Of course, also they might be inaccurate, but this is not yet of interest and hence we neglect this possibility here. Thus the first approximation occurs while computing the WKB–type transformation from § 3.3 (combine (3.25) and Corollary 3.3.4)

$$\tilde{u}(x) \;:=\; \widehat{T}_\varepsilon(x)\, \widehat{E}_\varepsilon(x)\, \widehat{R}(x)\, z(x)\,,$$

with

$$\widehat{E}_\varepsilon^*(x) \;=\; \exp\big(-\tfrac{i}{\varepsilon}\widehat{\Phi}(x)\big) \quad \text{and} \quad \widehat{T}_\varepsilon \;=\; \sum_{j=0}^{\vartheta_1} \varepsilon^j \widehat{T}_j\,.$$

Here $\widehat{\Phi}, \widehat{T}_\varepsilon$ are suitable interpolation functions which coincide with the derived numerical values at the grid. We assume

$$\begin{aligned} \widehat{\Phi}(x) &= \Phi(x) + \Delta\Phi(x)\,, & \widehat{T}_\varepsilon(x) &= T_\varepsilon(x) + \Delta T_\varepsilon(x)\,, \\ \widehat{R}_\varepsilon(x) &= R_\varepsilon(x) + \Delta R_\varepsilon(x)\,, & \widehat{E}_\varepsilon(x) &= E_\varepsilon(x) + \Delta E_\varepsilon(x)\,. \end{aligned}$$

Let z, u be the exact solutions of (3.46) and (3.23) respectively. It holds

$$\|\tilde{u}(x) - u(x)\| \;\le\; \|\widehat{T}_\varepsilon(x)\widehat{E}_\varepsilon(x)\widehat{R}(x) - T_\varepsilon(x)E_\varepsilon(x)R(x)\|\, \|z(x)\|\,.$$

In order to get an estimate for the first norm we rewrite the matrix by adding certain "zeros". I. e.

$$\begin{aligned}
\widehat{T}_\varepsilon\widehat{E}_\varepsilon\widehat{R} - T_\varepsilon E_\varepsilon R &= \widehat{T}_\varepsilon\widehat{E}_\varepsilon\widehat{R} - \widehat{T}_\varepsilon\widehat{E}_\varepsilon R + \widehat{T}_\varepsilon\widehat{E}_\varepsilon R - T_\varepsilon E_\varepsilon R \\
&= \big(\widehat{T}_\varepsilon\widehat{E}_\varepsilon\big)\big(\widehat{R} - R\big) + \big(\widehat{T}_\varepsilon\widehat{E}_\varepsilon - T_\varepsilon E_\varepsilon\big)R \\
&= \big(\widehat{T}_\varepsilon\widehat{E}_\varepsilon\big)\big(\widehat{R} - R\big) \\
&\quad + \big(\widehat{T}_\varepsilon\widehat{E}_\varepsilon - \widehat{T}_\varepsilon E_\varepsilon + \widehat{T}_\varepsilon E_\varepsilon - T_\varepsilon E_\varepsilon\big)R \\
&= \big(\widehat{T}_\varepsilon\widehat{E}_\varepsilon\big)\big(\widehat{R} - R\big) \\
&\quad + \widehat{T}_\varepsilon\big(\widehat{E}_\varepsilon - E_\varepsilon\big)R + \big(\widehat{T}_\varepsilon - T_\varepsilon\big)E_\varepsilon R\,.
\end{aligned}$$

This yields

$$\|\widehat{T}_\varepsilon\widehat{E}_\varepsilon\widehat{R} - T_\varepsilon E_\varepsilon R\| \;\leq\; \|\widehat{T}_\varepsilon\|\|\Delta R\| + \|\widehat{T}_\varepsilon\|\|R\|\|\Delta E_\varepsilon\| + \|\Delta T_\varepsilon\|\|R\|\,.$$

Here we used that E_ε is unitary and hence $\|E_\varepsilon\| = 1$. Since $\widehat{E}_\varepsilon, E_\varepsilon$ are diagonal, the norm of ΔE_ε is the maximum absolute value of its eigenvalues. Let $\varphi, \widehat{\varphi} \in \mathbb{R}$. A straight forward calculation shows that

$$\big|e^{\frac{i}{\varepsilon}\varphi} - e^{\frac{i}{\varepsilon}\widehat{\varphi}}\big|^2 \;=\; 2\big(1 - \cos(\tfrac{\varphi-\widehat{\varphi}}{\varepsilon})\big) \;=\; 4\sin^2\big(\tfrac{\varphi-\widehat{\varphi}}{2\varepsilon}\big)\,.$$

This yields

$$\|(\widehat{E}_\varepsilon - E_\varepsilon)(x)\| \;=\; 2\sup_{j=1,\dots,d}\left|\sin\left(\frac{\Delta\Phi_{jj}(x)}{2\varepsilon}\right)\right|.$$

From Lemma 3.3.1 and Remark 3.3.5 we deduce that T_ε and R_ε are at least C^0–bounded independently of ε. Hence there exists a constant $c \geq 0$, such that

$$\|\tilde{u}(x) - u(x)\| \;\leq\; c\Big(\sup_j\big|\sin\big(\tfrac{\Delta\Phi_{jj}(x)}{2\varepsilon}\big)\big| + \|\Delta R_\varepsilon\| + \|\Delta T_\varepsilon\|\Big)\|z(x)\|\,.$$

By Proposition 6.2.2 we know that $\|z(x)\|$ is bounded independently of ε. Thus the crucial part of the numerical transformation is the approximation of the phase function. Even if $\Delta T_\varepsilon = 0$ (e. g. for the second order ODE from § 2.2.1 with $\vartheta_1 = 1$) and $\|\Delta R_\varepsilon\| = \mathcal{O}(\varepsilon^n h^\gamma)$, the error is still of order $\mathcal{O}(\varepsilon^{-1}\|\Delta\phi\|)$.

Remark 4.3.1. *The crucial part of the transformation error originates from the approximation of the matrix valued phase function Φ. It is of order $\varepsilon^{-1}\|\Delta\Phi\|$. This has two consequences:*

(i) *To compute Φ we should use a quadrature with a very high accuracy.*

(ii) *Step size restriction. Let γ_Φ be the order of the quadrature used to approximate Φ, i. e.*

$$\|\Phi - \widehat{\Phi}\|_\infty \;\leq\; c\,h^{\gamma_\Phi}\,.$$

If we want to construct a scheme with convergence order τ, i. e.

$$\|u - \widehat{u}\|_\infty \;\leq\; c\,h^\tau\,,$$

we have to compensate the factor ε^{-1} with powers of h. This means

$$\frac{h^{\gamma_\Phi - \tau}}{\varepsilon} \quad \leq \quad c\,,$$

with a constant c independently of ε. Hence we get an ε dependent upper bound for h, i. e.

$$h_0 \quad = \quad (c\,\varepsilon)^{\frac{1}{\gamma_\Phi - \tau}}\,.$$

Furthermore we observe that $\gamma_\Phi \geq \tau + 2$ has to hold, if we want to use step sizes $h \geq \varepsilon$.

If Φ is not exactly given, we should choose the grid dependent on Φ and ε, in order to control the error of the primal quantity u.

4.4 Step size control

In this section we briefly discuss an approach to construct a step size controller for the computation of the WKB–type transformation. It is based on ideas from [27], which are adapted to our setting. With the strategy described in the sequel we hope to achieve an approximation error for the quantity z (approximated with the one–step methods (OSM) from § 6.4), which is at most as large as for an equidistant grid with the same number of abscissas. If the quantity B (cf. (4.7)) or its derivatives are very large (at some points in the interval), we expect that our step size control approach is much more accurate than an equidistant grids.

The transformation

$$z(x) \quad = \quad R^{-1}(x) E_\varepsilon^*(x) T_\varepsilon^{-1}(x)\, u(x)$$

maps the solution u of

$$u'(x) \;=\; \frac{i}{\varepsilon} L\, u \,+\, B\, u\,, \qquad u(x_0) \;=\; u_0\,, \tag{4.7}$$

to the solution z of (3.46), i. e.

$$z' \;=\; \varepsilon^n \left(E_\varepsilon^* \mathcal{S}_n E_\varepsilon\right) z\,, \qquad z(x_0) = z_0\,. \tag{4.8}$$

As discussed in § 4.3, the crucial part of the transformation is the approximation of matrix valued function Φ. Instead of Φ let us rather compute $\Phi_\varepsilon := \frac{1}{\varepsilon}\Phi$. This yields

$$E_\varepsilon(x) \;=\; \exp(i\Phi_\varepsilon(x)) \quad \text{with} \quad \Phi_\varepsilon'(x) \;=\; \frac{1}{\varepsilon} L(x)\,.$$

The OSM from § 6 used to compute z can yield poor results, if the eigenvalues of L get close to each other, i. e. $\delta \ll 1$. Another problem can be that the norm of the matrix B is getting large at a certain point (cf. § 7.3). Both events cause a growth of the norm of $\mathcal{S}_{\vartheta_1}$. Hence we should take also this into account. Our approach is an adaption of ideas from [27, §VIII.2]. Let us make a space transformation of the whole system, i. e.

$$\phi(t) \;:=\; \Phi_\varepsilon(\omega(t))\,, \qquad \zeta(t) \;:=\; z(\omega(t))\,,$$

with some differentiable strictly monotone function ω. Differentiation yields (where $\dot{}$ denotes the derivative with respect to t)

$$\begin{aligned}\dot{\phi}(t) &= \dot{\omega}(t)\,\frac{1}{\varepsilon}L(\omega(t))\,,\\ \dot{\zeta}(t) &= \dot{\omega}(t)\,\varepsilon^{\vartheta_1}(E_\varepsilon^*\mathcal{S}_{\vartheta_1}E_\varepsilon)(\omega(t))\zeta(t)\,.\end{aligned}$$

In order to control the norm of the system matrix for ζ, i. e. the norm of $\mathcal{S}_{\vartheta_1}$, we simply demand

$$\dot{\omega}(t) = \frac{1}{\|\mathcal{S}_{\vartheta_1}(\omega(t))\|}\,.$$

Since the IVP for ζ still fits to our Model Problem 2, we assume that we can use rather coarse grids to get a sufficient numerical solution. Hence the IVPs for ϕ and ω shall define the used grid. Thus one approach could be

(i) Use a standard integrator like the embedded Runge–Kutta routine `ode45` from *Matlab* to solve the nonlinear IVP

$$\dot{\phi}(t) = \dot{\omega}(t)\,\frac{1}{\varepsilon}L(\omega(t))\,,\quad \phi(x_0) = 0\,, \tag{4.9}$$

$$\dot{\omega}(t) = \frac{1}{\|\mathcal{S}_{\vartheta_1}(\omega(t))\|}\,,\quad \omega(0) = x_0\,, \tag{4.10}$$

with a prescribed accuracy. Hence we use the step size controller of the standard routine to generate our grid. But we have to save a lot of data at every grid point.

(ii) Use the stored data to solve the IVP

$$\dot{\zeta}(t) = \dot{\omega}(t)\,\varepsilon^{\vartheta_1}(E_\varepsilon^*\mathcal{S}_{\vartheta_1}E_\varepsilon)(\omega(t))\zeta(t)\,,\qquad \zeta(0) = z(x_0)\,.$$

with the OSM derived in § 6.

This is of course a non optimized method, but easy to implement, provided $\mathcal{S}_{\vartheta_1}$ is explicitly known. And if the method works, one can combine the standard ODE solver and the OSM, such that less data has to be stored in each step.

If the problem for u has conserved quantities, e. g. if it comes from a second order system, it should be possible to construct a *reversible controller* as suggested in [27]. For a first try one should think about the Crank-Nicolson like scheme. It is symmetric, if the supporting abscissas and multiplicities for the interpolation problems are symmetric with respect to the integration interval.

Since the IVP (4.10) for ω is independent of the used OSM, it is not necessary to transform the z IVP (4.8), i. e. we do not have to introduce a new variable ζ. Let $x_0 = \omega_0 < \omega_1 < \cdots < \omega_N = b$ be approximate solutions of (4.10). Hence we set $x_j = \omega_j$ and solve the IVP for z on this non equidistant grid. In the following subsection we present one approach, which solves (4.10), such that only one evaluation of $\mathcal{S}_{\vartheta_1}$ is needed per step. Once ω and $\mathcal{S}_{\vartheta_1}$ are known, one can directly apply the OSM. The grid is only determined by ω. We do not take the phase function Φ into account.

Computation of w for $\vartheta_1 = 1$ and exact T_0

In this subsection we develop a program to approximate w for $\vartheta_1 = 1$. We want to test this approach for the second order problems from § 7.3. Thus we assume that T_0 is explicitly computable. We remark that one can extend this ansatz also to the case where T_0 is not (explicitly) known.

Since the IVP (4.10) is highly nonlinear, we shall use an explicit integrator to speed up the calculations. Furthermore, this enables us to use only one function evaluation (i. e. computation of L, B) per step. Our method of choice (for the first try) is the following *Adams–Bashforth* (AB) scheme [63, 68], which is a multistep integrator.

Definition 4.4.1. *Let $f\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ and let $t_j = t_0 + j\Delta t$ with $\Delta t > 0$ and $j \in \mathbb{N}$. For given values y_0, y_1 we define the AB2 update routine for $n \geq 1$ by*

$$y_{n+1} \;=\; y_n + \frac{\Delta t}{2}(3f_n - f_{n-1}) \tag{4.11}$$

Here we use the notation $f_j = f(t_j, y_j)$. If y_0, y_1 are properly chosen, then the sequence (y_n) approximates the solution y of the IVP

$$y'(t) \;=\; f(t, y(t))\,, \qquad y(t_0) = y_0\,.$$

Remark 4.4.2. *The multistep scheme AB2* (4.11) *has a convergence order of* 2 *with respect to Δt. This shall be enough to find an appropriate solution of* (4.10)*. To determine the first value y_1 we simply use the* explicit Euler scheme*, which yields an error of order $\mathcal{O}((\Delta t)^2)$. If one wants to invest more in the accuracy of y_1, one may use the* method of Heun (4.12) *or the* modified Euler method (4.13)*. Since these methods have a convergence order of 2, the (local) error of y_1 is of order $\mathcal{O}((\Delta t)^3)$.*

The advantage of the multistep approach AB2 (4.11) compared to second order Runge–Kutta methods like the *method of Heun*

$$\begin{aligned} k_1 &= f(t_n,\, y_n)\,,\\ k_2 &= f(t_n + \Delta t,\, y_n + \Delta t k_1)\,,\\ y_{n+1} &= y_n \,+\, \Delta t\tfrac{1}{2}(k_1 + k_2)\,, \end{aligned} \tag{4.12}$$

or the *modified Euler method*

$$\begin{aligned} k_1 &= f(t_n,\, y_n)\,,\\ k_2 &= f(t_n + \tfrac{1}{2}\Delta t,\, y + \tfrac{1}{2}\Delta t f(t_n, y_n))\,,\\ y_{n+1} &= y_n \,+\, \Delta t\, k_2\,, \end{aligned} \tag{4.13}$$

(cf. [68] for both methods) is that we only have to evaluate the function f at the grid points. Hence we only need one function evaluation per time step. Especially for the IVP (4.10) we save a lot of computational effort, since the evaluation of $\mathcal{S}_{\vartheta_1}$ is expensive.

To increase the accuracy one can extend the AB2 scheme to a *predictor corrector* scheme. This is done by using the value y_{n+1} as a first guess (which should be quite good) for an implicit scheme of *Adams–Moulton* type and just make the first iteration to get a correction of y_{n+1} (cf. [63]). However, this yields an additional evaluation of f.

In the sequel we simply write $\mathcal{S}$ instead of $\mathcal{S}_{\vartheta_1}$. Let us fix an increment $\Delta t > 0$. Furthermore let $\Delta x > 0$ be our initial spatial step size. We use it for the finite difference approximation of T_1' at $x_0 = a$. Hence it should be small with respect to Δt. The crucial variables we have to approximate are $\mathcal{S}, R_\varepsilon, T_\varepsilon$. They are needed for the OSM. In contrast, ω is just an auxiliary variable, which is only used to determine the grid.

The function $\|\mathcal{S}\|$ (which determines the ODE for ω) can have very high, sharp peaks (cf. Figure 7.13, 7.14 for the example of § 7.3). If the step size gets very large, it can happen that the peaks (the crucial parts) of the function $\|\mathcal{S}\|^{-1}$ are not resolved. In this case the determined variables may have large approximation errors. Thus, in order to avoid insufficient (large) step sizes, we fix a maximum step size $h_{\max}$ for the OSM–grid $x_0 < \cdots < x_N$. For the computation of $\mathcal{S}$ we need R_ε. Since R_ε is given by an IVP (cf. (3.45) of Corollary 3.3.4), we make one RK step in each loop to determine R_ε. Thus we set $x_{2n} = \omega_n$ and $x_{2n+1} = \frac{\omega_n + \omega_{n-1}}{2}$ (the intermediate point for the RK method). Hence the distance between ω_n and ω_{n-1} has to be smaller than $2h_{\max}$. Also, we establish a lower bound $h_{\min}$ in order to avoid that the computation stalls. The (first try) program reads:

(i) Compute T_0, T_1, T_ε at $\xi_j = x_0 + j\Delta x$, $j = 0, \ldots, 4$.

(ii) use the values of T_1 at $\xi_0, \ldots, \xi_4$ to approximate T_1' at x_0.

  (a) compute the relative coordinates ($h = 1$)

$$\eta_0 := 0\,, \qquad \eta_j := \xi_j - \xi_0\,, \quad j = 1, \ldots, 4$$

  (b) solve the linear system (8.3) from Remark 8.1.3 with $r = 1$

  (c) approximate T_1' with the Finite Differences from Definition 8.1.1

(iii) compute S at $x_0 = \xi_0$ (Remark 3.3.2)

(iv) set $R_{\varepsilon,0} = \mathrm{Id}$ and compute $\mathcal{S}_0$ (Corollary 3.3.4)

(v) Set $\omega_0 := x_0 = a$ and use the explicit Euler method to compute ω_1, i. e.

$$\omega_1 \;=\; \omega_0 \,+\, \min\left(2h_{\max},\, \frac{\Delta t}{\|\mathcal{S}_0\|}\right).$$

(vi) set $x_1 = \frac{1}{2}(\omega_1 - x_0)$ and $x_2 = \omega_1$

(vii) `if` $\omega_1 - \omega_0 < \Delta x$ set $\Delta x = \frac{1}{2}(\omega_1 - \omega_0)$ and restart

(viii) `else` continue and compute T_0, T_1, T_ε at x_1, x_2

(ix) compute approximations of T_1' at x_1, x_2

  (a) choose the four nearest neighbors $y_1, \ldots, y_4$ of the abscissa x_1 from the set $\{\xi_0, \ldots, \xi_4, x_2\}$

  (b) compute the relative coordinates ($h = 1$)

$$\eta_0 := 0\,, \qquad \eta_j := y_j - \xi_0\,, \quad j = 1, \ldots, 4$$

(c) solve the linear system (8.3) from Remark 8.1.3 with $r = 1$

(d) approximate T_1' at x_1 with the Finite Difference from Definition 8.1.1

(e) exchange x_1 and x_2 and repeat the procedure to approximate T_1' at the abscissa x_2

(x) compute S at x_1, x_2

(xi) use one RK step to approximate R_ε at x_2

(xii) use interpolation to compute R_ε at x_1 (cf. Lemma 8.3.1)

(xiii) compute $\mathcal{S}_1$, $\mathcal{S}_2$ and set $n = 1$

(xiv) compute ω_2 with the AB2 scheme (4.11), i. e.

$$\begin{aligned} \text{inc} &= \max\left(2h_{\min}, \min\left(2h_{\max}, \tfrac{\Delta t}{2}\left(\tfrac{3}{\|\mathcal{S}_n\|} - \tfrac{1}{\|\mathcal{S}_{n-1}\|}\right)\right)\right), \\ \omega_{n+1} &= \omega_n + \text{inc} \,. \end{aligned}$$

(xv) `while` $\omega_n < b$

(a) set $x_{2n+1} = \omega_n + \frac{1}{2}(\omega_{n+1} - \omega_n)$ and $x_{2(n+1)} = \omega_{n+1}$

(b) compute T_0, T_1, T_ε at $x_{2n+1}, x_{2(n+1)}$

(c) use the values of T_1 at $x_{2(n-1)}, \ldots, x_{2(n+1)}$ to compute an approximation of T_1' at $x_{2n+1}, x_{2(n+1)}$

(d) compute S at $x_{2n+1}, x_{2(n+1)}$

(e) use one RK step to approximate R_ε at $x_{2(n+1)}$

(f) use interpolation to compute R_ε at x_{2n+1} (cf. Lemma 8.3.1)

(g) compute $\mathcal{S}_{2n+1}$, $\mathcal{S}_{2(n+1)}$ and set $n = n + 1$

(h) compute ω_2 with the AB2 scheme (4.11), i. e.

$$\begin{aligned} \text{inc} &= \max\left(2h_{\min}, \min\left(2h_{\max}, \tfrac{\Delta t}{2}\left(\tfrac{3}{\|\mathcal{S}_n\|} - \tfrac{1}{\|\mathcal{S}_{n-1}\|}\right)\right)\right), \\ \omega_{n+1} &= \omega_n + \text{inc} \,. \end{aligned}$$

(xvi) set $\omega_{n+1} = b$ and repeat (a)-(g) of (xvi)

It turns out that in some cases the AB2 scheme yields negative increments. Due to the max min restriction this does not make trouble, but creates unnatural artifacts in the grid. They are very good visible in Figure 7.17. To get rid of this problem we exchange the AB2 integrator by the simplest explicit one step integrator, i. e. the Euler scheme. Thus simply replace the variable inc in (xiv) and (xv)(h) by

$$\text{inc} = \max\left(2h_{\min}, \min\left(2h_{\max}, \tfrac{\Delta t}{\|\mathcal{S}_n\|}\right)\right).$$

The rest of the code remains unchanged. This reduces the accuracy for ω, but since we are not interested in this variable it does not matter.

Chapter 5

Approximation of highly oscillatory integrals

The preprocessing as discussed in § 3.3 yields an IVP for the new variable z of the form

$$z' \;=\; \varepsilon^n \, E_\varepsilon^* \mathcal{S}_n E_\varepsilon \, z \,, \qquad z(x_0) = z_0 \,. \tag{5.1}$$

Thus (for $\varepsilon \ll 1$) we have to deal with highly oscillatory entries of the system matrix. The one–step method we shall derive in chapter 6 is specially designed to numerically integrate (5.1). A key ingredient is a sophisticated quadrature for highly oscillatory integrals, which originate from the highly oscillatory entries of the system matrix. The advanced quadrature shall be discussed in this chapter.

In the sequel let $J \subset \mathbb{R}$ be a closed, bounded, non–trivial interval and let the numbers $\alpha, \beta \in \mathbb{R}$, $\alpha < \beta$, such that

$$[\alpha, \beta] \quad \subset \quad J \quad \subset \quad \mathbb{R} \,.$$

The subject of this chapter is the approximation of the integral

$$I[f] \quad := \quad \int_\alpha^\beta f(x) e^{-\frac{i}{\varepsilon}\varphi(x)} \, dx \,, \tag{5.2}$$

where $0 < \varepsilon \ll 1$, $\varphi \in C^s([\alpha, \beta])$ strictly monotone and f a suitable smooth function, such that $I[f]$ is well defined.

The first idea to approximate $I[f]$ might be to use the well understood and discussed Newton–Cotes or Gauss–Christoffel quadratures (cf. [28]). But it turns out, as illustrated by Iserles [36] for the Gauss quadrature, that this approach yields sufficiently good results only for large values of ε. On the contrary, if $\varepsilon \ll 1$ and hence the integrand is highly oscillatory, both approaches yield inefficient quadratures, since the number of nodes in the interval have to increase with decreasing ε in order to keep the error below a desired bound.

We shall use an interpolation approach to derive a quadrature rule for the highly oscillatory integral (5.2). It is closely related to the classical polynomial interpolation. This idea is very similar to a more general technique discussed by S. Olver [61] in 2007, a so called "Moment–free Filon–type method".

We start with a brief review of quadrature rules for highly oscillatory integrals in §5.1. For more methods and references we refer to the review article [35]. Afterwards, in §5.2 we derive the *modified Filon–type method* for (5.2). Furthermore we prove an upper bound for the quadrature error. The estimates we derive for the quadrature error explicitly depend on the length of the integration interval. This is not considered in [61]. Here the author focuses on the asymptotic behavior with respect to the small parameter ε. In § 5.4 we shall make some numerical experiments, which show even better error behaviors of the quadrature than predicted in § 5.2. We compare our modified Filon–type method to the *shifted asymptotic method* presented in [4], which we shall briefly discuss in § 5.3. In this section we also derive a symmetric version of the shifted asymptotic method, which we expect to yield better results than the original quadrature.

5.1 Review of some quadrature rules

We start with a simple computation which leads to a fundamental property of the integral $I[f]$ (see (5.2) for its definition). Let $f, \varphi\colon J \to \mathbb{R}$ be smooth functions and let φ be strictly monotone. Hence $|\varphi'|$ is bounded from below by a positive constant δ. We shall make one integration by parts. Ad hoc we rephrase the integrand of $I[f]$:

$$f(x)\, e^{-\frac{i}{\varepsilon}\varphi(x)} \quad = \quad i\varepsilon \frac{f(x)}{\varphi'(x)} \left(e^{-\frac{i}{\varepsilon}\varphi(x)} \right)',$$

which yields

$$I[f] \quad = \quad i\varepsilon \, \tfrac{f(x)}{\varphi'(x)} \, e^{-\frac{i}{\varepsilon}\varphi(x)} \big|_{x=\alpha}^{\beta} \, - \, i\varepsilon \int_\alpha^\beta \left(\tfrac{f(x)}{\varphi'(x)} \right)' e^{-\frac{i}{\varepsilon}\varphi(x)} \, dx \, .$$

Since all derivatives are well defined and smooth, a further integration by parts in the above sense shows that the remaining integral is of order $\mathcal{O}(\varepsilon^2)$. Thus the expression

$$Q_1^A[f] \quad := \quad i\varepsilon \, \tfrac{f(x)}{\varphi'(x)} \, e^{-\frac{i}{\varepsilon}\varphi(x)} \big|_{x=\alpha}^{\beta}$$

approximates $I[f]$ with an error of order $\mathcal{O}(\varepsilon^2)$. As long as all quantities are smooth we can continue this procedure and obtain an approximation of $I[f]$. By induction we get

Lemma 5.1.1 (Asymptotic method)**.** *Let $s \in \mathbb{N}$, $f \in C^s(J, \mathbb{C})$ and let the phase function $\varphi \in C^{s+1}(J, \mathbb{R})$ with $|\varphi'(x)| \geq \delta > 0$ for all $x \in J$. We inductively define for $j \in \{1, \ldots, s\}$*

$$f_0 := f \, , \qquad f_j := \left(\frac{f_{j-1}}{\varphi'} \right)' ,$$

and set

$$Q_s^A[f] \quad := \quad - \sum_{j=0}^{s-1} (-i\varepsilon)^{j+1} \, \frac{f_j(x)}{\varphi'(x)} \, e^{-\frac{i}{\varepsilon}\varphi(x)} \Bigg|_{x=\alpha}^{\beta} .$$

Then it holds

$$I[f] \;=\; Q_s^A[f] \;+\; (-i\varepsilon)^s\, I[f_s]\,. \tag{5.3}$$

The approximation $Q_s^A[f]$ is the *asymptotic method* presented by Iserles and Nørsett [37, 38]. Since φ is strictly monotone on J and since f_s is a continuous function, it holds:

$$I[f_s] \;=\; \int_\alpha^\beta f_s(x)\, e^{-\frac{i}{\varepsilon}\varphi(x)}\, dx \;=\; \int_{\varphi(\alpha)}^{\varphi(\beta)} \tfrac{f_s(\varphi^{-1}(\xi))}{\varphi'(\varphi^{-1}(\xi))}\, e^{-\frac{i}{\varepsilon}\xi}\, d\xi\,.$$

Due to the Riemann–Lebesgue Lemma [58] the last integral is of order $o(1)$ as $\varepsilon \to 0$. Hence we get for the quadrature error:

$$|I[f] - Q_s^A[f]| \;=\; \varepsilon^s\, |I[f_s]| \;=\; o(\varepsilon^s)\,.$$

Since an asymptotic expansion, like $Q_s^A[f]$ for $I[f]$, does not have to converge (cf. § 3.6) it is not ensured that for fixed ε and a given bound $c > 0$ there exists an index $S \in \mathbb{N}$, such that

$$\big|I[f] - Q_S^A[f]\big| \;<\; c\,.$$

In [4] the authors establish the *shifted asymptotic method*, which is (as the asymptotic method) only based on integration by parts. The presented approach additionally yields a "spatial" expansion of the integral, i. e. an expansion with respect to the length of the integration interval. We shall derive a symmetric version of the shifted asymptotic method in § 5.3 and thus refer to this section for more details.

To overcome the shortcoming of the asymptotic method we generalize the method with an ansatz based on the ideas of Filon (1928) [18].

Let $\xi_1, \ldots, \xi_\kappa \in J$ with $\xi_1 < \cdots < \xi_\kappa$, such that there are indices j_α, j_β with

$$\xi_{j_\alpha} \;=\; \alpha \quad \text{and} \quad \xi_{j_\beta} \;=\; \beta\,.$$

The complex valued function f is approximated by a Hermite interpolation polynomial[1] p at the pairwise distinct nodes $\xi_1, \ldots, \xi_\kappa$ with corresponding multiplicities $m_1, \ldots, m_\kappa \in \mathbb{N}$, yielding a quadrature of the form

$$Q_s^F[f] \;:=\; I[p]\,, \tag{5.4}$$

where s is equal to $\min\{m_{j_\alpha}, m_{j_\beta}\}$. This is the *Filon–type method* by Iserles and Nørsett [37]. The quadrature error of this approach depends on the accuracy of the Hermite interpolation. Additionally it has the asymptotic property of the asymptotic method, since p and f coincides up to the s-th derivative at the boundary points of the integral. Unfortunately, the moments $I[x^j]$ have to be known, which generally are not exactly computable.

To remove also this disadvantage we follow an idea of Sheehan Olver [60]. Instead of approximating f by a linear combination of the monomials $x^0, x^1, x^2, \ldots$ we rather use a set of functions $\psi_0, \psi_1, \psi_2, \ldots$ for which the oscillatory integrals can be computed explicitly.

[1] cf. [68]: p is the unique polynomial of degree $m = (\sum_{j=1}^{\kappa} m_j) - 1$ which satisfies the interpolation condition: $f^{(k)}(x_j) = p^{(k)}(x_j)$ for $j = 1, \ldots, \kappa$, $0 \le k \le m_j - 1$

One possibility (cf. [60]) to construct these functions is a simplified ansatz of Levin's approach [52]. Let Ψ and ψ be functions, such that

$$\Psi' - \tfrac{i}{\varepsilon}\varphi'\Psi = \psi\,. \tag{5.5}$$

Multiplication of (5.5) with the integrating factor $e^{-\frac{i}{\varepsilon}\varphi}$ yields

$$\frac{d}{dx}\big(\Psi(x)\,e^{-\frac{i}{\varepsilon}\varphi(x)}\big) = \psi(x)\,e^{-\frac{i}{\varepsilon}\varphi(x)}$$

and hence we immediately find by simply integrating that

$$I[\psi] = \Psi(x)\,e^{-\frac{i}{\varepsilon}\varphi(x)}\Big|_{x=\alpha}^{\beta}\,. \tag{5.6}$$

Thus for ψ we know how to compute $I[\psi]$ and hence we can use this relation to derive a set of functions which are exactly integrable. Let $\Psi_0,\ldots,\Psi_{m-1}$ be smooth functions and define for $j=0,\ldots,m-1$

$$\psi_j := \Psi_j' - \tfrac{i}{\varepsilon}\varphi'\Psi_j\,.$$

Since $I[\cdot]$ is a linear map, we also know how to compute the integral of

$$p(x) := \sum_{j=0}^{m-1} c_j\,\psi_j(x)\,, \tag{5.7}$$

with some $c_1,\ldots,c_{m-1}\in\mathbb{C}$. Thus we can write

$$I[f] = I[p] + I[f-p]$$

and hence p is a good candidate for an approximation of f. If we use for given support abscissas $x_1,\ldots,x_\kappa$ and corresponding multiplicities $m_1,\ldots,m_\kappa$ the generalized Hermite interpolation approach

$$p^{(k)}(x_l) = f^{(k)}(x_l)\,, \qquad k=0,\ldots,m_l-1,\quad l=1,\ldots,\kappa\,,$$

then we get the *Levin–type method*[2] of [60]:

$$Q^L[f] := I[p] = \sum_{j=0}^{m-1} c_j\,I[\psi_j]\,. \tag{5.8}$$

For $\psi_k=\varphi'\varphi^k$, $k\in\mathbb{N}\cup\{0\}$ it is possible to derive Ψ_k explicitly. This choice yields[3] the moment–free Filon–type method from [61]. We shall use this basis for our quadrature, which is discussed in § 5.2.

What properties should p have in order to yield a good quadrature? First of all it is clear that we get an error estimate which is proportional to the approximation error $f-p$. Hence we should choose p, such that this is small. Furthermore the integral $I[f-p]$ has an asymptotic expansion and hence it is also a good idea to take this into account. How to do this shows the following Lemma 5.1.2, established by the author of this thesis.

[2] The functions ψ_k in the article corresponds to our function Ψ_k.
[3] Therefor we (formally) have to set $r=1$ in Lemma 2.1 from [61]

Lemma 5.1.2. *Let the assumptions of Lemma 5.1.1 hold and let $p \in C^s(J,\mathbb{C})$ be given by* (5.7)*, such that for $k = 0,\dots,s-1$*

$$p^{(k)}(\alpha) = f^{(k)}(\alpha) \quad \textit{and} \quad p^{(k)}(\beta) = f^{(k)}(\beta)\,. \tag{5.9}$$

We set $\eta_0 := f - p$ and inductively define for $j = 1,\dots,s$

$$\eta_j \;:=\; \left(\frac{\eta_{j-1}}{\varphi'}\right)'.$$

Then the quadrature (5.8) *induced by p yields the error estimate*

$$\big|I[f] - Q^L[f]\big| \;\le\; \min_{k=0,\dots,s} \varepsilon^k \big|I[\eta_k]\big|\,. \tag{5.10}$$

If p, f, φ' are even in $C^{s+1}(J,\mathbb{C})$, then it additionally holds

$$\big|I[f] - Q^L[f]\big| \;\le\; \varepsilon^{s+1}\big(\max_{x=\alpha,\beta}\big|\tfrac{\eta_s(x)}{\varphi'(x)}\big| + |I[\eta_{s+1}]|\big)\,.$$

Proof. From Lemma 5.1.1 we get for[4] $k = 0,\dots,s$

$$I[f] - Q^L[f] \;=\; I[f-p] \;=\; Q_k^A[f-p] + (-i\varepsilon)^k I[\eta_k]\,.$$

Due to definition it is $\eta_0^{(0)}(\zeta) = \cdots = \eta_0^{(s-1)}(\zeta) = 0$, which yields with (5.17) from Lemma 5.2.3 that $\eta_0(\zeta) = \cdots = \eta_{s-1}(\zeta) = 0$ for $\zeta = \alpha, \beta$. Hence we get for all $k \in \{0,\dots,s\}$

$$Q_k^A[f-p] \;=\; 0\,.$$

This yields

$$\big|I[f] - Q^L[f]\big| \;\le\; \varepsilon^k \big|I[\eta_k]\big|\,.$$

For $p, f, \varphi' \in C^{s+1}(J,\mathbb{C})$ we additionally get from Lemma 5.1.1:

$$I[f] - Q^L[f] \;=\; -(-i\varepsilon)^{s+1}\tfrac{\eta_s(x)}{\varphi'(x)} e^{\frac{i}{\varepsilon}\varphi(x)}\Big|_{x=a}^{b} + (-i\varepsilon)^{s+1} I[\eta_{s+1}]\,.$$

□

Remark 5.1.3. *The estimate* (5.10) *is not only a statement about the asymptotic accuracy of the quadrature rule. Since $\eta_0 = f - p$, it also takes the approximation error of p with respect to f into account.*

5.2 The modified Filon–type method

The quadrature we shall use for our one–step method (see § 6) is based on the Levin–type method presented in § 5.1. Hence we have to specify how the functions $\Psi_0,\dots,\Psi_{m-1}$ and respectively $\psi_0,\dots,\psi_{m-1}$ should look like. In [61] Olver derives a basis for the more general problem of highly oscillatory integrals with a single stationary point, i. e. $\varphi^{(1)}(x) > 0$ for $x \in [\alpha,\beta]\backslash\{\zeta\}$ and

$$\varphi^{(j)}(\zeta) = 0\,, \quad 0 \le j \le r-1\,, \qquad \varphi^{(r)}(\zeta) > 0\,,$$

[4] $Q_0^A[f] = 0$ by definition, since it is an empty sum.

with $\zeta \in (\alpha, \beta)$. In our setting we have $r = 1$ and deduce from [61] that

$$\psi_k(x) \quad := \quad \varphi'(x)\,\varphi^k(x)\,, \qquad k \in \mathbb{N}_0\,. \tag{5.11}$$

The generalized moments $I[\psi_k]$, and hence the functions Ψ_k, can exactly be computed with repeated integrations by parts (see Lemma 5.2.2).

The Levin–type approach with the special set of functions from (5.11) was suggested to the author by Claudia Negulescu in 2006. In 2008 we became aware of the cited references [60, 61] due to a discussion with Markus Melenk. The articles only contain quadrature error estimates in terms of the small parameter ε. Since we want to use the quadrature for our one–step method, we are interested in estimates of the quadrature error with respect to the length of the integration interval. Hence we shall establish new results, which can not be derived from the articles. However, the quadrature coincides exactly with the moment–free Filon–type method from [61] for our special choice of functions.

The following Proposition 5.2.1 is the main result of this section. Since we want to apply Lemma 5.1.2 and benefit from the asymptotic structure of $I[f]$, the boundary points α, β have to be support abscissas. Let $\Omega := J \times (0, \varepsilon_1)$.

Proposition 5.2.1. *Let $f\colon \Omega \to \mathbb{C}$ and $\varphi\colon \Omega \to \mathbb{R}$, such that f, φ' are C^s–bounded independently of ε and such that $|\varphi'(x,\varepsilon)| \geq \delta > 0$ for all $(x,\varepsilon) \in \Omega$. Furthermore let $\xi_1, \ldots, \xi_\kappa \in J$ be support abscissas with corresponding multiplicities $1 \leq m_1, \ldots, m_\kappa \leq s+1$, such that there are indices j_α, j_β with*

$$\xi_{j_\alpha} \;=\; \alpha \quad \text{and} \quad \xi_{j_\beta} \;=\; \beta\,.$$

Then there exists a unique function

$$p(x,\varepsilon) \quad := \quad \varphi'(x,\varepsilon) \sum_{j=0}^{m-1} c_j(\varepsilon)\,\varphi(x,\varepsilon)^j\,, \tag{5.12}$$

with $m := \sum_{j=1}^{\kappa} m_j$ and $c_0(\varepsilon), \ldots, c_{m-1}(\varepsilon) \in \mathbb{C}$, such that

$$p^{(k)}(\xi_j,\varepsilon) \;=\; f^{(k)}(\xi_j,\varepsilon) \quad \text{for} \quad k = 0, \ldots, m_j - 1\,,\;\; j = 1, \ldots, \kappa\,. \tag{5.13}$$

If $s \geq m$, then the quadrature

$$Q[f] \quad := \quad I[p] \;=\; i\varepsilon\, e^{-\frac{i}{\varepsilon}\varphi(x,\varepsilon)} \sum_{k=0}^{m-1} \Bigg(\sum_{l=k}^{m-1} c_l(\varepsilon) \frac{l!}{k!} (-i\varepsilon)^{l-k} \Bigg) \varphi(x,\varepsilon)^k \Bigg|_{x=\alpha}^{\beta}$$

induced by p yields the error estimate

$$|I[f] - Q[f]| \quad \leq \quad c\,|\alpha - \beta|\,h^m \min\left(1, \gamma \left(\frac{\varepsilon}{h}\right)^{\mu+1}\right),$$

with

$$\mu := \min(m_{j_\alpha}, m_{j_\beta}) \quad \text{and} \quad h := \max\left(|\xi_\kappa - \alpha|, |\xi_1 - \beta|\right).$$

The constants $\gamma, c \geq 0$ depend on δ, $\|\varphi\|_{C^{m+1}(J)}$ and $\|f\|_{C^m(J)}$, but not on ξ. Furthermore the constants tend to infinity as $\delta \to 0$.

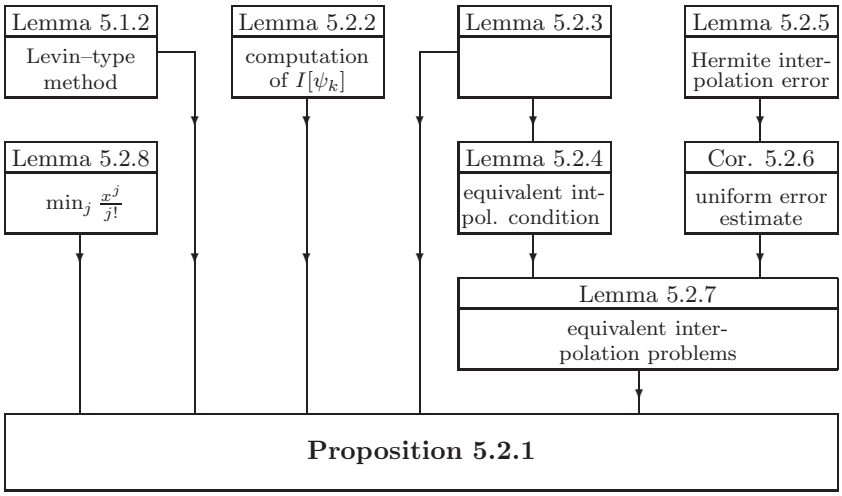

Figure 5.1: The arrows indicate the direction of dependence. For example Corollary 5.2.6 follows from Lemma 5.2.5.

Before we give a proof of Proposition 5.2.1 (see p.98) we derive some auxiliary results that help us to keep it more readable. The relations of these lemma are sketched in Figure 5.1.

We start with the derivation of an exact formula for $I[\psi_k]$ which is the content of Lemma 5.2.2. Next we prove the technical Lemma 5.2.3. With this we derive Lemma 5.2.4. The main idea of this result is that the interpolation problem (5.13) is equivalent to a Hermite interpolation problem in the ordinary polynomial sense. Lemma 5.2.5 gives an exact representation of the interpolation error for the ordinary Hermite interpolation problem for polynomials. This yields Corollary 5.2.6, which gives an uniform error estimate of the interpolation error. Then we establish Lemma 5.2.7, which already proves half of Proposition 6.3.1. Lemma 5.1.1 is already discussed in § 5.1.

Lemma 5.2.2. *Let $\varphi \in C^1(J, \mathbb{C})$ and $k \in \mathbb{N}$ and let ψ_k as in (5.11). It holds*

$$I[\psi_k] \quad = \quad i\varepsilon\, e^{-\frac{i}{\varepsilon}\varphi(x)} \sum_{l=0}^{k} \tfrac{k!}{l!}(-i\varepsilon)^{k-l}\varphi(x)^l \Big|_{x=\alpha}^{\beta} . \tag{5.14}$$

Proof. Due to (5.5), (5.6) we have to prove that ${\Psi'}_K - \frac{i}{\varepsilon}\varphi'\Psi_k = \psi_k$, with

$$\Psi_k(x) \quad := \quad i\varepsilon \sum_{l=0}^{k} \tfrac{k!}{l!}(-i\varepsilon)^{k-l}\varphi(x)^l .$$

It follows

$$\begin{aligned} &{\Psi'}_K - \tfrac{i}{\varepsilon}\varphi'\Psi_k \\ &= \quad -\varphi' \sum_{l=1}^{k} \tfrac{k!}{(l-1)!}(-i\varepsilon)^{k-(l-1)}\varphi^{l-1} + \varphi' \sum_{l=0}^{k} \tfrac{k!}{l!}(-i\varepsilon)^{k-l}\varphi(x)^l \\ &= \quad \varphi'\, \varphi^k . \end{aligned}$$

□

As we will see in the sequel, the interpolation problem (5.13) is closely related to ordinary polynomial interpolation. In order to prove existence and uniqueness of p and to derive an error bound let us rewrite the quadrature error. The change of variable $x = \varphi^{-1}(\xi)$ yields

$$\begin{aligned} I[f-p] &= \int_{\varphi(\alpha)}^{\varphi(\beta)} \frac{f\big(\varphi^{-1}(\xi)\big) - p\big(\varphi^{-1}(\xi)\big)}{\varphi'(\varphi^{-1}(\xi))}\, e^{-\frac{i}{\varepsilon}\xi}\, d\xi \\ &= \int_{\varphi(\alpha)}^{\varphi(\beta)} (g-\pi)(\xi)\, e^{-\frac{i}{\varepsilon}\xi}\, d\xi\,, \end{aligned}$$

with

$$g(\xi) := \frac{f(\varphi^{-1}(\xi))}{\varphi'(\varphi^{-1}(\xi))} \quad \text{and} \quad \pi(\xi) := \sum_{k=0}^{m-1} c_k\, \xi^k\,. \tag{5.15}$$

In the following Lemma 5.2.4 we prove that the interpolation conditions (5.13) are equivalent[5] to

$$\pi^{(k)}(\xi_j) = g^{(k)}(\xi_j)\,, \quad k = 0, \dots, m_j - 1\,,\ j = 1, \dots, \kappa\,, \tag{5.16}$$

where we set $\xi_j := \varphi^{-1}(\zeta_j)$ for $j = 1, \dots, \kappa$. Hence π is the unique Hermite interpolation polynomial of degree $m-1$ with respect to (5.16). Thus we get existence, uniqueness and error bounds from the ordinary Hermite interpolation theory.

To prove Lemma 5.2.4 we need

Lemma 5.2.3. *Let $\eta, \phi \in C^s(J, \mathbb{C})$ such that $|\phi(x)| \geq \delta > 0$ for all $x \in J$. We set $\eta_0 := \eta$ and inductively define for $j = 1, \dots, s$*

$$\eta_j := \left(\frac{\eta_{j-1}}{\phi}\right)'.$$

It holds for all $k \in \{0, \dots, s\}$

$$\eta_k = \frac{1}{\phi^k}\, \eta^{(k)} + \frac{1}{\phi^{2k}} \sum_{j=0}^{k-1} \gamma_{kj}\, \eta^{(j)}\,. \tag{5.17}$$

The functions γ_{kj} are multivariate polynomials in $\phi^{(0)}, \dots, \phi^{(k)}$ and independent of η.

Proof. Obviously, equation (5.17) holds for $k = 0$. Assume that the claim holds for k. We compute

$$\begin{aligned} \eta_{k+1} &= \frac{1}{\phi}\eta_k' - \frac{\phi'}{\phi^2}\eta_k \\ &= \frac{1}{\phi}\frac{\eta^{(k+1)}\phi^k - \eta^{(k)}(\phi^k)'}{\phi^{2k}} + \frac{1}{\phi}\sum_{j=0}^{k-1}\left(\frac{\gamma_{kj}\,\eta^{(j)}}{\phi^{2k}}\right)' - \frac{\phi'}{\phi^2}\eta_k \\ &= \frac{\eta^{(k+1)}}{\phi^{k+1}} - \frac{(k+1)\phi'\,\eta^{(k)}}{\phi^{k+2}} + \sum_{j=0}^{k-1}\frac{(\gamma_{kj}\,\eta^{(j)})'\phi - (2k+1)\phi'\gamma_{kj}\,\eta^{(j)}}{\phi^{2k+2}}\,. \end{aligned}$$

[5] Here "equivalent" means that both interpolation problems yield the same set of constants $c_0, \dots, c_{m-1}$.

Here we used two times (5.17) in order to write η_k in terms of $\eta^{(0)}, \dots, \eta^{(k)}$. By assumption γ_{kj} is a multivariate polynomial in $\phi^{(0)}, \dots, \phi^{(k)}$. Hence the coefficients of $\eta^{(j)}, \eta^{(j+1)}$ in $(\gamma_{kj}\, \eta^{(j)})'$ are multivariate polynomials in $\phi^{(0)}, \dots, \phi^{(k+1)}$. This completes the induction. □

Lemma 5.2.4. *Let the functions $f \in C^s(J, \mathbb{C})$ and $\varphi \in C^{s+1}(J, \mathbb{R})$, such that $|\varphi'(x)| \geq \delta > 0$ for all $x \in J$. Further let p, π and g be given by* (5.12), (5.15).

(i) *Let $\zeta \in J$. It is equivalent*

(a) $\forall\ k \in \{0, \dots, \mu\}\colon\ p^{(k)}(\zeta) = f^{(k)}(\zeta)$,

(b) $\forall\ k \in \{0, \dots, \mu\}\colon\ \pi^{(k)}\big(\varphi(\zeta)\big) = g^{(k)}(\varphi(\zeta))$.

(ii) *It holds for all $\xi \in [\varphi(\alpha), \varphi(\beta)]$ and $k = 0, \dots, s$*

$$g^{(k)}(\xi) \;=\; \frac{f_k(\varphi^{-1}(\xi))}{\varphi'(\varphi^{-1}(\xi))}\,, \qquad \pi^{(k)}(\xi) \;=\; \frac{p_k(\varphi^{-1}(\xi))}{\varphi'(\varphi^{-1}(\xi))}\,,$$

where we inductively define for $j = 1, \dots, s$

$$f_0 := f\,, \quad f_j := \big(\tfrac{f_{j-1}}{\varphi'}\big)' \quad \textit{and} \quad p_0 := p\,, \quad p_j := \big(\tfrac{p_{j-1}}{\varphi'}\big)'.$$

Proof. Since φ' is continuous and bounded away from zero, φ is strictly monotone. Hence the inverse φ^{-1} is well defined on J and it holds $\varphi^{-1} \in C^{s+1}(J, \mathbb{C})$. Due to definition of g (see (5.15))

$$\begin{aligned} g^{(1)}(\xi) \;&=\; \frac{d}{d\xi}\left(\frac{f(\varphi^{-1}(\xi))}{\varphi'(\varphi^{-1}(\xi))}\right) \;=\; \frac{d}{dx}\left(\frac{f(x)}{\varphi'(x)}\right)\Big|_{x=\varphi^{-1}(\xi)} \frac{d}{d\xi}\varphi^{-1}(\xi) \\ &=\; \frac{f_1(\varphi^{-1}(\xi))}{\varphi'(\varphi^{-1}(\xi))}\,. \end{aligned}$$

By induction we deduce

$$g^{(k)}(\xi) \;=\; \frac{f_k(\varphi^{-1}(\xi))}{\varphi'(\varphi^{-1}(\xi))} \quad \text{for} \quad k = 0, \dots, s\,.$$

From (5.12), (5.15) we get

$$\pi(\xi) \;=\; \frac{p(\varphi^{-1}(\xi))}{\varphi'(\varphi^{-1}(\xi))}$$

and analog to g we prove by induction

$$\pi^{(k)}(\xi) \;=\; \frac{p_k(\varphi^{-1}(\xi))}{\varphi'(\varphi^{-1}(\xi))} \quad \text{for} \quad k = 0, \dots, s\,.$$

This yields (ii).

Let $\zeta \in J$ and set $\xi := \varphi(\zeta)$. Since $\varphi'(\zeta) \neq 0$ we deduce form (ii) that (i)(b) is equivalent to

$$p_k(\zeta) \;=\; f_k(\zeta) \quad \text{for} \quad k \;=\; 0, \dots, \mu\,. \tag{5.18}$$

Since $|\varphi'| \geq \delta > 0$, we can use Lemma 5.2.3 and deduce for $k = 0, \ldots, \mu$ that $p_k(\zeta) = f_k(\zeta)$ is equivalent to

$$p^{(k)}(\zeta) - f^{(k)}(\zeta) \quad = \quad \big(\varphi'(\zeta)\big)^k \sum_{j=0}^{k-1} \gamma_{kj}(\zeta)\big(f^{(j)}(\zeta) - p^{(j)}(\zeta)\big)\,.$$

Hence we get by induction with respect to k that the above statement (5.18) is equivalent to (i)(a). □

Since our approximation relies on Hermite interpolation, we need an estimate for the approximation error. The following Lemma 5.2.5 is a result from [44].

Lemma 5.2.5 (Error representation)**.** *Let $J' = [\zeta_1, \zeta_\kappa] \subset \mathbb{R}$ and let the function $g \in C^s(J', \mathbb{C})$ and let $\zeta_1 < \cdots < \zeta_\kappa$ be supporting abscissas in J' with multiplicities $m_1, \ldots, m_\kappa \in \mathbb{N}$ and $m := \sum_{j=1}^{\kappa} m_j \leq s$. Further let π be the corresponding unique Hermite interpolation polynomial of degree $m-1$, i. e.*

$$\pi^{(k)}(\zeta_j) \;=\; g^{(k)}(\zeta_j)\,, \qquad k = 0, \ldots, m_j - 1\,, \quad j = 1, \ldots, \kappa\,. \tag{5.19}$$

Let $0 \leq r \leq m$. Then there exist $\zeta_1^r, \ldots, \zeta_{m-r}^r \in J'$, such that for each $x \in J'$ there exists a $\zeta^r = \zeta^r(x) \in J'$ with

$$g^{(r)}(x) - \pi^{(r)}(x) \quad = \quad \frac{g^{(m)}(\zeta^r)}{(m-r)!} \prod_{j=1}^{m-r} (x - \zeta_j^r)\,. \tag{5.20}$$

We immediately deduce from the previous Lemma 5.2.5

Corollary 5.2.6. *Let the assumptions of Lemma 5.2.5 hold and let $0 \leq r \leq m$. Then for all $x \in [\alpha', \beta'] \subset [\zeta_1, \zeta_\kappa]$ the following uniform estimate holds:*

$$|g^{(r)}(x) - \pi^{(r)}(x)| \quad \leq \quad \frac{h^{m-r}}{(m-r)!}\, \|g^{(m)}\|_\infty\,.$$

Here $h := \max\big(|\zeta_1 - \beta'|\,,\ |\zeta_\kappa - \alpha'|\big)$.

As in Proposition 5.2.1 let $\Omega := J \times (0, \varepsilon_0)$. The first part of the Proposition's proof is covered by

Lemma 5.2.7. *Let $f\colon \Omega \to \mathbb{C}$ and $\varphi\colon \Omega \to \mathbb{R}$, such that f, φ' are C^s–bounded independently of ε and such that $|\varphi'(x, \varepsilon)| \geq \delta > 0$ for all $(x, \varepsilon) \in \Omega$. Furthermore let $\xi_1, \ldots, \xi_\kappa \in J$ be support abscissas with corresponding multiplicities $1 \leq m_1, \ldots, m_\kappa \leq s+1$. Then there exists one and only one function*

$$p(x, \varepsilon) \quad := \quad \varphi'(x, \varepsilon) \sum_{j=0}^{m-1} c_j(\varepsilon)\, \varphi(x, \varepsilon)^j\,,$$

with $m := \sum_{j=1}^{\kappa} m_j$ and $c_0(\varepsilon), \ldots, c_{m-1}(\varepsilon) \in \mathbb{C}$, such that for all $\varepsilon \in (0, \varepsilon_0)$:

$$p^{(k)}(\xi_j, \varepsilon) \;=\; f^{(k)}(\xi_j, \varepsilon) \quad \textit{for} \quad k = 0, \ldots, m_j - 1\,,\ j = 1, \ldots, \kappa\,. \tag{5.21}$$

Let $[\alpha, \beta] \subset [\xi_1, \xi_\kappa]$. If additionally $s \geq m$, then there exists a constant $c \geq 0$ independent of ε and the support abscissas, such that for all $x \in [\alpha, \beta]$:

$$|p^{(k)}(x, \varepsilon) - f^{(k)}(x, \varepsilon)| \quad \leq \quad c\, h^{m-k}\,,$$

with $h := \max\big(|\xi_\kappa - \alpha|, |\xi_1 - \beta|\big)$. The constant $c \geq 0$ depends on δ, $\|\varphi\|_{C^{m+1}(J)}$ and $\|f\|_{C^m(J)}$, but not on ξ. Furthermore it tends to infinity as $\delta \to 0$.

Proof. Let $\varepsilon \in (0, \varepsilon_0)$ fix. For simplicity of notation we drop the second argument in all functions. I. e. whenever we write f, we mean $f(\cdot, \varepsilon)$.

Since φ' is (uniformly) bounded away from zero, φ is strictly monotone. Thus we can define g by (5.15). The points $\zeta_j := \varphi(\xi_j)$ for $j = 1, \ldots, \kappa$ are pairwise distinct. Hence the Hermite interpolation problem

$$\pi^{(k)}(\zeta_j) \;=\; g^{(k)}(\zeta_j) \quad \text{for} \quad k = 0, \ldots, m_j - 1\,, \quad j = 1, \ldots, \kappa\,, \qquad (5.22)$$

has a unique solution $\pi(x) = \sum_{j=0}^{m-1} c_j x^j$ [68]. From Lemma 5.2.4 we deduce that the (polynomial) interpolation problem (5.22) is equivalent to (5.21). This yields existence and uniqueness of p.

Let $x \in [\alpha, \beta]$, $\zeta := \varphi(x)$ and let f_k, p_k as in Lemma 5.2.4. Further we define the function $\eta := f - p$. Then

$$\eta_k(x) \;:=\; f_k(x) - p_k(x) \;=\; \varphi'(x)\big(g^{(k)}(\zeta) - \pi^{(k)}(\zeta)\big)\,.$$

This yields with Corollary 5.2.6[6] and the mean value theorem

$$\begin{aligned}
|\eta_k(x)| &\le \|\varphi'\|_\infty \left| g^{(k)}(\zeta) - \pi^{(k)}(\zeta) \right| \\
&\le \|\varphi'\|_\infty \frac{\max\big(|\varphi(\xi_\kappa) - \varphi(\alpha)|,\, |\varphi(\xi_1) - \varphi(\beta)|\big)^{m-k}}{(m-k)!}\, \|g^{(m)}\|_\infty \\
&\le \|\varphi'\|_\infty \frac{\|\varphi'\|_\infty^{m-k} \max\big(|\xi_\kappa - \alpha|,\, |\xi_1 - \beta|\big)^{m-k}}{(m-k)!}\, \left\| \frac{f_m}{\varphi'} \right\|_\infty .
\end{aligned}$$

Since $|\varphi'| \ge \delta$, we deduce from Lemma 5.2.3 that there exists a constant $\widehat{c} \ge 0$ independent of δ, such that

$$\left\| \frac{f_m}{\varphi'} \right\|_\infty \;\le\; \frac{\widehat{c}}{\delta^{2m+1}}\,.$$

Hence we get a constant $c \ge 0$, such that

$$|\eta_k(x)| \;\le\; \frac{\widehat{c}}{\delta^{2m+1}}\, \frac{\|\varphi'\|_\infty^{m-k+1}\, h^{m-k}}{(m-k)!} \;\le\; c\, h^{m-k}\,.$$

Furthermore we get from (5.17)

$$|\eta^{(k)}(x)| \;\le\; |\eta_k(k)| + \left\| \frac{1}{\varphi^k} \right\|_\infty \sum_{j=0}^{k-1} |\gamma_{kj}|\, |\eta^{(j)}(x)|\,.$$

For $k = 0$ this yields $|\eta^{(0)}(x)| \le c h^m$ and by induction we conclude that there exists a constant $c \ge 0$, such that

$$|f^{(k)}(x) - p^{(k)}(x)| \;=\; \eta^{(k)}(x)| \;\le\; c\, h^{m-k}$$

holds for $k = 0, \ldots, m$. □

Now we come to the last result before proving Proposition 5.2.1.

[6] Since φ is strictly monotone, it follows either $\varphi(\xi_1) \le \varphi(\alpha) < \varphi(\beta) \le \varphi(\xi_\kappa)$ or the reverse relation $\varphi(\xi_1) \ge \varphi(\alpha) > \varphi(\beta) \ge \varphi(\xi_\kappa)$.

Lemma 5.2.8. *Let $n, m \in \mathbb{N}_0$ with $n < m$. It holds for all $x \in \mathbb{R}^+$:*

$$\min_{j=n,\dots,m} \frac{x^j}{j!} \;=\; \begin{cases} \frac{x^m}{m!},\, x \le (\frac{m!}{n!})^{\frac{1}{m-n}} \\ \frac{x^n}{n!},\, x \ge (\frac{m!}{n!})^{\frac{1}{m-n}} \end{cases} .$$

The following proof is a revised version of our (more lengthy) previous one, where we act on a suggestion by Anton Arnold.

Proof. For $l \in \mathbb{N}_0$ we set $f(l) := \sum_{k=1}^{l} \ln k$. Let $x \in \mathbb{R}^+$ and define for $j \in \mathbb{N}_0$:

$$g(j) \;:=\; \ln\Big(\frac{x^j}{j!}\Big) \;=\; j\,\ln x \,-\, \sum_{k=1}^{j} \ln k \;=\; j\,\ln x \,-\, f(j)\,.$$

For $j \in \mathbb{N}$ it holds:

$$\begin{aligned} \frac{f(j-1)+f(j+1)}{2} - f(j) &= \frac{\ln j + \ln(j+1)}{2} - \ln j \\ &= \frac{\ln(j+1) - \ln j}{2} \;>\; 0\,. \end{aligned}$$

Thus $\frac{1}{2}f(j-1) + \frac{1}{2}f(j+1) > f(j) = f(\frac{1}{2}(j-1) + \frac{1}{2}(j+1))$, which can be interpreted as locally strict (midpoint) convexity of f. Since g is the sum of a linear function $(j \ln x)$ and $-f$ it holds

$$g(j-1) + g(j+1) \;<\; 2g(j)\,. \tag{5.23}$$

This means g is (in a discrete sense) strictly concave. Let $n < k < m$. Due to (5.23) it holds $g(k) > g(k-1)$ or $g(k) > g(k+1)$. In the first case it follows

$$g(k-2) - g(k) < 2g(k-1) \;\Leftrightarrow\; g(k-2) \;<\; g(k-1) - \big(g(k) - g(k-1)\big)\,.$$

Hence $g(k-2) < g(k-1)$ and by induction we get $g(n) < g(n+1) < \cdots < g(k)$. Analog we derive from $g(k) > g(k+1)$ that $g(k) > g(k+1) > \cdots > g(m)$. Hence $\min_{j=n,\dots,m} g(j) = \min\big(g(n), g(m)\big)$. It holds

$$\begin{aligned} g(m) - g(n) &= (m-n)\ln x \,-\, \big(\ln m! - \ln n!\big) \\ &= (m-n)\bigg(\ln x \,-\, \ln\Big(\frac{m!}{n!}\Big)^{\frac{1}{m-n}}\bigg)\,. \end{aligned}$$

Thus $g(m) \ge g(n)$, if and only if $x \ge \big(\frac{m!}{n!}\big)^{\frac{1}{m-n}}$ and consequently $g(m) \le g(n)$ for $x \le \big(\frac{m!}{n!}\big)^{\frac{1}{m-n}}$.

□

Proof of Proposition 5.2.1. By Lemma 5.2.7 we get existence and uniqueness of p. From Lemma 5.2.2 we deduce with Remark 6.2.8:

$$\begin{aligned} Q[f] &= i\varepsilon\, e^{-\frac{i}{\varepsilon}\varphi(x)} \sum_{k=0}^{m-1} \bigg(\sum_{l=k}^{m-1} c_l \tfrac{l!(-i\varepsilon)^{l-k}}{k!} \bigg) \varphi(x)^k \bigg|_{x=\alpha}^{\beta} \\ &= \sum_{k=0}^{m-1} c_k\; i\varepsilon\, e^{-\frac{i}{\varepsilon}\varphi(x)} \sum_{l=0}^{k} \tfrac{k!(-i\varepsilon)^{k-l}}{l!} \varphi(x)^l \bigg|_{x=\alpha}^{\beta} \\ &= \sum_{k=0}^{m-1} c_k I[\psi_k] \;=\; I[p]\,. \end{aligned}$$

Hence it is

$$I[f] - Q[f] \;=\; I[f] - I[p] \;=\; I[f-p]\,.$$

Let $\mu := \min(m_{j_\alpha}, m_{j_\beta})$. Since it holds $s \geq m \geq 2\mu$, we can set $s = \mu + 1$ in Lemma 5.1.2, which yields

$$|I[f] - Q[f]| \;\leq\; \begin{cases} \min_{l=0,\ldots,\mu} \varepsilon^l |I[\eta_l]|\,, \\ \varepsilon^{\mu+1}\big(\max_{x=\alpha,\beta} \left|\frac{\eta_\mu(x)}{\varphi'(x)}\right| + |I[\eta_{\mu+1}]|\big)\,. \end{cases} \tag{5.24}$$

The functions $\eta_0, \ldots, \eta_{\mu+1}$ are inductively defined by setting $\eta_0 := \eta := f - p$ and

$$\eta_k \;=\; \left(\frac{\eta_{k-1}}{\varphi'}\right)', \qquad k = 1, \ldots, \mu+1\,.$$

By Lemma 5.2.3

$$\eta_k \;=\; \frac{1}{(\varphi')^k}\,\eta^{(k)} + \frac{1}{(\varphi')^{2k}} \sum_{j=0}^{k-1} \gamma_{kj}\,\eta^{(j)}\,.$$

Thus, by Lemma 5.2.7, for each $k = 0, \ldots, m$ exists a polynomial q_k in h with positive coefficients, such that for all $x \in [\alpha, \beta]$

$$|\eta_k(x,\varepsilon)| \;\leq\; q_k(h)\,h^{m-k}\,.$$

Hence for $h_0 \geq 0$ exists a $c \geq 0$, such that for all $0 \leq h \leq h_0$ and all $x \in [\alpha, \beta]$

$$|\eta_k(x,\varepsilon)| \;\leq\; c_k\,h^{m-k}\,.$$

This yields with (5.24)

$$|I[f] - Q[f]| \;\leq\; \min_{l=0,\ldots,\mu} \varepsilon^l |\alpha-\beta|\,c_l\,h^{m-l}$$

and

$$\begin{aligned} |I[f] - Q[f]| \;&\leq\; \varepsilon^{\mu+1}\left(\frac{c_{\mu+1}}{\delta}\,h^{m-(\mu+1)} + |\alpha-\beta|\,c_{\mu+1}\,h^{m-(\mu+1)}\right) \\ &\leq\; c_{\mu+1}\big(\tfrac{1}{\delta} + |\alpha-\beta|\big)\,\varepsilon^{\mu+1}\,h^{m-(\mu+1)}\,. \end{aligned}$$

Combing both estimates yield (with a new constant $c_{\mu+1} \geq 0$)

$$|I[f] - Q[f]| \;\leq\; |\alpha-\beta| \min_{l=0,\ldots,\mu+1} c_l\,\varepsilon^l h^{m-l}\,.$$

By Lemma 5.2.8 we can restrict the min to the lowest and highest index. □

Remark 5.2.9. *Let us summarize the basic idea of the quadrature from Proposition 5.2.1. Since φ is continuously differentiable and strictly monotone, it holds*

$$\int_\alpha^\beta f(x)\,e^{-\frac{i}{\varepsilon}\varphi(x)}\,dx \;=\; \int_{\varphi(\alpha)}^{\varphi(\beta)} \tfrac{f(\varphi^{-1}(\xi))}{\varphi'(\varphi^{-1}(\xi))}\,e^{-\frac{i}{\varepsilon}\xi}\,d\xi\,.$$

Since the phase function of the right–hand side is linear, we can exactly compute the moments. This enables us to replace

$$g(\xi) \quad = \quad \tfrac{f(\varphi^{-1}(\xi)}{\varphi'(\varphi^{-1}(\xi))}$$

by a Hermite interpolation polynomial. Hence we use the Filon–type method to approximate the integral

$$\int_{\varphi(\alpha)}^{\varphi(\beta)} g(\xi)\, e^{-\frac{i}{\varepsilon}\xi}\, d\xi\,.$$

Thus, we can interpret the quadrature as a modified Filon–type method.

5.3 The symmetric shifted asymptotic method

In [4] the authors establish a quadrature for highly oscillatory integrals, which is called *shifted asymptotic method* (SAM). It is (as the asymptotic method) only based on integration by parts. Let us briefly point out the underlying idea. Therefor we make one integration by parts:

$$\begin{aligned}\int_\alpha^\beta f(x)\, e^{-\frac{i}{\varepsilon}\varphi(x)}\, dx \quad &= \quad i\varepsilon \int_\alpha^\beta \frac{f(x)}{\varphi'(x)} \left(e^{-\frac{i}{\varepsilon}\varphi(x)} - e^{-\frac{i}{\varepsilon}\varphi(a)}\right)' dx \qquad (5.25)\\ &= \quad i\varepsilon \frac{f(\beta)}{\varphi'(\beta)} \left(e^{-\frac{i}{\varepsilon}\varphi(\beta)} - e^{-\frac{i}{\varepsilon}\varphi(\alpha)}\right)\\ &\quad - i\varepsilon \int_\alpha^\beta \left(\frac{f}{\varphi'}\right)'(x) \left(e^{-\frac{i}{\varepsilon}\varphi(x)} - e^{-\frac{i}{\varepsilon}\varphi(\alpha)}\right) dx\,.\end{aligned}$$

Hence the integral in the last line (including the factor $i\varepsilon$) is of order $\mathcal{O}(|\alpha-\beta|^2)$ and

$$Q_\beta^{\mathrm{SAM}_1}[f] \quad := \quad i\varepsilon \frac{f(\beta)}{\varphi'(\beta)} \left(e^{-\frac{i}{\varepsilon}\varphi(\beta)} - e^{-\frac{i}{\varepsilon}\varphi(\alpha)}\right) \qquad (5.26)$$

is a quadrature of order $\mathcal{O}(|\alpha-\beta|^2)$. Further integration by parts leads to higher orders in $|\alpha-\beta|$. Thus the idea of the SAM is as follows: Use the asymptotic method (as described in Lemma 5.1.1) up to order n and then use integration by parts (as above) to approximate the remaining integral up to the desired order m in $|\alpha-\beta|$. This yields estimates of order $\mathcal{O}(\varepsilon^n|\alpha-\beta|^m)$.

In § 5.4 we shall (numerically) compare the SAM to our *modified Filon–type method* (MFM). There we shall use a SAM version, which has a quadrature error of at most $\mathcal{O}(|\alpha-\beta|^3)$. To derive it we have to make one more integration by parts in the above sense. Let $f_0 := \frac{f}{\varphi'}$ and $f_1 := \frac{f_0'}{\varphi'}$. Then

$$\begin{aligned}&\int_\alpha^\beta f_0'(x) \left(e^{-\frac{i}{\varepsilon}\varphi(x)} - e^{-\frac{i}{\varepsilon}\varphi(\alpha)}\right) dx\\ &= \quad i\varepsilon \int_\alpha^\beta \frac{f_0'(x)}{\varphi'(x)} \left(e^{-\frac{i}{\varepsilon}\varphi(x)} - e^{-\frac{i}{\varepsilon}\varphi(\alpha)} + \tfrac{i}{\varepsilon}(\varphi(x)-\varphi(\alpha))e^{-\frac{i}{\varepsilon}\varphi(\alpha)}\right)' dx\\ &= \quad i\varepsilon\, f_1(\beta) \left(e^{-\frac{i}{\varepsilon}\varphi(\beta)} - e^{-\frac{i}{\varepsilon}\varphi(\alpha)} + \tfrac{i}{\varepsilon}(\varphi(\beta)-\varphi(\alpha))e^{-\frac{i}{\varepsilon}\varphi(\alpha)}\right) + \dots\,.\end{aligned}$$

For simplicity of notation, in the sequel we use the abbreviation[7] g_α, g_β for $g(\alpha)$ and $g(\beta)$ respectively. This yields[8]

$$\begin{aligned} Q_\beta^{\mathrm{SAM}_2}[f] \;:=\; & i\varepsilon\; f_{0,\beta}\left(e^{-\frac{i}{\varepsilon}\varphi_\beta} - e^{-\frac{i}{\varepsilon}\varphi_\alpha}\right) \\ & -(i\varepsilon)^2 f_{1,\beta}\left(e^{-\frac{i}{\varepsilon}\varphi_\beta} - e^{-\frac{i}{\varepsilon}\varphi_\alpha} + \tfrac{i}{\varepsilon}(\varphi_\beta-\varphi_\alpha)e^{-\frac{i}{\varepsilon}\varphi_\alpha}\right). \end{aligned}$$

We arbitrarily choose the (added) constants, such that we (only) have to evaluate f_0, f_1 at $x=\beta$. But we also could have used other constants, such that we have to evaluate the functions at $x=\alpha$. Thus let us repeat the previous calculations, but replace α by β in the added constant terms. This yields

$$\begin{aligned} \int_\alpha^\beta f(x)\, e^{-\frac{i}{\varepsilon}\varphi(x)}\, dx \;=\; & -i\varepsilon f_0(\alpha)\left(e^{-\frac{i}{\varepsilon}\varphi(\alpha)} - e^{-\frac{i}{\varepsilon}\varphi(\beta)}\right) \\ & -i\varepsilon\int_\alpha^\beta f_0'(x)\left(e^{-\frac{i}{\varepsilon}\varphi(x)} - e^{-\frac{i}{\varepsilon}\varphi(\beta)}\right) dx\,. \end{aligned} \tag{5.27}$$

One more integration by parts of the remaining integral yields

$$\begin{aligned} & \int_\alpha^\beta f_0'(x)\left(e^{-\frac{i}{\varepsilon}\varphi(x)} - e^{-\frac{i}{\varepsilon}\varphi(\beta)}\right) dx \\ &= i\varepsilon\int_\alpha^\beta \frac{f_0'(x)}{\varphi'(x)}\left(e^{-\frac{i}{\varepsilon}\varphi(x)} - e^{-\frac{i}{\varepsilon}\varphi(\beta)} + \tfrac{i}{\varepsilon}(\varphi(x)-\varphi(\beta))e^{-\frac{i}{\varepsilon}\varphi(\beta)}\right)' dx \\ &= -i\varepsilon\, f_1(\alpha)\left(e^{-\frac{i}{\varepsilon}\varphi(\alpha)} - e^{-\frac{i}{\varepsilon}\varphi(\beta)} + \tfrac{i}{\varepsilon}(\varphi(\alpha)-\varphi(\beta))e^{-\frac{i}{\varepsilon}\varphi(\beta)}\right) + \dots\,. \end{aligned}$$

Hence we define

$$Q_\alpha^{\mathrm{SAM}_1}[f] \;:=\; i\varepsilon f_{0,\alpha}\left(e^{-\frac{i}{\varepsilon}\varphi_\beta} - e^{-\frac{i}{\varepsilon}\varphi_\alpha}\right) \tag{5.28}$$

and

$$\begin{aligned} Q_\alpha^{\mathrm{SAM}_2}[f] \;:=\; & i\varepsilon\; f_{0,\alpha}\left(e^{-\frac{i}{\varepsilon}\varphi_\beta} - e^{-\frac{i}{\varepsilon}\varphi_\alpha}\right) \\ & -(i\varepsilon)^2 f_{1,\alpha}\left(e^{-\frac{i}{\varepsilon}\varphi_\beta} - e^{-\frac{i}{\varepsilon}\varphi_\alpha} + \tfrac{i}{\varepsilon}(\varphi_\beta-\varphi_\alpha)e^{-\frac{i}{\varepsilon}\varphi_\beta}\right). \end{aligned}$$

In general, both quadratures (Q_α^{SAM}, Q_β^{SAM}) are equal. Hence, we expect that a symmetric[9] version of the SAM, which we shall call *symmetric shifted asymptotic method* (SSAM), yields a smaller approximation error than the SAM from [4]. For the lowest order we take the mean of (5.26) and (5.28) and define

$$Q_{\alpha,\beta}^{\mathrm{SSAM}_1}[f] \;:=\; \frac{i\varepsilon}{2}\big(f_{0,\alpha}+f_{0,\beta}\big)\left(e^{-\frac{i}{\varepsilon}\varphi_\beta} - e^{-\frac{i}{\varepsilon}\varphi_\alpha}\right).$$

Taylor expansion shows that this quadrature is of order $\mathcal{O}(\varepsilon^{-1}|\alpha-\beta|^3)$, while the non symmetric versions (5.26), (5.28) from [4] are of order $\mathcal{O}(|\alpha-\beta|^2)$.

To derive a symmetric quadrature based on two integration by parts, we could take the mean of $Q_\alpha^{\mathrm{SAM}_2}[f]$ and $Q_\alpha^{\mathrm{SAM}_2}[f]$. This scheme is denoted by

[7] Here g is of course a wild card the functions f_0, f_1, φ.

[8] The index 2 is the number of integrations by parts used to derive the quadrature.

