Diese Dissertation haben begutachtet:

....

DISSERTATION - DOCTORAL THESIS

On Finite Time Singularities in Unsteady Marginally Separated Flows

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines Doktors der Naturwissenschaften

unter der Leitung von

Ao. Prof. Dr. Stefan Braun

E322 Institut für Strömungsmechanik und Wärmeübertragung

eingereicht an der Technischen Universität Wien
Fakultät für Mathematik und Geoinformation

von

Mario J. Aigner
Mat. Nr.: 0201310

Wien, Oktober 2012

Kurzfassung

Diese Arbeit beschäftigt sich mit dem Phänomen der marginalen, d. h. lokalen, Ablösung von laminaren Grenzschichtströmungen entlang glatter Oberflächen. Da solche Zustände, in Form von Ablöseblasen, im Allgemeinen als instabil bezüglich gewisser Störungen gelten und daher den Prozess der *Transition zur Turbulenz* auslösen können, liegt das Hauptaugenmerk hier auf instationären, (lokal) dreidimensionalen Strömungen.

Durch die Assoziation der Strömungsablösung mit negativen Werten der Wandschubspannung ist es zielführend die Zeitentwicklung dieser genauer zu Untersuchen. Das dafür notwendige Cauchy Problem wird mittels der Methode der angepassten asymptotischen Entwicklungen aus den Navier-Stokes-Gleichungen bei unendlich hohen Reynoldszahlen hergeleitet. Dabei ergibt sich eine Integro-differentialgleichung die hier erstmals durch einen neuen, eleganteren Zugang über die Fredholmsche Alternative ermittelt wird.

Durch verwenden von Operatorsymbolen und der Dispersionsrelation lässt sich zeigen, dass das Cauchy Problem *nicht sachgemäß gestellt* ist. Für die deshalb notwendige Regularisierung werden zwei unterschiedliche Methoden gewählt – a) Diskretisierung des Problems und b) Verwenden von Regularisierungsoperatoren. Durch numerische Berechnungen kann hier gezeigt werden, dass, abgesehen von Diskretisierungsfehlern, beide Verfahren dieselben (regularisierten) Lösungen liefern. Ferner konvergieren solche Lösungen zu Lösungen des ursprünglichen Problems im Grenzwert verschwindender Regularisierung.

Um die regularisierten Lösungen in das physikalische Konzept einzupassen, werden Terme höherer Ordnung der asymptotischen Entwicklungen des Strömungsfeldes ermittelt, welche zeigen, dass sich die Stromlinienkrümmung in der Grenzschicht in, mit der äußeren Potentialströmung wechselwirkenden Druckstörungen wiederspiegelt. Durch deduktives Einarbeiten dieser in das Grundproblem, kann dann dessen *sachgemäße Gestelltheit* gezeigt werden.

Da das (regularisierte) Cauchy Problem im weiteren Sinne in die Klasse von Reaktions-Diffusions-Gleichungen fällt, wird für gewisse Anfangsbedingungen eine Singularität in endlicher Zeit angenommen, was durch numerische Experimente bestätigt wird. In der Nähe des Entstehungszeitpunkts der Singularität lässt sich weiters eine selbstähnliche Struktur finden, die, wie heuristisch gezeigt wird, sogar eindeutig ist.

Solch auftretende Singularitäten stehen im Zusammenhang mit dem Zusammenbruch (d. h. nicht gleichmäßigen Gültigkeit) der asymptotischen Entwicklungen. Als Folge werden neue Zeit- und Ortskalen induziert, welche in dem hier betrachteten Fall zu einem nichtlinearen "Triple-Deck" Problem führen, das die selbstähnliche Struktur als Anfangswert besitzt.

Der zweite Teil der Arbeit beschreibt in allen Details die verwendeten numerischen Methoden. Dazu werden sogenannte *rationale* Chebyshev Polynome auf $\mathbb{R}^n$ definiert, um Kollokationsschemata, basierend auf orthogonalen Projektionen, entwickeln zu können. Als Grundlage werden klassische Konvergenzresultate im L^2 und L^∞ Sinne für diese neue Klasse von vollständigen Orthogonalsystemen bewiesen. Weiters wird auf die notwendigen Beschränkt- und Kompaktheitseigenschaften der involvierten Operatoren eingegangen.

Abstract

This study deals with the phenomenon of marginally, i.e. locally separated laminar boundary layer flows along smooth surfaces. Such occurrences, in form of separation bubbles, are generally regarded as unstable with respect to certain perturbations and hence can describe what is known as *transition to turbulence*. Thus, this treatise is mainly concerned with unsteady, three-dimensional flows.

As separation regions can be associated with negative values of the wall shear stress, it is important to study its time evolution. In doing so, we deduce the according Cauchy problem from the Navier-Stokes equations at high Reynolds numbers by applying the method of matched asymptotic expansions. To derive the governing integro-differential equation a new, elegant approach utilizing the Fredholm alternative is presented.

Operator symbols and the dispersion relation then prove the general *ill-posedness* of the Cauchy problem. Therefore, it has to be regularized, which is done by a) discretization and b) using regularizing operators. Numerical calculations show further that, modulo discretization errors, these two methods yield the same solutions. Also, such solutions converge to solutions of the original problem for vanishing regularization.

To embed the regularized solutions into the physical concept, higher order terms of the original asymptotic expansions of the flow field are derived. Here the streamline curvature in the boundary layer appears in form of interacting (with the potential flow region) pressure disturbances. By deductively including these in the original problem one can again show its *well-posedness*.

Since the (regularized) problem belongs in principle to the class of reaction-diffusion equations, one can expect, for certain initial conditions, a *finite time singularity* to occur. This is then confirmed by numerical computations. Furthermore, near the blow-up time one can find a self-similar blow-up profile, which is heuristically shown to be unique.

Such singularities indicate the breakdown (i.e. non-uniform validity) of the asymptotic expansions. They also induce new spatio-temporal scales resulting here in a nonlinear "triple-deck" problem, for which the blow-up profile proves to be an initial condition.

The second part of this treatise describes in all detail the methods used for setting up the numerical schemes. Here, what are known as *rational* Chebyshev polynomials are defined on $\mathbb{R}^n$, such that one can develop collocation schemes based on orthogonal projections. As a basis, classical convergence results in L^2 and L^∞ are proved for this new type of complete orthogonal systems. Necessary boundedness and compactness results for the involved operators are presented as well.

Acknowledgments

These are intended to be purely scientific – so, to all my dear friends who have helped me through all those years of studying, please feel yourself thanked and highly esteemed.
Otherwise, I wish to express my utmost thanks to my main supervisor *Prof. Stefan Braun* for the scientific open-mindedness within the scope of the project objectives and despite the established traditions of the institute, thus creating an optimal basis a PhD student can wish for to gain the most knowledge out of such an undertaking. Also, this work was funded by the Austrian Science Fund, FWF project P21426-N22, were Prof. Braun was the principal investigator, for which I am very thankful as well.
Secondly, for being the mathematical conscience, so to speak, I thank *Prof. Christian Schmeiser* for always asking critical and the right questions, thus providing crucial contributions and hints.
As for further subject-specific discussions, the help of *Dr. Sabine Hittmeir* (regarding mathematical details) and of *Dr. Stefan Scheichl* (regarding principles of triple-deck theory) is greatly acknowledged.
Additionally, I would like to thank *Prof. Anatoly Ruban* from the mathematics department at the Imperial College London for the hospitality during my time as an academic visitor, where I also very much enjoyed the discussions with my colleagues *Sebastian Krumscheid* and *David Rottensteiner*.
Special thanks are addressed to *Dr. Harald Schmidt* and *Dr. Michael Pürrer* for constantly challenging my fairly mathematical approach to science per se, hence lifting the value (at least in my opinion) of the present treatise.
At long last, necessary words of thanks go to *Dr. Daniel Lanzerstorfer* and *Ernst Hofmann* for creating a more lively – scientifically, as well as humanly – work environment at the institute.

Dedicated to my parents, Hertha and Alfred,
and my sister Julia,
for their unconditional and never ending support.

Contents

1 General Introduction and Motivation

One might say, the study of fluid mechanics started as early as somewhere around 250 BC, when Archimedes discovered some basic principles of hydrostatics. Since then the subject of fluid mechanics has flourished to be as multifaceted as physics and mathematics themselves. As a consequence many renowned physicists and mathematicians published a vast number of textbooks and chapters introducing the principles of fluid mechanics from lots of different angles.

With the works of Bernoulli, Euler, Navier, Poisson and Stokes (to name just a few) the playing field became even larger by further developing the concepts of conservation laws and differential equations describing the underlying physics in a formal language. But not only this theoretical side, also the practical applications, experiments and engineering works lead to fruitful results, further ramifications and (most importantly for the present treatise) to the need of a deeper theoretical understanding.

Contemporary research areas dealing with or applying fluid mechanics in general go from, say designing aircrafts, studying combustion processes, conducting wind tunnel experiments, developing computer codes by applying numerical analysis and linear algebra, to using topology, algebra and functional analysis to prove existence and uniqueness results for certain Navier-Stokes or Euler problems or applying singular perturbation techniques diving deeper into boundary layer theory. Despite this huge variety of interests and utilized methods there is one common denominator, posing as an open problem, for all the afore-mentioned fields of study – *turbulence*.

It is yet to be fully understood how to properly describe turbulence per se or even to predict and characterize the onset of transition to turbulent flows. Common approaches, presented in, e.g. Landau & Lifshitz (1959) or Marchioro & Pulvirenti (1994), are stability theory, dynamical systems, ergodic and chaos theory on the one side and statistical mechanics, measure and probability theory and stochastic processes on the other. Unfortunately, they all do not necessarily yield satisfying results for all the areas of application. A major issue here might be the fact that we are still missing an answer from the very basis, namely the existence and uniqueness of smooth solutions of the three-dimensional initial-boundary value problem of the Navier-Stokes equations for smooth data. Other (e.g. regarding weak solutions or breakdown) descriptions of this problem can be found in the official problem description by Charles F. Fefferman at the *Clay Mathematics Institute*. Most interestingly here is the fact that in two dimensions one even has strong solutions for the Navier-Stokes equations for all times $0 < t < \infty$, whereas for the three-dimensional problem so far one can only show existence and uniqueness of strong solutions on $t \in [0, T)$, where T might be *finite* (see Sell & You (2002) for more details).

Although the approach in this work comes from a completely different angle – the separation of a laminar boundary layer from the surface – it is still aiming to gain insight into the phenomenon of *transition to turbulence*. As it is well-known, the notion of a *boundary layer* stems from Ludwig Prandtl, who, in his seminal work in 1904, heuristically and exper-

imentally discovered the appearance of a thin viscous layer near the surface in an otherwise irrotational flow. To be more specific, one considers viscous fluids at asymptotically large Reynolds numbers, where the velocity at the surface shall vanish entirely (i.e. no-slip condition). Since such a flow can be regarded as ideal, i.e. satisfying the Euler equations (cf. Landau & Lifshitz (1959)), the decrease of the velocity to zero happens in a thin layer adjacent to the surface. By (reasonably) arguing the vertical velocity component to be small within this layer, one immediately arrives at the equations describing the boundary layer presented by Prandtl. It was not until the development of the techniques of *singular perturbations* and *matched asymptotic expansions* (cf. Eckhaus (1973)) that one was able to base this arguments on profound mathematical grounds. Furthermore, owing to those techniques, the phenomenon of separation of flows past blunt bodies (well established experimentally), could now also be investigated much more systematically and deductively (see Sychev et al. (1998)).

Here we are interested in separation (from smooth surfaces), since it has been observed that it might trigger transition to turbulence. This question became much more prominent with the development and the design of airfoils. In this case, if not leading to turbulence, separation can still significantly influence the flight performance due to the increase of drag and loss of lift forces. In aerodynamics, where high velocities and low viscosities are the dominant fluid flow characteristics, the Reynolds number Re can be assumed to be high, such that it is reasonable to study the limit $Re \to \infty$. Consequently one obtains a singular perturbation problem, as this parameter appears in the viscous (highest order) term of the Navier-Stokes equations. This is the point where Prandtl's heuristic concept and the physical argumentation given, e.g. in Landau & Lifshitz (1959) can be shown to be equal to the results from the perturbation analysis. Eventually, it has been established that the flow outside the viscous layer is irrotational and inviscid (i.e. *potential* flow).

Increasing the angle of attack of the airfoil, the appearance of so-called laminar *separation bubbles* near the leading edge has been observed. It may have a significant influence on the overall flow past the airfoil and, as mentioned above, may lead to higher drag forces. In its initial stages such a bubble is short compared to the chord length of the airfoil, with almost negligible effect on the flow behavior. But as short bubbles only exist within small angle of attack variations, such situations are critical, since for larger angles of attack the bubbles burst, resulting in large, non-negligible separation regions.

These critical stages are commonly referred to as cases of *marginal separation*, which, for being at the verge of a bubble burst, are often seen to trigger the transition process to turbulence. It is thus of high importance to understand the physics underlying it. It has been shown, e.g. in Sychev et al. (1998) in a very deductive manner, that the theory of marginal separation is embedded into the classical (Prandtl) boundary layer concept. However, in some vicinity of the separation region, the hierarchical structure of the boundary layer breaks down due to a singularity occurring in the flow description. As a remedy, the

concept of *viscous-inviscid interactions* has been successfully introduced, using the method of matched asymptotic expansions, see Section 2.1 for more details and some references.

A typical aspect of asymptotic expansions is the search for points or regions where they might break down, i.e. being non-uniformly valid. These can be found by looking for singularities appearing in the solutions, which, as the title of this treatise suggests, shall be done in the following. To make this more accessible to the reader, we provide some guidelines to the structure of this work.

How to read this treatise

As we have mentioned above, transition to turbulence is one of the most important occurrences in the field of fluid dynamics. Since, owing to the mathematical analysis done for the Navier-Stokes equations, turbulence is strongly connected to unsteady, three-dimensional flows, the main aims of this treatise are to extend existing results for unsteady, planar marginally separated flows to unsteady, (locally) three-dimensional set-ups.

It is well-established that a finite time singularity occurs in the solutions of the asymptotic expansions of two-dimensional velocity fields within the theory of marginal separation. Also, this singularity has been shown to admit a unique self-similar structure, which can be utilized to obtain shorter spatio-temporal scales and an according non-linear triple-deck stage, emerging from the blow-up. We refer the reader to Sections 2.1, 2.3.3 and 2.4 for some references and details to the existing results.

Therefore, in very general, three main objectives are investigated in this work: (*i*) finding solutions, using special numerical techniques, to the fundamental Cauchy problem of marginal separation and determine conditions for the finite time blow-up, (*ii*) deriving the scales and equations associated with the blow-up scenario to compute, again numerically, the blow-up profile and see whether it is unique and self-similar and (*iii*) dealing with the question if, and in what sense, a flow description beyond blow-up can be established.

To start with (*i*), we consider the locally three-dimensional Cauchy problem and find sufficient conditions for solutions to terminate in a singularity. Thus, in Section 2.1 we present a novel technique to derive the equation governing this initial value problem. Finding solutions, where the local qualitative and quantitative behavior is of interest, is necessarily a substantial numerical task. Ergo, we put some emphasis on novel methods, deploying polynomial approximations, as derived in this treatise. They are even novel in the sense that the very mathematical basis has not yet been described in a comprehensive and applicable manner in the existing literature. Thus, we shall provide this in full detail in Section 3, which is written in a "stand-alone", self-contained and general fashion, such that it can be easily adapted to various other numerical problems in science and engineering.

To then demonstrate the applicability and convergence of the technique, in Section 2.2 we apply it to the according steady problem for the two- and three-dimensional flows, with the advantage of having independent reference solutions from previous works.

The most important issue is that of the ill-posedness of the Cauchy problem. Thus, we devote Section 2.3.1 to study its characteristics, ramifications and resolution in terms of regularizations. Furthermore, these rather theoretical aspects are then shown how to be utilized for numerical calculations, which further provide exact mathematical meaning to the solutions. Moreover, in Section 2.3.2 a whole new approach to a possible regularization is presented. It is shown that by expanding the flow field to higher orders and applying the matched asymptotic expansion technique, a connection of regularizing operators to the curvature of the streamlines in the boundary layer can be found. This concludes the longstanding discussions on the overall physical meaning of time dependent solutions in the theory of marginally separated flows.

Having found a well-posed problem, were solutions converge (in some sense) to solutions of the original Cauchy problem, the proposed finite time blow-up scenario can be studied. Therefore, in Section 2.3.3 we start with showing how, where and when this singularity appears. To proceed with (*ii*) of finding a self-similar structure, we use the (numerical) data from computing the time evolution as close as possible to the blow-up point and depict it in similarity variables. This yields the convergence of the time dependent solutions to some stationary profile. To gain more information on this structure, an equation is presented governing the exact shape of such a profile. Eventually, we establish the uniqueness of the complete self-similar structure and its blow-up profile.

Section 2.4 concludes with some further considerations regarding (*iii*), i.e. possibilities to extend the study beyond the finite time blow-up. Here one might be interested in continuing the solutions of the Cauchy problem for long times or in deriving a new triple-deck stage, valid after the blow-up. Also, some issues on the stability of the singularity shall be addressed.

Overall, Section 2 as a whole is aimed to be self-contained, with only weak connections made to the numerical analysis from Section 3, such that the reader not interested in the details of the computations may stop after the above mentioned conclusions. Still, a few comments on the structure of Section 3 are in order.

Therein, we start with extending the quasi-comprehensive knowledge on classical Chebyshev polynomials, defined on the unit hypercube, to the n-dimensional Euclidean space. Section 3.2 then continues with some basic properties of projection operators comprising the extended polynomials in $\mathbb{R}^n$ and some new results regarding convergence rates of both approximation and interpolation operators in some L^2 and L^∞ spaces.

In virtue of the equations dealt with in Section 2, especially with respect to the involved operators, Section 3.3 provides some theorems on boundedness and compactness of integral operators. The reader accepting the convergence of the approximation strategy and being rather interested in how to actually apply it to the equations from Section 2, or in general to integro-differential equations, is referred to Section 3.3.2. Therein we show in detail how the finite-dimensional system is obtained. Also, difficulties and special issues arising from the properties of the integral operators and possible remedies are explained. Finally, a Nyström based approach, used for the equation of the blow-up profile, is described in Section 3.4.

2 Marginal Separation Theory

Studies of flow separation from smooth surfaces can be connected to high Reynolds number aerodynamics and airfoil theory. In fact, if the airfoil is thin, i.e. the thickness is about one tenth of the chord length, we say β, the thickness parameter, shall be given by this ratio and additionally shall tend to zero. Assuming further a parallel, inviscid flow enclosing the airfoil and the angle of attack α to be of order β, then the velocities and the pressure (in two dimensions) may be written as

$$u \sim u_\infty + \beta u_1(x, y), \quad v \sim \beta v_1(x, y), \quad p \sim \beta p_1(x, y),$$

see Figure 1 in Section 2.1 for the relation of the velocities and the coordinates.

Remark 2.1. When studying Navier-Stokes dynamics from a theoretical or mathematical viewpoint, there is no need to mention that all involved quantities are, of course, non-dimensional, cf. Sell & You (2002). Nevertheless, such an assumption is not necessarily a trivial one. In fact, the Reynolds number is the result of substituting the suitably scaled, dimensional coordinates, the velocity and the pressure field into the original Navier-Stokes equations, see Equation (2.2). Consequently, the Reynolds number then reads

$$Re = \frac{u_\infty L}{\nu}, \tag{2.1}$$

with L being some characteristic length and ν the kinematic viscosity. It is due to this relation that one has to be careful when making assertions about the state of a flow at a certain Reynolds number. Ruban (1981), for example, uses the radius of the leading edge of an airfoil, while Stewartson et al. (1982) take the chord length for L. In aerodynamics, with the free-stream velocity u_∞ being comparably high and the viscosity small, one can argue to have high Reynolds number flows for both characteristic length scales.
N.b.: In all what follows we assume the appearing quantities to be non-dimensionalized in the above mentioned manner.

Experiments showed for α being small enough for the flow to be fully attached to the airfoil, one has an advantageous ratio of lift to drag forces, admitting good flight performances. By increasing α one can observe a sudden change of the flow situation at the leading edge – the appearance of a short separation bubble enclosing recirculating flow, i.e. the flow separates locally from the surface. This we shall call the separation angle α_s. The short bubble has almost no effect on the surrounding flow and hence could be neglected. But as this state is, on the one hand, highly unstable, meaning that small disturbances, such as e.g. sound waves, can cause this bubble to burst and to either form a larger bubble, extending over a significant part of the airfoil, or to result in full separation. On the other hand, short bubbles only exist in a short range of variation of the angle, i.e. if α exceeds some $\alpha_c > \alpha_s$, the bubble bursts without any further outer influences. Such a situation may have severe consequences on the flight performance (known as *aerodynamic stall*) or may lead to transition to turbulence.

It is thus necessary to theoretically understand the mechanisms behind flow separation and this is best done at critical conditions, i.e. $\alpha \in [\alpha_s, \alpha_c)$, which is called *marginal separation* (see the original works by Ruban (1981) and Stewartson et al. (1982)). A more precise definition for the cases of marginal separation will be given in Section 2.1. As we will demonstrate in the following sections, the advantage of studying the case of marginal separation is that one does not have to deal with turbulence models or have to include large separation regions. By being at the verge of separation, the classical laminar boundary layer theory, supplemented with interaction concepts, provides the necessary frame work. It shall be noted that leading edge separation at airfoils is not the only case where marginal separation can occur, this also happens in channel flows with suction slots or for backward facing steps, where the relative suction rate and the step height to length ratio, respectively, take the role of the angle of attack above. This is why in the next section, when deriving the fundamental equations, we will just assume the existence of such a control parameter and its critical value. Clearly, bubble bursting and transition to turbulence are inherently unsteady and three-dimensional effects, thus the main aim of this treatise is to generalize well-established results for planar flows to unsteady, (locally) three-dimensional problems in a way, such that the previous two-dimensional results are included as special cases.

2.1 The Triple-Deck of Marginal Separation

It has been established in the original papers (see Ruban (1981) and Stewartson et al. (1982)) that the asymptotic description of marginally separated boundary layer flows leads to a so-called *triple-deck interaction* structure. In the following, we shall paraphrase the main ideas presented in these works to demonstrate how the three decks or layers emerge. The aim is, among other things, to show that this (unique) structure does *not* come from any modeling assumptions but from purely physical and mathematical (using matched asymptotic expansions) reasoning. A thorough and detailed deduction can be found in Sychev et al. (1998) and Ruban (2010), and for comparison reasons, we adopted the notations used therein. Furthermore, we will deploy a recently developed, elegant approach to derive the fundamental equations governing the (local) flow properties, the deduction of which, so far, needed heavily involved technical procedures.