[9] Interchanging the integration boundaries of an integral is equal to multiplying it by -1. Hence, here symmetry of the quadrature means that if we interchange α and β, than we get the negative quadrature.

mean shifted asymptotic method (MSAM). In § 5.4 we shall see that this yields only a slight improvement of the constant compared to the original quadrature. Instead we take the mean value of (5.25) and (5.27) and thus

$$\begin{aligned}\int_\alpha^\beta f(x)\, e^{-\frac{i}{\varepsilon}\varphi(x)}\, dx &= \frac{i\varepsilon}{2}(f_0(\beta)+f_0(\alpha))\left(e^{-\frac{i}{\varepsilon}\varphi(\beta)}-e^{-\frac{i}{\varepsilon}\varphi(\alpha)}\right)\\ &\quad - i\varepsilon\int_\alpha^\beta f_0'(x)\left(e^{-\frac{i}{\varepsilon}\varphi(x)}-\tfrac{1}{2}\left(e^{-\frac{i}{\varepsilon}\varphi(\beta)}+e^{-\frac{i}{\varepsilon}\varphi(\alpha)}\right)\right) dx\,.\end{aligned}$$

We proceed with integration by parts:

$$\begin{aligned}&\int_\alpha^\beta f_0'(x)\left(e^{-\frac{i}{\varepsilon}\varphi(x)}-\tfrac{1}{2}\left(e^{-\frac{i}{\varepsilon}\varphi(\beta)}+e^{-\frac{i}{\varepsilon}\varphi(\alpha)}\right)\right) dx\\ &= i\varepsilon\int_\alpha^\beta \frac{f_0'(x)}{\varphi'(x)}\left(e^{-\frac{i}{\varepsilon}\varphi(x)}-\tfrac{1}{2}\left(e^{-\frac{i}{\varepsilon}\varphi(\beta)}+e^{-\frac{i}{\varepsilon}\varphi(\alpha)}\right)\right)' dx\\ &\quad -\tfrac{1}{2}\left(e^{-\frac{i}{\varepsilon}\varphi(\beta)}+e^{-\frac{i}{\varepsilon}\varphi(\alpha)}\right)\int_\alpha^\beta \frac{f_0'(x)}{\varphi'(x)}\left(\varphi(x)-\frac{\varphi(\beta)+\varphi(\alpha)}{2}\right)' dx\\ &= \frac{i\varepsilon}{2}\left(f_1(\beta)+f_1(\alpha)\right)\left(e^{-\frac{i}{\varepsilon}\varphi(\beta)}-e^{-\frac{i}{\varepsilon}\varphi(\alpha)}\right)+\dots\\ &\quad -\tfrac{1}{2}\left(e^{-\frac{i}{\varepsilon}\varphi(\beta)}+e^{-\frac{i}{\varepsilon}\varphi(\alpha)}\right)\left(f_1(\beta)+f_1(\alpha)\right)\left(\varphi(\beta)-\varphi(\alpha)\right)+\dots\,.\end{aligned}$$

Thus we can define the (with respect to α,β) symmetric quadrature

$$\begin{aligned}Q_{\alpha,\beta}^{\mathrm{SAM}_2}[f] &:= \frac{i\varepsilon}{2}\,(f_{0,\beta}+f_{0,\alpha})\left(e^{-\frac{i}{\varepsilon}\varphi_\beta}-e^{-\frac{i}{\varepsilon}\varphi_\alpha}\right)\\ &\quad -\frac{(i\varepsilon)^2}{2}\,(f_{1,\beta}+f_{1,\alpha})\left(e^{-\frac{i}{\varepsilon}\varphi_\beta}-e^{-\frac{i}{\varepsilon}\varphi_\alpha}\right)\\ &\quad -\frac{(i\varepsilon)^2}{2}\,(f_{1,\beta}+f_{1,\alpha})\,\tfrac{i}{\varepsilon}(\varphi_\beta-\varphi_\alpha)\,\frac{e^{-\frac{i}{\varepsilon}\varphi_\beta}+e^{-\frac{i}{\varepsilon}\varphi_\alpha}}{2}\,.\end{aligned}\tag{5.29}$$

We shall call this quadrature *revised shifted asymptotic method* (RSAM). It has a bit better approximation properties than the MSAM and hence the SAM, but is still of third order as $h\to 0$.

Now let us compare the quadrature error of $Q_\alpha^{\mathrm{SAM}_2}$, $Q_\beta^{\mathrm{SAM}_2}$ and $Q_{\alpha,\beta}^{\mathrm{SAM}_2}$. For this purpose we shall use Taylor expansion of the approximation error for all three methods. Let

$$I(h) := \int_\alpha^{\alpha+h} f(x)e^{-\frac{i}{\varepsilon}\varphi(x)}\,dx\,.\tag{5.30}$$

With *Maple14* we derive (with $\beta=\alpha+h$)

$$\begin{aligned}I(h)-Q_\alpha^{\mathrm{SAM}_2}[f] &= \frac{1}{6}\,c\,e^{-\frac{i}{\varepsilon}\varphi(\alpha)}\,h^3+\mathcal{O}(h^4)\,,\\ I(h)-Q_\beta^{\mathrm{SAM}_2}[f] &= \frac{1}{6}\,c\,e^{-\frac{i}{\varepsilon}\varphi(\alpha)}\,h^3+\mathcal{O}(h^4)\,,\\ I(h)-Q_{\alpha,\beta}^{\mathrm{SAM}_2}[f] &= -\frac{1}{12}\,c\,e^{-\frac{i}{\varepsilon}\varphi(\alpha)}\,h^3+\mathcal{O}(h^4)\,.\end{aligned}$$

The constant c depends on f,φ and their derivatives. Furthermore it is equal in all three equations. Hence we can combine the three methods to a quadrature

which is symmetric (with respect to α, β) and of higher order with respect to the interval length h. The SSAM (with two integration by parts) is defined by

$$\begin{aligned} Q_{\alpha,\beta}^{\mathrm{SSAM}_2}[f] &:= \frac{1}{6}\big(Q_{\alpha}^{\mathrm{SAM}_2}[f] + 4Q_{\alpha,\beta}^{\mathrm{SAM}_2}[f] + Q_{\beta}^{\mathrm{SAM}_2}[f]\big) \\ &= \frac{i\varepsilon}{2}\ (f_{0,\beta}+f_{0,\alpha})\left(e^{-\frac{i}{\varepsilon}\varphi_\beta} - e^{-\frac{i}{\varepsilon}\varphi_\alpha}\right) \\ &\quad - \frac{(i\varepsilon)^2}{2}\,(f_{1,\beta}+f_{1,\alpha})\left(e^{-\frac{i}{\varepsilon}\varphi_\beta} - e^{-\frac{i}{\varepsilon}\varphi_\alpha}\right) \\ &\quad + \frac{i\varepsilon}{6}\ (\varphi_\beta-\varphi_\alpha)\,(f_{1,\beta}+f_{1,\alpha})\left(e^{-\frac{i}{\varepsilon}\varphi_\beta} + e^{-\frac{i}{\varepsilon}\varphi_\alpha}\right) \\ &\quad + \frac{i\varepsilon}{6}\ (\varphi_\beta-\varphi_\alpha)\left(f_{1,\alpha}e^{-\frac{i}{\varepsilon}\varphi_\beta} + f_{1,\beta}e^{-\frac{i}{\varepsilon}\varphi_\alpha}\right). \end{aligned}$$

The structure in the first line reminds us to the Simpson rule. And in fact Taylor expansion (with the aid of *Maple 14*) shows

$$I(h) - Q_{\alpha,\beta}^{\mathrm{SSAM}_2}[f] = \mathcal{O}(\varepsilon^{-2}h^5).$$

Thus our SSAM_2 is of fifth order with respect to h, while the SAM_2 is of third order. However, both quadratures use the same set of data. Hence we expect our SSAM to be much more efficient then the SAM from [4].

The SAM is a way to derive an asymptotic expansion of (5.30) with respect to h (the interval length) of an arbitrary order (cf. [4]). In the previous discussion we have restricted ourself to the special case, where we use only two times integration by parts. Here we are able to (significantly) improve the asymptotic order (with respect to h) of the SAM. However, a systematic way to construct symmetric versions of the SAM of "maximal" order (as our derived SSAM_2) for arbitrary numbers of integration by parts is not yet available.

5.4 Numerical experiments

In this section we shall illustrate the approximation accuracy of our modified Filon–type method (MFM) from Proposition 5.2.1 and the quadratures discussed in § 5.3. Furthermore we shall test the classical trapezoidal rule on two highly oscillatory examples. Mainly, we are interested in the convergence behavior of the quadrature error as the step size (interval length) tends to zero. Thus, for the SAM from [4] we shall not apply the asymptotic method, which is usually the first step and increases the asymptotic properties with respect to ε. We want to test the spatial approximation abilities of the quadratures. Therefor we shall also use different values of ε in order to visualize the dependency of the quadrature errors on this small parameter.

Let $f, \varphi\colon [a,b] \to \mathbb{R}$, such that the assumptions of Proposition 5.2.1 hold on the whole interval. For $N \in \mathbb{N}$ we set $h = \frac{b-a}{N}$ and $x_n = a + nh$. Thus

$$I := \int_a^b f(x)\, e^{-\frac{i}{\varepsilon}\varphi(x)}\, dx = \sum_{n=0}^{N-1} \int_{x_n}^{x_{n+1}} f(x)\, e^{-\frac{i}{\varepsilon}\varphi(x)}\, dx =: \sum_{n=0}^{N-1} I_n\,.$$

Now each integral I_n is approximated with a quadrature Q_n based on Proposi-

tion 5.2.1. We set $Q := \sum_{n=0}^{N-1} Q_n$ and get

$$\begin{aligned} |I - Q| &\leq \sum_{n=0}^{N-1} \left| I_n - Q_n \right| \leq \sum_{n=0}^{N-1} c\, h^{m+1} \min\left(1, \gamma\left(\frac{\varepsilon}{h}\right)^{\mu+1}\right) \\ &= c\,|a-b|\, h^m \left(1, \gamma\left(\frac{\varepsilon}{h}\right)^{\mu+1}\right). \end{aligned}$$

Hence the h–order of the composed quadrature Q is reduced by one with respect to those of Proposition 5.2.1. Nevertheless, the asymptotic behavior with respect to ε remains unchanged. Thus the composed version Q of the quadrature is suitable for our purpose.

Now we specify the version of our MFM we shall use for the numerical experiments. We choose the interpolation abscissas $\xi_1 = x_n$, $\xi_2 = x_{n+1}$ and the corresponding multiplicities $m_1 = m_2 = 1$. This is the simplest admissible setting. Thus $m = 2$, $\mu = 1$, and (for the subinterval $[x_n, x_{n+1}]$)

$$c_1 = \frac{\frac{f(x_n)}{\varphi'(x_n)} - \frac{f(x_{n+1})}{\varphi'(x_{n+1})}}{\varphi(x_n) - \varphi(x_{n+1})}, \qquad c_0 = \frac{f(x_n)}{\varphi'(x_n)} - c_1\,\varphi(x_n).$$

This yields

$$Q_n^{\mathrm{MFM}_1} = i\varepsilon\, e^{-\frac{i}{\varepsilon}\varphi(x)} \big(c_0 - i\varepsilon c_1 + c_1\,\varphi(x)\big)\big|_{x=x_n}^{x_{n+1}}.$$

Hence, from Proposition 5.2.1 we get the (theoretical) upper bound for the (absolute) quadrature error of Q^{MFM_1}:

$$|I - Q^{\mathrm{MFM}_1}| \leq c\,|a-b|\, h^2 \left(1, \gamma\left(\frac{\varepsilon}{h}\right)^2\right). \tag{5.31}$$

In Figure 5.2 we plot for $\varepsilon = 10^{-1}, \ldots, 10^{-5}$ the absolute quadrature error of the MFM_1 (circles) for the integral $I = \int_0^1 \log(1+x) e^{-\frac{i}{\varepsilon}x}\, dx$. The exact value of the integral can be written down in terms of exponential integrals. They are evaluated with the Matlab function `mfun`. Additionally we plot the theoretical error bound (5.31) with fitted constants $c = 0.02$, $\gamma = 0.7$ (solid line). The color code we use here to mark the different ε values shall be used in the sequel for all plots.

We observe that the error estimate (5.31) from Proposition 5.2.1 (solid line) is close to the numerical results in the constant regime ($h > \varepsilon$). Also the predicted convergence error of order two with respect to the spatial step size h is visible. However, the bound of Proposition 5.2.1 over estimates the error when it starts to decrease monotonously ($h < \varepsilon$). In this regime, where the second order nature with respect to h dominates, the numerical error seems to decay when $\varepsilon \to 0$. Contrary, (5.31) predicts an ε–independent quadrature error here. This phenomena is also observed in the following examples.

Before we start to compare our MFM_1 with the SAM from [4] and the new SSAM from § 5.3, we shall test the performance of the classical trapezoidal rule (TR) for an highly oscillatory integral. In Figure 5.3 we plot the absolute approximation error of the composed MFM_1 (triangle) and the composed TR (circle) for the integral $I = \int_0^1 \cos x\, e^{\frac{i}{\varepsilon}(x+1)^2} dx$. The exact value of the integral

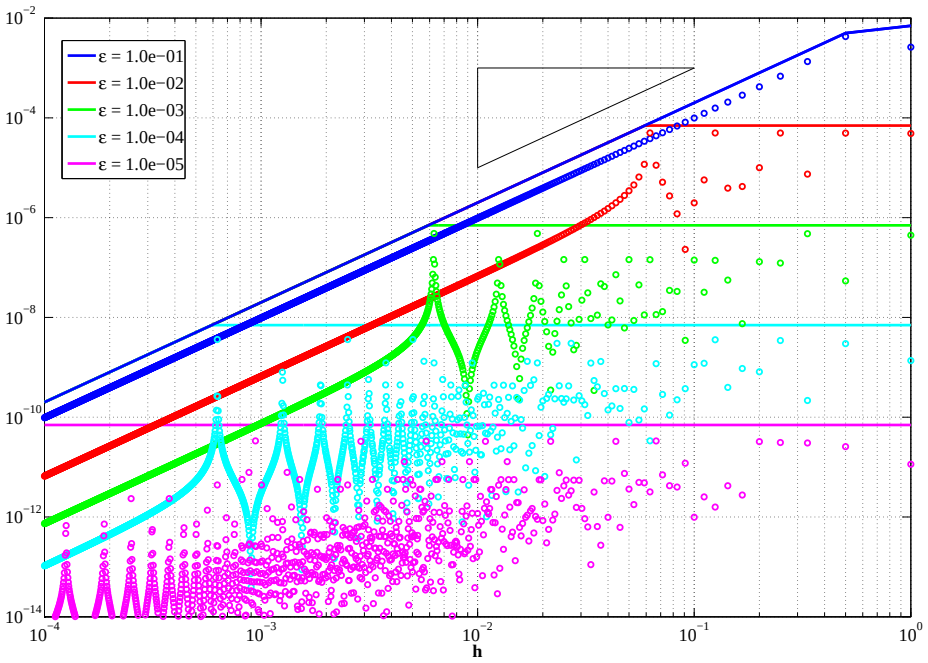

Figure 5.2: Error of the MFM (circles) and the estimate from Proposition 5.2.1 (solid line) with fitted constants ($c = 0.02$, $\gamma = 0.7$) for $I = \int_0^1 \log(1+x)e^{-\frac{i}{\varepsilon}x}\,dx$.

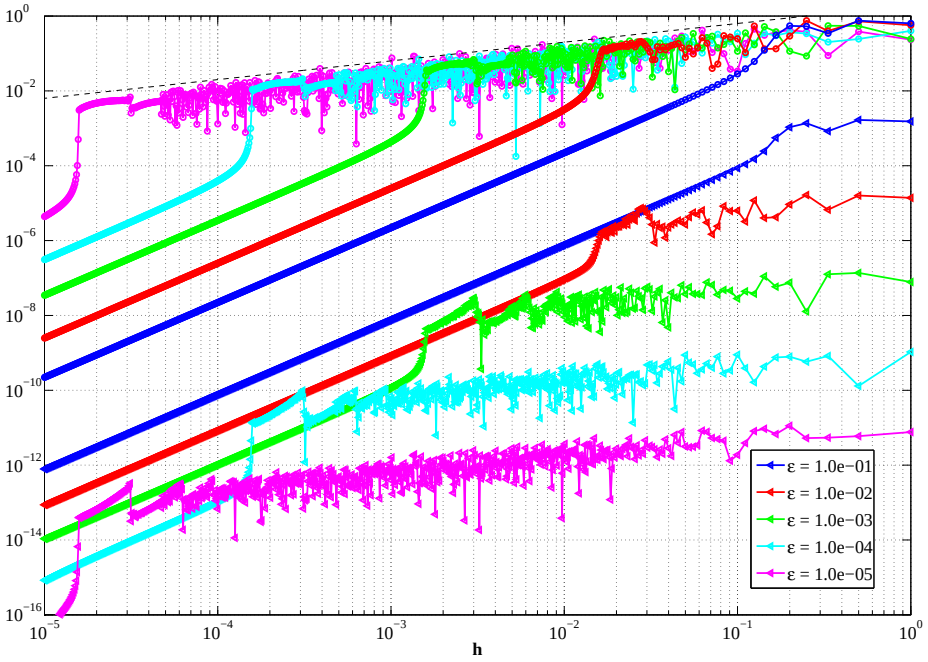

Figure 5.3: Absolute quadrature error for the TR (circles) and the MFM_1 (triangle) for $I = \int_0^1 \cos x\, e^{-\frac{i}{\varepsilon}(x+1)^2}\,dx$. The black reference line has slope $\frac{1}{2}$.

can be written down in terms of error functions. They are evaluated with the Matlab function `mfun`.

For small values of h ($h < \varepsilon$) the TR (circle) shows its predicted second order behavior with respect to h. Instead, for large values of h (i. e. $h > \varepsilon$) we observe only a slight decay of the quadrature error, almost independently of ε. The dashed black line in the upper part of the plot has slope $\frac{1}{2}$ and it seems to be a good upper bound, at least for this example. The point where the (smooth) second order behavior kicks in is proportional to ε. There we see (for fixed ε) a very steep descent with decreasing h. The height of this transition seems to increase when $\varepsilon \to 0$. Furthermore the error curves show an $\mathcal{O}(\varepsilon^{-1}h^2)$ behavior in the smooth regime, i. e. $h < \varepsilon$. However, the MFM_1 has a much better performance. We observe that the error is proportional to ε^2 for $h > \varepsilon$. Additionally for $h < \varepsilon$ the quadrature error is monotonously decreasing with second order in h. Furthermore we get from the plot that (for small h) the error of the MFM_1 is approximately ε^2 times smaller than the quadrature error of the TR, while using the same data. This is an amazing result, which shows how powerful our sophisticated quadrature is.

It is also remarkable that the error curves for both quadratures seem to have similar shape. At least the points where the oscillatory parts of the curves turns into a smooth monotone increasing line lie almost at the same position (for both quadratures). For the integral $I = \int_0^1 \log(x+1)e^{-\frac{i}{\varepsilon}x}dx$ the similarity of the error curve shapes is even stronger. For this example, in Figure 5.4 we plot the quadrature error of the TR (circle) for $\varepsilon = 10^{-1}, 10^{-2}, 10^{-3}$ and $h \in [10^{-3}, 1]$. We compute the results for all (1000) equidistant subdivisions of the interval $[0, 1]$, with subinterval length greater or equal to $h_{\min} = 10^{-3}$. Furthermore we plot the quadrature error of the MFM_1 (triangle) multiplied by the (approximate) factors $5.57 \cdot 10^1$, $8.61 \cdot 10^3$, $7.85 \cdot 10^5$ for $\varepsilon = 10^{-1}, 10^{-2}, 10^{-3}$ respectively. We choose the factors, such that the corresponding curves coincide at $h = 10^{-3}$. Here we numerically observe that the error of the TR is (approximately) $c\varepsilon^{-2}$ larger than that of the MFM_1. The vertically shifted error curves of the MFM_1 almost coincide with those of the TR. Differences of the curves are visible in the plot only for large values of h. We also observe this interesting coincidence for the integrals $I = \int_0^1 \cos x e^{-\frac{i}{\varepsilon}x}dx$ and $I = \int_0^1 x^3 e^{-\frac{i}{\varepsilon}x}dx$, which we do not plot here. For the example from Figure 5.3 the curves do not fit as good as in the cases with linear phase. However, if we create in $I = \int_0^1 \cos x \, e^{-\frac{i}{\varepsilon}(x+1)^2} \, dx$ a linear phase by the substitution $y = (x+1)^2$ and afterwards apply the quadrature rules, then we also get quite good (in shape) matching error curves.

Now let us consider the SAM[10]. In Figure 5.5 we plot the absolute quadrature error of the SAM_1 (solid line) and our SSAM_1 (circle). Both quadratures use the same set of data. We observe that the SAM_1 is of first order as $h \to 0$, while our SSAM_1 shows a second order behavior. The error curves for both quadratures are oscillating for $h > \varepsilon$ and it seems to be, that the error of SSAM_1 is bounded from above by that of SAM_1. However, the point where the oscillatory nature turns into a smooth, monotonously decreasing line is (approximately) the same for both quadratures and the graphs almost coincide at this position. Despite the SAM_1's asymptotic behavior of $\mathcal{O}(\varepsilon h)$ (which we deduce from the plot), the SSAM_1 is much more effective, even with its ε–independent convergence behavior.

[10] See § 5.3 for the definition of the SAM, MSAM, RSAM, SSAM.

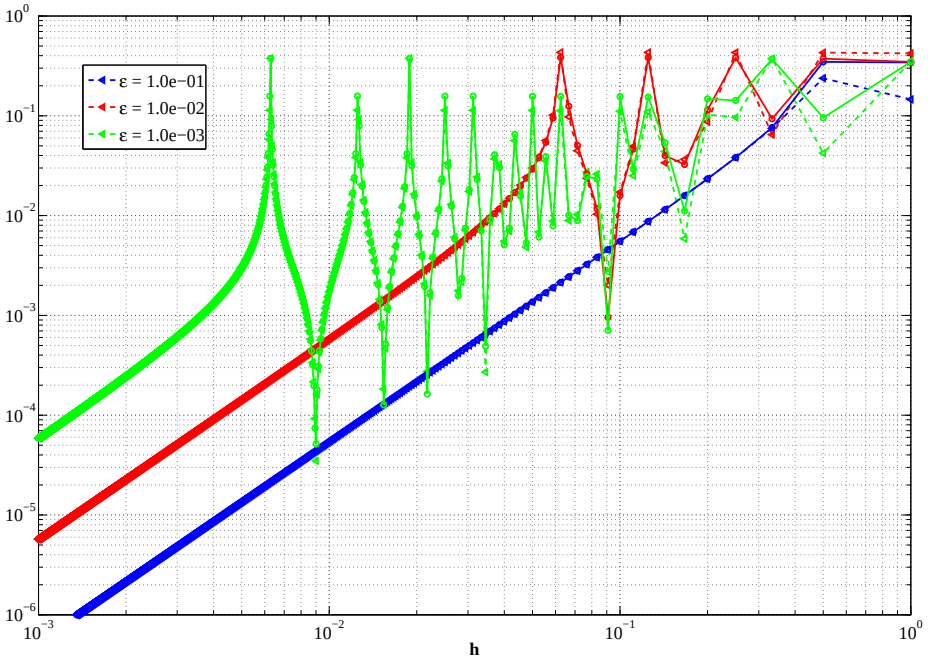

Figure 5.4: Absolute quadrature error of the TR (circles) and the vertically shifted error of the MFM_1 (triangle) for $I = \int_0^1 \log(x+1)\, e^{-\frac{i}{\varepsilon}x}\, dx$.

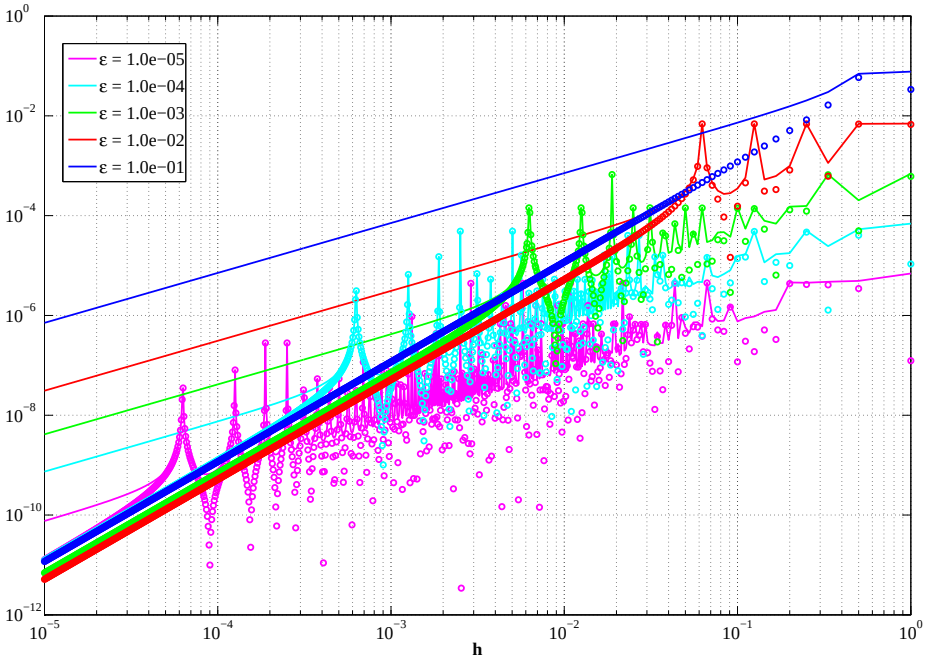

Figure 5.5: Absolute quadrature error of the SAM_1 from [4] with one integration by parts (solid line) and SSAM_1 from § 5.3 (circle) for $I = \int_0^1 \log(x+1)\, e^{-\frac{i}{\varepsilon}x}\, dx$.

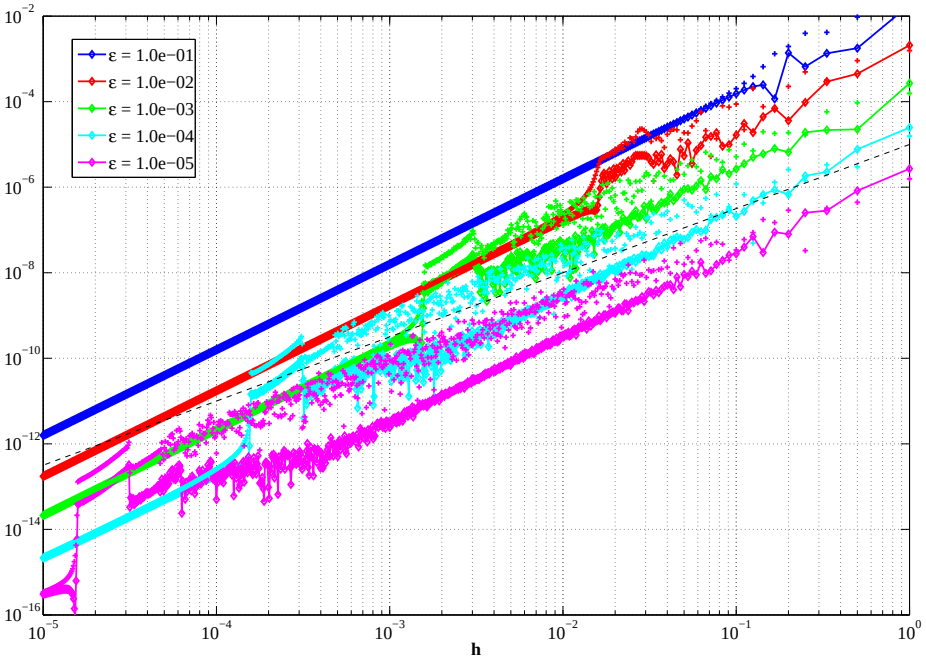

Figure 5.6: Absolute quadrature error of the SAM_2 (+) and MSAM (solid line) for $I = \int_0^1 \cos x \, e^{-\frac{i}{\varepsilon}(x+1)^2} \, dx$. The black dashed reference line has slope $\frac{3}{2}$.

In Figure 5.6 we plot the absolute quadrature error of the SAM_2 (+) and the MSAM (solid line) for $I = \int_0^1 \cos x \, e^{-\frac{i}{\varepsilon}(x+1)^2} \, dx$. Both quadratures use the same set of data. For $h \ll \varepsilon$ the error curves (almost) coincide. They show an $\mathcal{O}(\varepsilon h^2)$ behavior in this part. As observed for the MFM_1, SAM_1, SSAM_1 and TR, also the SAM_2 and MSAM are oscillatory for $h > \varepsilon$. However, the oscillations of the MSAM line are less distinct. Furthermore the MSAM shows a smaller error compared to the SAM_2. Except for a part where the oscillatory behavior turns into a smooth monotone growth, the MSAM quadrature error seems to decay (in the mean) with second order in h on the whole interval. For the SAM_2 instead we see a slower decay for $h > \varepsilon$. The thin black dashed line has slope $\frac{3}{2}$ which seems to be close to the decay rate. Since the MSAM shows less oscillatory error curves (compared to the SAM_2), we shall use it instead of the SAM_2 for further considerations. In all upcoming plots the error of the SAM_2 is always greater or equal than those of the MSAM.

Next we shall compare the MFM_1 with the MSAM. Furthermore we shall plot the quadrature error of the RSAM. In Figure 5.7 we plot the absolute quadrature error of the three methods for $I = \int_0^1 \cos x \, e^{-\frac{i}{\varepsilon}(x+1)^2} \, dx$. The same setting for $I = \int_0^1 \log(x+1) \, e^{-\frac{i}{\varepsilon}x} \, dx$ is plotted in Figure 5.8. We start with the discussion of Figure 5.7.

In the oscillatory part ($h > \varepsilon$) the error of the RSAM (circle) decays faster than that of the MSAM (dashed line). Also in the smooth regime the error of the RSAM has a smaller constant compared to the MSAM. For large h (i. e. $h > \varepsilon$) the MFM_1 yields the best performance. In this part the error is at

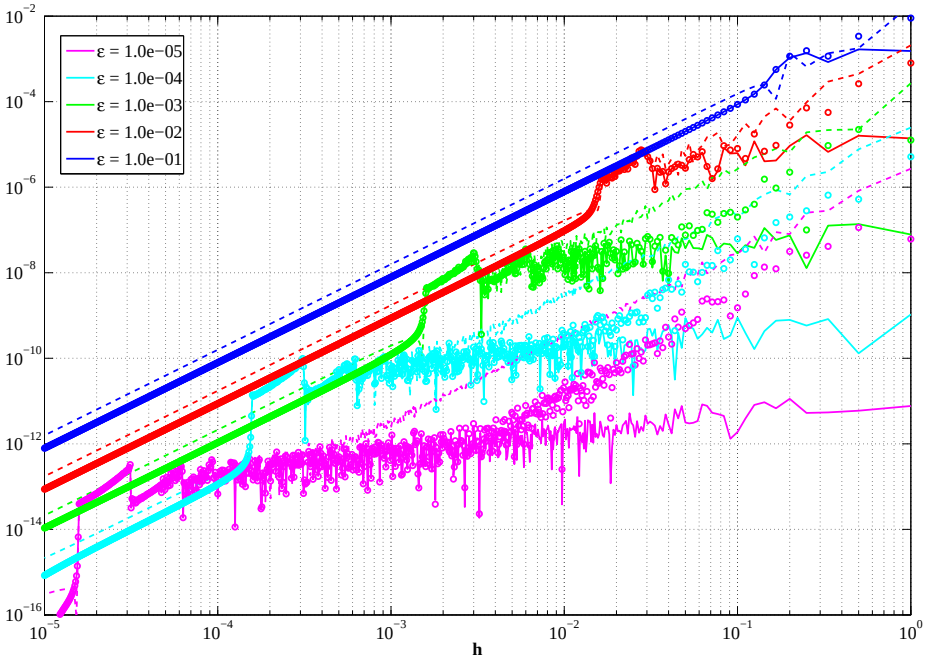

Figure 5.7: Absolute quadrature error of the MSAM (dashed line), RSAM (circle) and MFM_1 (solid line) for $I = \int_0^1 \cos x \, e^{-\frac{i}{\varepsilon}(x+1)^2} \, dx$.

most of order $\mathcal{O}(\varepsilon^2)$, even for (fixed) $h = 1$. Here the RSAM decays a bit faster than the already known $\mathcal{O}(\varepsilon h^2)$ behavior of the MSAM. Both methods show the described h–dependent decay until they reach the curve of the MFM_1. Nevertheless, in the transition region all three curves almost coincide. The three methods show an $\mathcal{O}(\varepsilon h^2)$ convergence behavior as $h \to 0$. The error of the RSAM and MFM_1 almost coincide for (very) small h.

In Figure 5.8 we see that the error curves do not always coincide at the transition region. Here the MSAM error shows a different shape compared to that of the RSAM and MFM_1, which are almost identical. The other observations from Figure 5.7 also hold here.

The RSAM and MSAM use the same data and hence yield comparable results with slight advantages for the RSAM. However, we do not need first derivatives of f for the MFM_1, contrary to the MSAM and RSAM. Hence the MFM_1 has the least numerical effort (of the three methods), while yielding the smallest approximation errors and the same asymptotic order when $h \to 0$.

In Figure 5.9 we plot the absolute quadrature error of the SAM_2 (from [4]) and our improved symmetric version SSAM_2 (derived in § 5.3) for the integral $I = \int_0^1 \cos x \, e^{-\frac{i}{\varepsilon}(x+1)^2} \, dx$. Both methods use the same set of data. We observe the predicted forth order of our SSAM_2 as $h \to 0$. On the whole interval the new method shows smaller approximation errors than the SAM_2. Consequently, due to the difference of two orders in the h–asymptotic, the error of the SSAM_2 is significantly smaller for $h < \varepsilon$. Furthermore we plot the error of the MFM_1 (dot). In the transition area the MFM_1 and SSAM_2 yield comparable results. Nevertheless, for $h > \varepsilon$ the MFM_1 once again shows the best performance.

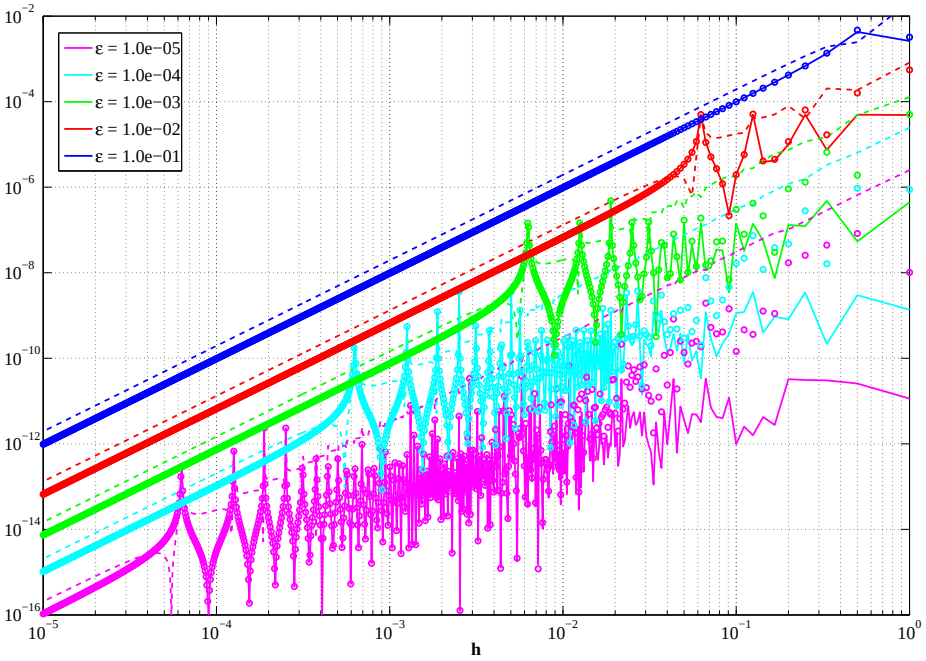

Figure 5.8: Absolute quadrature error of the MSAM (dashed line), RSAM (circle) and MFM_1 (solid line) for $I = \int_0^1 \log(x+1)\, e^{-\frac{i}{\varepsilon}x}\, dx$.

In Figure 5.10 we have the same setting as for Figure 5.9, except for the integral which is $I = \int_0^1 \log x\, e^{-\frac{i}{\varepsilon}x}\, dx$ here. The previously described observations from Figure 5.9 also hold for this Figure. In both Figures we observe oscillations of the SAM_2 curves in the lower left corner. A reason for this may be the machine precision of Matlab (which is approximately 10^{-16}). Furthermore, also the accuracy of the reference solution may be reached in this regime.

In Figure 5.7 and Figure 5.8 we observe that the RSAM and and MFM_1 have a similar asymptotic behavior as $h \to 0$. Using Taylor expansion (which is done with Maple 14) we (locally) get

$$Q^{\mathrm{SAM}}_{\alpha,\beta,n} - Q^{\mathrm{MFM}_1}_n \quad = \quad c\, e^{-\frac{i}{\varepsilon}\varphi(x_n)}\, \varepsilon^{-1} h^5 + \mathcal{O}(\varepsilon^{-2} h^6)\,.$$

Hence we expect that the difference $Q^{\mathrm{SAM}}_{\alpha,\beta} - Q^{\mathrm{MFM}_1}$ is of order $\mathcal{O}(\varepsilon^{-1}h^4)$. In Figure 5.11 we plot its absolute value for the integral $I = \int_0^1 \cos x\, e^{-\frac{i}{\varepsilon}(x+1)^2} dx$. For large values of ε (10^{-1}, 10^{-2}, 10^{-3}) we observe the predicted forth order behavior with respect to h. However, the ε^{-1} behavior is not present for small h. Here the lines almost lie on each other. A reason for this may be that summing up the local quadrature errors yields a highly oscillatory sum. As discussed in [4, §3.3] this may yield a higher order in ε as locally predicted. The oscillations in the lower left corner may be do to machine precision.

In the previous examples we always used 1000 abscissas to create the error plots. The resulting curves show (for large h) an oscillatory behavior, which is more or less good resolved. We can use such a high resolution, because the approximation of one integral is quite fast. However, the solution of the

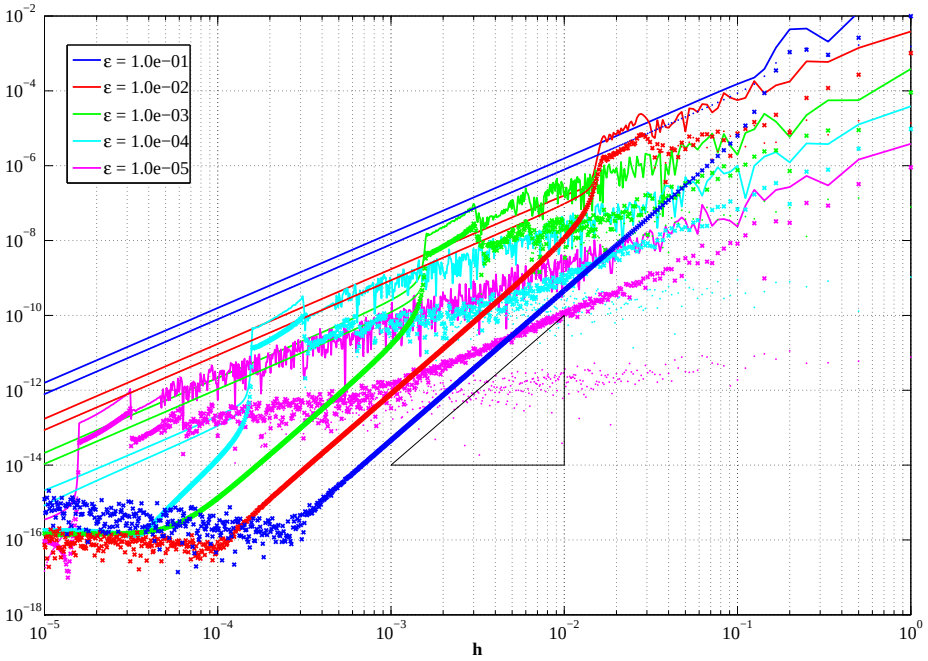

Figure 5.9: Absolute quadrature error of the SAM_2 (solid line), $SSAM_2$ (+) and MFM_1 (dot) for $I = \int_0^1 \cos x \, e^{-\frac{i}{\varepsilon}(x+1)^2} \, dx$.

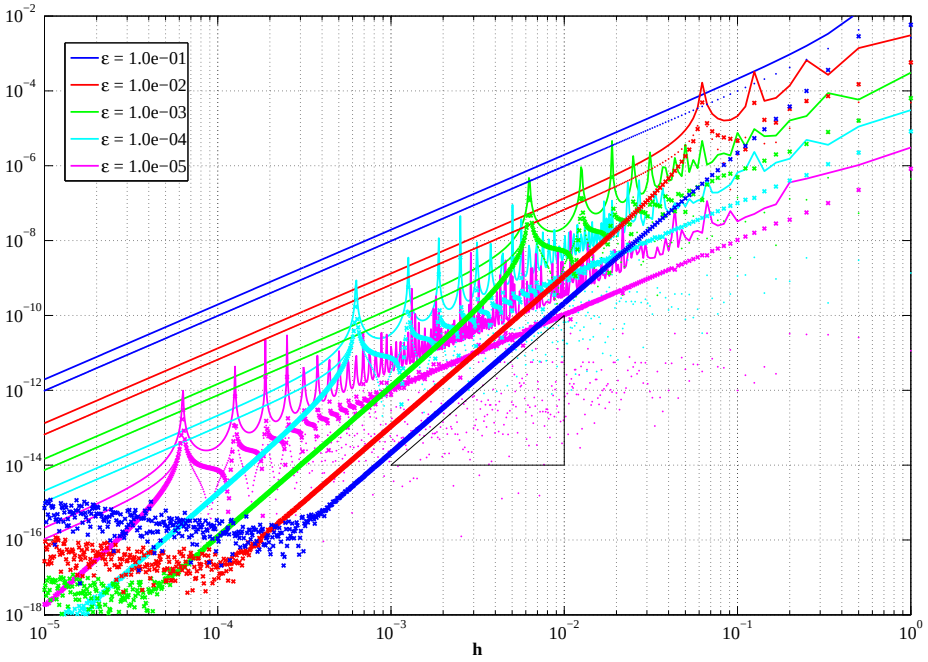

Figure 5.10: Absolute quadrature error of the SAM_2 (solid line), $SSAM_2$ (+) and MFM_1 (dot) for $I = \int_0^1 \log x \, e^{-\frac{i}{\varepsilon}x} \, dx$.

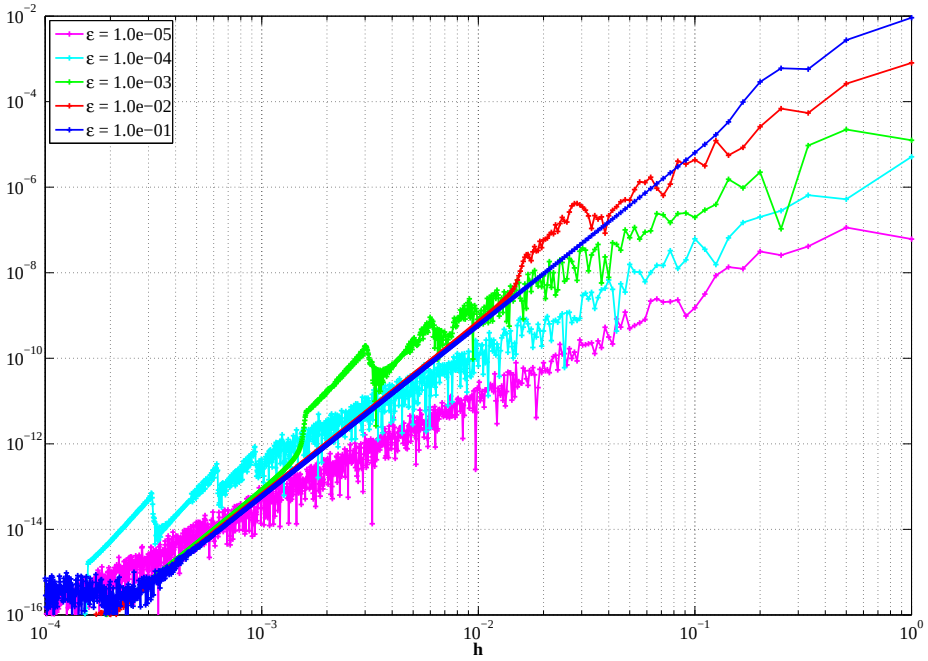

Figure 5.11: Absolute value of $Q^{\mathrm{MFM}_1} - Q^{\mathrm{SAM}_2}_{\alpha,\beta}$ for $I = \int_0^1 \cos x \, e^{-\frac{i}{\varepsilon}(x+1)^2} \, dx$.

numerical examples in chapter 7 is much more involved and hence much slower. Hence, there we shall use only a few grid points (approx. 13) to create the error plots (cf. Figure 7.2– 7.10). The MFM_1 is used by the schemes to approximate the occurring highly oscillatory integrals. To get an idea, how the error plots look like with only a few abscissas, we shall once more plot the examples from Figure 5.7 and Figure 5.8. Now we use only 17 grid points. As for the error plots in § 7, the subinterval lengths are given by $h = (b-a)2^{-n}$ (here $n = 0, \ldots, 16$), which yields an equidistant distribution on the logarithmic axis. The results are plotted in Figure 5.12 for $I = \int_0^1 \cos(x) e^{-\frac{i}{\varepsilon}(x+1)^2} \, dx$ and Figure 5.13 for $I = \int_0^1 \log(x+1) e^{-\frac{i}{\varepsilon}x} \, dx$. Hence we may identify effects in the error plots of the numerical integrators derived in § 6, which probably originate from the used quadrature.

In the end of this section we shall summarize the results. Let us start with the methods, which only uses values of f and no derivatives. These are the classical trapezoidal rule, the SAM_1 from [4], our new SSAM_1 derived in § 5.3 and the MFM_1 discussed in § 5.2. The numerical effort of this four methods is comparable. We have observed that the SAM_1 is only of first order as $h \to 0$, while the other three methods are of second order with respect to h. Furthermore, the TR is not well suited for highly oscillatory problems as we have seen in Figure 5.3. Comparing the error, especially for large $h > \varepsilon$ non of the other three methods is competitive to the MFM_1. The asymptotic accuracy with respect ε of the MFM_1 is to dominant here. Shortly we can write (for the error)

$$\mathrm{MFM}_1 \;\leq\; \mathrm{SSAM}_1 \;\leq\; \mathrm{SAM}_1 \; .$$

Allowing also the use of the first derivatives yields the SAM_2 from [4] and

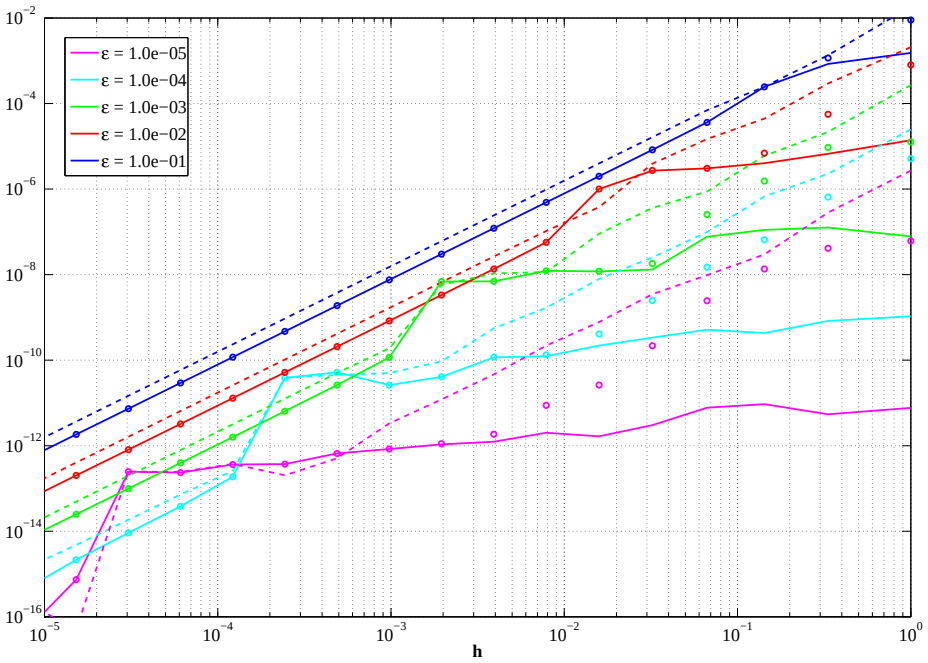

Figure 5.12: Absolute quadrature error of the SAM_2 (dashed lines), MFM_1 (solid lines), and RSAM (circles) for $I = \int_0^1 \cos(x) e^{-\frac{i}{\varepsilon}(x+1)^2}\, dx$.

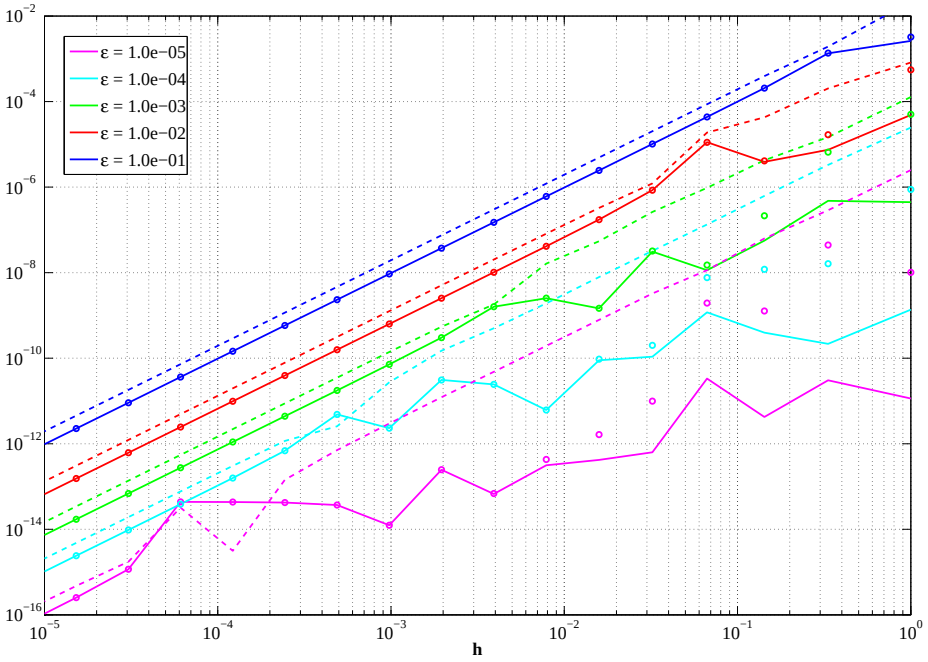

Figure 5.13: Absolute quadrature error of the SAM_2 (dashed lines), MFM_1 (solid lines), and RSAM (circles) for $I = \int_0^1 \log(1+x) e^{-\frac{i}{\varepsilon}x}\, dx$.

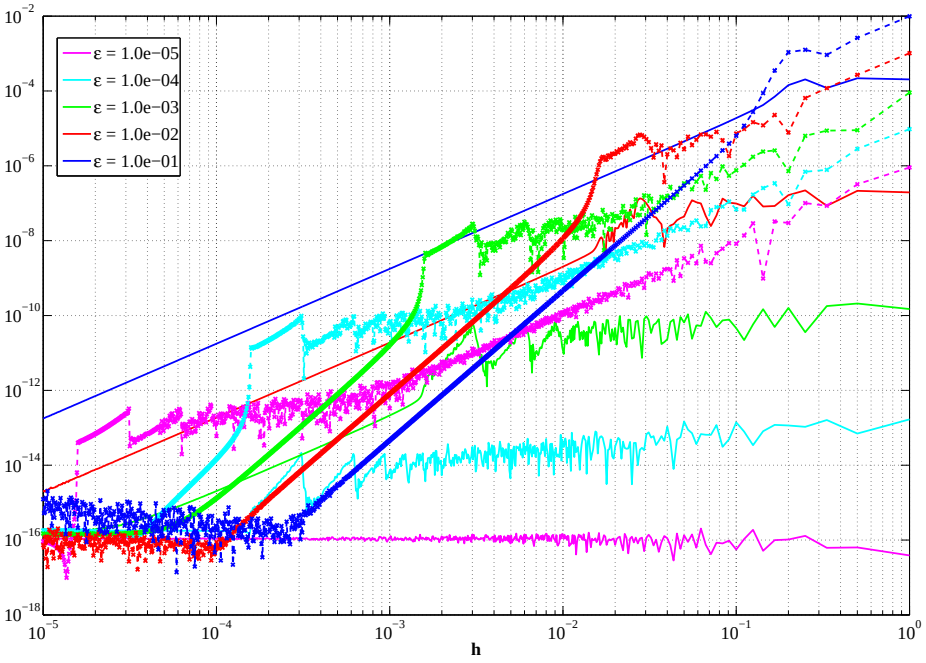

Figure 5.14: Absolute quadrature error of the SSAM_2 (dashed lines), and the hybrid method HQ (solid lines) for $I = \int_0^1 \log(1+x)e^{-\frac{i}{\varepsilon}x}\,dx$.

MSAM, RSAM and SSAM_2 from § 5.3. While the first three methods are of second order as $h \to 0$, the SSAM_2 shows its predicted fourth order behavior. The SAM_2, MSAM, and RSAM are less efficient for $h > \varepsilon$ compared to the MFM_1, which is (surprisingly) the method with the least numerical effort. Shortly we could summarize our experience in the following relations:

$$\mathrm{MFM}_1 \;\leq\; \mathrm{RSAM} \;\leq\; \mathrm{MSAM} \;\leq\; \mathrm{SAM}_2\ .$$

Since the SSAM_2 is of fourth order with respect to h, it is a bit complicate to directly compare it to our favorite method, the MFM_1, which is of second order and use a different set of date. But we could compare them, if the numerical effort of the two methods would be similar. Hence we shall make one "step" of the asymptotic method and than use our MFM_1. The resulting hybrid quadrature (HQ) now use the same data as SSAM_2 and SAM_2. We plot the absolute quadrature error of the HQ and SSAM_2 in Figure 5.14 for the integral $I = \int_0^1 \log(1+x)e^{-\frac{i}{\varepsilon}x}\,dx$. We observe a third order accuracy with respect to ε of the HQ. Hence the HQ is much more efficient for $h > \varepsilon$ than the SSAM_2. The point, where the fourth order behavior of the SSAM_2 beats the second order accuracy of the HQ (i. e. where the two curves cross), rapidly decreases with ε.

Chapter 6

Efficient one–step methods

As motivated in chapter 3 by the two–band $k \cdot p$–model, we are interested in an efficient scheme to approximate the solution of the IVP (3.21)

$$u' \;=\; \tfrac{i}{\varepsilon} L u \,+\, B u\,, \qquad u(x_0) = u_0\,. \tag{6.1}$$

The matrix valued functions $L, B\colon I \times (0,\varepsilon_0) \to \mathbb{C}^{d\times d}$ are C^r–bounded independently of ε. Furthermore $L(x,\varepsilon)$ is real and diagonal for all $(x,\varepsilon) \in I \times (0,\varepsilon_0)$. Here "efficient" means that the numerical method do not have to resolve all oscillations of the solution u in order to yield a good approximation. I.e. that the used discretization grid is (in the best case) independently of ε.

By Proposition 3.3.1 (WKB–type transformation) and Corollary 3.3.4 we know that (6.1) is equivalent to ($r \geq n$)

$$z' \;=\; \varepsilon^n \, E_\varepsilon^* \mathcal{S}_n E_\varepsilon \; z\,, \qquad z(x_0) = z_0\,. \tag{6.2}$$

This ODE has a system matrix of order $\mathcal{O}(\varepsilon^n)$. Hence it is possible to get an approximation error of at most the same order. Since E_Φ^ε is a highly oscillatory matrix function, a naive discretization of this problem will lead to a reduced (or negative) error order with respect to ε.

The first discretization step consists of truncating a series representation (limit of the Picard iteration) for the solution z of (6.2). This is discussed in § 6.1 for a general linear first order initial value problem. Here we derive a (semi discretized) pre–version of our one–step method (cf. p.119f). In the following section § 6.2 we focus on the special case (6.2) and attune the "pre–method" from § 6.1 to it. The result shall be a raw version of our efficient one–step method (see p.127f). Subsection § 6.2.1 is of interest for programming. Here we reformulate the derived coefficients of the one–step method, such that it is more convenient to implement them in the code. In § 6.3 we derive the quadrature for the oscillatory integrals as well as upper bounds for the quadrature defect. These are the missing ingredients, which makes the raw version of § 6.2 a completely discretized one–step method. We merge the results of § 6.2 and § 6.3 in § 6.4 and get an explicit description of our one–step method. It is summarized in the end of the section on page 137ff.

In order to prove convergence of our one–step method we show two things. First we prove in § 6.5 that the matrices and vectors, which determine the method, are bounded independently of ε and the spatial step size. Furthermore the derived bound guarantees stability of the scheme. Secondly, in § 6.6,

we derive an upper bound for the local error. It is of the form $\theta(\varepsilon, h_n)h_n^{\tau+1}$, where $\theta(\varepsilon, h_n) \to 0$ as $\varepsilon \to 0$. How to construct methods with maximum asymptotic order with respect to the small parameter ε is discussed in § 6.6.1. With the results from § 6.5 and § 6.6 we shall prove convergence of the one–step method in 6.7. For this purpose we use standard arguments.

6.1 Picard iteration: truncation error and iterated integrals

Let $I := [a, b]$ be a non–trivial compact interval and let $A\colon I \to \mathbb{C}^{d\times d}$ and $f\colon I \to \mathbb{C}^d$ be continuous. The goal of this section is the derivation of ("pre") one–step methods (OSM) for the linear initial value problem

$$\begin{aligned} y'(x) &= \rho\, A(x)\, y(x) + \lambda f(x)\,, \qquad x \in I \\ y(\xi) &= y_\xi\,. \end{aligned} \tag{6.3}$$

with $\xi \in I$ and $\rho, \lambda \geq 0$. It is given in (6.11).

We use the objects from the following Definition 6.1.1 to write down an exact series representation of y (cf. Lemma 6.1.4).

Definition 6.1.1. *Let $\zeta \in I$ and let $M \in C(I, \mathbb{C}^{d\times d})$. We define a linear map $\mathcal{I}_\zeta\colon C(I, \mathbb{C}^{d\times m}) \to C^1(I, \mathbb{C}^{d\times m})$ by*

$$(\mathcal{I}_\zeta U)(x) := \int_\zeta^x M(s)\, U(s)\, ds\,.$$

We denote the j–times application of $\mathcal{I}_\zeta$ to U by $\mathcal{I}_\zeta^j U$. As usual $\mathcal{I}_\zeta^0$ is the identity operator. For constant $U = \mathrm{Id}$ we use the notation

$$(\mathcal{I}_\zeta^j \,\mathrm{Id})(x) = \mathcal{I}_\zeta^j(x)\,.$$

Remark 6.1.2 (Iterated integrals)**.** *We call $(\mathcal{I}_\zeta^j U)$ the j^{th}* iterated integral. *Due to Definition 6.1.1 it holds for all $r, s \in \mathbb{N}_0$ and all $x \in I$:*

$$(\mathcal{I}_\zeta^r \mathcal{I}_\zeta^s)(x) = \mathcal{I}_\zeta^{r+s}(x)\,.$$

We will frequently use the iterated integrals and therefore we need a priori estimates.

Lemma 6.1.3. *Let the assumption of Definition 6.1.1 hold. Then we get for all $j \in \mathbb{N} \cup \{0\}$ and for all $x \in I$:*

$$\|(\mathcal{I}_\zeta^j U)(x)\| \leq \frac{\|M\|_\infty^j |x-\zeta|^j}{j!}\, \|U\|_\infty\,. \tag{6.4}$$

Proof. By definition it is $(\mathcal{I}_\zeta^0 U)(x) = U(x)$. Hence estimate (6.4) holds for

$j = 0$. We proceed by induction. Let (6.4) hold for $j \in \mathbb{N} \cup \{0\}$.

$$\begin{aligned}
\|(\mathcal{I}_\zeta^{j+1}U)(x)\| &= \left\| \int_\zeta^x M(s)(\mathcal{I}_\zeta^j U)(s)\, ds \right\| \\
&\leq \operatorname{sgn}(x-\zeta) \int_\zeta^x \|M(s)\|\, \|(\mathcal{I}_\zeta^j U)(s)\|\, ds \\
&\leq \operatorname{sgn}(x-\zeta)^{j+1} \|M\|_\infty \int_\zeta^x \frac{(s-\zeta)^j}{j!}\, ds\, \|M\|_\infty^j\, \|U\|_\infty \\
&= \frac{\|M\|_\infty^{j+1} |x-\zeta|^{j+1}}{j+1!}\, \|U\|_\infty .
\end{aligned}$$

□

The unique solution y of the linear IVP (6.3) is given by the limit of the Picard iteration. As we will see soon, the Picard limit y can be represented as a von Neumann series.

Lemma 6.1.4 (Series Representation)**.** *In Definition 6.1.1 let $M = A$. Then the unique solution y of the IVP* (6.3) *is given by*

$$y(x) = \sum_{j=0}^{\infty} \rho^j \big(\mathcal{I}_\xi^j\, (y_\xi + \lambda f_\xi)\big)(x)\,, \tag{6.5}$$

with $f_\xi(x) := \int_\xi^x f(s)\, ds$. Here we interpret y_ξ as a constant function on the interval I.

Proof. Let

$$\begin{aligned}
y_n(x) &:= \sum_{j=0}^{n} \rho^j \big(\mathcal{I}_\xi^j\, (y_\xi + \lambda f_\xi)\big)(x)\,, \\
y_n^\dagger(x) &:= y_n'(x) = \sum_{j=0}^{n} \rho^j \big(\mathcal{I}_\xi^j\, (y_\xi + \lambda f_\xi)\big)'(x)\,.
\end{aligned}$$

Due to Lemma 6.1.3 we get

$$\sum_{j=0}^{\infty} \rho^j \|\mathcal{I}_\xi^j (y_\xi + \lambda f_\xi)\|_\infty \leq e^{\rho\|A\|_\infty |a-b|} (\|y_\xi\| + \lambda \|f_\xi\|_\infty) =: c < \infty$$

and

$$\sum_{j=0}^{\infty} \rho^j \big\|\big(\mathcal{I}_\xi^j (y_\xi + f_\xi)\big)'\big\|_\infty \leq \rho\, \|A\|_\infty\, c + \lambda \|f_\xi\|_\infty < \infty\,.$$

This yields uniform convergence of both sequences (cf. [23] "Konvergenzkriterium von Weierstraß"), i. e. $y_n \to y$ and $y_n^\dagger \to y^\dagger$. Hence y is continuously differentiable with $y' = y^\dagger$ (cf. [49, XIII, §9]). It holds for $n \in \mathbb{N}$

$$y_n^\dagger(x) = \rho A(x) y_{n-1}(x) + \lambda f(x)\,. \tag{6.6}$$

Passing on both sides of (6.6) to the limit finishes the proof. □

From the series representation of y in Lemma 6.1.4 we directly deduce, with the estimate from Lemma 6.1.3,

Corollary 6.1.5. *Let y be the unique solution of the IVP (6.3). Than it holds for all $x \in I$*

$$\|y(x)\| \quad \leq \quad e^{\rho|x-\xi|\|A\|_\infty} \|y_\xi + \lambda f_\xi\| .$$

Hence y continuously depend on the data of the IVP. From the ODE it directly follows that

$$\|y'(x)\| \quad \leq \quad \rho\|A(x)\|\|y(x)\| + \lambda\|f_\xi(x)\| .$$

For our highly oscillatory model problem 2 in § 6.2 this means, that z is C^1–bounded independently of ε, since the function $A(x) = E_\varepsilon^*(x)S(x,\varepsilon)E_\varepsilon(x)$ is at least C^0–bounded independently of ε.

Remark 6.1.6. *Lemma 6.1.4 is a tool that helps us to derive our OSM in § 6.1. But it is also useful to get another point of view of the difficulties (highly oscillatory behavior, exponential growth and decay of the solution) of the second order IVP*

$$\begin{aligned} \varepsilon^2\psi'' + A(x)\psi &= 0 , \\ \psi(\xi) &= \psi_0 , \\ \psi'(\xi) &= \psi_1 . \end{aligned}$$

We can rewrite it as a first order IVP by setting $u_1 = \psi$, $u_2 = \varepsilon\psi'$, which yields

$$u'(x) \;=\; \frac{1}{\varepsilon}\begin{pmatrix} 0 & \mathrm{Id} \\ -A(x) & 0 \end{pmatrix} u , \qquad u(\xi) \;=\; u_0 .$$

For the following consideration the structure of the system matrix is not relevant. Hence we shall generalize the first order system as follows. Let $z \in \mathbb{C}\backslash\{0\}$ and let M be a continuous matrix valued function. By Lemma 6.1.4 the solution of the IVP (on the interval I)

$$u'(x) \;=\; \frac{1}{z}M(x)u , \qquad u(\xi) \;=\; u_0 ,$$

is given by

$$u(x,z) \quad = \quad \sum_{j=0}^{\infty} \left(\mathcal{I}_\xi^j(x)u_0\right) z^{-j} .$$

Hence, for every fixed $x \in I$ each component f_j of the function $f(z) := u(x,z)$ is analytic in z. Further it is obvious that it has an essential singularity at $z = 0$. Now the complex analysis yields an explanation for the difficulties u makes when approaching $z = 0$. Due to the theorem of Casorati–Weierstraß (cf. [64]) we know, that the image of any origin–neighborhood under f_j is dense in $\mathbb{C}$. Hence the function f_j shows a lot of different behaviors depending of the direction one approaches $z = 0$. Consequently the nature of the IVP solution u sensitively depends on the matrix M and on the parameter $z = \varepsilon > 0$. It can (e. g.) easily switch from highly oscillatory to exponential growth depending on M.

Now let $a = x_{n_a} < x_{n_a+1} < \cdots < x_{n_b} = b$ be a grid on $[a,b]$ with $n_a \le 0 \le n_b$ and $x_0 = \xi$ and let y be the unique solution of (6.3). From the above Lemma 6.1.4 we immediately get (with $\xi = x_n$)

$$y(x_{n+1}) \;=\; \sum_{j=0}^{\infty} \rho^j \big(\mathcal{I}^j_{x_n}(y(x_n) \,+\, \lambda f_{x_n})\big)(x_{n+1})\,. \tag{6.7}$$

We can also integrate 'backwards' in x which yields

$$y(x_n) \;=\; \sum_{k=0}^{\infty} \rho^k \big(\mathcal{I}^k_{x_{n+1}}(y(x_{n+1}) \,+\, \lambda f_{x_{n+1}})\big)(x_n)\,. \tag{6.8}$$

A linear combination of both equations is the starting point for our numerical method. Before we proceed with this we introduce some notation.

Notation.

$$\mathbb{I}^j_n \;:=\; \mathcal{I}^j_{x_n}(x_{n+1})\,, \qquad \mathbb{F}^j_n \;:=\; \big(\mathcal{I}^j_{x_n} f_{x_n}\big)(x_{n+1})\,,$$

$$\mathbb{J}^k_n \;:=\; \mathcal{I}^k_{x_{n+1}}(x_n)\,, \qquad \mathbb{G}^k_n \;:=\; (\mathcal{I}^k_{x_{n+1}} f_{x_{n+1}})(x_n)\,.$$

With this notation equations (6.7) and (6.8) read

$$y(x_{n+1}) \;=\; \sum_{j=0}^{\infty} \rho^j \mathbb{I}^j_n \; y(x_n) \,+\, \lambda \sum_{j=0}^{\infty} \rho^j \mathbb{F}^j_n\,, \tag{6.9}$$

$$y(x_n) \;=\; \sum_{k=0}^{\infty} \rho^k \mathbb{J}^k_n \; y(x_{n+1}) \,+\, \lambda \sum_{k=0}^{\infty} \rho^k \mathbb{G}^k_n\,. \tag{6.10}$$

Assumption 6. *Let $\sigma_e, \sigma_i \in [0,1]$ with $\sigma_e + \sigma_i = 1$.*

Due to (6.9), (6.10) it holds

$$\begin{aligned}\left(\sigma_e \,\mathrm{Id} + \sigma_i \sum_{k=0}^{\infty} \rho^k \mathbb{J}^k_n\right) y(x_{n+1}) \;=\; &\left(\sigma_i \,\mathrm{Id} + \sigma_e \sum_{j=0}^{\infty} \rho^j \mathbb{I}^j_n\right) y(x_n) \\ &+ \lambda\sigma_e \sum_{j=0}^{\infty} \rho^j \mathbb{F}^j_n \;-\; \lambda\sigma_i \sum_{k=0}^{\infty} \rho^k \mathbb{G}^k_n\,.\end{aligned}$$

Numerical Method 1. *To get a numerical method we truncate the series at both sides after the same number of summands. By Assumption 6 $\sigma_e + \sigma_i = 1$ and it holds $\mathbb{J}^0_n = \mathbb{I}^0_n = \mathrm{Id}$. This yields the* one–step method

$$\begin{aligned}\left(\mathrm{Id} + \sigma_i \sum_{k=1}^{\tau} \rho^j \mathbb{J}^k_n\right) y_{n+1} \;=\; &\left(\mathrm{Id} + \sigma_e \sum_{j=1}^{\tau} \rho^j \mathbb{I}^j_n\right) y_n \\ &+ \lambda \sum_{k=0}^{\tau} \rho^k \big(\sigma_e \,\mathbb{F}^k_n - \sigma_i \,\mathbb{G}^k_n\big)\end{aligned} \tag{6.11}$$

For the exact solution y of (6.3) *it holds*

$$\begin{aligned}\Big(\mathrm{Id}+\sigma_i\sum_{k=1}^{\tau}\rho^j\,\mathbb{J}_n^k\Big)\,y(x_{n+1}) \;&=\; \Big(\mathrm{Id}+\sigma_e\sum_{j=1}^{\tau}\rho^j\mathbb{I}_n^j\Big)\,y(x_n)\\ &\quad+\lambda\sum_{k=0}^{\tau}\rho^k\big(\sigma_e\,\mathbb{F}_n^k-\sigma_i\,\mathbb{G}_n^k\big)\;+\;\mathrm{err}_{\mathrm{trunc}}^{n,\tau}\end{aligned}$$

with

$$\begin{aligned}\mathrm{err}_{\mathrm{trunc}}^{n,\tau} \;:=\;& \sigma_e\sum_{j=\tau+1}^{\infty}\rho^j\mathbb{I}_n^j\;y(x_n)\;+\;\lambda\sigma_e\sum_{j=\tau+1}^{\infty}\rho^j\mathbb{F}_n^j\\ &-\sigma_i\sum_{k=\tau+1}^{\infty}\rho^k\mathbb{J}_n^k\;y(x_{n+1})\;-\;\lambda\sigma_i\sum_{k=\tau+1}^{\infty}\rho^k\mathbb{G}_n^k\,.\end{aligned}\tag{6.12}$$

For $\sigma_i=0$ we get an explicit and for $\sigma_e=0$ a 'pure' implicit scheme. Due to $\mathbb{J}_n^1=-\mathbb{I}_n^1$ we get for $\tau=1$ and $\sigma_e=\sigma_i=\frac{1}{2}$ a Crank–Nicolson like Scheme.

Lemma 6.1.7. *The local truncation error* (6.12) *of* (6.11) *can be written as*

$$\mathrm{err}_{\mathrm{trunc}}^{n,\tau}\;=\;\rho^{\tau+1}\big(\sigma_e(\mathcal{I}_{x_n}^{\tau+1}y)(x_{n+1})\;-\;\sigma_i(\mathcal{I}_{x_{n+1}}^{\tau+1}y)\big)(x_n)\big)\,.\tag{6.13}$$

Furthermore it holds

$$\|\,\mathrm{err}_{\mathrm{trunc}}^{n,\tau}\,\|\;\le\;\frac{\rho^{\tau+1}\|A\|_\infty^{\tau+1}\,h_n^{\tau+1}}{(\tau+1)!}\;\|y\|_\infty\,.\tag{6.14}$$

with $h_n=x_{n+1}-x_n$.

Proof. In (6.13) we simply replace $y(x)$ by its series representation (6.5). We do the calculation only for one term. The other one is analogously treated. With Remark 6.1.2 we compute:

$$\begin{aligned}&\rho^{\tau+1}\big(\mathcal{I}_{x_{n+1}}^{\tau+1}y\big)(x_n)\\ &=\;\Big(\mathcal{I}_{x_{n+1}}^{\tau+1}\sum_{k=0}^{\infty}\rho^{k+\tau+1}\big(\mathcal{I}_{x_{n+1}}^{k}\,y(x_{n+1})\;+\;\lambda\,\mathcal{I}_{x_{n+1}}^{k}f_{x_{n+1}}\big)\Big)(x_n)\\ &=\;\sum_{k=0}^{\infty}\rho^{k+\tau+1}\mathcal{I}_{x_{n+1}}^{k+\tau+1}(x_n)\,y(x_{n+1})\;+\;\lambda\sum_{k=0}^{\infty}\rho^{k+\tau+1}\big(\mathcal{I}_{x_{n+1}}^{k+\tau+1}f_{x_{n+1}}\big)(x_n)\\ &=\;\sum_{k=\tau+1}^{\infty}\rho^k\mathbb{J}_n^k\;y(x_{n+1})\;+\;\lambda\sum_{k=\tau+1}^{\infty}\rho^k\mathbb{G}_n^k\,.\end{aligned}$$

The error estimate is a consequence of Assumption 6, Lemma 6.1.3 and (6.13). □

It is well known that the Crank–Nicolson scheme has a local error of third order (cf. [28]). In § 6.2 we will see that $\mathbb{I}_n^0$ has a highly oscillatory integrand. Hence the trapezoid rule, which would give us the exact Crank–Nicolson scheme from literature, should not be applied.

The explicit expression for the constant c in the following Lemma is needed in the upcoming section, in order to determine the asymptotic error behavior.

Lemma 6.1.8. *Let $A \in C^1([a,b],\mathbb{C}^{d\times d})$. For the local truncation error of the Crank–Nicolson like scheme* (i. e. $\tau = 1$ and $\sigma_e = \sigma_i = \frac{1}{2}$) *it holds:*

$$\| \operatorname{err}^{n,1}_{\text{trunc}} \| \quad \leq \quad c\,\rho^2 h_n^3\,,$$

with

$$c \quad = \quad \lambda \frac{\|A\|_\infty^2}{4}\,\|f\|_\infty \,+\, \frac{|\rho|\|A\|_\infty^3 + 2\|A\|_\infty\|A'\|_\infty}{4}\,\|y\|_\infty\,.$$

Proof. Due to Lemma 6.1.7 it is

$$\begin{aligned}\tfrac{2}{\rho^2}\ \operatorname{err}^{n,1}_{\text{trunc}} \quad &= \quad \big((\mathcal{I}^2_{x_n} y)(x_{n+1}) \,-\, (\mathcal{I}^2_{x_{n+1}} y)(x_n)\big) \\ &= \quad \int_{x_n}^{x_{n+1}} A(x) \int_{x_n}^{x} A(s)y(s)\,ds \,-\, \int_{x_{n+1}}^{x_n} A(x) \int_{x_{n+1}}^{x} A(s)y(s)\,ds\,.\end{aligned}$$

Using the fundamental theorem of calculus we get:

$$\begin{aligned}&\tfrac{2}{\rho^2}\operatorname{err}^{n,1}_{\text{trunc}} \\ &= \quad \int_{x_n}^{x_{n+1}} A(x_n)\bigg(\int_{x_n}^{x} A(x_n)y(x_n)\,ds + \int_{x_{n+1}}^{x} A(x_n)y(x_n)\,ds\bigg)\,dx \,+ \\ &\qquad \int_{x_n}^{x_{n+1}}\int_{x_n}^{x} A'(s)\,ds\bigg(\int_{x_n}^{x} A(x_n)y(x_n)\,ds + \int_{x_{n+1}}^{x} A(x_n)y(x_n)\,ds\bigg)\,dx \,+ \\ &\qquad \int_{x_n}^{x_{n+1}} A(x)\bigg(\int_{x_n}^{x}\int_{x_n}^{s} (Ay)'(t)\,dt\,ds + \int_{x_{n+1}}^{x}\int_{x_n}^{s} (Ay)'(t)\,dt\,ds\bigg)\,dx\,.\end{aligned}$$

Since the first integral is zero we get

$$\begin{aligned}&\tfrac{2}{|\rho|^2}\|\operatorname{err}^{n,1}_{\text{trunc}}\| \\ &\leq \quad \|A'\|_\infty\|A\|_\infty\|y\|_\infty\bigg|\int_{x_n}^{x_{n+1}}\int_{x_n}^{x} ds\bigg(\int_{x_n}^{x} ds + \int_{x}^{x_{n+1}} ds\bigg)\,dx\bigg| \,+ \\ &\qquad \|A\|_\infty\|(Ay)'\|_\infty\bigg|\int_{x_n}^{x_{n+1}}\bigg(\int_{x_n}^{x}\int_{x_n}^{s} dt\,ds + \int_{x}^{x_{n+1}}\int_{x_n}^{s} dt\,ds\bigg)dx\bigg| \\ &= \quad \|A\|_\infty\big(\|A'\|_\infty\|y\|_\infty \,+\, \|(Ay)'\|_\infty\big)\frac{h_n^3}{2}\,.\end{aligned}$$

To finish the proof we remark that $(Ay)' = A'y \,+\, \rho\,A^2 y \,+\, \lambda A f$. □

6.2 The highly oscillatory case: raw version of the method

While the setting in the previous section § 6.1 is quite general, we now focus on the highly oscillatory problem as mentioned in the introduction. We use the results from § 6.1 and a special strategy for the highly oscillatory integrals to derive a numerical scheme. In this section we do not approximate the iterated integrals. Instead we rewrite them in a favorable way, which is motivated by the quadrature technique from § 5. This procedure results in Lemma 6.2.5 and Lemma 6.2.6 and finally yields the Numerical Scheme 2 (cf. 127).

Now let us specify the oscillatory IVP. Let $[a,b] \subset \mathbb{R}$ be a non–empty bounded interval and let $\Omega := [a,b] \times (0,\varepsilon_1)$ for some $\varepsilon_1 > 0$. Further let the matrix valued function $\Phi : \Omega \to \mathbb{R}^{d\times d} \subset \mathbb{C}^{d\times d}$ be, such that for all $(x,\varepsilon) \in \Omega$:

$$\Phi(x,\varepsilon) \quad = \quad \operatorname{diag}(\varphi_1(x,\varepsilon)\operatorname{Id}_{\nu_1}, \ldots, \varphi_{\nu_q}(x,\varepsilon)\operatorname{Id}_{\nu_q}) \in \mathbb{R}^{d\times d}\,,$$

where $\operatorname{Id}_{\nu_j}$ denotes the identity matrix on $\mathbb{C}^{\nu_j}$. The number $q \in \mathbb{N}$ is assumed to be constant on Ω. Since $\Phi(x,\varepsilon) \in \mathbb{R}^{d\times d}$, it has to hold $\nu_1 + \cdots + \nu_q = d$. We denote the vector of the geometric multiplicities of the eigenvalues by

$$\nu \quad := \quad (\nu_1, \ldots, \nu_q)^T \in \mathbb{N}^q\,.$$

Further we define on Ω the matrix valued function E_ε by

$$E_\varepsilon(x) \quad := \quad \exp\left(\tfrac{i}{\varepsilon}\Phi(x,\varepsilon)\right). \tag{6.15}$$

The matrix $E_\varepsilon(x)$ is unitary for all $(x,\varepsilon) \in \Omega$. Due to the definition of E_ε we call Φ the *phase function*.

Assumption 7. *The eigenvalue functions $\varphi_1, \ldots, \varphi_q$ are C^{s+1}–bounded independently of ε. Further there exists a constant $\delta > 0$, such that for all $(x,\varepsilon) \in \Omega$ and $k \neq j$*

$$|\varphi_k'(x,\varepsilon) - \varphi_j'(x,\varepsilon)| \quad \geq \quad \delta\,.$$

Now we can write down the initial value problem we are interested in.

Model Problem 2. *Let $S\colon \Omega \to \mathbb{C}^{d\times d}$ and $f\colon \Omega \to \mathbb{C}^d$ be C^s–bounded independently of ε and let Assumption 7 hold. Furthermore let E_ε be given by* (6.15)*. The IVP we want to approximate for fixed $\varepsilon \in (0,\varepsilon_1)$ and $x \in [a,b]$ is given by*

$$\begin{aligned} z'(x,\varepsilon) \quad &= \quad \rho\, E_\varepsilon^*(x) S(x,\varepsilon) E_\varepsilon(x)\; z(x,\varepsilon) \;+\; \lambda E_\varepsilon^*(x) f(x,\varepsilon)\,, \\ z(\overline{x},\varepsilon) \quad &= \quad z^0 \in \mathbb{C}^d\,, \end{aligned} \tag{6.16}$$

with $\overline{x} \in [a,b]$ and $\rho, \lambda \geq 0$.

From Corollary 3.3.4 we know that it is always possible to remove the ν–block diagonal part of S with a C^s–bounded linear transformation. Since the ν–block diagonal entries of the system matrix are not highly oscillatory, one can use a standard integrator, like Runge-Kutta methods, to solve this problem. The integrator can use an ε independent grid. In some special cases the integration can be done by hand (cf. p. 18 ff). Hence we assume that $\operatorname{diag}_\nu(S) = 0$.

Assumption 8. *For all $(x,\varepsilon) \in \Omega$ it holds:* $\operatorname{diag}_\nu\left(S(x,\varepsilon)\right) = 0$.

Due to Lemma 3.3.1 the IVP (6.1) can be reformulated, such that it fits in the above setting. Hence we should keep in mind that ρ is a very small constant.

In § 6.3 we derive the quadratures for the highly oscillatory iterated integrals. These quadratures are designed for phase functions which do not have stationary points. Since this has to hold also for the inhomogeneity we make

Assumption 9. *If $f \neq 0$, then we additionally assume*

$$|\varphi_j'(x,\varepsilon)| \quad \geq \quad \delta$$

for all $(x,\varepsilon) \in \Omega$ and all $j = 1, \ldots, q$.

Remark 6.2.1. *There are also quadratures for highly oscillatory integrals with stationary points. If one wants to neglect Assumption 9 one can use for example the method of Olver [61]. For simplicity we restrict ourself to the non–stationary point case.*

Before we continue with numerics let us have a look on the behavior of the solution z as $\varepsilon \to 0$.

Proposition 6.2.2. *Let $\rho(\varepsilon), \lambda(\varepsilon)\colon (0, \varepsilon_1) \to \mathbb{R}^+$ be bounded functions and let $z_\varepsilon = z(\cdot, \varepsilon)$ be the unique solution of (6.16). There exists a constant $c \geq 0$, such that*

$$\|z_\varepsilon - z^0\|_\infty \quad \leq \quad c\,\varepsilon\,.$$

Here we interpret the initial condition z^0 as a constant function.

Proof. This proof is based on the series representation of z_ε from Lemma 6.1.4. Let $\xi := \overline{x}$. It holds for all $(x, \varepsilon) \in \Omega$

$$\begin{aligned} z(x,\varepsilon) - z^0 \quad &= \quad \sum_{j=1}^{\infty} \rho^j \left(\mathcal{I}_\xi^j\ (y_\xi\ +\ \lambda f_\xi)\right)(x) \\ &= \quad \rho\left(\mathcal{I}_\xi^1 \sum_{j=0}^{\infty} \rho^j \mathcal{I}_\xi^j\ (y_\xi\ +\ \lambda f_\xi)\right)(x) \quad = \quad \rho\,(\mathcal{I}_\xi^1 z_\varepsilon)(x)\,. \end{aligned}$$

Using integration by parts we get

$$(\mathcal{I}_\xi^1 z_\varepsilon)(x) \quad = \quad \int_\xi^x (E_\varepsilon^* S E_\varepsilon)(r)\ dr\, z_\varepsilon(x) \ - \ \int_\xi^x \int_\xi^t (E_\varepsilon^* S E_\varepsilon)(r)\ dr\, z_\varepsilon'(t)\ dt\,.$$

Furthermore $\mathrm{diag}_\nu(S) = 0$ and hence we deduce from § 3.1.2 (especially (3.20)), that there exists a constant $c \geq 0$ such that

$$\left\| \int_\xi^t (E_\varepsilon^* S E_\varepsilon)(r)\ dr \right\| \quad \leq \quad c\varepsilon\,,$$

for all $x \in I$. With the triangle inequality we conclude

$$\big\|z(x,\varepsilon) - z^0\big\| \quad \leq \quad \rho\, c\, \varepsilon(\|z(x)\| \ +\ |x - \xi|\, \|z'\|_\infty)\,.$$

Since ρ, λ are bounded we get from Corollary 6.1.5 that z is C^1–bounded independently of ε, which finishes the proof. □

Remark 6.2.3. *The relevance of Proposition 6.2.2 for the upcoming construction of the numerical integrators is as follows. Since the exact solution z_ε of our Model Problem 2 tends to a constant as $\varepsilon \to 0$, the numerical schemes we construct shall have this behavior too. I. e. the convergence error of the integrators are at most of $\mathcal{O}(\varepsilon h_n^\tau)$, for some $\tau \in \mathbb{N}$. Hence in the limit $\varepsilon \to 0$ these schemes yield the exact (constant) solution, even for a fixed spatial grid.*

In § 6.1 we have derived a family of semi discretized one–step methods for linear first order IVP. The only missing part is a suitable quadrature for the iterated integrals, which are highly oscillatory now. Due to the highly oscillatory

integrands we cannot efficiently apply standard quadratures like Trapezoid or Simpson rule (cf. [37]). In this section we do not directly discretize the oscillatory integrals. In Lemma 6.2.5 we prove an exact representation for $\mathcal{I}_\xi^j$, where we set $M = E_\varepsilon^* S E_\varepsilon$ in its Definition 6.1.1. From the derived formula one can get a quadrature by dropping the error terms. Before we continue with Lemma 6.2.5 we need

Definition 6.2.4. *Let $\xi \in [a,b]$.*

(i) *For $F \in C^0([a,b], \mathbb{C}^{d\times d})$ we define*

$$I_\xi[F](x) \;:=\; \int_\xi^x E_\varepsilon^*(s) F(s) E_\varepsilon(s)\, ds\,. \tag{6.17}$$

(ii) *For $g \in C^0([a,b], \mathbb{C}^d)$ we define*

$$I_\xi^v[g](x) \;:=\; \int_\xi^x E_\varepsilon^*(s) g(s)\, ds\,. \tag{6.18}$$

The following considerations are a motivation for Lemma 6.2.5. The first iterated integral $\mathcal{I}_\xi^1$ can be written as

$$\mathcal{I}_\xi^1(x) \;=\; I_\xi[S](x) \;=\; I_\xi[P_1](x) \,+\, I_\xi[S-P_1](x)\,,$$

with an arbitrary matrix function P_1. The basic idea for the quadrature of $\mathcal{I}_\xi^1$ is to choose P_1, such that the integral $I_\xi[P_1]$ can exactly be integrated and such that the remainder is small. As we will see in § 6.3 the function P_1 can be chosen, such that

$$I_\xi[P_1](x) \;=\; E_\varepsilon^*(s) P_1^\diamond(s) E_\varepsilon(s)\big|_{s=\xi}^x\,, \tag{6.19}$$

where the matrix function $P_1^\diamond$ is C^k–bounded independently of ε and can explicitly be computed. We iteratively apply the above idea to the iterated integrals and set

$$S_1 \;:=\; S\,, \qquad C_\xi^1 \;:=\; E_\varepsilon^*(\xi) P_1^\diamond(\xi) E_\varepsilon(\xi)$$

and denote the quadrature error by

$$\mathrm{Err}_\xi^1(x) \;:=\; I_\xi[S_1 - P_1](x)\,.$$

Thus we have

$$\mathcal{I}_\xi^1(x) \;=\; I_\xi[P_1](x) \,+\, \mathrm{Err}_\xi^1(x)\,, \tag{6.20}$$

which yields with (6.19)

$$\begin{aligned}\mathcal{I}_\xi^2(x) \;&=\; \int_\xi^x E_\varepsilon^*(t) S(t) E_\varepsilon(t) \mathcal{I}_\xi^1(t)\, dt \\ &=\; I_\xi[SP_1^\diamond](x) \,-\, I_\xi[S](x) C_\xi^1 \,+\, \big(\mathcal{I}_\xi^1\, \mathrm{Err}_\xi^1\big)(x)\,.\end{aligned}$$

We set

$$S_2 \;:=\; SP_1^\diamond$$

and choose a suitable P_2 with

$$I_\xi[P_2](x) \;=\; E_\varepsilon^*(s)P_2^\diamond(s)E_\varepsilon(s)\big|_{s=\xi}^x\,,$$

which yields

$$\begin{aligned}\mathcal{I}_\xi^2(x) \;=\;& I_\xi[P_2](x) \,-\, I_\xi[P_1](x)C_\xi^1\\ &+\, \mathrm{err}_\xi^2(x) \,-\, \mathrm{Err}_\xi^1(x)C_\xi^1 \,+\, \big(\mathcal{I}_\xi^1\,\mathrm{Err}_\xi^1\,\big)(x)\,.\end{aligned}$$

Obviously it is possible to continue with this procedure.

Lemma 6.2.5. *Let $P_1,\dots,P_\tau\in C^0(\Omega,\mathbb{C}^{d\times d})$, such that for $j=1,\dots,\tau$*

$$I_\xi[P_j](x) \;=\; E_\varepsilon^*(s)P_j^\diamond(s)E_\varepsilon(s)\big|_{s=\xi}^x\,,$$

where $P_1^\diamond,\dots,P_\tau^\diamond$ are C^0–bounded independently of ε. Further we set $\mathcal{C}_\xi^0:=\mathrm{Id}$, $P_0^\diamond:=\mathrm{Id}$ and inductively define for $j=1,\dots,\tau$:

$$\begin{aligned}\mathcal{C}_\xi^j \;&:=\; -\sum_{l=1}^j P_l^\diamond(\xi)\,\mathcal{C}_\xi^{j-l}\,,\\ C_\xi^j \;&:=\; E_\varepsilon^*(\xi)\mathcal{C}_\xi^j E_\varepsilon(\xi)\,,\\ S_j(x) \;&:=\; S(x)P_{j-1}^\diamond(x)\,,\\ \mathrm{Err}_\xi^j(x) \;&:=\; I_\xi[S_j-P_j](x)\,.\end{aligned}$$

With this definitions it holds for $j=1,\dots,\tau$:

$$\mathcal{I}_\xi^j(x) \;=\; \sum_{k=1}^j I_\xi[P_k](x)\,C_\xi^{j-k} \,+\, \sum_{k=1}^j\sum_{l=1}^k \big(\mathcal{I}_\xi^{j-k}\,\mathrm{Err}_\xi^l\,\big)(x)\,C_\xi^{k-l}\,. \quad (6.21)$$

Proof. We prove (6.21) by induction. For $j=1$ the right–hand side of (6.21) reads

$$I_\xi[P_1](x)C_\xi^0 \,+\, \big(\mathcal{I}_\xi^0\,\mathrm{Err}_\xi^1\,\big)(x)C_\xi^0\,,$$

which is equal to $\mathcal{I}_\xi^1(x)$ due to (6.20) and $C_\xi^0=\mathrm{Id}$. To simplify the following computations we set

$$\Sigma_j(x) \;:=\; \sum_{k=1}^j\sum_{l=1}^k \big(\mathcal{I}_\xi^{j-k}\,\mathrm{Err}_\xi^l\,\big)(x)\,C_\xi^{k-l}\,.$$

Hence we get

$$\begin{aligned}
\mathcal{I}_\xi^{j+1}(x) &= \int_\xi^x \big(E_\varepsilon^* S\, E_\varepsilon\big)(t)\, \mathcal{I}_\xi^j(t)\, dt \\
&= \sum_{k=1}^{j} \int_\xi^x \big(E_\varepsilon^* S E_\varepsilon\, I_\xi[P_k]\big)(t)\, dt\, C_\xi^{j-k} + \big(\mathcal{I}_\xi \Sigma_j\big)(x) \\
&= \sum_{k=1}^{j} \big(I_\xi[S_{k+1}](x) - I_\xi[S](x)\,(E_\varepsilon^* P_k^\diamond E_\varepsilon)(\xi)\big)\, C_\xi^{j-k} + \big(\mathcal{I}_\xi \Sigma_j\big)(x) \\
&= \sum_{k=2}^{j+1} I_\xi[S_k](x)\, C_\xi^{j+1-k} + I_\xi[S_1](x)\, C_\xi^j + \big(\mathcal{I}_\xi \Sigma_j\big)(x) \\
&= \sum_{k=1}^{j+1} I_\xi[P_k](x)\, C_\xi^{j+1-k} + \sum_{k=1}^{j+1} \mathrm{Err}_\xi^k(x)\, C_\xi^{j+1-k} + \big(\mathcal{I}_\xi \Sigma_j\big)(x) \\
&= \sum_{k=1}^{j+1} I_\xi[P_k](x)\, C_\xi^{j+1-k} + \sum_{k=1}^{j+1} \sum_{l=1}^{k} \big(\mathcal{I}_\xi^{j+1-k}\, \mathrm{Err}_\xi^l\big)(x)\, C_\xi^{k-l}\,.
\end{aligned}$$

□

With a similar ansatz we can rephrase $\mathcal{I}_\xi^j f_\xi$.

Lemma 6.2.6. *Let the assumptions and definitions of Lemma 6.2.5 hold and let $u_0, \dots, u_\tau \in C^0(\Omega, \mathbb{C}^d)$, such that for $j = 0, \dots, \tau$*

$$I_\xi^v[u_j](x) = E_\varepsilon^*(s) u_j^\diamond(s)\big|_{s=\xi}^x\,,$$

where $u_0^\diamond, \dots, u_\tau^\diamond$ are C^0–bounded independently of ε. We set for $j = 0, \dots, \tau$

$$\begin{aligned}
c_\xi^j &:= E_\varepsilon^*(\xi) u_j^\diamond(\xi)\,, \\
s_j(x) &:= S(x) u_{j-1}^\diamond(x)\,, \quad (j \geq 1) \\
\mathrm{err}_\xi^j(x) &:= I_\xi^v[s_j - u_j](x)
\end{aligned}$$

and $s_0 := f_\xi$. With this definitions it holds for $j = 0, \dots, \tau$

$$\begin{aligned}
(\mathcal{I}_\xi^j f_\xi)(x) &= I_\xi^v[u_j](x) - \sum_{k=1}^{j} \mathcal{I}_\xi^k(x)\, c_\xi^{j-k} \\
&\quad + \sum_{k=0}^{j} \big(\mathcal{I}_\xi^{j-k}\, \mathrm{err}_\xi^k\big)(x)\,.
\end{aligned} \tag{6.22}$$

Here $\sum_{k=1}^0$ is the empty sum and hence zero.

Proof. For $j = 0$ equation (6.22) holds by definition. To simplify the following computation we set

$$\Sigma_k(x) := \sum_{k=0}^{j} \big(\mathcal{I}_\xi^{j-k}\, \mathrm{err}_\xi^k\big)(x)$$

and proceed by induction.

$$\begin{aligned}
(\mathcal{I}_\xi^{j+1} f_\xi)(x) &= \Big(\mathcal{I}_\xi\Big(I_\xi^v[u_j] - \sum_{k=1}^{j} \mathcal{I}_\xi^k\, c_\xi^{j-k} + \Sigma_k\Big)\Big)(x) \\
&= I_\xi^v[Su_j^\diamond](x) - \sum_{k=0}^{j} \mathcal{I}_\xi^{k+1}(x)\, c_\xi^{j-k} + \big(\mathcal{I}_\xi \Sigma_j\big)(x) \\
&= I_\xi^v[s_{j+1}](x) - \sum_{k=1}^{j+1} \mathcal{I}_\xi^{k}(x)\, c_\xi^{j+1-k} + \big(\mathcal{I}_\xi \Sigma_j\big)(x) \\
&= I_\xi^v[u_{j+1}](x) - \sum_{k=1}^{j+1} \mathcal{I}_\xi^{k}(x)\, c_\xi^{j+1-k} + \sum_{k=0}^{j+1} \big(\mathcal{I}_\xi^{j+1-k}\, \mathrm{err}_\xi^k\big)(x)\,.
\end{aligned}$$

□

Now we combine the results of Lemma 6.2.5 and Lemma 6.2.6 with (6.11) from § 6.1 to write down a one–step method for our Model Problem 2.

Notation.

$$\mathcal{Q}_\xi^j(x) := \sum_{k=1}^{j} I_\xi[P_k](x) C_\xi^{j-k}\,, \tag{6.23}$$

$$\mathcal{E}_\xi^j(x) := \sum_{k=1}^{j}\sum_{l=1}^{k} \big(\mathcal{I}_\xi^{j-k}\, \mathrm{Err}_\xi^l\big)(x)\, C_\xi^{k-l}\,. \tag{6.24}$$

From Lemma 6.2.5 we immediately get

$$\mathcal{I}_\xi^j(x) = \mathcal{Q}_\xi^j(x) + \mathcal{E}_\xi^j(x)\,.$$

Numerical Method 2. *Let $n \in \{n_a, \ldots, n_b - 1\}$ and let the assumptions and definitions of Lemma 6.2.5 and Lemma 6.2.6 hold. Let z be the unique solution of* (6.16)*. We set*

$$A_n := \sum_{j=1}^{\tau} \rho^j\, \mathcal{Q}_{x_n}^j(x_{n+1}) \tag{6.25}$$

$$B_n := \sum_{j=1}^{\tau} \rho^j\, \mathcal{Q}_{x_{n+1}}^j(x_n) \tag{6.26}$$

$$v^n := \sum_{j=0}^{\tau} \rho^j\, I_{x_n}^v[u_k](x_{n+1}) - \sum_{j=1}^{\tau} \rho^j \sum_{k=1}^{j} \mathcal{Q}_{x_n}^k(x_{n+1})\, c_{x_n}^{j-k} \tag{6.27}$$

$$w^n := \sum_{j=0}^{\tau} \rho^j\, I_{x_{n+1}}^v[u_k](x_n) - \sum_{j=1}^{\tau} \rho^j \sum_{k=1}^{j} \mathcal{Q}_{x_{n+1}}^k(x_n)\, c_{x_{n+1}}^{j-k} \tag{6.28}$$

and

$$\begin{aligned}
\mathrm{err}_\xi(x) := {} & \sum_{j=1}^{\tau} \rho^j \mathcal{E}_\xi^j(x) \left(z(\xi) - \lambda \sum_{k=0}^{\tau-j} \rho^k\, c_\xi^k\right) \\
& + \lambda \sum_{j=0}^{\tau} \rho^j \sum_{k=0}^{j} \big(\mathcal{I}_\xi^{j-k}\, \mathrm{err}_\xi^k\big)(x)\,.
\end{aligned} \tag{6.29}$$

Than the one step method *(OSM) is given by*

$$\begin{aligned} \big(\mathrm{Id} + \sigma_i B_n\big) z^{n+1} + \lambda\sigma_i\, w^n &= \big(\mathrm{Id} + \sigma_e A_n\big) z^n + \lambda\sigma_e\, v^n\,, \\ z^0 &= z(x_0)\,. \end{aligned} \tag{6.30}$$

For the exact solution z it holds

$$\begin{aligned} \big(\mathrm{Id} + \sigma_i B_n\big) z(x_{n+1}) + \lambda\sigma_i\, w_n &= \big(\mathrm{Id} + \sigma_e A_n\big) z(x_n) + \lambda\sigma_e\, v_n \\ &\quad + \mathrm{err}^n\,, \end{aligned} \tag{6.31}$$

with

$$\mathrm{err}^n := \mathrm{err}^{n,\tau}_{\mathrm{trunc}} + \mathrm{err}^n_{\mathrm{int}} \tag{6.32}$$

and

$$\mathrm{err}^n_{\mathrm{int}} := \sigma_e\, \mathrm{err}_{x_n}(x_{n+1}) - \sigma_i\, \mathrm{err}_{x_{n+1}}(x_n)\,. \tag{6.33}$$

It remains to specify the functions $P_1, \dots, P_\tau$ and $u_0, \dots, u_\tau$, which is part of § 6.3. Let us summarize the properties that these functions are supposed to have:

Remark 6.2.7. *For the functions $P_1, \dots, P_\tau$ and $u_0, \dots, u_\tau$ it holds*

(i) *there exists matrix valued functions $P_1^\diamond, \dots, P_\tau^\diamond$ C^0–bounded independently of ε, such that*

$$I_\xi[P_j](x) = E_\varepsilon^*(s) P_j^\diamond(s) E_\varepsilon(s)\big|_{s=\xi}^x$$

(ii) *there exists vector valued functions $u_0^\diamond, \dots, u_\tau^\diamond$ C^0–bounded independently of ε, such that*

$$I_\xi^v[u_j](x) = E_\varepsilon^*(s) u_j^\diamond(s)\big|_{s=\xi}^x$$

(iii) *the error terms $\mathrm{Err}_\xi^1, \dots, \mathrm{Err}_\xi^\tau$ and $\mathrm{err}_\xi^0, \dots, \mathrm{err}_\xi^\tau$ are "small".*

Since the error integrals $\mathrm{Err}_\xi^j(x)$ are of the form

$$I_\xi[\tilde{S} - \tilde{P}](x) = \int_\xi^x (E_\varepsilon^* \tilde{S} - \tilde{P} E_\varepsilon)(s)\, ds = \int_\xi^x E_\Phi^\varepsilon(s) \odot (\tilde{S} - \tilde{P})(s)\, ds$$

and since they do not contain a matrix multiplication and all off_ν–diagonal entries can be written as (cf. § 3.1.2)

$$I[f] := \int_\alpha^\beta f(x)\, e^{-\frac{i}{\varepsilon}\varphi(x)}\, dx\,,$$

we deduce that it is enough to find an appropriate approximation rule for the above scalar valued integral. Then the derived quadrature can be applied one–to–one to the integral matrix.

6.2.1 Reprocessing of the raw version

We use this section to reprocess the quantities A_n, B_n, v^n, w^n, such that it is more convenient to implement the numerical scheme. The following Remark 6.2.8 helps us to do manipulations of multiple sums.

Remark 6.2.8. *Let $(G,+)$ be a commutative monoid and let $\theta,\tau \in \mathbb{Z}$ with $\theta \leq \tau$. Furthermore let $\{A_{i,j} \in G \mid j = \theta,\dots,i\,, i = \theta,\dots,\tau\}$. Then*

$$\sum_{i=\theta}^{\tau}\sum_{j=\theta}^{i} A_{i,j} \;=\; \sum_{j=\theta}^{\tau}\sum_{i=j}^{\tau} A_{i,j}\,. \tag{6.34}$$

Proof. On the left–hand side of (6.34) one sum up the row sums and on the right–hand side one adds together the column sums of the following scheme:

$$\begin{array}{llll} A_{\tau,\theta} & A_{\tau,\theta+1} & A_{\tau,\theta+2} & \cdots \quad A_{\tau,\tau} \\ \vdots & \vdots & \vdots & \iddots \\ A_{\theta+2,\theta} & A_{\theta+2,\theta+1} & A_{\theta+2,\theta+2} & \\ A_{\theta+1,\theta} & A_{\theta+1,\theta+1} & & \\ A_{\theta,\theta} & & & \end{array}\,.$$

□

To simplify the upcoming computations we us the following

Notation.

$$Q_n^k \;:=\; I_{x_n}[P_k](x_{n+1})\,, \qquad q_n^k \;:=\; I^v_{x_n}[u_k](x_{n+1})\,.$$

Now let us start with the reprocessing.

Lemma 6.2.9. *Let A_n, B_n be the matrices defined by* (6.25) *and* (6.26)*. It holds for all $n \in \{n_a,\dots,n_b-1\}$*

$$A_n \;=\; \sum_{k=1}^{\tau} \rho^k\, Q_n^k \sum_{l=0}^{\tau-k} \rho^l\, C^l_{x_n}\,, \tag{6.35}$$

$$B_n \;=\; -\sum_{k=1}^{\tau} \rho^k\, Q_n^k \sum_{l=0}^{\tau-k} \rho^l\, C^l_{x_{n+1}}\,. \tag{6.36}$$

Proof. Let

$$X(\xi,x) \;:=\; \sum_{j=1}^{\tau} \rho^j\, \mathcal{Q}^j_\xi(x)\,.$$

Due to definition of A_n, B_n we get

$$A_n \;=\; X(x_n,x_{n+1}) \quad\text{and}\quad B_n \;=\; X(x_{n+1},x_n)\,.$$

We further compute, using Remark 6.2.8,

$$\begin{aligned} X(\xi,x) &= \sum_{j=1}^{\tau} \rho^j \sum_{k=1}^{j} I_\xi[P_k](x) C_\xi^{j-k} \\ &= \sum_{k=1}^{\tau} \rho^k \sum_{j=k}^{\tau} \rho^{j-k} I_\xi[P_k](x)\, C_\xi^{j-k} \\ &= \sum_{k=1}^{\tau} \rho^k\, I_\xi[P_k](x) \sum_{j=0}^{\tau-k} \rho^j C_\xi^j \,. \end{aligned}$$

Replacing ξ by x_n and x by x_{n+1} yields (6.35). To get (6.36) we set $\xi = x_{n+1}$, $x = x_n$, and use $I_{x_{n+1}}[P_k](x_n) = -I_{x_n}[P_k](x_{n+1})$. □

Remark 6.2.10. *Lemma 6.2.9 yields a representation of A_n, B_n which allows to compute the quantities (each) with a single loop. The* `pseudo code` *reads:*

$\Sigma = 0;\ \Gamma = 0;$
`for` $k = \tau : -1 : 1$ `do`
  $\Gamma = \Gamma + \rho^{\tau-k} C_{x_n}^{\tau-k};$
  $\Sigma = \Sigma + \rho^k Q_n^k \Gamma;$
`end`
$A_n = \Sigma;$

If we replace x_n by x_{n+1} and replace the last equation by $B_n = -\Sigma$, then we get the `pseudo code` *for B_n.*

Similar computations can be done for the vectors v^n and w^n.

Lemma 6.2.11. *Let v^n, w^n be the vectors defined by* (6.27) *and* (6.28)*. It holds for all $n \in \{n_a, \ldots, n_b - 1\}$:*

$$v^n = \sum_{j=0}^{\tau} \rho^j q_n^k - \sum_{l=1}^{\tau} \rho^l Q_n^l \sum_{k=0}^{\tau-l} \rho^k C_{x_n}^k \sum_{j=0}^{\tau-k-l} \rho^j c_{x_n}^j \,, \tag{6.37}$$

$$w^n = -\sum_{j=0}^{\tau} \rho^j q_n^k + \sum_{l=1}^{\tau} \rho^l Q_n^l \sum_{k=0}^{\tau-l} \rho^k C_{x_{n+1}}^k \sum_{j=0}^{\tau-k-l} \rho^j c_{x_{n+1}}^j \,. \tag{6.38}$$

Proof. Let us define

$$X(\xi,x) := \sum_{j=1}^{\tau} \rho^j \sum_{k=1}^{j} \mathcal{Q}_\xi^k(x)\, c_\xi^{j-k}$$

and

$$Q_\xi^k(x) := I_\xi[P_k](x)\,.$$

Inserting the definition (6.23) of $\mathcal{Q}_\xi^k(x)$ yields

$$X(\xi,x) = \sum_{j=1}^{\tau} \rho^j \sum_{k=1}^{j} \left(\sum_{l=1}^{k} Q_\xi^l(x)\, C_\xi^{k-l} \right) c_\xi^{j-k} \,.$$

We reformulate X by repeated application of Remark 6.2.8 and index shifting:

$$\begin{aligned} X(\xi,x) &= \sum_{l=1}^{\tau} \rho^l Q_\xi^l(x) \sum_{k=l}^{\tau} \rho^{k-l} C_\xi^{k-l} \sum_{j=k}^{\tau} \rho^{j-k} c_\xi^{j-k} \\ &= \sum_{l=1}^{\tau} \rho^l Q_\xi^l(x) \sum_{k=l}^{\tau} \rho^{k-l} C_\xi^{k-l} \sum_{j=0}^{\tau-k} \rho^{j} c_\xi^{j} \\ &= \sum_{l=1}^{\tau} \rho^l Q_\xi^l(x) \sum_{k=0}^{\tau-l} \rho^{k} C_\xi^{k} \sum_{j=0}^{\tau-k-l} \rho^{j} c_\xi^{j} \,. \end{aligned}$$

To complete the proof we have to remark that

$$\begin{aligned} v^n &= \sum_{j=0}^{\tau} \rho^j \, q_n^k \, - \, X(x_n, x_{n+1}) \,, \\ w^n &= - \sum_{j=0}^{\tau} \rho^j \, q_n^k \, - \, X(x_{n+1}, x_n) \,, \end{aligned}$$

and $Q_\xi(x)^l = -Q_x(\xi)^l$ hold. □

We use the recurrence relation from the following Lemma 6.2.12 to write down a pseudo code for the computation of v^n and w^n.

Lemma 6.2.12. *For $l = 1, \ldots, \tau$ let*

$$\gamma_\xi^l := \sum_{k=0}^{\tau-l} \rho^k C_{x_{n+1}}^k \sum_{j=0}^{\tau-k-l} \rho^j c_{x_{n+1}}^j \,.$$

It holds for $l = 1, \ldots, \tau - 1$

$$\gamma_\xi^l = \gamma_\xi^{l+1} + \rho^{\tau-l} \sum_{k=0}^{\tau-l} C_\xi^k \, c_\xi^{\tau-l-k} \,.$$

Proof. Let $l \in \{1, \ldots, \tau-1\}$. It holds

$$\begin{aligned} \gamma_\xi^l &= \sum_{k=0}^{\tau-l} \rho^k C_\xi^k \sum_{j=0}^{\tau-k-l} \rho^j c_\xi^j \\ &= \sum_{k=0}^{\tau-l-1} \rho^k C_\xi^k \sum_{j=0}^{\tau-k-l} \rho^j c_\xi^j + \rho^{\tau-l} C_\xi^{\tau-l} \rho^0 c_\xi^0 \\ &= \sum_{k=0}^{\tau-(l+1)} \rho^k C_\xi^k \sum_{j=0}^{\tau-k-(l+1)} \rho^j c_\xi^j + \rho^{\tau-l} C_\xi^{\tau-l} \rho^0 c_\xi^0 \\ &\quad + \sum_{k=0}^{\tau-l-1} \rho^k C_\xi^k \rho^{\tau-k-l} c_\xi^{\tau-k-l} \\ &= \gamma_\xi^{l+1} + \rho^{\tau-l} \sum_{k=0}^{\tau-l} C_\xi^k \, c_\xi^{\tau-k-l} \,. \end{aligned}$$

□

Remark 6.2.13. *To compute v^n and w^n one can use the following* `pseudo code` *based on Lemma 6.2.11 and Lemma 6.2.12:*

$\Sigma = 0;\ \gamma = 0;$
`for` $l = \tau : -1 : 1$
  $\sigma = 0;$
  `for` $k = 0 : \tau - l$
    $\sigma = \sigma + C^k_{x_n} c^{\tau-l-k}_{x_n};$
  `end`
  $\gamma = \gamma + \rho^{\tau-l}\sigma;$
  $\Sigma = \Sigma + \rho^l \left(q^l_n - Q^l_n\gamma\right);$
`end`
$v^n = \Sigma;$

To get the `pseudo code` *for w^n we have to replace x_n by x_{n+1} and replace the last equation in the outer loop by $\Sigma = \Sigma - \rho^l \left(q^l_n - Q^l_n\gamma\right)$.*

6.3 A quadrature for the highly oscillatory iterated integrals

It remains to write down the quadratures for the matrix and vector valued integrals

$$\begin{aligned} I_\alpha[F](\beta) &= \int_\alpha^\beta E^*_\varepsilon(x)F(x)E_\varepsilon(x)\,dx = \int_\alpha^\beta E^\varepsilon_\Phi(x) \odot F(x)\,dx\,, \\ I^v_\alpha[g](\beta) &= \int_\alpha^\beta E^*_\varepsilon(x)g(x)\,dx \end{aligned}$$

from § 6.2. Since $\operatorname{diag}_\nu(E^\varepsilon_\Phi) = \operatorname{diag}(\mathbb{1}_{\nu_1}, \ldots, \mathbb{1}_{\nu_q})$ is independently of ε and Φ, we can use standard Hermite interpolation to derive a quadrature rule for the ν–diagonal elements of $I_\alpha[F]$. For the highly oscillatory ν–off diagonal elements we use Proposition 5.2.1.

Proposition 6.3.1 (OSM quadrature)**.** *Let $F, \Phi\colon \Omega \to \mathbb{C}^{d\times d}$ be, such that F, Φ' are C^s–bounded independently of ε and such that for all $(x,\varepsilon) \in \Omega$ it holds*

(i) $\Phi(x,\varepsilon) = \operatorname{diag}(\varphi_1(x,\varepsilon)\operatorname{Id}_{\nu_1}, \ldots, \varphi_q(x,\varepsilon)\operatorname{Id}_{\nu_q}) \in \mathbb{R}^{d\times d}$,

(ii) $\forall\ k \neq l\colon\ |\varphi'_k(x,\varepsilon) - \varphi'_l(x,\varepsilon)| \geq \delta$.

Further let $\xi_1, \ldots, \xi_\kappa \in J$ be support abscissas with corresponding multiplicities $1 \leq m_1, \ldots, m_\kappa \leq s+1$, such that there are indices j_α, j_β with

$$\xi_{j_\alpha} = \alpha \quad \text{and} \quad \xi_{j_\beta} = \beta\,.$$

We define (cf. § 3.1 for definition of D_Φ)

$$M(x,\varepsilon) := D_\Phi(x,\varepsilon) + \operatorname{diag}(\mathbb{1}_{\nu_1}, \ldots, \mathbb{1}_{\nu_q})\,x\,.$$

There exists one and only one matrix valued function

$$P(x,\varepsilon) := M'(x,\varepsilon) \odot \sum_{j=0}^{m-1} K_j(\varepsilon) \odot M(x,\varepsilon)^{\odot j}$$

with $m = \sum_{j=1}^{\kappa} m_j$ *and* $K_0(\varepsilon), \dots, K_{m-1}(\varepsilon) \in \mathbb{C}^{d\times d}$, *such that*

$$P^{(k)}(\xi_j, \varepsilon) \quad = \quad F^{(k)}(\xi_j, \varepsilon) \tag{6.39}$$

for $k = 0, \dots, m_j - 1$ *and* $j = 1, \dots, \kappa$. *If* $s \geq m$, *then there are constants* $c_d, c, \gamma \geq 0$ *(independently of* ε*), such that the quadrature*

$$\begin{aligned} & I_\alpha[P](\beta) \\ & := \quad i\varepsilon\, E_\Phi^\varepsilon(x) \odot \sum_{j=0}^{m-1} \left(\sum_{l=j}^{m-1} K_l(\varepsilon)\, \tfrac{l!}{j!}(-i\varepsilon)^{l-j} \right) \odot D_\Phi(x,\varepsilon)^{\odot j} \Big|_{x=\alpha}^{\beta} \\ & \qquad + \sum_{j=0}^{m-1} \operatorname{diag}_\nu(K_j(\varepsilon))\, \frac{x^{j+1}}{j+1} \Big|_{t=\alpha}^{\beta} \end{aligned} \tag{6.40}$$

induced by P *yields the error estimate*

$$\|I_\alpha[F](\beta) - I_\alpha[P](\beta)\| \quad \leq \quad c_d\, \theta(\varepsilon, h)\, |\alpha - \beta|\, h^m\,. \tag{6.41}$$

The term $\theta(\varepsilon, h)$ *is given by*

$$\theta(\varepsilon, h) \quad := \quad \max\left(\frac{\|\operatorname{diag}_\nu(F^{(m)})\|_\infty}{m!}\,,\; c \min\left(1,\, \gamma \frac{\varepsilon^{\mu+1}}{h^{\mu+1}}\right)\right),$$

with

$$\mu := \min(m_{j_\alpha}, m_{j_\beta}) \quad \textit{and} \quad h := \max\left(|\xi_\kappa - \alpha|, |\xi_1 - \beta|\right).$$

The constants $c, \gamma \geq 0$ *depend on* δ, $\|\varphi\|_{C^{m+1}(J)}$ *and* $\|f\|_{C^m(J)}$, *but not on* ξ. *Furthermore they tend to infinity as* $\delta \to 0$. c_d *only depend on the space dimension* d.

Proof. Let $I[F]_{kr} := \big(I_\alpha[F](\beta)\big)_{(k,r)}$ bet the $(k,r)^{\text{th}}$ matrix element of the integral matrix, i. e.

$$I[F]_{kr} \quad = \quad \int_\alpha^\beta F_{kl}(x,\varepsilon) e^{-\frac{i}{\varepsilon}(\Phi_{kk}(x,\varepsilon) - \Phi_{rr}(x,\varepsilon))}\, dx\,. \tag{6.42}$$

By definition there exist $\tilde{k}, \tilde{r} \in \{1, \dots, q\}$, such that $\varphi_{\tilde{k}} = \Phi_{kk}$ and $\varphi_{\tilde{r}} = \Phi_{rr}$. For a ν–off diagonal element it is $\tilde{k} \neq \tilde{r}$ and hence $\varphi := \varphi_{\tilde{k}} - \varphi_{\tilde{r}} \neq 0$. Due to assumption (ii) it is $|\varphi'(x,\varepsilon)| \geq \delta > 0$ for all $(x,\varepsilon) \in \Omega$. If we set $f := F_{kr}$, then f, φ fulfill the assumptions of Proposition 5.2.1. Hence we get a unique function P_{kr} with

$$I[P]_{kr} \quad := \quad i\varepsilon\, e^{-\frac{i}{\varepsilon}(\varphi(x,\varepsilon))} \sum_{j=0}^{m-1} \left(\sum_{l=j}^{m-1} K_l(\varepsilon)_{kr} \frac{l!}{j!}(-i\varepsilon)^{l-j} \right) \varphi(x,\varepsilon)^j \Big|_{x=\alpha}^{\beta},$$

which is the kr^{th} element of (6.40). The estimate from Proposition 5.2.1 yields

$$|I[F]_{kr} - I[P]_{kr}| \quad \leq \quad c_{kr}\, |\alpha - \beta|\, h^m \min\left(1,\, \gamma_{kr} \left(\frac{\varepsilon}{h}\right)^{\mu+1}\right),$$

with $\mu := \min(m_{j_\alpha}, m_{j_\beta})$ and $h := \max\big(|\xi_\kappa - \alpha|, |\xi_1 - \beta|\big)$. For the ν–diagonal elements we set $c_{kr} = \gamma_{kr} = 0$ and define $c := \sup_{kr} c_{kr}$ and $\gamma := \sup_{kr} \gamma_{kr}$. Hence

$$|I[F]_{kr} - I[P]_{kr}| \quad \leq \quad c\,|\alpha - \beta|\,h^m \min\left(1,\, \gamma\left(\frac{\varepsilon}{h}\right)^{\mu+1}\right).$$

Since $\|\cdot\|_{\sup}$ is a norm on $\mathbb{C}^{d\times d}$, there exists a constant $\widehat{c} \geq 1$, such that

$$\frac{1}{\widehat{c}}\,\|A\| \quad \leq \quad \|A\|_{\sup} \quad \leq \quad \widehat{c}\,\|A\|$$

holds for all $A \in \mathbb{C}^{d\times d}$. If $I[F]_{kr}$ is a ν–diagonal element, then it is $\Phi_{kk} = \Phi_{ll}$ (cf. (6.42)) and we obtain

$$I[F]_{kr} \quad = \quad \int_\alpha^\beta F_{kr}(x,\varepsilon)\,dx\,.$$

The kr^{th}–component of the matrix P is the uniquely determined Hermite interpolation polynomial corresponding to F_{kr} and (6.39) (cf. [68]). Since the kr^{th}–component of (6.40) is nothing but

$$I[P]_{kr} \quad := \quad \int_\alpha^\beta P_{kr}(x,\varepsilon)\,dx\,,$$

we deduce with Corollary 5.2.6

$$\begin{aligned} |I[F]_{kr} - I[P]_{kr}| \quad &\leq \quad \int_\alpha^\beta |F_{kr}(x,\varepsilon) - P_{kr}(x,\varepsilon)|\,dx \\ &\leq \quad |\alpha-\beta|\,h^m\,\frac{\|F_{kr}^{(m)}\|_\infty}{m!} \\ &\leq \quad \widehat{c}\,|\alpha-\beta|\,h^m\,\frac{\|\operatorname{diag}_\nu(F^{(m)})\|_\infty}{m!}\,. \end{aligned}$$

Hence we get

$$\begin{aligned} &\|I_\alpha[F-P](\beta)\| \quad \leq \quad \widehat{c}\,\|I_\alpha[F-P]\|_{\sup} \\ &\leq \quad \widehat{c}^{\,2}\,|\alpha-\beta|\,h^m\ \max\left(\frac{\|\operatorname{diag}_\nu(F^{(m)})\|_\infty}{m!},\, \frac{c}{\widehat{c}}\,\min\left(1,\,\gamma\left(\frac{\varepsilon}{h}\right)^{\mu+1}\right)\right). \end{aligned}$$

□

Remark 6.3.2. *Let the assumptions of Corollary 6.3.1 hold. If* $\operatorname{diag}_\nu(F)$ *is componentwise a polynomial of order* $m-1$*, then we get an error estimate which is of order* $\mathcal{O}(\varepsilon^{\mu+1})$*. To be more precise we have*

$$\|I_\alpha[F-P](\beta)\| \quad \leq \quad c\,|\alpha-\beta|\,h^m\,\min\left(1,\,\gamma\left(\frac{\varepsilon}{h}\right)^{\mu+1}\right). \tag{6.43}$$

Remark 6.3.3. *Since* $\operatorname{diag}_\nu(F) = \operatorname{diag}_\nu(F) \odot E_\Phi^\varepsilon$ *holds for all* $F \in \mathbb{C}^{d\times d}$*, we can write*

$$I_\alpha[P](\beta) \quad = \quad E_\varepsilon^*(x)P^\diamond(x)E_\varepsilon(x)\big|_{x=\alpha}^\beta\,, \tag{6.44}$$

with

$$\begin{aligned} P^{\diamond}(x) &:= i\varepsilon \sum_{k=0}^{m-1} \left(\sum_{l=k}^{m-1} K_l(\varepsilon)\, \tfrac{l!}{k!}(-i\varepsilon)^{l-k} \right) \odot D_\Phi(x,\varepsilon)^{\odot k} \\ &\quad + \sum_{l=0}^{m-1} \operatorname{diag}_\nu(K_l(\varepsilon))\, \frac{x^{l+1}}{l+1}\,. \end{aligned} \tag{6.45}$$

The quadrature for the vector valued integral $I_\alpha^v[g](\beta)$ is much easier to derive. In each component we can directly apply Proposition 5.2.1.

Corollary 6.3.4. *Let $g\colon \Omega \to \mathbb{C}^d$ and $\Phi\colon \Omega \to \mathbb{C}^{d\times d}$, such that g, Φ' are C^s–bounded independently of ε and such that for all $(x,\varepsilon) \in \Omega$ it holds*

(i) $\Phi(x,\varepsilon) = \operatorname{diag}(\varphi_1(x,\varepsilon)\operatorname{Id}_{\nu_1}, \dots, \varphi_q(x,\varepsilon)\operatorname{Id}_{\nu_q}) \in \mathbb{R}^{d\times d}$

(ii) $\forall\ k = 1,\dots,q\colon\ |\varphi_k'(x,\varepsilon)| \geq \delta.$

Further let $\xi_1, \dots, \xi_\kappa \in J$ be support abscissas with corresponding multiplicities $1 \leq m_1, \dots, m_\kappa \leq s+1$, such that $\xi_1 < \dots < \xi_\kappa$ and such that there are indices j_α, j_β with

$$\xi_{j_\alpha} = \alpha \quad \text{and} \quad \xi_{j_\beta} = \beta\,.$$

Then there exists one and only one function

$$u(x,\varepsilon) := \Phi'(x,\varepsilon) \sum_{j=0}^{m-1} \Phi(x,\varepsilon)^j\, c_j(\varepsilon)\,, \tag{6.46}$$

with $m := \sum_{j=1}^{\kappa} m_j$ and $c_0, \dots, c_{m-1} \in \mathbb{C}^d$, such that

$$u^{(k)}(\xi_j) = g^{(k)}(\xi_j) \quad \text{for} \quad k = 0, \dots, m_j - 1\,, \quad j = 1, \dots, \kappa\,. \tag{6.47}$$

It holds

$$I_\alpha^v[u](\beta) = i\varepsilon\, e^{-\frac{i}{\varepsilon}\Phi(x,\varepsilon)} \sum_{k=0}^{m-1} \Phi(x,\varepsilon)^k \left(\sum_{l=k}^{m-1} c_l(\varepsilon) \frac{l!}{k!}(-i\varepsilon)^{l-k} \right) \Bigg|_{x=\alpha}^{\beta} \tag{6.48}$$

If $s \geq m$, then the quadrature induced by u yields the error estimate

$$|I_\alpha^v[g-u](\beta)| \leq c\,|\alpha-\beta|\,h^m \min\left(1,\, \gamma\left(\frac{\varepsilon}{h}\right)^{\mu+1}\right), \tag{6.49}$$

with

$$\mu := \min(m_{j_\alpha}, m_{j_\beta}) \quad \text{and} \quad h := \max\left(|\xi_\kappa - \alpha|, |\xi_1 - \beta|\right).$$

The constant $c, \gamma > 0$ depend on δ, $\|\Phi\|_{C^{m+1}(J)}$ and $\|g\|_{C^m(J)}$, but not on ξ. Furthermore they tend to infinity as $\delta \to 0$.

Remark 6.3.5. *Let u be given by* (6.46)*. From* (6.48) *we deduce that*

$$I_\alpha^v[u](\beta) = E_\varepsilon^*(s) u^{\diamond}(s)\Big|_{s=\alpha}^{\beta} \tag{6.50}$$

with

$$u^{\diamond}(x) := i\varepsilon \sum_{k=0}^{m-1} \Phi(x,\varepsilon)^k \left(\sum_{l=k}^{m-1} c_l(\varepsilon) \frac{l!}{k!}(-i\varepsilon)^{l-k} \right).$$

Remark 6.3.6. *Let $\epsilon \in \mathbb{C}$ and let u, g be the functions from the previous Corollary 6.3.4. Due to the linearity of the integral and the interpolation problem* (6.47) *it obviously holds*

$$|I_\alpha^v[\epsilon g - \epsilon u](\beta)| \quad \leq \quad |\epsilon| \, \tfrac{c}{\delta^{2m+1}} \, |\alpha - \beta| \, \min_{k=0,\mu} \left(\tfrac{\|\varphi'\|_\infty^{m-k}}{(m-k)!} \, h^{m-k} \, \varepsilon^k \right) .$$

6.4 The one–step method

This section contains the essence of the discretization steps from § 6.1, § 6.2, and § 6.3, which is our one–step method (OSM).

Let $a = x_{n_a} < x_{n_a+1} \cdots < x_{n,b}$ be a grid on $[a, b]$, with $n_a \leq 0 \leq n_b$ and let $x_0 = \xi$. We use the quadrature rules from § 6.3 to approximate the highly oscillatory integrals. For this purpose we have to define the support abscissas $\xi_1^n < \cdots < \xi_\kappa^n$ used for the quadrature. We only want to use elements of the interval $I = [a, b]$. Hence we have to distinguish between subintervals $[x_n, x_{n+1}]$ in the "proper" interior of I and such subintervals which are "close" to the boundary points a, b. To construct the support abscissas we use the following rule.

Support abscissas. *Let $\kappa^\circ, \kappa^a, \kappa^b \in \mathbb{N}$ and let*

$$\begin{array}{rcccl} & \iota_1^\circ & < & \ldots & < \; \iota_{\kappa^\circ}^\circ \in \mathbb{R} \, , \\ 0 = & \iota_1^a & < & \ldots & < \; \iota_{\kappa^a}^a \, , \\ & \iota_1^b & < & \ldots & < \; \iota_{\kappa^b}^b = 1 \, . \end{array}$$

For $n \in \{n_a, \ldots, n_b - 1\}$ we define $h_n := x_{n+1} - x_n$ and set $\kappa = \kappa^\circ$, if

$$\forall \; j = 1, \ldots, \kappa^\circ : \qquad \xi_j^n \;:=\; x_n \, + \, \iota_j^\circ \, h_n \;\in\; I \, .$$

Else we set

$$\diamondsuit \;:=\; \left\{ \begin{array}{l} a \, , \; x_n \leq \frac{a+b}{2} \\ b \, , \; x_n > \frac{a+b}{2} \end{array} \right. , \qquad \kappa \;:=\; \kappa^\diamondsuit$$

and define

$$\xi_j^n \quad := \quad x_n \, + \, \iota_j^\diamondsuit \, h_n \, , \qquad j = 1, \ldots, \kappa \, .$$

Remark 6.4.1. *We implicitly assume that the support abscissas constructed above are elements of the interval I. This can always be ensured by a refinement of the grid, if necessary. In the sequel (in the majority of cases) we drop the marks $\circ, a, b$ and simply write*

$$\xi_j^n \quad = \quad x_n \, + \, \iota_j \, h_n \, , \qquad j = 1, \ldots, \kappa \, .$$

But we should keep in mind that κ and also $\iota_1, \ldots, \iota_\kappa$ depend on n. By definition the distances between the support abscissas $\xi_1^n, \ldots, \xi_\kappa^n$ tends to zero as $h_n \to 0$.

As we have seen during the discussion of the oscillatory integrals in § 5.1, it is exceedingly useful to incorporate the boundary points x_n, x_{n+1} in the support abscissas.

Assumption 10. *We assume that there exists $j_\alpha, j_\beta \in \{1,\dots,\kappa\}$ with*

$$\iota_{j_\alpha} \;=\; 0\,, \qquad \iota_{j_\beta} \;=\; 1\,.$$

That is $\xi^n_{j_\alpha} = x_n$ and $\xi^n_{j_\beta} = x_{n+1}$.

The only missing ingredients to use Proposition 6.3.1 and Corollary 6.3.4 are the multiplicities corresponding to the support abscissas $\xi^n_1,\dots,\xi^n_\kappa$.

Remark 6.4.2. *Since we have a hierarchy of at maximum $\tau+1$ (iterated) integrals, we fix $\tau+1$ sets of multiplicities and denote them by*

$$m_{j,1}\,,\dots,m_{j,\kappa}\,, \qquad j=0,\dots,\tau\,.$$

Here again, we implicitly define different sets of multiplicities corresponding to the three types of support abscissas. Furthermore we set for $k=0,\dots,\tau$:

$$\mu_k \;:=\; \min\big(m_{k,j_\alpha},\, m_{k,j_\beta}\big) \quad \textit{and} \quad |m_k| \;:=\; \sum_{j=1}^{\tau} m_{k,j}\,.$$

The multiplicities are integers and hence it holds $\mu_k \geq 1$ for $k=0,\dots,\tau$.

Now we have defined all quantities needed for the description of the OSM.

Remark 6.4.3. *We shortly write (τ,κ,ι,m) for the set of parameters that determine the OSM, i. e. τ and*

$$\begin{array}{ccc}
\kappa^a & & \\
\iota^a_1, & \dots & ,\iota^a_{\kappa^a} \\
m^a_{0,1}, & \dots & ,m^a_{0,\kappa^a} \\
\vdots & \vdots & \vdots \\
m^a_{\tau,1}, & \dots & ,m^a_{\tau,\kappa^a}
\end{array}
\qquad
\begin{array}{ccc}
\kappa^\circ & & \\
\iota^\circ_1, & \dots & ,\iota^\circ_{\kappa^\circ} \\
m^\circ_{0,1}, & \dots & ,m^\circ_{0,\kappa^\circ} \\
\vdots & \vdots & \vdots \\
m^\circ_{\tau,1}, & \dots & ,m^\circ_{\tau,\kappa^\circ}
\end{array}
\qquad
\begin{array}{ccc}
\kappa^b & & \\
\iota^b_1, & \dots & ,\iota^b_{\kappa^b} \\
m^b_{0,1}, & \dots & ,m^b_{0,\kappa^b} \\
\vdots & \vdots & \vdots \\
m^b_{\tau,1}, & \dots & ,m^b_{\tau,\kappa^b}
\end{array}$$

Numerical Method 3. *The set of vectors $z^{n_a},\dots,z^{n_b}\in\mathbb{C}^d$ is called a solution of the OSM (τ,κ,ι,m), if and only if $z^0 = z(\xi)$ and for all $n\in\{n_a,\dots,n_b-1\}$*

$$\big(\mathrm{Id} + \sigma_i B_n\big) z^{n+1} \;+\; \lambda\sigma_i\, w^n \;\;=\;\; \big(\mathrm{Id} + \sigma_e A_n\big) z^n \;+\; \lambda\sigma_e\, v^n\,. \qquad (6.51)$$

The matrices A_n, B_n and the vectors w^n, v^n are constructed as follows:

(i) *Set $S_1 := S$ and $\mathcal{C}^0_{n,\alpha} = \mathcal{C}^0_{n,\beta} = C^0_{n,\alpha} = C^0_{n,\beta} = \mathrm{Id}$.*
`For` $j=1,\dots,\tau$ `do`

(a) *Compute the unique solution*

$$P_j(x) \;\;=\;\; M'(x,\varepsilon) \odot \sum_{l=0}^{|m_j|-1} K_{j,l}(\varepsilon) \odot M(x,\varepsilon)^{\odot l}$$

of the generalized Hermite interpolation problem

$$P_j^{(k)}(\xi^n_l) \;=\; S_j^{(k)}(\xi^n_l)\,, \qquad k=0,\dots,m_{j,l}-1,\;\; l=1,\dots,\kappa\,.$$

(b) *Compute* $P_j^\diamond$ *by Remark 6.3.3, i. e.*

$$\begin{aligned} P_j^\diamond(x) &= i\varepsilon \sum_{k=0}^{|m_j|-1} \left(\sum_{l=k}^{|m_j|-1} K_{j,l}(\varepsilon)\, \tfrac{l!}{k!}(-i\varepsilon)^{l-k} \right) \odot D_\Phi(x,\varepsilon)^{\odot k} \\ &\quad + \sum_{l=0}^{|m_j|-1} \mathrm{diag}_\nu(K_{j,l}(\varepsilon))\, \frac{x^{l+1}}{l+1}\,. \end{aligned}$$

(c) *Compute* $Q_n^j = I_{x_n}[P_j](x_{n+1})$ *by* (6.40)*, i. e. (cf. Remark 6.3.3)*

$$Q_n^j = E_\Phi^\varepsilon(x_{n+1}) \odot P_j^\diamond(x_{n+1}) - E_\Phi^\varepsilon(x_n) \odot P_j^\diamond(x_n)\,.$$

(d) *Set*

$$\mathcal{C}_{n,\alpha}^j := -\sum_{l=1}^{j} P_l^\diamond(x_n)\, \mathcal{C}_{n,\alpha}^{j-l}\,, \qquad C_{n,\alpha}^j := E_\varepsilon^*(x_n) \mathcal{C}_{n,\alpha}^j E_\varepsilon(x_n)\,,$$

$$\mathcal{C}_{n,\beta}^j := -\sum_{l=1}^{j} P_l^\diamond(x_{n+1})\, \mathcal{C}_{n,\beta}^{j-l}\,, \qquad C_{n,\beta}^j := E_\varepsilon^*(x_{n+1}) \mathcal{C}_{n,\beta}^j E_\varepsilon(x_{n+1})\,.$$

(e) *Set* $S_{j+1} = SP_j^\diamond$.

(f) *Continue with* (a).

`end`

(ii) *Set* $s_0 = f_\xi$.
`For` $j = 0, \ldots, \tau$ `do`

(a) *Compute the unique solution*

$$u_j(x,\varepsilon) = \Phi'(x,\varepsilon) \sum_{l=0}^{|m_j|-1} \Phi(x,\varepsilon)^l\, c_{j,l}(\varepsilon)\,,$$

of the generalized Hermite interpolation problem

$$u_j^{(k)}(\xi_l) = s_j^{(k)}(\xi_l) \quad \textit{for} \quad k = 0, \ldots, m_{j,l} - 1\,, \quad l = 1, \ldots, \kappa\,.$$

(b) *Compute* $u_j^\diamond$ *by Remark 6.3.5, i. e.*

$$u_j^\diamond(x) = i\varepsilon \sum_{k=0}^{|m_j|-1} \Phi(x,\varepsilon)^k \left(\sum_{l=k}^{|m_j|-1} c_{j,l}(\varepsilon) \frac{l!}{k!}(-i\varepsilon)^{l-k} \right).$$

(c) *Compute* $q_n^j = I_{x_n}^v[u_j](x_{n+1})$ *by* (6.50)*, i. e.*

$$q_n^j = E_\varepsilon^*(x_{n+1}) u_j^\diamond(x_{n+1}) - E_\varepsilon^*(x_n) u_j^\diamond(x_n)\,.$$

(d) *Set*

$$c_{n,\alpha}^j = E_\varepsilon^*(x_n) u^\diamond(x_n)\,, \qquad c_{n,\beta}^j = E_\varepsilon^*(x_{n+1}) u^\diamond(x_{n+1})\,.$$

(e) *Set* $s_{j+1} = Su_j^\diamond$.

(f) *Continue with* (a).

`end`

(iii) *Compute* A_n, B_n *by* (6.35), (6.36), *i. e.*

$$A_n = \sum_{k=1}^{\tau} \rho^k Q_n^k \sum_{l=0}^{\tau-k} \rho^l C_{n,\alpha}^l , \tag{6.52}$$

$$B_n = -\sum_{k=1}^{\tau} \rho^k Q_n^k \sum_{l=0}^{\tau-k} \rho^l C_{n,\beta}^l . \tag{6.53}$$

(iv) *Compute* v^n, w^n *by* (6.37), (6.38), *i. e.*

$$v^n = \sum_{j=0}^{\tau} \rho^j q_n^k - \sum_{l=1}^{\tau} \rho^l Q_n^l \sum_{k=0}^{\tau-l} \rho^k C_{n,\alpha}^k \sum_{j=0}^{\tau-k-l} \rho^j c_{n,\alpha}^j ,$$

$$w^n = -\sum_{j=0}^{\tau} \rho^j q_n^k + \sum_{l=1}^{\tau} \rho^l Q_n^l \sum_{k=0}^{\tau-l} \rho^k C_{n,\beta}^k \sum_{j=0}^{\tau-k-l} \rho^j c_{n,\beta}^j .$$

6.5 Boundedness of the coefficients

As we will see in the proof of Lemma 6.7.1, a sufficient criteria for existence and uniqueness of a OSM solutions is that the matrices

$$\mathrm{Id} + \sigma_i B_n \quad \text{and} \quad \mathrm{Id} + \sigma_e A_n$$

are regular for all $n \in \{n_a, \dots, n_b - 1\}$. This is the case, if

$$\|A_n\| < 1 \quad \text{and} \quad \|B_n\| < 1$$

holds for all $n \in \{n_a, \dots, n_b - 1\}$ (cf. [68, p.188]). In the sequel we derive estimates for the norm of the matrices, which enables us to ensure the above condition. Furthermore we need these estimates to prove convergence of the numerical scheme.

For the following discussion we fix one $n \in \{n_a, \dots, n_b - 1\}$ and only discuss the estimate for the matrix A_n. The argumentation for B_n is exactly the same. From (6.52) we get

$$A_n = \sum_{k=1}^{\tau} \rho^k Q_n^k \sum_{l=0}^{\tau-k} \rho^l C_{n,\alpha}^l . \tag{6.54}$$

To construct the matrices Q_n^k and $C_{n,\alpha}^l$ we use the Hermite interpolation based quadratures from § 6.3. Here the essential quantities are the matrix valued coefficients $K_{j,l}(\varepsilon)$, $l = 0, \dots, |m_j| - 1$, $j = 1, \dots, \tau$, which determine the matrix valued functions $P_1, \dots, P_\tau$ and $P_1^\diamond, \dots, P_\tau^\diamond$.

To prove convergence of the OSM one has to refine the discretization more and more. Hence the distances between the supporting abscissas of the interpolation problems are getting smaller and smaller. As the following example illustrates it is not self–evident that the coefficients of the interpolation polynomial, which corresponds to $K_{j,l}$ in our numerical scheme, stay bounded.

Example 6.5.1. *Let $f(x) = |x|$ and let $\xi_1 = -h$, $\xi_2 = 0$, $\xi_3 = h$. Hence the polynomial $p(x) = \frac{x^2}{h}$ is the unique solution of the interpolation problem*

$$p(\xi_j) \;=\; f(\xi_j)\,, \qquad j = 1, 2, 3\,.$$

Obviously, the coefficient of the leading order tends to ∞ as $h \to 0$.

Is the function f sufficiently smooth, which is not the case in the previous example, we can prove that the coefficients are bounded independently of the support abscissas. This is the content of the following Lemma 6.5.2.

Lemma 6.5.2. *Let $m_1, \ldots, m_\kappa \in \mathbb{N}$ with $m = \sum_{j=1}^{\kappa} m_j$ and let the function $f\colon [\alpha, \beta] \times (0, \varepsilon_0) \to \mathbb{C}$ be C^m–bounded independently of ε. For $\xi \in [\alpha, \beta]^\kappa$ with $\xi_1 < \xi_2 < \cdots < \xi_\kappa$ we further denote by*

$$p(x, \varepsilon, \xi) \;:=\; \sum_{j=0}^{m-1} c_j(\varepsilon, \xi)\, x^j$$

the unique Hermite interpolation polynomial of degree $m-1$ with[1]

$$p^{(k)}(\xi_j, \varepsilon, \xi) \;=\; f^{(k)}(\xi_j, \varepsilon)\,, \qquad k = 0, \ldots, m_j - 1\,, \quad j = 1, \ldots, \kappa\,.$$

There exists a constant $c > 0$, independently of ε and ξ, such that

$$|c_j(\varepsilon, \xi)| \;\leq\; c\,, \quad j = 0, \ldots, m-1\,.$$

Proof. Let us fix one vector $\xi \in [\alpha, \beta]^\kappa$ with $\xi_1 < \cdots < \xi_\kappa$ and one $\varepsilon \in (0, \varepsilon_0)$. Furthermore we set $J' := [\xi_1, \xi_\kappa]$. Due to Lemma 5.2.5 (set $r = m-1$) there exists a $\zeta_1^{m-1} \in J'$ and a function $\zeta^{m-1}(x)\colon J' \to J'$, such that for all $x \in J'$ it holds

$$(m-1)!\, c_{m-1}(\xi, \varepsilon) \;=\; f^{(m-1)}(x, \varepsilon) \,-\, f^{(m)}(\zeta^{m-1}(x), \varepsilon)\,(x - \zeta_1^{m-1})\,.$$

We choose $x = \zeta_1^{m-1}$ which yields

$$|c_{m-1}(\xi, \varepsilon)| \;\leq\; \tfrac{1}{(m-1)!} \sup_{\varepsilon \in (0, \varepsilon_0)} \|f^{(m)}(\cdot, \varepsilon)\|_{C([\alpha,\beta])} \;=:\; c\,.$$

Since f is C^m–bounded independently of ε, the constant c is finite. Hence $c_{m-1}(\xi, \varepsilon)$ is bounded.

We continue by induction. Let $c_{m-1}, \ldots, c_{j+1}$ be bounded. We set $r = j$ and again Lemma 5.2.5 yields

$$\begin{aligned} &j!\, c_j(\xi, \varepsilon) \\ &\quad= \; f^{(j)}(x, \varepsilon) \,-\, \sum_{l=j+1}^{m-1} c_l(\xi, \varepsilon) \tfrac{l!}{(l-j)!} x^{l-j} \,-\, \tfrac{f^{(m)}(\zeta^j(x), \varepsilon)}{(m-j)!} \prod_{l=1}^{m-j} (x - \zeta_l^j)\,. \end{aligned}$$

Since $\zeta_l^j \in J' \subset [\alpha, \beta]$ and due to the assumptions the right–hand side is bounded independently of ε and ξ. Hence the same holds for $c_j(\xi, \varepsilon)$. □

[1] Here, of course, we differentiate the functions p, f with respect to the spatial variable x.

The matrix valued functions P_j from our Numerical Scheme 3 (p.137f) are determined by a generalized Hermite interpolation problem. Hence we have to extend the previous Lemma 6.5.2.

Lemma 6.5.3. *Let $m_1, \dots, m_\kappa \in \mathbb{N}$ with $m = \sum_{j=1}^{\kappa} m_j$ and let the functions $f, \varphi' \colon [\alpha, \beta] \times (0, \varepsilon_0) \to \mathbb{C}$ be C^m–bounded independently of ε. Furthermore let $|\varphi'| > 0$. For $\xi \in [\alpha, \beta]^\kappa$ with $\xi_1 < \xi_2 < \cdots < \xi_\kappa$ we denote by*

$$p(x, \varepsilon, \xi) \;:=\; \varphi'(x, \varepsilon) \sum_{j=0}^{m-1} c_j(\varepsilon, \xi)\, \varphi(x, \varepsilon)^j$$

the unique function p with

$$p^{(k)}(\xi_j, \varepsilon, \xi) \;=\; f^{(k)}(\xi_j, \varepsilon)\,, \qquad k = 0, \dots, m_j - 1\,, \quad j = 1, \dots, \kappa\,.$$

There exists a constant $c > 0$, independently of ε and ξ, such that

$$|c_j(\varepsilon, \xi)| \;\leq\; c\,, \quad j = 0, \dots, m-1\,.$$

Proof. Since $|\varphi'| > 0$ for all $(x, \varepsilon) \in [\alpha, \beta] \times (0, \varepsilon_0)$, the function $\varphi(\cdot, \varepsilon)$ is invertible for all $\varepsilon \in (0, \varepsilon_0)$. Let

$$g(x, \varepsilon) \;:=\; \frac{f(\varphi^{-1}(x, \varepsilon), \varepsilon)}{\varphi'(\varphi^{-1}(x, \varepsilon), \varepsilon)} \quad \text{and} \quad \pi(x, \varepsilon, \xi) \;:=\; \frac{p(\varphi^{-1}(x, \varepsilon), \varepsilon, \xi)}{\varphi'(\varphi^{-1}(x, \varepsilon), \varepsilon)}\,.$$

Due to Lemma 5.2.4 the polynomial π solves the Hermite interpolation problem

$$\pi^{(k)}(\varphi(\xi_j, \varepsilon), \varepsilon, \xi) = g^{(k)}(\varphi(\xi_j, \varepsilon), \varepsilon)\,, \qquad k = 0, \dots, m_j - 1\,, \quad j = 1, \dots, \kappa\,.$$

Furthermore the degree of π is $m - 1$. By assumptions g is C^m–bounded independently of ε and hence we can apply Lemma 6.5.2 to π. □

As we have seen above, it is important that the function we interpolate has a certain regularity.

Assumption 11. *The matrix valued functions $S, \Phi' \colon [a, b] \times (0, \varepsilon_1) \to \mathbb{C}^{d \times d}$ from Model Problem 2 are C^s–bounded independently of ε. We assume that the multiplicities are chosen, such that*

$$s \;\geq\; \max_{j=1,\dots,\kappa} |m_j|\,.$$

Of course, the above inequality should hold for all three types of multiplicities.

Now we can estimate the quantities of the Numerical Scheme 3.

Proposition 6.5.4. *Let Assumptions 6–11 hold. There exists a constant $c > 0$ independent of n, ε, and h_n, such that*

$$\|A_n\|,\ \|B_n\| \;\leq\; c\,\rho\,h_n\,, \tag{6.55}$$

$$\|v^n\|,\ \|w^n\| \;\leq\; c\,h_n\,. \tag{6.56}$$

Furthermore it holds $\|C^j_{n,\alpha}\|, \|C^j_{n,\beta}\|, \|c^j_{n,\alpha}\|, \|c^j_{n,\beta}\| \leq c$ for all $j = 0, \dots, \tau$.

Proof. We give the proof only for A_n and[2] $C^\bullet_{n,\alpha}$. For the remaining corresponding sets of quantities $\{B_n, C^\bullet_{n,\beta}\}$, $\{v^n, c^\bullet_{n,\alpha}\}$ and $\{w^n, c^\bullet_{n,\beta}\}$ it is (completely) analogue.

First we prove by induction that the matrices $K_{j,l}(\varepsilon)$ are bounded independently of ε and $\zeta^n_1, \ldots, \zeta^n_\kappa$. In the following we (simply) write ζ instead of $\zeta^n_1, \ldots, \zeta^n_\kappa$.

The matrix function

$$P_1(x,\varepsilon) \;=\; M'(x,\varepsilon) \odot \sum_{l=0}^{|m_1|-1} K_{1,l}(\varepsilon) \odot M(x,\varepsilon)^{\odot l}$$

is componentwise defined by an interpolation problem, cf. Numerical Scheme 3 p. 137. By assumption the matrix valued function $S_1 = S$ is C^s–bounded independently of ε. Hence we can apply Lemma 6.5.3 for each component of $P_1(x,\varepsilon)$, which yields that each component of the matrices

$$K_{1,0}(\varepsilon)\,,\; \ldots\;,\; K_{1,|m_1|-1}(\varepsilon)$$

are bounded independently of the support abscissas ζ and ε. Hence there exists a constant $c > 0$ independently of ε and ζ, such that

$$\|K_{1,l}\|_{\sup} \;\le\; c\,, \qquad l = 0, \ldots, |m_j| - 1\,.$$

Since all norms on $\mathbb{C}^{d\times d}$ are equivalent the same holds for $\|\cdot\|$.

Now assume $K_{j,l}(\varepsilon), \ldots, K_{j,|m_j|-1}(\varepsilon)$ are bounded independently of the support abscissas ζ and ε. Since D_Φ is C^s–bounded independently of ε, the same holds for

$$\begin{aligned} P^\diamond_j(x,\varepsilon) \;&=\; i\varepsilon \sum_{j=0}^{|m_j|-1} \left(\sum_{l=j}^{|m_j|-1} K_{j,l}(\varepsilon)\, \tfrac{l!}{j!} (-i\varepsilon)^{l-j} \right) \odot D_\Phi(x,\varepsilon)^{\odot j} \\ &\quad + \sum_{l=0}^{|m_j|-1} \operatorname{diag}_\nu(K_{j,l}(\varepsilon))\, \frac{x^{l+1}}{l+1}\,, \end{aligned}$$

with $(x,\varepsilon) \in [a,b] \times (0,\varepsilon_1)$. Thus the matrix valued function

$$S_{j+1} \;=\; S\, P^\diamond_j$$

is C^s–bounded independently of ε. Since all norms on $\mathbb{C}^{d\times d}$ are equivalent this holds in particular for $\|\cdot\|_{\sup}$. Hence we can apply Lemma 6.5.3 for each component of $P_{j+1}(x,\varepsilon)$. This yields that the $K_{j+1,0}(\varepsilon), \ldots, K_{j+1,|m_j+1|-1}(\varepsilon)$ are componentwise bounded independently of the support abscissas ζ and ε. Again by equality of norms we find a $c > 0$ independently of ε and ζ, such that

$$\|K_{1,l}\| \;\le\; c\,, \qquad l = 0, \ldots, |m_j| - 1\,.$$

An immediate consequence of the previous calculation is that $P^\diamond_1, \ldots, P^\diamond_\tau$ are C^s–bounded independently of ε. Inductively we deduce from the definition that

[2] Here $C^\bullet_{n,\beta}$ is an abbreviation for $C^0_{n,\beta}, \ldots, C^\tau_{n,\beta}$.

$\mathcal{C}^0_{n,\alpha}, \dots, \mathcal{C}^\tau_{n,\alpha}$ are bounded independently of ε and ζ. Since $E_\varepsilon(x_n)$ is a unitary matrix we get

$$\|C^0_{n,\alpha}\| \;=\; \|E^*_\varepsilon(x_n)\mathcal{C}^0_{n,\alpha}E_\varepsilon(x_n)\| \;\le\; \|\mathcal{C}^0_{n,\alpha}\|\,.$$

Thus $C^0_{n,\alpha}$ is bounded independently of ε and ζ. By definition we have

$$Q^j_{\alpha,n} \;=\; I_{x_n}[P_j](x_{n+1}) \;=\; \int_{x_n}^{x_{n+1}} E^*_\varepsilon(x)P_j(x,\varepsilon)E_\varepsilon(x)\;dx\,.$$

If follows

$$\begin{aligned}\|Q^j_{\alpha,n}\| &\le h_n \sup_{x\in[x_n,x_{n+1}]} \|E^*_\varepsilon(x)P_j(x,\varepsilon)E_\varepsilon(x)\| \\ &\le h_n \sup_{x\in[x_n,x_{n+1}]} \|P_j(x,\varepsilon)\|\,.\end{aligned}$$

Since $K_{j,0}(\varepsilon), \dots, K_{j,|m_j|-1}$ are bounded independently of ε and ζ, the same holds for P_j. Hence there exists a $c>0$ independently of ε and ζ, such that

$$\|Q^j_{\alpha,n}\| \;\le\; c\,h_n\,.$$

Finally we compute

$$\|A_n\|_{\sup} \;\le\; \sum_{j=1}^{\tau} \|Q^j_{\alpha,n}\|\rho^j \sum_{k=0}^{\tau-j} \rho^k \|C^k_{n,\alpha}\| \;\le\; c\;\rho\;h_n\,.$$

□

6.6 The local error

In this section we derive an estimate for the local error of the OSM (τ,κ,ι,m) from § 6.4 on the interval $[x_n, x_{n+1}]$. For the whole section Assumptions 6–11 shall hold. Let us start with a recall of the local error. By (6.32)

$$\mathrm{err}^n \;=\; \mathrm{err}^{n,\tau}_{\mathrm{trunc}} + \mathrm{err}^n_{\mathrm{int}}$$

with (cf. (6.33))

$$\mathrm{err}^n_{\mathrm{int}} \;=\; \sigma_e\;\mathrm{err}_{x_n}(x_{n+1}) \;-\; \sigma_i\;\mathrm{err}_{x_{n+1}}(x_n)\,. \tag{6.57}$$

and (cf. (6.29))

$$\begin{aligned}\mathrm{err}_\xi(x) \;=\; &\sum_{j=1}^{\tau} \rho^j \mathcal{E}^j_\xi(x)\left(z(\xi) \;-\; \lambda\sum_{k=0}^{\tau-j}\rho^k\,c^k_\xi\right) \\ &+\lambda\sum_{j=0}^{\tau}\rho^j\sum_{k=0}^{j}\left(\mathcal{I}^{j-k}_\xi\,\mathrm{err}^k_\xi\right)(x)\,.\end{aligned}$$

Furthermore we get from (6.24)

$$\mathcal{E}^j_\xi(x) \;=\; \sum_{k=1}^{j}\sum_{l=1}^{k}\left(\mathcal{I}^{j-k}_\xi\,\mathrm{Err}^l_\xi\right)(x)\,C^{k-l}_\xi\,.$$

For the definition of $\mathrm{Err}_\xi^\bullet$, $\mathrm{err}_\xi^\bullet$, $C_\xi^\bullet$ and $c_\xi^\bullet$ we refer to Lemma 6.2.5 and Lemma 6.2.6. Since we already derived an estimate for the truncation error in Lemma 6.1.7, we focus the following discussion on $\mathrm{err}_{\mathrm{int}}^n$. For this purpose we continue to estimate $\mathrm{err}_\xi(x)$. We should keep in mind that the variables ξ, x take the values $\xi = x_n$, $x = x_{n+1}$ and vice versa.

Notation. *In the sequel we use the notation*

$$\|F\|_{\infty,n} \quad := \quad \sup_{x\in[x_n,x_{n+1}]} \|F(x)\|\,.$$

The triangle inequality yields

$$\begin{aligned}
&\|\,\mathrm{err}_\xi(x)\| \\
&\quad\le \quad \sum_{j=1}^{\tau}\rho^j\Big(\sum_{k=1}^{j}\sum_{l=1}^{k}\|\big(\mathcal{I}_\xi^{j-k}\,\mathrm{Err}_\xi^l\,\big)(x)\|\,\|C_\xi^{k-l}\|\Big)\Big(\|z(\xi)\| \;+\; \lambda\sum_{k=0}^{\tau-j}\rho^k\,\|c_\xi^k\|\Big) \\
&\qquad + \lambda\sum_{j=0}^{\tau}\rho^j\sum_{k=0}^{j}\|\big(\mathcal{I}_\xi^{j-k}\,\mathrm{err}_\xi^k\,\big)(x)\|\,.
\end{aligned}$$

Due to Proposition 6.5.4 there exists a constant c independently of ε, n, h_n, such that $\|C_\xi^j\|, \|c_\xi^j\| < c$ for $j = 0,\dots,\tau$. By Proposition 6.2.2 $\|z(\xi)\|$ is bounded by a constant, which is independently of ε an ξ.