Let us first set up the coordinate system, see Figure 1, and define the notions *upstream* and *downstream*. We say, for a given velocity field $\underline{u}^*$ the components (u^*, v^*, w^*) are functions of the coordinates (x^*, y^*, z^*). Since in many theoretical studies, but also in real applications the oncoming, unperturbed flow is considered uni-directional, e.g. $\underline{u}^* = (u^*, 0, 0)$ (or has only comparably small v and w components), *downstream* means in direction of u and vice versa for the notion *upstream*. Consequently we call x the *streamwise* and z the *spanwise* coordinates. Note that the asterisk has no special meaning here and in the following and was chosen purely for the sake of readability of the upcoming asymptotic expansions.

For the following deduction we assume $\underline{u}^* := (u^*, v^*, w^*) = \underline{u}^*(x^*, y^*, z^*, t^*)$ and $p^* = p^*(x^*, y^*, z^*, t^*)$ satisfy the Navier-Stokes equations for incompressible, transient flows on

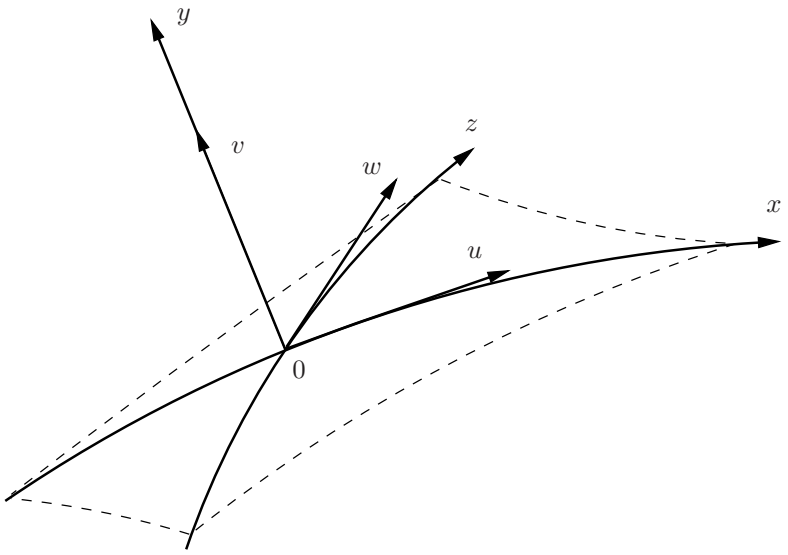

Figure 1: The orthogonal coordinate system with origin at the surface and an according flow field.

$(x^*, y^*, z^*) \in \Omega \subset \mathbb{R}^3$ with a given suction or blowing velocity $v_w^* = v_w^*(x^*, z^*, t^*)$ at the surface $y^* = 0$ and a Reynolds number Re, defined as in (2.1), assumed to be large. The governing equations together with the initial and boundary conditions hence read

$$\left.\begin{aligned} &\partial_t \underline{u}^* + \underline{u}^* \cdot \nabla \underline{u}^* = -\nabla p^* + Re^{-1} \Delta \underline{u}^* \\ &\operatorname{div} \underline{u}^* = 0 \end{aligned}\right\} \quad \text{on} \quad \Omega \times [0,T]$$

$$\begin{aligned} &\underline{u}^* = (0, v_w^*, 0) && \text{at} \quad y^* = 0 \quad \forall x^*, z^*, t^* \\ &|\underline{u}^*| \to 1, \ p^* \to 0 && \text{as} \quad y^* \to \infty \\ &\underline{u}^* = u_0^* && \text{at} \quad t^* = 0 \quad \text{on} \quad \Omega \end{aligned} \tag{2.2}$$

For the sake of comprehensibility and conciseness (and also for historical reasons) say, for the moment, $\underline{u}^* = (u^*, v^*) = \underline{u}^*(x^*, y^*)$ on $\Omega = [0, \infty)^2$, meaning steady and planar flows. In the limit $Re \to \infty$ this then yields the *Euler equations* (i.e. the Laplace term is canceled out), which cannot satisfy the no-slip boundary conditions at the surface $y^* = 0$. We have thus obtained a singularly perturbed problem, where applying the techniques provided in Eckhaus (1973) suggests to find the so-called *significant degeneration*. The resulting scaling factor for the boundary layer variable and the appropriate asymptotic expansions is well-known and given as

$$u^* \sim u_0(x^*, y), \quad v^* \sim Re^{-1/2} v_0(x^*, y), \quad p^* \sim p_0(x^*, y), \quad y = Re^{1/2} y^*,$$

where u_0, v_0 and p_0 satisfy the classical boundary layer equations in (x^*, y), subject to the no-slip condition. By denoting $U_e = U_e(x^*)$ to be the velocity at $y^* = 0$, the Bernoulli equation and the matching condition with u_0 yields $u_0 = U_e$ as $y \to \infty$ and $\partial_{x^*} p_0 = -U_e \partial_{x^*} U_e$. This shows the hierarchical structure of classical boundary layer theory, meaning that the

outer flow solution can be determined independently and its solution is then imposed via the matching procedure onto the boundary layer.

Remark 2.2. In case of non-flat surfaces bounding the flow (cf. Figure 1), e.g. an airfoil, one would have to use curvilinear coordinates for the Navier-Stokes equations, which then include the local curvature. It can be easily seen, in case of moderately curved surfaces, the curvature terms to be a higher order effect in the according boundary layer equations.

A very useful characteristic to study boundary layer flows is known as the *wall shear stress* or *skin friction*, given as

$$\tau(x^*) \propto \partial_y u_0(x^*, y)\big|_{y=0}, \tag{2.3}$$

which is generally regarded to be positive along x^* for an attached boundary layer, whereas (for steady flows) the situation of $\tau \leq 0$ (in some regions) is seen as equivalent to *separation* of the boundary layer from the surface. The seminal work by Goldstein (1948) shows, under the circumstances of a sufficiently strong adverse pressure gradient $\partial_x p > 0$, there exists a point x_0, such that $\tau \propto \sqrt{x_0 - x}$ as $x \to x_0$. The flow description develops a square-root-singularity and the solution ceases to exist $\forall x > x_0$ (e.g. it becomes imaginary). Note that singularity here includes discontinuities, unbounded derivatives and so forth. In essence, one can conclude that near a point of separation the hierarchical structure of the boundary layer concept is no longer valid. This goes even further, for Stewartson (1970) showed the singularity is inevitably present in strongly adverse pressure situations and cannot be removed by including the viscous-inviscid interaction technique.

Suppose we have a parameter k, connected to the geometry or given flow conditions (e.g. the angle α or the height of a backward facing step, as mentioned above), such that $\tau = \tau(x^*, k)$ and

$$\exists\, k_0, x_0^* : \quad \tau(x_0^*, k_0) = 0 \quad \text{and} \quad \tau(x^*, k_0) > 0 \;\; \forall x^* \neq x_0^*. \tag{2.4}$$

Additionally say $\forall k < k_0$, $\tau(x^*, k)$ is everywhere positive, such that a solution exists for all x^* and the boundary layer concept holds, meaning we have a fully attached flow. We refer to the limiting, critical case $k = k_0$ as *marginal separation* (with immediate reattachment). This now provides a precise definition for the marginal separation cases at the leading edge of an airfoil, i.e. for the existence of short separation bubbles when $\alpha \in [\alpha_s, \alpha_c)$, as mentioned in the introduction. As a consequence, the corresponding singularity occurring in the solution is weaker than the Goldstein singularity. In fact, it has been shown in the original works by Ruban (1981) and Stewartson et al. (1982), in situations where k is close to, but still greater than k_0 it is possible to extend the solution continuously through the point of zero skin friction (using the interaction concept). If k is significantly larger than k_0 the flow is fully (or largely) separated (with the appearance of the Goldstein singularity), where the point of separation lies upstream of x_0^* (cf. $\alpha > \alpha_c$ for an airfoil).

The investigation is hence continued with assuming

$$\Delta k := k - k_0 \to 0,$$

i.e. being infinitesimally below critical conditions, where solutions of the boundary layer equations exist along the whole x-axis and passing to the limit $k \to k_0$. Since the velocity and pressure field are sufficiently smooth for $k < k_0$, we can expand both into a Taylor series for small Δk, where the zeroth order terms shall satisfy the classical boundary layer equations, subject to the no-slip condition. It has been presented in very detail by Ruban (2010) that the boundary layer has to be split into various regions when approaching the point of zero skin friction in order to satisfy the no-slip condition at $y = 0$ and the matching condition $u_0 = U_e$ at the outer edge of the boundary layer. What happens is that the viscous part diminishes as the separation point is approached and a mainly inviscid boundary layer remains, with the thickness of the viscous sublayer decreasing according to

$$y = O\big((x_0^* - x^*)^{1/4}\big).$$

Let us further consider the linear term in the Taylor series (with respect to Δk) in the vicinity of x_0^* in the form of an asymptotic series as $x^* \to x_0^*$. It has been shown in the aforementioned works that the linear term for $x^* \to x_0^*$ is not asymptotically small compared to the zeroth order solution. Thus, one can argue a reinvestigation of the viscous sublayer connected to the main part of the boundary layer to be needed in a vicinity of x_0^*. Additionally, the pressure gradient perturbation can be seen to become unbounded in the present set-up as $x^* \to x_0^*$. But this means that the boundary layer induces a pressure perturbation, which starts to influence the leading order boundary layer solution (so far assumed to be given via the outer potential flow). Also, the streamlines in the boundary layer at x_0^* (due to the singular behavior of the solution) experience a kink (or a discontinuity in their gradient), which creates infinitely large perturbations in the outer flow. At this point, *viscous-inviscid interaction* needs to be introduced, meaning that in some region around the point of zero skin friction the boundary layer starts to interact with the outer flow regime. Taking into account the gradient of the induced pressure perturbations order of magnitude estimates yield for this *interaction region*

$$|x_0^* - x^*| = O(Re^{-1/5}), \quad \text{with} \;\; \Delta k = Re^{-2/5} k_1, \tag{2.5}$$

where k_1 remains an order one quantity as $Re \to \infty$, see Ruban (1981).

As mentioned above, the boundary layer near the point of separation consists of a main part and a viscous sublayer, whereas within the proposed interaction region, the outer flow has to be included as well, since only therein the displacement of the streamlines transfers into (simultaneously present) pressure perturbations (yielding the viscous-inviscid interaction concept). Thus, this interaction region consists of three layers, which is termed a *triple-deck structure*. Note that outside this interaction region the hierarchical concept of boundary layer theory remains valid. The crucial conclusion of this whole deduction is, that the interaction

only takes place between the viscous sublayer and the outer flow. Also, due to this procedure, *solutions can be continuously extended through the point of zero skin friction* in the case of the weak singularity for k close to, but greater than k_0.

In knowing the triple-deck structure to play the main role in our treatise, we will state the scalings and expansions in all three layers explicitly, since this builds the basis of deriving the fundamental equations of marginal separation theory. To be more general, this shall be done for the unsteady case, which was first studied in Ruban (1982) and Smith (1982) for planar flows. Also, as transition to turbulence is always viewed as a three-dimensional process, z dependency shall be taken into account as well. This may be realized via locally three-dimensional perturbation devices, such as a surface mounted hump $h = h(x, z, t)$ and a blowing/suction slot with $v_w = v_w(x, z, t)$. Note that these devices enter the problem by modifying the solid boundary and the boundary conditions, cf. Braun & Kluwick (2002) and references therein.

Obviously, with the rescaled original independent variables x^*, y^*, z^* in the interaction region, the domain Ω for the Navier-Stokes equations (2.2) is given henceforth as $\Omega := \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}$. We set the perturbation parameter $\epsilon := Re^{-1/20}$, such that the coordinates are scaled as (indices $1, 2, 3$ denoting the upper, main and lower deck, respectively and the point of zero skin friction is shifted into the origin)

$$t^* = \epsilon^{-1} t, \quad x^* = \epsilon^4 x, \quad z^* = \epsilon^4 z, \quad y^* = \begin{cases} \epsilon^4 y_1 \\ \epsilon^{10} y_2 + \epsilon^{14} h \\ \epsilon^{11} y_3 + \epsilon^{14} h \end{cases} \tag{2.6}$$

Remark 2.3. The scaling for the proposed hump, or say local, smooth alteration of the surface, regarding its height, is chosen, as stated in Braun & Kluwick (2002), such that the resulting pressure perturbations are of the same order as the perturbations stemming from the interaction process.

Figure 2 shows a sketch of the triple-deck structure in accordance to the coordinate scalings above. If not otherwise stated, in what follows the individual expansion terms in these decks are assumed to depend on (x, y_i, z, t) and the asymptotic expansions are taken in principle from Braun & Kluwick (2004) and references therein.

What has been presented so far in this section was merely an excerpt of the comprehensive and physically deductive derivation of how and in what form the three decks actually emerge from the singular perturbation problem for the Navier-Stokes equations, as presented in very detail in Sychev et al. (1998) and even more so in Ruban (2010) (although only for the steady, planar case). The same holds for the explicit, individual description of the flow field expansions in the decks, as given below. This means that we refrain from stating all the equations with their boundary and/or matching conditions for each deck, and rather provide only those, necessary to deduce the fundamental problems of marginal separation.

***(i)* The upper deck.** Here one essentially has a potential flow region. That is at the leading order a constant (within the interaction region), uni-directional velocity field is

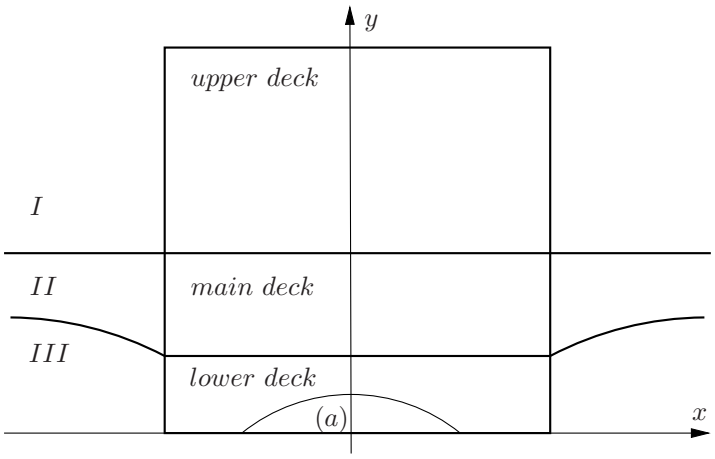

Figure 2: The triple-deck structure of the interaction region. *I–III* indicate the potential flow, the main part of the boundary layer and the viscous sublayer, respectively, upstream of $x_0^* = 0$, with a recirculation region (local separation bubble) (a).

prescribed, i.e. $(u_1, v_1, w_1) = (U_{00}, 0, 0)$ and all spatial coordinates scale with the same power of the Reynolds number. Note that U_{00} is U_e evaluated at x_0^*. The pressure at leading order has to be constant, which can be easily calculated applying Bernoulli's equation to the upper deck and the far field given in (2.2). Thus, we state the expansions to be

$$\begin{aligned} u_1 &\sim u_{10} + \epsilon^4 u_{11} + \epsilon^{10} u_{12} \\ v_1 &\sim \epsilon^4 v_{11} + \epsilon^{10} v_{12} \\ w_1 &\sim \epsilon^{10} w_{12} \\ p_1 &\sim p_{10} + \epsilon^4 p_{11} + \epsilon^{10} p_{12} \end{aligned} \tag{2.7}$$

where, by substituting (2.6) and (2.7) into (2.2) and taking into account the afore-mentioned, one obtains

$$\begin{aligned} u_{10} &= U_{00} & u_{11} &= -U_{01} x \\ p_{10} &= \frac{1 - U_{00}^2}{2} & v_{11} &= U_{01} y_1 \\ & & p_{11} &= p_{00} x. \end{aligned}$$

The imposed pressure gradient p_{00} (at $y_1 = 0$, to be precise) is obviously constant within the interaction region and consequently $U_{01} = p_{00}/U_{00}$. The next higher order terms, induced through the interaction, then have to satisfy

$$\operatorname{div}(u_{12}, v_{12}, w_{12}) = 0, \qquad \begin{aligned} U_{00}\partial_x u_{12} &= -\partial_x p_{12} \\ U_{00}\partial_x v_{12} &= -\partial_{y_1} p_{12} \\ U_{00}\partial_x w_{12} &= -\partial_z p_{12} \end{aligned} \quad \Rightarrow \quad \Delta p_{12} = 0.$$

As expected, from the potential flow assumption, the induced pressure p_{12} has to satisfy the Laplace equation on the half space $y_1 > 0$ subject to Neumann boundary conditions, i.e.

$$\Delta p_{12} = 0 \ \text{ on } \ \Omega, \qquad \partial_{y_1} p_{12} = -U_{00}\partial_x v_{12} \ \text{ at } \ y_1 = 0. \tag{2.8}$$

Green's function for this Neumann problem is given as

$$G(\xi_1,\xi_3,\xi_2;x,y_1,z)=-\frac{1}{4\pi}\left(\frac{1}{|(x,y_1,z)-(\xi_1,\xi_3,\xi_2)|}+\frac{1}{|(x,-y_1,z)-(\xi_1,\xi_3,\xi_2)|}\right).$$

In general a solution can be derived to be $p_{12}=\int\left(p_{12}\partial_n G-G\partial_n p_{12}\right)d\xi_1 d\xi_2$, such that with $\partial_n=-\partial_{y_1}$ and $\partial_{y_1}G=0$

$$p_{12}(x,y_1,z,t)=\int\limits_{\mathbb{R}^2}G(\xi_1,0,\xi_2;x,y_1,z)\partial_{y_1}p_{12}(\xi_1,0,\xi_2,t)\,d\xi_1 d\xi_2$$

and consequently

$$p_{12}(x,0,z,t)=\frac{U_{00}}{2\pi}\int\limits_{\mathbb{R}^2}\frac{1}{|(x,z)-(\xi_1,\xi_2)|}\partial_{\xi_1}v_{12}(\xi_1,0,\xi_2,t)\,d\xi_1 d\xi_2. \tag{2.9}$$

To arrive at the governing equations and problems addressed in this treatise, one does not need to calculate the remaining terms of the upper deck expansions, meaning that having the above description of the pressure perturbation p_{12}, shall be sufficient, as will become clear later.

***(ii)* The main deck.** Although here the expansions assume the form of the classical boundary layer description (at the leading order), that is, the leading term of the streamwise component u_2 only depends on y_2, while the vertical component v_2 and y_2 scale with $Re^{-1/2}$, thus leading to Prandtl's equations (with no w_2 component present), this layer is actually *inviscid*, as deduced in Sychev et al. (1998). Consequently, u_{20} represents the boundary layer velocity profile at the verge of separation and, as can be inferred from the coordinate scalings for t and z in (2.6) and the expansions below, the main deck remains two-dimensional and steady at the leading order. This is in agreement with the assumptions of a two-dimensional outer flow and only locally three-dimensional perturbations. The expansions thus read

$$\begin{aligned}
u_2 &\sim u_{20}+\epsilon^4 u_{21}\\
v_2 &\sim \epsilon^{10}(v_{21}+h_{21})\\
w_2 &\sim \epsilon^{10}w_{21}\\
p_2 &\sim p_{20}+\epsilon^4 p_{21}+\epsilon^{10}p_{22}.
\end{aligned} \tag{2.10}$$

The velocity profile $u_{20}=:U_0(y_2)$ is the same as for the planar flow case, where one can use the usual stream function and the matching condition to obtain its asymptotic behavior (see Sychev et al. (1998))

$$\begin{aligned}
U_0(y_2) &\sim \frac{p_{00}}{2}y_2^2 \quad \text{as } y_2\to 0\\
U_0(y_2) &\to U_{00} \qquad \text{as } y_2\to\infty.
\end{aligned} \tag{2.11}$$

The function h_{21} can be determined, as done in Braun & Kluwick (2002), by using Prandtl's transposition "theorem". It asserts the boundary layer equations to remain in-

variant if the confining wall is shifted by a smooth function with small enough curvature. This assertion has been extended in Glauert (1957) to three-dimensional, compressible flows. In essence we have from these results that for an original boundary layer solution (u, v, w) on $(x, y, z) \in \Omega$, a shifted velocity field $(\bar{u}, \bar{v}, \bar{w})$ on $\bar{y} = y - h(x, z, t)$ given as

$$\bar{u} = u, \quad \bar{v} = v + \partial_t h + u\partial_x h + w\partial_z h, \quad \bar{w} = w$$

satisfies the same equations as (u, v, w) on $(x, \bar{y}, z) \in \Omega$. Thus, by applying the coordinate scalings, which transfer the time and z derivative of h into higher order terms, we have $h_{21} = u_{20}\partial_x h$. Since $u_{20} = U_0(y_2)$,

$$h_{21}(x, y_2, z, t) = U_0(y_2)\partial_x h(x, z, t). \tag{2.12}$$

As for the pressure expansions one can readily see the well-known fact of p_2 being constant across the boundary layer (at every x) and from the matching requirement $p_1 \overset{!}{=} p_2$ as $y_1 \to 0$ and $y_2 \to \infty$ the expansion terms of p_1 and p_2 are mutually equal in this limit. Furthermore, p_1 at $y_1 = 0$ thus determines p_2 for all y_2.

Having established the leading order term U_0 and the behavior of the pressure, substitution of the expansions (2.10) into (2.2) reveals the equations for the next higher order terms

$$\begin{aligned} &\operatorname{div}(u_{21}, v_{21}) = 0 \\ &U_0\partial_x u_{21} + v_{21}U_0' = U_0'' - p_{00} \\ &U_0\partial_x w_{21} = -\partial_z p_{22}, \end{aligned}$$

where combining the conservation of mass and the first momentum equation gives

$$v_{21}(x, y_2, z, t) = -U_0(y_2)\left(\frac{\partial_x A(x, z, t)}{p_{00}} + \int_0^{y_2} \frac{U_0''(s) - p_{00}}{U_0^2(s)} ds\right). \tag{2.13}$$

Remark 2.4. At this point, the function $A = A(x, z, t)$ above represents an integration constant stemming from generally solving the according momentum equation. Therefore, it remains undetermined. Here we enter the crucial part of marginal separation and the whole present treatise, namely finding and solving the governing problem for A. As has been mentioned in the original works, e.g. Stewartson et al. (1982), and can also be easily seen from the definition (2.3), the integration constant A possesses a physical interpretation, that is it is proportional to the wall shear stress (its streamwise component, to be precise)

$$\underline{\tau} = (\tau_x, \tau_y, \tau_z), \qquad \tau_x(x, z, t) \propto \partial_y u\big|_{y=0} \propto A(x, z, t).$$

Furthermore, there is also a connection to the so-called *displacement thickness*, more precisely, A also describes (negative corrections) of the local displacement thickness and is therefore also called *displacement function*. Ergo, within the context of the flow being at the verge of

separation, or at a potential bubble bust, the (local) structure and time evolution of A can qualitatively describe these crucial processes.

Mentioned at the end of the description of the upper deck problem, for deriving the governing equations not all expansion terms have to be given in closed formulae, hence the main deck is sufficiently characterized at this point.