Remark 6.6.1. *In the sequel $c \ge 0$ always denotes a constant, which is independently of ε, h_n, and n.*

Hence there exists a $c \ge 0$, such that (use Remark 6.2.8 to rearrange the sum)

$$\begin{aligned}
\|\,\mathrm{err}_\xi(x)\| \quad\le\quad & c\sum_{l=1}^{\tau}\sum_{j=l}^{\tau}\sum_{k=l}^{j}\rho^j\|\big(\mathcal{I}_\xi^{j-k}\,\mathrm{Err}_\xi^l\,\big)(x)\| \\
& + \lambda\sum_{k=0}^{\tau}\sum_{j=k}^{\tau}\rho^j\|\big(\mathcal{I}_\xi^{j-k}\,\mathrm{err}_\xi^k\,\big)(x)\|\,.
\end{aligned} \tag{6.58}$$

It turns out that the summand with $l = k = j = 1$ in the first line of the right–hand side is a crucial term. This is the only terms which is just multiplied with ρ. All other summands contain a factor of $\mathcal{O}(\rho^2)$ (keep in mind that ρ is a small parameter, e. g. $\rho = \varepsilon^\alpha$, $\alpha > 0$). Thus, we need a more sophisticated estimate for it. Also the other summands with $l = 1$ have to be treated separately. From Lemma 6.1.3 we get an estimate for the iterated integrals, which yields

$$\begin{aligned}
\|\,\mathrm{err}_\xi(x)\| \quad\le\quad & c\sum_{j=1}^{\tau}\rho^j\|\,\mathrm{Err}_\xi^1(x)\| \;+\; c\sum_{j=2}^{\tau}\sum_{k=1}^{j-1}\rho^j\|\big(\mathcal{I}_\xi^{j-1-k}\mathcal{I}_\xi^1\,\mathrm{Err}_\xi^1\,\big)(x)\|, \\
& + c\sum_{l=2}^{\tau}\sum_{j=l}^{\tau}\rho^j\sum_{k=l}^{j}\frac{(|\xi-x|\,\|S\|_\infty)^{j-k}}{(j-k)!}\|\,\mathrm{Err}_\xi^l\,\|_{\infty,n} \\
& + \lambda\sum_{k=0}^{\tau}\rho^k\sum_{j=k}^{\tau}\rho^{j-k}\frac{(|\xi-x|\,\|S\|_\infty)^{j-k}}{(j-k)!}\|\,\mathrm{err}_\xi^k\,\|_{\infty,n}\,.
\end{aligned}$$

For $0 \le \rho < 1$ it holds $\sum_{j=1}^{\tau} \rho^j \le \frac{\rho}{1-\rho}$. Furthermore $|\xi - x| = h_n \le h$ holds for all n. Let $\gamma := h\|S\|_\infty$. In the first line we substitute $i = j-1$. Further we use Remark 6.2.8 and the estimate $\sum_{j=r}^{s} \frac{z^j}{j!} \le e^z$ $(z \in \mathbb{R}_0^+)$, which yields

$$\begin{aligned}
\|\operatorname{err}_\xi(x)\| &\le \tfrac{c\,\rho}{1-\rho}\|\operatorname{Err}_\xi^1(x)\| + c\sum_{i=1}^{\tau-1}\sum_{k=1}^{i}\rho^{i+1}\frac{\gamma^{i-k}}{(i-k)!}\|\big(\mathcal{I}_\xi^1 \operatorname{Err}_\xi^1\big)(x)\|,\\
&\quad + c\sum_{l=2}^{\tau}\sum_{k=l}^{\tau}\rho^k\sum_{j=k}^{\tau}\frac{(\rho\,\gamma)^{j-k}}{(j-k)!}\|\operatorname{Err}_\xi^l\|_{\infty,n}\\
&\quad + \lambda\sum_{k=0}^{\tau}\rho^k\sum_{j=k}^{\tau}\frac{(\rho\,\gamma)^{j-k}}{(j-k)!}\|\operatorname{err}_\xi^k\|_{\infty,n}\\
&\le \tfrac{c\,\rho}{1-\rho}\|\operatorname{Err}_\xi^1(x)\| + c\sum_{k=1}^{\tau-1}\rho^{k+1}\sum_{i=k}^{\tau-1}\frac{(\rho\,\gamma)^{i-k}}{(i-k)!}\|\big(\mathcal{I}_\xi^1 \operatorname{Err}_\xi^1\big)(x)\|,\\
&\quad + c\sum_{l=2}^{\tau}\sum_{k=l}^{\tau}\rho^k e^{\rho\,\gamma}\|\operatorname{Err}_\xi^l\|_{\infty,n} + \lambda\sum_{k=0}^{\tau}\rho^k e^{\rho\,\gamma}\|\operatorname{err}_\xi^k\|_{\infty,n}\\
&\le \tfrac{c}{1-\rho}\,\rho\,\|\operatorname{Err}_\xi^1(x)\| + \tfrac{c\,e^{\rho\,\gamma}}{1-\rho}\,\rho^2\,\|\big(\mathcal{I}_\xi^1 \operatorname{Err}_\xi^1\big)(x)\|,\\
&\quad + \tfrac{c\,e^{\rho\,\gamma}}{1-\rho}\,\rho^2\sum_{l=2}^{\tau}\rho^{l-2}\|\operatorname{Err}_\xi^l\|_{\infty,n} + \lambda e^{\rho\,\gamma}\sum_{k=0}^{\tau}\rho^k\|\operatorname{err}_\xi^k\|_{\infty,n}.
\end{aligned}$$

Hence there is a constant $c \ge 0$, such that

$$\begin{aligned}
\|\operatorname{err}_\xi(x)\| &\le c\,\rho\,\|\operatorname{Err}_\xi^1(x)\| + c\,\rho^2\,\|\big(\mathcal{I}_\xi^1 \operatorname{Err}_\xi^1\big)(x)\|,\\
&\quad + c\,\rho^2\sum_{l=2}^{\tau}\|\operatorname{Err}_\xi^l\|_{\infty,n} + \lambda c\sum_{k=0}^{\tau}\rho^k\|\operatorname{err}_\xi^k\|_{\infty,n}.
\end{aligned} \tag{6.59}$$

Remark 6.6.2. *To derive* (6.59) *from* (6.58) *we assumed* $0 \le \rho < 1$. *This is just for simplicity of the calculation. Since* $\tau \in \mathbb{N}$ *is finite, the appearing geometric sums are always finite and hence* (6.59) *holds for all* $\rho \in \mathbb{R}_0^+$.

Since we use Proposition 6.3.1 to determine the functions $P_1, \dots, P_\tau$ of the OSM (q. v. Lemma 6.2.5), the quadrature errors Err_ξ^j can be estimated by (6.41). There the error is estimated in terms of $h := \max\big(|\xi_\kappa - \alpha|, |\xi_1 - \beta|\big)$, which is related, but in general not equal to the local spatial step size h_n. In order to avoid confusion in the proofs of the following Lemma 6.6.4 and Lemma 6.6.5 we denote the quantity h from Proposition 6.3.1 by Δ_n.

Remark 6.6.3. *The variable* ξ *takes the values* $\{x_n, x_{n+1}\}$ *and* $x \in [x_n, x_{n+1}]$. *Comparing Proposition 6.3.1 with the quantities of this section we get*

$$\alpha = \xi\,,\ \beta = x \quad \text{or} \quad \alpha = x\,,\ \beta = \xi\,.$$

Hence it is

$$\Delta_n = \max\big(|\xi_1^n - \xi|, |\xi_1^n - x|, |\xi_\kappa^n - \xi|, |\xi_\kappa^n - x|\big)\,. \tag{6.60}$$

There exists a constant $c > 0$, *which only depend on*[3] ι, *such that*

$$\tfrac{1}{c}\,h_n \le \Delta_n \le c\,h_n\,. \tag{6.61}$$

[3]The set of parameters ι determines the support abscissa for the interpolation problems.

Proof. We get an upper estimate for Δ_n by replacing in (6.60) each ξ and x with x_n and x_{n+1}. Hence we search for the maximum of eight absolute values. Since some of the terms coincide, we have

$$\Delta_n \quad \leq \quad \max\left(|\xi_1^n - x_n|,\, |\xi_1^n - x_{n+1}|,\, |\xi_\kappa^n - x_n|,\, |\xi_\kappa^n - x_{n+1}|\right).$$

Due to the definition of the support abscissas (see p.136) and Assumption 10 we get

$$\Delta_n \quad \leq \quad \max\left(|\iota_1 - \iota_{j_\alpha}|,\, |\iota_1 - \iota_{j_\beta}|,\, |\iota_\kappa - \iota_{j_\alpha}|,\, |\iota_\kappa - \iota_{j_\beta}|\right) h_n\,.$$

On the other hand side we have

$$\Delta_n \quad \geq \quad \left\{ \begin{array}{rcl} \xi_\kappa^n - \xi & = & \iota_\kappa\, h_n\,,\ \xi = x_n\,, \\ \xi - \xi_1^n & = & (1-\iota_1)\, h_n\,,\ \xi = x_{n+1} \end{array} \right. .$$

By Assumption 10 ι_κ is strictly positive and ι_1 is non–positive. Hence there exists a constant $c > 0$ independently of the grid, such that

$$\tfrac{1}{c}\, h_n \;\leq\; \Delta_n \;\leq\; c\, h_n\,.$$

□

Now we are prepared to estimate $\|\operatorname{Err}_\xi^l\|$.

Lemma 6.6.4. *There are constants $c, \gamma_{\mathrm{Err}} \geq 0$ independently of ε, h_n, and n, such that for all $x \in [x_n,\, x_{n+1}]$ it holds:*

$$\begin{aligned} \|\operatorname{Err}_\xi^1(x)\| &\leq c\, h_n^{|m_1|+1} \min\left(1,\, \gamma_{\mathrm{Err}} \left(\frac{\varepsilon}{h_n}\right)^{\mu_1+1}\right), \\ \|(\mathcal{I}_\xi^1 \operatorname{Err}_\xi^1)(x)\| &\leq c\,\varepsilon\, h_n^{|m_1|+1}\,, \end{aligned}$$

and

$$\|\operatorname{Err}_\xi^k(x)\| \quad \leq \quad c\,\varepsilon\, h_n^{|m_k|+1}\,, \qquad k = 2, \ldots, \tau\,.$$

The constants c, γ_{Err} depend on δ and tend to infinity as $\delta \to 0$.

Proof. By definition (cf. Lemma 6.2.5) we get

$$\operatorname{Err}_\xi^1(x) \quad = \quad I_\xi[S_1 - P_1](x)\,.$$

Since the ν–diagonal entries of the matrix function $S_1 = S$ from ODE (6.16) vanish identically and since P_1 is constructed as in Proposition 6.3.1[4], we get from (6.43) of Remark 6.3.2 the error estimate

$$\|\operatorname{Err}_{x_n}^1(x_{n+1})\| \quad \leq \quad c\,|x_{n+1} - x_n|\, \Delta_n^{|m_1|} \min\left(1\,, \gamma_1 \left(\frac{\varepsilon}{\Delta_n}\right)^{\mu_1+1}\right). \quad (6.62)$$

Since the constants c, γ_1 do not depend on ε and h_n or n, the first estimate of Lemma 6.6.4 follows with (6.61) of Remark 6.6.3. The constants $c, \gamma_1 \geq 0$ depend on δ and tends to ∞ as $\delta \to 0$.

[4]replace: $\alpha \mapsto \min(x, \xi)$, $\beta \mapsto \max(x, \xi)$, $F \mapsto S_1$, $\xi_j \mapsto \zeta_j$ and $m_j \mapsto m_{0,j}$ for $j = 1, \ldots, \kappa$

In order to estimate $\|\mathcal{I}_\xi^1 \operatorname{Err}_\xi^1(x)\|$ let us make one integration by parts as done in § 3.1.2. This is possible, because $\operatorname{diag}_\nu(S - P_1) = 0$. By construction it holds $S(\xi) - P_1(\xi) = 0$. Also remember that $E_\varepsilon^* B E_\varepsilon = E_\Phi^\varepsilon \odot B$ for every matrix $B \in \mathbb{C}^{d\times d}$. Hence

$$\begin{aligned}
\mathcal{I}_\xi^1 \operatorname{Err}_\xi^1(x) &= i\varepsilon \int_\xi^x \big(E_\varepsilon^* S E_\varepsilon\big)(t) \int_\xi^t \big(\big(E_\Phi^\varepsilon\big)' \odot D_\Phi^- \odot (S - P_1)\big)(r)\, dr\, dt \\
&= i\varepsilon \int_\xi^x \big(E_\varepsilon^* S E_\varepsilon\big)(t)\big(\big(E_\Phi^\varepsilon\big) \odot D_\Phi^- \odot (S - P_1)\big)(t)\, dt \\
&\quad + i\varepsilon \int_\xi^x \big(E_\varepsilon^* S E_\varepsilon\big)(t) \int_\xi^t \big(E_\Phi^\varepsilon \odot (D_\Phi^-)' \odot (S - P_1)\big)(r)\, dr\, dt \\
&\quad + i\varepsilon \int_\xi^x \big(E_\varepsilon^* S E_\varepsilon\big)(t) \int_\xi^t \big(E_\Phi^\varepsilon \odot D_\Phi^- \odot (S - P_1)'\big)(r)\, dr\, dt
\end{aligned}$$

It holds $E_\varepsilon^* B E_\varepsilon (E_\Phi^\varepsilon \odot A) = E_\varepsilon^* B A E_\varepsilon$ for $A, B \in \mathbb{C}^{d\times d}$ (cf. § 3.1.2). This yields the estimate

$$\begin{aligned}
\|\mathcal{I}_\xi^1 \operatorname{Err}_\xi^1(x)\| \;\le\;& \varepsilon|\xi - x| \|E_\varepsilon^* S\big(D_\Phi^- \odot (S - P_1)\big) E_\varepsilon\|_{\infty,n} \\
&+ \varepsilon|\xi - x|^2 \|E_\varepsilon^* S E_\varepsilon\|_{\infty,n} \big\|E_\varepsilon^*\big((D_\Phi^-)' \odot (S - P_1)\big) E_\varepsilon\big\|_{\infty,n} \\
&+ \varepsilon|\xi - x|^2 \|E_\varepsilon^* S E_\varepsilon\|_{\infty,n} \big\|E_\varepsilon^*\big(D_\Phi^- \odot (S - P_1)'\big) E_\varepsilon\big\|_{\infty,n} .
\end{aligned}$$

Let $J_n := [x_n, x_{n+1}]$. Since $\|\cdot\|_{\sup}$ and $\|\cdot\|$ are equivalent norms, there exists a constant $\widehat{c} > 0$, such that $\frac{1}{\widehat{c}}\|A\| \le \|A\|_{\sup} \le \widehat{c}\|A\|$ and hence

$$\begin{aligned}
\|A \odot B\|_{\infty,n} &= \sup_{x\in J_n} \|A(x) \odot B(x)\| \;\le\; \widehat{c} \sup_{x\in J_n} \|A(x) \odot B(x)\|_{\sup} \\
&\le \widehat{c} \sup_{x\in J_n} \|A(x)\|_{\sup} \|B(x)\|_{\sup} \\
&\le \widehat{c}^2 \sup_{x\in J_n} \|A(x)\| \sup_{y\in J_n} \|B(y)\| \;=\; \widehat{c}^2 \|A\|_{\infty,n} \|B\|_{\infty,n} .
\end{aligned}$$

The matrix $E_\varepsilon(x)$ is unitary and thus

$$\begin{aligned}
\|\mathcal{I}_\xi^1 \operatorname{Err}_\xi^1(x)\| \;\le\;& \widehat{c}^2\, \varepsilon\, |\xi - x|\, \|S\|_{\infty,n}\, \|D_\Phi^-\|_{\infty,n}\, \|S - P_1\|_{\infty,n} \\
&+ \widehat{c}^2\, \varepsilon\, |\xi - x|^2\, \|S\|_{\infty,n}\, \|(D_\Phi^-)'\|_{\infty,n}\, \|S - P_1\|_{\infty,n} \\
&+ \widehat{c}^2\, \varepsilon\, |\xi - x|^2\, \|S\|_{\infty,n}\, \|D_\Phi^-\|_{\infty,n}\, \|(S - P_1)'\|_{\infty,n} .
\end{aligned}$$

By Lemma 5.2.7 exists a constant $c \ge 0$, such that

$$\begin{aligned}
\|\mathcal{I}_\xi^1 \operatorname{Err}_\xi^1(x)\| \;\le\;& c\, \varepsilon\, h_n\, \|S\|_{\infty,n}\, \|D_\Phi^-\|_{\infty,n}\, h_n^{|m_1|} \\
&+ c\, \varepsilon\, h_n^2\, \|S\|_{\infty,n}\, \|(D_\Phi^-)'\|_{\infty,n}\, \|S\|_{\infty,n}\, h_n^{|m_1|} \\
&+ c\varepsilon\, |\xi - x|^2\, \|S\|_{\infty,n}\, \|D_\Phi^-\|_{\infty,n}\, h_n^{|m_1|-1} .
\end{aligned}$$

This yields the second estimate.

Further we get from (i)(b) of Numerical Scheme 3 ($\operatorname{diag}_\nu(S_1) = 0$)

$$P_1^\diamond(x) \;=\; i\varepsilon \sum_{k=0}^{|m_1|-1} \Bigg(\sum_{l=1}^{|m_1|-1} K_{1,l}(\varepsilon)\, \tfrac{l!}{k!} (-i\varepsilon)^{l-k} \Bigg) \odot D_\Phi(x,\varepsilon)^{\odot k} .$$

Hence we can write $P_1^\diamond = i\varepsilon\,\tilde{P}_1^\diamond$, with $\tilde{P}_1^\diamond$ C^s–bounded independently of ε and the spatial grid. This yields

$$S_2 \;=\; S\,P_1^\diamond \;=\; i\varepsilon\,S\,\tilde{P}_1^\diamond \;=:\; i\varepsilon\tilde{S}_2\,,$$

with $\tilde{S}_2$ C^s–bounded independently of ε. Let

$$\tilde{P}_2(x) \;=\; M'(x,\varepsilon)\odot\sum_{l=0}^{|m_2|-1}\tilde{K}_{2,l}(\varepsilon)\odot M(x,\varepsilon)^{\odot l}$$

be the unique solution of the generalized Hermite interpolation problem

$$\tilde{P}_2^{(k)}(\xi_l^n) \;=\; \tilde{S}_2^{(k)}(\xi_l^n)\,,\qquad k=0,\ldots,m_{2,l}-1,\;\; l=1,\ldots,\kappa\,.$$

Since the interpolation problem is linear and uniquely solvable, we get

$$P_2 \;=\; i\varepsilon\,\tilde{P}_2\,,\qquad P_2^\diamond \;=\; i\varepsilon\,\tilde{P}_2^\diamond\,,$$

and it holds

$$\mathrm{Err}_\xi^2(x) \;=\; I_\xi[S_2-P_2](x) \;=\; i\varepsilon\,I_\xi[\tilde{S}_2-\tilde{P}_2](x)\,.$$

Due to construction we can apply Proposition 6.3.1[5] for $\tilde{P}_2$, which yields

$$\|\,\mathrm{Err}_\xi^2(x)\| \;=\; \varepsilon\,\|I_\xi[\tilde{S}_2-\tilde{P}_2](x)\| \;\le\; \varepsilon\,c\,\theta\,|x-\xi|\,\Delta_n^{|m_2|}\,,$$

with a constant c independent of $\varepsilon,\Phi,\tilde{S}_2$ and δ. The constant θ is given by

$$\theta \;:=\; \max\left(\frac{\|\,\mathrm{diag}_\nu(\tilde{S}_2^{(|m_2|)})\|_\infty}{|m_2|!}\,,\; c_2\min\left(1,\,\gamma_2\left(\frac{\varepsilon}{\Delta_n}\right)^{\mu_2+1}\right)\right).$$

In general $\mathrm{diag}_\nu(\tilde{S}_2)\neq 0$. Hence θ is of order $\mathcal{O}(1)$ with respect to ε and h_n. Again we use (6.61) of Remark 6.6.3, which yields

$$\|\,\mathrm{Err}_\xi^2(x)\| \;\le\; \varepsilon\,c\,h_n^{|m_2|+1}\,.$$

Furthermore we get from (i)(b) of Numerical Scheme 3

$$\begin{aligned}\tilde{P}_2^\diamond(x) \;=\; & i\varepsilon\sum_{k=0}^{|m_j|-1}\left(\sum_{l=k}^{|m_2|-1}\tilde{K}_{2,l}(\varepsilon)\,\tfrac{l!}{k!}(-i\varepsilon)^{l-k}\right)\odot D_\Phi(x,\varepsilon)^{\odot k}\\ & +\sum_{l=0}^{|m_2|-1}\mathrm{diag}_\nu(\tilde{K}_{2,l}(\varepsilon))\,\frac{x^{l+1}}{l+1}\,.\end{aligned}$$

Hence it is $\tilde{P}_2^\diamond=\mathcal{O}(1)$ with respect to ε and C^s–bounded independently of ε. This yields

$$S_3 \;=\; S\,P_2^\diamond \;=\; i\varepsilon\,S\,\tilde{P}_2^\diamond \;=:\; i\varepsilon\tilde{S}_3\,,$$

with $\tilde{S}_3$ C^s–bounded independently of ε. Inductively continue this procedure to derive the remaining estimates. □

[5]replace: $\alpha\mapsto\min(x,\xi)$, $\beta\mapsto\max(x,\xi)$, $F\mapsto\tilde{S}_2$, $\xi_j\mapsto\zeta_j$ and $m_j\mapsto m_{0,j}$ for $j=1,\ldots,\kappa$

A similar result holds for the vector valued integrals.

Lemma 6.6.5. *There are positive constants $c > 0$ and $\widehat{\gamma}_0, \dots, \widehat{\gamma}_\tau > 0$ independently of ε and h_n, such that for all $x \in [x_n, x_{n+1}]$ it holds:*

$$\| \operatorname{err}_\xi^k(x) \| \quad \leq \quad c\, \varepsilon^k\, h_n^{|m_k|+1} \min\left(1, \widehat{\gamma}_k \left(\frac{\varepsilon}{h_n}\right)^{\mu_k+1}\right), \qquad k = 0, \dots, \tau\,.$$

The constants depend on δ and tends to infinity as $\delta \to 0$.

Proof. By definition (cf. Lemma 6.2.6) we get for $k = 0$

$$\operatorname{err}_\xi^0(x) \quad = \quad I_\xi^v[s_0 - u_0](x)\,.$$

Since u_0 is constructed as in Corollary 6.3.4[6], we get

$$\| \operatorname{err}_\xi^0(x) \| \quad \leq \quad c\, h_n\, \Delta_n^{|m_0|+1} \min\left(1, \widehat{\gamma}_0 \left(\frac{\varepsilon}{\Delta_n}\right)^{\mu_0+1}\right).$$

Thus (6.61) of Remark 6.6.3 yields (with a new constant $c \geq 0$)

$$\| \operatorname{err}_\xi^0(x) \| \quad \leq \quad c\, h_n^{|m_0|+1}\ \min\left(1, \widehat{\gamma}_0 \left(\frac{\varepsilon}{h_n}\right)^{\mu_0+1}\right).$$

Due to the used estimates the constants $c, \widehat{\gamma}_0$ depend on δ and tends to infinity as $\delta \to 0$. Furthermore we get from the OSM or Remark 6.3.5 that

$$u_0^\diamond(x) \quad = \quad i\varepsilon \sum_{k=0}^{|m_0|-1} \Phi(x,\varepsilon)^k \Bigg(\sum_{l=k}^{|m_0|-1} c_{0,l}(\varepsilon) \frac{l!}{k!} (-i\varepsilon)^{l-k} \Bigg)\,.$$

This yields

$$s_1 \quad = \quad S\, u_0^\diamond \ = \ i\varepsilon\, S\, \tilde{u}_0^\diamond \ =: \ i\varepsilon \tilde{s}_1\,,$$

with $\tilde{s}_1$ C^s–bounded independently of ε. Let

$$\tilde{u}_1(x,\varepsilon) \quad = \quad \Phi'(x,\varepsilon) \sum_{l=0}^{|m_1|-1} \Phi(x,\varepsilon)^l\, \tilde{c}_{1,l}(\varepsilon)\,,$$

be the unique solution of the generalized Hermite interpolation problem

$$\tilde{u}_j^{(k)}(\xi_l) \ = \ \tilde{s}_j^{(k)}(\xi_l) \quad \text{for} \quad k = 0, \dots, m_{j,l} - 1\,, \quad l = 1, \dots, \kappa\,.$$

Since the interpolation problem is linear and uniquely solvable, we get

$$u_1 \ = \ i\varepsilon\, \tilde{u}_1\,, \qquad u_1^\diamond \ = \ i\varepsilon\, \tilde{u}_1^\diamond\,,$$

and it holds

$$\operatorname{err}_\xi^1(x) \quad = \quad I_\xi^v[s_1 - u_1](x) \ = \ i\varepsilon\, I_\xi^v[\tilde{s}_1 - \tilde{u}_1](x)\,.$$

[6] replace: $\alpha \mapsto \min(x,\xi)$, $\beta \mapsto \max(x,\xi)$ $g \mapsto f_\xi$, $\xi_j \mapsto \zeta_j$ and $m_j \mapsto m_{0,j}$ for $j = 1, \dots, \kappa$

Due to construction we can apply Corollary 6.3.4[7] for $\tilde{u}_1$, which yields

$$\|\operatorname{err}^1_\xi(x)\| \;\leq\; \varepsilon\,\|I_\xi[\tilde{s}_1-\tilde{u}_1](x)\| \;\leq\; \varepsilon\, c\, h_n \Delta_n^{|m_1|}\, \min\left(1,\, \widehat{\gamma}_1\left(\frac{\varepsilon}{\Delta_n}\right)^{\mu_1+1}\right).$$

As before we use (6.61) of Remark 6.6.3 to estimate Δ_n by h_n. Further we get

$$u_1^\diamond(x) \;=\; (i\varepsilon)^2 \sum_{k=0}^{|m_1|-1} \Phi(x,\varepsilon)^k \Bigg(\sum_{l=k}^{|m_1|-1} \tilde{c}_{1,l}(\varepsilon)\frac{l!}{k!}(-i\varepsilon)^{l-k}\Bigg).$$

Hence we can write $u_1^\diamond = (i\varepsilon)^2\,\tilde{u}_1^\diamond$, with $\tilde{u}_1^\diamond$ C^s–bounded independently of ε and the spatial grid. The remaining estimates follow by induction. □

Now we continue to estimate (6.59). By Lemma 6.6.4 and Lemma 6.6.5 we get

$$\begin{aligned}\|\operatorname{err}_\xi(x)\| \;\leq\;& c\,\rho\, h_n^{|m_1|+1} \min\left(1,\, \gamma_{\mathrm{Err}}\left(\frac{\varepsilon}{h_n}\right)^{\mu_1+1}\right) \\ &+ c\,\rho^2\,\varepsilon \sum_{l=1}^{\tau} h_n^{|m_l|+1} \\ &+ \lambda c \sum_{k=0}^{\tau} \rho^k \varepsilon^k\, h_n^{|m_k|+1} \min\left(1,\, \widehat{\gamma}_k\left(\frac{\varepsilon}{h_n}\right)^{\mu_k+1}\right).\end{aligned} \tag{6.63}$$

From Lemma 6.1.7 we already know that the truncation error is $\mathcal{O}(\rho^{\tau+1}h_n^{\tau+1})$. Hence the quadratures we use should be at least of order $\tau+1$ with respect to the spatial step size h_n.

Assumption 12. *The multiplicities $m_{j,1},\ldots,m_{j,\kappa}$, are chosen, such that*

$$|m_j| \;=\; \sum_{l=1}^{\kappa} m_{j,l} \;\geq\; \tau\,, \qquad j=0,\ldots,\tau\,.$$

Let $\widehat{\gamma} := \max(\gamma_{\mathrm{Err}}, \widehat{\gamma}_1)$. Then it holds (with a new constant $c \geq 0$)

$$\begin{aligned}\|\operatorname{err}_\xi(x)\| \;\leq\;& c\,\rho\,(1+\lambda\varepsilon) h_n^{|m_1|+1} \min\left(1,\, \widehat{\gamma}\left(\frac{\varepsilon}{h_n}\right)^{\mu_1+1}\right) \\ &+ c\,\rho^2\,\varepsilon\, h_n^{\tau+1} \\ &+ \lambda\, c\, h_n^{|m_0|+1} \min\left(1,\, \widehat{\gamma}_0\left(\frac{\varepsilon}{h_n}\right)^{\mu_0+1}\right).\end{aligned} \tag{6.64}$$

Hence we have proven

Proposition 6.6.6. *Let Assumptions 6–12 hold. Then there are non–negative constants $c, \gamma_1, \gamma_0 \geq 0$, such that for all $\varepsilon \in (0, \varepsilon_1)$*

$$\|\operatorname{err}_\xi(x)\| \;\leq\; c\,\theta(\varepsilon, h_n)\, h_n^{\tau+1}\,,$$

with $\theta(\varepsilon, h_n)$ given by

$$\begin{aligned}\theta(\varepsilon,h_n) \;=\;& \lambda\, \min\left(1,\, \gamma_0\left(\frac{\varepsilon}{h_n}\right)^{\mu_0+1}\right) h_n^{|m_0|-\tau} \\ &+ \rho \min\left(1,\, \gamma_1\left(\frac{\varepsilon}{h_n}\right)^{\mu_1+1}\right) h_n^{|m_1|-\tau} \;+\; \varepsilon\rho^2\,.\end{aligned}$$

[7] replace: $\alpha \mapsto \min(x,\xi)$, $\beta \mapsto \max(x,\xi)$ $g \mapsto \tilde{s}_1$, $\xi_j \mapsto \zeta_j$ and $m_j \mapsto m_{0,j}$ for $j = 1,\ldots,\kappa$

Combing Proposition 6.6.6, Lemma 6.1.7 and (6.57) yields the main result of this section.

Corollary 6.6.7 (Local Error)**.** *Let Assumptions 6–12 hold. There are constants $c, \gamma_0, \gamma_1 > 0$ independently of ε and n, such that*

$$\| \operatorname{err}^n \| \quad \leq \quad c \left(\rho^{\tau+1} + \theta(\varepsilon, h_n) \right) h_n^{\tau+1} ,$$

with $\theta(\varepsilon, h_n)$ given by

$$\begin{aligned} \theta(\varepsilon, h_n) \quad = \quad & \lambda \min \left(1, \gamma_0 \left(\frac{\varepsilon}{h_n} \right)^{\mu_0+1} \right) h_n^{|m_0|-\tau} \\ & + \rho \min \left(1, \gamma_1 \left(\frac{\varepsilon}{h_n} \right)^{\mu_1+1} \right) h_n^{|m_1|-\tau} + \varepsilon \rho^2 . \end{aligned}$$

The constants tend to infinity as $\delta \to 0$.

6.6.1 Schemes of maximum order

The estimate for the local error in Corollary 6.6.7 holds for the whole "zoo" of one–step methods that fit to Assumption 6–12. In this section we shall have a closer look on the local error in order to construct schemes with a high asymptotic order with respect to ε. Up to now we have not specified the parameters ρ and λ. The form of ODE (6.16) is mainly motivated by Lemma 3.3.1 from § 3.3. Hence (in the sequel) these constants are supposed to be non–negative powers of ε.

Assumption 13. *There are constants $\vartheta_0, \vartheta_1 \geq 0$, such that*

$$\lambda \; = \; \varepsilon^{\vartheta_0} \quad \textit{and} \quad \rho \; = \; \varepsilon^{\vartheta_1} .$$

Let us review the estimate of Proposition 6.6.6. It says

$$\| \operatorname{err}_\xi(x) \| \quad \leq \quad c \, \theta(\varepsilon, h_n) \, h_n^{\tau+1} ,$$

with

$$\begin{aligned} \theta(\varepsilon, h_n) \quad = \quad & \varepsilon^{\vartheta_0} \min \left(1, \gamma_0 \left(\frac{\varepsilon}{h_n} \right)^{\mu_0+1} \right) h_n^{|m_0|-\tau} \\ & + \varepsilon^{\vartheta_1} \min \left(1, \gamma_1 \left(\frac{\varepsilon}{h_n} \right)^{\mu_1+1} \right) h_n^{|m_1|-\tau} + \varepsilon^{2\vartheta_1+1} . \end{aligned} \tag{6.65}$$

Here we have replaced ρ and λ with respect to Assumption 13. Since ε is a small constant, we are interested in a maximal asymptotic order of θ with minimal (numerical) effort. Since ϑ_1 is prescribed by the initial value problem the maximal achievable order is $\mathcal{O}(\varepsilon^{2\vartheta_1+1})$. In an optimal case the exponents of ε coincide in all three terms of (6.65). Since we do not assume $\vartheta_1, \vartheta_2 \in \mathbb{N}$, in general equality can not hold. Instead we get the (desired) inequalities

$$\vartheta_0 + \mu_0 + 1 \quad \geq \quad 2\vartheta_1 + 1 , \tag{6.66}$$

$$\vartheta_1 + \mu_1 + 1 \quad \geq \quad 2\vartheta_1 + 1 . \tag{6.67}$$

Since the choice of μ_0, μ_1 should not reduce the order with respect to the spatial step size h_n, we get the additional constraints

$$|m_0| - \tau - (\mu_0 + 1) \;\geq\; 0\,, \tag{6.68}$$
$$|m_1| - \tau - (\mu_1 + 1) \;\geq\; 0\,. \tag{6.69}$$

From (6.66) and (6.67) we immediately derive

$$\mu_1 \;\geq\; \vartheta_1\,, \qquad \mu_0 \;\geq\; 2\vartheta_1 - \vartheta_0\,, \tag{6.70}$$

Due to definition (cf. Remark 6.4.2) it further holds

$$|m_0| \;\geq\; 2\mu_0\,, \qquad |m_1| \;\geq\; 2\mu_1\,. \tag{6.71}$$

The values μ_0, μ_1 are defined as the minimum of the multiplicities at the boundary of the integration interval (see Remark 6.4.2). Hence the numerical effort grows with μ_0 and μ_1, since more and more derivatives have to be approximated in order to solve the Hermite interpolation problem. Also $|m_0|$ and $|m_1|$ should be as small as possible, since they are the degree of the generalized Hermite interpolation polynomials. Hence we are interested in the smallest natural numbers, such that the above derived constraints (6.68)–(6.71) hold.

Definition 6.6.8. *For $x \in \mathbb{R}$ let $\lceil x \rceil \in \mathbb{Z}$ denote the unique integer (cf. [23]), such that*

$$\lceil x \rceil - 1 \;<\; x \;\leq\; \lceil x \rceil\,.$$

In some literature (and in Matlab*) this map is also denoted as* ceil(x)*. Further we denote by $\lfloor x \rfloor \in \mathbb{Z}$ the unique integer, such that*

$$\lfloor x \rfloor \;\leq\; x \;<\; \lfloor x \rfloor + 1\,.$$

This map is also denoted as floor(x)*.*

Now it is simple to write down the optimal values.

Definition 6.6.9. *For given $\vartheta_0, \vartheta_1 \geq 0$ we set*

$$\mu_0^* \;:=\; \lceil 2\vartheta_1 - \vartheta_0 \rceil \quad \textit{and} \quad \mu_1^* \;:=\; \lceil \vartheta_1 \rceil\,.$$

Furthermore we define

$$m_0^* \;:=\; \max\left(\mu_0^*,\, \tau + 1\right) + \mu_0^* \quad \textit{and} \quad m_1^* \;:=\; \max\left(\mu_1^*,\, \tau + 1\right) + \mu_1^*\,.$$

Remark 6.6.10. *Obviously, μ_0^*, μ_1^* are the smallest integers which solve* (6.70)*. Combining* (6.68)*,* (6.69) *and* (6.71) *yields* $(j = 0, 1)$

$$|m_j| \;\;\geq\;\; \max\left(2\mu_j^*,\, \tau + \mu_j^* + 1\right) \;\;=\;\; \max\left(\mu_j^*,\, \tau + 1\right) + \mu_j^*\,.$$

Hence $m_0^*, m_1^* \in \mathbb{N}$ are the optimal (minimal) integers.

With Definition 6.6.9 and the previous discussion we deduce from Proposition 6.6.6 and Lemma 6.1.7

Corollary 6.6.11. *Let (τ, κ, ι, m) be the set of parameters of our OSM, such that*

$$\begin{aligned} \mu_0 &= \mu_0^*, \qquad |m_0| = m_0^*, \\ \mu_1 &= \mu_1^*, \qquad |m_1| = m_1^*, \end{aligned}$$

hold for all three types of support abscissas. Than there exists a constant $c > 0$ independently of ε and h_n, such that

$$\| \operatorname{err}^n \| \leq c\, \varepsilon^{2\vartheta_1+1} h_n^{\tau+1} + \operatorname{err}_{\text{trunc}}^{n,\tau} .$$

By Lemma 6.1.7 it holds

$$\| \operatorname{err}_{\text{trunc}}^{n,\tau} \| \leq c\, \varepsilon^{\vartheta_1(\tau+1)} h_n^{\tau+1} .$$

From the previous Corollary 6.6.11 we deduce that we get the maximal[8] possible order with respect to ε, if $\vartheta_1 > 0$ and

$$(\tau+1)\,\vartheta_1 \geq 2\vartheta_1 + 1 \quad \Leftrightarrow \quad \tau \geq \frac{\vartheta_1 + 1}{\vartheta_1} > 1 . \tag{6.72}$$

Remark 6.6.12. *Let* (6.72) *hold. Since τ is a natural number, it holds $\tau \geq 2$.*

Now one can ask, what is the minimal effort in our setting to get a "maximal" scheme with respect to ε? It is clear that we have to choose τ as small as possible, since τ is also the number of interpolation problems we have to solve in each step. Furthermore it is evident that we set (cf. Assumption 12 p.150)

$$|m_j| = \tau , \qquad j = 2, \ldots, \tau .$$

Remark 6.6.13. *For $\vartheta_1 \in [1, \infty)$ we can choose $\tau = 2$ in order to get a scheme with "maximal" asymptotic order with respect to ε. In the case of $\vartheta_1 \in (0, 1)$ we have to set*

$$\tau = 1 + \left\lceil \frac{1}{\vartheta_1} \right\rceil$$

to ensure an asymptotic behavior of order $\mathcal{O}(\varepsilon^{2\vartheta_1+1})$ as $\varepsilon \to 0$.

As we have seen the "highest" asymptotic order with respect to ε can be achieved for $\tau \geq 2$. For the first order schemes, i. e. $\tau = 1$, the situation is a bit different. In this case the ε–order of the truncation error for the scheme from Corollary 6.6.11 is smaller than the order of the quadrature error. We get

$$\| \operatorname{err}^n \| \leq \left(c_{\text{trunc}}\, \varepsilon^{2\vartheta_1} + c_{\text{quadr}}\, \varepsilon^{2\vartheta_1+1} \right) h_n^2 .$$

If we want to improve the asymptotic behavior of the first order scheme, we have to decrease the truncation error. The iterated integrals are highly oscillatory and from § 3.1.2 we get that they are of order $\mathcal{O}(\varepsilon)$. Hence we can use integration by parts (as done to construct the asymptotic method in § 5.1) to gain an additional ε in the truncation error estimate . Unfortunately integration by parts is a trade

[8] Here "maximal" has to be understood in the context of the derived error estimates. Since we do not know if they are sharp or not, we can not exclude the existence of one step methods with better asymptotic order.

off between ε and the local step size h_n. This means that we get a vector $\chi^{n,1}$ (cf. Lemma 6.6.14) which is of order $\mathcal{O}(\varepsilon h_n)$. Thus, if we want a first order scheme, i. e. a scheme with a global convergence error of order $\mathcal{O}(\varepsilon^{2\vartheta_1+1}h_n)$, we have to take this vector into account.

Also the case $\vartheta_1 = 0$ is special, since the previously derived estimates for the truncation error only give an $\mathcal{O}(1)$ behavior with respect to ε. But also here it is possible to modify the schemes in order to get truncation error estimates of order $\mathcal{O}(\varepsilon)$. The following Lemma 6.6.14 holds for the whole "zoo" of OSM.

Lemma 6.6.14. *There exists a constant $c > 0$ independently of ε and n, such that it holds for all $n \in \{n_a, \ldots, n_b - 1\}$*

$$\left\| \operatorname{err}_{\text{trunc}}^{n,\tau} \right\| \;\leq\; c\,\varepsilon^{(\tau+1)\vartheta_1+1}\, h_n^{\tau}\,, \tag{6.73}$$

$$\left\| \operatorname{err}_{\text{trunc}}^{n,\tau} - \rho^{\tau+1}\chi^{n,\tau} \right\| \;\leq\; c\,\varepsilon^{(\tau+1)\vartheta_1+1}\, h_n^{\tau+1}\,, \tag{6.74}$$

with

$$\begin{aligned}
\chi^{n,\tau} \;:=\; & \sigma_e\, i\varepsilon\, \rho^{\tau+1} \left(\mathcal{I}_{x_n}^{\tau-1} I_{x_n}[S(D_{\Phi'}^{-} \odot S)]\right)(x_{n+1})\; z(x_n) \\
& - \sigma_e\, i\varepsilon\, \rho^{\tau+1}\, \mathcal{I}_{x_n}^{\tau}(x_{n+1})\; G(x_{n+1})\; z(x_n) \\
& - \sigma_i\, i\varepsilon\, \rho^{\tau+1} \left(\mathcal{I}_{x_{n+1}}^{\tau-1} I_{x_{n+1}}[S(D_{\Phi'}^{-} \odot S)]\right)(x_n)\; z(x_{n+1}) \\
& + \sigma_i\, i\varepsilon\, \rho^{\tau+1}\, \mathcal{I}_{x_{n+1}}^{\tau}(x_n)\; G(x_n)\; z(x_{n+1})
\end{aligned}$$

and $G(x) := (E_{\Phi}^{\varepsilon} \odot D_{\Phi'}^{-} \odot S)(x)$.

Proof. By definition of the truncation error in (6.12) we obviously can write

$$\operatorname{err}_{\text{trunc}}^{n,\tau} \;=\; \rho^{\tau+1}\, v^{n,\tau} \;+\; \operatorname{err}_{\text{trunc}}^{n,\tau+1}\,,$$

with some vector $v^{n,\tau}$. In order to get an estimate for $\operatorname{err}_{\text{trunc}}^{n,\tau+1}$ we do the following integration by parts:

$$\begin{aligned}
(\mathcal{I}_{\xi}^{1} z)(x) \;&=\; \int_{\xi}^{x} (E_{\Phi}^{\varepsilon} \odot S)(t) z(t)\, dt \\
&=\; \int_{\xi}^{x} (E_{\Phi}^{\varepsilon} \odot S)(t)\, dt\, z(x) \\
&\quad - \int_{\xi}^{x} \int_{\xi}^{t} (E_{\Phi}^{\varepsilon} \odot S)(r)\, dr\, z'(t)\, dt\,.
\end{aligned} \tag{6.75}$$

Property (iv) of § 3.1.2 yields an ε independent constant $\widehat{c} \geq 0$, such that for all $\zeta \in [\xi, x]$ it holds

$$\left\| \int_{\xi}^{\zeta} (E_{\Phi}^{\varepsilon} \odot S)(t)\, dt \right\| \;\leq\; \widehat{c}\,\varepsilon\,.$$

Due to Corollary 6.1.5 the solution z of (6.16) is C^1–bounded independently of ε and hence we get from (6.75)

$$\left\| (\mathcal{I}_{\xi}^{1} z)(x) \right\| \;\leq\; \widehat{c}\,\varepsilon\, \|z\|_{\infty} \;+\; |x - \xi| \widehat{c}\,\varepsilon\, \|z'\|_{\infty} \;\leq\; c\,\varepsilon\,.$$

This yields with Lemma 6.1.3 and Lemma 6.1.7

$$\begin{aligned} \|\operatorname{err}_{\text{trunc}}^{n,\tau+1}\| &\leq \sigma_e\,\rho^{\tau+2}\big\|\mathcal{I}_{x_n}^{\tau+1}\big(\mathcal{I}_{x_n}^{1}z\big)(x_{n+1})\big\| \\ &\quad + \sigma_i\,\rho^{\tau+2}\big\|\mathcal{I}_{x_{n+1}}^{\tau+1}\big(\mathcal{I}_{x_{n+1}}^{1}z\big)(x_n)\big\| \\ &\leq c\,\varepsilon\,\rho^{\tau+2}h_n^{\tau+1}\,. \end{aligned} \tag{6.76}$$

From (6.12) we get

$$\begin{aligned} v^{n,\tau} &= \sigma_e\,\mathcal{I}_{x_n}^{\tau+1}(x_{n+1})z(x_n) \;-\; \sigma_i\,\mathcal{I}_{x_{n+1}}^{\tau+1}(x_n)z(x_{n+1}) \\ &\quad + \sigma_e\,\lambda(\mathcal{I}_{x_n}^{\tau+1}f_{x_n})(x_{n+1}) \;-\; \sigma_i\,\lambda(\mathcal{I}_{x_{n+1}}^{\tau+1}f_{x_{n+1}})(x_n)\,. \end{aligned}$$

To derive an estimate for $v^{n,\tau}$ we shall rephrase the matrix and vector valued integrals from the above equation. In the following computation we use x, ξ as wild cards for x_n, x_{n+1} and vice verse. Further we use the symbol $\circledast$ to mark the place of a functions free variable when assigned to an operator. By Definition 6.1.1 it holds ($\tau \geq 1$)

$$\mathcal{I}_\xi^{\tau+1}(x) = \left(\mathcal{I}_\xi^{\tau-1}\int_\xi^{\circledast}(E_\Phi^\varepsilon S)(t)\int_\xi^t(E_\Phi^\varepsilon S)(r)\,dr dt\right)(x)\,.$$

Since $\operatorname{diag}_\nu(S) = 0$, we can use property (iv) of § 3.1.2 for the inner integral, which yields

$$\begin{aligned} \mathcal{I}_\xi^{\tau+1}(x) &= \left(\mathcal{I}_\xi^{\tau-1}\int_\xi^{\circledast}(E_\Phi^\varepsilon \odot S)(t)\; i\varepsilon\big(E_\Phi^\varepsilon \odot D_{\Phi'}^- \odot S\big)(t)\,dt\right)(x) \\ &\quad -\left(\mathcal{I}_\xi^{\tau-1}\int_\xi^{\circledast}(E_\Phi^\varepsilon \odot S)(t)\; i\varepsilon\big(E_\Phi^\varepsilon \odot D_{\Phi'}^- \odot S\big)(\xi)\,dt\right)(x) \\ &\quad -\left(\mathcal{I}_\xi^{\tau-1}\int_\xi^{\circledast}(E_\Phi^\varepsilon \odot S)(t) i\varepsilon\int_\xi^t E_\Phi^\varepsilon(r)\odot\big(D_{\Phi'}^-\odot S\big)'(r)\,dr dt\right)(x)\,. \end{aligned}$$

With property (ii) of § 3.1.2 we can simplify the first integral. The second one is just $\mathcal{I}_\xi^1$ times a constant matrix. Since we want to construct a τ^{th}–order scheme[9] with respect to the spatial step size, the remainder has to be of order $\mathcal{O}(h_n^{\tau+1})$. Hence we can neglect the third integral. This yields

$$\begin{aligned} \mathcal{I}_\xi^{\tau+1}(x) &= i\varepsilon\,\big(\mathcal{I}_\xi^{\tau-1}I_\xi[S(D_{\Phi'}^-\odot S)]\big)(x) \\ &\quad - i\varepsilon\,\mathcal{I}_\xi^{\tau}(x)\big(E_\Phi^\varepsilon\odot D_{\Phi'}^-\odot S\big)(\xi) \;+\; \mathcal{O}(\varepsilon\,h_n^{\tau+1})\,. \end{aligned} \tag{6.77}$$

The vector valued integrals are similarly treated. By Definition 6.1.1 and with the definition of f_ξ in Lemma 6.1.4 we get

$$\big(\mathcal{I}_\xi^2 f_\xi\big)(x) = \int_\xi^x(E\odot S)(t)\int_\xi^t(E\odot S)(r)\int_\xi^r f(u)\,du\,dr\,dt\,.$$

Integration by parts yields

$$\begin{aligned} \big(\mathcal{I}_\xi^2 f_\xi\big)(x) &= \int_\xi^x(E\odot S)(t)\int_\xi^t(E\odot S)(r)\,dr\int_\xi^t f(u)\,du\,dt \\ &\quad -\int_\xi^x(E\odot S)(t)\int_\xi^t\int_\xi^r(E\odot S)(u)\,du\,f(r)\,dr\,dt\,. \end{aligned}$$

[9] Here the convergence order is meant.

By property (iv) of § 3.1.2 we know that there exists a constant $c > 0$ independently of ε, such that

$$\left\| \int_\xi^r (E \odot S)(u)\, du \right\| \;\leq\; c\,\varepsilon\,, \quad \text{which yields} \quad \left\| \big(\mathcal{I}_\xi^2 f_\xi\big)(x) \right\| \;=\; \mathcal{O}(\varepsilon h_n^2)\,.$$

Since we can write

$$(\mathcal{I}_\xi^{\tau+1} f_\xi)(x) \;=\; \big(\mathcal{I}_\xi^{\tau-1}(\mathcal{I}_\xi^2 f_\xi)\big)(x)\,,$$

Lemma 6.1.3 yields the estimate

$$\|(\mathcal{I}_\xi^{\tau+1} f_\xi)(x)\| \;\leq\; c\,\varepsilon\, h_n^{\tau+1}\,. \tag{6.78}$$

Combining (6.77) and (6.78) yield

$$\begin{aligned} v^n \;=\;& \sigma_e\, i\varepsilon\, \rho^{\tau+1}\, \big(\mathcal{I}_{x_n}^{\tau-1} I_{x_n}[S(D_{\Phi'}^- \odot S)]\big)(x_{n+1})\; z(x_n) \\ & - \sigma_e\, i\varepsilon\, \rho^{\tau+1}\, \mathcal{I}_{x_n}^{\tau}(x_{n+1})\; G(x_{n+1})\; z(x_n) \\ & - \sigma_i\, i\varepsilon\, \rho^{\tau+1}\, \big(\mathcal{I}_{x_{n+1}}^{\tau-1} I_{x_{n+1}}[S(D_{\Phi'}^- \odot S)]\big)(x_n)\; z(x_{n+1}) \\ & + \sigma_i\, i\varepsilon\, \rho^{\tau+1}\, \mathcal{I}_{x_{n+1}}^{\tau}(x_n)\; G(x_n)\; z(x_{n+1}) \\ & + \mathcal{O}(\varepsilon\rho^{\tau+1}\, h_n^{\tau+1})\,. \end{aligned}$$

Again using Lemma 6.1.3 we deduce the estimate $\|v^{n,\tau}\| \leq c\varepsilon\rho^{\tau+1}h_n^\tau$, which yields the first estimate (6.73). Furthermore, by definition of $\chi^{n,\tau}$, it holds

$$v^{n,\tau} \;=\; \chi^{n,\tau} \,+\, \mathcal{O}(\varepsilon\rho^{\tau+1}\, h_n^{\tau+1})\,.$$

This yields with (6.76)

$$\mathrm{err}_{\mathrm{trunc}}^{n,\tau} - \rho^{\tau+1}\chi^{n,\tau} \;=\; \mathcal{O}(\varepsilon\rho^{\tau+1}\, h_n^{\tau+1}) + \mathrm{err}_{\mathrm{trunc}}^{n,\tau+1} \;=\; \mathcal{O}(\varepsilon\rho^{\tau+1}\, h_n^{\tau+1})\,,$$

which is the second estimate (6.74). □

The proof of Lemma 6.6.14 yields

Corollary 6.6.15.

$$\begin{aligned} \mathrm{err}_{trunc}^{n,\tau} \;=\;& \varepsilon^{\vartheta_1(\tau+1)}\big(\sigma_e\, \mathcal{I}_{x_n}^{\tau+1}(x_{n+1}) z(x_n) \,-\, \sigma_i\, \mathcal{I}_{x_{n+1}}^{\tau+1}(x_n) z(x_{n+1})\big) \\ & + \mathcal{O}\big(\varepsilon^{\vartheta_1(\tau+1)+1} h_n^{\tau+1}\big)\,. \end{aligned}$$

A first consequence of Lemma 6.6.14 is a refined error estimate for the schemes of Corollary 6.6.11

Corollary 6.6.16. *For the scheme from Corollary 6.6.11 it holds*

$$\|\,\mathrm{err}^n\,\| \;\leq\; c\left(\varepsilon^{\vartheta_1(\tau+1)} \min\left(c_*,\, \frac{\varepsilon}{h_n}\right) \,+\, \varepsilon^{2\vartheta_1+1}\right) h_n^{\tau+1}\,.$$

with non–negative constants $c, c_* \geq 0$ independently of ε.

Remark 6.6.17. *Hence even for $\vartheta_1 = 0$ the schemes from Corollary 6.6.11 are asymptotically correct with respect to ε. But with a reduced spatial order.*

Another consequence of Lemma 6.6.14 is that we have to incorporate the vector $\rho^2\chi^{n,1}$ in our first order scheme, if we want a local discretization error of $\mathcal{O}(\varepsilon^{2\vartheta_1+1}h_n^2)$. I. e. we take our first order scheme from Corollary 6.6.11 and add an approximation of $i\varepsilon\rho^2\chi^{n,1}$.

Remark 6.6.18. *Since it holds $\mathcal{I}^1_\xi(x) = -\mathcal{I}^1_x(\xi)$ and $I_\xi[F](x) = -I_x[F](\xi)$, we get for $\tau = 1$*

$$\begin{aligned}\chi^{n,1} \;=\;& i\varepsilon\, I_{x_n}[S(D^-_{\Phi'}\odot S)](x_{n+1})\,\big(\sigma_e y(x_n) + \sigma_i y(x_{n+1}\big)\\ &- i\varepsilon\,\mathcal{I}^1_{x_n}(x_{n+1})\,\big(\sigma_e G(x_{n+1})\,y(x_n) + \sigma_i G(x_n)\,y(x_{n+1})\big)\,.\end{aligned}$$

The first iterated integral $\mathcal{I}^1_{x_n}(x_{n+1})$ is already approximated by the Corollary 6.6.11 scheme and hence we only have to find a suitable quadrature for the other integral that shows up in $\chi^{n,1}$. Since $\chi^{n,1}$ is multiplied by $\varepsilon^{2\vartheta_1+1}$ we only need a second order approximation with respect to h_n. Thus a first order approximation with respect to h_n of the integrand is enough. Since the integral is highly oscillatory we use the technique from § 6.3 to find a quadrature.

Corollary 6.6.19 (modified first order scheme)**.** *Let $(1,\kappa,\iota,m)$ be the set of parameters of our OSM, such that*

$$\begin{aligned}\mu_0 &= \mu_0^*\,, \qquad |m_0| = m_0^*\,,\\ \mu_1 &= \mu_1^*\,, \qquad |m_1| = m_1^*\,,\end{aligned}$$

hold for all three types of support abscissas and let the matrices A_n, B_n and the vectors w^n, v^n be given by our Numerical Scheme 3. Furthermore let (on each subinterval $[x_n, x_{n+1}]$) the function $\widehat{P}_1 \in C^0(\Omega, \mathbb{C}^{d\times d})$ be, such that

$$I_\xi[\widehat{P}_1](x) \;=\; E^*_\varepsilon(s)\widehat{P}^\diamond_1(s)E_\varepsilon(s)\big|^x_{s=\xi}\,,$$

where $\widehat{P}^\diamond_1$ is C^0–bounded independently of ε and such that

$$\|\big(S(D^-_{\Phi'}\odot S)\big)(x) - \widehat{P}_1(x)\| \;=\; \mathcal{O}(\varepsilon^0 h_n^1)\,.$$

Than the local error err^n *of the modified first order scheme*

$$\big(\mathrm{Id} + \sigma_i B^\star_n\big)z^{n+1} + \lambda\sigma_i\, w^n \;=\; \big(\mathrm{Id} + \sigma_e A^\star_n\big)z^n + \lambda\sigma_e\, v^n\,,$$

with

$$\begin{aligned}A^\star_n &:= A_n + i\varepsilon\rho^2\big(I_{x_n}[\widehat{P}_1](x_{n+1}) - Q^1_n G(x_{n+1})\big)\,,\\ B^\star_n &:= B_n - i\varepsilon\rho^2\big(I_{x_n}[\widehat{P}_1](x_{n+1}) - Q^1_n G(x_n)\big)\,,\end{aligned}$$

is of order $\mathcal{O}(\varepsilon^{2\vartheta_1+1}h_n^2)$. See Numerical Scheme 3 (p.137ff) for the definition of Q^1_n and see Lemma 6.6.14 for the definition of G.

Remark 6.6.20. *The simplest choice for $\widehat{P}_1$ is a generalized "constant" Hermite interpolation polynomial, i. e.*

$$\widehat{P}_1(x) \;:=\; D_{\Phi'}(x)\odot D^-_{\Phi'}(\zeta)\odot\big(S(D^-_{\Phi'}\odot S)\big)(\zeta)\,,$$

with $\zeta \in [x_n, x_{n+1}]$. In this case it holds

$$I_\xi[\widehat{P}_1](x) \;=\; i\varepsilon E^\varepsilon_\Phi(t)\big|^x_{t=\xi}\odot D^-_{\Phi'}(\zeta)\odot\big(S(D^-_{\Phi'}\odot S)\big)(\zeta)\,.$$

The first order scheme in Corollary 6.6.19 shows that it is possible to improve the ε–order of the truncation error and hence the asymptotic order of the OSM. This is, of course, not limited to this special case and can be done for every $\tau \in \mathbb{N}$. Let us fix a OSM (from Corollary 6.6.11) we want to modify, which we shall refer to as the *underlying (one–step) method*. In the sequel we discuss two approaches of modifying the underlying method.

(i) As we have seen in Corollary 6.6.19, we can use a suitable quadrature for the highly oscillatory integral

$$\big(\mathcal{I}_\xi^{\tau-1} I_\xi[S(D_{\Phi'}^- \odot S)]\big)(x)\,, \tag{6.79}$$

(where ξ, x are wild cards for x_n, x_{n+1} and vice versa) to reduce the truncation error. With the idea of Lemma 6.2.5 combined with Corollary 6.3.4 one can construct such a quadrature. Let us define

$$\widehat{S}_1 \;:=\; S(D_{\Phi'}^- \odot S)\,.$$

Since the integral (6.79) is multiplied by $\rho^{\tau+1}\varepsilon$ (cf. definition of $\chi^{n\tau}$ in Lemma 6.6.14), we only need an $\mathcal{O}(\varepsilon^0 h_n^{\tau+1})$ approximation. Hence it is enough to approximate $\widehat{S}_1$ up to $\mathcal{O}(h_n)$. Thus let $\widehat{P}_1$ be, such that

$$\|\widehat{S}_1 - \widehat{P}_1\|_\infty \;\le\; c\,h_n\,,$$

with a constant $c \ge 0$ independently of ε, h_n and such that

$$I_\xi[\widehat{P}_1](x) \;=\; \big(E_\varepsilon^* \widehat{P}_1^\diamond E_\varepsilon\big)(t)\big|_{t=\xi}^x\,,$$

with $\widehat{P}_1^\diamond$ C^τ–bounded independently of ε. For example choose the function $\widehat{P}_1$ as in Remark 6.6.20. We further compute

$$\begin{aligned}
&\big(\mathcal{I}_\xi^{\tau-1} I_\xi[\widehat{S}_1]\big)(x)\\
&\quad= \big(\mathcal{I}_\xi^{\tau-1} I_\xi[\widehat{P}_1]\big)(x) - \big(\mathcal{I}_\xi^{\tau-1} I_\xi[\widehat{P}_1 - \widehat{S}_1]\big)(x)\\
&\quad= \big(\mathcal{I}_\xi^{\tau-2} I_\xi[S\widehat{P}_1^\diamond]\big)(x) \;-\; \mathcal{I}_\xi^{\tau-1}(x)\, E_\Phi^\varepsilon(\xi) \odot \widehat{P}_1^\diamond(\xi) \;+\; \mathcal{O}(\varepsilon^0 h_n^{\tau+1})\,.
\end{aligned}$$

To approximate $\mathcal{I}_\xi^{\tau-1}(x)$ one can use the quadrature of the underlying OSM. Now we can repeat the previous steps to approximate the first integral. The only difference to the first cycle is the approximation order of $\widehat{P}_2$. This has to be increased by one. Thus we can describe the whole procedure with the following loop, where $j = 1, \dots, \tau$.

(a) Choose the function $\widehat{P}_j$ such that

$$\|\widehat{S}_j - \widehat{P}_j\|_\infty \;\le\; c\,h_n^j\,,$$

with a constant $c \ge 0$ independently of ε and h_n, and such that

$$I_\xi[\widehat{P}_j](x) \;=\; \big(E_\varepsilon^* \widehat{P}_j^\diamond E_\varepsilon\big)(t)\big|_{t=\xi}^x\,,$$

with $\widehat{P}_j^\diamond$ C^τ–bounded independently of ε.

(b) It follows

$$\begin{aligned}&\big(\mathcal{I}_\xi^{\tau-j} I_\xi[\widehat{S}_j]\big)(x)\\ &\quad= \big(\mathcal{I}_\xi^{\tau-j} I_\xi[\widehat{P}_j]\big)(x) - \big(\mathcal{I}_\xi^{\tau-j} I_\xi[\widehat{P}_j - \widehat{S}_j]\big)(x)\\ &\quad= \big(\mathcal{I}_\xi^{\tau-j-1} I_\xi[S\widehat{P}_j^\diamond]\big)(x) \;-\; \mathcal{I}_\xi^{\tau-j}(x)\, E_\Phi^\varepsilon(\xi) \odot \widehat{P}_j^\diamond(\xi) \;+\; \mathcal{O}(\varepsilon^0 h_n^{\tau+1})\,.\end{aligned}$$

(c) Set $\widehat{S}_{j+1} := S\widehat{P}_j^\diamond$ and continue with (i).

We see that the numerical effort to approximate $\chi^{n,\tau}$ growth with τ. If one uses the quadrature form Corollary 6.3.4, one has to solve the same number of interpolation problems as for the underlying scheme. But with lower order of the generalize Hermite interpolation problem. Nevertheless we only gain a little benefit from the data computed for the underlying OSM.

(ii) Another ansatz to increase the order of the OSM with respect to ε is as follows. Corollary 6.6.15 yields

$$\begin{aligned}\mathrm{err}_{\mathrm{trunc}}^{n,\tau} &= \varepsilon^{\vartheta_1(\tau+1)}\big(\sigma_e\, \mathcal{I}_{x_n}^{\tau+1}(x_{n+1}) z(x_n) \;-\; \sigma_i\, \mathcal{I}_{x_{n+1}}^{\tau+1}(x_n) z(x_{n+1})\big)\\ &\quad + \mathcal{O}\big(\varepsilon^{\vartheta_1(\tau+1)+1} h_n^{\tau+1}\big)\,.\end{aligned}$$

Thus we simply have to incorporate a suitable approximation of the right hand side vector in our underlaying OSM. This can be done by solving one additional generalized Hermite interpolation problem. In the sequel we use the notation and quantities from the Numerical Scheme 3 (cf. p.137ff). Let $m_{\tau+1,1}, \ldots, m_{\tau+1,\kappa}$ be additional multiplicities (for all three types of intervals), such that $|m_{\tau+1}| \geq \tau$. Than replace in Numerical Scheme 3 (i) (the first loop) and (iii) τ by $\tau+1$. I. e. we additionally have to make the following computations:

(a) Compute the unique solution

$$P_{\tau+1}(x) \;=\; M'(x,\varepsilon) \odot \sum_{l=0}^{|m_{\tau+1}|-1} K_{\tau+1,l}(\varepsilon) \odot M(x,\varepsilon)^{\odot l}$$

of the generalized Hermite interpolation problem

$$P_{\tau+1}^{(k)}(\xi_l^n) \;=\; S_{\tau+1}^{(k)}(\xi_l^n)\,, \qquad k = 0, \ldots, m_{\tau+1,l} - 1,\;\; l = 1, \ldots, \kappa\,,$$

with $S_{\tau+1} = SP_j^\diamond$.

(b) compute $P_{\tau+1}^\diamond$ by Remark 6.3.3

(c) compute $Q_n^j = I_{x_n}[P_j](x_{n+1})$ by (6.40), i. e. (cf. Remark 6.3.3)

$$Q_n^j \;=\; E_\Phi^\varepsilon(x_{n+1}) \odot P_j^\diamond(x_{n+1}) \;-\; E_\Phi^\varepsilon(x_n) \odot P_j^\diamond(x_n)\,.$$

(d) Set

$$\begin{aligned}A_n^\star &= A_n \;+\; \rho^{\tau+1} \sum_{k=1}^{\tau+1} Q_n^k C_{n,\alpha}^{\tau+1-k}\\ B_n^\star &= B_n \;-\; \rho^{\tau+1} \sum_{k=1}^{\tau+1} Q_n^k C_{n,\beta}^{\tau+1-k}\,.\end{aligned}$$

Hence the modified scheme is a $\tau+1$ scheme (provided $f=0$) with a not maximized spatial convergence order.

Remark 6.6.21. *As we have seen it is possible to modify our OSM, such that we gain the local error estimate*

$$\| \operatorname{err}^n \| \quad \leq \quad c\, \varepsilon^{2\vartheta_1+1}\, h_n^{\tau+1}\,.$$

Hence even in the "critical" case $\vartheta_1 = 0$ the modified schemes yields the right convergence behavior as $\varepsilon \to 0$ (cf. Remark 6.2.3). Due to Lemma 6.6.14 this holds for all schemes of this section, but only the modified ones additionally guarantee spatial convergence. For $\vartheta_1 \geq 1$ there is no need to construct a modified scheme, if $\tau \geq 2$.

6.7 Convergence

For this section let us fix one numerical method with parameter (τ, κ, ι, m). In the sequel we refer to it as the OSM. Let $a = x_{n_a} < \cdots < x_{n_b} = b$ be a grid. We define the *global step size* h as

$$h \quad := \quad \max_{n_a \leq n \leq n_b-1} h_n \quad = \quad \max_{n_a \leq n \leq n_b-1} (x_{n+1} - x_n)\,.$$

Furthermore, (in the sequel) we assume that the grids we consider are chosen, such that $x_0 = \overline{x}$. Existence and uniqueness of a solution is guaranteed under certain (weak) assumptions on the grid.

Lemma 6.7.1. *Let Assumptions 6–12 hold. There is a constant $h_0 > 0$ independently of ε, such that for all grids $a = x_{n_a} < \cdots < x_{n_b} = b$ with $0 < h < h_0$ the OSM has a unique solution. Further exists a constant $c_s \geq 0$ independently of ε and n, such that*

$$\|A_n\|, \|B_n\| \;\leq\; c_s \rho h_n \;\leq\; c_s \rho h < 1\,.$$

Hence the matrices $\mathrm{Id} + \sigma_e A_n$ *and* $\mathrm{Id} + \sigma_i B_n$ *are regular for* $n \in \{n_a, \ldots, n_b - 1\}$.

Proof. The OSM (6.51) reads

$$(\mathrm{Id} + \sigma_i B_n) z^{n+1} \,+\, \sigma_i\, w^n \quad = \quad (\mathrm{Id} + \sigma_e A_n) z^n \,+\, \sigma_e\, v^n\,.$$

Assume that the matrices

$$\mathrm{Id} + \sigma_i B_n \quad \text{and} \quad \mathrm{Id} + \sigma_e A_n$$

are regular for all $n \in \{n_a, \ldots, n_b - 1\}$. Then we can split the coupled system of equations into two subproblems.

(i) For $n_a \leq n \leq 0$ we write

$$z^n \quad = \quad (\mathrm{Id} + \sigma_i A_n)^{-1} \big((\mathrm{Id} + \sigma_e B_n) z^{n+1} \,-\, \sigma_e\, v^n \,+\, \sigma_i\, w^n \big)\,,$$

(ii) and for $0 \leq n \leq n_b - 1$

$$z^{n+1} \quad = \quad (\mathrm{Id} + \sigma_i B_n)^{-1} \big((\mathrm{Id} + \sigma_e A_n) z^n \,+\, \sigma_e\, v^n \,-\, \sigma_i\, w^n \big)\,.$$

For both problems we have the initial condition $z^0 = z(x_0) = z(\xi)$. Since (i) and (ii) are explicit difference equations, they have unique solutions, which are compatible at $n = 0$. Hence we get existence of a unique OSM solution.

A sufficient criteria for regularity of $\mathrm{Id} + \sigma_e A_n$ and $\mathrm{Id} + \sigma_i B_n$ is that

$$\|A_n\| < 1 \quad \text{and} \quad \|B_n\| < 1$$

holds for all $n \in \{n_a, \dots, n_b - 1\}$ (cf. [68] p. 188). Due to Proposition 6.5.4 there exists a constant c independently of ε, h_n, and n, such that for all indices $n \in \{n_a, \dots, n_b - 1\}$

$$\|A_n\|, \|B_n\| \;\leq\; c\,\rho\,h_n \;\leq\; c\,\rho\,h\,.$$

Hence we set $h_0 = (c\rho)^{-1}$. □

Proposition 6.7.2. *Let Assumptions 6–12 hold. Let z be the unique solution of the IVP (6.16) and let $a = x_{n_a} < \cdots < x_{n_b} = b$ be a grid with $0 < h < h_0$. Furthermore we denote the unique solution of the OSM from Lemma 6.7.1 by $z^{n_a}, \dots, z^{n_b}$. Then there are constants $c_e, c_s, \gamma_0, \gamma_1 \geq 0$ independently of ε, such that for all $n \in \{n_a, \dots, n_b - 1\}$*

$$\|z(x_n) - z^n\| \;\leq\; e^{c_s \rho (x_n - x_0)} \left(\|\eta^0\| \,+\, c_e\,(x_n - x_0)\,\theta(\varepsilon, h)\,h^\tau \right),$$

with

$$\begin{aligned} \theta(\varepsilon, h) \;=\; & \lambda \min\left(1,\, \gamma_0 \left(\frac{\varepsilon}{h}\right)^{\mu_0+1}\right) h^{|m_0|-\tau} \\ & + \rho \min\left(1,\, \gamma_1 \left(\frac{\varepsilon}{h}\right)^{\mu_1+1}\right) h^{|m_1|-\tau} \,+\, \varepsilon\rho^2\,. \end{aligned} \tag{6.80}$$

Proof. By assumptions and Lemma 6.7.1 there exists a unique solution of the OSM and the matrices

$$\mathrm{Id} + \sigma_e A_n \quad \text{and} \quad \mathrm{Id} + \sigma_i B_n$$

are regular for $n \in \{n_a, \dots, n_b - 1\}$. For $n > 0$ we reformulate the OSM as an explicit scheme (as in the proof of Lemma 6.7.1), i. e. it holds for all $n \in \{0, \dots, n_b - 1\}$

$$z^{n+1} \;=\; (\mathrm{Id} + \sigma_i B_n)^{-1} \big((\mathrm{Id} + \sigma_e A_n) z^n \,+\, \sigma_e v^n \,-\, \sigma_i w^n \big)\,.$$

From (6.31) of Numerical Scheme 2 (cf. p.127f) we know that for all indices $n \in \{0, \dots, n_b - 1\}$ it holds

$$z(x_{n+1}) \;=\; (\mathrm{Id} + \sigma_i B_n)^{-1} \big((\mathrm{Id} + \sigma_e A_n) z(x_n) \,+\, \sigma_e v^n \,-\, \sigma_i w^n \,+\, \mathrm{err}^n \big)\,.$$

Hence the quantity $\eta^n := z(x_n) - z^n$ solves the inhomogeneous explicit difference equation

$$\eta^{n+1} \;=\; (\mathrm{Id} + \sigma_i B_n)^{-1} \big((\mathrm{Id} + \sigma_e A_n) \eta^n \,+\, \mathrm{err}^n \big) \;=:\; \widehat{A}_n\,\eta^n \,+\, \widehat{\mathrm{err}}^n\,.$$

This yields (by induction)

$$\eta^n \;=\; \prod_{j=0}^{n-1} \widehat{A}_j\,\eta^0 \,+\, \sum_{j=0}^{n-1} \Big(\prod_{k=j+1}^{n-1} \widehat{A}_k \Big)\, \widehat{\mathrm{err}}^{\,j}\,. \tag{6.81}$$

Furthermore, by Lemma 6.7.1 it holds $\|A_j\|, \|B_j\| \le c_s\rho h_n \le c_s\rho h < 1$. This yields ($\sigma_i \in [0,1]$)

$$\begin{aligned} \|(\mathrm{Id} + \sigma_i B_j)^{-1}\| \;&=\; \left\| \sum_{k=0}^{\infty} (-1)^k \sigma_i^k B_j^k \right\| \;\le\; \sum_{k=0}^{\infty} (\sigma_i c_s \rho h_j)^k \\ &\le\; \frac{1}{1-\sigma_i c_s \rho h_j} \;\le\; \frac{1}{1-\sigma_i c_s \rho h}\,. \end{aligned}$$

Since $\sigma_e = 1 - \sigma_i$, we get the following estimate for $\widehat{A}_j$:

$$\begin{aligned} \|\widehat{A}_j\| \;&\le\; \|(\mathrm{Id} + \sigma_i B_j)^{-1}\|\, \|(\mathrm{Id} + \sigma_e A_j)\| \\ &\le\; \frac{1 + \sigma_e c_s \rho h_j}{1 - \sigma_i c_s \rho h_j} \;=\; 1 \,+\, \frac{c_s \rho h_j}{1 - \sigma_i c_s \rho h_j} \;\le\; 1 \,+\, \frac{c_s \rho h_j}{1 - \sigma_i c_s \rho h}\,. \end{aligned}$$

By Corollary 6.6.7 exist $c_e, \gamma_0, \gamma_1 \ge 0$ independently of ε and n, such that

$$\|\widehat{\mathrm{err}}^{\,j}\| \;\le\; \frac{c_e\, \theta(\varepsilon, h_n)\, h_n^{\tau+1}}{1 - \sigma_i c_s \rho h}\,.$$

The function θ is given by

$$\begin{aligned} \theta(\varepsilon, h_n) \;=\;& \lambda \,\min\left(1,\, \gamma_0 \left(\frac{\varepsilon}{h_n}\right)^{\mu_0+1}\right) h_n^{|m_0|-\tau} \\ &+ \rho \min\left(1,\, \gamma_1 \left(\frac{\varepsilon}{h_n}\right)^{\mu_1+1}\right) h_n^{|m_1|-\tau} \,+\, \varepsilon\rho^2\,. \end{aligned}$$

Since $h_n \le h$, it holds (all exponents are non–negative)

$$\begin{aligned} \theta(\varepsilon, h_n) h_n^{\tau} \;=\;& \lambda \,\min\left(h_n^{\mu_0+1},\, \gamma_0\, \varepsilon^{\mu_0+1}\right) h_n^{|m_0|-(\mu_0+1)} \\ &+ \rho \min\left(h_n^{\mu_1+1},\, \gamma_1\, \varepsilon^{\mu_1+1}\right) h_n^{|m_1|-(\mu_1+1)} \,+\, \varepsilon\rho^2\, h_n^{\tau} \\ \le\;& \lambda \,\min\left(h^{\mu_0+1},\, \gamma_0\, \varepsilon^{\mu_0+1}\right) h^{|m_0|-(\mu_0+1)} \\ &+ \rho \min\left(h^{\mu_1+1},\, \gamma_1\, \varepsilon^{\mu_1+1}\right) h^{|m_1|-(\mu_1+1)} \,+\, \varepsilon\rho^2\, h^{\tau} \\ =\;& \theta(\varepsilon, h)\, h^{\tau}\,. \end{aligned} \tag{6.82}$$

Furthermore we set $\widehat{c}_s := c_s(1 - \sigma_i c_s \rho h)^{-1}$ and $\widehat{c}_e := c_e(1 - \sigma_i c_s \rho h)^{-1}$. With

the triangle inequality we derive from (6.81):

$$\begin{aligned}
\|\eta^n\| &\leq \prod_{j=0}^{n-1} \|\widehat{A}_j\| \, \|\eta^0\| + \sum_{j=0}^{n-1} \prod_{k=j+1}^{n-1} \|\widehat{A}_k\| \, \|\widehat{\mathrm{err}}^j\| \\
&\leq \prod_{j=0}^{n-1} \left(1 + \widehat{c}_s \rho h_j\right) \|\eta^0\| + \sum_{j=0}^{n-1} \prod_{k=j+1}^{n-1} \left(1 + \widehat{c}_s \rho h_k\right) \widehat{\mathrm{err}}^j \\
&\leq \prod_{j=0}^{n-1} e^{\widehat{c}_s \rho h_j} \, \|\eta^0\| + \sum_{j=0}^{n-1} \left(\prod_{k=j+1}^{n-1} e^{\widehat{c}_s \rho h_k} \right) \widehat{c}_e \, \theta(\varepsilon, h_j) \, h_j^{\tau+1} \\
&= e^{\widehat{c}_s \rho \sum_{j=0}^{n-1} h_j} \, \|\eta^0\| + \widehat{c}_e \sum_{j=0}^{n-1} h_j \left(e^{\widehat{c}_s \rho \sum_{k=j+1}^{n-1} h_k} \right) \theta(\varepsilon, h_j) \, h_j^{\tau}
\end{aligned}$$

By (6.82) we can replace $\theta(\varepsilon, h_j) h_j^\tau$ by $\theta(\varepsilon, h) h^\tau$. Since $h_j = x_{j+1} - x_j$ it holds

$$\sum_{k=j+1}^{n-1} h_k \;=\; x_n - x_{j+1} \;\leq\; x_n - x_0 \;=\; \sum_{j=0}^{n} h_j \,.$$

Thus $e^{\widehat{c}_s \rho \sum_{k=j+1}^{n-1} h_k} \leq e^{\widehat{c}_s \rho \sum_{j=0}^{n-1} h_j} = e^{\widehat{c}_s \rho (x_n - x_0)}$ and hence

$$\begin{aligned}
\|\eta^n\| &\leq e^{\widehat{c}_s \rho (x_n - x_0)} \left(\|\eta^0\| + \widehat{c}_e \, \theta(\varepsilon, h) h^\tau \sum_{j=0}^{n-1} h_j \right) \\
&\leq e^{\widehat{c}_s \rho (x_n - x_0)} \left(\|\eta^0\| + \widehat{c}_e \, (x_n - x_0) \, \theta(\varepsilon, h) \, h^\tau \right).
\end{aligned}$$

Due to definition of the grid we have $x_0 = \overline{x}$. For $n < 0$ we rephrase the problem as an explicit scheme for z^n and do the same computations as above. □

A consequence of Proposition 6.7.2 is the pointwise convergence of the OSM.