***(iii)* The lower deck.** In classical boundary layer theory, seen as a singular perturbation of the Navier-Stokes equations (2.2), the boundary layer plays the crucial part, i.e. taking care of satisfying the no-slip condition at the surface. With the proposed triple-deck structure, the boundary layer, that is the main deck, acts rather as a "transfer-layer", meaning that the interaction actually takes place between the lower deck (as a viscous sublayer) and the outer flow. This becomes immediately clear from the y_2 independence of the pressure p_2, as mentioned above.

We shall write the lower deck expansions as

$$\begin{aligned} u_3 &\sim \epsilon^2 u_{30} + \epsilon^5 u_{31} + \epsilon^8 u_{32} \\ v_3 &\sim \epsilon^{12}(v_{31} + h_{31}) + \epsilon^{15}(v_{32} + h_{32}) \\ w_3 &\sim \epsilon^8 w_{32} \\ p_3 &\sim p_{30} + \epsilon^4 p_{31} + \epsilon^{10} p_{32}, \end{aligned} \tag{2.14}$$

where it is now necessary to find the governing equations for the terms subscripted with 32, especially since p_{32} represents the induced pressure (perturbations), which are transmitted to the upper deck. For $y_3 \to \infty$ and $y_2 \to 0$ the matching condition gives $p_3 = p_2$ and, as done above, substitution into (2.2) yields

$$u_{30} = p_{00}\frac{y_3^2}{2} \qquad \begin{aligned} u_{31} &= A y_3 \\ v_{31} &= -\frac{y_3^2}{2}\partial_x A \\ h_{31} &= p_{00}\frac{y_3^2}{2}\partial_x h \end{aligned} \qquad h_{32} = A y_3 \partial_x h. \tag{2.15}$$

The problem governing the shape and evolution of the function A can be derived by formulating the equations for (u_{32}, v_{32}, w_{32}). The new and crucial part is that no knowledge of the solution for this higher order terms is needed explicitly, or in other words, the governing equation for A is a solvability condition for the equations of these velocities.

To proceed, we utilize the idea presented in Braun et al. (2012) (for the two-dimensional case) to derive the solvability condition for the lower deck problem in terms of A, thus avoiding the rather tedious procedure followed by Stewartson (1970), Stewartson et al. (1982), Smith (1982) and Ruban (1982) (for planar flows). So far, to the authors knowledge, only the method of these works was used to gain the steady and unsteady problem for A (even for the three-dimensional case studied in e.g. Braun & Kluwick (2004) and Duck (1990)). The new idea, as was further shown, is generic in the sense that it need not be modified to

obtain equations for higher order correction terms of the wall shear stress in the triple deck expansions.

From (2.14) one immediately has $\partial_x u_{31} + \partial_y v_{31} = 0$ and thus there exists a streamfunction Ψ_1 with $\partial_{y_3}\Psi_1 = u_{31}$, $-\partial_x \Psi_1 = v_{31}$ ($w_{31} \equiv 0$). Obviously Ψ_1 is the same as in the planar flow case, satisfying the same problem, that is

$$\begin{aligned} &\mathcal{L}\Psi_1 = 0, \quad \mathcal{L} := \partial_{y_3}^3 - p_{00}\frac{y_3^2}{2}\partial_{xy_3}^2 + p_{00}y_3\partial_x \\ &\left.\begin{aligned} &\Psi_1 = \partial_{y_3}\Psi = 0 \qquad \text{at} \;\; y_3 = 0, \\ &\Psi_1 = A\frac{y_3^2}{2} \qquad\quad \text{as} \;\; y_3 \to \infty \end{aligned}\right\} \;\; \forall x, z, t. \end{aligned} \tag{2.16}$$

Assume now $\psi = \psi(x, \cdot) \in L^2(\mathbb{R})$ with respect to x, such that the Fourier transform (see (2.54)) $\hat{\psi} = \hat{\psi}(k, \cdot)$ and its inverse exist. Then

$$\mathcal{F}(\mathcal{L}\psi) = \widehat{\mathcal{L}}\hat{\psi}(k, \cdot) = \left(\frac{d^3}{dy_3^3} - p_{00}\frac{y_3^2}{2}(ik)\frac{d}{dy_3} + p_{00}y_3(ik)\right)\hat{\psi},$$

where $\widehat{\mathcal{L}}$ is now an ordinary differential operator with respect to y_3.

Define for some $a, b \in L^2(\mathbb{R}^+)$, not necessarily real valued (hence the bar denotes the complex conjugate),

$$\langle a, b\rangle_y := \int_0^\infty a(y)\bar{b}(y)dy.$$

The equation in problem (2.16) obviously gives

$$\mathcal{F}(\mathcal{L}\psi) = \widehat{\mathcal{L}}\hat{\psi} = 0, \tag{2.17}$$

where the boundary condition for large y_3 require A to be Fourier transformable with respect to x (which might have to be understood in a distributional sense).
Let $\phi \in L^2(\mathbb{R}^+)$, as smooth as necessary, then

$$\langle\widehat{\mathcal{L}}\hat{\psi}, \phi\rangle_y = \langle\hat{\psi}, \widehat{\mathcal{L}}^*\phi\rangle_y + B(\bar{\phi}, \hat{\psi}) = 0, \tag{2.18}$$

denoting the function comprising the boundary terms from the integration by parts as B, see Braun et al. (2012) for some more details. It is thus obvious that the homogeneous problem (2.17) and its adjoint have at least one non-trivial solution. Furthermore, we formally claim that the (finite) number of linearly independent solutions for both problems is equal.

Let us now continue with the formulation of the governing equations for the next higher order terms in (2.14). The conservation of mass

$$\partial_x u_{32} + \partial_{y_3} v_{32} + \partial_z w_{32} = 0 \tag{2.19}$$

can be readily deduced, as well as the momentum equations

$$\begin{aligned} p_{00}\frac{y_3^2}{2}\partial_x u_{32} + p_{00}y_3 v_{32} - \partial_{y_3}^2 u_{32} &= -\partial_x p_{32} - y_3(\partial_t A - p_{00}\partial_t h) - \frac{y_3^2}{2}A\partial_x A \\ p_{00}\frac{y_3^2}{2}\partial_x w_{32} - \partial_{y_3}^2 w_{32} &= -\partial_z p_{32} \end{aligned} \tag{2.20}$$

subject to

$$\left.\begin{aligned} u_{32} &= 0 \\ w_{32} &= 0 \\ v_{32} &= v_w \end{aligned}\right\} \quad \text{at} \quad y_3 = 0, \quad \forall x, z, t, \tag{2.21}$$

where v_w is the scaled suction or blowing velocity introduced in (2.2).

One of the advantages of two-dimensional flows, from a technical viewpoint, is the existence of a streamfunction, satisfying the conservation of mass identically and yielding only one momentum equation. Therefore, in order to deploy the idea in Braun et al. (2012), we define

$$\breve{w}_{32} := \int_{-\infty}^{x} \partial_z w_{32} \quad \text{and} \quad \breve{u}_{32} := u_{32} + \breve{w}_{32} \quad \overset{(2.19)}{\Rightarrow} \quad \partial_x \breve{u}_{32} + \partial_{y_3} v_{32} = 0,$$

where differentiation with respect to z and integration with respect to x in the second equation in (2.20) yields

$$p_{00}\frac{y_3^2}{2}\partial_x \breve{w}_{32} - \partial_{y_3}^2 \breve{w}_{32} = -\int_{-\infty}^{x} \partial_z^2 p_{32} d\xi =: -\breve{p}_{32}.$$

By adding this to the first equation in (2.20) we obtain

$$p_{00}\frac{y_3^2}{2}\partial_x \breve{u}_{32} + p_{00}y_3 v_{32} - \partial_{y_3}^2 \breve{u}_{32} = \underbrace{-\partial_x p_{32} - \breve{p}_{32} - y_3(\partial_t A - p_{00}\partial_t h)}_{=:-b_2} - \frac{y_3^2}{2}A\partial_x A. \tag{2.22}$$

A shift of the unknowns $\breve{u}_{32}$ and v_{32} (cf. Sychev et al. (1998)) of the form

$$\begin{aligned} \breve{u}_{32} &= \frac{a_0^2}{24}x y_3^4 + \frac{a_0^2 p_{00}}{4480}y_3^8 + \frac{A^2 - 2a_0 a_1 k_1 - a_0^2 x^2}{2p_{00}} + \widetilde{u_{32}} \\ v_{32} &= -\frac{a_0^2}{120}y_3^5 - \frac{2A\partial_x A - 2a_0^2 x}{2p_{00}}y_3 + \widetilde{v_{32}} \end{aligned}$$

and substitution into (2.22) gives

$$p_{00}\frac{y_3^2}{2}\partial_x \widetilde{u_{32}} + p_{00}y_3\widetilde{v_{32}} - \partial_{y_3}^2 \widetilde{u_{32}} = -b_2 \tag{2.23}$$

where

$$\widetilde{u_{32}} = -\frac{A^2 - 2a_0 a_1 k_1 - a_0^2 x^2}{2p_{00}}, \quad \widetilde{v_{32}} = v_w \quad \text{at} \quad y_3 = 0.$$

The constant k_1 represents the rescaled difference of the control parameter for separation to its critical value, see (2.4) and (2.5), and a_0 and a_1 are (undetermined) integration constants

appearing in higher order expansion terms of the solution of the viscous sublayer ahead of the point of zero skin friction, see Figure 2, region III and Sychev et al. (1998).

Next we introduce a function Ψ_2, mimicking a streamfunction, i.e.

$$\begin{array}{l}\partial_{y_3}\Psi_2 = \widetilde{u_{32}} \\ -\partial_x\Psi_2 = \widetilde{v_{32}}\end{array} \overset{(2.23)}{\Rightarrow} \quad p_{00}\frac{y_3^2}{2}\partial^2_{xy_3}\Psi_2 - p_{00}y_3\partial_x\Psi_2 - \partial^3_{y_3}\Psi_2 = -b_2,$$

and consequently we arrive at the problem for the next higher order unknowns in terms of Ψ_2 reading

$$\begin{array}{l}\mathcal{L}\Psi_2 = b_2 \\ \left.\begin{array}{ll}\partial_{y_3}\Psi_2 & = -\dfrac{A^2 - 2a_0a_1k_1 - a_0^2x^2}{2p_{00}} \\ -\partial_x\Psi_2 & = v_w\end{array}\right\} \quad \text{at} \quad y_3 = 0,\ \forall(x,z,t),\end{array} \tag{2.24}$$

which is in complete accordance to the two-dimensional case. Moreover, reasonably assuming $b_2 \in L^2(\mathbb{R})$ with respect to x gives the inhomogeneous problem $\widehat{\mathcal{L}}\hat{\psi} = \hat{b}_2$ for some $\psi \in L^2(\mathbb{R})$. Say, for the moment, such a ψ exists, then from (2.18), necessarily

$$\langle \hat{b}_2, \phi\rangle_y = \langle \hat{\psi}, \widehat{\mathcal{L}}^*\phi\rangle_y + B, \tag{2.25}$$

where the right hand side equals $B(\bar{\phi}, \hat{\psi})$, for all $\bar{\phi}$ solving the homogeneous adjoint problem and $\hat{\psi}$ solving the inhomogeneous problem above. By viewing $\mathcal{L} = \sum_{|k|} a_k(x,y)\partial^k_{xy}$, where the principal part has $a_3 \equiv 1$, yields $\mathcal{L}$ to be elliptic and thus the *Fredholm alternative* for elliptic operators shows (2.25) to be also sufficient for solvability of the inhomogeneous problem.

Remark 2.5. As we have pointed out earlier, the homogeneous equation $\widehat{\mathcal{L}}\hat{\psi} = 0$ and its adjoint $\widehat{\mathcal{L}}^*\bar{\phi} = 0$ have at least one non-zero solution with respect to y_3, for all k, z, t, where consequently the question of uniqueness of the inhomogeneous problem $\widehat{\mathcal{L}}\hat{\psi} = b$ for a given $b = b(y,\cdot) \in L^2(\mathbb{R}^+)$ remains. In fact, one has that if the homogeneous, adjoint problem has only the trivial solution, there exists a unique solution for $\widehat{\mathcal{L}}\hat{\psi} = b$. Since we are only interested in the mere existence of a solution, the Fredholm alternative provides a sufficient criterion for the present case.

One can find, as done in Braun et al. (2012), a solution to $\widehat{\mathcal{L}}^*\bar{\phi} = 0$ in closed form given as (the bar again denotes the complex conjugate)

$$\bar{\phi}(k, y_3, z, t) = c(k, z, t)(ik)^{1/8}\, y_3^{3/2}\, K_{1/4}\big(y_3^2\sqrt{p_{00}ik/8}\big) =: c(k, z, t)\tilde{\phi}(k, y_3),$$

with $K_{1/4}$ denoting the modified Bessel function of the second kind and c being the undetermined integration "constant". It is straight forward to see the L^2 integrability of $\tilde{\phi}$ on $y_3 \in [0, \infty)$, as required (in fact, $\tilde{\phi}$ decays exponentially as $y_3 \to \infty$). Therefore, in virtue of the solvability condition (2.25) considered for (2.24) and taking the expression for the

boundary terms B from Braun et al. (2012) yields

$$\langle \hat{b_2}, \phi \rangle_y = \int_0^\infty \hat{b_2}\, c\tilde{\phi}\, dy_3 = B(c\tilde{\phi}, \hat{\Psi}_2). \tag{2.26}$$

Remark 2.6. Having said nothing about L^2 integrability of b_2 with respect to y_3, the meaning of the inner product on the left hand side above is still to be considered. From (2.22) it becomes clear that the violation of the integrability, if so, can only come from insufficient decay of b_2 at zero and/or infinity. On the other hand, the existence (per se) of the integral defining this inner product is clear from the behavior of $\tilde{\phi}$ at the boundaries. Following the overall deduction this existence is in fact sufficient for the result presented below to have the correct meaning.

With this remark, we assert a necessary and sufficient condition for the existence of Ψ_2 (satisfying (2.24)) is given by the second equality in (2.26), which certainly holds if one equates the left and right hand side excluding the undetermined function c. Furthermore, we additionally assumed all modifications leading to this condition to be set in $L^2(\mathbb{R})$, ergo the inverse Fourier transform and the convolution theorem can be applied with respect to x, leading to

$$\int_0^\infty b_2 * \mathcal{F}^{-1}\tilde{\phi}\, dy_3 = \mathcal{F}^{-1}\big(B(\tilde{\phi}, \hat{\Psi}_2)\big|_{y_3=0}\big).$$

We reasonably assume the functions on the left hand side to be smooth enough, such that the order of the Fourier transform, the convolution and the inner product integral can be arbitrarily interchanged. On the right hand side it is important to evaluate B (i.e. its arguments) at $y_3 = 0$ before applying the inverse Fourier transform. In concrete, that is substituting b_2 from (2.22) and using

$$\Phi(x, y_3) := \mathcal{F}^{-1}\tilde{\phi} = c_1 \frac{y_3^2}{x^{5/4}} \exp\left(-c_2 \frac{y_3^4}{x}\right) \mathbb{1}_{[0,\infty)}(x)$$

with again referring to Braun et al. (2012) for the details and the values of the all the constants c_i appearing, this reads

$$\begin{aligned} c_3 \partial_{y_3}\Psi_2\big|_{y_3=0} - c_4 \int_{-\infty}^{x} \frac{1}{(x-\xi)^{1/4}} \partial_\xi \Psi_2\big|_{y_3=0}\, d\xi = \\ = c_5 \int_{\mathbb{R}} \int_0^\infty \Phi(x-\xi, y_3)\big(\partial_x p_{32} + \breve{p_{32}} + y_3(\partial_t A - p_{00}\partial_t h)\big)\, dy_3 d\xi. \end{aligned} \tag{2.27}$$

Remark 2.7. The fractional integral on the left hand side stems from, given $a < 1$, $\mathcal{F}^{-1}\big((ik)^a \hat{f}\big) = \mathcal{F}^{-1}\big((ik)^{1-a}(ik)\hat{f}\big) = \mathcal{F}^{-1}(ik)^{1-a} * \partial_x f$. In the following sections we will

provide a precise meaning of such integrals in terms of Abel operators and their symbols (see Sections 2.3.1 and 3.3.1).

Substituting further the boundary conditions from (2.24), using

$$\int_0^\infty y^2 e^{-y^4/x} dy = \frac{1}{4} x^{3/4} \Gamma(3/4) \quad \text{and} \quad \int_0^\infty y^3 e^{-y^4/x} dy = \frac{x}{4},$$

and deploying the affine transforms for the unknown functions and variables presented in Sychev et al. (1998) and Braun & Kluwick (2004) eventually yields

$$\begin{aligned} A^2 - x^2 + \Gamma = -\lambda \int_{-\infty}^{x} \frac{\partial_x p_{32} + \int_{-\infty}^{\xi} \partial_z^2 p_{32} d\zeta}{(x-\xi)^{1/2}} d\xi - \gamma \int_{-\infty}^{x} \frac{\partial_t (A-h)}{(x-\xi)^{1/4}} d\xi - \\ - \gamma \int_{-\infty}^{x} \frac{v_w}{(x-\xi)^{1/4}} \, d\xi, \end{aligned} \tag{2.28}$$

with λ and γ being positive and $\Gamma \in \mathbb{R}$ denoting the rescaled control parameter, i.e. k_1 in (2.5). Note that the fractional integral for the pressure stems from the characteristic function $\mathbb{1}_{[0,\infty)}$ in Φ and the convolution of the pressure terms in (2.27) with $x^{1/2}$.

In order to obtain a problem purely in terms of A as the unknown, one needs to relate A to the pressure p_{32}. The result for p_{12}, Equation (2.9), has been established to match with p_{22} (in the limit $y_1 \to 0$, $y_2 \to \infty$) and consequently with p_{32} (as $y_3 \to \infty$, $y_2 \to 0$), such that we will henceforth call p_{32} the *interaction pressure* p_i. Since v_{12} therein is undetermined, the matching condition $v_{12} \overset{!}{=} v_{21} + h_{21}$ as $y_1 \to 0$ and $y_2 \to \infty$, which consequently has to hold for the according derivatives with respect to x as well, i.e.

$$\partial_x v_{12}(x, 0, z, t) \overset{!}{=} \partial_x v_{21}(x, y_2, z, t) + \partial_x h_{21}(x, z, t) \quad \text{as} \quad y_2 \to \infty,$$

provides an appropriate relation. From (2.13) and (2.12), taking into account $U_0(y_2) \to U_{00}$ as $y_2 \to \infty$ and using again the affine transform, the interaction pressure reads

$$p_i(x, z, t) = -\frac{1}{2\pi} \int_{\mathbb{R}^2} \frac{1}{|(x, z) - (\xi_1, \xi_2)|} \partial_{\xi_1}^2 (A - h) \, d\xi_1 d\xi_2. \tag{2.29}$$

Remark 2.8. It is important to note at this point that in all the considerations above initial conditions (for the velocity and pressure field) were never taken into account. From a physical and historical viewpoint this relates to the initial studies of the steady problem, which were later extended to unsteady situations. Hence, one might claim the time evolution to only make sense when applying unsteady perturbations to steady states. Mathematically, obviously, an initial condition for (2.28) remains as arbitrary as initial conditions for (2.2), cf. Section 2.3.

Overall, what has been deduced as a solvability condition for the velocity field of the viscous wall layer is thus regarded as the fundamental problem of marginally separated flows. By providing all necessary theoretical instruments, such as a control parameter Γ and flow control devices h and v_w, solving this problem yields criteria for detecting when and where the flow breaks down.

It is easily seen from the triple-deck structure and the resulting fundamental equations derived in this section that, in principle, the three-dimensionality stems from involving z dependent perturbations (in form of the hump and the suction slot). Otherwise, from a heuristic viewpoint, there is no reason for the viscous sublayer to develop z dependency. This becomes even more obvious when considering the problems and results presented in Sections 2.2 and 2.3. Such a (naive) conclusion might not hold for values of Γ being near critical conditions, by which we mean the existence of an upper bound for Γ (cf. the bifurcation diagram in Figure 4), above which no real, planar steady state solutions exist and three-dimensionality might be inherently present (if we allow for pressure perturbations with respect to z, cf. $p_{32}^{\breve{}}$ in (2.27)).

Gaining insight into the local behavior of solutions of (2.28) is necessarily a numerical task, even in the case of the according steady problems. Hence, in the next section we present a novel computation technique based on polynomial approximations to establish sufficient accuracy and convergence results, such that for the main investigation, regarding the Cauchy problems, a consistent spatial discretization can be assumed. Also, from here on, references to (spatially) *two-dimensional* (*planar*) and *three-dimensional* problems are always in virtue of the according flow field, even though the actual problems with respect to the unknown A are independent of y.

2.2 Steady Problems

It is common practice in fluid dynamical research to start theoretically investigating certain physical set-ups and the according problems by considering steady state solutions. This is partly due to an easier comparison to experiments. Also, it is not hard to imagine the difficulties arising in experiments when studying unsteady effects by simultaneously avoiding the involuntary inclusion of disturbances. Moreover, stationary solutions, mostly given as numerically computed data, are always the starting point of all sorts of stability analysis. Some spacial kinds of instabilities will be presented in Section 2.3.

It is quite obvious, when considering the structure of (2.28) that a good working numerical method is the almost exclusive way to establish some knowledge on the important characteristics of A, such as the domain of its negative values. Although numerical solutions of the steady problems are well-known, we shall reinvestigate them in the following, mainly to demonstrate the qualities of the novel numerical technique developed and analyzed in Section 3. In doing so, it is preferable to combine the fundamental problem (2.28) and (2.29) into one equation. From the right hand side in (2.28) it is obvious that the unknown A assumes an at most linear growth at infinity. Thus, with the decay of the kernel in (2.29)

the boundary terms when integrating by parts vanish and consequently, assuming sufficient differentiability, the pressure terms in (2.28) can be replaced by the accordingly modified term from (2.29). Defining polar coordinates $(x,z) \to (r,\phi)$ and setting $\partial_t(A-h)=0$, we finally say $A=A(x,z)$ shall satisfy the following problem in $\mathbb{R}^2$

$$\begin{aligned} A^2 - x^2 + \Gamma &= \frac{\lambda}{2\pi} \int\limits_{-\infty}^{x} \frac{1}{(x-s)^{1/2}} \int\limits_{\mathbb{R}^2} \frac{\partial_{\xi_1}^3 + \partial_{\xi_1}\partial_{\xi_2}^2}{|(s,z)-(\xi_1,\xi_2)|} A(\xi_1,\xi_2) d\xi_1 d\xi_2 \, ds + g(x,z) \\ A(x,z) &\sim c(\phi) r \quad \text{as} \quad r \to \infty. \end{aligned} \tag{2.30}$$

The far field condition above just indicates an at most linear growth behavior, depending on ϕ. Alternatively, considering the x and z coordinate separately, we write this condition as

$$A(x,z) = O(|x|) \quad \text{as} \quad |x| \to \infty, \quad A(x,z) < \infty \quad \text{as} \quad |z| \to \infty. \tag{2.31}$$

One can identify the one-dimensional integral above as an Abel operator and the double integral as the two-dimensional Riesz potential. For the general definitions and some basic boundedness and compactness properties of such operators, we refer to Section 3.3.1 (e.g. Theorems 3.32 through 3.38). Thus, in the following we will regard the fundamental problem as an integro-differential equation written as

$$A^2 - x^2 + \Gamma = \frac{\lambda}{2\pi} \mathcal{J}_{-\infty}^{1/2} \mathcal{R}^1 \left([\partial_x^3 + \partial_x \partial_z^2] A \right)(x,z,t) + g(x,z), \tag{2.32}$$

Remark 2.9. The function g above contains the hump and/or the suction/blowing device (cf. the deduction in Section 2.1), and hence one can view g as a general inhomogeneity or forcing term. Due to the linearity of the right hand side operators in (2.32) it follows immediately that the argument $A-h$, as used in the deduction for (2.28) and (2.29), can be separated. Viewing such a hump as a surface mounted obstacle it can act as a flow control device, i.e. shifting or delaying separation of the laminar boundary layer. More details on these subjects can be found in Braun & Kluwick (2002).