Corollary 6.7.3. *Let Assumptions 6–12 hold and let z be the unique solution of the IVP* (6.16)*. Additionally let $\xi \in [a, b]$ and let $\{x^r : r \in \mathbb{N}\}$ be a family of grids, with*

$$a = x^r_{n_a(r)} < \cdots < x^r_{n_b(r)} = b \,, \qquad h_r := \sup_{n_a^r \leq n \leq n_b^r} (x^r_{n+1} - x^r_n) < h_0 \,.$$

Here $h_0 \geq 0$ is the constant from Lemma 6.7.1. Furthermore we assume that $\lim_{r\to\infty} h_r = 0$ and that for every $r \in \mathbb{N}$ there is a $N_r \in \{n_a(r), n_b(r)\}$, such that $x_{N_r} = \xi$.

For $r \in \mathbb{N}$ let $z^{r,n_a(r)}, \ldots, z^{r,n_b(r)}$ be the unique solution of the OSM from Lemma 6.7.1 corresponding to the grid x^r. We assume that there exists a $c_0 \geq 0$ independently of ε and r, such that

$$\|z(\overline{x}) - z^0\| \;\leq\; c_0 \, h_r^\tau \,.$$

Let

$$z(\xi, r) \;:=\; z^{r, N_r} \,.$$

There are constants $\widehat{c}_e, \widehat{c}_s, \gamma_0, \gamma_1 \geq 0$ *independently of* ε*, such that for all* $r \in \mathbb{N}$

$$\|z(x) - z(x,r)\| \quad \leq \quad \widehat{c}_s \left(c_0 \, + \, \widehat{c}_e \; \theta(\varepsilon, h_r)\right) h_r^\tau \, .$$

The function θ *is given by* (6.80)*. Hence the OSM converges as* $h_r \to 0$*.*

Proof. By Proposition 6.7.2 it holds for all $r \in \mathbb{N}$

$$\|z(\xi) - z(\xi,r)\| \quad \leq \quad e^{c_s \rho(\xi - x_0)} \left(\|z(\overline{x}) - z(x_0,r)\| \, + \, c_e \, (\xi - x_0) \, \theta(\varepsilon, h_r) \, h_r^\tau \right).$$

The constants are independent of r. Thus

$$\|z(\xi) - z(\xi,r)\| \quad \leq \quad e^{c_s \rho(b-a)} \left(c_0 \, + \, c_e \, (b-a) \, \theta(\varepsilon, h_r)\right) h_r^\tau \, .$$

□

Chapter 7

Numerical experiments for the one–step method

In this chapter we present some numerical results for the efficient one–step methods from § 6.4. The first section § 7.1 is dedicated for the introduction of a reference example from [54]. In the article it is used to illustrate the performance of certain integrators discussed there. The problems they are designed for can be transformed, such that they fit into our setting from § 3.2. Hence the subsequently described problem from § 7.1 is an ideal candidate to compare our new efficient one–step methods with an existing method from literature. Numerical results for an explicit and the Crank–Nicolson like setting of our one–step method are discussed in § 7.2. They are compared with the *adiabatic midpoint–rule* from [54], which is a symmetric two step integrator of order $\mathcal{O}(\varepsilon^0 h^2)$.

The example discussed in § 7.3 is from [27]. It is used to illustrate the problems of the super–adiabatic transformation of lowest order close to avoided eigenvalue crossings of the matrix L from (3.21) (i. e. $\delta \ll 1$). The same problems appear for our WKB–type transformation from § 3.3. In the textbook it is mentioned that the problem was studied by Clarance Zener [75] in 1932. An alternative formulation (which is closer to its origin in quantum mechanics) of the example can be found in [73]. Hence we start § 7.3 with the derivation of the example from [27] from the more general problem stated in [73]. Afterwards we derive a formula for T_ε from § 3.3 for $n = 1$. Furthermore we are able to compute an explicit expression for $\|\mathcal{S}_1\|$. This quantity is of interest, because it is the crucial variable of the step size algorithm discussed in § 4.4. In § 7.4 we solve the Zener problem from § 7.3 with our step size control algorithm from § 4.4. Here we use the same setting as in [27], in order to generate comparable results to the textbook ones.

7.1 A vector valued reference example by Lorenz et al. [54]

We use this section to introduce a reference example from [54]. Let $\delta > 0$ be a real parameter and let

$$A(x) \;=\; \begin{pmatrix} x+3 & \delta \\ \delta & 2x+3 \end{pmatrix}^2 .$$

A diagonalization $A = U^*\Lambda U$ is given by (cf. [54])

$$\Lambda(x) \;=\; \begin{pmatrix} \frac{3}{2}x+3+\frac{1}{2}\sqrt{x^2+4\delta^2} & 0 \\ 0 & \frac{3}{2}x+3-\frac{1}{2}\sqrt{x^2+4\delta^2} \end{pmatrix}^2 ,$$

$$U(x) \;=\; \begin{pmatrix} \cos\xi(x) & \sin\xi(x) \\ -\sin\xi(x) & \cos\xi(x) \end{pmatrix} \quad \text{with } \xi(x) \;=\; \tfrac{\pi}{4}+\tfrac{1}{2}\arctan(\tfrac{x}{2\delta}) .$$

Then the initial value problem we shall solve is given by

$$\begin{aligned} \varepsilon^2\Psi''(x) \,+\, A(x)\Psi(x) &= 0 , \\ \Psi(x_0) &= \Psi_0 \in \mathbb{C}^2 , \\ \Psi'(x_0) &= \Psi_1 \in \mathbb{C}^2 . \end{aligned}$$

for some initial conditions and $x \in I := [-1,1]$. (For the numerical examples we use $\Psi_0 = (1,0)^T$ and $\Psi_1 = (0,1)^T$.) Thus the equivalent first order IVP as derived in § 2.2.1 reads

$$u'(x) \;=\; \frac{i}{\varepsilon}\begin{pmatrix} \Lambda^{\frac{1}{2}}(x) & 0 \\ 0 & -\Lambda^{\frac{1}{2}}(x) \end{pmatrix} u(x) \;+\; B(x)\,u(x) \tag{7.1}$$

$$u(x_0) \;=\; \tfrac{1}{\sqrt{2}}\begin{pmatrix} i & 1 \\ 1 & i \end{pmatrix} \otimes U(x_0)\begin{pmatrix} A^{\frac{1}{2}}\Psi_0 \\ \varepsilon\Psi_1 \end{pmatrix}, \tag{7.2}$$

with

$$B \;=\; \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes (U'U^*) + \frac{1}{2}\begin{pmatrix} 1 & i \\ -i & 1 \end{pmatrix} \otimes \big(UA^{\frac{1}{2}\prime}A^{-\frac{1}{2}}U^*\big) .$$

A straight forward calculation shows

$$\big(U'U^*\big)(x) \;=\; \xi'(x)\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad \text{with} \quad \xi'(x) \;=\; \frac{\delta}{x^2+4\delta^2} .$$

Using the identity (2.30) from § 2.2 we further compute:

$$\begin{aligned} &\big(UA^{\frac{1}{2}\prime}A^{-\frac{1}{2}}U^*\big) \\ &= -\xi'\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} + \xi'\begin{pmatrix} 0 & (\frac{\lambda_1}{\lambda_2})^{\frac{1}{2}} \\ -(\frac{\lambda_2}{\lambda_1})^{\frac{1}{2}} & 0 \end{pmatrix} + (\Lambda^{\frac{1}{2}})'\Lambda^{-\frac{1}{2}} \\ &= \Big[\xi'\begin{pmatrix} 0 & -\lambda_2^{\frac{1}{2}} \\ \lambda_1^{\frac{1}{2}} & 0 \end{pmatrix} + \xi'\begin{pmatrix} 0 & \lambda_1^{\frac{1}{2}} \\ -\lambda_2^{\frac{1}{2}} & 0 \end{pmatrix} + (\Lambda^{\frac{1}{2}})'\Big]\Lambda^{-\frac{1}{2}} \\ &= \Big[\xi'(\lambda_1^{\frac{1}{2}}-\lambda_2^{\frac{1}{2}})\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + (\Lambda^{\frac{1}{2}})'\Big]\Lambda^{-\frac{1}{2}} . \end{aligned}$$

There exists a representation in terms of ξ and Λ. To derive this we need the following identities, which hold since $-\frac{\pi}{2} < \arctan\frac{x}{2\delta} < \frac{\pi}{2}$.

$$\begin{aligned}\cos^2\xi(x) &= \frac{1}{2}\frac{\sqrt{x^2+4\delta^2}-x}{\sqrt{x^2+4\delta^2}}\,,\\ \sin^2\xi(x) &= \frac{1}{2}\frac{\sqrt{x^2+4\delta^2}+x}{\sqrt{x^2+4\delta^2}}\,,\\ \sin\xi(x)\cos\xi(x) &= \frac{\delta}{\sqrt{x^2+4\delta^2}}\,.\end{aligned}$$

This yields

$$\xi'(x)(\lambda_1^{\frac{1}{2}}(x)-\lambda_2^{\frac{1}{2}}(x)) = \frac{\delta}{\sqrt{x^2+4\delta^2}} = \sin\xi(x)\cos\xi(x)\,.$$

Furthermore it holds

$$(\lambda_1^{\frac{1}{2}})'(x) = \frac{3\sqrt{x^2+4\delta^2}+x}{2\sqrt{x^2+4\delta^2}} = 1+\sin^2\xi(x)\,.$$

Analog we find $(\lambda_2^{\frac{1}{2}})' = 1+\cos^2\xi$, which yields

$$\begin{aligned}UA^{\frac{1}{2}\prime}A^{-\frac{1}{2}}U^* &= \begin{pmatrix}1+\sin^2\xi & \sin\xi\cos\xi\\ \sin\xi\cos\xi & 1+\cos^2\xi\end{pmatrix}\Lambda^{-\frac{1}{2}}\\ &= \begin{pmatrix}1+\sin^2\xi & \frac{1}{2}\sin 2\xi\\ \frac{1}{2}\sin 2\xi & 1+\cos^2\xi\end{pmatrix}\Lambda^{-\frac{1}{2}}\,.\end{aligned}$$

This is a nice compact formula which can easily be implemented in *Matlab*. Furthermore we get with respect to the notation of § 3.3

$$L(x) = \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}\otimes\begin{pmatrix}\frac{3}{2}x+3+\frac{1}{2}\Delta(x) & 0\\ 0 & \frac{3}{2}x+3-\frac{1}{2}\Delta(x)\end{pmatrix},$$

where we set

$$\Delta(x) := \sqrt{x^2+4\delta^2}\,.$$

Using the function Δ we can also write

$$UA^{\frac{1}{2}\prime}A^{-\frac{1}{2}}U^* = \frac{1}{2\Delta(x)}\begin{pmatrix}3\Delta(x)+x & 2\delta\\ 2\delta & 3\Delta(x)-x\end{pmatrix}\Lambda^{-\frac{1}{2}}(x)\,.$$

In our *Matlab* code we use this representation, since also Λ can be built up from the function Δ.

7.2 Convergence behavior

In this section[1] we illustrate the convergence behavior (stated in Proposition 6.7.2) of our numerical approximation to the solution z of the IVP (6.16). The

[1]The author published parts of this section in [25].

results are derived with the methods from § 7.5. The procedure how to approximate $T_\varepsilon, R_\varepsilon, S, \Phi$ is discussed in § 4. Anyhow, the numerical integration of the phase Φ usually incurs an additional error for the original, oscillatory function u from (3.23) or (6.1) (cf. § 4.3). This situation is the same also for scalar ODEs (cf. Th. 3.1 in [4]).

We shall compare our one step method (OSM) to the *Adiabatic Midpoint Rule* (AMPR) from [54]. This integrator is a space–symmetric two–step method, which yields a convergence error of order $\mathcal{O}(\varepsilon^0 h^2)$ for the function η defined in Lemma 3.2.2. If we want to have the same error behavior for the original function u, we also have to impose the step size restriction $h \leq \sqrt{\varepsilon}$, if we use the Simpson rule to approximate the matrix valued phase Φ, see Remark 4.3.1. Using a higher order quadrature rule for Φ would weaken this restriction on h.

Let us choose a family of equidistant grids. Let $g \in \mathbb{N}$ and define for $n = 0, \ldots, 2^g =: N_g$ the grid points

$$x_n^g := a + nh_g \quad \text{with} \quad h_g := \frac{b-a}{N_g}\,.$$

For integers $g_1 < g_2$

$$h_{g_1} \quad = \quad \frac{b-a}{2^{g_1}} = \frac{b-a}{2^{g_2}}\frac{2^{g_2}}{2^{g_1}} = 2^{g_2-g_1}\, h_{g_2}$$

and hence it holds for all $x_n^{g_1}$ with $n = 0, \ldots, N_{g_1}$:

$$x_n^{g_1} = a + n2^{g_2-g_1}\, h_{g_2} = x_m^{g_2} \quad \text{with} \quad m = n\, 2^{g_2-g_1}\,. \tag{7.3}$$

Thus the grid corresponding to g_1 is a (coarser) sub–grid of the g_2–grid. Hence no interpolation is needed when comparing solutions on two different grids. To generate error plots we fix a finite number of indices, e. g. $g = 2, \ldots, 16$, and use the numerical solution on the finest grid as reference solution. To illustrate the convergence behavior of the OSM (w.r.t. the step size h, and in dependence of ε) we shall give the relative L^1–error.

Let us denote the solutions corresponding to the grid g by z^g and denote the reference solution by z^*. By (7.3) we know that $z_n^g \approx z_{m_n}^*$, with

$$m_n \quad := \quad n\, 2^{g^*-g}\,.$$

Hence the (discrete) relative L^1–error is defined by

$$\operatorname{Err} z^g \;:=\; \frac{h_g \sum_{n=0}^{N_g} \left| z_n^g - z_{m_n}^* \right|}{\Sigma} \quad \text{with} \quad \Sigma \;:=\; h_{g^*} \sum_{j=0}^{N_{g^*}} \left| z_j^* \right|. \tag{7.4}$$

The quantity Σ is the discrete L^1–norm of z^*. Since h_g is reciprocal proportional to N_g (which is approximately the number of summands in (7.4) if $N_g \gg 1$), we can also interpret $\operatorname{Err} z^g$ as a scaled average error.

Figures 7.2–7.11 show the relative L^1–error of our OSM for z (or η or u) for the example discussed in § 7.1. It is already used in [54] to illustrate the performance of the AMPR. In all Figures we plot the relative L^1–error of the AMPR (for η) as reference curve. We use an explicit scheme (ES) (i. e. $\sigma_i = 0$) with $\tau = 2$ and the Crank–Nicolson like scheme (CNS) (cf. § 7.5), as well as different stages of discretization of S.

In Figure 7.1 we plot the theoretical error prediction of Proposition 6.7.2, with a fitted leading constant. Due to our experience with the quadratures (cf. § 5.4) we choose an ε–dependent constant c. This is of course not exactly the formulation from the Proposition, but this behavior is reflected quite well in Figure 7.2 (ES), 7.7 (CNS). Here we use almost exact values for the coefficients appearing in the IVP (6.16). They are derived via interpolation from the approximation of S, L, Φ for the finest grid as described in § 4.1. The interpolation is done with the Matlab function `interp1` (with the method 'pchip'; piecewise cubic interpolation). This is of course not necessary, but simpler to implement then figuring out the right indices by hand. Furthermore, in Figure 7.2–7.5, we also observe the error threshold at about 10^{-14}, probably resulting from the Matlab computations in double precision.

The graphs in Figure 7.12 are the relative L^1–errors of the variable z, computed with the Kane model of § 2.1.1. We used the following data: $a = 0$, $b = p(x) = 1$, $E = 2$, $V(x) = 10x(\frac{3}{4} - x)$, $E_g(x) = \frac{1}{2}\sin^2(2\pi x) + \frac{1}{2}$. As for Figure 7.2, 7.7 we use almost exact data for S.

For the simulation of Figure 7.2 (ES) and 7.7 (CNS) the coefficients S, L, Φ are approximated (as it will be done in practice) on the same grid that is used for the solution of the IVP. We use the algorithm described in § 4.1. For small values of ε one observes the influence of the approximation for rather large h.

In order to compare our results with the AMPR we transform z into η. The resulting errors are plotted in Figure 7.4–7.5 (ES), 7.9–7.10 (CNS). Since we do not use the exact transformation, the accuracy is reduced, but still significantly better than those of the AMPR. For the full discretized schemes (Figure 7.5, 7.10) we lose the asymptotic correctness with respect to ε. We observe quite good the fourth order convergence of the transformation. However, if T_0 is exactly given (as for the considered example), then the approximation of the variable η with our OSM is (globally) asymptotically correct, as we can see Figure 7.4, 7.9.

The error of the full discretized schemes for the original variable u is plotted in Figure 7.6 (ES) and 7.11 (CNS).

The numerical experiments confirm the theoretical results. We observe the $\mathcal{O}(\varepsilon^0 h^2)$ convergence behavior for the AMPR as discussed in [54]. So, the error of that scheme (for the variable η from Lemma 3.2.2) is uniform in ε, but it does not decrease as $\varepsilon \to 0$. However, our OSM shows an even better error behavior than predicted in Proposition 6.7.2. While for large step sizes h the graphs of the z–error behave like $\mathcal{O}(\varepsilon^3 h^0)$ (which coincides with the theoretical estimate), they seem to turn to an $\mathcal{O}(\varepsilon^2 h^2)$ behavior, if h gets small enough (see Fig. 7.3, 7.7, 7.12). This is a "better" convergence property than the predicted $\mathcal{O}(\varepsilon^1 h^2)$ behavior from Proposition 6.7.2. This behavior is also described in §3.3 of [4], and it is due to cancellation effects in successive integration steps. We also observed it in the numerical examples for the quadratures in § 5.4. The Figures 7.2–7.4, 7.7–7.9, 7.12 also illustrate the asymptotic correctness of our OSM as $\varepsilon \to 0$, even for rather large values of h.

The two methods (OSM and AMPR) use the same set of data from the original IVP (6.1) for u. Since the computation of the matrix valued functions L, B can be computational expensive (e.g. one has to derive the eigenvalues and eigenvectors for a large matrix in each step) compared to the computation of $\Phi, T_\varepsilon, R_\varepsilon, S_\varepsilon$, our OSM is an improvement of the AMPR on the level of the transformed quantities z and η.

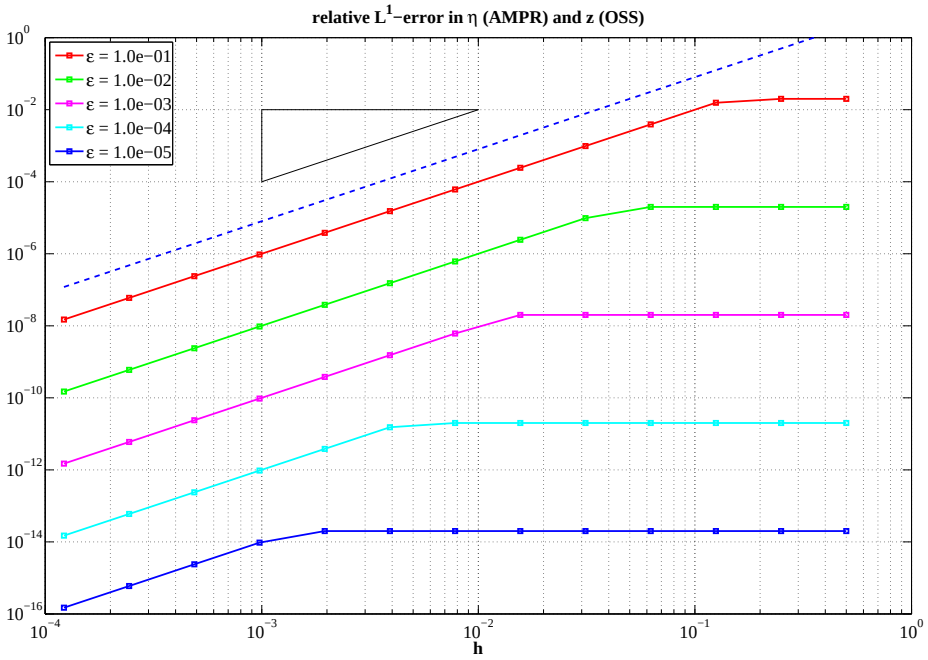

Figure 7.1: Plot of the functions $20\min(5\varepsilon^2 h^2, \varepsilon^3)$ (solid lines) and $8h^2$ (dashed lines) for different values of ε.

Both methods, the OSM and the AMPR are subject to the fact that (in general) the transformation back to the original variable u introduces an error of the order $\mathcal{O}(\varepsilon^{-1})$, as discussed in 4.3. This is due to the multiplication of z (and η) with the highly oscillatory matrix $E_\varepsilon(x) = \exp(\frac{i}{\varepsilon}\Phi(x))$. Since Φ is approximated with the Simpson rule (which yields an error of $\mathcal{O}(h^4)$ for Φ) we get an transformation error of $\mathcal{O}(\varepsilon^{-1}h^4)$. This explains the step size restriction mentioned in the beginning of this section. But if the matrix valued phase function Φ is exactly known[2], the error behavior of z, η carries over to u. In this situation our OSM yields much better results for u than the AMPR – with approximately the same numerical effort. One can observe the influence of the transformation quite good for the ES in Figure 7.6 and for the CNS 7.11. The AMPR is not affected by the "back" transformation, because its errors are larger then the induced transformation error.

7.3 An example of avoided eigenvalue crossing

The effects of avoided eigenvalue crossings are (shortly) discussed in [27] by an example which was studied by Clarence Zener in 1932 [75]. It is also discussed (with different focus) in [73]. Here the author considers the second order differential equation ($\alpha \in \mathbb{R}$, $\delta > 0$)

$$\varepsilon^2\psi''(x) + 2i\,\varepsilon\,\alpha\,x\,\psi'(x) + \delta^2\psi(x) = 0\,. \tag{7.5}$$

[2] E. g., piecewise linear functions V, E_g, p in the Kane model lead to an exactly integrable phase.

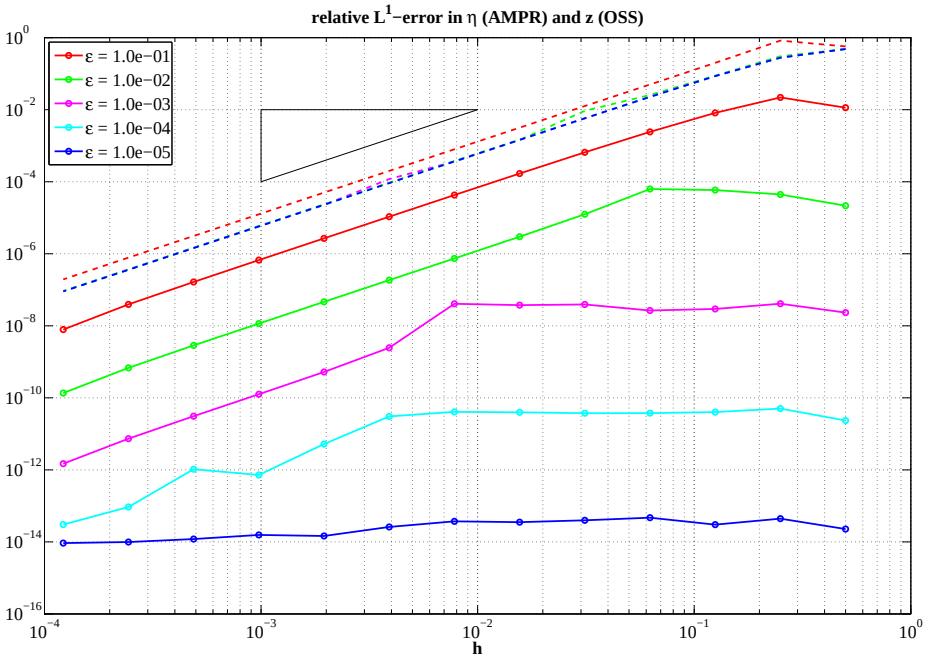

Figure 7.2: Relative L^1–error of the (explicit) OSM for z (solid lines) and the AMPR [54] for η (dashed lines) for different values of ε. "Exact" evaluation of S via interpolation is used.

We slightly changed the notation with respect to [73] in order to derive the example discussed in [27]. A general solution (computed with *Maple14*) of (7.5) can be expressed in terms of *Kummer functions* KummerM, KummerU (*Maple14* notation, see also *confluent hypergeometric function*) by

$$\begin{aligned}\psi(x) \quad = \quad & c_1\, x\, e^{-\frac{i\alpha x^2}{\varepsilon}}\, \mathrm{KummerM}\left(1+\tfrac{i\delta^2}{4\varepsilon\alpha}, \tfrac{3}{2}, \tfrac{i\alpha x^2}{\varepsilon}\right) \\ & + c_2\, x\, e^{-\frac{i\alpha x^2}{\varepsilon}}\, \mathrm{KummerU}\left(1+\tfrac{i\delta^2}{4\varepsilon\alpha}, \tfrac{3}{2}, \tfrac{i\alpha x^2}{\varepsilon}\right).\end{aligned} \tag{7.6}$$

Now we use the approach from § 2.2 to reformulate (7.5) as a first order system of differential equations. I. e. we set

$$\tilde{v}(x) \quad := \quad \begin{pmatrix} \delta\psi(x) \\ \varepsilon\psi'(x) \end{pmatrix},$$

which yields

$$\tilde{v}'(x) \quad = \quad \frac{1}{\varepsilon}\begin{pmatrix} 0 & \delta \\ -\delta & -2i\alpha x \end{pmatrix}\tilde{v}(x).$$

The final transformation

$$v(x) \quad = \quad \exp\left(\frac{i\alpha x^2}{2\varepsilon}\right)\begin{pmatrix} 0 & \frac{1+i}{\sqrt{2}} \\ \frac{1-i}{\sqrt{2}} & 0 \end{pmatrix}\tilde{v}(x)$$

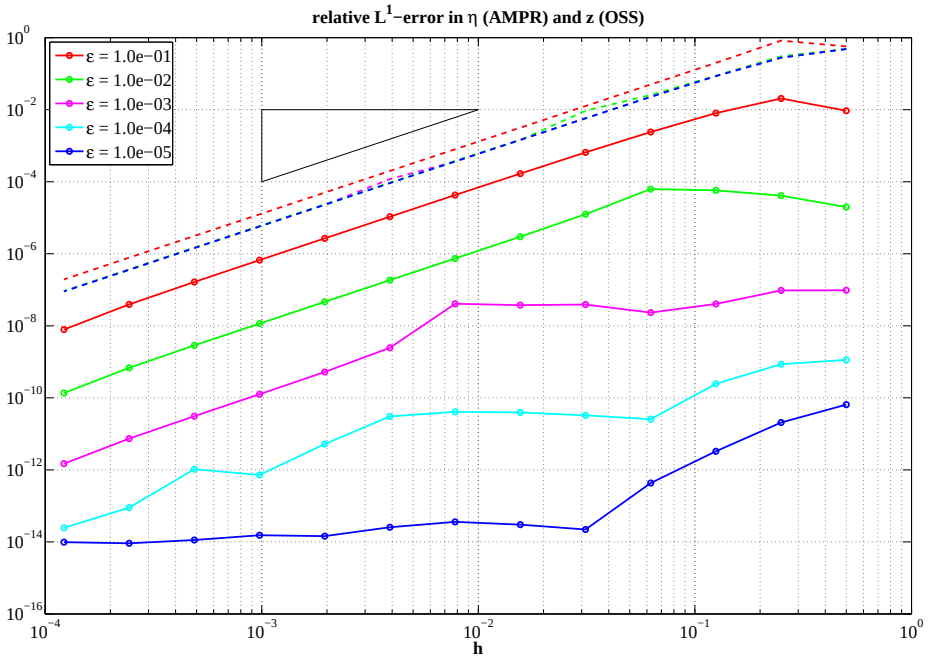

Figure 7.3: Relative L^1–error of the (explicit) OSM for z (solid lines) and the AMPR from [54] for η (dashed lines) for different values of ε. The function S is for every value of h separately approximated as described in § 4.

yields

$$v'(x) \;=\; -\frac{i}{\varepsilon}\begin{pmatrix} \alpha x & \delta \\ \delta & -\alpha x \end{pmatrix} v(x) \;=:\; \frac{i}{\varepsilon}\, A(x)\, v(x)\,,$$

which is the example from [27, p.535f], if we set $\alpha = 1$, what is assumed from now on. From the textbook we get $A(x) = Q^*(x) i L(x) Q(x)$, with

$$Q(x) \;:=\; \begin{pmatrix} \cos\xi(x) & \sin\xi(x) \\ -\sin\xi(x) & \cos\xi(x) \end{pmatrix} \quad\text{and}\quad L(x) \;:=\; \begin{pmatrix} -\Delta(x) & 0 \\ 0 & \Delta(x) \end{pmatrix},$$

where we set

$$\xi(x) \;:=\; \frac{\pi}{4} - \frac{1}{2}\arctan\frac{x}{\delta} \quad\text{and}\quad \Delta(x) \;:=\; \sqrt{x^2+\delta^2}\,.$$

This yields

$$\begin{aligned} Q'(x)Q^*(x) \;&=\; \xi'(x)\begin{pmatrix} -\sin\xi(x) & \cos\xi(x) \\ -\cos\xi(x) & -\sin\xi(x) \end{pmatrix}\begin{pmatrix} \cos\xi(x) & -\sin\xi(x) \\ \sin\xi(x) & \cos\xi(x) \end{pmatrix} \\ &=\; -\frac{1}{2}\frac{\delta}{x^2+\delta^2}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \end{aligned}$$

Hence the new variable

$$u(x) \;:=\; Q(x)v(x)$$

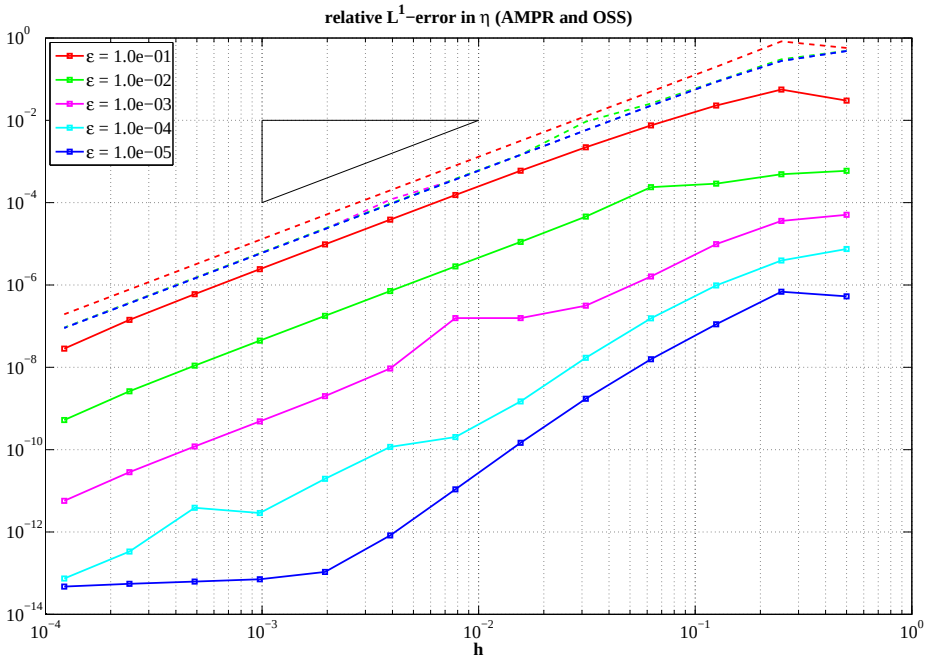

Figure 7.4: Relative L^1–error of the (explicit) OSM (solid lines) and the AMPR from [54] for η (dashed lines) for different values of ε. T_0 is exactly computed by (2.31).

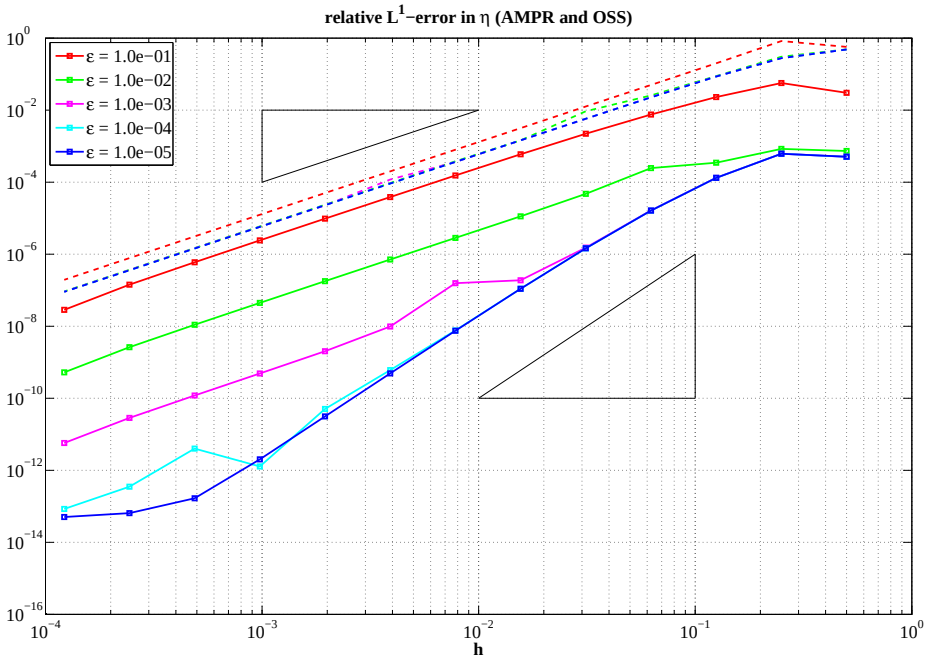

Figure 7.5: Relative L^1–error of the (explicit) OSM (solid lines) and the AMPR from [54] for η (dashed lines) for different values of ε.

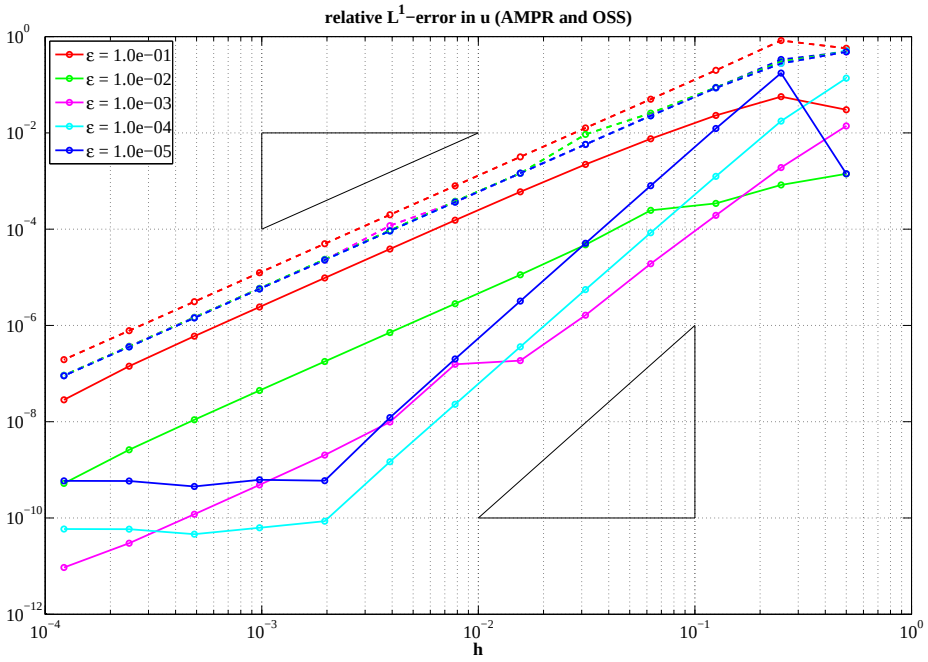

Figure 7.6: Relative L^1–error of the (explicit) OSM (solid lines) and the AMPR from [54] (dashed lines) for u for different values of ε.

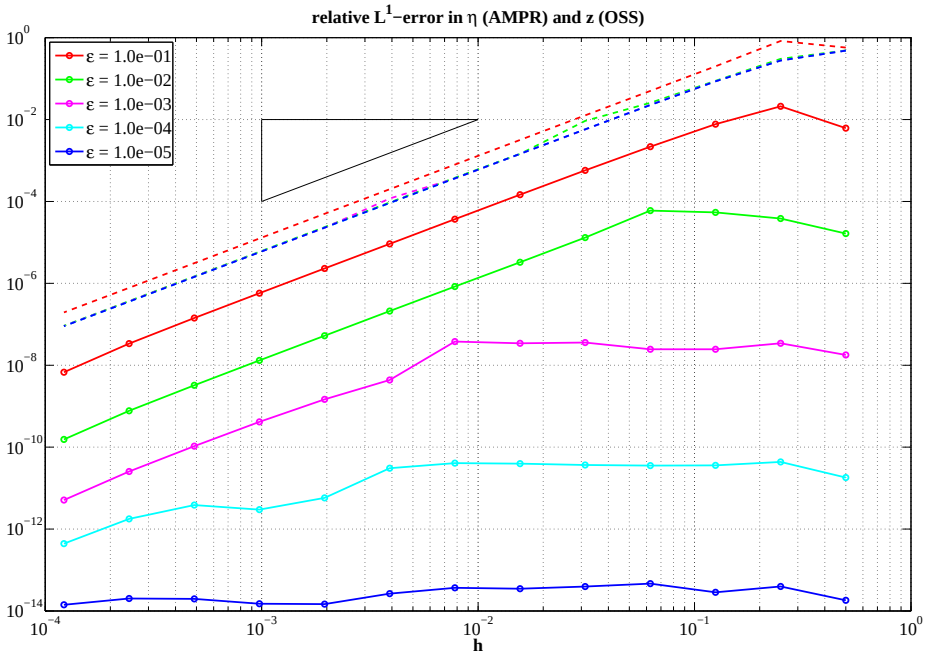

Figure 7.7: Relative L^1–error of the (Crank–Nicolson type) OSM for z (solid lines) and the AMPR [54] for η (dashed lines) for different values of ε. "Exact" evaluation of S via interpolation is used.

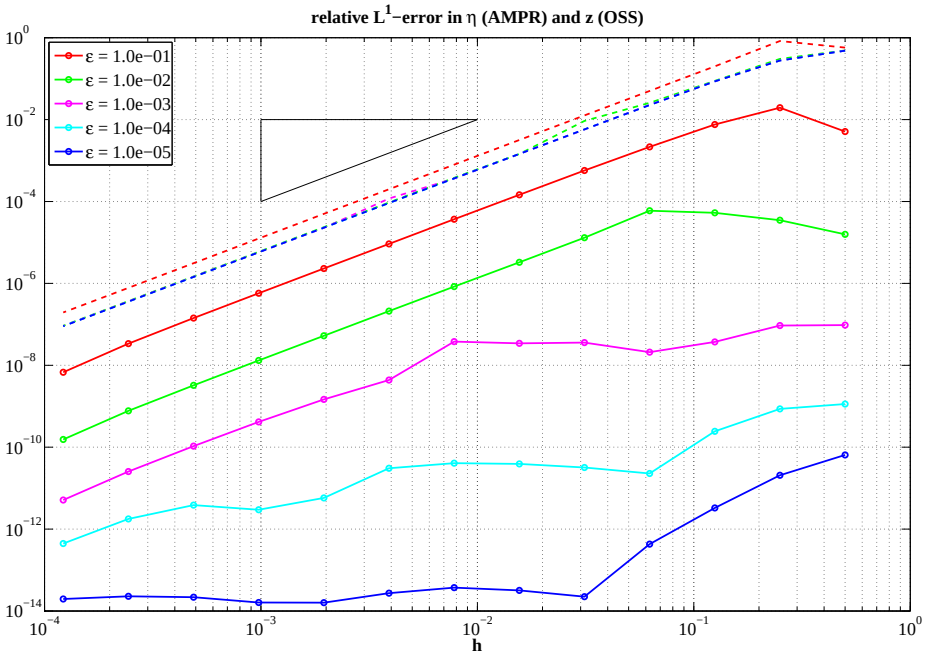

Figure 7.8: Relative L^1–error of the (Crank–Nicolson type) OSM for z (solid lines) and the AMPR [54] for η (dashed lines) for different values of ε. S is approximated as described in § 4.

solves the ODE

$$u'(x) \quad = \quad \frac{i}{\varepsilon} L(x)u(x) \, + \, B(x)u(x)\,,$$

with

$$B(x) \quad := \quad Q'(x)Q^*(x) \quad = \quad -\frac{1}{2}\frac{\delta}{\Delta^2(x)} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Let us proceed with the transformation from Remark 3.3.2 p.39 (with $n = 1$). The matrix valued phase function Φ can explicitly be computed. A primitive of Δ is (see [8] p.309)

$$\phi(x) \quad := \quad \frac{1}{2}\left(x\,\Delta(x) \, + \, \delta^2 \ln\left(x + \Delta(x)\right)\right)$$

and hence

$$\Phi(x) \quad = \quad \left(\phi(x) - \phi(x_0)\right) \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}.$$

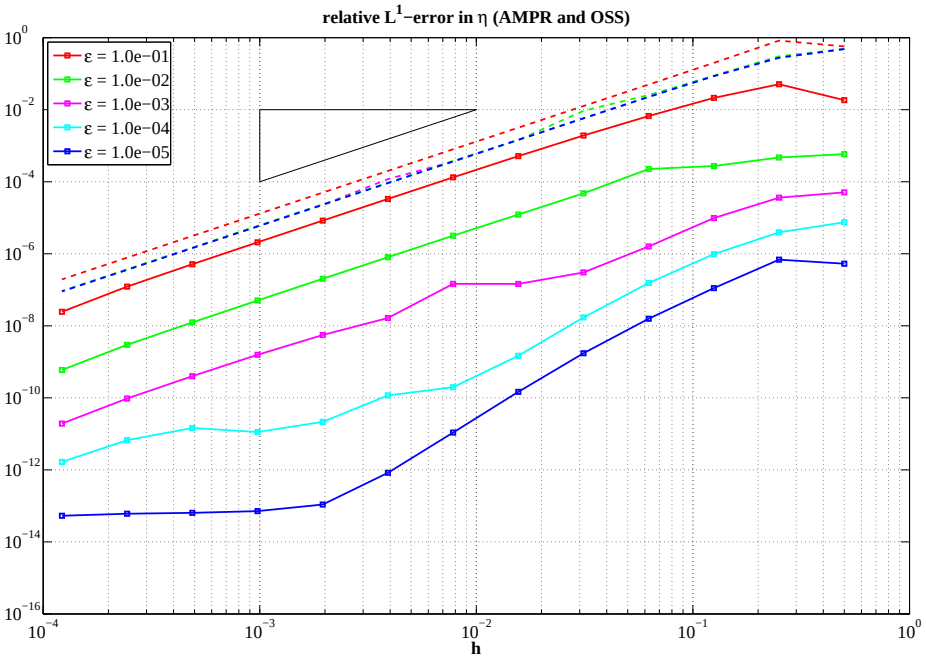

Figure 7.9: Relative L^1–error of the (Crank–Nicolson type) OSM (solid lines) and the AMPR [54] (dashed lines) for η for different values of ε. S is approximated as described in § 4, but with exact T_0 (cf. (2.31)).

Since L has two distinct eigenvalues we have $\nu = (1,1)^T$ and it follows from (3.37) that $T_0(x) = \mathcal{T}(x) = \mathrm{Id}$. This yields (cf. (3.38))

$$\begin{aligned} T_1(x) &= iD_L^-(x) \odot B(x) \\ &= -\frac{i}{2\Delta(x)} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \odot \left(-\frac{\delta}{2\Delta^2(x)} \right) \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \\ &= \frac{i\delta}{4\Delta^3(x)} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \end{aligned}$$

Thus the matrix valued function T_ε is given by (cf. (3.24))

$$T_\varepsilon(x) = T_0(x) + \varepsilon T_1(x) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \frac{i\delta\varepsilon}{4\Delta^3(x)} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

and hence

$$\begin{aligned} T_\varepsilon(x)^{-1} &= \frac{1}{1 + \frac{\delta^2\varepsilon^2}{16\Delta^6(x)}} \left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \frac{i\delta\varepsilon}{4\Delta^3(x)} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right) \\ &= \frac{16\Delta^6(x)}{16\Delta^6(x) + \delta^2\varepsilon^2} \begin{pmatrix} 1 & -\frac{i\delta\varepsilon}{4\Delta^3(x)} \\ -\frac{i\delta\varepsilon}{4\Delta^3(x)} & 1 \end{pmatrix}. \end{aligned}$$

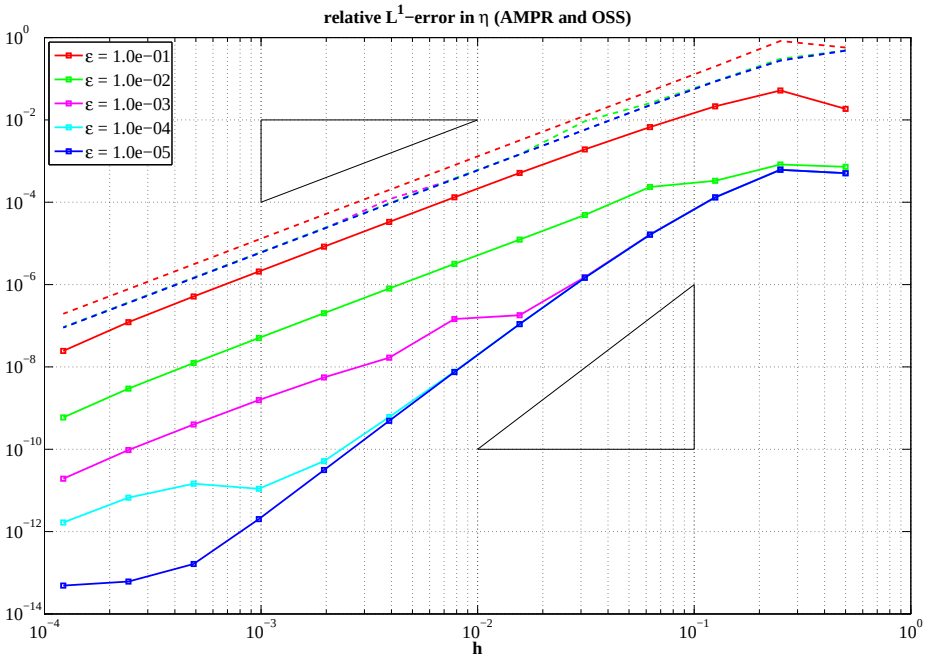

Figure 7.10: Relative L^1–error of the (Crank–Nicolson type) OSM (solid lines) and the AMPR [54] (dashed lines) for η for different values of ε. S is approximated as described in § 4.

Furthermore it holds

$$\begin{aligned} B(x)T_1(x) - T_1'(x) &= -\frac{i\delta^2}{8\Delta^5(x)} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} + \frac{3i\delta x}{4\Delta^5(x)} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \\ &= \frac{i\delta}{8\Delta^5(x)} \begin{pmatrix} -\delta & 6x \\ 6x & \delta \end{pmatrix} \end{aligned}$$

and hence we get (cf. (3.41))

$$\begin{aligned} S_1(x) &= T_\varepsilon(x)^{-1}\big(B(x)T_1(x) - T_1'(x)\big) \\ &= \frac{16\Delta^6(x)}{16\Delta^6(x) + \delta^2\varepsilon^2} \begin{pmatrix} 1 & -\frac{i\delta\varepsilon}{4\Delta^3(x)} \\ -\frac{i\delta\varepsilon}{4\Delta^3(x)} & 1 \end{pmatrix} \frac{i\delta}{8\Delta^5(x)} \begin{pmatrix} -\delta & 6x \\ 6x & \delta \end{pmatrix} \\ &= \frac{2i\delta\,\Delta(x)}{16\Delta^6(x) + \delta^2\varepsilon^2} \begin{pmatrix} -\delta - \frac{6i\delta\varepsilon x}{4\Delta^3(x)} & 6x - \frac{i\delta^2\varepsilon}{4\Delta^3(x)} \\ 6x + \frac{i\delta^2\varepsilon}{4\Delta^3(x)} & \delta - \frac{6i\delta\varepsilon x}{4\Delta^3(x)} \end{pmatrix}. \end{aligned}$$

The step size control approach from § 4.4 uses the norm of $\mathcal{S}_1$ to determine the grid. Hence we shall compute $\|\mathcal{S}_1\|$. By definition (cf. Corollary 3.3.4) it holds

$$\mathcal{S}_1(x) = R_\varepsilon(x)^{-1}\big(S_1(x) - \mathrm{diag}_\nu(S_1)\big)R_\varepsilon(x)\,,$$

where the matrix valued function R_ε solves the IVP (cf. (3.45))

$$R_\varepsilon'(x) = \mathrm{diag}(S_1(x))R_\varepsilon(x)\,, \qquad R_\varepsilon(x_0) = \mathrm{Id}\,.$$

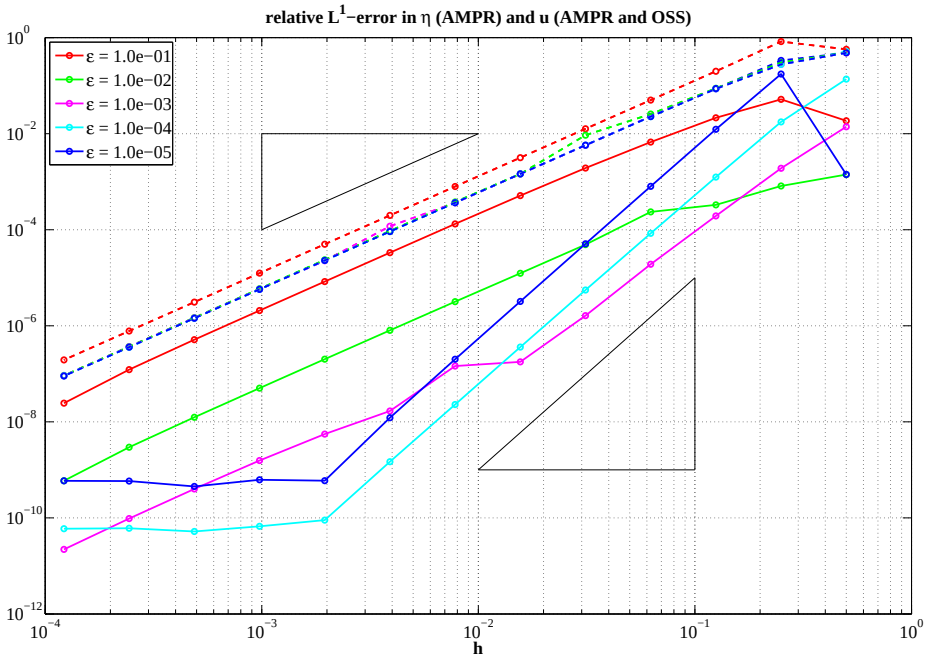

Figure 7.11: Relative L^1–error of the (Crank–Nicolson type) OSM (solid lines) and the AMPR [54] (dashed lines) for u for different values of ε. S is approximated as described in § 4.

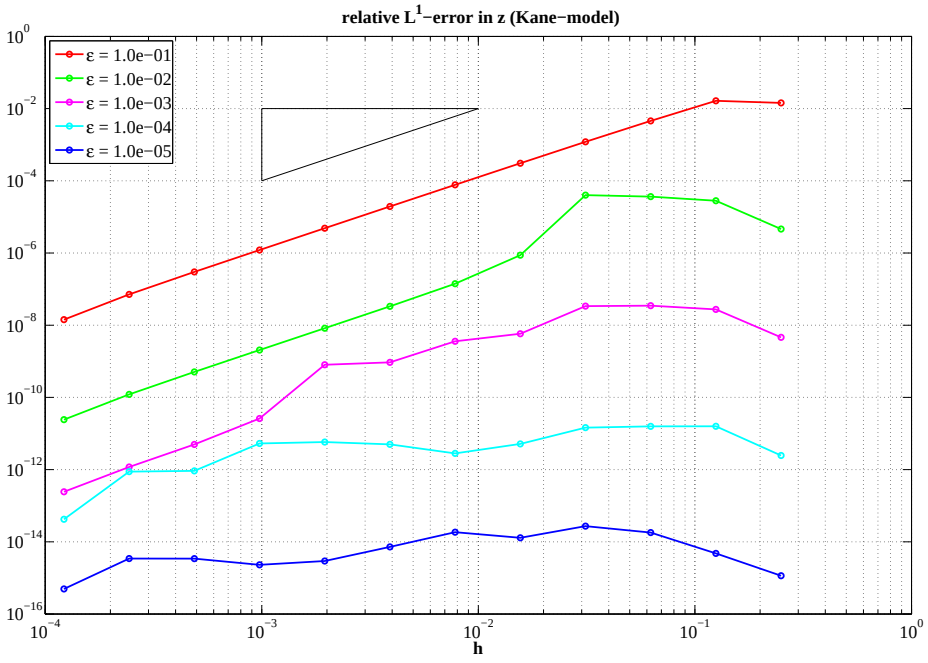

Figure 7.12: Relative L^1–error of the OSM for z related to the Kane model of § 2.1.1. "Exact" evaluation of S via interpolation is used.

Furthermore the function $\operatorname{diag}(S_1)$ can be written as

$$\operatorname{diag}(S_1) \;=\; \begin{pmatrix} f(x)+ig(x) & 0 \\ 0 & f(x)-ig(x) \end{pmatrix}$$

with real valued functions f, g. This yields

$$\begin{aligned} R_\varepsilon(x) &= \begin{pmatrix} \exp\big(\int_{x_0}^x f(s)+ig(s)\,ds\big) & 0 \\ 0 & \exp\big(\int_{x_0}^x f(s)-ig(s)\,ds\big) \end{pmatrix} \\ &= e^{\int_{x_0}^x f(s)\,ds} \begin{pmatrix} e^{i\int_{x_0}^x g(s)\,ds} & 0 \\ 0 & e^{-i\int_{x_0}^x g(s)\,ds} \end{pmatrix} \end{aligned}$$

Hence the diagonal matrix R_ε factorizes into an unitary part and a strictly positive real (exponentially increasing/decreasing) part. I. e. we can write $R_\varepsilon = UR$ with $U^*U = \mathrm{Id}$ and R real. Furthermore R is nothing but the multiplication with a certain scalar function r. This yields

$$\begin{aligned} \|\mathcal{S}_1(x)\| &= \|U(x)^*R(x)^{-1}\big(S_1(x)-\operatorname{diag}_\nu(S_1)\big)R(x)U(x)\| \\ &= \|r(x)^{-1}\big(S_1(x)-\operatorname{diag}_\nu(S_1)\big)r(x)\| \\ &= \|S_1(x)-\operatorname{diag}_\nu(S_1)\| \\ &= \left|\frac{2i\delta\,\Delta(x)}{16\Delta^6(x)+\delta^2\varepsilon^2}\right| \left\| \begin{pmatrix} 0 & 6x-\frac{i\delta^2\varepsilon}{4\Delta^3(x)} \\ 6x+\frac{i\delta^2\varepsilon}{4\Delta^3(x)} & 0 \end{pmatrix} \right\|. \end{aligned}$$

In order to compute the remaining "norm"–term let us have a look on the matrix

$$M \;:=\; \begin{pmatrix} 0 & z \\ \overline{z} & 0 \end{pmatrix} \quad \text{with} \quad z \in \mathbb{C}\,.$$

It is well known (cf. [68, p.186f]) that $\|M\|^2$ is given by the largest eigenvalue of M^*M. Since it holds

$$M^*M \;=\; \begin{pmatrix} 0 & z \\ \overline{z} & 0 \end{pmatrix} \begin{pmatrix} 0 & z \\ \overline{z} & 0 \end{pmatrix} \;=\; \begin{pmatrix} |z|^2 & 0 \\ 0 & |z|^2 \end{pmatrix}$$

we immediately get $\|M\| = |z|$. This yields

$$\|\mathcal{S}_1(x)\| \;=\; \frac{2\delta\,\Delta(x)}{16\Delta^6(x)+\delta^2\varepsilon^2} \sqrt{(6x)^2+\left(\frac{\delta^2\varepsilon}{4\Delta^3(x)}\right)^2}\,.$$

For $\varepsilon = 0.01$ the function $\log_{10}\|\mathcal{S}_1(x)\|$ is plotted in Figure 7.13. In order to get a better imagination of the surface shape we plotted some cross–sections of Figure 7.13 in Figure 7.14. We use the same values of δ and ε for the step size control algorithm (cf. Figure 7.15, 7.17).

7.4 Step size control

Let us test the two (Euler and AB2 based scheme) step size control algorithms from § 4.4 on the avoided eigenvalue crossing example from § 7.3. We compute the solution of (7.5) with $\alpha = 1$ on the interval $[-1, 1]$ for the values $\varepsilon = 0.01$

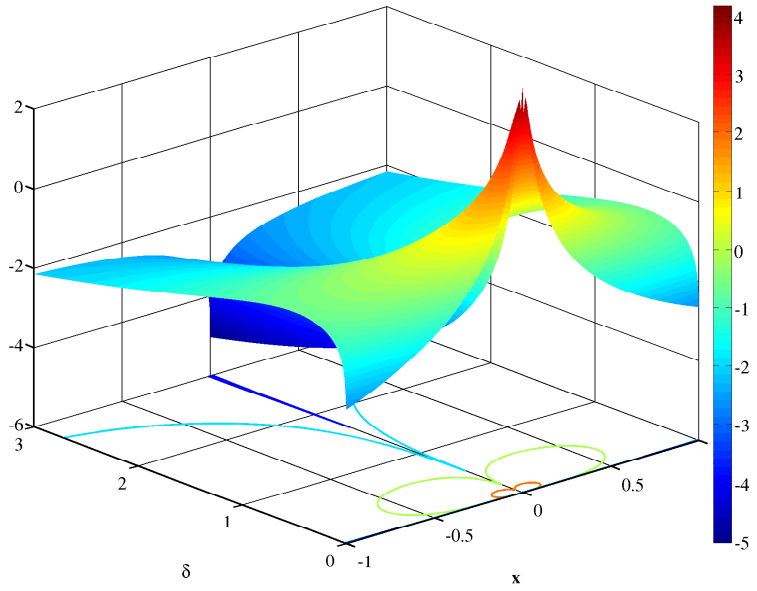

Figure 7.13: The plot shows the function $\log_{10} \|\mathcal{S}_1(x)\|$ for $\varepsilon = 0.01$.

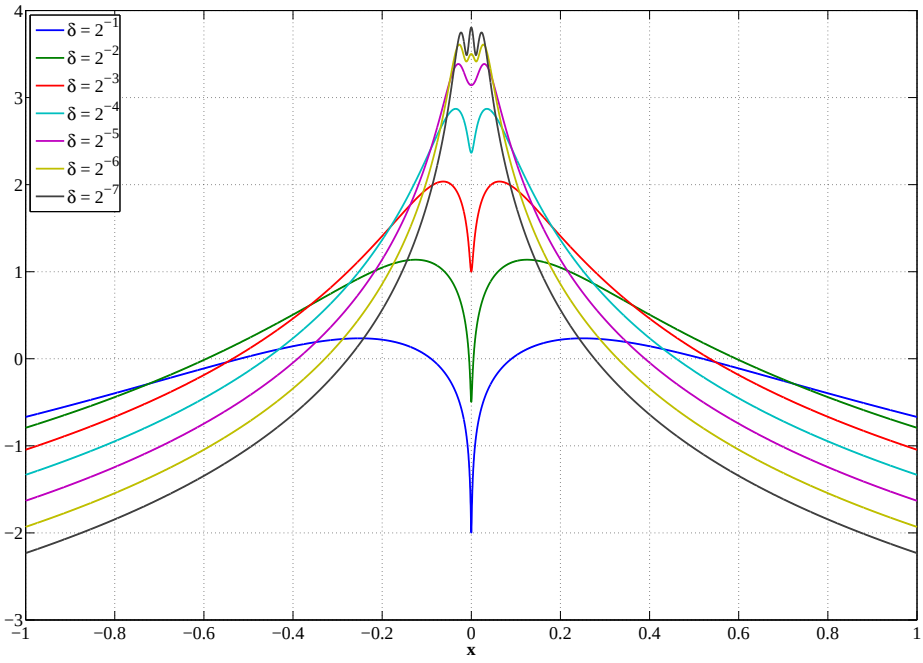

Figure 7.14: Cross sections of figure 7.13 along lines with $\delta = \text{const.}$ in the δ–x plane.

and $\delta = 2^{-1}, 2^{-3}, 2^{-5}, 2^{-7}$. We choose this setting, because it is used in [27] for the same purpose. Hence the results are comparable to the textbook ones.

We choose the initial condition for ψ at $x = 0$, such that the exact solution is given by (7.6), with $c_1 = 1, c_2 = 0$. Since there is no standard routine in Matlab to compute the Kummer functions, which are parts of the general solution of our problem (cf. (7.6)), we used Maple 14 to approximate initial conditions and the solution ψ at $x = 1$.

In the following table we collect the data of the Euler (upper half) and AB2 (lower half) based algorithms.

δ	Δt	#points	error	error (equidistant)		
2^{-1}	0.5	23	0.0002	0.0004		
2^{-3}	0.5	47	0.0024	0.0297	(Euler)	
2^{-5}	0.5	85	0.0014	0.2832		
2^{-7}	0.75	471	0.0046	3.5245		(7.7)
2^{-1}	0.5	25	0.0003	0.0002		
2^{-3}	0.5	139	0.0012	0.0175	(AB2)	
2^{-5}	0.5	1093	0.0012	0.9445		
2^{-7}	0.1	7901	0.0010	6.6888		

The third column contains the number of grid points the generated (non–equidistant) grids have. In the fourth column we present the approximation error of the OSM, where the generated non–equidistant grid is used. As reference problem we also solve the same problem on an equidistant grid with the same number of grid points and the same parameter set as for the OSM. The obtained errors are given in the last column.

In Figure 7.15 and Figure 7.17 we see the step sizes as function of x, generated with the step size control algorithms from § 4.4. While in Figure 7.15 we see the results of the algorithm based on the explicit Euler scheme to solve ω, we plot in Figure 7.17 the AB2 based scheme. For both algorithms we use the same set of parameters. In detail for $\delta = 2^{-1}, 2^{-3}, 2^{-5}$ we set $\Delta t = 0.5$. For $\delta = 2^{-7}$ it turns out (for the Euler scheme) that $\Delta t = 0.5$ is to large. Hence we reduce it to $\Delta t = 0.075$. For the AB2 based scheme we can use a coarser t–grid with $\Delta t = 0.1$. In all cases we set $\Delta x = 0.1\Delta t$.

If we compare the number of grid points (cf. (7.7)) the two algorithms generate, we observe that the Euler scheme based algorithm is significantly faster for small values of δ. Furthermore, in Figure 7.17 we observe that the AB2 based scheme produces unnatural peaks in the step sizes. They are the result of negative step sizes, which are adjusted by the max–min restriction of the algorithm to admissible increments (minimal and maximal admissible step size). Hence at this points we find step sizes equal to $h_{\min} = 10^{-6}$.

The generated grids are used to solve the problem (7.5) or rather the related IVP for z (cf. § 3.3) with the OSM from § 7.5, for the the parameters $\tau = 1$ and $\sigma_e = \sigma_i = \frac{1}{2}$. I. e. we use the Crank–Nicolson like scheme. We compare the numerical solution z of the OSM at $x = 1$ with the exact solution (approximation from Maple 14). The maximum of the errors for z and ψ are listed in table (7.7) (fourth column). For both algorithms they are between $0.2 \cdot 10^{-3}$ and $5 \cdot 10^{-3}$. This are similar values as in [27], where the error (for the same problem, but most likely different initial conditions) is between $0.5 \cdot 10^{-3}$ and $2 \cdot 10^{-3}$. We also observe that the step sizes from the approach (briefly) discussed in [27]

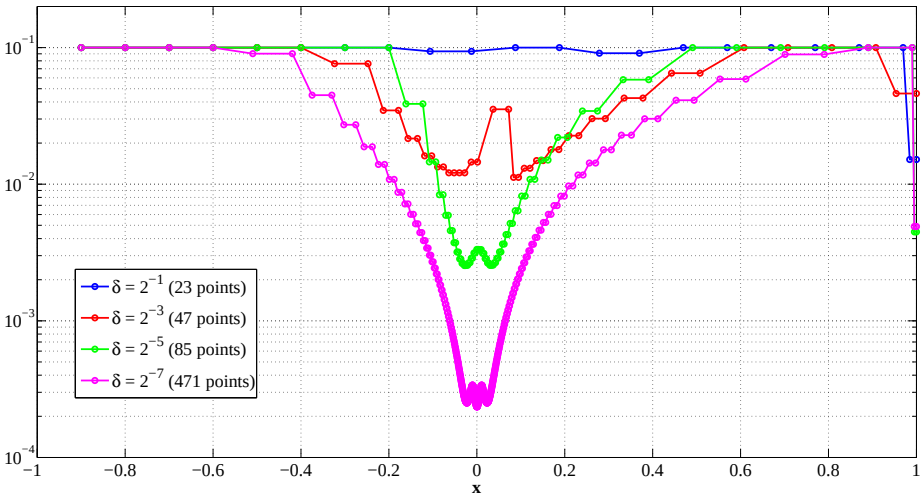

Figure 7.15: Step sizes of the non equidistant grids, derived with the Euler scheme based algorithm from § 4.4 for $\varepsilon = 0.01$ and $\delta = 2^{-1}, 2^{-3}, 2^{-5}, 2^{-5}$.

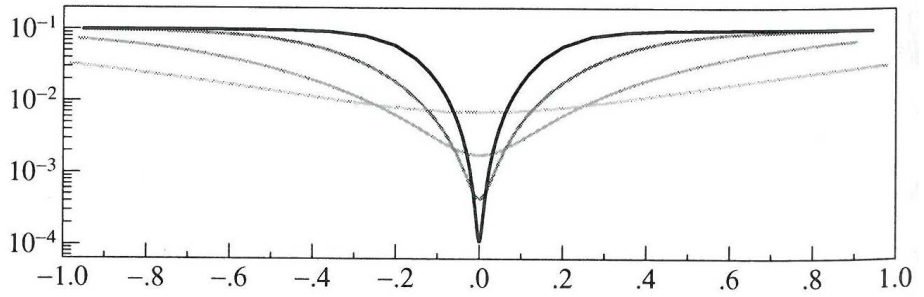

Figure 7.16: [Origin: [27] Figure 1.2 p. 538] Step sizes as function of t for $\varepsilon = 0.01$ and $\delta = 2^{-1}, 2^{-3}, 2^{-5}, 2^{-5}$ (increasing darkness). "In each case the error at the end-point $t = 1$ was between $0.5 \cdot 10^{-3}$ and $2 \cdot 10^{-3}$." [27] XIV.1.2 p.538.

(cf. Figure 7.16) are comparable (may be a bit smaller) to those derived with the Euler based scheme. In contrast the relative errors for the equidistant grid problems are competitive only for large values of δ ($\delta = \frac{1}{2}$, cf. (7.7)). If δ gets smaller, the errors increase and yield unusable results for $\delta = 2^{-7}$. In this case, the relative difference to the exact value is at least 3.5.

We have seen that our ansatz yields comparable results with respect to the approach from [27]. Since we did not spent much time on optimizing our program (and ansatz) the author believes that there is still some space for improvement.

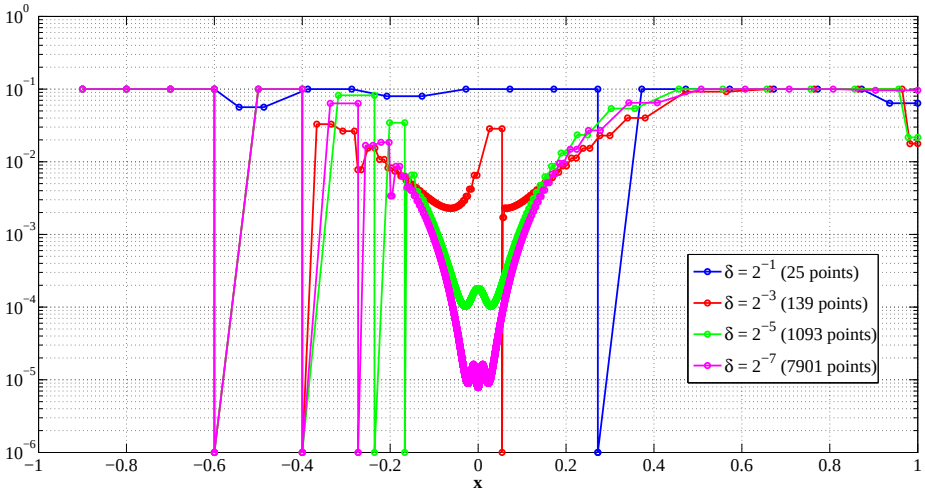

Figure 7.17: Step sizes of the non equidistant grids, derived with the AB2 scheme based algorithm from § 4.4 for $\varepsilon = 0.01$ and $\delta = 2^{-1}, 2^{-3}, 2^{-5}, 2^{-5}$

7.5 Used schemes

For the numerical examples our OSM shall only use informations at the given grid in order to avoid additional function evaluations of the ODE quantities. Additionally the multiplicities at both boundary points of the considered local subinterval $[x_n, x_{n+1}]$ should be equal. Hence we set $\kappa = 2$ and end up with the two supporting abscissas $\alpha := x_n = \zeta_1 < \zeta_2 = x_{n+1} =: \beta$.

Since we want to compare our OSM with the AMPR from [54], the desired convergence order for the one–step method is two. This yields $\tau = 2$ and hence we set $|m_1| = |m_2| = \tau$ (cf. Assumption 12 p.150). By definition of $|m_1|, |m_2|$ (cf. Remark (6.4.2)) it follows $m_{i,j} = 1$ for $i, j = 1, 2$, which yields $\mu_1 = \mu_2 = 1$. Let h be the maximal step size of the used grid. Hence, by Proposition 6.7.2 we expect that there are constants $c, \gamma \geq 0$, such that

$$c\,\rho\,\min\left(1, \gamma\left(\frac{\varepsilon}{h}\right)^2\right) h^2$$

is an upper bound of the convergence error (for our examples $\lambda = 0$).

It remains to compute the unique solution of the (generalized) interpolation problems (see p.137, (i)(a)). Since $m_1 = m_2$ we have to derive only one formula. Let[3] $P = M' \odot (K_0 + K_1 \odot M)$. Then the interpolation conditions read:

$$M'(\alpha) \odot \big(K_0 + K_1 \odot M(\alpha)\big) = F(\alpha)\,, \tag{7.8}$$
$$M'(\beta) \odot \big(K_0 + K_1 \odot M(\beta)\big) = F(\beta)\,, \tag{7.9}$$

By Assumption 7 (p.122) $M(x,\varepsilon)'_{ij} \neq 0$ for all $1 \leq i, j \leq d$ and all $(x, \varepsilon) \in \Omega$. Thus, the unique solution of (7.8), (7.9) is

$$K_1 = \Big(F(x) \odot M'(x)^{\odot -1}\Big|_{x=\alpha}^{\beta}\Big) \odot \big(M(\alpha) - M(\beta)\big)^{\odot -1} \tag{7.10}$$
$$K_0 = F(\alpha) \odot M'(\alpha)^{\odot -1} - C_1 \odot M(\alpha)\,. \tag{7.11}$$

[3]Here $M(x,\varepsilon) = D_\Phi(x) + \mathrm{diag}(\mathbb{1}_{\nu_1}, \ldots, \mathbb{1}_{\nu_q})x$, cf. Proposition 6.3.1

We use our Matlab function `fun_Pd_linear.m` to compute K_0, K_1. The crucial variables for the OSM are $P_1^\diamond(\alpha)$ and $P_1^\diamond(\beta)$ (see p.137ff). These values are returned by the function. The code reads[4]:

```
function [Pd1, Pd2] = fun_Pd_linear(L1, L2, Phi1, Phi2, ...
                                    F1, F2, x1, x2, epsilon)

    global eins I_nu

    DPhi1 = kron(Phi1, eins')-kron(Phi1, eins')';
    DPhi2 = kron(Phi2, eins')-kron(Phi2, eins')';

    M1 = DPhi1 + I_nu*x1;
    M2 = DPhi2 + I_nu*x2;

    Mx1 = kron(L1, eins')-kron(L1, eins')' + I_nu;
    Mx2 = kron(L2, eins')-kron(L2, eins')' + I_nu;

    K1 = (F1./Mx1 - F2./Mx2)./(M1-M2);
    K0 = F1./Mx1 - K1.*M1;

    Pd1 = 1i*epsilon*( (K0-1i*epsilon*K1) + K1.*DPhi1) + ...
             I_nu.*(x1*K0 + x1^2/2*K1);

    Pd2 = 1i*epsilon*( (K0-1i*epsilon*K1) + K1.*DPhi2) + ...
             I_nu.*(x2*K0 + x2^2/2*K1);
end
```

Now let us write down the program for the OSM. The complete Matlab Code is written on page 185ff.

We do not replace τ by its specified value, because the program (as written down below) is valid for all $\tau \in \mathbb{N}$. Assume we have a more sophisticated function `fun_Pd.m`, which solves an arbitrary generalized interpolation problem. Then, replacing `fun_Pd_linear.m` by `fun_Pd.m` yields a program which includes all OSM from p.137ff with $\lambda = 0$.

Assume we have already computed the quantities[5] $S_n, L_n, \Phi_n, E_{\varepsilon,n}$, and z^n.

(i) compute S_{n+1}, L_{n+1}, Φ_{n+1}, $E_{\varepsilon,n+1}$ and set $S_{1,n} := S_n$, $S_{1,n+1} := S_{n+1}$ and $\mathcal{C}^0_{n,\alpha} = \mathcal{C}^0_{n,\beta} = C^0_{n,\alpha} = C^0_{n,\beta} = \mathrm{Id}$.

(ii) `For` $j = 1, \ldots, \tau$ `do`

(a) compute $P^\diamond_{j,n}, P^\diamond_{j,n+1}$ with `fun_Pd_linear.m` $(F_1 = S_{j,n}, F_2 = S_{j,n+1})$

(b) $Q^j_n = E^\varepsilon_\Phi(x_{n+1}) \odot P^\diamond_j(x_{n+1}) \;-\; E^\varepsilon_\Phi(x_n) \odot P^\diamond_j(x_n)$

[4] Here the variables at α and β are marked with 1 and 2 respectively.
[5] Here the lower index n denotes the exact quantity evaluated at the grid point x_n.

(c) compute

$$\mathcal{C}^j_{n,\alpha} := -\sum_{l=1}^{j} P^\diamond_l(x_n)\,\mathcal{C}^{j-l}_{n,\alpha}\,, \qquad C^j_{n,\alpha} := E^*_\varepsilon(x_n)\mathcal{C}^j_{n,\alpha}E_\varepsilon(x_n)\,,$$

$$\mathcal{C}^j_{n,\beta} := -\sum_{l=1}^{j} P^\diamond_l(x_{n+1})\,\mathcal{C}^{j-l}_{n,\beta}\,, \qquad C^j_{n,\beta} := E^*_\varepsilon(x_{n+1})\mathcal{C}^j_{n,\beta}E_\varepsilon(x_{n+1})\,.$$

(d) set $S_{j+1,n} = S_n P^\diamond_{j,n}$ and $S_{j+1,n+1} = S_{n+1}P^\diamond_{j,n+1}$

`end`

(iii) update quantities for the next interval, i. e.

$$S_n = S_{n+1}\,, \quad L_n = L_{n+1}\,, \quad \Phi_n = \Phi_{n+1}\,, \quad E_{\varepsilon,n} = E_{\varepsilon,n+1}\,.$$

(iv) compute A_n, B_n by (6.35), (6.36), i. e.

$$A_n = \sum_{k=1}^{\tau} \rho^k\, Q_n^k \sum_{l=0}^{\tau-k} \rho^l\, C^l_{n,\alpha}\,, \qquad B_n = -\sum_{k=1}^{\tau} \rho^k\, Q_n^k \sum_{l=0}^{\tau-k} \rho^l\, C^l_{n,\beta}$$

(v) solve

$$\big(\mathrm{Id} + \sigma_i B_n\big) z^{n+1} \;=\; \big(\mathrm{Id} + \sigma_e A_n\big) z^n\,.$$

If we set $\tau = 1$ and $\sigma_e = \sigma_i = \frac{1}{2}$ we get the Crank–Nicolson like scheme.