The according two-dimensional or planar problem, originally derived and investigated in Ruban (1981) and Stewartson et al. (1982) (with $g \equiv 0$), for $A=A(x)$ in $\mathbb{R}$ reads

$$\begin{aligned} A^2 - x^2 + \Gamma &= \lambda \int\limits_{x}^{\infty} \frac{1}{(s-x)^{1/2}} A''(s) ds + g(x) \\ A(x) &= O(|x|) \quad \text{as} \quad |x| \to \infty, \end{aligned} \tag{2.33}$$

where the dash denotes derivatives with respect to x and again an Abel operator appears on the right-hand side above, such that the equation can be rewritten as

$$A^2 - x^2 + \Gamma = \lambda \, \mathcal{J}_{\infty}^{1/2} \big(A'' \big)(x) + g(x). \tag{2.34}$$

Remark 2.10. It does not come as a coincidence that the left hand sides of (2.33) and (2.30) are identical, since in both problems the outer flow field (cf. the leading order of the upper and main deck descriptions in Section 2.1) is assumed to be parallel and planar. In fact, the only difference, which can be most easily seen from the novel deduction of (2.28) and (2.29), between the two- and three-dimensional problem here lies in the term $p_{32}^{\breve{}}$, which, by containing derivatives with respect to z, vanishes for planar flows. The necessary change in the relation for the interaction pressure (2.29) when independent of z can be readily seen from the modifications done in Remark 2.14.

Remark 2.11. The previous remark stands in contrast to what has been studied in Duck (1989), where an overall three-dimensional (although z-symmetric) set-up is prescribed. The fundamental equations therein are given similarly to (2.28) and (2.29), where the latter is identical in both cases. What distinguishes locally and globally three-dimensional flows lies exclusively in the left hand side of the fundamental problem, which explicitly depends on x *and* z in Duck (1989). Although the far field condition in (2.30) remains the same, in the case of global three-dimensionality a relation similar to (2.31) cannot be derived. For further (numerical) treatment of the problem, Duck (1989) reformulated the combination of the Abel operator and the Riesz potential into one double integral with a more involved kernel function (see Remark 3.66). For the approach in this treatise such a description does not lead to any simplification and hence was not adopted.

Remark 2.12. Form the partial derivatives on the right hand side in (2.32) we assume classical solutions of the equation to be at least three times continuously differentiable on $\mathbb{R}^2$. The at most linear growth given as a far field condition in (2.30) or (2.31) thus renders the argument $[\partial_x^3 + \partial_x \partial_z^2]A(x,z)$ of the integral operators *bounded and continuous* on $\mathbb{R}^2$, with a decay rate of r^{-2} (in principle). Therefore, using the fact that combinations of compact and bounded operators are compact, one can claim permissible (classical) solutions to satisfy the requirements of Theorems 3.32, 3.33 and 3.38 and assert the right hand side operators to form a compact mapping (cf. the argumentation on compactness of singular integro-differential operators on Lyapunov curves between Sobolev and L^p spaces presented in Mikhlin & Prößdorf (1980)). The two-dimensional case (2.34) can be argued analogously.

Remark 2.13. The knowledge whether the operators involved are bounded or compact can be crucial to the existence and uniqueness of solutions and to the convergence of approximations. Equations of the first kind, for example, are ill-posed if the operator is compact, but for equations of the second kind, compactness implies existence and uniqueness (in most cases). Additionally, regarding approximation procedures, boundedness can often be necessary and compactness even sufficient for convergence (cf. the general results in Hackbusch (1995)).
Most authors, when treating singular integral equations, deal mainly with Abel or Cauchy type equations. Hence, it is worth mentioning the more general approach used in Prößdorf (1974). Therein results regarding solutions, i.e. invertibility of the involved operators, are found in terms of operator algebras and the notion of the *symbol*. The latter will be utilized in Section 2.3.1 to deal with the well-posedness of Cauchy problems.

Remark 2.14. The problems (2.33) and (2.30) are consistent in the sense that assuming $A = A(x)$ equivalence of (2.30) and (2.33) can be proved. Obviously, for the far field condition, this is trivially satisfied. Considering the right-hand sides in both equations, we have (for an argument function depending only on x)

$$\begin{aligned}
&\frac{\lambda}{2\pi}\int_{-\infty}^{x}\frac{1}{(x-s)^{1/2}}\int_{\mathbb{R}^2}\frac{1}{\sqrt{(s-\xi_1)^2+(z-\xi_2)^2}}\,\partial_{\xi_1}^3 f(\xi_1)d\xi\,ds = \\
&= -\frac{\lambda}{2\pi}\int_{-\infty}^{x}\frac{1}{(x-s)^{1/2}}\int_{\mathbb{R}^2}\frac{s-\xi_1}{((s-\xi_1)^2+(z-\xi_2)^2)^{3/2}}\,\partial_{\xi_1}^2 f(\xi_1)d\xi\,ds = \\
&= -\frac{\lambda}{2\pi}\int_{-\infty}^{x}\frac{1}{(x-s)^{1/2}}\int_{\mathbb{R}}(s-\xi_1)\partial_{\xi_1}^2 f(\xi_1)\underbrace{\int_{\mathbb{R}}\frac{1}{((s-\xi_1)^2+(z-\xi_2)^2)^{3/2}}d\xi_2}_{=2/(s-\xi_1)^2}\,d\xi_1\,ds = \\
&= -\frac{\lambda}{\pi}\int_{-\infty}^{x}\frac{1}{(x-s)^{1/2}}\int_{\mathbb{R}}\frac{\partial_{\xi_1}^2 f(\xi_1)}{s-\xi_1}d\xi_1\,ds = -\frac{\lambda}{\pi}\int_{x}^{\infty}\partial_{\xi_1}^2 f(\xi_1)\underbrace{\int_{-\infty}^{x}\frac{1}{(s-\xi_1)(x-s)^{1/2}}d\xi_1}_{=-\pi/\sqrt{\xi_1-x}}\,ds = \\
&= \lambda\int_{x}^{\infty}\frac{1}{(\xi_1-x)^{1/2}}\partial_{\xi_1}^2 f(\xi_1)d\xi_1,
\end{aligned}$$

where we used integration by parts in the second line. By assuming f to be twice continuously differentiable and its second derivative to decay to zero, Theorem 3.40 shows the existence of the resulting integral and hence the modifications for the third line are justified, where the appearing integrals are regarded as their Cauchy principal value, if necessary. □

Numerical Solutions

Solutions of (2.33) and (2.30) are almost comprehensively treated, using various kinds of numerical schemes (see e.g. Stewartson et al. (1982), Brown & Stewartson (1983), Sychev et al. (1998), Scheichl et al. (2008), as well as, Duck (1989), Braun & Kluwick (2002) and references therein). In the following we shall present a novel technique based on polynomial approximation and spectral collocation, which works equally well for both situations and, as will be shown, the method for the three-dimensional problem can be directly applied to the two-dimensional case. In principle, we use the steady problems to establish convergence of the polynomial scheme, with the advantage of having independent reference solutions.

The very basis of the approach is presented in Section 3. Therein, Section 3.1 states the main properties of *rational Chebyshev polynomials*, Section 3.2 then provides convergence rates and results in terms of truncated sums and projections, as well as interpolation operators. Section 3.3 contains necessary characteristics of the involved operators and the

essentials of setting up the actual collocation algorithm as well as some general consistency, stability and convergence considerations.

Problems (2.33) and (2.30) are defined in an unbounded domain, where the unknown function admits an algebraic far field behavior. Thus, the above mentioned *rational Chebyshev polynomials* are the most appropriate basis functions. As shown in Theorem 3.1 and Lemma 3.4, these polynomials form a complete orthogonal set in the weighted Lebesgue space $L^2_u(\mathbb{R}^n)$, with $u(x) = \prod_{i=1}^n 1/(1+x_i^2)$. Although such spaces allow for functions to grow (weakly) at infinity, due to the given asymptotic behavior in (2.33) and (2.30), one cannot expect the function A to lie in these spaces. From (2.31) it is clear that $A = A(\cdot, z)$ remains bounded and hence we can subtract the growth with respect to x, such that

$$A(x,\cdot) = B(x,\cdot) + \sqrt{1+x^2}, \quad B(x,\cdot) = O(|x|^{-1}) \ \text{ as } \ |x| \to \infty. \tag{2.35}$$

It is now reasonable to assume $B \in L^2_u(\mathbb{R}^2)$, satisfying the modified equation (subject to the far field above)

$$B^2 + 2\sqrt{1+x^2}B + \Gamma + 1 = \frac{\lambda}{2\pi}\,\mathcal{J}^{1/2}_{-\infty}\,\mathcal{R}^1\left([\partial_x^3 + \partial_x\partial_z^2]B\right)(x,z) + f(x) + g(x,z), \tag{2.36}$$

where the function f results from substituting (2.35) into the right-hand side of (2.32). Remark 2.14 shows, for functions independent of z, the right-hand side operators in (2.32) to be reduced to the right-hand side in (2.34) and consequently

$$f(x) = \lambda \mathcal{J}^{1/2}_{\infty}\big((1+x^2)^{-3/2}\big), \tag{2.37}$$

which can be given in terms of hypergeometric functions or elliptic integrals using some (computer) algebra.

According to Theorem 3.1, Lemma 3.4 and Theorems 3.5 and 3.6 one can approximate B by B_N using the rational Chebyshev polynomials R_i as defined in Section 3.1 in (3.1), such that

$$B_N(x,z) = \mathcal{P}_N B(x,z) = \sum_{i=0}^{N_x}\sum_{k=0}^{N_z} a_{ik} R_i(x) R_k(z), \tag{2.38}$$

with B_N reasonably expected to converge to B. The operator $\mathcal{P}_N$ is the orthogonal projection given in Lemma 3.7. We also applied here the definition of multivariate polynomials (3.10), which can be seen as a tensor basis description.

Note that by saying $N_z = 0$ in (2.38) we immediately obtain the one-dimensional expansion, since $R_0 \equiv 1$. Additionally, in virtue of computational costs, we will treat N_x, N_z independently (in contrast to the argumentation in Section 3). With Theorem 3.6 showing the coefficients a_{ik} to be unique and independent of N, we have a good, heuristic and easy-to-test convergence criterion.