The Matlab code of the OSM

In this section we present the complete Matlab code for the OSM, which is used to solve the numerical examples. We only erased some unimportant comments (line 1-15) at the beginning of the file and rearranged some lines, such that the program fits to the pages. The conversion of the program from a Matlab m-file to LaTeX is done with the "free" m–file `highlight.m` by Guillaume Flandin.

```
016 function z = fun_oss(x, Phi, L, S, nu, epsilon, rho, z0)
017
018 global Id eins I_nu
019
020 %-------------------------------------------------------
021 % Schemaparameter
022 %-------------------------------------------------------
023
024 tau = 1;
025
026 sigma   = 1/2;
027 sigma_i = 1-sigma;
028 sigma_e = sigma;
029
030 % kappa^a = 2;
```

```
031 % kappa^o = 2;
032 % kappa^b = 2;
033 %
034 % iota^a = [0 1];
035 % ioat^o = [0 1];
036 % iota^b = [0 1];
037 %
038 % m^a = [1 1; 1 1; 1 1];
039 % m^o = [1 1; 1 1; 1 1];
040 % m^b = [1 1; 1 1; 1 1];
041
042 %-------------------------------------------------------
043 % Technische Größen
044 %-------------------------------------------------------
045
046 N    = max(size(x));
047 d    = sum(nu);
048 Id   = eye(d);
049 eins = ones(d,1);
050 O    = zeros(d,d);      % hier steht der Buchstabe O
051                         % und nicht die Ziffer 0
052 I_nu = ones(nu(1),nu(1));
053 for j = 2:max(size(nu))
054      I_nu = blkdiag( I_nu, ones(nu(j),nu(j)) );
055 end
056
057 %-------------------------------------------------------
058 % Speicherreservierung
059 %-------------------------------------------------------
060
061 z = zeros(d,N);
062
063 Pdnm1 = zeros(d,d,tau);
064 Pdn   = zeros(d,d,tau);
065
066 calCa = zeros(d,d,tau+1);
067 calCb = zeros(d,d,tau+1);
068
069 Ca = zeros(d,d,tau+1);
070 Cb = zeros(d,d,tau+1);
071
072 Qn = zeros(d,d,tau);
073
074 %-------------------------------------------------------
075 % Anfangswerte
076 %-------------------------------------------------------
077
078 calCa(:,:,1) = Id;
079 calCb(:,:,1) = Id;
080
```

```
081 Ca(:,:,1) = Id;
082 Cb(:,:,1) = Id;
083
084 z(:,1) = z0;
085
086 xnm1   = x(1);
087 Phinm1 = Phi(:,1);
088 Lnm1   = L(:,1);
089 Snm1   = S(:,:,1);
090
091 Enm1   = diag(exp(1i/epsilon*Phinm1));
092
093 %-------------------------------------------------------
094 % Evolutionsschleife
095 %-------------------------------------------------------
096
097 for n=2:N
098 %--------------------------(i)-------------------------
099     xn   = x(n);
100     Phin = Phi(:,n);
101     Ln   = L(:,n);
102     Sn   = S(:,:,n);
103
104     jSnm1 = Snm1;
105     jSn   = Sn;
106     En    = diag(exp(1i/epsilon*Phin));
107
108 %--------------------------(ii)------------------------
109     for j=1:tau
110
111         %-------------------------(a)-----------------
112         [Pdm, Pd] = fun_Pd_linear(Lnm1, Ln, Phinm1, Phin, ...
113                                  jSnm1, jSn,xnm1, xn, epsilon);
114         Pdnm1(:,:,j) = Pdm;
115         Pdn(:,:,j)   = Pd;
116
117         %-------------------------(b)-----------------
118         Qn(:,:,j) = En'*Pd*En - Enm1'*Pdm*Enm1;
119
120         %-------------------------(c)-----------------
121         clCa = 0;       % hier steht der Buchstabe O
122         clCb = 0;       % und nicht die Ziffer 0
123
124         for l=1:j
125             clCa = clCa - Pdnm1(:,:,l)*calCa(:,:,j-l+1);
126             clCb = clCb - Pdn(:,:,l)  *calCb(:,:,j-l+1);
127         end
128
129         calCa(:,:,j+1) = clCa;
130         calCb(:,:,j+1) = clCb;
```

```
131
132         Ca(:,:,j+1) = Enm1'*clCa*Enm1;
133         Cb(:,:,j+1) = En'  *clCb*En;
134
135         %--------------------------(d)----------------
136         jSnm1 = Snm1*Pdm;
137         jSn   = Sn  *Pd;
138
139     end
140 %-------------------------(iii)----------------------
141
142     Phinm1 = Phin;
143     Snm1   = Sn;
144     Lnm1   = Ln;
145     Enm1   = En;
146     xnm1   = xn;
147
148 %-------------------------(iv)-----------------------
149     An = 0;              % hier steht der Buchstabe
150     Bn = 0;              % O und nicht die Ziffer 0
151     Gmma = 0;
152     Gmmb = 0;
153
154     for k = tau:-1:1
155         Gmma = Gmma + rho^(tau-k)*Ca(:,:,tau-k+1);
156         Gmmb = Gmmb + rho^(tau-k)*Cb(:,:,tau-k+1);
157         An   = An   + rho^k*Qn(:,:,k)*Gmma;
158         Bn   = Bn   - rho^k*Qn(:,:,k)*Gmmb;
159     end
160
161 %-------------------------(iv)-----------------------
162     z(:,n) = (Id+sigma_i*Bn)\((Id+sigma_e*An)*z(:,n-1));
163
164 end
165 end
```

Chapter 8

Miscellaneous

For the computation of the WKB–type transformation from § 3.3 (cf. § 4) we need numerical approximations of derivatives. Furthermore (for non–equidistant grids) we also have to approximate certain values at of grid points. In this chapter we shall discuss our strategies for this problems.

We choose a finite difference approach to approximate derivatives. This is discussed in § 8.1. Here we prove a quite general statement about the approximation error of finite difference approximations and derive an inhomogeneous linear system whose solution gives the "best" choice of weights for prescribed abscissas. This approach can be used to derive approximations of derivatives for non–equidistant spaced abscissas. In this case, one has to solve the linear system. We also compute the formulas for equidistant finite differences for the first derivative, which are of fourth order.

In § 8.2 we make some numerical experiments with the finite differences discussed in § 8.1. Here we quite good observe a superposition of the theoretical error bound and numerical noise which (most likely) originates from *Matlabs* machine accuracy.

The section § 8.3 is dedicated to the approximation of non–grid values for the numerical solution of a first order initial value problem. We use Hermite interpolation of the numerical solution to determine an approximation between to grid points. As long as the integrator used to compute the solution of the IVP has an order of four or less, the interpolation approach shall (approximately) be as accurate as the solutions at the grid points (cf. Lemma 8.3.1).

In the final section § 8.4 we collect some (frequently used) classical results for ODEs from literature.

8.1 Finite differences

In this section we discuss the finite differences we use to approximate the first derivative of a given function f. We start with a general setting where the support abscissas are (in a certain framework) arbitrary. Afterwards we discuss the special case of equidistant grids and write down the explicit formulas.

Definition 8.1.1. *Let $r, s \in \mathbb{N}$ with $r < s$ and let $\eta_1 < \cdots < \eta_s \in \mathbb{R}$ and*

$v \in \mathbb{R}^s$. For $h > 0$, $x \in \mathbb{R}$ and $f \in C(\mathbb{R})$ we define

$$\mathrm{FD}^r_{[\eta,v]}(f,x,h) \quad := \quad \frac{1}{h^r}\sum_{l=1}^{s} v_l f(x+\eta_l h)\,.$$

Lemma 8.1.2. *Let $s \in \mathbb{N}$ and let $\eta_1 < \cdots < \eta_s \in \mathbb{R}$. Furthermore let $x \in \mathbb{R}$ and let $h_0 > 0$. We set*

$$I \quad := \quad [\min(x, x+\eta_1 h_0), \max(x, x+\eta_s h_0)]\,.$$

Than there exists for every $r = 1, \ldots, s-1$ a unique vector $v^r \in \mathbb{R}^s$, such that for all $f \in \mathbb{C}^s(I)$ there exists a constant $c \geq 0$, such that for all (admissible) $0 < h \leq h_0$ it holds

$$|f^{(r)}(x) - \mathrm{FD}^r_{[\eta,v^r]}(f,x,h))| \quad \leq \quad c\,h^{s-r}\,. \tag{8.1}$$

The constant c depends on p and f. More precise one finds

$$c \quad \leq \quad \tilde{c}(p)\,\|f^{(s)}\|_{L^\infty(I)}\,.$$

Proof. Firstly it is clear that $x + \eta_j h \in I$ for $0 < h \leq h_0$ and $j = 1, \ldots, s$. The proof is based (as usual) on Taylor expansion. Since $f \in C^s(I)$ it holds for $x, x+\delta \in I$

$$f(x+\delta) \quad = \quad \sum_{k=0}^{s-1} f^{(k)}(x)\frac{\delta^k}{k!} \;+\; R(x,\delta,s)\,. \tag{8.2}$$

The remainder can be written down in its Lagrangian form (cf. [23]):

$$R(x,\delta,s) \quad = \quad f^{(s)}(\xi(x,\delta))\frac{\delta^s}{s!}\,,$$

with some $\xi(x,\delta) \in [\min(x, x+\delta), \max(x, x+\delta)]$. This yields

$$\begin{aligned}
\mathrm{FD}^r_{[\eta,v]}(f,x,h) &= \frac{1}{h^r}\sum_{l=1}^{s} v_l f(x+\eta_l h) \\
&= \frac{1}{h^r}\sum_{l=1}^{s} v_l \sum_{k=0}^{s-1} f^{(k)}(x)\frac{(\eta_l h)^k}{k!} \;+\; R(x,\eta_l h,s) \\
&= \frac{1}{h^r}\sum_{k=0}^{s-1} f^{(k)}(x)\frac{h^k}{k!}\sum_{l=1}^{s} v_l \eta_l^k \;+\; \frac{1}{h^r}\sum_{l=1}^{s} v_l R(x,\eta_l h,s)\,.
\end{aligned}$$

Since v does not depend on h, the sum of the remainder terms is of order $\mathcal{O}(h^{s-r})$. Hence to fulfill (8.1) it is necessary and sufficient that v solves the h–independent linear system:

$$\begin{pmatrix} \eta_1^0 & \cdots & \eta_s^0 \\ \vdots & & \vdots \\ \eta_1^{s-1} & \cdots & \eta_s^{s-1} \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_s \end{pmatrix} \quad = \quad r!\,\left(\delta_{i,r}\right)_{i=1,\ldots,s-1}\,.$$

The transposed of the above matrix is known as Vandermonde matrix (cf. [33]). Since $\eta_1 < \cdots < \eta_s$ the linear system is uniquely solvable. □

Remark 8.1.3. *From the proof of Lemma 8.1.2 we deduce that the unique vector $v(r) \in \mathbb{R}^n$ from Lemma 8.1.2 is given as the unique solution of the linear system*

$$\begin{pmatrix} 1 & \dots & 1 \\ \eta_1^1 & \dots & \eta_s^1 \\ \vdots & & \vdots \\ \eta_1^{s-1} & \dots & \eta_s^{s-1} \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_s \end{pmatrix} = r! \left(\delta_{i,r}\right)_{i=0,\dots,s-1}. \tag{8.3}$$

Let A be the matrix on the left hand side and let w be the unique solution of $A^T w = b$. Than (cf. [33]) $p(x) := \sum_{j=0}^{s-1} w_j x^j$ solves the interpolation problem $p(\eta_j) = b_j$ for $j = 1, \dots, s$. Hence finding the coefficient vector v for the finite difference scheme is (in some sense) an adjoint problem to the polynomial interpolation problem. If $w_1, \dots, w_s \in \mathbb{R}^s$ are given such that $A^T w_j = e_j$ for $j = 1, \dots, s$ with $e_j = (\delta_{ij}) \in \mathbb{R}^s$ (i. e. $w_1, \dots, w_s$ are the coefficients of the Lagrange interpolation polynomials corresponding to the support abscissas η), than it holds

$$v_j = e_j^T v = (A^T w_j)^T v = w_j^T A v = w_j^T r!\, e_{r+1}.$$

This yields $v = r!(w_{1,r+1}, \dots, w_{s,r+1})^T$.

Corollary 8.1.4. *The FD schemes from Lemma 8.1.2 are exact on the space of polynomials of degree less than s.*

Proof. For polynomials of degree less than s the remainder of the Taylor approximation in (8.2) is zero. Hence the linear system (8.3) is equivalent to the FD. □

For certain numerical examples we use equidistant grids. Since we want to approximate first derivatives with order $\mathcal{O}(h^4)$, we set $s = 5$. We should stay as close as possible to the point where the derivative is approximated. Hence there shall be an index j_0 such that $\eta_{j_0} = 0$. This yields the following five vectors:

$$\eta = \begin{pmatrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} -2 \\ -1 \\ 0 \\ 1 \\ 2 \end{pmatrix}, \begin{pmatrix} -3 \\ -2 \\ -1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -4 \\ -3 \\ -2 \\ -1 \\ 0 \end{pmatrix}.$$

For each of them we have derived the corresponding coefficients v with *Maple14*. This yields the finite difference schemes of Definition 8.1.5.

Definition 8.1.5. *Let f be continuous on $[a, b]$ and let x_n be grid a point of the equidistant grid $a = x_{n_a} < \dots < x_{n_b} = b$ with $n_a \le 0 \le n_b$ and step size h, i. e.*

$x_j = x_0 + jh$ for $j = n_a, \dots, n_b$. For (in each case) properly chosen n we define

$$\begin{aligned}
\mathrm{FD}^1_{(0,4)}[f]_n &:= \frac{1}{12h}\big(-25f_n + 48f_{n+1} - 36f_{n+2} + 16f_{n+3} - 3f_{n+4}\big)\,,\\
\mathrm{FD}^1_{(1,3)}[f]_n &:= \frac{1}{12h}\big(-3f_{n-1} - 10f_n + 18f_{n+1} - 6f_{n+2} + f_{n+3}\big)\,,\\
\mathrm{FD}^1_{(2,2)}[f]_n &:= \frac{1}{12h}\big(f_{n-2} - 8f_{n-1} + 8f_{n+1} - f_{n+2}\big)\,,\\
\mathrm{FD}^1_{(3,1)}[f]_n &:= \frac{1}{12h}\big(-f_{n-3} + 6f_{n-2} - 18f_{n-1} + 10f_n + 3f_{n+1}\big)\,.\\
\mathrm{FD}^1_{(4,0)}[f]_n &:= \frac{1}{12h}\big(3f_{n-4} - 16f_{n-3} + 36f_{n-2} - 48f_{n-1} + 25f_n\big)\,.
\end{aligned}$$

Here we use the notation f_j for $f(x_j)$. The index tuple describes the numerical stencil used for the scheme. More precise (r, s) means one uses the values at the grid points $x_{n-r}, \dots, x_{n+s}$.

8.2 Numerical experiments for the finite differences from § 8.1

In this section we visualize the numerical behavior of the Finite Differences from Lemma 8.1.2. The estimate (8.1) yields the convergence of the FD as $h \to 0$. But it is well known that this theoretically result breaks down for numerical approximations on a computer, when the step size gets to small (cf. [29]). A reason for this is the influence of the machine accuracy and the related round off error. Hence the step size we use for the FD should not be to small or to large. In order to get an idea of a "good" interval let us approximate some derivatives and compare them with the exact solution. The result shall be a superposition of the theoretical error from Lemma 8.1.2 and the effect of machine accuracy and accuracy of solving the occurring linear system (8.3). The latter errors we call numerical noise, which is plotted in Figure 8.1 (p.194) and Figure 8.1 (p.194) for general FDs from Lemma 8.1.2, Remark 8.1.3, and the equidistant FDs from Definition 8.1.5 respectively. The (theoretical) general FDs from Lemma 8.1.2 are exact on polynomials up to a degree of order s (see Corollary 8.1.4). Hence an approximation of the first derivative of a polynomial of degree less than s makes the numerical noise visible. Since we use second and fourth order FDs we choose $f(x) = x$ and approximate $f'(x) = 1$ at $x = 1$.

Let f be a smooth function (for simplicity of notation on the whole real line) and let $x \in \mathbb{R}$ be a point at which we want to approximate the r^{th} derivative of f. Further let $\xi_1 < \dots < \xi_n \in \mathbb{R}$ $(n > r)$ be our support abscissas and let $f_1, \dots, f_n$ be the corresponding values of f. Since the largest distance from x to the ξ_j determines h, the abscissas should be chosen close to x. The program to compute the general FDs is based on Remark 8.1.3:

(i) compute the relative coordinates, i. e. $\tilde{\eta} = (\xi_j - x)_{j=1,\dots,n}$

(ii) determine $h = \texttt{max}(\texttt{abs}(\eta))$

(iii) rescale $\tilde{\eta}$: $\eta = \tilde{\eta}/h$

(iv) compute the Vandermonde Matrix A corresponding to η, i. e. $A_{ij} = \eta_i^{j-1}$

(v) solve the linear system $A^T v = r!(\delta_{r+1,j})_{j=1,\dots,n}$

(vi) set $FD = \frac{1}{h^r} \sum_{j=1}^{n} v_j f_j$.

In *matlab* exists the routine `vander` to derive the Vandermonde Matrix A from the vector η. But its definition differs from our. There the calculation rule is $A_{ij} = v_i^{n-j}$. Hence the linear system one has to solve reads $Av = r!(\delta_{n-r,j})$. The rescaling of η in (iii) yields moderate coefficients of A which are of order $\mathcal{O}(1)$. Without rescaling the smallest (non zero) entries of A could be at machine precision, which strongly reduces the accuracy of the linear system solution.

In order to compute the numerical noise we first fix a vector η for each scheme. Than we derive for every value of h the corresponding abscissas ξ and plug in the (needed) corresponding values in the program. The result is compared with the exact solution $f'(1) = 1$. We use the following general FDs:

scheme	η
4^{th}-order symmetric	$\frac{1}{2}(-2,\ -1,\ 0,\ 1,\ 2)$
4^{th}-order left	$\frac{1}{8}(-8,\ -5,\ -2,\ -1,\ 0)$
4^{th}-order right	$\frac{1}{8}(0,\ 1,\ 2,\ 5,\ 8)$
2^{th}-order symmetric	$(-1,\ 0,\ 1)$
2^{th}-order left	$\frac{1}{4}(-4,\ -3,\ 0)$
2^{th}-order right	$\frac{1}{3}(0,\ 1,\ 3)$

In Figure 8.2 and 8.2 we plot the numerical noise for non–equidistant and equidistant FD as described before. We observe that the approximation error of the FDs from Definition 8.1.5 is a bit smaller than those of the general FDs. This is reasonable, since one additionally has to solve a linear problem, compared to the equidistant schemes. However, in both Figures we observe a $\mathcal{O}(h^{-1})$ behavior of the numerical noise.

In Figure 8.3 we plot the numerical error the equidistant FDs from Definition 8.1.5. We approximate the first derivative of $f(x) = \sqrt{x^2 + \delta^2}$ at $x = 1$ with $\delta = 10^{-7}$. This function appears in the § 7.3 and is denoted by Δ. The triangles have a slope of 2 (upper triangle) and slope 4 (lower triangle). The (black) dashed and solid lines are the approximate upper bounds of the numerical noise for equidistant and non–equidistant FDs respectively, as drawn in Figure 8.1 and Figure 8.2.

Furthermore, in Figure 8.4 we plot the numerical error for FDs with equidistant abscissas from Definition 8.1.5. We approximate the first derivative of $f(x) = 1/\sqrt{x^2 + \delta^2}$ at $x = 1$ with $\delta = 10^{-7}$. The triangles have a slope of 2 (upper triangle) and slope 4 (lower triangle). The (black) dashed and solid lines are the approximate upper bounds of the numerical noise for equidistant and non–equidistant FDs respectively, as drawn in figure 8.1 and Figure 8.2.

In both Figures we observe that the FDs with support abscissas at the right–hand side yield (for large h) the "smoothest" results. This is reasonable, since the function Δ has a sharp "peak" at $x = 0$. Thus the other FDs use function values beyond or close to the peak, which of course reduces the accuracy.

8.3 Intermediate values

Let y be the unique solution of the initial value problem (on $[a, b] \subset \mathbb{R}$)

$$y'(x) \ = \ A(x)y(x) \ + \ f(x)\,, \qquad y(x_0) \ = \ y_0 \in \mathbb{C}^d\,,$$

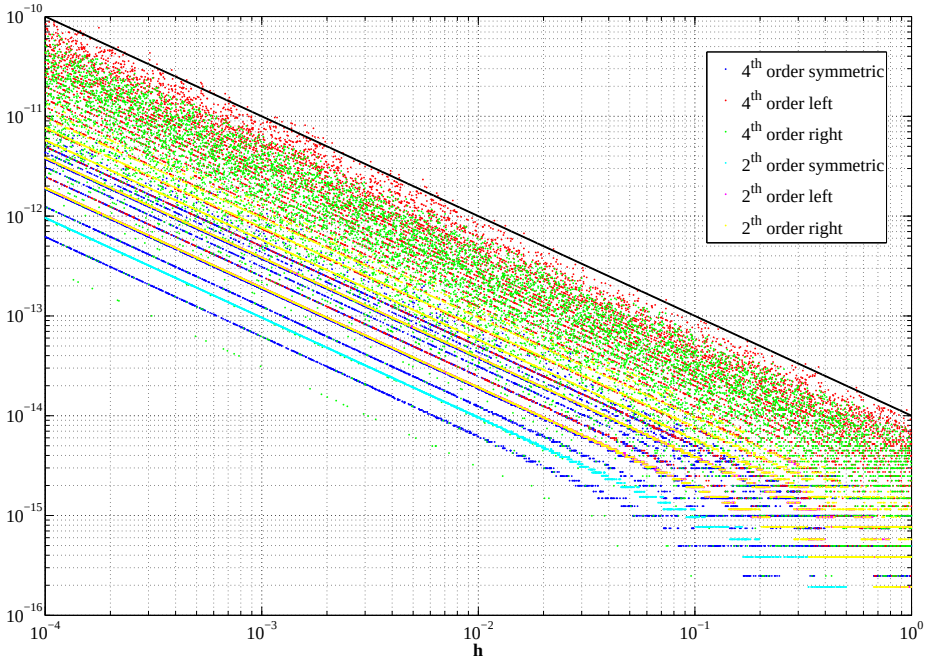

Figure 8.1: Numerical noise for some FDs with non–equidistant abscissas. The first derivative of $f(x) = x$ at $x = 1$ is approximated. The solid black line has slope -1, which indicates an $\mathcal{O}(h^{-1})$ behavior of the approximation error.

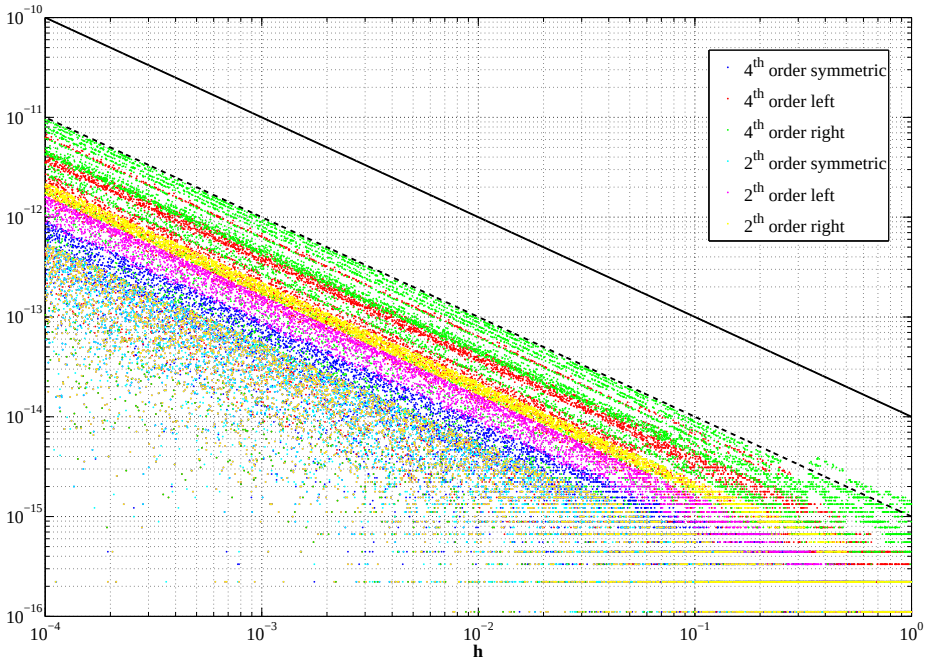

Figure 8.2: Numerical noise for the FDs with equidistant abscissas from Definition 8.1.5. The first derivative of $f(x) = x$ at $x = 1$ is approximated.

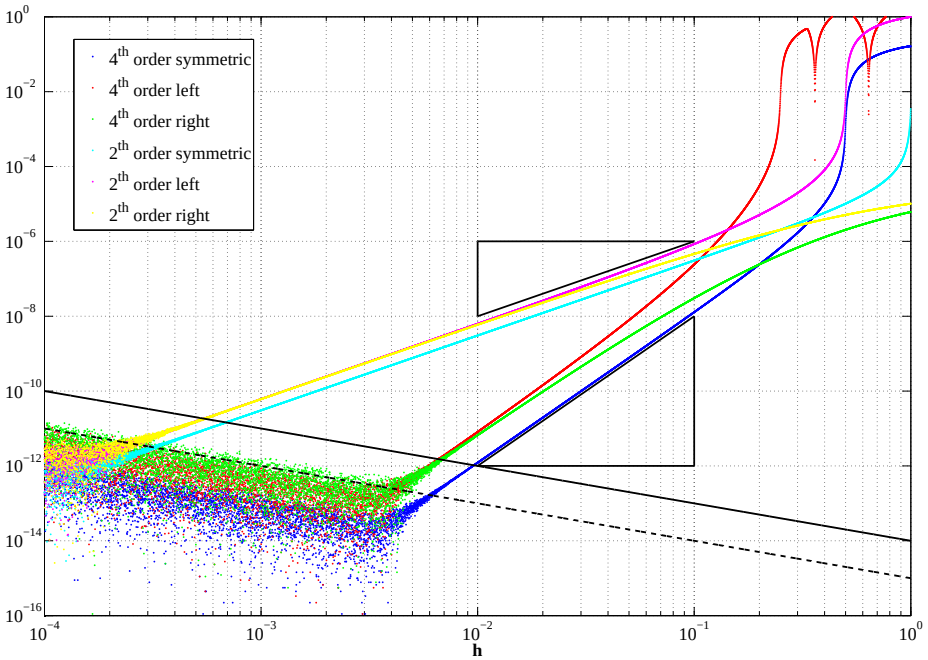

Figure 8.3: Numerical error the equidistant FDs from Definition 8.1.5. The first derivative of $f(x) = \sqrt{x^2 + \delta^2}$ at $x = 1$ with $\delta = 10^{-7}$ is approximated. The triangles have a slope of 2 (upper triangle) and slope 4 (lower triangle). The (black) dashed and solid lines are the approximate upper bounds of the numerical noise for equidistant and non–equidistant FDs respectively, as drawn in Figure 8.1 and Figure 8.2.

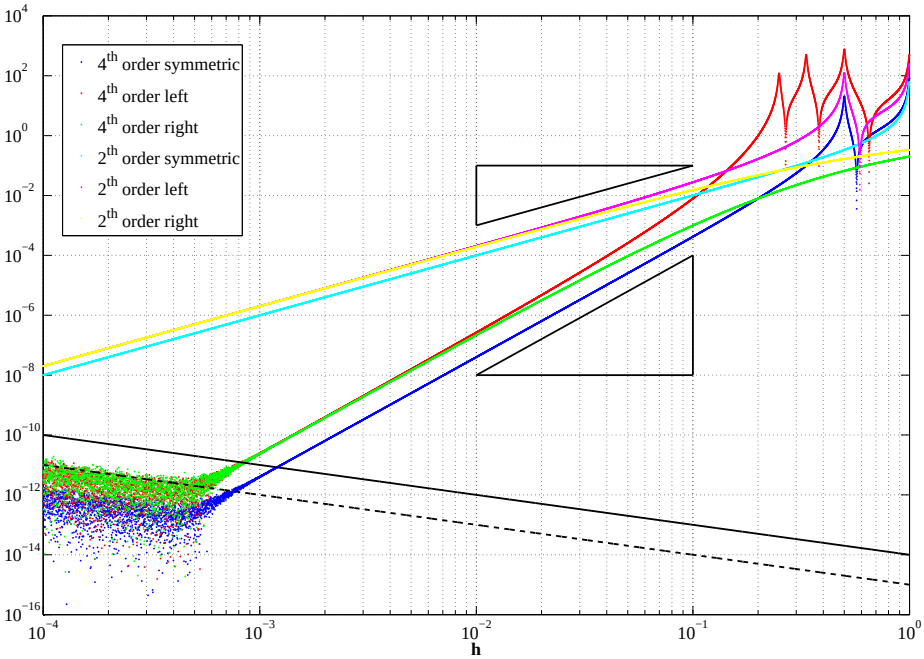

Figure 8.4: Numerical error for FDs with equidistant abscissas from Definition 8.1.5. The first derivative of $f(x) = 1/\sqrt{x^2 + \delta^2}$ at $x = 1$ with $\delta = 10^{-7}$ is approximated. The triangles have a slope of 2 (upper triangle) and slope 4 (lower triangle). The (black) dashed and solid lines are the approximate upper bounds of the numerical noise for equidistant and non–equidistant FDs respectively, as drawn in figure 8.1 and Figure 8.2.

with $x_0 \in [a,b]$. Further let y_n be a numerical approximation of $y(x_n)$, computed with a numerical integrator on the grid $a = x_{n_a} < x_{n_a+1} \cdots < n_{n_b} = b$. We assume that the used method is of order $\mathcal{O}(h_n^\gamma)$, i. e. there exists a constant $c > 0$ independently of n and the grid, such that

$$|y(x_n) - y_n| \;\leq\; c\, h_n^\gamma \quad \text{with} \quad h_n \;:=\; x_{n+1} - x_n \,.$$

If we want to compute values of y at a non grid abscissa $\overline{x}$ we use the following interpolation approach.

Let $\overline{x} \in [x_n, x_{n+1}]$ and let p be the unique third order polynomial which solves the Hermite interpolation problem

$$\begin{aligned} p(x_n) &= y(x_n)\,, \qquad & p(x_{n+1}) &= y(x_{n+1})\,,\\ p'(x_n) &= y'(x_n)\,, \qquad & p'(x_{n+1}) &= y'(x_{n+1})\,. \end{aligned}$$

It holds (cf. [29, p.311f])

$$\begin{aligned} p(x) \;=\; & \quad y(x_n)H_n(x) \,+\, y(x_{n+1})H_{n+1}(x) \\ & + y'(x_n)\widehat{H}_n(x) \,+\, y'(x_{n+1})\widehat{H}_{n+1}(x)\,. \end{aligned}$$

The polynomials $H_n, H_{n+1}, \widehat{H}_{n+1}, \widehat{H}_{n+1}$ are given by

$$H_n(x) \;=\; \left(1 - 2\frac{x - x_n}{x_n - x_{n+1}}\right)\left(\frac{x - x_{n+1}}{x_n - x_{n+1}}\right)^2, \tag{8.4}$$

$$H_{n+1}(x) \;=\; \left(1 - 2\frac{x - x_{n+1}}{x_{n+1} - x_n}\right)\left(\frac{x - x_n}{x_{n+1} - x_n}\right)^2, \tag{8.5}$$

$$\widehat{H}_n(x) \;=\; (x - x_n)\left(\frac{x - x_{n+1}}{x_n - x_{n+1}}\right)^2, \tag{8.6}$$

$$\widehat{H}_{n+1}(x) \;=\; (x - x_{n+1})\left(\frac{x - x_n}{x_{n+1} - x_n}\right)^2. \tag{8.7}$$

Hence $p(\overline{x})$ is a suitable approximation for $y(\overline{x})$.

Lemma 8.3.1. *Let $A \in C^3([a,b], \mathbb{C}^{n\times n})$, $f \in C^3([a,b], \mathbb{C}^d)$ and let the relative coordinates $\theta_l, \theta_r \in [0,1]$, such that*[1]

$$\overline{x} = x_n + \theta_l h_n \quad \textit{and} \quad \overline{x} = x_{n+1} - \theta_r h_n \,.$$

Further let y, y_n, y_{n+1} be as described in the beginning of this section and let

$$\begin{aligned} \overline{y} \;:=\; & \quad \theta_r^2\big[(1 + 2\theta_l)\,\mathrm{Id} + \theta_l h_n\, A_n\big]\, y_n \,+\, \theta_r^2\theta_l h_n\, f_n \\ & + \theta_l^2\big[(1 + 2\theta_r)\,\mathrm{Id} - \theta_r h_n\, A_{n+1}\big]\, y_{n+1} \,-\, \theta_l^2\theta_r h_n\, f_{n+1}\,. \end{aligned}$$

Here A_n, A_{n+1}, f_n, f_{n+1} are short notations for $A(x_n)$, $A(x_{n+1})$ and $f(x_n)$, $f(x_{n+1})$ respectively. It holds

$$|y(\overline{x}) - \overline{y}| \;\leq\; \tfrac{\|y^{(4)}\|_\infty}{256}\, h_n^4 \,+\, \big(\|\,\mathrm{Id}\,\| \,+\, \tfrac{h_n}{4}\|A\|_\infty\big) c\, h_n^\gamma\,.$$

The constant $c \geq 0$ is (only) determined by the numerical integrator used to compute y_n, y_{n+1}.

[1] I. e. $\theta_r = 1 - \theta_l$.

Proof. Inserting the two representations of $\overline{x}$ in (8.4),...,(8.7) yields

$$\begin{aligned} H_n(\overline{x}) &= (1+2\theta_l)\theta_r^2\,, & H_{n+1}(\overline{x}) &= (1+2\theta_r)\theta_l^2\,,\\ \widehat{H}_n(\overline{x}) &= \theta_l h_n \theta_r^2\,, & \widehat{H}_{n+1}(\overline{x}) &= -\theta_r h_n \theta_l^2\,. \end{aligned}$$

Hence we get,

$$\begin{aligned} p(\overline{x}) \;=\;& y(x_n)(1+2\theta_l)\theta_r^2 \,+\, y(x_{n+1})(1+2\theta_r)\theta_l^2\\ &+ y'(x_n)\theta_l h_n \theta_r^2 \,-\, y'(x_{n+1})\theta_r h_n \theta_l^2\,. \end{aligned}$$

The derivatives of y can be replaced, using the ODE, by $Ay+f$. This yields

$$\begin{aligned} p(\overline{x}) \;:=\;& \theta_r^2\big[(1+2\theta_l)\,\mathrm{Id} + \theta_l h_n\, A(x_n)\big]\, y(x_n) \,+\, \theta_r^2\theta_l h_n\, f(x_n)\\ &+ \theta_l^2\big[(1+2\theta_r)\,\mathrm{Id} - \theta_r h_n\, A(x_{n+1})\big]\, y(x_{n+1}) \,-\, \theta_l^2\theta_r h_n\, f(x_{n+1})\,. \end{aligned}$$

Thus we get

$$\begin{aligned} \|p(\overline{x})-\overline{y}\| \;\le\;& \big\|(\theta_r^2(1+2\theta_l)\,\mathrm{Id} + \theta_r^2\theta_l h_n A_n\big\|\|y(x_n)-y_n\|\\ &+ \big\|(\theta_l^2(1+2\theta_r)\,\mathrm{Id} + \theta_l^2\theta_r h_n A_{n+1}\big\|\|y(x_{n+1})-y_{n+1}\|\\ \le\;& \big((\theta_r^2(1+2\theta_l)\|\,\mathrm{Id}\,\| \,+\, \theta_r^2\theta_l h_n\|A_n\|\big)\, c\, h_n^\gamma\\ &+ \big((\theta_l^2(1+2\theta_r)\|\,\mathrm{Id}\,\| \,+\, \theta_l^2\theta_r h_n\|A_{n+1}\|\big)\, c\, h_n^\gamma\\ =\;& \big(\theta_r^2(1+2\theta_l) + \theta_l^2(1+2\theta_r)\big)\|\,\mathrm{Id}\,\|\, c\, h_n^\gamma\\ &+ \theta_l\theta_r(\theta_r\|A_n\| + \theta_l\|A_{n+1}\|)\, c\, h_n^{\gamma+1}\,. \end{aligned}$$

Since $\theta_r = 1-\theta_l$ and $\theta_l \in [0,1]$, we get good estimates of the constants in front of the norms. A straight forward calculation/discussion shows

$$\begin{aligned} \sup_{\theta\in[0,1]} (1-\theta)^2(1+2\theta) + \theta^2(1+2(1-\theta)) &= 1\,,\\ \sup_{\theta\in[0,1]} \theta(1-\theta) &= \frac{1}{4}\,. \end{aligned}$$

Hence these estimates yield ($\theta_l+\theta_r=1$)

$$\|p(\overline{x})-\overline{y}\| \;\le\; \big(\|\,\mathrm{Id}\,\| \,+\, \tfrac{h_n}{4}\|A\|_\infty\big)\, c\, h_n^\gamma\,.$$

Since A and f are C^3, we have $y \in C^4([a,b],\mathbb{C}^d)$. Thus by Lemma 5.2.5 or [29, (6.51)] we know that there exists a $\xi \in [x_n, x_{n+1}]$, such that

$$y(\overline{x}) - p(\overline{x}) \;=\; y^{(4)}(\xi)\frac{(\overline{x}-x_n)^2(\overline{x}-x_{n+1})^2}{4!} \;=\; y^{(4)}(\xi)\tfrac{\theta_l^2\theta_r^2}{16}h_n^4\,.$$

This yields the error estimate

$$\begin{aligned} \|y(\overline{x})-\overline{y}\| &\le \|y(\overline{x})-p(\overline{x})\| \,+\, \|p(\overline{x})-\overline{y}\|\\ &\le \tfrac{\|y^{(4)}\|_\infty}{256}\, h_n^4 \,+\, \big(\|\,\mathrm{Id}\,\| \,+\, \tfrac{h_n}{4}\|A\|_\infty\big) c\, h_n^\gamma\,. \end{aligned}$$

Here we have used the estimate $\theta_r\theta_l = \theta_r(1-\theta_r) \le \frac{1}{4}$. □

Remark 8.3.2. *The explicit error estimate in Lemma 8.3.1 shows that the approximation error of y_n, y_{n+1} with respect to $y(x_n), y(x_{n+1})$ is moderately amplified. Thus the interpolation approach has some kind of robustness with respect to the perturbed data. As long as $\gamma \le 4$, the data error will be the main contribution, provided $\|y^{(4)}\|_\infty$ is not too large.*

8.4 Some classical results for linear ODEs

In this section we review the *variation of constants* approach and a *Gronwall* estimate, which we use several times in the previous chapters.

Lemma 8.4.1 (Variation of constants)**.** *Let $I \subset \mathbb{R}$ be an open interval and $A \in C(I, \mathbb{C}^{\nu\times\nu})$, $b \in C(I, \mathbb{C}^{\nu})$. The unique solution of the IVP*

$$y' \;=\; Ay + b\,, \qquad y(x_0) = y_0 \in \mathbb{C}^{\nu}$$

is given by

$$y(x) \;=\; U(x, x_0)y_0 \;+\; \int_{x_0}^{x} U(x, s)b(s)\, ds\,.$$

The evolution operator $U \in C^1(I \times I, C^{\nu\times\nu})$ fulfills for all $x, s \in I$

$$\frac{d}{dx}U(x, s) \;=\; A(x)U(x, s)\,, \qquad U(s, s) = \mathrm{Id}$$

and for all $x, s \in I$

$$U(x, s)U(s, x) \;=\; \mathrm{Id}\,.$$

This lemma is [2, Theorem 11.13].

Remark 8.4.2. *Let $\alpha := \sup_{x\in I} \|A(x)\| < \infty$. Then, by a simple Gronwall argument (cf. Lemma 8.4.3), there exists a constant β only depending on α, such that for all $x, s \in I$ it holds $\|U(x, s)\| < \beta$. Hence $\|\frac{d}{dx}U(x, s)\| \le \alpha + \beta$. Since $U(x, s)U(s, x) = \mathrm{Id}$ and $U \in C^1(I \times I, \mathbb{C}^{d\times d})$,*

$$\frac{d}{dx}U(s, x) \;=\; -U(s, x)\frac{d}{dx}U(x, s)U(s, x)\,,$$

which yields for all $x, s \in I$

$$\left\|\frac{d}{ds}U(x, s)\right\| \;\le\; \beta(\alpha)^2(\alpha + \beta(\alpha))\,.$$

Lemma 8.4.3 (Gronwall)**.** *Let $I \subset \mathbb{R}$ be an interval, $x_0 \in I$ and let the functions $\mu, \sigma, f \in C(I, \mathbb{R})$ be nonnegative. Additionally, let for all $x \in I$*

$$f(x) \;\le\; \mu(x) + \left|\int_{x_0}^{x} \sigma(s)f(s)\, ds\right|.$$

Then it holds for all $x \in I$:

$$f(x) \;\le\; \mu(x) + \left|\int_{x_0}^{x} \mu(s)\sigma(s)e^{\left|\int_s^x \sigma(\xi)\, d\xi\right|}\, ds\right|.$$

This is also from [2].

Chapter 9

Conclusion and open problems

The two–band Schrödinger models from chapter 2 have been our motivation to discuss numerical integrators for the highly oscillatory model problem from § 3.2, i. e.

$$u'(x) \;=\; \frac{1}{\varepsilon} L(x)u(x) \,+\, B(x)u(x)\,, \qquad u(a) \;=\; u_0 \in \mathbb{C}^d\,. \tag{9.1}$$

Here $L, B\colon [a,b] \to \mathbb{C}^{n\times n}$ are smooth, such that $L(x)$ is real and diagonal for all $x \in [a,b]$. We have introduced a new *analytic preprocessing* for the vector valued initial value problem (9.1), which generalizes the 'reformulation' approach from [4]. The derived transformation removes the dominant high oscillations with frequency $\sim \frac{1}{\varepsilon}$ and amplitude of $\mathcal{O}(1)$ as $\varepsilon \to 0$ of the solution u. Despite the fact that (in general) high oscillations are still present in the transformed variable z, the gained equivalent initial value problem

$$z'(x) \;=\; \varepsilon^n A_n(x) z(x)\,, \qquad z(a) \;=\; z_0 \in \mathbb{C}^d \tag{9.2}$$

is much better suited for numerical treatment than (9.1). Here the solution z oscillates (also with frequency $\sim \frac{1}{\varepsilon}$) around the initial condition z_0, but with amplitudes of $\mathcal{O}(\varepsilon^n)$ as $\varepsilon \to 0$, instead of $\mathcal{O}(1)$ as for u.

We have also shown that the transformation can be derived from an *asymptotic expansion* of the solution u from (9.1), which we established in § 3.5. Our expansion can be interpreted as the vector valued analogon of the WKB–method, which is well known for the scalar stationary Schrödinger equation (9.3). The scalar case is incorporated in our model problem (9.1) for strictly positive potentials V of the ODE (9.3).

In § 3.3.1 we have shown that our approach is a generalization of the ansatz used in [4]. Furthermore, in § 3.3.2 we discussed the differences to the super adiabatic transformations (SAT) introduced in [27]. Our new transformation approach and the SAT yield very good results when staying away from *avoided eigenvalue crossing* of the matrix $L(x)$ from (9.1). Away from this points the asymptotic expansions, upon which the methods are based, are quite accurate. But close to an avoided crossing, a new smaller scale has to be introduced, possibly combined with an adaptive choice of the step size for the numerical

computations (cf. § 4.4 and [27, §XIV]). However, the limiting case of an eigenvalue crossing can be understood with the *Landau–Zener formula* [26].

Furthermore, the transformation of the (one dimensional) scalar *stationary Schrödinger* equation

$$\psi''(x) + \frac{1}{\varepsilon^2} V(x)\,\psi(x) = 0\,, \tag{9.3}$$

to (9.1) (as discussed in § 2.2) and the asymptotic expansion of the related solution u from (9.1) break down, if we approach a *turning point* of the differential equation (9.3). A turning point is a zero of the function V from (9.3). Also here it is not yet clear how to modify the discussed transformation approach and the asymptotic expansion of the solution, such that we can deal with this situation. For a simple turning point (i. e. a simple zero of V) there exists an asymptotic expansion of the solution ψ from (9.3), which is based on *Airy functions*. It seems to be a globally valid asymptotic approximation, i. e. it is asymptotically correct in the oscillatory regime ($V > 0$), in the part where ψ exponentially growths and decays ($V < 0$, also called *evanescent region*), and in the transition layer as $\varepsilon \to 0$. An explicit formula for the expansion can be found in [32, p.179]. Based on this one may derive a transformation approach for (9.3) in the presence of a simple turning point. It is also worth to notice, that the scalar second order equation for ψ can be transformed into a semi linear first order differential equation, a so called *Riccati equation*. This can be done[1] by the ansatz $\psi = e^{\frac{1}{\varepsilon}\int \rho\, dx}$, which yields

$$\varepsilon\rho' + \rho^2 + V = 0\,. \tag{9.4}$$

If is ψ highly oscillatory, then the solution ρ of (9.4) has a large imaginary part, while in the evanescent region the real part of ρ dominates. Beside this difference, we expect both regions to be equal in this representation. I. e. we expect that the function ρ shows a similar growth behavior for the evanescent and oscillatory parts. Hence this may allow a uniform approximation of ψ via ρ as $\varepsilon \to 0$.

In § 6 we have derived efficient marching methods for the initial value problem (9.2) and proved convergence of the methods. The derived schemes are *asymptotically correct* as $\varepsilon \to 0$. To be more precise, the approximation error is at most of order $\mathcal{O}(\varepsilon^n)$, even for a fixed spatial grid. Moreover we have shown that, in the whole "zoo" of one–step methods, there are integrators with an approximation error of maximal asymptotic order (with respect to ε) of $\mathcal{O}(\varepsilon^{2n+1}h^m)$. Here h is the maximal step size of the grid. While n is a prescribed value, coming from the initial value problem, $m \in \mathbb{N}$ is (more or less) arbitrary. In applications ε is a very small constant and hence these integrators can use very large grids compared to the local wavelength which is $\sim \varepsilon$. In this theses we have not considered conservative, reversible, or symplectic methods. But since our approach is quite general, it should be possible to construct such integrators with the discussed tools.

One essential ingredient for the construction of the efficient one–step meth-

[1] Since V is real valued, there exists a real valued fundamental system of solutions ψ_1, ψ_2 of (9.3). Hence $\psi = \psi_1 + i\psi_2$ has no zeros and thus we can write $\psi = e^{\frac{1}{\varepsilon}\int \rho\, dx}$

ods in § 6 is an *advanced quadrature* for highly oscillatory integrals of the form

$$I \;:=\; \int_{\alpha}^{\beta} f(x) e^{-\frac{i}{\varepsilon}\varphi(x)}\, dx \;=\; \int_{\varphi(\alpha)}^{\varphi(\beta)} \frac{f(\varphi^{-1}(\xi))}{\varphi'(\varphi^{-1}(\xi))}\, e^{-\frac{i}{\varepsilon}\xi}\, d\xi\,. \qquad (9.5)$$

We have chosen an approach from literature (cf. [60, 61]). For this ansatz we have established an error analysis, which yields error estimates in terms of the interval length $|\alpha - \beta|$. As far as we know this has not yet been done. We have also been able to improve the quadrature used in [4]. Our newly derived version yields a much better asymptotic accuracy (with respect to the interval length) than the original, but with the same numerical effort. However, the discussed quadratures are not designed for integrals with *stationary points*, which appear if we allow crossing eigenvalues of L in (9.1). To deal with this case other methods have to be used, which are already available (cf. [35]). It is also important to analyze this quadratures in the presence of "avoided stationary points", which show up when we approach an avoided eigenvalue crossing of L. One idea to create a more robust quadrature, i. e. a quadrature which admits error estimates which are valid for functions without and with (avoided) stationary points, is to approximate the term $\frac{f(\varphi^{-1})}{\varphi'(\varphi^{-1})}$ of the integrand in (9.5) by rational functions instead of polynomials, as we have done for our method of choice.

The approximation procedures to discretize the WKB–type transformation from § 3.3, as discussed in § 4, are derived in a straight forward way. Hence we believe that there is some space for improvement. Also we have not yet established a rigorous error analysis of the described methods. This is dedicated to future work.

Chapter 10

The one way wave equation

The final chapter of this thesis is not (directly) related to the previous work. It is a collection of few results the author derived during the beginning of his doctoral studies. The chapter heading is just a part of the topic we were dealing with. We intended to derive (possible finite) difference schemes for a *one way wave equation* (cf. (10.18)), which originates form a two dimensional Helmholtz (–type) equation (cf. (10.14)-(10.18)). Unfortunately, it turns out that the intended techniques and ideas do not work. Hence, this chapter mainly contains results from literature. A large part are revised lecture notes of a short course the author taught at the Wissenschaftskolleg Differential Equations[1] Summer Camp in 2007. Nevertheless, it turns out that the numerical methods discussed in the previous chapters may be used to derive integrators for the one way wave equation. Due to lack of time this is not discussed in this work. We shall only state the basic ideas.

We shall continue to point out the basic idea of a one way wave equation. Wave phenomena in electrodynamics or acoustics are often well described by the *Wave Equation* ($x \in \mathbb{R}^d$, $t \in \mathbb{R}$)

$$\Delta_x \psi(x,t) \; - \; \frac{1}{c^2(x,t)} \frac{\partial^2}{\partial t^2} \psi(x,t) \quad = \quad 0 \, .$$

The quantity c is the local wave propagation velocity and ψ (e. g.) describes the strength of the electric potential (electrodynamics) or the pressure (acoustics). If c does not depend on time, the separation ansatz[2] $\psi(x,t) = e^{i\omega t} \widehat{\psi}(x)$ yields the *Helmholtz equation* (HE)

$$\Delta_x \widehat{\psi}(x) \; + \; \frac{\omega^2}{c^2(x)} \widehat{\psi}(x) \quad = \quad 0 \, . \tag{10.1}$$

Let $z = (x_2, \ldots, x_d) \in \mathbb{R}^{d-1}$ and define

$$A(x_1, z) \quad := \quad \Delta_z \; + \; \frac{\omega^2}{c^2(x_1, z)} \, .$$

If the waves, we are modeling, mainly propagate in x_1–direction, it may be more convenient to reformulate (10.1) as a first order evolution problem (with respect

[1] A PhD program of the Vienna University of Technology and the University of Vienna.
[2] Alternatively one can apply the Fourier–transformation in t.

to x_1). This can be (formally) done by using the ansatz (2.24) from § 2.2.1:

$$v_1 \;:=\; A^{\frac{1}{2}}\psi\,, \qquad v_2 \;:=\; \frac{\partial}{\partial x_1}\psi\,. \tag{10.2}$$

Let us denote the partial derivate with respect to x_1 by $'$, i. e. $\psi' = \frac{\partial}{\partial x_1}\psi$. Then

$$v' \;=\; \begin{pmatrix} 0 & A^{\frac{1}{2}} \\ -A^{\frac{1}{2}} & 0 \end{pmatrix} v + \begin{pmatrix} [\partial_{x_1}, A^{\frac{1}{2}}] & 0 \\ 0 & 0 \end{pmatrix} v\,. \tag{10.3}$$

Since zero may be part of the spectrum of the operator A, the reformulation (10.2) may work only in one direction. However, it is well defined as long as the square root of A and its commutator with ∂_{x_1} are well defined. Thus we should regard (10.3) as a necessary condition for ψ to be a solution of the HE.

The first operator–matrix of (10.3) can be given a block diagonal structure by setting

$$u \;=\; \frac{1}{\sqrt{2}} \begin{pmatrix} i & 1 \\ 1 & i \end{pmatrix} v\,.$$

This yields

$$u' \;=\; \begin{pmatrix} iA^{\frac{1}{2}} & 0 \\ 0 & -iA^{\frac{1}{2}} \end{pmatrix} u + \frac{1}{2} \begin{pmatrix} 1 & i \\ -i & 1 \end{pmatrix} \otimes [\partial_{x_1}, A^{\frac{1}{2}}]\, u\,. \tag{10.4}$$

A similar ansatz is described in the final part of [69, §3].

If $[\partial_{x_1}, A^{\frac{1}{2}}]$ is zero (e. g. c does not depend on x_1), then (10.4) decouples in the forward and backward *one way wave equation* (OWWE)

$$u_1' \;=\; iA^{\frac{1}{2}}u_1 \quad \text{and} \quad u_2' \;=\; -iA^{\frac{1}{2}}u_2\,. \tag{10.5}$$

In this case it holds $\Delta_x + \frac{\omega^2}{c^2(x)} = (\partial_{x_1} + iA^{\frac{1}{2}})(\partial_{x_1} - iA^{\frac{1}{2}})$. Hence each solution u_1, u_2 of (10.5) is also a solution of the HE (10.1). Thus it is of interest to construct efficient integrators for equations of type (10.5). Furthermore, the structure of (10.4) is similar to the problem (3.23) from § 3.3. Thus, if we have efficient integrators for (10.5), it may be possible to adapt the approach for the matrix valued problem from § 3.3 to (10.4) .

In order to get an idea at which state of model reduction the OWWE may arise, in § 10.1 we (briefly) discuss its derivation for acoustic waves in an inviscid, compressible fluid. Here we start with physical conservation laws and an equation of state. During the course of discussion we shall make certain assumptions, which simplifies the problem and shall end up with the OWWE.

Most difficulties of dealing with the full OWWE (10.5) arise from the fact that it contains a square root of a self–adjoint, indefinite differential operator. In literature (cf. [21, 22, 15]) it is often treated as *pseudo differential* or *Fourier integral operator*. In § 10.2 we give a (very brief) definition of pseudo differential operators and briefly discuss some strategies to discretize the "fractional" differential operator $A^{\frac{1}{2}}$, which is sometimes called *square root Helmholtz operator* (SRHO) (cf. [22, 20]).

An alternative (and from the authors point of view more natural) way to the pseudo differential operator ansatz is the definition of $A^{\frac{1}{2}}$ by functional calculus.

The drawback of the standard approach via C^*–algebras is that it is not well suited for numerical treatment. In § 10.3 we establish an elegant method from [71] for the computation of functions of self–adjoint operators. This ansatz seems to be more useful for numerics. We try to give a self consistent prove. In the last section § 10.4 we (most times formally) apply the method from § 10.3.5 for some problems arising from the OWWE; computing the square root of an self–adjoint operator or a formal solution of the OWWE. We are also able to deduce *De Santo's transformation* [12], which is a crucial tool for constructing the exact operator symbol (needed for the representation of $A^{\frac{1}{2}}$ as a pseudo differential operator) of the SRHO as described in [21].

10.1 From physics to the one way wave equation

In the sequel we formally derive a OWWE for wave propagation in an inviscid, compressible fluid. We sketch the way from the basic equations of fluid mechanics to a OWWE. This in done in several model reduction steps. The whole procedure outlined below (strongly) follow the textbook [24].

The fluid is described via the density ρ, pressure P, entropy S, and particle velocity $\mathbf{v} \in \mathbb{R}^3$ and the governing equations:

- *Euler's Equation* (momentum balance)

$$\rho \frac{d\mathbf{v}}{dt} \quad = \quad -\nabla_x P\,, \tag{10.6}$$

- *Equation of Continuity* (mass conservation)

$$\frac{\partial \rho}{\partial t} + \operatorname{div}_x \rho\mathbf{v} \quad = \quad 0\,, \tag{10.7}$$

- *Adiabatic Condition* (no heat transfer)

$$\frac{\partial S}{\partial t} + \mathbf{v} \cdot \nabla_x S \quad = \quad 0\,, \tag{10.8}$$

- Equation of State

$$P \quad = \quad P(\rho, S)\,. \tag{10.9}$$

As usual t denotes time and $x \in \mathbb{R}^3$ is the free spatial variable. Under the assumption, that the quantities ρ, P, and S vary only very little around a steady state (with respect to time, but may be spatial dependent) it is reasonable to make the ansatz

$$\begin{aligned} P(x,t) &= P_0(x) + p(x,t)\,, \qquad \rho(x,t) = \rho_0(x) + \overline{\rho}(x,t)\,, \\ S(x,t) &= S_0(x) + s(x,t)\,. \end{aligned}$$

The velocity $\mathbf{v}$ is not treated in this fashion, because we do not assume a mean flow of the particles. Now the system of equations (10.6)–(10.9) is linearized at ρ_0, P_0, and S_0. For this we insert the ansatz into the equations and omit all

quadratic and higher terms. Afterwards we eliminate the entropy of the linear equations for the first order terms, which finally leads to[3]

$$\rho_0(x)\,\mathrm{div}_x\left(\frac{1}{\rho_0(x)}\nabla_x\,p(x,t)\right) - \frac{1}{c^2(x,t)}\frac{\partial^2 p(x,t)}{\partial t^2} \quad = \quad 0\,. \qquad (10.10)$$

This is the "time–dependent acoustic wave equation with density and sound velocity stratification and no sources" [24]. The quantity c is the sound velocity and is given by

$$c^2(x,t) \quad = \quad \left.\frac{\partial P_0}{\partial \rho_0}\right|_S .$$

If c is independent of time t, we can apply the Fourier transformation $(t \to \omega)$ at (10.10), which yields the *Helmholtz Equation* (HE) in the following form

$$\rho_0(x)\nabla_x\left(\frac{1}{\rho_0(x)}\nabla_x\,\widehat{p}(x,w)\right) + \frac{w^2}{c^2(x)}\,\widehat{p}(x,w) \quad = \quad 0\,. \qquad (10.11)$$

By the ansatz $\psi = \frac{\widehat{p}}{\sqrt{\rho_0}}$ the HE (10.11) is transformed to *standard form*

$$\Delta\psi(x,w) \,+\, V^2(x,w)\,\psi(x,w) \quad = \quad 0\,. \qquad (10.12)$$

The quantity V (total acoustic wave number, cf. [24]) is given by

$$V(x,w) \quad = \quad \frac{w^2}{c^2(x)} \,+\, \frac{1}{2\rho_0(x)}\Delta\rho_0(x) \,-\, \frac{3}{4}\left(\frac{1}{\rho_0(x)}|\nabla\rho_0(x)|\right)^2 .$$

We assume cylindrical symmetry (further model reduction) and introduce cylindrical coordinates:

$$x(r,\varphi,z) \quad = \quad \begin{pmatrix} r\,\cos\varphi \\ r\,\sin\varphi \\ z \end{pmatrix} .$$

Thus the HE (10.12) reads

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\right)\psi \,+\, \frac{1}{r^2}\frac{\partial^2\psi}{\partial\varphi} \,+\, \frac{\partial^2\psi}{\partial z^2} \,+\, V^2\psi \quad = \quad 0\,.$$

Due to cylindrical symmetry, i. e. $\psi = \psi(r,z)$, the PDE simplifies to

$$\frac{1}{r}\frac{\partial\psi}{\partial r} \,+\, \frac{\partial^2\psi}{\partial r^2} \,+\, \frac{\partial^2\psi}{\partial z^2} \,+\, V^2\psi \quad = \quad 0\,. \qquad (10.13)$$

To remove the singularity at $r = 0$ in front of the first derivative with respect to r, we make the ansatz (cf. [69, 41, 24])

$$\psi(r,z) \quad = \quad u(r,z)\,H_0^{(1)}(\kappa r)\,,$$

where $\kappa \in \mathbb{R}$ is a positive constant and $H_0^{(1)}$ is the *Hankel function* of first kind first order. It solves the *Bessel differential equation* (cf. [51])

$$y'' \,+\, \frac{1}{z}y' \,+\, y \quad = \quad 0\,.$$

[3]For more (but not all) details we refer to [24] p.13ff.

Thus we deduce from (10.13)

$$\frac{\partial^2 u}{\partial r^2} + \left(\frac{2}{H_0^{(1)}(\kappa r)} \frac{\partial H_0^{(1)}}{\partial r}(\kappa r) + \frac{1}{r} \right) \frac{\partial u}{\partial r} + \frac{\partial^2 u}{\partial z^2} + \kappa^2 (V^2 - 1) u \quad = \quad 0 \,.$$

The next model reduction step is the *far field approximation* (cf. [69, 24]). Here we assume that $\kappa r \gg 1$. From the asymptotic expansion of the Hankel function $H_0^{(1)}$ we get [51, 69]

$$\frac{2}{H_0^{(1)}} \frac{\partial H_0^{(1)}}{\partial r}(\kappa r) + \frac{1}{r} \quad = \quad 2i\kappa \big(1 + \mathcal{O}(\kappa^{-2} r^{-2}) \big) \,.$$

In the far field, i. e. $\kappa r \gg 1$, we obtain $\frac{2}{H_0^{(1)}} \frac{\partial H_0^{(1)}}{\partial r}(\kappa r) + \frac{1}{r} \sim 2i\kappa$ and hence

$$\frac{\partial^2 \tilde{u}}{\partial r^2} + 2i\kappa \frac{\partial \tilde{u}}{\partial r} + \frac{\partial^2 \tilde{u}}{\partial z^2} + \kappa^2 (V^2 - 1)\, \tilde{u} \quad = \quad 0 \,. \tag{10.14}$$

If V does not depend on r (this is the last model reduction step) equation (10.14) can be (formally) factorized. For this let

$$\partial_r \;:=\; \frac{\partial}{\partial r} \quad \text{and} \quad A \;:=\; \frac{\partial^2}{\partial z} + \kappa^2 V^2(z) \,. \tag{10.15}$$

Provided $\sqrt{A}$ exists, it holds

$$\begin{aligned} &[\partial_r + i\kappa + i\sqrt{A}][\partial_r + i\kappa - i\sqrt{A}]\, \tilde{u} \\ &\quad = \quad \big[\partial_r^2 + 2i\kappa\partial_r - \kappa^2 + A\big] \tilde{u} \;-\; i\kappa [\partial_r, \sqrt{A}] \tilde{u} \,. \end{aligned} \tag{10.16}$$

Since A commutes with ∂_r, the same holds for $\sqrt{A}$. Hence the commutator in (10.16) is zero, i. e. the right hand side of (10.16) coincides with the left hand side operator of (10.14). Thus a solution $\tilde{y}$ of the *one way wave equation* (OWWE)

$$\frac{\partial \tilde{y}}{\partial r}(r, z) \quad = \quad i(\sqrt{A} - \kappa) \tilde{y} \tag{10.17}$$

is also a solution of the far field approximation (10.14). The constant term $-i\kappa$ can easily be removed by the final transformation

$$y(r, z) \quad = \quad e^{i\kappa r} \tilde{y}(r, z) \,,$$

which yields

$$\frac{\partial y}{\partial r}(r, z) \quad = \quad i\sqrt{A(z)}\, y(r, z) \,. \tag{10.18}$$

Up to now we only considered the PDE without specifying boundary conditions (BC). Instead of discussing explicit examples we rather assume that the BC are, such that A is self–adjoint on a suitable Hilbert space. In this case the operator $\sqrt{A}$ is defined via the functional calculus for self–adjoint unbounded operators. In literature the operator $\sqrt{A}$ is called the *square root Helmholtz operator* [21].

It is usually interpreted as a *pseudo differential* or *Fourier integral operators* (cf. § 10.2).

Since the solution y of (10.18) should be bounded for all $r \geq 0$, we have to specify the right branch of the complex square root. Assume y_0 is an eigenvector of A to the eigenvalue λ. Since A is self–adjoint, λ is real. Further $y(r) = e^{i\sqrt{\lambda}r}y_0$ is a solution of the OWWE with initial condition y_0. Hence we need $\sqrt{x} = i\sqrt{|x|}$ for $x < 0$.

10.2 Some remarks on $\sqrt{A}$ and the OWWE

In Literature the square root Helmholtz operator (SRHO)

$$A \quad := \quad \sqrt{\frac{\partial^2}{\partial z} + \kappa^2 V^2(z)}$$

is considered as a *pseudo differential operator* (PDO) [21, 22, 15]. An introduction to the theory of pseudo differential operators can be found in [71]. They are defined via the Fourier transformation. Here we shall give only a brief (formal) definition for the one dimensional case. Let $\Omega_B \colon \mathbb{R} \times \mathbb{R} \to \mathbb{C}$ be a given function of class C^∞, which fulfills a growth condition of the form

$$|D_x^\beta D_\xi^\alpha \Omega_B(x,\xi)| \quad \leq \quad C_{\beta,\alpha}(1+|\xi|^2)^{\frac{m-\rho|\alpha|+\delta|\beta|}{2}}$$

for some $\rho, \delta \in [0,1]$ and $m \in \mathbb{R}$. Then a PDO B can be defined by

$$(Bu)(z) \quad = \quad \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \Omega_B(z,\zeta)\, \widehat{u}(\zeta)\, e^{i\zeta z}\, d\zeta\,.$$

As usual we denote by $\widehat{u}$ the Fourier transform of u (see (10.24) in § 10.3.2). The function Ω_B is called the symbol (or left symbol) of the operator B. For example the symbol of $B = \partial_z^2$ is $\Omega_B = -|\zeta|^2$. There are also other representations of PDOs. Another important one is the *Weyl representation* [71]

$$(Bu)(z) \quad = \quad \frac{1}{2\pi} \int_{\mathbb{R}} \int_{\mathbb{R}} \Omega_B^W \left(\frac{z+y}{2}, \zeta \right) u(y) e^{i(z-y)\zeta}\, dy\, d\zeta\,,$$

which is often used for the analysis of PDO. The Weyl representation is used in [22] for a discussion of the SRHO.

If we denote by $\mathcal{F}$ the Fourier transformation on $L^2(\mathbb{R})$, then we can write

$$(Bu)(z) \quad = \quad \mathcal{F}^{-1}\Omega_B(z,\cdot)\mathcal{F}u\,,$$

or (with $u = \mathcal{F}^{-1}g$)

$$(\mathcal{F}B\mathcal{F}^{-1}g)(z,\zeta) \quad = \quad \Omega_B(z,\zeta)g(\zeta)\,.$$

Hence the Fourier transformation (in some sense) "diagonalizes" the operator B. If Ω_B does not depend on z, then $\Omega_{B^2} = \Omega_B^2$ and hence $\Omega_{\sqrt{B}} = \sqrt{\Omega_B}$.

In [20] one finds the outline of the exact symbol construction procedure for the SRHO. It is based on the Greens function for the Helmholtz equation or a fundamental solution of a related Schrödinger equation (SE) respectively. In general both, the Greens function and the SE solution, are unknown which

makes this construction approach not very useful for explicit computations. Nevertheless it is successfully used in [21] to derive a uniform high–frequency approximation of the operator symbol, using the WKB[4] approximation of the SE solution.

To numerically solve the OWWE we have to derive a (finite) discretization of the SRHO. Since we theoretically know its symbol (cf. [20]), we can use the ansatz from [56]. The approach to approximate a PDO, as presented in the article, is based on the following (simple) idea: Let $\Omega_{\tau_a}(z,\zeta) = e^{ia\zeta}$ for some $a \in \mathbb{R}$. Then

$$(\tau_a u)(z) \;=\; \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \widehat{u}(\zeta)\, e^{i(z+a)\zeta}\, d\zeta \;=\; u(z+a)\,.$$

Hence, to discretize the PDO B, one seeks a good approximation of the operator symbol Ω_B by trigonometric polynomials. In general Ω_B is not a periodic function and thus there are two approximation steps to do. Firstly truncate the operator symbol by multiplying it with a suitable cut–off function $\eta^h(\zeta) = \eta(h\zeta)$. Here $\eta \in C^\infty(\mathbb{R})$ is bounded from above an below by $0 \le \eta \le 1$ and

$$\eta(\zeta) \;=\; \begin{cases} 1, |\zeta| < \frac{1}{2}\pi\,, \\ 0, |\zeta| > \frac{3}{4}\pi\,. \end{cases}$$

Hence for every fixed $z \in \mathbb{R}$ we have $\mathrm{supp}(\Omega_B(z,\cdot)\eta^h) \subset [-\frac{3}{4}\frac{\pi}{h}, \frac{3}{4}\frac{\pi}{h}] \subset [-\frac{\pi}{h}, \frac{\pi}{h}]$. Thus, the $\frac{2\pi}{h}$ periodic extension $\Omega_B^h(z,\cdot)$ of $\Omega_B(z,\cdot)\eta^h$ can be expressed in a Fourier series:

$$\Omega_B^h(z,\zeta) \;=\; \sum_{j\in\mathbb{Z}} c_{j,h}(z)\, e^{i\zeta jh} \quad \text{with} \quad c_{j,h}(z) \;=\; \frac{h}{2\pi} \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \Omega_B(z,\zeta)\, e^{i\zeta jh}\, d\zeta\,.$$

This yields

$$\begin{aligned} (Bu)(z) \;&=\; \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \Big(\sum_{j\in\mathbb{Z}} c_{j,h}(z)\, e^{i\zeta jh} \Big) \widehat{u}(\zeta)\, e^{i\zeta z}\, d\zeta \\ &=\; \sum_{j\in\mathbb{Z}} c_{j,h}(z) u(z+jh)\,. \end{aligned}$$

Thus the operator B is approximated by a (in general non–finite) difference operator. Hence the second approximation step is to truncate the Fourier series. This yields a finite difference operator. Let us consider a simple example, related to the SRHO, in order to point out possible difficulties of this approach. Let $\Omega_B(\zeta) := 1 - \zeta^2$ be the symbol of $B = 1 + \partial_z^2$ and let

$$\Omega_R(\zeta) \;:=\; \chi_{[-1,1]}(\zeta)\sqrt{1-\zeta^2}\,.$$

Thus Ω_R is the real part of $\sqrt{\Omega_B} = \Omega_{\sqrt{B}}$. Since $\mathrm{supp}(\Omega_R) = [-1,1]$ it holds for all $0 < h < \frac{\pi}{2}$

$$\Omega_R^h(\zeta) \;=\; \eta(h\zeta)\Omega_R(\zeta) \;=\; \Omega_R(\zeta)\,,$$

[4] An asymptotic approximation technique for the Schrödinger equation, named after the physicists Wentzel, Kramers and Brillouin. See § 3.5 for the basic concept of it.

for all $\zeta \in [-\frac{\pi}{h}, \frac{\pi}{h}]$. Hence we can compute the Fourier coefficients $c_{j,h}$ of Ω_R^h without specifying the truncation function η. We get (with *Maple14*)

$$\Omega_R^h(\zeta) \quad = \quad \frac{h}{4} + \sum_{j \in \mathbb{Z}\setminus\{0\}} \frac{\mathrm{BesselJ}_1(jh)}{2j}\, e^{i\zeta jh}\,.$$

Here $\mathrm{BesselJ}_1$ denotes the *Bessel* function of first kind first order (cf. [51]). To get a finite difference approximation we have to truncate the series. To this end we define for $N \in \mathbb{N}$

$$\Omega_R^{h,N}(\zeta) \quad = \quad \frac{h}{4} + \sum_{0 \neq j = -N}^{N} \frac{\mathrm{BesselJ}_1(jh)}{2j}\, e^{i\zeta jh}\,.$$

Thus, the corresponding PDO is

$$(R^{h,N}u)(z) \quad = \quad \frac{h}{4}\, u(z) + \sum_{0 \neq j = -N}^{N} \frac{\mathrm{BesselJ}_1(jh)}{2j}\, u(z + jh)\,.$$

Since $\mathrm{BesselJ}_1(x) \to 0$ as $x \to 0$ it follows for all continuous u

$$\lim_{h \to 0} (R^{h,N}u)(z) \quad = \quad 0\,.$$

Hence N and h have to be coupled in the right way in order to get the correct limit as $h \to 0$. This makes it even more difficult to use this approach, beside computing the operator symbol.

An alternative procedure to derive an approximate solution to the OWWE is to replace $\sqrt{A}$ by a formal Taylor or Padé approximation (see [14] for further references). This leads to so called (wide) angle parabolic equations, which are (ordinary) partial differential equations. In the case of first order Taylor approximation one gets an equation of Schrödinger type. How the approximation error can be quantified in not yet clear to the author.

Another strategy to discretize the OWWE may be based on a discretization of A. This means one firstly approximate A by a suitable finite difference A^h, which yields a second order system of ordinary differential equations for the Helmholtz equation. Afterwards one determines a square root of the finite difference operator, which is in general a non–local difference operator, i. e. it has an unbounded numerical stencil. Since the factorization of the discrete Helmholtz equation is analogue to the continuous case, we only have to replace $\sqrt{A}$ by $\sqrt{A^h}$ in the OWWE. This yields a (probably infinite) system of ordinary differential equations. Let us consider this approach for the partial differential operator $A = \partial_z^2 + c^2$, with a fixed positive constant c. Furthermore let the difference operator $A^h := D_h^2 + c^2$ be a discretization of A with the central second difference. The identity

$$\begin{aligned} \frac{u(z+h) - 2u(z) + u(z-h)}{h^2} \quad &= \quad \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \frac{e^{i\zeta h} - 2 + e^{-i\zeta h}}{h^2} \widehat{u}(\zeta) e^{i\zeta z}\, d\zeta \\ &= \quad -\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \frac{4\sin^2(\zeta\frac{h}{2})}{h^2} \widehat{u}(\zeta) e^{i\zeta z}\, d\zeta \end{aligned}$$

yields

$$\begin{aligned}
&(\sqrt{A}u)(z) - (\sqrt{A^h}u)(z) \\
&= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \left(\sqrt{c^2 - \zeta^2} - \sqrt{c^2 - \frac{\sin^2(\zeta \frac{h}{2})}{(\frac{h}{2})^2}} \right) \widehat{u}(\zeta)\, e^{i\zeta z}\, d\zeta \\
&=: \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \Delta(h,\zeta)\, \widehat{u}(\zeta)\, e^{i\zeta z}\, d\zeta\,.
\end{aligned}$$

At the crucial points $\zeta_0 = \pm c$ it holds

$$\lim_{h\to 0} \frac{\Delta(h,\pm c)}{c^2 h} \;=\; -\frac{c^2}{2\sqrt{3}}\,,$$

while

$$\lim_{h\to 0} \frac{\Delta(h,\zeta)}{\zeta^3 h^2} \;=\; -\frac{1}{24}\frac{\zeta}{\sqrt{c^2-\zeta^2}}$$

holds for $\zeta \in \mathbb{R}\backslash\{\pm c\}$. Hence if $\pm c \notin \operatorname{supp}(\widehat{u})$, then the difference is of order $\mathcal{O}(h^2)$. Otherwise the order reduces to $\mathcal{O}(h)$, especially for functions $\widehat{u}$ "strongly" located at $\pm c$.

The problem of order reduction can be fixed by using the difference operator

$$\tilde{A}^h \;:=\; D_h^2 \,+\, \frac{\sin^2(c\frac{h}{2})}{(\frac{h}{2})^2}\,.$$

We get for all $\zeta \in \mathbb{R}$:

$$\lim_{h\to 0} \frac{\tilde{\Delta}(h,\zeta)}{h^2} \;=\; \frac{1}{24}(c^2+\zeta^2)\sqrt{c^2-\zeta^2}\,.$$

Hence $\delta(h,\zeta) := \frac{\Delta(h,\zeta)}{(1+\zeta^2)^{3/2}h^2}$ is a bounded function and for $u \in H^3(\mathbb{R})$ it holds

$$\begin{aligned}
\left\| \sqrt{A}u - \sqrt{\tilde{A}^h}u \right\|_{L^2} &= \|\mathcal{F}^{-1}(1+\zeta^2)^{\frac{3}{2}} h^2 \delta(h,\zeta) \mathcal{F}u\|_{L^2} \\
&\le h^2 \|\delta(h,\zeta)\|_{L^\infty} \|(1+\zeta^2)^{\frac{3}{2}}\widehat{u}\|_{L^2} \;\le\; c\,h^2\,\|u\|_{H^3}\,.
\end{aligned}$$

This uniform estimate is due to the fact that the symbol of the difference operator coincide with the symbol of the differential operator at the crucial points $\zeta = \pm c$. Thus, this last procedure (choose a discretization of ∂_z^2 and then compute the square root) is also not a reliable method with respect to the expected convergence order, as we have seen in the very simple case of a constant potential.

Due to the described problems of the different approaches we want to derive an approximation of $\sqrt{A}$ directly from the operator itself. Therefore we first have to find a way to compute $\sqrt{A}$. In the application we have in mind (OWWE) the operator A is self–adjoint. Hence $\sqrt{A}$ is defined via functional calculus. Thus, we shall discuss in § 10.3 a non standard approach (this means an ansatz which does not use the theory of Banach algebras) for the functional calculus of self–adjoint operators. It seems to be well suited for numerical considerations.

10.3 Functions of self–adjoint operators

The following subsections are revised lecture notes of a short course the author taught at the Wissenschaftskolleg Differential Equations[5] Summer Camp in 2007.

The aim of this section is the proof of a spectral theorem for self–adjoint operators on a separable Hilbert space. We reproduce an approach from [71], especially § 10.3.5 is very close to the textbook. The underlying idea is as follows. Let the continuous function $f\colon \mathbb{R} \to \mathbb{R}$ be in $L^1(\mathbb{R})$, such that its Fourier transformation

$$\widehat{f}(t) \;=\; \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(x)\, e^{-itx}\, dx$$

is in $L^1(\mathbb{R})$ too. Then for any $a \in \mathbb{R}$ we can evaluate f at $x = a$ by the formula

$$f(a) \;=\; \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \widehat{f}(t)\, e^{iat}\, dt\,.$$

The oscillatory part of the integrand

$$u(t) \;:=\; e^{iat}$$

is the unique solution of the initial value problem

$$u_t \;=\; ia\, u(t)\,, \qquad u(0) \;=\; 1\,.$$

Now assume that $A \in \mathbb{R}^{d\times d}$ is a real valued, self–adjoint matrix and let $U(t)$ be the unique solution of the IVP

$$U_t \;=\; iA\, U(t)\,, \qquad U(0) \;=\; \mathrm{Id}\,. \tag{10.19}$$

Here Id denotes the identity matrix on $\mathbb{R}^{d\times d}$. Since A is self–adjoint, there exists an orthogonal matrix Q (cf. [19, 7]), such that

$$A \;=\; Q^*\Lambda Q\,.$$

The diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ contains the eigenvalues of A, which are real. Hence it holds for all $t \in \mathbb{R}$

$$U(t) \;=\; Q^* \exp(i\Lambda t) Q \;=\; Q^* \,\mathrm{diag}(e^{i\lambda_1 t}, \ldots, e^{i\lambda_d t}) Q\,.$$

Since $\lambda_1, \ldots, \lambda_d$ are real, $U(t)$ is unitary for all $t \in \mathbb{R}$. Thus the integrals

$$\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \widehat{f}(t) \exp(i\lambda_j t)\, dt$$

are well defined for $j = 1, \ldots, d$ and it follows

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} \int \widehat{f}(t)\, U(t)\, dt \;&=\; Q^* \frac{1}{\sqrt{2\pi}} \int \widehat{f}(t)\, \exp(i\Lambda t)\, dt Q & (10.20)\\ &=\; Q^* \,\mathrm{diag}(f(\lambda_1), \ldots, f(\lambda_d)) Q\,. & (10.21) \end{aligned}$$

[5] A PhD program of the Vienna University of Technology and the University of Vienna.

The right–hand side of (10.21) is nothing but $f(A)$ as discussed in [34] §6.2. Hence the left hand side of (10.20) yields a way to compute $f(A)$ without diagonalizing the matrix A, provided one can solve (10.19). As a next step one can generalize this ansatz to bounded, self–adjoint operators. It turns out that the main property for the approach to work is that $U(t)$ is a "continuous" family of unitary operators. By *Stone's Theorem* 10.3.7 any self–adjoint operator is the generator of a C_0 group of unitary operators. Thus it is not surprising that one can extend the ideas to general self–adjoint operators.

In § 10.3.1 we give a short introduction into the theory of semigroups in order to explain Stone's Theorem 10.3.7, which is stated without a proof. Before we can start to prove the spectral theorem in § 10.3.5 we need some technical results about distributions which are collected in § 10.3.2. Furthermore we shall prove a version of *Riesz–representation theorem* in § 10.3.3 and in § 10.3.4 we shall show that the test–functions are dense in any L^p space ($1 \leq p < \infty$) corresponding to a Borel measure on an open subset of $\mathbb{R}^n$. These two results are also crucial parts of the proof of the spectral theorem.

10.3.1 Semigroups of Linear Operators

The main part of this section is based on [62].

Let $A \in \mathbb{R}^{n\times n}$ be a quadratic matrix and for $t \in \mathbb{R}$ let $T(t) := e^{At}$. Then the unique solution of the initial value problem $u' = Au$, $u(0) = u_0 \in \mathbb{R}^n$ is given by $u(t) = T(t)u_0$. The same argument holds, if A is a bounded operator on an arbitrary Banach Space. On the other hand one can derive the operator A from the family $T(t)$ by the formula

$$\lim_{t\downarrow 0} \frac{T(t) - I}{t} \quad = \quad A\,. \tag{10.22}$$

The theory of semigroups investigates and generalize these concepts.

In the following let $(X, \|\cdot\|)$ be a Banach space and denote by $\mathbb{R}^+$ ($\mathbb{R}_0^+$) the positive real axis excluded (included) zero.

Definition 10.3.1. *A set of bounded linear operators $\{T(t)\}_{t\in\mathbb{R}_0^+}$ on X is called a* strongly continuous semigroup of bounded linear operators *(or simply C_0* semigroup*) if*

(i) $T(0) = I$ *(the identity operator on X)*

(ii) $\forall\ s,t \geq 0\colon\ T(s+t) = T(s)\,T(t)$ *(semigroup property)*

(iii) $\forall\ x \in X\colon\ \lim_{t\downarrow 0} T(t)x = x$.

If (ii) *is valid for all $s,t \in \mathbb{R}$ and* (iii) *could be replaced by $\lim_{t\to 0} T(t)x = x$, then $T(t)$ is called a C_0 group.*

Equation (10.22) motivates the following Definition 10.3.2 of a linear Operator. As the Example 10.3.6 (see p. 217) shows, this operator is in general not bounded, even if we consider a C_0 group of unitary operators.

Definition 10.3.2. *Let $T(t)$ be a C_0 semigroup. The linear operator A defined by*

$$D(A) \quad = \quad \left\{ x \in X \,\middle|\, \lim_{t\downarrow 0} \tfrac{T(t)x-x}{t} \ \textit{exists} \right\} \quad , \quad Ax = \lim_{t\downarrow 0} \tfrac{T(t)x-x}{t}\,,$$

is called the infinitesimal generator *of the semigroup $T(t)$. $D(A)$ is called the domain of A.*

In the case of a bounded operator A the norm of the semigroup $T(t)$ is bounded by $e^{\|A\|t}$. A similar equation holds for a C_0 semigroup.

Proposition 10.3.3. *Let $T(t)$ be a C_0 semigroup. There exist real constants $w \geq 0$ and $M \geq 1$, such that*

$$\|T(t)\| \quad \leq \quad M\, e^{wt} \tag{10.23}$$

for all $t \in \mathbb{R}_0^+$.

Proof. There exists an $\eta > 0$ such that $\|T(t)\|$ is bounded for $0 \leq t \leq \eta$. If this claim is false, then there exists a sequence of positive $t_n \to 0$, such that $\|T(t_n)\| \geq n$. Thus, by the resonance theorem[6], there exists an $x \in X$ such that $\|T(t_n)x\|$ is unbounded in contradiction to property (iii) of Definition 10.3.1. Hence, $\|T(t)\| \leq M$ for $0 \leq t \leq \eta$. $M \geq 1$ is true because $T(0)$ is equal to I. Define $w := \eta^{-1} \ln M \geq 0$. For $t > 0$ one computes

$$\|T(t)\| = \|T(n\eta + \delta)\| = \|T(\eta)^n T(\delta)\| \leq MM^n = Me^{wn\eta} \leq Me^{wt}\,,$$

with $0 \leq \delta \leq \eta$. □

Corollary 10.3.4. *For every fixed $x \in X$ the assignment $t \to T(t)x$ is a continuous map from $\mathbb{R}_0^+$ into X.*

Proof. Let $t \geq s \geq 0$, $t - s = \delta$.

$$\|T(t)x - T(s)x\| \quad = \quad \|T(t)(T(\delta)x - x)\| \quad \leq \quad Me^{wt}\|T(\delta)x - x\| \stackrel{\delta \to 0}{\longrightarrow} 0\,.$$

Analogue for $s \geq t \geq 0$. □

The next Proposition 10.3.5 contains the main properties of C_0 semigroups.

Proposition 10.3.5. *Let $T(t)$ be a C_0 semigroup on X and let A be its infinitesimal generator. Then it holds:*

(i) *For $x \in X$ it holds*

$$\lim_{h \to 0} \tfrac{1}{h} \int_t^{t+h} T(s)x\, ds \quad = \quad T(t)x\,.$$

(ii) *For $x \in X$ it holds $\int_0^t T(s)x\, ds \in D(A)$ and*

$$A\bigg(\int_0^t T(s)x\, ds\bigg) \quad = \quad T(t)x - x\,.$$

(iii) *For $x \in D(A)$ it holds $T(t)x \in D(A)$ and*

$$\frac{d}{dt}T(t)x \;=\; AT(t)x \;=\; T(t)Ax\,.$$

[6] see [74] p. 69: Let X Banach space, Y normed space and $\{T_a \in L(X,Y) | a \in A\}$ be a family of bounded linear operators. The set $\{\|T_a\| | a \in A\}$ is bounded, if $\{\|T_a x\| | a \in A\}$ is bounded for all $x \in X$.

(iv) *For* $x \in D(A)$

$$T(t)x - T(s)x \;=\; \int_s^t T(\tau)Ax\,d\tau \;=\; \int_s^t AT(\tau)x\,d\tau\ .$$

Proof. (i) follows from the continuity of $t \to T(t)x$, see Corollary 10.3.4[7].
(ii) Let $h > 0$, then

$$\begin{aligned}\tfrac{T(h)-I}{h}\int_0^t T(s)x\,ds &= \tfrac{1}{h}\int_0^t (T(s+h)x - T(s)x)\,ds \\ &= \tfrac{1}{h}\int_h^{t+h} T(s)x\,ds - \tfrac{1}{h}\int_0^h T(s)x\,ds\ .\end{aligned}$$

If $h \downarrow 0$, then the right–hand side tends to $T(t)x - x$. Thus, by definition of A, the left-hand side tends to $A\int_0^t T(s)x\,ds$.
(iii) Let $h > 0$, then

$$\tfrac{T(h)-I}{h}T(t)x \;=\; \tfrac{T(t+h)-T(t)}{h}x \;=\; T(t)\tfrac{T(h)-I}{h}x \;\overset{h\to 0}{\longrightarrow}\; T(t)Ax\ .$$

Thus $\lim_{h\to 0}\frac{T(h)-I}{h}T(t)x$ exists and we get $AT(t)x = T(t)Ax$. Especially it holds $T(t)x \in D(A)$. The middle term also converges to $\frac{d^+}{dt}T(t)x$ which proves (iii) for the right derivative. To show the same equation for the left derivative one calculates ($t > 0$)

$$\begin{aligned}&\frac{T(t)x - T(t-h)x}{h} - T(t)Ax \\ &\quad= T(t-h)\left(\frac{T(h)x - x}{h} - Ax\right) + \big(T(t-h)Ax - T(t)Ax\big)\ .\end{aligned}$$

The first term of the right–hand side tends to 0 as $h \to 0$, since $T(t-h)$ is bounded (see Theorem 10.3.3), $x \in D(A)$, and A is the infinitesimal generator of the C_0 semigroup (see Definition 10.3.2). By the strong continuity of $T(t)$ (property (iii) of Definition 10.3.1) also the second term tends to zero.
(iv) Integrate (iii). □

Point (iii) states that $T(t)$ maps X to $D(A)$ and is (Frechet-) differentiable with derivative A. This leads to an existence-result for the solution of the abstract initial-value problem ($t \geq 0$)

$$\frac{d}{dt}u(t) = Au\ , \qquad u(0) = u_0 \in X\ ,$$

if A is the infinitesimal generator of a C_0 semigroup.

Example 10.3.6. *One interesting example is the group of translations in* $X = L^p(\mathbb{R})$, $1 \leq p < +\infty$ *see [9] p. 302. For* $u \in L^p(\mathbb{R})$ *define* $T(t)u$ *by*

$$T(t)u(x) \;=\; u(x+t)\ \textit{a.e. in}\ \mathbb{R}.$$

It follows Immediately that $\|T(t)u\|_{L^p(\mathbb{R})} = \|u\|_{L^p(\mathbb{R})}$, *i. e.* $T(t)$ *is a bounded operator for all* $t \in \mathbb{R}$. *Moreover the properties (i) and (ii) from Definition*

[7] $\|\frac{1}{h}\int_t^{t+h} T(s)x\,ds - T(t)x\| = \|\frac{1}{h}\int_t^{t+h}(T(s)x - T(t)x)\,ds\| \leq \sup_{s\in[0,h]}\|T(t+s)x - T(t)x\|$

10.3.1 obviously are fulfilled. In order to show that $T(t)$ is a C_0 semigroup it remains to prove the strong continuity. If ϕ is a continuous function with compact support, i. e. $\phi \in C_c(\mathbb{R})$ we have[8]*:*

$$\|T(t)\phi - \phi\|_{L^p(\mathbb{R})} \leq (\lambda(\operatorname{supp}\phi))^{\frac{1}{p}} \sup_{x\in\mathbb{R}} |\phi(x+t) - \phi(x)| \overset{t\to 0}{\longrightarrow} 0 \,.$$

Because $C_c(\mathbb{R})$ is dense in $L^p(\mathbb{R})$, see [66] p. 69, there exists for a given $f \in L^p(\mathbb{R})$ a $g \in C_c(\mathbb{R})$ such that $\|f-g\|_{L^p(\mathbb{R})} \leq \frac{\varepsilon}{3}$. It follows

$$\|T(t)f - f\| \quad \leq \quad \underbrace{\|T(t)(f-g)\|}_{=\|f-g\|\leq\frac{\varepsilon}{3}} + \underbrace{\|f-g\|}_{\leq\frac{\varepsilon}{3}} + \underbrace{\|T(t)g-g\|}_{\leq\frac{\varepsilon}{3}\,(t\ll 1)} \,.$$

The infinitesimal operator of the C_0 group $T(t)$ is given by (ϕ suitable)

$$A\phi(x) \quad = \quad \lim_{t\to 0} \frac{T(t)\phi(x) - \phi(x)}{t} \quad = \quad \lim_{t\to 0} \frac{\phi(x+t) - \phi(x)}{t} \quad = \quad \frac{d}{dx}\phi(x) \,.$$

For a more rigorous discussion see [9] p. 311.

The above example shows that unbounded operators can generate a group of bounded or even isometric operators. A full characterization of the infinitesimal-generator of a C_0 group of unitary operators on a Hilbert-space is given by

Theorem 10.3.7. (Stone)
A is the infinitesimal generator of a C_0 group of unitary operators on a Hilbert space H if and only if iA is self–adjoint.

Proof. See e. g. [62] p.41. □

Once again consider the translation group $T(t)$ on the complex Hilbert–space $L^2(\mathbb{R})$. We have seen that the infinitesimal–generator is given by $\frac{d}{dx}$. An easy calculation shows, that the operator $iA := i\frac{d}{dx}$ is self–adjoint. Hence by Stones Theorem $T(t)$ is a C_0 group of unitary operators.

Remark 10.3.8. *Let X be a Hilbert space with the inner product $(\cdot,\cdot)$ and let $U(t)$ be a C_0-group of unitary operators with infinitesimal generator iA. Furthermore let $v, w \in H$, with $v \in D(A)$. Since the sequence $\frac{1}{h}(U(h) - \mathrm{Id})v \to iAv$ in H, we get for $h, t \in \mathbb{R}$:*

$$\begin{aligned} &\frac{1}{h}\big[(U(t+h)v, w) - (U(t)v, w)\big] \\ &\quad = \quad \big(\tfrac{1}{h}(U(h) - \mathrm{Id})v, U(t)^*w\big) \overset{h\to 0}{\longrightarrow} (iAv, U(t)^*w) \quad = \quad (iAU(t)v, w) \,. \end{aligned}$$

Hence the map $t \mapsto (U(t)v, w)$ is continuously differentiable for all $t \in \mathbb{R}$ and it holds $\frac{d}{dt}(U(t)v, w) = (iAU(t)v, w)$.

10.3.2 Distributions

We refer to [1, 10] for a more detailed discussion. Let $\Omega \subset \mathbb{R}^n$ be an open set and let $\mathcal{D}(\Omega)$ be the set of test functions on Ω. I. e. $f\colon \Omega \to \mathbb{R}$ is an element of $\mathcal{D}(\Omega)$ if and only if $f \in \mathbb{C}^\infty(\Omega)$ and $\operatorname{supp}(f)$ is compact.

[8] Here λ denotes the Lebesgue measure.

Definition 10.3.9. *A sequence $(\phi_n)_{n\in\mathbb{N}}$ of elements from $\mathcal{D}(\Omega)$ converges to $\phi \in \mathcal{D}(\Omega)$ if and only if the following conditions are satisfied:*

(i) *there exists a compact set $K \subset \Omega$ such that* $\operatorname{supp}\phi_n \subset K$ *for all $n \in \mathbb{N}$,*

(ii) *for all multi–indices $\alpha \in \mathbb{N}_0^n$, the sequence $(D_x^\alpha \phi_n)_{n\in\mathbb{N}}$ converges to $D^\alpha\phi$ uniformly on K.*

With $\mathcal{D}'(\Omega)$ we denote the space of linear functionals on $\mathcal{D}(\Omega)$ which are pointwise continuous with respect to the convergence concept from Definition 10.3.9. The elements of $\mathcal{D}'(\Omega)$ are called *distributions*. The application of a distribution $T \in \mathcal{D}'(\Omega)$ on a test function $\phi \in \mathcal{D}(\Omega)$ is denoted by $T(\phi) =< T, \phi >$.

Definition 10.3.10. *A sequence of distributions $(T_n)_{n\in\mathbb{N}}$ converges to the distribution T, if for every $\phi \in \mathcal{D}(\Omega)$*

$$< T_n - T, \phi > \longrightarrow 0 \text{ in } \mathbb{C}, \text{ when } n \to \infty.$$

Let L be a continuous linear operator on $\mathcal{D}(\Omega)$ and L^* its adjoint with respect to the L^2-scalar product[9].

Definition 10.3.11. *Let $T \in \mathcal{D}'(\Omega)$ and L as above. Then $LT \in \mathcal{D}'(\Omega)$ is pointwise defined by*

$$(LT)(\phi) \;=< LT, \phi > := < T, L^*\phi > \;.$$

That the extension of L is well defined is left to the reader. Some examples:

(i) Let $\alpha \in \mathbb{N}$ and $D^\alpha = \left(\frac{\partial}{\partial x_j}\right)^\alpha$ as above. The adjoint operator of D^α is $(-1)^\alpha D^\alpha$, this means

$$D^\alpha T(\phi) \;=< D^\alpha T, \phi > = < T, (-1)^\alpha D^\alpha \phi > \;.$$

(ii) Let $\Omega = \mathbb{R}^n$, $\psi \in \mathcal{D}(\Omega)$ and let $L\phi(x) := \int_{\mathbb{R}^n} \psi(x-y)\phi(y)\,dy$ be the convolution with ψ. The adjoint operator of L is given by the formula $L^*\phi(x) = \int_{\mathbb{R}^n} \psi(y-x)\phi(y)\,dy$. Denote $R(\psi)(x) = \psi(-x)$, then

$$(\psi * T)(\phi) \;=< \psi * T, \phi > = < T, R(\psi) * \phi > \;.$$

Another important tool is the Fourier transformation. For a Schwartz function $\varphi \in \mathcal{S}(\mathbb{R}^n)$ (cf. [1, p.219f], [65][10]) we define

$$\widehat{\varphi}(\xi) \;=\; \tfrac{1}{(2\pi)^{\frac{n}{2}}} \int_{\mathbb{R}^n} \varphi(x)\, e^{-i\xi^T x}\, dx \;. \tag{10.24}$$

The Fourier transformation is an isomorphism of $\mathcal{S}(\mathbb{R}^n)$ to $\mathcal{S}(\mathbb{R}^n)$ and continuous with respect to the convergence concept from Definition 10.3.13.

Since any function in $f \in \mathcal{S}(\mathbb{R}^n)$ defines a distribution T_f by

$$< T_f, \phi > \;:=\; \int_{\mathbb{R}^n} f(x)\,\phi(x)\,dx \;, \tag{10.25}$$

it is natural to define the Fourier transformation of a distribution T by[11]

$$\left\langle \widehat{T}, \phi \right\rangle \;:=\; \left\langle T, \widehat{\phi} \right\rangle . \tag{10.26}$$

Another important fact is, that the set of test functions[12] is dense in $\mathcal{D}'(\mathbb{R}^n)$.

[9] $\forall\ \phi, \psi \in \mathcal{D}(\Omega)\colon\ (L\phi, \psi) = \int_\Omega (L\phi)\overline{\psi}dx = \int_\Omega \phi\overline{(L^*\psi)}dx = (\phi, L^*\psi)$.

[10] In both books the *Schwartz* functions are called "Rapidly decreasing functions".

[11] $< T_f, \widehat{\phi} >= \int_{\mathbb{R}^n} f\widehat{\phi}\,dx = \frac{1}{(2\pi)^{\frac{n}{2}}} \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} f(x)\phi(y)e^{-ix^Ty}\,dydx = \int_{\mathbb{R}^n} \widehat{f}\phi\,dy =< T_{\widehat{f}}, \phi >$

[12] one identifies $\phi \in \mathcal{D}(\mathbb{R}^n)$ with the distribution T_ϕ

The following Proposition 10.3.12 is from [53, p.147].

Proposition 10.3.12. *Let $T \in \mathcal{D}'(\mathbb{R}^n)$ and $\varphi \in \mathcal{D}(\mathbb{R}^n)$. Then there exists a function $t \in C^\infty(\mathbb{R}^n)$ such that for all $\phi \in \mathcal{D}(\mathbb{R}^n)$*

$$< \varphi * T, \phi > \quad = \quad \int_{\mathbb{R}^n} t(y)\phi(y)\, dy \ .$$

If further $\int_{\mathbb{R}^n} \varphi\, dy = 1$, then $\varphi_\epsilon * T \to T$ in $\mathcal{D}'(\mathbb{R}^n)$ as $\epsilon \to 0$ and the function φ_ϵ is defined by $\varphi_\epsilon(x) = \frac{1}{\epsilon^n}\varphi(\frac{x}{\epsilon})$.

The proof shows that that $t(y) = T(\varphi(y - \cdot))$.

Of course the Schwartz functions are not the only objects which define a distribution by (10.25). By the same construction one can define distributions for $f \in L^\infty(\Omega)$ for example. Also any Borel measure μ on $\mathbb{R}^n$ defines a distribution (on $\Omega = \mathbb{R}^n$) via

$$\mu(\phi) \quad := \quad \int_{\mathbb{R}^n} \phi\, d\mu \ .$$

The reverse is true for nonnegative distributions, cf. Corollary 10.3.16. This means that for a nonnegative distribution T there exists a measure μ such that

$$T(\phi) \quad = \quad \int_{\mathbb{R}^n} \phi\, d\mu \ .$$

holds for all $\phi \in \mathcal{D}(\Omega)$. Corollary 10.3.16 is related to the *Riesz' representation theorem* (cf. [16], VIII§2).

On the set of Schwartz functions $\mathcal{S}(\mathbb{R}^n)$ on can establish a convergence concept similar to those of the test function.

Definition 10.3.13. *A sequence $(\phi_n)_{n\in\mathbb{N}}$ of elements from $\mathcal{S}(\mathbb{R})$ converges to $\phi \in \mathcal{S}(\mathbb{R}^n)$ if and only if for all multi indices $\alpha, \beta \in \mathbb{N}_0^n$*

$$\lim_{j\to\infty} x^\alpha D_x^\beta \phi_j(x) \ = \ 0 \qquad \textit{uniformly on } \mathbb{R}.$$

The dual space of the Schwartz functions (with respect to the convergence concept from Definition 10.3.13) is denoted by $\mathcal{S}'(\mathbb{R}^n)$. Its elements are called *tempered distributions.*

10.3.3 A variant of Riesz' representation theorem

Let $\Omega \subset \mathbb{R}^d$ be a non empty open set. By $C_c(\Omega)$ we denote the space of continuous functions $f\colon \Omega \to \mathbb{R}$ which have compact support. In this section we prove a variant of *Riesz' representation theorem* for positive linear forms $I\colon V \to \mathbb{R}$, with a suitable function space $V \subset C_c(\Omega)$. The main result is Proposition 10.3.15.

Let us fix some notations and recall some definitions.

Notation. *We denote by $\mathcal{B}(\Omega)$ the Borel σ–algebra of Ω, i. e. the σ–algebra generated by the open sets of Ω* (cf. [16, p.310]).

By $\mathcal{K}(\Omega)$ we denote the set of all compact subsets of Ω. Further we denote the characteristic function of a set A by χ_A. I. e. $\chi_A|_A = 1$ and $\chi_A|_{A^c} = 0$.

A measure $\mu\colon \mathcal{B}(\Omega) \to [0,\infty]$ is called a *Borel measure*, if and only if μ is locally finite. I. e. for every $x \in \Omega$ there exists an open neighborhood $U \subset \Omega$, such that $\mu(U) < \infty$.

Remark 10.3.14. *Since Ω is locally compact, a measure μ is locally finite, if and only if $\mu(K) < \infty$ for all $K \in \mathcal{K}(\Omega)$* (see [16, Folgerung 1.2 c), p.331]). *Hence any Borel measure on Ω has this property.*

Let $\mathcal{A} \supset \mathcal{B}(\Omega)$ be a σ–algebra on Ω. A corresponding measure $\mu\colon \mathcal{A} \to [0,\infty]$ is called *inner regular*, if and only if for all $A \in \mathcal{A}$

$$\mu(A) \quad = \quad \sup \mu(K) | K \subset A,\, K \in \mathcal{K}(\Omega)\,.$$

If additionally for all $A \in \mathcal{A}$

$$\mu(A) \quad = \quad \inf\{\mu(U) | A \subset U,\, U \subset \Omega \text{ open}\}\,,$$

then μ is called a *regular* measure. An inner regular Borel measure is called a *Radon measure* (see [16, p.310]).

Proposition 10.3.15. *Let V be a subspace of $C_c(\Omega)$ such that $\mathcal{D}(\Omega) \subset V$. Furthermore, let $I\colon V \to \mathbb{R}$ be a nonnegative linear form, i. e. for all $f \in V$ with $f \geq 0$ it holds $I(f) \geq 0$. Then there exists a unique Radon measure μ on Ω, such that*

$$I(f) \quad = \quad \int_\Omega f \, d\mu \tag{10.27}$$

holds for all $f \in V$.

In reference [16] different versions of Riesz' representation theorem are given. Most of the theorems are formulated for Ω to be a locally compact or complete regular Hausdorff space. Hence from this point of view Proposition 10.3.15 is a special case of the results from the textbook, since we restrict ourself to open subsets of $\mathbb{R}^n$. Due to the general setting of Ω in [16], there exist no test functions and hence all theorems treat linear forms $I\colon V \to \mathbb{K}$, with[13] $V = C_c(\Omega), C_0(\Omega), C(\Omega)$. Thus our restriction to a subspace $V \subset C_c(\Omega)$ which contains the test functions is a refinement of the results from [16] for the special choice of Ω.

A direct consequence of Proposition 10.3.15 is

Corollary 10.3.16. *Let $\Omega \subset \mathbb{R}^n$ be open. A nonnegative distribution*[14] *$T \in \mathcal{D}'(\Omega)$ is a Radon measure. I. e. there exists a unique Radon measure μ on Ω, such that for all $\varphi \in \mathcal{D}(\Omega)$*

$$T(\varphi) \quad = \quad \int_\Omega \varphi \, d\mu\,.$$

[13] Here $C_0(\Omega)$ denotes the space of continuous functions $f\colon \Omega \to \mathbb{K}$ which vanish at the boundary of Ω, i. e. $\forall\, f \in C_0(\Omega)\ \forall\, \varepsilon > 0\ \exists\, K \subset \Omega$ compact: $|f|\big|_{\Omega\setminus K} < \varepsilon$.

[14] T positive $\Leftrightarrow \forall \varphi \in \mathcal{D}(\Omega),\, \varphi \geq 0 :\ < T, \varphi > \geq 0$

Example. *Let $\lambda \in \mathbb{R}$ and let $T = \delta_\lambda \in \mathcal{D}(\mathbb{R})'$ be the Delta–distribution with pole in λ. Obviously T is a nonnegative distribution. Hence, by Corollary 10.3.16 exists a unique measure μ, such that $T(\varphi) = \int_{\mathbb{R}} \varphi \, d\mu$ for all $\varphi \in \mathcal{D}(\mathbb{R})$. As discussed in Example 10.3.22 it holds for all $A \in \mathcal{B}(\mathbb{R})$:*

$$\mu(A) \quad = \quad \begin{cases} 1\,, \lambda \in A\,, \\ 0\,, \quad else\,. \end{cases}$$

As we have seen, nonnegative distributions can be identified with Radon measures. The same result holds for nonnegative tempered distributions.

Corollary 10.3.17. *Let $W = \mathcal{S}(\mathbb{R}^n)$ be the space of rapidly decreasing functions* (Schwartz functions) *and let $T\colon W \to \mathbb{C}$ be a nonnegative tempered distribution. I. e. for all $f \in W$ with $f \geq 0$ it holds $T(f) \geq 0$. Then there exists one and only one Radon measure μ on $\mathbb{R}$, such that*

$$T(f) \quad = \quad \int_\Omega f \, d\mu \tag{10.28}$$

holds for all $f \in W$. Furthermore the measure μ coincides with the Radon measure from Corollary 10.3.16, if we set $\Omega = \mathbb{R}^n$.

Proof. We restrict T to the subspace $V := \mathcal{D}(\mathbb{R}^n)$. By Proposition 10.3.15 exists a unique Radon measure μ on $\mathbb{R}^n$, such that $T(\varphi) = \int_{\mathbb{R}^n} \varphi \, d\mu$ holds for all test functions $\varphi \in V$.

Let $f \in \mathcal{S}(\mathbb{R}^n)$. By [65, Theorem 7.10, p.189] the test functions are dense in $\mathcal{S}(\mathbb{R}^n)$. Hence there exists a sequence $\{f_n\}$ in V such that $f_n \to f$ in $\mathcal{S}(\mathbb{R}^n)$. Further the proof of [65, Theorem 7.10] shows, that we can use the sequence

$$f_n(x) \quad = \quad f(x)\varphi\left(\frac{x}{n}\right) \in \mathcal{S}(\mathbb{R}^n)\,,$$

with $\varphi \in \mathcal{D}(\mathbb{R}^n)$, such that $0 \leq \varphi \leq 1$ and[15] $\varphi|_{B_1(0)} = 1$. If we decompose f_n in its positive and negative part, i. e. $f_n = f_n^+ - f_n^-$, we find that $f_n^\pm$ are monotonously increasing. By definition the tempered distribution T is continuous, which yields

$$T(f) \quad = \quad \lim_{n\to\infty} T(f_n) \quad = \quad \lim_{n\to\infty} \int_{\mathbb{R}^n} f_n \, d\mu\,.$$

Due to the Monotone convergence theorem of Levi (see [16] p.124) we can interchange the limit and integration. I. e..

$$\begin{aligned} \lim_{n\to\infty} \int_{\mathbb{R}^n} f_n \, d\mu \quad &= \quad \lim_{n\to\infty} \int_{\mathbb{R}^n} f_n^+ \, d\mu - \lim_{n\to\infty} \int_{\mathbb{R}^n} f_n^- \, d\mu \\ &= \quad \int_{\mathbb{R}^n} \lim_{n\to\infty} f_n^+ \, d\mu - \int_{\mathbb{R}^n} \lim_{n\to\infty} f_n^- \, d\mu\,. \end{aligned}$$

Since $\lim_{n\to\infty} f_n^\pm \to f^\pm$ we are done. □

[15] Here $B_1(0)$ denotes the unit ball in $\mathbb{R}^n$.

Proof of Proposition 10.3.15

In order to prove Proposition 10.3.15 we mimic the proof of *Darstellungssatz von Riesz* [16, VIII§2, p.331ff]. The crucial property of the space $\mathcal{D}(\Omega)$ is Lemma 10.3.19, i. e. that the characteristic functions of compact sets can be "well" approximated by test functions. In order to prove this we need Lemma 10.3.18 and hence shall start with it.

Let $g\colon \mathbb{R} \to \mathbb{R}$ be piecewise defined by

$$g(x) \quad := \quad \begin{cases} e^{-\frac{1}{x^2}}\,, & 0 < x \\ 0 \quad , & x \leq 0 \end{cases} .$$

From [23, p.229f] we know that g is arbitrarily often differentiable. Hence

$$\tilde{g}(x) \quad := \quad g(x)\, g(1-x) \quad = \quad \begin{cases} e^{-\frac{1}{x^2}} e^{-\frac{1}{(1-x)^2}}\,, & 0 < x < 1 \\ 0 \qquad\quad , & \text{else} \end{cases}$$

is in $C^\infty(\mathbb{R})$, strictly positive in the interval $(0,1)$ and zero at $\mathbb{R}\backslash(0,1)$. Let

$$c(x) \quad := \quad \int_0^x e^{-\frac{1}{s^2}} e^{-\frac{1}{(1-s)^2}}\, ds\,. \tag{10.29}$$

Obviously it holds $c \in C^\infty([0,1])$, $c(0) = 0$ and c is strictly monotone increasing on the interval $[0,1]$.

Lemma 10.3.18. *The piecewise defined function*[16]

$$\mathrm{cut}_1(x) \quad := \quad \begin{cases} 0 \quad , & x \leq 0 \\ \frac{c(x)}{c(1)}\,, & 0 < x < 1 \\ 1 \quad , & 1 \leq x \end{cases}$$

is in $C^\infty(\mathbb{R})$, nonnegative and bounded from above by 1.

Proof. Since it is $c > 0$ for $x > 0$, the function cut_1 is well defined. Obviously it holds, that cut_1' coincides with $\frac{\tilde{g}}{c(1)}$ on $\mathbb{R}\backslash\{0,1\}$. Thus we have to show, that cut_1 is continuously differentiable at $x = 0, 1$.

(i) Let $x < 1$. Then it holds

$$\frac{\mathrm{cut}_1(x) - \mathrm{cut}_1(0)}{x - 0} \quad = \quad \begin{cases} \frac{c(x)}{xc(1)}\,, & 0 < x \\ 0 \quad , & x \leq 0 \end{cases} .$$

By L'Hôpital's rule (cf. [23]) we deduce from the definition of c (10.29) that $\lim_{x\to 0} \frac{c(x)}{x} = 0$. This yields the differentiability at $x = 0$.

(ii) Let $x > 0$. Then it holds

$$\frac{\mathrm{cut}_1(x) - \mathrm{cut}_1(1)}{x - 1} \quad = \quad \begin{cases} 0 \quad , & 1 \leq x \\ \frac{\frac{c(x)}{c(1)} - 1}{(x-1)}\,, & 0 < x < 1 \end{cases}$$

Again we use L'Hôpital's rule and find

$$\lim_{x\to 1} \frac{\frac{c(x)}{c(1)} - 1}{(x-1)} \quad = \quad 0\,.$$

Hence cut_1 is differentiable at $x = 1$.

[16] Thanks to Dominik Stürzer and Jan Sprenger.

□

The next Lemma 10.3.19 is the essential property of the space of test functions $\mathcal{D}(\Omega)$. It states that characteristic functions of compact subsets of Ω can be quite good approximated by test functions.

Lemma 10.3.19. *Let $\Omega \subset \mathbb{R}^n$ open and let $K \subset \Omega$ be compact and let $U \subset \Omega$ be an open neighborhood*[17] *of K. Then there exists a function $\varphi \in C_c^\infty(\Omega) = \mathcal{D}(\Omega)$, such that $\varphi|_K = 1$ and $\operatorname{supp}(\varphi) \subset U$, i. e. $\varphi|_{\Omega\setminus U} = 0$.*

Proof. We mimic the proof of [16, 2.1 Lemma, p.327] .

Since U is open, there exists for every $x \in K$ a radius $r_x > 0$, such that the open Ball

$$B_{r_x}(x) \quad := \quad \{y \in \Omega |\ \|x - y\| < r_x\}$$

with center at x and radius r_x is a subset of U (cf. [31]). Hence the function (cf. Lemma 10.3.18 for definition of cut_1)

$$\varphi_x(y) \quad := \quad \operatorname{cut}_1\left(\frac{3}{r_x}\left(\frac{2r_x}{3} - \|y - x\|\right)\right)$$

has compact support $\operatorname{supp}(\varphi_x) = \overline{B_{\frac{2}{3}r_x}(x)} \subset U$ and it holds $\varphi_x|_{B_{\frac{1}{3}r_x}(x)} = 1$. Since the map $y \mapsto \|y - x\|$ is $C^\infty(\mathbb{R}^n\setminus\{0\})$, we further deduce that $\varphi_x \in C_c^\infty(\mathbb{R}^n)$. Let $V_x := B_{\frac{1}{3}r_x}(x)$. Since K is compact, there are finitely many $x_1, \ldots, x_m$, such that $K \subset \bigcup_{j=1}^m V_{x_j}$. Hence the function $\widehat{\varphi} = \sum_{j=1}^m \varphi_{x_j}$ is of class $C_c^\infty(\mathbb{R}^n)$ and it holds $\widehat{\varphi}\big|_K \geq 1$ and $\operatorname{supp}(\widehat{\varphi}) \subset U$. The support of $\widehat{\varphi}$ is compact and hence it holds $\widehat{\varphi} \in \mathcal{D}(\Omega)$. Thus we can set $\varphi = \operatorname{cut}_1(\widehat{\varphi})$. □

Lemma 10.3.20. *Let the assumptions of Proposition 10.3.15 hold. We define for all compact subsets $K \subset \Omega$*

$$\mu_0(K) \quad := \quad \inf\{I(f)\big|\ f \in V,\ f \geq \chi_K\}\,. \tag{10.30}$$

It holds:

(i) $0 \leq \mu_0(K) \leq \mu_0(L) < \infty$ *for all $K, L \in \mathcal{K}(\Omega)$ with $K \subset L$.*

(ii) $\mu_0(K \cup L) \leq \mu_0(K) + \mu_0(L)$ *for all $K, L \in \mathcal{K}(\Omega)$.*

(iii) $\mu_0(K \cup L) = \mu_0(K) + \mu_0(L)$ *for all $K, L \in \mathcal{K}(\Omega)$ with if $K \cap L = \emptyset$.*

(iv) *For every $K \in \mathcal{K}(\Omega)$ and $\varepsilon > 0$ exists an open neighborhood U of K, such that for all compact $L \subset U$ it holds*

$$\mu_0(L) \ \leq \ \mu_0(K) + \varepsilon\,.$$

Proof. The following proof is a (free) translation of the proof of [16, 2.2 Lemma, p.327f]. We only replace C_c by V or $\mathcal{D}(\Omega)$ and add some calculations and comments.

[17] We consider the topology induced by the euclidean norm on $\mathbb{R}^n$.

(i): By Lemma 10.3.19 there exists a function $f \in \mathcal{D}(\Omega)$, such that $f \geq \chi_L$. Obviously it holds $\chi_K \leq \chi_L$, which yields

$$\emptyset \neq S_L := \{\varphi \in V |\, f \geq \chi_L\} \subset \{\varphi \in V |\, f \geq \chi_K\} =: S_K$$

Since I is nonnegative we get

$$0 \leq \mu_0(K) = \inf\{I(f) \big|\, f \in S_K\} \leq \inf\{I(f) \big|\, f \in S_L\} = \mu_0(K)\,.$$

(ii): Now let $f, g \in V$ with $f \geq \chi_K$, $g \geq \chi_L$. Hence it is $f + g \in V$ and it holds $f + g \geq \chi_{K \cup L}$. This yields

$$\mu(K \cup L) \leq I(f + g) = I(f) + I(g)\,.$$

Passing to the infimum at the right–hand side yields (ii).

(iii): Due to (ii) we only have to show "$\geq$". Let $h \in \mathcal{D}(\Omega)$, such that $h \geq \chi_{K \cup L}$. The set $U := \Omega \backslash L$ is an open neighborhood of K. By Lemma 10.3.19 there exists a function $\varphi \in \mathcal{D}(\Omega)$, such that $0 \leq \varphi \leq 1$, $\varphi|_K = 1$, $\varphi|_{\Omega \backslash U = L} = 0$. Thus the functions $f := h\varphi$ and $g := h(1 - \varphi)$ are of class $\mathcal{D}(\Omega) \subset V$ and it holds

$$f \geq \chi_K\,, \quad g \geq \chi_L \quad \text{and} \quad f + g = h\,.$$

This yields

$$I(h) = I(f) + I(g) \geq \mu_0(K) + \mu_0(L)\,.$$

Hence it holds $\mu_0(K \cup L) \geq \mu_0(K) + \mu_0(L)$.

(iv): Let $K \in \mathcal{K}(\Omega)$. By Lemma 10.3.19 there exists a function $f \in \mathcal{D}(\Omega)$ with $f \geq \chi_K$. Hence $\mu_0(K)$ is well defined and there exists a sequence $\{f_n\}$ in $\mathcal{D}(\Omega) \subset V$, such that $\mu_0(K) = \lim_{n \to \infty} I(f_n)$. Thus for every $\delta > 0$ there exists a function $f \in V$, such that $I(f) \leq \mu_0(K) + \delta$. The set $U := \{x \in \Omega |\, f(x) > \frac{1}{\delta + 1}\}$ is an open neighborhood of K. For every compact $L \subset U$ it holds $(1+\delta)f \geq \chi_L$, which yields

$$\mu_0(L) \leq (1 + \delta) I(f) \leq (1 + \delta)(\mu_0(K) + \delta)\,.$$

If we choose δ, such that $\delta(\mu_0(K) + \delta + 1) < \varepsilon$, we get (iv). □

Lemma 10.3.21. *Let the assumptions of Lemma 10.3.20 hold and let μ_0 be defined by* (10.30)*. There exists one and only one continuation of μ_0 to an inner regular measure $\mu\colon \mathcal{B}(\Omega) \to [0, \infty]$. Further it holds for all $A \in \mathcal{B}(\Omega)$:*

$$\mu(A) = \sup\{\mu_0(K) |\, K \subset A,\, K \in \mathcal{K}(\Omega)\}\,. \tag{10.31}$$

Proof. Due to Lemma 10.3.20 we can apply [16, 2.3 Lemma, p.328]. This yields the requirements for [16, 2.4 Fortsetzungssatz, p.329], which finishes the proof. □

Example 10.3.22. *Let $V = \mathcal{D}(\mathbb{R})$ and let $I = \delta_\lambda \in \mathcal{D}(\mathbb{R})'$ be the Delta-distribution with pole in $\lambda \in \mathbb{R}$. For $K \in \mathcal{K}(\mathbb{R})$ we get*

$$\mu_0(K) = \inf\{f(\lambda) \big|\, f \in V,\, f \geq \chi_K\} = \begin{cases} 1\,, \lambda \in K\,, \\ 0\,, \quad \textit{else}\,. \end{cases}$$

Now let $A \in \mathcal{B}(\mathbb{R})$. We remark that the set $\{\lambda\}$ is compact. Thus if $\lambda \in A$ there exists a $K \in \mathcal{K}(\mathbb{R})$ with $\lambda \in K \subset A$. Hence by Lemma 10.3.21 it holds

$$\mu(A) \;=\; \sup\{\mu_0(K)|\; K \subset A,\, K \in \mathcal{K}(\Omega)\} \;=\; \begin{cases} 1\,, \lambda \in A\,, \\ 0\,, \quad else\,. \end{cases}$$

Now we can prove Proposition 10.3.15.

Proof of Proposition 10.3.15

We mimic the proof of [16, 2.4 Fortsetzungssatz, p.332f].

Uniqueness of μ is one to one as in the proof of [16, 2.4 Fortsetzungssatz, p. 332f]. One only has to replace "X" by "Ω", "$C_c(X)$" by "V" and "Lemma 2.1" by "Lemma 10.3.19".

Existence: The measure μ is defined by (10.31) of Lemma 10.3.21. Hence μ is an inner regular measure. Since $\mathbb{R}^n$ is locally compact (cf. [66, p.36]), the same holds for Ω. As in [16, p.332] we find that μ is locally finite and hence a Radon measure. It remains to prove (10.27).

(i) We prove: $\forall\; f \in V$ with $f \geq 0$ it holds $I(f) \geq \int_\Omega f\; d\mu$.

To prove the claim we also have to verify the existence of the integral. Since f is continuous, it is Borel measurable (cf. [16, 1.4 Korollar, p.86]). Hence there exists a sequence of step functions $\{u_n\}$, such that $u_n \nearrow f$ (cf. [16, 4.13 Satz, p.108]). Since the integral $\int_\Omega f\; d\mu$ is defined by the limit of the integrals of these step functions (and of any other sequence with the same convergence property), we shall derive the claimed estimate for these u_n. To be more precise we shall prove that $I(f) \geq \int_\Omega v\; d\mu$ holds for all nonnegative step functions with $f \geq v$. The function v is given by

$$v \;=\; \sum_{j=1}^{n} a_j\, \chi_{A_j}\,.$$

The numbers a_j are positive and the sets $A_1, \ldots, A_n \in \mathcal{B}(\Omega)$ are disjoint. Due to $v \leq f$ it holds $A_1, \ldots, A_n \subset \mathrm{supp}(f)$. Since $\mathrm{supp}(f)$ is compact, Lemma 10.3.20 yields

$$\forall\, K \in \mathcal{K}(\Omega),\, K \subset \mathrm{supp}(f)\colon\; 0 \leq \mu(K) \leq \mu(\mathrm{supp}\, f) < \infty\,.$$

Thus by definition of μ (cf. (10.31)) $\mu(A_j)$ is finite for all $j = 1, \ldots, n$. Hence for prescribed $0 < \varepsilon < \min(a_1, \ldots, a_n)$ there exist compact $K_j \subset A_j$, such that $\mu(A_j) - \varepsilon \leq \mu(K_j)$ $(j = 1, \ldots, m)$. The disjoint compact sets $K_1, \ldots, K_n$ have disjoint open neighborhoods $U_1, \ldots, U_n$. Since f is continuous the set $O_j := \{x \in \Omega|\; f(x) > a_j - \varepsilon\}$ is open. Further it holds $K_j \subset A_j \subset O_j$. Hence we can choose U_j such that $U_j \subset O_j$ holds. By Lemma 10.3.19 there exists a $\varphi_j \in \mathcal{D}(\Omega)$ such that $\chi_{K_j} \leq \varphi_j \leq \chi_{U_j}$. This yields

$$g \;:=\; \sum_{j=1}^{n} (a_j - \varepsilon)\varphi_j \;\in \mathcal{D}(\Omega) \subset V\,, \qquad g \;\leq\; f$$

and hence

$$\begin{aligned} I(f) \;\geq\; I(g) \;&=\; \sum_{j=1}^{n}(a_j-\varepsilon)I(\varphi_j) \;\geq\; \sum_{j=1}^{n}(a_j-\varepsilon)\mu(K_j) \\ &\geq\; \sum_{j=1}^{n}(a_j-\varepsilon)(\mu(A_j)-\varepsilon) \\ &=\; \int_\Omega v\,d\mu \,-\, \varepsilon\sum_{j=1}^{n}(a_j+\mu(A_j)-\varepsilon)\,. \end{aligned}$$

Since $\varepsilon > 0$ can be arbitrarily small it follows $I(f) \geq \int_\Omega v\,d\mu$.

(ii) We prove: $\forall\ f \in V$ with $f \geq 0$ it holds $I(f) = \int_\Omega f\,d\mu$.

Without restriction of generality we can assume $0 \leq f \leq 1$. From [16, p.310f][18] we get that for every $\varepsilon > 0$ there exists a relative compact neighborhood U of $K := \mathrm{supp}(f)$, such that $\mu(U) \leq \mu(K) + \varepsilon$. By Lemma 10.3.19 there exists a function $\varphi \in \mathcal{D}(\Omega) \subset V$ such that $\chi_K \leq \varphi \leq \chi_U$. Since $\varphi - f \in V$ is nonnegative, it follows from (i):

$$\begin{aligned} I(\varphi) - I(f) \;&=\; I(\varphi - f) \;\geq\; \int_\Omega (\varphi - f)\,d\mu \\ &=\; \int_\Omega \varphi\,d\mu \,-\, \int_\Omega f\,d\mu\,. \end{aligned} \tag{10.32}$$

Let $g \in V$, such that $g \geq \chi_{\mathrm{supp}(\varphi)}$. Hence $g - \varphi \geq 0$, which yields $I(g) \geq I(\varphi)$. Since $\mathrm{supp}(\varphi)$ is compact, we get from (10.30):

$$\mu(\mathrm{supp}(\varphi)) \;\geq\; I(g) \;\geq\; I(\varphi)\,.$$

This yields with (10.32) and (i)

$$\begin{aligned} 0 \;\leq\; I(f) \,-\, \int_\Omega f\,d\mu \;&\leq\; I(\varphi) \,-\, \int_\Omega \varphi\,d\mu \\ &\leq\; \mu\big(\mathrm{supp}(\varphi)\big) \,-\, \mu(K) \;\leq\; \mu(U) \,-\, \mu(K) \;\leq\; \varepsilon\,. \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, we have proven (ii).

(iii) Let $f \in V$. Since f is continuous and has compact support, it holds

$$c \;:=\; \inf_{x\in\Omega} f(x) > -\infty\,.$$

If $c \geq 0$, then nothing is to do, since $f \geq 0$ holds.

Let $c < 0$. As in (ii) there exists a relative compact neighborhood U of $K := \mathrm{supp}(f)$ and by Lemma 10.3.19 we get a function $\varphi \in \mathcal{D}(\Omega)$, such that $\chi_K \leq \varphi \leq \chi_U$. Hence the function $g := f - c\varphi \in V$ is nonnegative. Thus we get from (ii)

$$I(f) \,-\, c\,I(\varphi) \;=\; I(g) \;=\; \int_\Omega g\,d\mu \;=\; \int_\Omega f\,d\mu \,-\, c\int_\Omega \varphi\,d\mu\,.$$

Since $\varphi \in V$ is nonnegative, it holds $I(\varphi) = \int_\Omega \varphi\,d\mu$, which yields $I(f) = \int_\Omega f\,d\mu$.

[18] Here the crucial part is 1.2 Folgerung g).

10.3.4 $\mathcal{D}(\Omega)$ is dense in $L^p(\Omega, \mathcal{B}(\Omega), \mu)$

In this section we proof the following result:

Proposition 10.3.23. *Let $\Omega \subset \mathbb{R}^n$ be open and let μ be a Borel measure on $\mathcal{B}(\Omega)$. Then $\mathcal{D}(\Omega)$ is dense in $L^p(\Omega, \mathcal{B}(\Omega), \mu)$ for $1 \leq p < \infty$.*

In the sequel we use the notation $L^p(\Omega)$ for $L^p(\Omega, \mathcal{B}(\Omega), \mu)$.

Remark 10.3.24. *Let $\Omega = \mathbb{R}$ and let μ be the measure associated with the Delta-distribution. Hence any function on $\mathbb{R}$ which is continuous on a neighborhood of $x = 0$ is μ integrable and its $L^p(\Omega)$ norm is finite. One might get the impression that $L^p(\Omega)$ is much larger than the standard L^p space where μ is the Lebesgue measure. But $L^p(\Omega)$ consists of equivalence classes and hence $L^p(\Omega) \cong \mathbb{R}$.*

To prove Proposition 10.3.23 we need the following results.

Lemma 10.3.25. *Let $(X, \mathcal{A}, \mu)$ be a measure space. Then*

$$\mathcal{T}_e \quad := \quad \operatorname{span}\{\chi_A |\, A \in \mathcal{A},\, \mu(A) < \infty\}$$

is dense in $L^p(X, \mathcal{A}, \mu)$ for $1 \leq p < \infty$. $\mathcal{T}_e$ is the linear space of simple functions which take only a finite number of different values and whose support has finite measure, i. e. $\mu\big(\{x \in X | f(x) \neq 0\}\big) < \infty$.

A proof can be found in [16, Satz 2.28, p.240f] or in [66, Theorem 3.13, p.69].

Lemma 10.3.26. *Let X be a locally compact Hausdorff space with countable basis. Then any Borel measure on $\mathcal{B}(X)$ is regular and in particular a Radon measure.*

For a proof we refer to [16, 1.12 Korollar, p.316]. The next result, Lemma 10.3.27, is a special case of Lusin's Theorem as stated in [66, Theorem 2.24, p.55]. The theorem shows the close relation between Borel measurable and continuous functions.

Lemma 10.3.27. *Let X be a locally compact Hausdorff space with countable basis and let μ be a Borel measure on $\mathcal{B}(X)$. Further let $f \in \mathcal{T}_e$. Then there exists for every $\delta > 0$ a function $\varphi \in C_c(X)$, such that*

$$\mu\big(\{x \in \Omega | f(x) \neq \varphi(x)\}\big) \leq \delta \quad \textit{and} \quad \|\varphi\|_\infty \leq \|f\|_\infty \,.$$

Proof. By Lemma 10.3.26 μ, is a regular Borel measure on $\mathcal{B}(X)$ and hence the assumptions for Lusin's Theorem (see [66, Theorem 2.24, p.55]) are fulfilled. □

Lemma 10.3.28. *Let the assumptions of Proposition 10.3.23 hold. The space $C_c(\Omega)$ is dense in $L^p(\Omega)$ for $1 \leq p < \infty$.*

Proof. Let $f \in L^p(\Omega)$ and let $\varepsilon > 0$. By Lemma 10.3.25 exists a $h \in \mathcal{T}_e$, such that $\|f - h\|_{L^p} \leq \frac{\varepsilon}{2}$. Since $\mathbb{R}^n$ (with the Euclidean topology) is a locally compact Hausdorff space with countable basis, the same holds for the topological subspace Ω. Hence, by Lemma 10.3.27 there exists for every $\delta > 0$ a $g \in C_c(\Omega)$, such that

$$\mu(A) \leq \delta \text{ with } A := \{x \in \Omega | g(x) \neq h(x)\} \text{ and } \|g\|_\infty \leq \|h\|_\infty \,.$$

This yields

$$\begin{aligned}\|h-g\|_{L^p(\Omega)}^p &= \int_\Omega |h(x)-g(x)|^p\,dx \\ &= \int_A |h(x)-g(x)|^p\,dx \;\le\; (2\,\|h\|_\infty)^p\,\delta\,.\end{aligned}$$

For $\delta = \big(\frac{\varepsilon}{4\|h\|_\infty}\big)^p$ we get

$$\|f-g\|_{L^p(\Omega)} \;\le\; \|f-h\|_{L^p(\Omega)} + \|h-g\|_{L^p(\Omega)} \;\le\; \varepsilon\,.$$

Since ε is arbitrary, $C_c(\Omega)$ is dense in $L^p(\Omega)$. □

Proof of Proposition 10.3.23

The following discussion is an adaption of the proof of [1, Lemma 2.18, p.29ff].

Let $J \in \mathcal{D}(\mathbb{R}^n)$ be a nonnegative test function, such that $J(x) = 0$ holds for all $x \in \mathbb{R}^n$ with $|x| \ge 1$ and let $\int_{\mathbb{R}^n} J(x)\,dx = 1$. For $r > 0$ we define the nonnegative function $J_r(x) := r^{-n} J(\frac{x}{r})$. Obviously $J_r \in \mathcal{D}(\mathbb{R}^n)$ and satisfies

(i) $J_r(x) = 0$ for all $x \in \mathbb{R}^n$ with $|x| \ge r$.

(ii) $\int_{\mathbb{R}^n} J_r(x)\,dx \;=\; 1$.

Let $\varepsilon > 0$ and let $f \in L^p(\Omega)$. By Lemma 10.3.28 exists an $h \in C_c(\Omega)$, such that $\|f-h\|_{L^p(\Omega)} \le \frac{\varepsilon}{2}$. Since h has compact support in Ω, there exists an $R > 0$, such that for all $0 < r \le R$ the function $g_r := J_r * h \in C_c(\Omega)$. Furthermore $g_r \in C^\infty(\mathbb{R}^n)$ [1]. Hence $g_r \in \mathcal{D}(\Omega)$. We compute (using $\int_{\mathbb{R}^n} J_r(x)\,dx = 1$)

$$\begin{aligned}|g_r(x)-h(x)| &= \Big|\int_{\mathbb{R}^n} J_r(x-y)\big(h(y)-h(x)\big)\,dy\Big| \\ &= \Big|\int_{|x-y|<r} J_r(x-y)\big(h(y)-h(x)\big)\,dy\Big| \\ &\le \sup_{|y-x|\le r} |h(y)-h(x)|\,.\end{aligned}$$

Since h is continuous and has compact support, it is uniformly continuous on the whole domain Ω. Hence for every $\delta > 0$ exists an $0 < r_\delta < R$, such that for all $x \in \Omega$

$$\sup_{|y-x|\le r_\delta} |h(y)-h(x)| \;\le\; \delta$$

For $0 < r \le R$ let $K_r := \operatorname{supp}(g_r) \cup \operatorname{supp}(h)$. Both functions have compact support in Ω. Hence $K_R \subset \Omega$ is compact and thus has finite measure (see Lemma 10.3.20). Due to definition of g_r, $\operatorname{supp}(g_r) \subset \operatorname{supp}(g_R)$ for all $0 < r \le R$. Hence $\mu(K_r) \le \mu(K_R) < \infty$ for all $0 < r \le R$. This yields

$$\|g_{r_\delta}-h\|_{L^p(\Omega)}^p \;\le\; \int_{K_{r_\delta}} |g_{r_\delta}(x)-h(x)|^p\,d\mu \;\le\; \mu(K_R)\,\delta^p\,.$$

If $\mu(K_R) = 0$ we can choose an arbitrary δ. Otherwise set $\delta = \frac{\varepsilon}{2}\frac{1}{\mu(K_R)^{\frac{1}{p}}}$ and get

$$\|f-g_{r_\delta}\|_{L^p(\Omega)} \;\le\; \|f-h\|_{L^p(\Omega)} + \|h-g_{r_\delta}\|_{L^p(\Omega)} \;\le\; \varepsilon\,.$$

Since $\varepsilon > 0$ and f are arbitrary $\mathcal{D}(\Omega)$, is dense in $L^p(\Omega)$.

10.3.5 A spectral theorem

This part is very close to [71, p.75ff]. We only added some comments and computations and slightly changed the notation. A further difference is the order of argumentation. We start with Stone's theorem (cf. Proposition 10.3.7) and use it to directly derive the functional calculus for (unbounded) self–adjoint operators. In contrast, in [71] the author first derives the functional calculus for bounded self–adjoint operators, extend it to bounded normal operators and finally derives with Neumann's unitary trick the calculus for unbounded self–adjoint operators. From this he derives Stone's theorem. Nevertheless, the following discussion is analogue to the first step from [71], which is the derivation of the functional calculus of bounded self–adjoint operators.

Let A be a self–adjoint operator on a separable Hilbert space H. By Stone's theorem (cf. Proposition 10.3.7) iA is the infinitesimal generator of a unitary C_0 group $U(t)$ on H which satisfies ($t \in \mathbb{R}$)

$$\frac{d}{dt}U(t) \;=\; iAU(t)\,, \qquad U(0) = I\,.$$

For a given $v \in H$ the closed linear subspace $H_v := \overline{\text{span}\{U(t)v | t \in \mathbb{R}\}}$ is called the cyclic space generated by v. The vector v is called cyclic vector of a subspace V, if and only if $H_v = V$.

Example. *Let v be a normed eigenvector of A corresponding to the eigenvalue λ. Hence v lies in the domain of A and by Proposition 10.3.5 the function $\psi(t) := U(t)v$ solves the IVP*

$$\begin{aligned} \psi_t &= iA\psi = iU(t)Av = i\lambda U(t)v = i\lambda\psi\,, \\ \psi(0) &= v\,. \end{aligned}$$

Thus $\psi(t) = e^{i\lambda t}v$ and H_v is the one dimensional subspace spanned by v. From time to time we shall consider this setting during the course of argumentation.

Further we can write $H = H_v \oplus H_v^\perp$ (cf. [31]). Here $H_v^\perp$ is the orthogonal complement of H_v with respect to the Hilbert space structure of H. Let $w \in H_v^\perp$ and let $(\cdot,\cdot)$ be the scalar product on H. Then it holds for all $u \in H_v$ and all $t \in \mathbb{R}$

$$(U(t)w, u) \;=\; (w, U(t)^* u) \;=\; (w, U(-t)u)\,.$$

Since u is an element of H_v, the same holds for the vector $U(-t)u$. This yields $(U(t)w, u) = 0$. Hence $U(t)w \in H_v^\perp$, which means that the orthogonal complement of H_v is invariant under $U(t)$.

Lemma 10.3.29. *H is the closure of an orthogonal direct sum of countable many cyclic subspaces.*

Proof. Let $\{w_j\}_{j\in\mathbb{N}}$ be a countable, dense subset of H. Set $v_1 = w_1$ and $H_1 = H_{v_1}$. If $H_1 \neq H$, let P_1 be an orthogonal projection of H onto $H_1^\perp$ and $j \in \mathbb{N}$ the lowest index with $P_1 w_j \neq 0$. Set $v_2 = P_1 w_j$ and define $H_2 = H_{v_2}$. Continue. □

Due to Lemma 10.3.29 we shall restrict our discussion to a cyclic subspace H_v of H with cyclic vector v. Since H_v is closed, it is (as H) a separable Hilbert

space. Hence we directly could have started with a Hilbert space, having a cyclic vector and thus (to simplify notation) we write H instead of H_v. Furthermore, we define for all $t \in \mathbb{R}$

$$\zeta(t) \quad := \quad \frac{1}{\sqrt{2\pi}} \left(U(t)v, v\right) . \tag{10.33}$$

Example. *For our eigenvector v we immediately get $\zeta(t) = \frac{1}{\sqrt{2\pi}} e^{i\lambda t}$.*

By the Cauchy–Schwarz inequality (cf. [31]) and since $U(t)$ is an unitary C_0 group, we immediately observe that ζ defined by (10.33) is in $C(\mathbb{R}) \cap L^\infty(\mathbb{R})$. Next we define a map W from the space of Schwartz functions $\mathcal{S}(\mathbb{R})$ to H via (cf. [71, p.77])

$$Wf \quad := \quad \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \widehat{f}(t) U(t)v \; dt \, . \tag{10.34}$$

Since $U(t)v$ is a function from $\mathbb{R}$ to H the expression Wf is defined by the theory of Bochner integrals (cf. [1, p.178f]). It follows from Corollary 10.3.4 that $t \mapsto U(t)v$ is continuous on $\mathbb{R}$. Since the Fourier transformation is a continuous mapping from $\mathcal{S}(\mathbb{R})$ to $\mathcal{S}(\mathbb{R})$ (cf. [65]), $\widehat{f} \in \mathcal{S}(\mathbb{R})$. Thus the integrand $\tilde{f} \colon t \mapsto \widehat{f}(t)U(t)v$ is a continuous map from $\mathbb{R} \to H$. Hence $\tilde{f}$ is measurable w.r.t the Borel algebras on $\mathbb{R}$ and H. Thus the integral exists, if and only if $t \mapsto \|\tilde{f}(t)\|$ is Lebesgue integrable (cf. [1, p.179]). Since $U(t)$ is unitary $\|\tilde{f}(t)\| = |\widehat{f}(t)| \|v\|$, which is integrable if and only if $\widehat{f} \in L^1(\mathbb{R})$. In this case

$$\|Wf\| \quad \leq \quad \frac{\|v\|}{\sqrt{2\pi}} \, \|\widehat{f}\|_{L^1(\mathbb{R})} \, . \tag{10.35}$$

Example. *Let v be our eigenvector from the previous examples and let the function $f \in \mathcal{S}(\mathbb{R})$. Then $Wf = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \widehat{f}(t) e^{i\lambda t} v \; dt = f(\lambda) v$.*

Let $\varphi \colon \mathbb{R} \to H$ be a Bochner integrable function. From [1] we know that the Bochner integral of f is defined (as the Lebesgue integral) by the limit of integrals of simple functions. I. e. there exists a sequence of simple functions ($n \in \mathbb{N}$)

$$\varphi_n \quad = \quad \sum_{j=1}^{m_n} b_{j,n} \, \chi_{A_{j,n}}$$

with[19] $\lambda(A_{j,n}) < \infty$, $b_{j,n} \in H$ for all $j = 1, \ldots, m_n$ and $\varphi_n(t) \to \varphi(t)$ a.e. on $\mathbb{R}$ and

$$\int_{\mathbb{R}} \varphi(t) \; dt \quad := \quad \lim_{n\to\infty} \int_{\mathbb{R}} \varphi_n(t) \; dt \quad = \quad \lim_{n\to\infty} \sum_{j=1}^{m_n} b_{j,n} \, \lambda(A_{j,n}) \, .$$

Since the sequence on the right–hand side converges with respect to the norm on H, we can interchange integration with any linear continuous map. Especially

$$\left(\int_{\mathbb{R}} \varphi(t) \; dt, \, g \right) \quad = \quad \int_{\mathbb{R}} (f(t), \, g) \; dt$$

[19] Here λ is a measure on $\mathbb{R}$, which is in our case the Lebesgue measure.

holds for all $g \in H$. Now let $f, g \in \mathcal{S}(\mathbb{R})$. Then it holds

$$\begin{aligned}
\big(Wf, Wg\big) &= \tfrac{1}{2\pi}\Big(\int_{\mathbb{R}} \widehat{f}(t) U(t)v \, dt, \int_{\mathbb{R}} \widehat{g}(s) U(s)v \, ds \Big) \\
&= \tfrac{1}{2\pi} \int_{\mathbb{R}}\int_{\mathbb{R}} \widehat{f}(t)\, \overline{\widehat{g}(s)}\, \big(U(t)v, U(s)v\big)\, dt\, ds \\
&= \tfrac{1}{2\pi} \int_{\mathbb{R}}\int_{\mathbb{R}} \widehat{f}(t)\, \overline{\widehat{g}(s)}\, \big(U(t-s)v, v\big)\, dt\, ds \\
&= \tfrac{1}{\sqrt{2\pi}} \int_{\mathbb{R}}\int_{\mathbb{R}} \widehat{f}(t)\, \overline{\widehat{g}(s)}\, \zeta(t-s)\, dt\, ds \\
&= \tfrac{1}{\sqrt{2\pi}} \int_{\mathbb{R}}\int_{\mathbb{R}} \widehat{f}(t)\, \overline{\widehat{g}(t-s)}\, \zeta(s)\, dt\, ds \\
&= \tfrac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \zeta(s) \Big(\tfrac{1}{\sqrt{2\pi}} \int_{\mathbb{R}}\int_{\mathbb{R}} \widehat{f}(t)\, \overline{g(x)} e^{i(t-s)x}\, dx\, dt \Big)\, ds \\
&= \tfrac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \zeta(s) \Big(\int_{\mathbb{R}} f(x)\, \overline{g(x)} e^{-isx}\, dx \Big)\, ds \\
&= \int_{\mathbb{R}} \zeta(s)\, \widehat{(f\,\overline{g})}(s)\, ds \\
&= \big\langle T_\zeta, \widehat{f\overline{g}} \big\rangle \,. \qquad (10.36)
\end{aligned}$$

Since $\zeta \in L^\infty(\mathbb{R})$ it defines a (tempered) distribution T_ζ by (10.25) (cf. [65]). Hence its Fourier transform (in the distributional sense) $\widehat{T}_\zeta$ is given by (10.26) and is a tempered distribution too. For $f \in \mathcal{S}(\mathbb{R})$

$$< \widehat{T}_\zeta, f > \;=\; < T_\zeta, \widehat{f} > \,.$$

Example. *It holds* $\widehat{T}_\zeta = \delta_\lambda$.

In the sequel we simply write T instead of $\widehat{T}_\zeta$. For $f = g$ one gets from (10.36)

$$< T, |f|^2 > \;=\; (Wf, Wf) \;\geq\; 0\,, \qquad \text{for all } f \in \mathcal{S}(\mathbb{R}).$$

With this property we can prove Lemma 10.3.30.

Lemma 10.3.30. *There exists a unique Radon measure μ on $\mathbb{R}$, such that*

$$\big\langle T, \phi \big\rangle \;=\; \int_{\mathbb{R}} \phi \, d\mu \,, \qquad \textit{for all } \phi \in \mathcal{S}(\mathbb{R}).$$

Proof. Let $F(t,x,y) := \frac{1}{\sqrt{4\pi t}} e^{-\frac{(x-y)^2}{4t}}$ and let $f(t,x,y) := \sqrt{F(t,x,y)}$. Hence for all fixed $t > 0$ and $x \in \mathbb{R}$ we have $f(t,x,\cdot) \in \mathcal{S}(\mathbb{R})$ which yields

$$< T, F(t,x,\cdot) > \;=\; < T, |f(t,x,\cdot)|^2 > \;=\; (Wf(t,x,\cdot), Wf(t,x,\cdot)) \;\geq\; 0\,.$$

Furthermore $u(t,x) :=< T, F(t,x,\cdot) >$ is a solution of the heat equation (in the tempered distributional sense)

$$\begin{aligned}
\Big(\frac{\partial}{\partial t} - \frac{\partial^2}{\partial x^2} \Big) u(t,x) &= 0\,, \\
u(t=0) &= T\,.
\end{aligned}$$

By [70, Proposition 5.1, p.217] $u(t) \xrightarrow{t\to 0} T$. Thus, T is a nonnegative tempered distribution. Hence, by Corollary 10.3.17 there exists a unique Radon measure μ on $\mathbb{R}$, such that $\langle T, \phi\rangle = \int_{\mathbb{R}} \phi \, d\mu$ holds for all $\phi \in \mathcal{S}(\mathbb{R})$. □

Up to now we have shown

$$\big(Wf, Wg\big) \;=\; \int_{\mathbb{R}} f\overline{g}\, d\mu \;=\; (f,g)_{L^2(\mathbb{R},\mu)}\,, \qquad \forall\ f,g \in \mathcal{S}(\mathbb{R})\,, \quad (10.37)$$

Hence $W\colon \mathcal{S}(\mathbb{R}) \to H$ is an isometry with respect to the $L^2(\mathbb{R},\mu)$ scalar product (cf. [72, Lemma V.5.4, p.234]). Since $\mathcal{D}(\mathbb{R}) \subset \mathcal{S}(\mathbb{R})$, by Proposition 10.3.23 we get that $\mathcal{S}(\mathbb{R})$ is dense in $L^2(\mathbb{R},\mu)$. Hence there exists a unique, continuous extension $\widehat{W}\colon L^2(\mathbb{R},\mu) \to H$ of W, with $\|\widehat{W}\| = \|W\|$ (cf. [72, Satz II.1.5, p.48]). In the sequel we shall denote the extension by W.

Lemma 10.3.31. *The map $W\colon L^2(\mathbb{R},\mu) \to H$ defined by continuous extension of* (10.34) *is unitary.*

Example. *In our example the measure μ is the measure corresponding to the Delta distribution with pole in λ. As we have seen before $Wf = f(\lambda)v$ holds for all $f \in \mathcal{S}(\mathbb{R})$. For a moment let us denote this map by $W_{\mathcal{S}}$. By* (10.37) *the $L^2(\mathbb{R},\mu)$-norm of any Schwartz function is finite. Hence there exists a natural embedding $\iota\colon \mathcal{S}(\mathbb{R}) \to \tilde{L}^2(\mathbb{R},\mu)$ into the space of all functions on $\mathbb{R}$ with finite $L^2(\mathbb{R},\mu)$-norm. Further there exists a surjective map $\sigma\colon \tilde{L}^2(\mathbb{R},\mu) \to L^2(\mathbb{R},\mu)$, namely the quotient map with respect to the equivalence relation induced by $\|\cdot\|_{L^2(\mathbb{R},\mu)}$. For our example $\sigma\circ\iota$ is surjective and $W_{\mathcal{S}} = W\circ\sigma\circ\iota$, which means that for any $[f] \in L^2(\mathbb{R},\mu)$ there exists a representative $f \in \mathcal{S}(\mathbb{R})$, such that $W[f] = W_{\mathcal{S}} f = f(\lambda)v$.*

Proof of Lemma 10.3.31. Since $\mathcal{S}(\mathbb{R})$ is a dense subspace of $L^2(\mathbb{R},\mu)$ (see Proposition 10.3.23), property (10.37) holds on the whole space $L^2(\mathbb{R},\mu)$. In particular it follows that W is injective. Hence it remains to show that W is surjective and thus bijective (cf. [72]).

Firstly we show: $\mathrm{Im}(W)$ is closed. Let $y \in \overline{\mathrm{Im}(W)}$. Hence there exists a sequence $y_n \to y$ with $y_n \in \mathrm{Im}\, W$. Further we remark that $\{y_n\}$ is a Cauchy sequence. Since W is injective, there exist unique $x_n \in L^2$ with $Wx_n = y_n$ and

$$\|x_n - x_m\| \;=\; \|W(x_n - x_m)\| \;=\; \|y_n - y_m\|\,.$$

Hence $\{x_n\}$ is a Cauchy sequence and thus converge to a unique $x \in L^2$. This yields (W is continuous)

$$Wx \;=\; \lim_{n\to\infty} Wx_n \;=\; \lim_{n\to\infty} y_n \;=\; y\,.$$

Hence $y \in \mathrm{Im}\, W$ and thus $\mathrm{Im}\, W = \overline{\mathrm{Im}\, W}$.

Since $\mathrm{Im}\, W$ is closed, we get $H = \mathrm{Im}\, W \oplus (\mathrm{Im}\, W)^{\perp}$. Thus we have to show, that $\mathrm{Im}\, W^{\perp} = 0$.

Let $w \in (\mathrm{Im}\, W)^{\perp}$. Furthermore let $s \in \mathbb{R}$ and let $\widehat{f_n}(x) = n\varphi(n(x-s))$ with $\varphi(x) = \frac{1}{\sqrt{\pi}} e^{-x^2}$. Hence $\widehat{f_n}$ and its inverse Fourier transform f_n are in $\mathcal{S}(\mathbb{R})$. Furthermore it holds for all $g \in C(\mathbb{R}) \cap L^{\infty}(\mathbb{R})$:

$$\begin{aligned}\int_{\mathbb{R}} \widehat{f_n}(t) g(t)\, dt \;&=\; \frac{1}{\sqrt{\pi}} \int_{\mathbb{R}} n e^{-(n(t-s))^2} g(t)\, dt \\ &=\; \frac{1}{\sqrt{\pi}} \int_{\mathbb{R}} e^{-y^2} g\big(\tfrac{y}{n} + s\big)\, dy\,.\end{aligned}$$

Since g is bounded, we can apply the *dominated convergence theorem of Lebesgue* and pass to the limit $n \to \infty$. Hence (g is continuous)

$$\lim_{n\to\infty} \int_{\mathbb{R}} \widehat{f}_n(t) g(t)\, dt \;=\; g(s)\,.$$

This yields

$$\begin{aligned} 0 \;=\; (w, Wf_n) \;&=\; \Big(w, \tfrac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \widehat{f}_n(t) U(t)v\, dt\Big) \\ &=\; \tfrac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \widehat{f}_n(t) \big(w, U(t)v\big)\, dt \overset{n\to\infty}{\longrightarrow} \big(w, U(s)v\big)\,. \end{aligned}$$

Here we use that the map $t \mapsto (w, U(t)v)$ is bounded and continuous. Since $s \in \mathbb{R}$ is arbitrary,

$$(w, U(t)v) \;=\; 0 \qquad \text{for all } t \in \mathbb{R}.$$

Furthermore v is a cyclic vector of H, i. e. $H = \overline{\operatorname{span}\{U(t)v | t \in \mathbb{R}\}}$. Thus there exists a sequence $w_n \in \operatorname{span}\{U(t)v | t \in \mathbb{R}\}$, such that $w_n \to w$ in H. Since any w_n is a finite linear combination of vectors from $\{U(t)v | t \in \mathbb{R}\}$, it follows $(w, w_n) = 0$ for all $n \in \mathbb{N}$. Consequently

$$\|w\|^2 \;=\; (w, w) \;=\; \lim_{n\to\infty} (w, w_n) \;=\; 0\,.$$

Here we use the continuity of the scalar product and the convergence with respect to the norm. Thus w is zero and hence $(\operatorname{Im} W)^\perp = 0$, which means $H = \operatorname{Im} W$. □

As we have seen, $W\colon L^2(\mathbb{R}, \mu) \to H$ is an unitary map. Let us compute its adjoint map. Therefore let $u \in \operatorname{span}\{U(t)v | t \in \mathbb{R}\} \subset H$, i. e. $u = \sum_{j=1}^n c_j U(t_j)v$ with $c_j \in \mathbb{C}$, and let $g \in \mathcal{S}(\mathbb{R}) \subset L^2(\mathbb{R}, \mu)$. Both subspaces are dense respectively. Furthermore we define $f(x) := \sum_{j=1}^n c_j e^{it_j x}$. It holds

$$\begin{aligned} (Wg, u) \;&=\; \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \widehat{g}(t) \big(U(t)v, u\big)\, dt \\ &=\; \sum_{j=1}^n \int_{\mathbb{R}} \overline{c}_j\, \widehat{g}(t)\, \frac{1}{\sqrt{2\pi}} \big(U(t)v, U(t_j)v\big)\, dt \\ &=\; \sum_{j=1}^n \int_{\mathbb{R}} \overline{c}_j\, \widehat{g}(t)\, \zeta(t - t_j)\, dt \\ &=\; \sum_{j=1}^n \int_{\mathbb{R}} \overline{c}_j\, \widehat{g}(s + t_j)\, \zeta(s)\, ds \\ &=\; \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} g(x) \overline{\Big(\sum_{j=1}^n c_j e^{it_j x}\Big)} e^{-isx}\, dx\, \zeta(s)\, ds \\ &=\; \int_{\mathbb{R}} \widehat{(g\overline{f})}(s)\, \zeta(s)\, ds\,. \end{aligned}$$

Since $g \in \mathcal{S}(\mathbb{R})$, also $g\overline{f} \in \mathcal{S}(\mathbb{R})$ and hence the last integral is nothing but $\widehat{\zeta}$ (in the tempered distributional sense) applied to $g\overline{f}$. Due to the construction of the measure μ we get

$$(Wg, u)_H \;=\; \langle T_{\widehat{\zeta}}, g\overline{f}\rangle \;=\; \int_{\mathbb{R}} g\overline{f}\, d\mu \;=\; (g, f)_{L^2(\mathbb{R},\mu)}\,.$$

Hence $f = W^*u = W^{-1}u$. Since $H = \overline{\operatorname{span}\{U(t)v | t \in \mathbb{R}\}}$, there are for every $u \in H$ a countable number of points $t_j \in \mathbb{R}$ and coefficients $c_j \in \mathbb{C}$, such that $u = \sum_{j\in\mathbb{N}} c_j U(t_j)v$. Hence $W^*u \;=\; W^{-1}u = \sum_{j\in\mathbb{N}} c_j e^{it_j x}$.

Let $f_s(x) := e^{ixs}f(x)$. Hence for every $f \in \mathcal{S}(\mathbb{R})$ it holds

$$\begin{aligned}\widehat{f}(t-s) &= \tfrac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} f(x)e^{-i(t-s)x}\,dx\\ &= \tfrac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{isx}f(x)e^{-itx}\,dx \;=\; \widehat{f_s}(t)\,.\end{aligned}$$

Thus we get for all Schwartz functions f

$$\begin{aligned}U(s)Wf &= \tfrac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \widehat{f}(t)U(t+s)v\,dt\\ &= \tfrac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \widehat{f}(y-s)U(y)v\,dy \;=\; Wf_s\,,\end{aligned}$$

which yields

$$\big(W^{-1}U(t)Wf\big)(x) \;=\; f_t(x) \;=\; e^{ixt}f(x)\,.$$

Since $\mathcal{S}(\mathbb{R})$ is dense in $L^2(\mathbb{R}, d\mu)$ we have proven

Proposition 10.3.32. *Let A be a self–adjoint operator on a separable Hilbert space H, having a cyclic vector v. Then there exists a Borel measure μ on $\mathbb{R}$ and a unitary map $W\colon L^2(\mathbb{R},\mu) \longrightarrow H$, such that*

$$\big(W^{-1}U(t)Wf\big)(x) \;=\; e^{itx}f(x)$$

holds for all $f \in L^2(\mathbb{R},\mu)$.

Corollary 10.3.33. *Let $f \in \mathcal{S}(\mathbb{R})$. Then it holds*

$$(W^{-1}AWf)(x) \;=\; x\,f(x)\,.$$

Proof. The boundary terms of the integration by parts (see the computation below) vanish since $\widehat{f}$ is a Schwartz function. Since we do not prove that integration by parts holds for the Bochner integral, the computations are only formal. We get

$$\begin{aligned}(W^{-1}AWf)(x) &= \left(W^{-1}\tfrac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \widehat{f}(t)\; AU(t)v\;dt\right)(x)\\ &= \left(W^{-1}\tfrac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \widehat{f}(t)\left(-i\frac{d}{dt}U(t)v\right)dt\right)(x)\\ &= \left(W^{-1}\tfrac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \left(i\frac{d}{dt}\widehat{f}(t)\right)U(t)v\;dt\right)(x)\,.\end{aligned}$$

Let $f^\dagger(x) := xf(x)$. Since it holds

$$\frac{d}{dt}\widehat{f}(t) \;=\; -i\,\tfrac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} xf(x)e^{-ixt}\,dx \;=\; -i\widehat{f^\dagger}(t)\,,$$

we get

$$(W^{-1}AWf)(x) \;=\; (W^{-1}Wf^\dagger)(x) \;=\; xf(x)\,.$$

□

By Corollary 10.3.33 we know, that A acts on $\mathcal{S}(\mathbb{R}) \subset L^2(\mathbb{R},\mu)$ via W like the multiplication with the identity function $x \mapsto x$. Thus, by induction A^n is nothing but the multiplication of f with the function $x \mapsto x^n$. Hence it is reasonable to define

Definition 10.3.34. *Let $f \in \mathcal{S}(\mathbb{R})$. Then $f(A)\colon H \to H$ is defined by*

$$f(A) \;:=\; \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \widehat{f}(t)\,U(t)\,dt\,.$$

Since $\widehat{f} \in \mathcal{S}(\mathbb{R}) \subset L^1(\mathbb{R})$ and $U(t)$ is continuous and bounded, the integral is well–defined.

The Definition 10.3.34 is not very precise, since we do not specify the domain of the operator $f(A)$. In general $D(f(A)) \neq H$. The following computation lead to an alternative definition of $f(A)$, which also allow a description of $D(f(A))$. For $f, g \in \mathcal{S}(\mathbb{R})$ it holds

$$\begin{aligned}
f(A)(Wg) &= \tfrac{1}{2\pi}\int_{\mathbb{R}} \widehat{f}(s)U(s)\,ds \int_{\mathbb{R}} \widehat{g}(t)U(t)v\,dt \\
&= \tfrac{1}{2\pi}\int_{\mathbb{R}}\int_{\mathbb{R}} \widehat{f}(s)\widehat{g}(t)U(t+s)v\,dt\,ds \\
&= \tfrac{1}{2\pi}\int_{\mathbb{R}}\int_{\mathbb{R}} \widehat{f}(s)\widehat{g}(r-s)U(r)v\,dr\,ds \\
&= \tfrac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \widehat{(fg)}(r)U(r)v\,dr \qquad (10.38)\\
&= W(fg)\,. \qquad (10.39)
\end{aligned}$$

Applying W^{-1} to (10.39) yields $W^{-1}f(A)Wg = fg$. Hence, the map $f(A)$ acts on $L^2(\mathbb{R},\mu)$ as the multiplication with f. This is well defined for all $g \in L^2(\mathbb{R},\mu)$, such that $fg \in L^2(\mathbb{R},\mu)$. Since this is a much weaker assumption than $f, g \in \mathcal{S}(\mathbb{R})$ or $\widehat{f} \in L^1(\mathbb{R})$ (as in Definition 10.3.34) we can interpret (10.39) as a weak definition of $f(A)$, which makes sense for all Borel measurable functions (as long as g is sufficiently smooth).

Definition 10.3.35. *Let $f\colon \mathbb{R} \to \mathbb{R}$ be Borel measurable and let*

$$D\big(f(A)\big) \;:=\; \{u = Wg \in H |\, g \in L^2(\mathbb{R},\mu),\, fg \in L^2(\mathbb{R},\mu)\} \;\subset\; H\,.$$

Then we define $f(A)\colon D(f(A)) \to H$ by

$$f(A)u \;:=\; W(fW^{-1}u) \;=\; W(fg)\,,$$

with $g = W^{-1}u \in L^2(\mathbb{R},\mu)$.

Remark 10.3.36. *Since $U(\xi)$ is unitary, it holds $\|U(\xi)v\| = \|v\|$ and hence* (10.38) *holds iff $\widehat{fg} \in L^1(\mathbb{R})$ (see comments on the Bochner integral on p.231). In this case*

$$f(A)u \;=\; \tfrac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \widehat{(fg)}(\xi) U(\xi) v \, d\xi \;=\; W(fg)\,.$$

The property $\widehat{fg} \in L^1(\mathbb{R})$ implies[20] *$fg \in C_0(\mathbb{R})$ and $\|fg\|_{L^\infty(\mathbb{R})} \leq \|\widehat{fg}\|_{L^1(\mathbb{R})}$* (cf. [65, 7.5 Theorem, p.185]). *This yields $fg \in L^\infty(\mathbb{R}, \mu)$. Applying $(\cdot, v)$ to the equation yields*

$$\mathbb{C} \ni (f(A)u, v) \;=\; \int_{\mathbb{R}} \widehat{(fg)}(\xi) \zeta(\xi) \, d\xi \;=\; < T_\zeta, \widehat{fg} > \;.$$

Provided fg is in the domain of the tempered distribution $T_{\widehat{\zeta}}$, then

$$\mathbb{C} \ni (f(A)u, v) \;=\; < T_{\widehat{\zeta}}, fg > \;=\; \int_{\mathbb{R}} fg \, d\mu\,.$$

Hence $fg \in L^1(\mathbb{R}, \mu) \cap L^\infty(\mathbb{R}, \mu)$.

Remark 10.3.37. *If $f \in L^\infty(\mathbb{R}, \mu)$, then $fg \in L^2(\mathbb{R}, \mu)$ for all $g \in L^2(\mathbb{R}, \mu)$. Hence $D(f(A)) = H$ and*

$$\begin{aligned} \|f(A)u\|_H \;&=\; \|W(fW^{-1}u)\| \;=\; \|fW^{-1}u\|_{L^2(\mathbb{R},\mu)} \\ &\leq\; \|f\|_{L^\infty(\mathbb{R},\mu)} \|W^{-1}u\|_{L^2(\mathbb{R},\mu)} \;=\; \|f\|_{L^\infty(\mathbb{R},\mu)} \|u\|_H\,. \end{aligned}$$

Thus $f(A)$ is a bounded operator.

Let us shortly discuss some properties (and quantities) of the approach for a simple example. Let $A = -i\partial_x$ on $H_0^1(\mathbb{R}) \subset L^2(\mathbb{R})$. The operator is self adjoint and hence (by Stone's theorem) it is the infinitesimal generator of an unitary group $U(t)$. From Example 10.3.6 on page 217 we know that $(U(t)u)(x) = u(x+t)$. Hence

$$\zeta(t) \;=\; \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} (U(t)v)(x) \overline{v}(x) \, dx \;=\; \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} v(x+t) \overline{v}(x) \, dx\,.$$

Let $v \in \mathcal{S}(\mathbb{R}) \subset H_0^1(\mathbb{R})$. Thus the Fourier transform of ζ is given by

$$\begin{aligned} \widehat{\zeta}(s) \;&=\; \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} v(x+t) \overline{v}(x) \, dx \, e^{-its} \, dt \\ &=\; \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \overline{v}(x) \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} v(\tau) e^{-i(\tau - x)s} \, d\tau \, dx \\ &=\; \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \overline{v}(x) e^{ixs} \, dx \, \widehat{v}(s) \\ &=\; \overline{\widehat{v}(s)} \, \widehat{v}(s) \;=\; \left|\widehat{v}(s)\right|^2 \end{aligned}$$

Hence the measure μ from Lemma 10.3.30 is a density measure with respect to the Lebesgue measure, with density $|\widehat{v}|^2$. Thus $L^2(\mathbb{R}, \mu)$ is a weighted L^2 space. Furthermore it holds for all $w \in \mathcal{S}(\mathbb{R})$ that

$$(U(t)v, w) \;=\; \int_{\mathbb{R}} v(x+t) w(x) \, dx \;=\; \int_{\mathbb{R}} v(y) w(y-t) \, dy \;=\; (v * \check{w})(t)\,,$$

[20] "Here $C_0(\mathbb{R})$ is the supremum–normed Banach space of all complex continuous functions that vanish at infinity." [65, p.185]

with $\check{w}(x) = w(-x)$. Hence $\check{w} \in \mathcal{S}(\mathbb{R})$ and [65, 7.8 Theorem, p.188] yields $(U(t)v, w) \in \mathcal{S}(\mathbb{R})$.

For the map $W\colon L^2(\mathbb{R},\mu) \to H_v$ we get ($f \in L^1(\mathbb{R})$)

$$\begin{aligned}(Wf)(x) &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \widehat{f}(t)\,(U(t)v)(x)\;dt \;=\; \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \widehat{f}(t)\,v(x+t)\;dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(s)\,\widehat{v}(s)\,e^{isx}\;ds \;=\; \mathcal{F}^{-1}(f\widehat{v})(x)\,.\end{aligned}$$

This is a nice explicit formula of the map W.

Remark 10.3.38. *During the course of the above discussion we only needed the group properties $U(t)U(s) = U(t+s)$, $U(-t) = U(t)^{-1} = U(t)^*$ and that the map $t \mapsto U(t)u$ is continuous and bounded for all $u \in H$, where H is a separable Hilbert space, having a cyclic vector.*

Now let $A_1, \dots, A_k$ be commutating self–adjoint operators on H and let $U_1, \dots, U_k$ be the corresponding unitary groups. For $t \in \mathbb{R}^k$ we define $U(t) = U_1(t_1) \dots U_k(t_k)$. Since $A_1, \dots, A_k$ commute, U satisfies $U(t)U(s) = U(t+s)$ for all $t, s \in \mathbb{R}^k$. Furthermore $U(-t) = U(t)^{-1} = U(t)^*$ and the map $t \mapsto U(t)v$ is continuous and bounded. Thus we can repeat the above construction and define $f(A)$ for the linear map $A = (A_1, \dots, A_k)\colon H^k \to H^k$

$$f(A) \;=\; \frac{1}{\sqrt{2\pi}^k} \int_{\mathbb{R}^k} \widehat{f}(t)U(t)\;dt$$

for all $f \in \mathcal{S}(\mathbb{R}^k)$. For example this ansatz can be used to construct a functional calculus for normal operators. See [71, p.78f] *for more details.*

10.4 Application

In this section we shall (often formally) apply the derived functional calculus for some examples, which are connected to the OWWE. We start with the derivation of the square root of an self–adjoint operator in § 10.4.1. Next, in § 10.4.2, we formally derive a solution formula for the OWWE. In § 10.4.3 we shall show, how the transformation from [12] can be obtained from the presented ansatz. The mentioned transformation is a crucial tool for the exact symbol contraction procedure of the SRHO as proposed in [21].

10.4.1 The square root of a self–adjoint operator

Let A be a self–adjoint operator on a separable Hilbert space H, having a cyclic vector v. The aim of this subsection is to derive the square root of A. It turns out, that it is easier to compute $A^{-\frac{1}{2}}$ instead of $A^{\frac{1}{2}}$. Thus we shall derive a formula for $A^{-\frac{1}{2}}$ and use $A^{\frac{1}{2}} = AA^{-\frac{1}{2}}$.

Our motivation is the OWWE from § 10.1. There one has to compute the square root of the differential operator (10.15), which reads $A = \partial_z^2 + \kappa^2 V(t)^2$. In the first part of the thesis we derived efficient numerical tools to derive the solution of $\Psi_{zz} + \kappa^2(V^2 + \xi^2)\Psi = f$, for $\xi \in \mathbb{R}$ and $V \geq \delta > 0$. As we will see, this equation appears in the construction of $A^{-\frac{1}{2}}$.

We start this section with a technical result and we shall define the branch of the complex square-root we use for our purpose (i. e. we define the "right"

branch for the OWWE). Afterwards we compute the Fourier transform of the map $x \mapsto x^{-\frac{1}{2}}$ and finally derive the desired formula for $A^{-\frac{1}{2}}$.

Lemma 10.4.1. *Let $f \in L^1([0,\infty),\mathbb{R})$ and let*[21] $\mathbb{C}_s := \{z \in \mathbb{C} | \operatorname{Im} z < 0\}$. *Then $\widehat{f}\colon \overline{\mathbb{C}_s} \to \mathbb{C}$, pointwise defined by*

$$\widehat{f}(z) \;:=\; \int_0^\infty f(t)e^{-itz}\,dt\,,$$

is well–defined and complex differentiable on $\mathbb{C}_s$. It holds for all $z \in \mathbb{C}_s$

$$\frac{d}{dz}\widehat{f}(z) \;=\; -i\int_0^\infty tf(t)e^{-itz}\,dt\,.$$

Proof. Since $f \in L^1([0,\infty),\mathbb{R})$, this also holds for $|f|$. For $z \in \overline{\mathbb{C}_s}$ it further holds for (almost every) $t \in [0,\infty)$:

$$|f(t)e^{-itz}| \;\le\; |f(t)|e^{-t|\operatorname{Im} z|} \;\le\; |f(t)|\,.$$

Hence the function $|f|$ is an integrable upper bound for $g(t,z) := f(t)e^{-itz}$, which yields $g(\cdot,z) \in L^1([0,\infty),\mathbb{C})$. Thus $\widehat{f}$ is well defined.

Now let $z = a - ib \in \mathbb{C}_s$, i. e. $b < 0$. Since for all (real) $x \ge 0$ it holds $x \le e^x$, it follows $xe^{-x} \le 1$, which yields

$$te^{-tb} \;\le\; \frac{1}{b}\,.$$

Thus $\frac{1}{b}|f|$ is an integrable upper bound for the function $t \mapsto tg(t,z)$. Hence

$$\widehat{f}^{\dagger}(z) \;:=\; -i\int_0^\infty tf(t)e^{-itz}\,dt$$

exists. Let $h \in \mathbb{C}$ with $|h| < b = |\operatorname{Im} z|$. Hence it holds $z + h \in \mathbb{C}_s$. Further we get (e. g. with the rule of L'Hospital)

$$\lim_{h\to 0} \tfrac{1}{h}(e^{-ith} - 1) \;=\; -it\,.$$

Thus for every $\varepsilon > 0$ exists $\delta \ge 0$, such that for all $|h| \le \delta$ it holds:

$$\left|\tfrac{1}{h}(e^{-ith} - 1) + it\right| \;\le\; \varepsilon.$$

With the lower triangle inequality we conclude

$$\left|\tfrac{1}{h}(e^{-ith} - 1)\right| \;\le\; \varepsilon + t\,.$$

Hence $|f|(\varepsilon + \frac{1}{b})$ is and integrable upper bound for $\frac{1}{h}(e^{-ith}-1)g(t,z)$. Thus the dominated convergence theorem of Lebesgue yields ($|h| \le \min(\delta, \operatorname{Im} z)$)

$$\tfrac{1}{h}\big(\widehat{f}(z+h) - \widehat{f}(z)\big) \;=\; \int_0^\infty f(t)\tfrac{1}{h}(e^{-ith} - 1)e^{-itz}\,dt \;\overset{h\to 0}{\longrightarrow}\; \widehat{f}^{\dagger}(z)\,.$$

Hence $\widehat{f}$ is complex differentiable with $\widehat{f}' = \widehat{f}^{\dagger}$. □

[21] We name the most common half planes of $\mathbb{C}$ (upper, lower, left, right) after the cardinal points. With the convention that the imaginary unit i is in the norther plane we have fixed the compass rose. Thus the lower index s is an abbreviation for "south".

Remark 10.4.2. *In the proof of Lemma 10.4.1 we only use the dominated convergence theorem of Lebesgue. Since the theorem holds for all measures μ on $[0,\infty)$, the lemma holds for all $f \in L^1([0,\infty),\mu)$.*

Corollary 10.4.3. *Lemma 10.4.1 also holds for $f \in L^1([0,\infty),\mathbb{C})$.*

Proof. Just split f into real and imaginary part and apply Lemma 10.4.1 to the gained real valued functions. □

Now we define the branch of the complex square root we shall use. It has to be, such that $\operatorname{Im}\sqrt{z}$ is nonnegative for all $z \in \mathbb{C}$. Since z^α (with $\alpha \in \mathbb{C}$ and $z \in \mathbb{C}^*$) is defined by the formula

$$z^\alpha \quad := \quad e^{\alpha \log z}\,, \tag{10.40}$$

we have to specify the appropriate branch of the logarithm. To this end let us restrict the exponential function to the strip $D = \{z \in \mathbb{C} |\, 0 \le \operatorname{Im} z < 2\pi\}$. Hence $\exp|_D$ is bijective and its inverse function reads[22]

$$\log(z) \quad := \quad \ln|z| \,+\, i\arg(z)\,,$$

with ($z = a + ib$)

$$\arg(z) \quad := \quad \begin{cases} \arctan\left(\frac{b}{a}\right) & ,\ a>0,\ b\ge 0\,,\\ \frac{\pi}{2} & ,\ a=0, b>0\,,\\ \arctan\left(\frac{b}{a}\right)+\pi\,, & a<0\,,\\ \arctan\left(\frac{b}{a}\right)+2\pi, & a>0,\ b<0\,,\\ -\frac{\pi}{2} & ,\ a=0, b<0\,. \end{cases}$$

The function arg is continuous on $\mathbb{C}\backslash\mathbb{R}_0^+$ and jumps when crossing the positive real line. It is also (one–sided) continuous when approaching $\mathbb{R}^+$ for the upper half plane. Since exp is complex differentiable on $D^o = \{z \in \mathbb{C} |\, 0 < \operatorname{Im} z < 2\pi\}$ (the interior of D) it is biholomorphic (cf. [64, p.221]) and hence log is complex differentiable on $\exp(D^o) = \mathbb{C}\backslash\mathbb{R}_0^+$. Hence all power functions given by (10.40) are holomorphic on $\mathbb{C}\backslash\mathbb{R}_0^+$ too.

To apply the functional calculus of § 10.3.5 we have to compute the Fourier transform of the function $f\colon \mathbb{R}^* \to \mathbb{C}$, $f(x) = x^{-\frac{1}{2}}$. To this end we rewrite f in the form (for $x \in \mathbb{R}^*$)

$$\begin{aligned} f(x) \;=\; x^{-\frac{1}{2}} \quad &= \quad \tfrac{1+\operatorname{sgn}(x)}{2}|x|^{-\frac{1}{2}} \,-\, i\tfrac{1-\operatorname{sgn}(x)}{2}|x|^{-\frac{1}{2}} \\ &= \quad \tfrac{1-i}{2}|x|^{-\frac{1}{2}} + \tfrac{1+i}{2}\operatorname{sgn}(x)|x|^{-\frac{1}{2}}\,. \end{aligned}$$

Since the Fourier transform is a linear map, we only need the Fourier transform of $x \mapsto |x|^{-\frac{1}{2}}$ and $x \mapsto \operatorname{sgn}(x)|x|^{-\frac{1}{2}}$. The following table is from the textbook [50, p.86] ($\nu \notin \mathbb{Z}$):

$g(x)$	$\int_{\mathbb{R}} g(x)\, e^{-i2\pi\xi x} dx$
$\lvert x\rvert^\nu$	$-2\,\Gamma(\nu+1)\,\sin(\nu\frac{\pi}{2})\,(2\pi\lvert\xi\rvert)^{-\nu-1}$
$\operatorname{sgn}(x)\lvert x\rvert^\nu$	$-2\,\Gamma(\nu+1)\,\cos(\nu\frac{\pi}{2})\,(2\pi\lvert\xi\rvert)^{-\nu-1}\,\operatorname{sgn}(\xi)\,i$

[22] As usual we denote by ln the real logarithm.

Here $\operatorname{sgn}\xi$ is the real signum of ξ and Γ denotes the Gamma function. From the previously given table one gets

$$\begin{aligned}\int_{\mathbb{R}} f(x)\, e^{-i2\pi\xi x}\, dx &= -(1-i)\,\Gamma\big(\tfrac{1}{2}\big)\,\sin\big(-\tfrac{\pi}{4}\big)\,(2\pi|\xi|)^{-\frac{1}{2}} \\ &\quad -(1+i)\,\Gamma\big(\tfrac{1}{2}\big)\,\cos\big(-\tfrac{\pi}{4}\big)\,(2\pi|\xi|)^{-\frac{1}{2}}\,\operatorname{sgn}(\xi)\,i \\ &= \frac{1}{2\sqrt{|\xi|}}\big((1-i)-i(1+i)\operatorname{sgn}\xi\big) \\ &= (1-i)|\xi|^{-\frac{1}{2}}\,H(\xi)\,,\end{aligned}$$

where H denotes the Heaviside function. We set $\xi = \frac{t}{2\pi}$ which yields

$$\widehat{f}(t) \;=\; \frac{1}{\sqrt{2\pi}}\,(1-i)\left|\frac{t}{2\pi}\right|^{-\frac{1}{2}} H\big(\tfrac{t}{2\pi}\big) \;=\; (1-i)|t|^{-\frac{1}{2}}\,H(t)\,.$$

The function $\widehat{f}$ is the tempered distributional Fourier transform of f.

Let $W\colon L^2(\mathbb{R},\mu)\to H$ be the unitary map from Proposition 10.3.32 and let $U(t)$ be the group of unitary operators with infinitesimal generator A, i. e. U solves

$$U_t \;=\; iAU\,, \qquad U(t=0) \;=\; \mathrm{Id}\,.$$

Furthermore let $g\in\mathcal{S}(\mathbb{R})$ with $g=0$ in an open neighborhood of zero and let $u := Wg \in H$. For $w\in H$ we define the function $h\colon\mathbb{R}\to\mathbb{C}$ by

$$h(t) \;:=\; (U(t)v, w).$$

Since $U(t)$ is unitary, the Cauchy–Schwarz inequality yields boundedness of h. Furthermore, by Remark 10.3.8 the function h is also continuously differentiable. Since g vanishes identically in an open neighborhood of zero and since $f\in C^\infty(\mathbb{R}^*,\mathbb{C})$, $fg\in\mathcal{S}(\mathbb{R})$. By Definition 10.3.35 it holds

$$\begin{aligned}(f(A)u,w) &= (W(fg),w) \;=\; \tfrac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \widehat{(fg)}(t)\big(U(t)v,w\big)\,dt \\ &= \tfrac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \widehat{(fg)}(t)\,h(t)\,dt\,.\end{aligned}$$

We can interpret f as a tempered distribution. Thus, by [65, 7.19 Theorem, p.195] it holds $\sqrt{2\pi}\widehat{(fg)} = (\widehat{f}*\widehat{g})$. Since $\widehat{f}$ is a function,

$$(f(A)u,w) \;=\; \tfrac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\left(\tfrac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\widehat{f}(\xi)\widehat{g}(t-\xi)\,d\xi\right)h(t)\,dt\,.$$

Since $g\in\mathcal{S}(\mathbb{R})$, there exists a constant $\sqrt{2}\le c\in\mathbb{R}$, such that for all $x\in\mathbb{R}$: $g(x)\le c(1+x^2)^{-1}$. Hence

$$\begin{aligned}\int_{\mathbb{R}} |\widehat{f}(\xi)||\widehat{g}(t-\xi)|\,d\xi &\le \int_{[-1,1]} |\widehat{f}(\xi)|\,c\,d\xi \;+\; \int_{\mathbb{R}\setminus[-1,1]} c\,|\widehat{g}(t-\xi)|\,d\xi \\ &\le c^2 2\int_0^1 \xi^{-\frac{1}{2}}\,d\xi \;+\; c^2\int_{\mathbb{R}}\frac{1}{1+(t-\xi)^2}\,d\xi \\ &= 4c^2 \;+\; c^2\int_{\mathbb{R}}\frac{1}{1+x^2}\,dx \;=\; c^2(4+\pi) < \infty\,.\end{aligned}$$

Thus, if we assume $h \in L^1(\mathbb{R})$, then

$$\int_{\mathbb{R}} \left(\int_{\mathbb{R}} |\widehat{f}(\xi)||\widehat{g}(t-\xi)||h(t)| \, d\xi \right) dt$$

exists and we can apply Fubini's theorem (cf. [16, 2.1 Satz von G. Fubini, p.173f]) to interchange the order of integration. This yields

$$(f(A)u, w) \quad = \quad \tfrac{1}{2\pi} \int_{\mathbb{R}} \widehat{f}(\xi) \left(H(\xi) \int_{\mathbb{R}} \widehat{g}(t-\xi) h(t) \, dt \right) d\xi \, .$$

Furthermore we assume that we can shift the Fourier transform from f to the other factor, i. e.

$$(f(A)u, w) \quad = \quad \tfrac{1}{2\pi} \int_{\mathbb{R}} f(s) \, \tfrac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \left(H(\xi) \int_{\mathbb{R}} \widehat{g}(t-\xi) h(t) \, dt \right) e^{-i\xi s} \, d\xi \, ds \, .$$

For $s \in \mathbb{R}$ we (formally) define

$$\psi(s) \quad := \quad \tfrac{1}{\sqrt{2\pi}} \int_0^\infty \int_{\mathbb{R}} \widehat{g}(t-\xi) U(t) v \, dt \, e^{-i\xi s} \, d\xi \, .$$

This yields $(f(A)u, w) = \frac{1}{2\pi} \int_{\mathbb{R}} f(s)(\psi(s), w) \, ds$. By Remark 10.3.8, h is continuously differentiable with $h'(t) = (iAU(t)v, w)$. Integration by parts yields

$$\begin{aligned}
(iA\psi(s), w) \quad &= \quad \tfrac{1}{\sqrt{2\pi}} \int_0^\infty \int_{\mathbb{R}} \widehat{g}(t-\xi)(iAU(t)v, w) \, dt \, e^{-i\xi s} \, d\xi \\
&= \quad \tfrac{1}{\sqrt{2\pi}} \int_0^\infty \int_{\mathbb{R}} \widehat{g}(t-\xi) \, h'(t) \, dt \, e^{-i\xi s} \, d\xi \\
&= \quad -\tfrac{1}{\sqrt{2\pi}} \int_0^\infty \int_{\mathbb{R}} \frac{\partial \widehat{g}(t-\xi)}{\partial t} \, h(t) \, dt \, e^{-i\xi s} \, d\xi \, .
\end{aligned}$$

Since $\widehat{g} \in \mathcal{S}(\mathbb{R})$, the boundary terms are zero. Furthermore $\frac{\partial \widehat{g}(t-\xi)}{\partial t} = -\frac{\partial \widehat{g}(t-\xi)}{\partial \xi}$, which yields with integration by parts:

$$\begin{aligned}
(iA\psi(s), w) \quad &= \quad \tfrac{1}{\sqrt{2\pi}} \int_0^\infty \frac{d}{d\xi} \left(\int_{\mathbb{R}} \widehat{g}(t-\xi) h(t) \, dt \right) e^{-i\xi s} \, d\xi \\
&= \quad -\tfrac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \widehat{g}(t) h(t) \, dt \\
&\qquad -\tfrac{1}{\sqrt{2\pi}} \int_0^\infty \int_{\mathbb{R}} \widehat{g}(t-\xi) h(t) \, dt \, (-is) \, e^{-i\xi s} \, d\xi \\
&= \quad -\left(\tfrac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \widehat{g}(t) U(t) v \, dt, \, w \right) \\
&\qquad + \left(is \, \tfrac{1}{\sqrt{2\pi}} \int_0^\infty \int_{\mathbb{R}} \widehat{g}(t-\xi) \, U(t) v \, dt \, e^{-i\xi t} \, d\xi \, , w \right) .
\end{aligned}$$

By definition of W (cf. (10.34)) it holds $u = Wg = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \widehat{g}(\xi) U(\xi) v \, d\xi$. Hence

$$(iA\psi(s), w) \quad = \quad (-u + is \, \psi(s), w) \, .$$

Let S be the set of all $w \in H$ for which the above computations hold. If we assume that S is dense in H, then $\psi(s)$ is the weak solution of the inhomogeneous linear equation

$$(A - s\,\mathrm{Id})\psi(s) \quad = \quad i\,u\,.$$

As the following discussion shows it is possible to restrict the integral

$$(f(A)u, w) \quad = \quad \tfrac{1}{2\pi}\int_{\mathbb{R}} f(s)(\psi(s), w)\; ds$$

to the positive or negative real line (of course we also have to multiply it with an appropriate constant factor). The following calculations are motivated by [21], where a similar trick, as we shall use in moment, is used. We assume that $w \in H$, such that $h \in L^1(\mathbb{R})$. Then Integration by parts yields for $s \in \overline{\mathbb{C}_s}\backslash\{0\}$:

$$\begin{aligned}
(\psi(s), w) &= \tfrac{1}{\sqrt{2\pi}}\int_0^\infty \int_{\mathbb{R}} \widehat{g}(t-\xi)h(t)\; dt \left(\frac{i}{s}\frac{\partial}{\partial\xi}e^{-i\xi s}\right) d\xi \\
&= \tfrac{1}{\sqrt{2\pi}}\frac{i}{s}\lim_{\xi\to\infty}\int_{\mathbb{R}} \widehat{g}(t-\xi)h(t)\; dt\, e^{-i\xi s} \\
&\quad - \tfrac{1}{\sqrt{2\pi}}\frac{i}{s}\int_{\mathbb{R}} \widehat{g}(t)h(t)\; dt \\
&\quad + \tfrac{1}{\sqrt{2\pi}}\frac{i}{s}\int_0^\infty \int_{\mathbb{R}} \widehat{g}'(t-\xi)h(t)\; dt\, e^{-i\xi s}\; d\xi\,.
\end{aligned}$$

Since $\widehat{g}' \in \mathcal{S}(\mathbb{R}) \subset L^1(\mathbb{R})$ it holds

$$\begin{aligned}
\int_{\mathbb{R}}\left(\int_0^\infty |\widehat{g}'(t-\xi)|\; d\xi\right)|h(t)|\; dt &= \int_{\mathbb{R}}\left(\int_{-\infty}^t |\widehat{g}'(y)|\; dy\right)|h(t)|\; dt \\
&\leq \|\widehat{g}'\|_{L^1(\mathbb{R})}\,\|h\|_{L^1(\mathbb{R})}\,.
\end{aligned}$$

Hence the integral on the left hand side exists and is finite. By Fubini's theorem it follows

$$\begin{aligned}
\int_0^\infty \int_{\mathbb{R}} |\widehat{g}'(t-\xi)||h(t)|\; dt\; d\xi &= \int_{\mathbb{R}}\left(\int_0^\infty |\widehat{g}'(t-\xi)|\; d\xi\right)|h(t)|\; dt \\
&\leq \|\widehat{g}'\|_{L^1(\mathbb{R})}\,\|h\|_{L^1(\mathbb{R})}\,.
\end{aligned}$$

Thus there exists a constant $0 \leq c < \infty$, such that for all $s \in \overline{\mathbb{C}_s}\backslash\{0\}$

$$|(\psi(s), w)| \quad \leq \quad \frac{c}{s}\,.$$

Furthermore we get $\int_{\mathbb{R}} \widehat{g}(t-\xi)h(t)\; dt \in L^1(\mathbb{R})$. By Lemma 10.4.1

$$\varphi(s, w) \quad := \quad (\psi(s), w)$$

is holomorphic on the lower half plane $\mathbb{C}_s$. Since also $f(s) = s^{-\frac{1}{2}}$ is holomorphic on the lower half plane $\mathbb{C}_s$, we get from Cauchy's Integral theorem (our branch of the square root is discontinuous at the positive real line)

$$\int_{-R}^0 f(s)\varphi(s, w)\; ds + \int_0^R \lim_{\varepsilon\to 0_+} f(s - i\varepsilon)\varphi(s, w)\; ds + \int_{\Gamma_R} f(s)\varphi(s, w)\; ds \quad = \quad 0\,,$$

with $R > 0$ and $\Gamma_R = \{Re^{-i\varphi} | \, 0 \le \varphi \le \pi\}$. Let us estimate the Γ_R integral:

$$\begin{aligned}
\left| \int_{\Gamma_R} f(s)\, \varphi(s,w)\, ds \right| &= \left| \int_0^\pi f(Re^{-i\varphi})\, \varphi(Re^{-i\varphi}, w)(-iR)e^{-i\varphi}\, ds \right| \\
&= R^{\frac{1}{2}} \left| \int_0^\pi \varphi(Re^{-i\varphi}, w) e^{-\frac{i}{2}\varphi}\, ds \right| \\
&\le R^{\frac{1}{2}} \int_0^\pi |\varphi(Re^{-i\varphi}, w)|\, ds \\
&\le \pi c R^{-\frac{1}{2}} .
\end{aligned}$$

Hence (in the limit $R \to \infty$)

$$\int_{-\infty}^0 f(s)\varphi(s,w)\, ds + \int_0^\infty \lim_{\varepsilon \to 0_+} f(s - i\varepsilon)\varphi(s,w)\, ds = 0 .$$

It holds for all $s \in \mathbb{R}^+$:

$$\lim_{\varepsilon \to 0_+} f(s - i\varepsilon) = -f(s) ,$$

which yields

$$\int_{-\infty}^0 f(s)\, \varphi(s,w)\, ds = \int_0^\infty f(s)\varphi(s,w)\, ds .$$

Finally we get

$$\begin{aligned}
(A^{-\frac{1}{2}} u, w) &= \tfrac{1}{2\pi} \int_{\mathbb{R}} f(s)(\psi(s), w)\, ds = \tfrac{1}{2\pi} \int_{\mathbb{R}} f(s)\varphi(s,w)\, ds \\
&= \tfrac{1}{2\pi}\, 2 \int_{-\infty}^0 f(s)\varphi(s,w)\, ds \\
&= \left(\tfrac{1}{\pi} \int_{-\infty}^0 f(s)\psi(s)\, ds,\, w \right) .
\end{aligned}$$

Let S be the set of all $w \in H$, such that the discussion and calculations of this section (up to this point) hold. If $S \subset H$ is a dense subset, then it is reasonable to write

$$A^{-\frac{1}{2}} u = \frac{1}{\pi} \int_{-\infty}^0 \frac{1}{\sqrt{s}}\, \psi(s)\, ds \tag{10.41}$$

$$= \frac{1}{\pi} \int_0^\infty \frac{1}{\sqrt{s}}\, \psi(s)\, ds . \tag{10.42}$$

Let $\kappa > 0$. With the substitution $s = -\kappa^2\xi^2$ and $\Psi(\xi) = -i\psi(-\kappa^2\xi^2)$ equation (10.41) simplifies to

$$A^{-\frac{1}{2}} u = \frac{2\kappa}{\pi} \int_0^\infty \Psi(\xi)\, d\xi .$$

If we set $s = \kappa^2\xi^2$ and $\Psi(\xi) = \psi(\kappa^2\xi^2)$, then we obtain the same integral from equation (10.42).

Remark 10.4.4. *Assume, that*

$$(A + \kappa^2\xi^2 \,\mathrm{Id})\Psi(\xi) \;=\; u \quad \text{or} \quad (A - \kappa^2\xi^2 \,\mathrm{Id})\Psi(\xi) \;=\; i\,u$$

has a unique solution for all $\xi \in \mathbb{R}_0^+$*, and assume that* S *is dense in* H*. Then* $A^{-\frac{1}{2}}u$ *is given (in a weak sense) by*

$$A^{-\frac{1}{2}}u \;=\; \frac{2\kappa}{\pi}\int_0^\infty \Psi(\xi)\,d\xi\,.$$

For $A \in \mathbb{R}_+$ *we get* $\Psi(\xi) = \frac{u}{A+\kappa^2\xi^2}$ *and*[23]

$$\frac{2\kappa}{\pi}\int_0^\infty \Psi(\xi)\,d\xi \;=\; \frac{u}{\kappa\pi}\int_{\mathbb{R}} \frac{1}{\frac{A}{\kappa^2} + \xi^2}\,d\xi \;=\; \frac{u}{A^{\frac{1}{2}}}\,.$$

Is $A < 0$ *we have to use the other formula, which yields* $\Psi(\xi) = -\frac{iu}{|A|+\kappa^2\xi^2}$*. Carrying out the integration yields the right result.*

The computations for scalar $A \in \mathbb{R}$ *indicate that the derived equations may hold for (strictly) positive or negative self–adjoint operators. However for* $A = 0$ *both possibilities do not work.*

In the end let us consider a special case, which originates from the factorization of the far–field equation as described in § 10.1. Let $A = \partial_z^2 + \kappa^2 V(z)$ on $L^2((a,b))$ with suitable boundary conditions, such that A is self–adjoint. Thus it formally holds for $f \in D(A^{-\frac{1}{2}})$:

$$\big(A^{-\frac{1}{2}}f\big)(z) \;=\; \frac{2\kappa}{\pi}\int_0^\infty \Psi(\xi, z)\,d\xi\,, \tag{10.43}$$

with

$$\Psi_{zz}(\xi,z) \,+\, \kappa^2(\xi^2 + V(z))\Psi(\xi,z) \;=\; f(z) \quad (+\,\mathrm{BC})\,. \tag{10.44}$$

The solution Ψ is oscillatory (with respect to z) and can be efficiently solved with the schemes discussed in the first part of this thesis. It remains to derive a suitable quadrature for the integral (10.43).

From the WKB analysis we can derive good approximations of Ψ for large values of ξ. This can yield an ansatz for the desired (missing) quadrature of (10.43).

In some cases also unbounded domains may be considered with this approach. For example, if the potential V is constant on the complement of a compact set, then we can impose non reflecting or transparent boundary conditions (TBC), to restrict the ODE (10.44) to a finite domain. Several approaches of artificial boundary conditions for the time dependent Schrödinger equation are discussed in the review article [3]. Even if the problems discussed in the article are time dependent, we can learn how to construct TBC for our setting.

[23] [8, p.285]: $\int \frac{1}{a^2+x^2}\,dx = \frac{1}{a}\arctan\frac{x}{a}$

10.4.2 A formal solution of the OWWE

Let A be a self–adjoint operator on the separable Hilbert space X, having a cyclic vector v. Our goal is to find a solution of the initial value problem

$$\begin{aligned} \frac{d}{dx}u(x) &= i\sqrt{A}\,u(x)\,, \qquad x \in \mathbb{R}^+\,, \\ u(x=0) &= u_0 \in H\,. \end{aligned} \tag{10.45}$$

Here we use the same branch of the complex square root as in § 10.4.1. If u_0 is an eigenvector of A with respect to the eigenvalue a, we get $u(x) = e^{i\sqrt{a}x}u_0$. Thus we expect that

$$u(x) \;=\; e^{i\sqrt{A}x}u_0$$

is a solution of the IVP (10.45) for suitable $u_0 \in H$. Hence let us define

$$f(x,y) \;:=\; e^{i\sqrt{y}\,x}\,.$$

Furthermore let $U(t)$ be the C_0 group of unitary operators with infinitesimal generator iA, i. e. the function $t \mapsto U(t)$ solves the initial value problem

$$\begin{aligned} \frac{d}{dt}U(t) &= iA\,U(t)\,, \qquad t \in \mathbb{R}\,, \\ U(t=0) &= \mathrm{Id}\,. \end{aligned}$$

By Definition 10.3.35 we have

$$u(x) \;:=\; f(x,A)u_0 \;=\; \tfrac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \gamma(x,\xi)\,U(\xi)v\;d\xi\,,$$

with $\gamma(x,\xi) := (\widehat{f(x,\cdot)g})(\xi)$. Here $g := W^{-1}u_0 \in L^2(\mathbb{R},\mu)$, where W is the unitary map from Proposition 10.3.32. The following Proposition is the main result of this section.

Proposition 10.4.5. *Let A be a self–adjoint operator on a separable Hilbert space H, having a cyclic vector v and let μ be the unique Radon measure from Proposition 10.3.32. Further let $g \in L^1(\mathbb{R}) \cap L^2(\mathbb{R},\mu)$, such that $t \mapsto \sqrt{t}g(t)$ is in $L^1(\mathbb{R})$ and let $u_0 := Wg$. Is $w \in H$, such that the map $t \mapsto (U(t)v, w) \in \mathbb{C}$ is in $L^1(\mathbb{R})$, then*

$$\begin{aligned} \frac{d}{dx}\big(u(x),w\big) &= \big(i\sqrt{A}u(x),w\big)\,, \qquad x \in \mathbb{R}^+\,, && (10.46) \\ \lim_{x\to 0}(u(x),w) &= (u_0,w)\,. && (10.47) \end{aligned}$$

Proof. Since $g \in L^1(\mathbb{R})$

$$\gamma(x,\xi) \;=\; \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{i\sqrt{t}x}\,g(t)\,e^{-i\xi t}\;d\xi$$

is well defined. Furthermore the function $|g|$ is an integrable upper bound for the integrand for all $x \in \mathbb{R}^+$. Thus, by the dominated convergence theorem

$$\lim_{x\to 0}\gamma(x,\xi) \;=\; \widehat{g}(\xi)\,.$$

Let $w \in H$, such that $h(t) := \frac{1}{\sqrt{2\pi}}(U(t)v, w)$ is in $L^1(\mathbb{R})$. Then, for all $x \in \mathbb{R}^+$, $\|g\|_{L^1(\mathbb{R})}|h|$ is an integrable upper bound for $\xi \mapsto \gamma(x,\xi)h(\xi)$ and hence

$$\begin{aligned}\lim_{x\to 0}(u(x), w) &= \lim_{x\to 0}\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\gamma(x,\xi)\,h(\xi)\,d\xi = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\widehat{g}(\xi)\,h(\xi)\,d\xi \\ &= \left(\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\widehat{g}(\xi)\,U(\xi)v\,d\xi,\, w\right) = (Wg, w) = (u_0, w)\,.\end{aligned}$$

Hence we have proven (10.47).

Let $\psi(x,\xi,t) := f(x,t)g(t)e^{-i\xi t}$. Since $g \in L^1(\mathbb{R})$ it holds $\psi(x,\xi,\cdot) \in L^1(\mathbb{R})$ for all $(x,\xi) \in \mathbb{R}^+\times\mathbb{R}$. Furthermore ψ is differentiable with respect to x for all $(x,\xi,t) \in \mathbb{R}^+\times\mathbb{R}\times\mathbb{R}$ and it holds $\psi_x(x,\xi,t) = i\sqrt{t}\psi(x,\xi,t)$. Since

$$|\psi_x(x,\xi,t)| = |\sqrt{t}|\,|\psi(x,\xi,t)| \le |\sqrt{t}|\,|g(t)| \in L^1(\mathbb{R})\,,$$

$t \mapsto |\sqrt{t}||g(t)|$ is an integrable upper bound for $\partial_x\psi(x,\xi,\cdot)$ for all $(x,\xi) \in \mathbb{R}^+\times\mathbb{R}$. Thus, by [16, Satz 5.7, p.146] we can interchange integration an differentiation, which yields

$$\frac{\partial}{\partial x}\gamma(x,\xi) = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} i\sqrt{t}e^{i\sqrt{t}x}g(t)\,e^{-i\xi t}\,dt\,.$$

Now let $\varphi(x,\xi) := \gamma(x,\xi)\,h(\xi)$. Since $h \in L^1(\mathbb{R})$ and $|\gamma(x,\xi)| \le \|g\|_{L^1(\mathbb{R})}$ $\varphi(x,\cdot) \in L^1(\mathbb{R})$ for all $x \in \mathbb{R}^+$. Furthermore $\partial_x\varphi(x,\xi) = h(\xi)\partial_x\gamma(x,\xi)$ exists for all $(x,\xi) \in \mathbb{R}^+\times\mathbb{R}$. Additionally it holds

$$|\partial_x\varphi(x,\xi)| = |h(\xi)||\partial_x\gamma(x,\xi)| \le |h(\xi)|\,\|\sqrt{\cdot}g(\cdot)\|_{L^1(\mathbb{R})}\,.$$

Again we can apply [16, Satz 5.7, p.146] and hence

$$\begin{aligned}\frac{d}{dx}(u(x), w) &= \frac{d}{dx}\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\gamma(x,\xi)\,h(\xi)\,d\xi \\ &= \frac{1}{2\pi}\int_{\mathbb{R}}\int_{\mathbb{R}} i\sqrt{t}e^{i\sqrt{t}x}g(t)\,e^{-i\xi t}\,dt\,h(\xi)\,d\xi \\ &= \frac{i}{\sqrt{2\pi}}\int_{\mathbb{R}}\widehat{\big(\sigma f(x,\cdot)g\big)}(\xi)\,h(\xi)\,d\xi\,,\end{aligned}$$

where $\sigma(t) = \sqrt{t}$ is the branch of the complex square root as defined in § 10.4.1. By Definition 10.3.35 it holds $W^{-1}u(x) = f(x,\cdot)g$. Hence

$$\begin{aligned}\frac{d}{dx}(u(x), w) &= \frac{i}{\sqrt{2\pi}}\int_{\mathbb{R}}\widehat{\big(\sigma W^{-1}u(x)\big)}(\xi)\,h(\xi)\,d\xi \\ &= \left(i\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\widehat{\big(\sigma W^{-1}u(x)\big)}(\xi)\,U(\xi)v\,d\xi,\, w\right) \\ &= \big(iW(\sigma W^{-1}u(x)), w\big) \\ &= \big(i\sqrt{A}u(x), w\big)\,.\end{aligned}$$

□

Remark 10.4.6. *Let the assumption of Proposition 10.4.5 hold. If the subspace* $\operatorname{span}\{w \in H |\, \big(U(\cdot)v, w\big) \in L^1(\mathbb{R})\}$ *is dense in* H*, then* u *is a weak solution of the initial value problem* (10.45)*.*

In the end of this section we prove the following estimate: $\|u(x)\| \leq \|u_0\|$. Therefore we frequently use that W is an unitary map from $L^2(\mathbb{R},\mu) \to H$.

$$\begin{aligned} \|u(x)\| &= \|f(x,A)u_0\| = \|W^{-1}f(x,A)u_0\|_{L^2(\mathbb{R},\mu)} \\ &= \|f(x,\cdot)g\|_{L^2(\mathbb{R},\mu)} \leq \|f(x,\cdot)\|_{L^\infty(\mathbb{R},\mu)}\|g\|_{L^2(\mathbb{R},\mu)} \\ &= \|Wg\| = \|u_0\|\,. \end{aligned}$$

This yields

Remark 10.4.7. *The evolution operator generated by $i\sqrt{A}$ is a semi group of contractions.*

The Fourier transform of $e^{i\sqrt{\tau}x}$

To compute the Fourier transform of $e^{i\sqrt{\tau}x}$, we define for $b \in \mathbb{C}$ and $\tau \in \mathbb{R}$:

$$e^{-b\sqrt{\tau}} := \tfrac{1+\mathrm{sgn}(\tau)}{2}e^{-b\sqrt{|\tau|}} + \tfrac{1-\mathrm{sgn}(\tau)}{2}e^{-ib\sqrt{|\tau|}}\,.$$

We use the same branch of the complex square root as in § 10.4.1. The following table is from [50]:

$f(\tau)$	$\int_{\mathbb{R}} f(\tau)\, e^{-i2\pi\xi\tau}\, d\tau$
$e^{-a\sqrt{\lvert\tau\rvert}}$	$\frac{a}{2\pi\xi\sqrt{\lvert\xi\rvert}}\big[\ \big(\frac{1}{2}-S(\rho)\big)\sin(\frac{a^2}{8\pi\xi}) + \eta\big(\frac{1}{2}-C(\rho)\big)\cos(\frac{a^2}{8\pi\xi})\big]$
$\mathrm{sgn}(\tau)\, e^{-a\sqrt{\lvert\tau\rvert}}$	$\frac{-ia}{2\pi\xi\sqrt{\lvert\xi\rvert}}\big[\eta\big(\frac{1}{2}-C(\rho)\big)\sin(\frac{a^2}{8\pi\xi}) - \ \big(\frac{1}{2}-S(\rho)\big)\cos(\frac{a^2}{8\pi\xi})\big] - \frac{i}{\pi\xi}\,,$

with $\eta := \mathrm{sgn}(\xi)$, $\rho := \frac{a}{2\pi\sqrt{|\xi|}}$, and

$$S(\rho) := \int_0^\rho \sin(\tfrac{\pi}{2}u^2)\, du\,, \qquad C(\rho) := \int_0^\rho \cos(\tfrac{\pi}{2}u^2)\, du\,.$$

One can extend the *Fresnel functions* S, C to the entire complex plane by

$$S(\zeta) = \int_0^1 \sin(\tfrac{\pi}{2}t^2\zeta^2)\zeta\, dt\,, \qquad C(\zeta) = \int_0^1 \cos(\tfrac{\pi}{2}t^2\zeta^2)\zeta\, dt\,.$$

This yields

$$S(i\zeta) = -iS(\zeta) \quad \text{and} \quad C(i\zeta) = iC(\zeta)\,.$$

Hence

$$
\begin{aligned}
& 2\int_{\mathbb{R}} e^{-b\sqrt{\tau}} e^{-i2\pi\xi\tau}\, d\tau \\
&= \tfrac{b}{2\pi\xi\sqrt{|\xi|}}\big[\ \big(\tfrac12 - S(\rho)\big)\sin(\tfrac{b^2}{8\pi\xi}) + \eta\big(\tfrac12 - C(\rho)\big)\cos(\tfrac{b^2}{8\pi\xi})\big] \\
&+ \tfrac{-ib}{2\pi\xi\sqrt{|\xi|}}\big[\eta\big(\tfrac12 - C(\rho)\big)\sin(\tfrac{b^2}{8\pi\xi}) - \ \big(\tfrac12 - S(\rho)\big)\cos(\tfrac{b^2}{8\pi\xi})\big] - \tfrac{i}{\pi\xi} \\
&+ \tfrac{ib}{2\pi\xi\sqrt{|\xi|}}\big[\ \big(\tfrac12 - S(i\rho)\big)\sin(\tfrac{-b^2}{8\pi\xi}) + \eta\big(\tfrac12 - C(i\rho)\big)\cos(\tfrac{-b^2}{8\pi\xi})\big] \\
&- \tfrac{b}{2\pi\xi\sqrt{|\xi|}}\big[\eta\big(\tfrac12 - C(i\rho)\big)\sin(\tfrac{-b^2}{8\pi\xi}) - \ \big(\tfrac12 - S(i\rho)\big)\cos(\tfrac{-b^2}{8\pi\xi})\big] + \tfrac{i}{\pi\xi} \\
&= \tfrac{(1-i)b}{2\pi\xi\sqrt{|\xi|}}\ \tfrac{1+\mathrm{sgn}(\xi)}{2}\ \sin(\tfrac{b^2}{8\pi\xi}) + \tfrac{(1+i)b}{2\pi\xi\sqrt{|\xi|}}\ \tfrac{1+\mathrm{sgn}(\xi)}{2}\ \cos(\tfrac{b^2}{8\pi\xi}) \\
&= \tfrac{b}{2\pi\xi\sqrt{|\xi|}}\ H(\xi)\ \big((1-i)\sin(\tfrac{b^2}{8\pi\xi}) + (1+i)\cos(\tfrac{b^2}{8\pi\xi})\big) \\
&= \tfrac{(1+i)b}{2\pi\xi\sqrt{|\xi|}}\ H(\xi)\ e^{-i\frac{b^2}{8\pi\xi}}\ .
\end{aligned}
$$

It follows from the previous calculation (with $t = 2\pi\xi$, $-b = ix$):

$$
\begin{aligned}
\tfrac{2}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{i\sqrt{\tau}x}\, e^{it\tau}\, d\tau &= \tfrac{1}{\sqrt{2\pi}}\tfrac{(1+i)(-ix)}{2\pi\frac{t}{2\pi}\sqrt{\left|\frac{t}{2\pi}\right|}}\ H(\tfrac{t}{2\pi})\ e^{-i\frac{(-ix)^2}{8\pi\frac{t}{2\pi}}} \\
&= \tfrac{(1-i)x}{t\sqrt{|t|}}\ H(t)\ e^{i\frac{x^2}{4t}}\ .
\end{aligned}
$$

Hence, if we use Definition 10.3.34, than the (formal) solution of the OWWE (10.45) is given by

$$
u(x) \ = \ \sqrt{\tfrac{1}{i\pi}}\,\tfrac{x}{2}\int_0^\infty \frac{1}{t^{\frac32}}\,\psi(t)\,e^{i\frac{x^2}{4t}}\,dt\,, \tag{10.48}
$$

with

$$
\psi_t \ = \ i\,A\,\psi\,, \qquad \psi(t=0) \ = \ u_0\,. \tag{10.49}
$$

Despite that (10.48) is derived by formal calculations, the integral on the right hand side is well defined for a certain class of scalar functions ψ.

Proposition 10.4.8. *Let $\psi \in L^\infty(\mathbb{R}^+)$, such that ψ is continuously differentiable in the semi open interval $[0,\varepsilon)$ for some $\varepsilon \in \mathbb{R}^+$. Then*

$$
u(x) \ := \ \sqrt{\tfrac{1}{i\pi}}\,\tfrac{x}{2}\int_0^\infty \frac{1}{t^{\frac32}}\,\psi(t)\,e^{i\frac{x^2}{4t}}\,dt\,, \tag{10.50}
$$

exists for all $x \in \mathbb{R}^+$. Furthermore it holds $\lim_{x\to 0} u(x) = \psi(0)$.

Proof. For $z \in \mathbb{C}$ let

$$
\Phi(z) \ := \ \frac{2}{\sqrt{\pi}}\int_0^z e^{-\zeta^2}\,d\zeta\,,
$$

where we integrate in the complex plane along an arbitrary path which connects 0 and z. The entire function Φ is called *error function* (cf. [51, p.36]). For $|\arg z| < \frac{\pi}{2}$ we find the asymptotic representation (see [51, p.39])

$$1 - \Phi(z) \quad \approx \quad \frac{e^{-z^2}}{\sqrt{\pi} z} \quad \text{as } |z| \to \infty.$$

Let $R \in \mathbb{R}^+$ and let $z := \frac{1-i}{\sqrt{2}} R = -\sqrt{-i} R$. Hence

$$\frac{e^{-z^2}}{\sqrt{\pi} z} \;=\; -\frac{e^{iR^2}}{\sqrt{\pi}\sqrt{-i} R} \to 0 \quad \text{as } R \to \infty \quad \Rightarrow \quad \lim_{R\to\infty} \Phi(-\sqrt{-i} R) \;=\; 1\,.$$

Furthermore it holds $\Phi(z) = -\Phi(-z)$ and thus

$$1 \;=\; \lim_{R\to\infty} \Phi(-\sqrt{-i} R) \;=\; \lim_{R\to\infty} -\Phi(\sqrt{-i} R)\,.$$

Let $\gamma(t) := \frac{\sqrt{-i}}{2} \frac{x}{t^{\frac{1}{2}}}$ be a parametrization of the path from 0 to $\sqrt{-i} R$. Hence

$$\begin{aligned} -\Phi(\sqrt{-i} R) \;&=\; -\frac{2}{\sqrt{\pi}} \int_{\infty}^{(\frac{x}{2R})^2} \big(-\tfrac{1}{2}\big) \frac{\sqrt{-i}}{2} \frac{x}{t^{\frac{3}{2}}} e^{i\frac{x^2}{4t}}\, dt \\ &=\; \sqrt{\tfrac{1}{i\pi}}\, \tfrac{x}{2} \int_{(\frac{x}{2R})^2}^{\infty} \frac{1}{t^{\frac{3}{2}}}\, e^{i\frac{x^2}{4t}}\, dt\,, \end{aligned}$$

which yields

$$\sqrt{\tfrac{1}{i\pi}}\, \tfrac{x}{2} \int_0^{\infty} \frac{1}{t^{\frac{3}{2}}}\, e^{i\frac{x^2}{4t}}\, dt \;=\; \lim_{R\to\infty} -\Phi(\sqrt{-i} R) \;=\; 1\,.$$

Therefore it holds

$$u(x) \;=\; \psi(0) + \sqrt{\tfrac{1}{i\pi}}\, \tfrac{x}{2} \int_0^{\infty} \frac{\psi(t) - \psi(0)}{t^{\frac{3}{2}}}\, e^{i\frac{x^2}{4t}}\, dt\,. \tag{10.51}$$

Let $g(t) := \frac{\psi(t)-\psi(0)}{t^{\frac{3}{2}}}$. Since ψ is continuously differentiable in an open neighborhood of zero, it holds for t small enough:

$$g(t) \;=\; \frac{1}{t^{\frac{3}{2}}} \Big(\int_0^t \psi'(s)\, ds \Big)\,.$$

It follows $\sqrt{t} g(t) \to \psi'(0)$ as $t \to 0$ and hence g has an integrable singularity at $t = 0$. Since ψ is bounded, there exists a constant $c > 0$ and a $t_0 > 0$, such that $|g(t)| < c\, t^{-\frac{3}{2}}$ for almost all $t > t_0$. Thus $g \in L^1(\mathbb{R}^+)$ and hence the integral in (10.50) exists. Furthermore $|g|$ is an integrable upper bound for the absolute value of the integrand in (10.51) and hence $\lim_{x\to 0} u(x) = \psi(0)$. □

Remark 10.4.9. *With Maple14 we also find that*

$$\int_0^{\infty} \frac{1}{t^{\frac{3}{2}}}\, e^{-t}\, e^{i\frac{x}{4t}}\, dt \;=\; 2\sqrt{\pi}\, \frac{e^{-\sqrt{-ix}}}{\sqrt{-ix}}\,.$$

Hence we can use a linear combination of the functions 1, e^{-t} *to manipulate* ψ *at* $t = 0$.

10.4.3 DeSanto's Transformation

The key tool of constructing the PDO symbol of the SRHO from § 10.1, as suggested in [21], is a integral transformation that John A. DeSanto established in [12]. It connects the solution of the Helmholtz equation

$$\begin{aligned} \frac{\partial^2}{\partial x^2}\, u(x,z) + \Big(\frac{\partial^2}{\partial z^2} + \kappa^2 V(z)\Big) u(x,z) &= 0\,, \\ u(x=0) &= u_0 \end{aligned} \tag{10.52}$$

and the parabolic equation for sound propagation

$$\begin{aligned} 2i\kappa \frac{\partial}{\partial s} p(s,z) + \frac{\partial^2}{\partial z^2} p(s,z) + \kappa^2 \big(V(z)-1\big)\, p(s,z) &= 0\,, \\ p(s=0) &= u_0 \end{aligned} \tag{10.53}$$

by the integral transformation

$$u(x,z) = \sqrt{\tfrac{\kappa}{2\pi i}}\, x \int_0^\infty \frac{1}{s^{\frac{3}{2}}}\, p(s,z)\, e^{i\kappa \frac{x^2+s^2}{2s}}\, ds\,. \tag{10.54}$$

In the sequel we shall briefly discuss the connection of this transformation approach with our results derived in the previous sections. Therefor let

$$A := \frac{\partial^2}{\partial z^2} + \kappa^2 V(z)$$

acting on a suitable Hilbert space H, having a cyclic vector v. We assume that A, with the not specified boundary conditions for the Helmholtz equation (10.52), is self–adjoint on H. Now the Helmholtz equation reads

$$\frac{\partial^2}{\partial x^2} u(x) + A u(x) = 0\,, \qquad u(x=0) = u_0\,.$$

Formally we can factorize the differential operator:

$$\Big(\frac{\partial}{\partial x} + i\sqrt{A}\Big)\Big(\frac{\partial}{\partial x} - i\sqrt{A}\Big) u(x) = 0\,.$$

Hence a solution u_* of the one way wave equation

$$\frac{\partial}{\partial x} u_*(x) = i\sqrt{A} u_*(x)\,, \qquad u_*(x=0) = u_0\,,$$

is also a solution of the Helmholtz equation. From (10.48) and (10.49) we get

$$u_*(x,z) = \sqrt{\tfrac{1}{i\pi}}\, \tfrac{x}{2} \int_0^\infty \frac{1}{t^{\frac{3}{2}}}\, \psi(t,z)\, e^{i\frac{x^2}{4t}}\, dt\,,$$

with

$$\begin{aligned} \psi_t &= i\,A\,\psi = i\Big(\frac{\partial^2}{\partial z^2} + \kappa^2 V(z)\Big)\psi\,, \\ \psi(t=0) &= u_0\,. \end{aligned} \tag{10.55}$$

In (10.54) we substitute $s = 2\kappa t$, which yields

$$\begin{aligned} u(x,z) &= \sqrt{\tfrac{\kappa}{2\pi i}}\, x \int_0^\infty \frac{2k}{(2kt)^{\frac{3}{2}}}\, p(2kt,z)\, e^{ik^2 t}\, e^{i\frac{x^2}{4t}}\, dt \\ &= \sqrt{\tfrac{1}{\pi i}}\, \tfrac{x}{2} \int_0^\infty \frac{1}{t^{\frac{3}{2}}}\, \rho(t,z)\, e^{i\frac{x^2}{4t}}\, dt\,, \end{aligned}$$

with $\rho := p(2kt,z)e^{ik^2 t}$. Differentiating ρ with respect to t and using (10.53) yields

$$\begin{aligned} \frac{\partial}{\partial t}\rho &= e^{i\kappa^2 t}\, 2\kappa \frac{\partial}{\partial s} p + i\kappa^2 \rho \\ &= i\frac{\partial^2}{\partial z^2} p(2\kappa t, z) e^{i\kappa^2 t} + i\kappa^2 V\rho - i\kappa^2\rho + i\kappa^2\rho \\ &= i\left(\frac{\partial^2}{\partial z^2} + \kappa^2 V\right)\rho\,. \end{aligned}$$

Additionally it holds $\rho(t=0) = u_0$. Hence ρ is a solution of the initial value problem (10.55) and thus $u = u_*$. Therefore is the solution u of the Helmholtz equation, constructed by the approach of deSanto, is also a solution of the one way way equation solution. This connection is not mentioned in [12].

For our approach the operator A only has to be self–adjoint. Hence we can interpret our ansatz as a generalization of transformation (10.54).

10.5 Summary and conclusions

We introduce the one way wave equation (OWWE) and present one example which gives an idea on which level of model reduction the OWWE may appear. Afterwards, in § 10.2, we discuss some difficulties occurring when discretizing the square root Helmholtz operator (SRHO) (treated as pseudo differential operator). Due to the stated problems we decide to consider the SRHO in the framework of functional calculus of self–adjoint operators, i. e. we define it as as function of a differential operator. In § 10.3 we present a non standard approach from [71] to derive a spectral theorem. It seems to be more suitable for numerical computations, than the ansatz by Banach algebras. This part is a revised lecture note from the author. During the course we prove a version of Riesz' representation theorem, which is a special case and a generalization of the results from [16]. Our derived version is used for the proof of the spectral theorem. Furthermore we prove, that the space of test functions $\mathcal{D}(\Omega)$ is dense in $L^p(\Omega,\mu)$, where μ is a Radon measure on the open set $\Omega \subset \mathbb{R}^n$. Since also this result is needed for the proof of the spectral theorem, and since we have found it only for the case where μ is the Lebesgue measure, we decided to carry out the computations.

With the functional calculus we (not completely rigorous) derive a formula to compute the image of a vector mapped by the inverse square root of a self–adjoint operator. It is based on the solution of a general (linear) Schrödinger equation. In the final version the derived formula is quite simple (see Remark 10.4.4). One just has to integrate the solution of a linear system depending on a parameter. For the special case that the operator is a linear second order

differential operator, one may use the numerical methods presented in the first part of this thesis to efficiently solve the linear equations, for large values of the parameter.

Furthermore we formally prove a formula to compute the solution of the one way wave equation. It turns out that the computational effort to compute the solution of the one way wave equation at a certain point is comparable to just applying the inverse square root operator to the initial condition. We are also able to derive deSanto's transformation from our theoretical approach. This transformation is connects the solution of a Schrödinger type equation with the solution of the Helmholtz equation. Due to our discussion, we find out that deSanto's solution is also a solution of the one way wave equation.

Bibliography

[1] Robert A. Adams. *Sobolov Spaces*. Acad. Press, Boston, 1992.

[2] Herbert Amann. *Gewöhnliche Differentialgleichungen*. de Gruyter, Berlin, 2. edition, 1995.

[3] Xavier Antoine, Anton Arnold, Christophe Besse, Matthias Ehrhardt, and Achim Schädle. A review of transparent and artificial boundary conditions techniques for linear and nonlinear Schrödinger equations. *Commun. Comput. Phys.*, 4(4):729–796, 2008.

[4] Anton Arnold, Naoufel Ben Abdallah, and Claudia Negulescu. WKB–based schemes for the Schrödinger equation in the semi–classical limit. *SIAM J. Numer. Anal.*, 49(4):1436–1460.

[5] Naoufel Ben Abdallah and Jihene Kefi-Ferhane. Mathematical analysis of the two-band Schrödinger model. *Math. Methods Appl. Sci.*, 31(10):1131–1151, 2008.

[6] Carl M. Bender. *Advanced mathematical methods for scientists and engineers*. McGraw-Hill, New York, 1978.

[7] Siegfried Bosch. *Lineare Algebra*. Springer, Berlin, 4. edition, 2008.

[8] Il'ja N. Bronštejn. *Taschenbuch der Mathematik*. Teubner, BSB, Leipzig, 10. edition, 1969.

[9] Robert Dautray. *Evolution problems*. Number 5. Springer, Berlin, 1992.

[10] Robert Dautray. *Functional and variational methods*. Springer, Berlin, 2000.

[11] Philip I. Davies and Nicholas J. Higham. A Schur-Parlett algorithm for computing matrix functions. *SIAM J. Matrix Anal. Appl.*, 25(2):464–485 (electronic), 2003.

[12] John A. DeSanto. Relation between the solutions of the Helmholtz and parabolic equations for sound propagation. *J. Acoust. Soc. Amer.*, 62(2):295–297, 1977.

[13] Harry Dym. *Lineare algebra in action*. American Math. Soc., Providence, RI, 2007.

[14] Matthias Ehrhardt and Anton Arnold. Discrete transparent boundary conditions for wide angle parabolic equations in underwater acoustics. *J. Comp. Phys*, 145:611–638, 1998.

[15] Matthias Ehrhardt and Andrea Zisowsky. Discrete non-local boundary conditions for split-step Padé approximations of the one-way Helmholtz equation. *J. Comput. Appl. Math.*, 200(2):471–490, 2007.

[16] Jürgen Elstrodt. *Maß- und Integrationstheorie*. Springer, Berlin, 1999.

[17] Michail V. Fedorjuk. *Asymptotic analysis*. Springer, Berlin, 1993.

[18] Louis Napoleon George Filon. On a quadrature formula for trigonometric integrals. *Proc. Roy. Soc. Edin.*, 49:38–47, 1928.

[19] Gerd Fischer. *Lineare Algebra*. Vieweg, Braunschweig, 11. edition, 1997.

[20] Louis Fishman. Exact and operator rational approximate solutions of the Helmholtz, Weyl composition equation in underwater acoustics—the quadratic profile. *J. Math. Phys.*, 33(5):1887–1914, 1992.

[21] Louis Fishman, A. K. Gautesen, and Zhiming Sun. Uniform high-frequency approximations of the square root Helmholtz operator symbol. *Wave Motion*, 26(2):127–161, 1997.

[22] Louis Fishman and Stephen C. Wales. Phase space methods and path integration: the analysis and computation of scalar wave equations. In *Proceedings of the 2nd international conference on computational and applied mathematics (Leuven, 1986)*, volume 20, pages 219–238, 1987.

[23] Otto Forster. *Analysis 1*. Vieweg, Braunschweig, 5. edition, 1999.

[24] George V. Frisk. *Ocean and seabed acoustics*. Prentice-Hall, Englewood Cliffs, NJ, 1994.

[25] Jens Geier and Anton Arnold. WKB–based schemes for two–band Schrödinger equations in the highly oscillatory regime. In *Nanosystems: Physics, Chemistry, Mathematics*, volume 2, pages 7–28, 2011.

[26] George A. Hagedorn. Proof of the Landau-Zener formula in an adiabatic limit with small eigenvalue gaps. *Comm. Math. Phys.*, 136(3):433–449, 1991.

[27] Ernst Hairer. *Geometric numerical integration*. Springer, Berlin, 2. edition, 2006.

[28] Martin Hanke-Bourgeois. *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. Teubner, Wiesbaden, 2. edition, 2006.

[29] Martin Hermann. *Numerische Mathematik*. Oldenbourg Wissenschaftsverlag, Munich, 2. edition, 2006.

[30] Nicholas J. Higham. The scaling and squaring method for the matrix exponential revisited. *SIAM Rev.*, 51(4):747–764, 2009.

[31] Friedrich Hirzebruch. *Einführung in die Funktionalanalysis*. BI Hochschultaschenbücher-Verl., Mannheim; Wien, 1971.

[32] Mark H. Holmes. *Introduction to Perturbation Methods*. Springer, New York, 1995.

[33] Roger A. Horn. *Matrix Analysis*. Cambridge Univ. Press, Cambridge, 1985.

[34] Roger A. Horn. *Topics in Matrix Analysis*. Cambridge Univ. Press, Cambridge, 1991.

[35] Daan Huybrechs and Sheehan Olver. Highly oscillatory quadrature. In *Highly oscillatory problems*, volume 366 of *London Math. Soc. Lecture Note Ser.*, pages 25–50. Cambridge Univ. Press, Cambridge, 2009.

[36] Arieh Iserles. On the numerical quadrature of highly-oscillating integrals. I. Fourier transforms. *IMA J. Numer. Anal.*, 24(3):365–391, 2004.

[37] Arieh Iserles and Syvert P. Nørsett. On quadrature methods for highly oscillatory integrals and their implementation. *BIT*, 44(4):755–772, 2004.

[38] Arieh Iserles and Syvert P. Nørsett. Efficient quadrature of highly oscillatory integrals using derivatives. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 461(2057):1383–1399, 2005.

[39] Tobias Jahnke. Long-time-step integrators for almost-adiabatic quantum dynamics. *SIAM J. Sci. Comput.*, 25(6):2145–2164 (electronic), 2004.

[40] Tobias Jahnke and Christian Lubich. Numerical integrators for quantum dynamics close to the adiabatic limit. *Numerische Mathematik*, 94:289–314, 2003. 10.1007/s00211-002-0421-1.

[41] Finn B. Jensen. *Computational ocean acoustics*. AIP Press, New York, NY, 1994.

[42] Evan O. Kane. Energy band structure in p–type germanium and silicon. *Journal of Physics and Chemistry of Solids*, 1(1-2):82–99, 1956.

[43] Evan O. Kane. Energy band theory. In W. Paul, editor, *Handbook on Semiconductors*, volume 1, pages 193–217. North-Holland, Amsterdam, 1982.

[44] K. Kansy. Elementare Fehlerdarstellung für Ableitungen bei der Hermite-Interpolation. *Numer. Math.*, 21:350–354, 1973/74.

[45] Jihene Kefi. *Analyse mathématique et numérique de modèles quantiques pour les semiconducteurs*. PhD thesis, Université Toulouse III - Paul Sabatier, 2003.

[46] Dirk Klindworth. *Discrete Transparent Boundary Conditions for Multiband Effective Mass Approximations*. PhD thesis, Technische Universität Berlin, 2009.

[47] H. A. Kramers. Wellenmechanik und halbzahlige Quantisierung. *Zeitschrift für Physik A Hadrons and Nuclei*, 39:828–840, 1926. 10.1007/BF01451751.

[48] Lev D. Landau and Eewgeni M. Lifschitz. *Quantenmechanik*. Akademie-Verlag, Berlin, 1985.

[49] Serge Lang. *Real and Functional Analysis*. Springer, New York, 3. edition, 1993.

[50] Jean Lavoine. *Transformation de Fourier des pseudo–fonctions*. CNRS, Paris, 1963.

[51] Nikolaj N. Lebedev. *Spezielle Funktionen und ihre Anwendung*. BI-Wiss.-Verl., Mannheim; Wien, 1973.

[52] David Levin. Analysis of a collocation method for integrating rapidly oscillatory functions. *J. Comput. Appl. Math.*, 78(1):131–138, 1997.

[53] Elliott H. Lieb. *Analysis*. American Math. Soc., Providence, RI, 2001.

[54] Katina Lorenz, Tobias Jahnke, and Christian Lubich. Adiabatic integrators for highly oscillatory second-order linear differential equations with time-varying eigendecomposition. *BIT*, 45(1):91–115, 2005.

[55] R. Magno, A. S. Bracker, B. R. Bennett, B. Z. Nosho, and L. J. Whitman. Barrier roughness effects in resonant interband tunnel diodes. *Journal of Applied Physics*, 90(12):6177–6181, dec 2001.

[56] Peter A. Markowich and Frederic Poupaud. The pseudo-differential approach to finite differences revisited. *Calcolo*, 36(3):161–186, 1999.

[57] Claudia Negulescu. Numerical analysis of a multiscale finite element scheme for the resolution of the stationary Schrödinger equation. *Numer. Math.*, 108(4):625–652, 2008.

[58] Frank W. J. Olver. *Introduction to asymptotics and special functions*. Acad. Press, New York, 1974.

[59] Sheehan Olver. Gmres for oscillatory matrix-valued differential equations. In Jan S. Hesthaven and Einar M. Rønquist, editors, *Spectral and High Order Methods for Partial Differential Equations*, volume 76 of *Lecture Notes in Computational Science and Engineering*, pages 267–274. Springer Berlin Heidelberg.

[60] Sheehan Olver. Moment-free numerical integration of highly oscillatory functions. *IMA J. Numer. Anal.*, 26(2):213–227, 2006.

[61] Sheehan Olver. Moment-free numerical approximation of highly oscillatory integrals with stationary points. *European J. Appl. Math.*, 18(4):435–447, 2007.

[62] Amnon Pazy. *Semigroups of linear operators and applications to partial differential equations*. Springer, New York, NY, 1983.

[63] Hans-Jürgen Reinhardt. *Numerik gewöhnlicher Differentialgleichungen*. de Gruyter, Berlin, 2. edition, 2008.

[64] Reinhold Remmert. *Funktionentheorie 1*. Springer, Berlin, 4. edition, 1995.

[65] Walter Rudin. *Functional analysis*. McGraw-Hill, New York, NY, 1991.

[66] Walter Rudin. *Real and complex analysis*. McGraw-Hill, New York, NY, 2005.

[67] Robert Schaback. *Numerische Mathematik*. Springer, Berlin, 5. edition, 2005.

[68] Josef Stoer. *Introduction to Numerical Analysis*. Springer, New York, NY, 2. edition, 1993.

[69] Fred D. Tappert. The parabolic approximation method. In *Wave propagation and underwater acoustics (Workshop, Mystic, Conn., 1974)*, pages 224–287. Lecture Notes in Phys., Vol. 70. Springer, Berlin, 1977.

[70] Michael Eugene Taylor. *Basic theory*. Number 1. Springer, New York, NY, 1996.

[71] Michael Eugene Taylor. *Qualitative studies of linear equations*. Number 2. Springer, New York, NY, 1996.

[72] Dirk Werner. *Funktionalanalysis*. Springer, Berlin, 5. edition, 2005.

[73] Curt Wittig. The Landau–Zener Formula. *J. Phys. Chem. B*, 109(17):8428–8430, 2005.

[74] Kôsaku Yoshida. *Functional analysis*. Springer, Berlin, 1974.

[75] Clarence Zener. Non-adiabatic crossing of energy levels. *Proc. R. Soc. Lond. A*, 137:696–702, 1932.

Curriculum vitae

Personal Data

Name	Jens Geier
Date of birth	04.04.1979
Place of birth	Höxter, Germany
Nationality	german
Marital status	married, one child

Education

1995 - 1998	Städtisches Anne–Frank Gymnasium Werne School–leaving qualification: Abitur (general qualification for university entrance)
1989 - 1995	Realschule der Gemeinde Ascheberg
1985 - 1989	Marien-Grundschule Herbern

Military service

1998 - 1999	corpsman, 5./gem. Lazarettregiment 11, Dülmen

Academic studies

since April 2006	PhD student of technical mathematics, Vienna University of Technology
August 2004	diploma examination in mathematics thesis: The $\overline{\partial}$–Poincaré lemma on foliated spaces
1999 - 2004	studies of mathematics and physics, Westfälische Wilhelms–Universität Münster

Work Experience

since Sept. 2005	research associate, Institute for Analysis and Scientific Computing, Vienna University of Technology
2004 - 2005	research associate, Institute for Computational and Applied Mathematics, Westfälische Wilhelms–Universität Münster