Remark 2.15. As established in Remark 2.12 one can expect the operators on the right hand side in (2.36) to be compact between the space of three times continuously differen-

tiable functions with at most linear growth at infinity and the space of bounded, continuous functions where a limit exists at infinity.
This certainly holds for a function B as in (2.35) when assuming A to be a classical solution (i.e. satisfying all differentiability requirements). The convergence rate of the series expansion, also taking into account the decay rate of B, can then be estimated using Lemmas 3.9 and 3.10 and Theorem 3.11, which show, that one actually works on the Sobolev-type space $H^r_{u,A}$, defined in (3.19). Consequently, due to the comparably slow decay at infinity of B, with respect to the requirement to lie in $H^r_{u,A}$, we cannot expect the usual exponentially fast convergence such expansions provide on bounded domains. This is supported by Theorem 3.15, where the decrease in the absolute value of the expansion coefficients also depends on the far field behavior of the function. Still, uniform convergence of the approximation, according to Theorem 3.13, can be assumed. We might regard this as even more important than fast convergence rates in Sobolev spaces, since for physical interpretations of the solutions, pointwise and uniform accuracy is crucial.
In virtue of the spectral collocation scheme set up below, one normally assumes functions to be continuous and to have finite L^∞ norm. Thus we shall note, that the involved operators are compact (or at least bounded) in such function spaces, where then consistency follows immediately (see Section 3.3).

Remark 2.16. Problems (2.33) and (2.30) are, above all, *nonlinear*. It is well established in the field of spectral methods that nonlinear equations (or nonlinear terms) are best treated using interpolation, meaning that Galerkin and collocation methods are not the best choice (see Sections 3.2.2 and 3.3.2). On the other hand, interpolation introduces the so-called *aliasing error* (cf. Section 3.2.2), which is essentially the difference between evaluating the coefficients in (2.38) exactly and approximately. But, as proved in Theorems 3.22 through 3.24, this error is of the same order of magnitude as the error made by truncating the expansion series.

The advantage of using interpolation (with function values being the discrete unknowns) lies in the evaluation of the nonlinearity, where this is done by simple pointwise calculations. It is obvious, having the additional integrals, that the Galerkin approach becomes heavily involved, even for weak nonlinearities (as the quadratic term here). Collocation lies somewhere between these two, as the coefficients are the unknowns, but the equation system is set up by pointwise evaluation.

We start with the two-dimensional case, where (with respect to (2.38))

$$B(x) \approx \mathcal{P}_N B(x) = \sum_{i=0}^{N} a_i R_i(x)$$

and substitution into (2.36) yields

$$\big(\mathcal{P}_N B(x)\big)^2 + 2\sqrt{1+x^2}\sum_{i=0}^{N} a_i R_i(x) + \Gamma + 1 = \lambda \sum_{i=0}^{N} a_i \mathcal{J}_\infty^{1/2}\big(R_i''\big)(x) + f(x) + g(x).$$

For reasons explained in Remarks 3.34 and 3.40 and Section 3.3.2, we take the zeros x_j of R_{N+1}, given via Lemma 3.2(ix), as the collocation points to obtain further

$$\big(\mathcal{P}_N B(x_j)\big)^2 + \sum_{i=0}^{N} a_i \underbrace{2\sqrt{1+x_j^2}}_{=:s_j} \underbrace{R_i(x_j)}_{=:C_{ji}} + \Gamma + 1 = \sum_{i=0}^{N} a_i \underbrace{\lambda \mathcal{J}_\infty^{1/2}\big(R_i''\big)(x_j)}_{=:K_{ji}} + f(x_j) + g(x_j), \quad (2.39)$$

for $j = 1, \ldots, N+1$, which reads in matrix vector form, denoting the vector $\underline{a} := (a_i)$ etc.,

$$\big(\underline{C}\,\underline{a}\big)^2 + (\underline{Cs})\,\underline{a} + \Gamma + 1 = \underline{K}\,\underline{a} + \underline{f} + \underline{g}, \quad (2.40)$$

with the matrix $\underline{Cs}$ defined via its elements $Cs_{ij} := s_j C_{ji}$. Here, apart from dealing with the nonlinearity, the obvious substantial task is obtaining the matrix $\underline{K}$ in a fast and accurate way. The according procedures are presented in detail in Section 3.3.2. Analogously, substituting (2.38), with $N = (N_x, N_z)$, into (2.36) and evaluating at the zeros (x_j, z_l) of $R_{N_x+1}R_{N_z+1}$ yields

$$\begin{aligned} &\big(B_N(x_j,z_l)\big)^2 + 2\sqrt{1+x_j^2}\sum_{i=0}^{N_x}\sum_{k=0}^{N_z} a_{ik}\underbrace{R_i(x_j)R_k(z_l)}_{=:C_{ijkl}} + \Gamma + 1 = \\ &= \sum_{i=0}^{N_x}\sum_{k=0}^{N_z} a_{ik}\underbrace{\frac{\lambda}{2\pi}\,\mathcal{J}_{-\infty}^{1/2}\,\mathcal{R}^1\left([\partial_{\xi_1}^3 + \partial_{\xi_1}\partial_{\xi_2}^2]R_i(\xi_1)R_k(\xi_2)\right)(x_j,z_l)}_{=:K_{ijkl}} + f(x_j) + g(x_j,z_l), \end{aligned} \quad (2.41)$$

where one can immediately see, by abbreviating the matrices as in the two-dimensional case, we obtain the same description for the discrete equations, i.e. (2.40).

Note that $\underline{a} = (a_{ik})$ is now a (possibly rectangular) matrix. By writing $\underline{a}$ as a vector, one can arrange the C_{ijkl} in a $(N_x+1)(N_z+1) \times (N_x+1)(N_z+1)$-matrix, such that $\underline{C}\,\underline{a} = B_N(x_j, z_l)$, as required. This is done analogously for K_{ijkl}, where the entries are obtained by the method described in Section 3.3.2. Thus, in (2.40) one just has to substitute $\underline{C}$, $(\underline{Cs})$, $\underline{K}$ and $\underline{g}$ with the matrices approximating the operators in (2.33) or (2.30) and gains the according coefficients $\underline{a}$. Furthermore, methods for solving (2.40), derived in the following, hold for arbitrarily sized (quadratic) matrices.

For an easy and efficient handling of the matrix-vector terms and equations presented in the whole treatise, the numerical linear algebra packages provided by the *NAG* ("Numerical Algebra Group". *www.nag.co.uk*) have been utilized.

As mentioned on several occasions in the introductory paragraphs and remarks of this subsection, the necessary results regarding the properties of the operators and the spectral collocation method are provided in Section 3.3.2, such that consistency of the numerical schemes described here follows immediately.

Remark 2.17. An important advantage of spectral collocation methods can be observed at this point. In setting up the equations system above we never have used any type of boundary

conditions or imposed the far field behavior of the unknown function B. This is in virtue of the fact that the expansion (2.38) is assumed to be uniformly convergent, implying that a consistent scheme yields the same behavior at infinity for B_N ($N \gg 1$) as in B. For the sake of completeness it shall be noted that if one would want to impose certain values at infinity, this can be done by either using another set of collocation points (the extrema of R_N, for example, are distributed up to the boundary) or by modifying the polynomials themselves, such that every polynomial satisfies the boundary conditions individually.

Remark 2.18. Solving (2.40) requires an iteration scheme for the quadratic nonlinearity. By taking $A_{n+1} = A_n + \delta A$ for the original unknown function, while assuming δA to be small in some sense, then substituting this ansatz into (2.32) or (2.34) and canceling the $(\delta A)^2$ term, yields an iteration procedure for A_n (the same holds for B, where $\delta B = \delta A$). This can be formally interpreted as using the Frechét derivative of the nonlinearity, which results in Newton's method. If we then repeat the modifications above, assuming B as the new unknown (i.e. substitution into the expansion and evaluation of the equation at the collocation points), we arrive at the exact same system as if starting from (2.40) and substituting $\underline{a}_{n+1} = \underline{a}_n + \underline{\delta a}$ and eventually defining the iteration procedure. The commutativity of linearization (or Frechét differentiation) and discretization, as we have just argued, has been mentioned in Golberg (1979), with a general treatment given by Ortega & Rheinboldt (1966).

The iteration scheme can thus be obtained to read

$$\begin{aligned} &\big(\underline{C}\,\underline{a}_n\big)^2 + (\underline{Cs})\,\underline{a}_n + (\underline{Cs})\,\underline{\delta a} + 2(\underline{C}\,\underline{a}_n)(\underline{C}\,\underline{\delta a}) + \Gamma + 1 = \underline{K}\,\underline{a}_n + \underline{K}\,\underline{\delta a} + \underline{f} \\ \Rightarrow\quad &\underline{\delta a} = \big[2\underline{C}\,\underline{a}_n\underline{C} + (\underline{Cs}) - \underline{K}\big]^{-1}\big(\underline{K}\,\underline{a}_n - \big(\underline{C}\,\underline{a}_n\big)^2 - (\underline{Cs})\,\underline{a}_n + \underline{f} - \Gamma - 1\big) \\ &\text{iterate}\ \ \underline{a}_n \to \underline{a}_n + \underline{\delta a}. \end{aligned} \tag{2.42}$$

Consistency of the scheme, as has been mentioned above, is clear at this point and hence we shall now make a few comments on convergence, whereas stability, although being a non-negligible topic in numerical analysis, will not be addressed, since a consistent method, if it converges, is stable (cf. also Lemma 3.28). A classical convergence proof will not be given, instead we heuristically show convergence via computing concrete solutions.

The projection method applied here transfers the unknowns from infinite dimensional function spaces to sequences of coefficients. Define, for the multi-index $i = (i_1, \dots, i_n)$ (see Remark 3.7) and $a := (a_i)_{|i|\geq 0}$

$$\ell^1(\mathbb{N}^n) := \{a : \|a\|_{\ell^1} = \sum_{|i|=0}^{\infty} |a_i| < \infty\} \ \text{ and } \ \ell^2(\mathbb{N}^n) := \{a : \|a\|_{\ell^2}^2 = \sum_{|i|=0}^{\infty} |a_i|^2 < \infty\},$$

then, from *Parseval's identity*, cf. Theorem 3.8, we have that if $f \in L^2_u$, the coefficients a_i, (uniquely) defining $\mathcal{P}_N f$, are in ℓ^2. In other words, if the components of the solution vector $\underline{a}$ from (2.40) (obtained via (2.42)) are square summable for all N, then the discretization converges in L^2_u. Furthermore, if the components are *absolute* summable, we even have

uniform convergence (cf. Remark 3.21). Necessarily, for $\underline{a} \in \ell^1$ (and consequently in ℓ^2), the a_i have to form a null sequence.

Yet another way is to utilize the fact that a sequence converges if and only if it is a Cauchy sequence. Thus, with respect to the L^2 norm and without loss of generality say $N > M$, then

$$\|\mathcal{P}_N B - \mathcal{P}_M B\|_u^2 = \|\sum_{i=M}^{N} a_i R_i\|_u^2 \overset{Parseval}{\leq} c \sum_{i=M}^{N} a_i^2, \tag{2.43}$$

such that, if we find an N_0, where, given an arbitrary $\epsilon > 0$, $\sum_{i=M}^{N} a_i^2 < \epsilon$ for all $N, M > N_0$, then (formally) the scheme converges in L^2. The same can be done in the supremum norm. Finally, one can assert that by having a consistent scheme, if it converges in the sense that calculated solutions for various N satisfy the criteria derived above, it converges to a (or the) solution of the original problem.

As for the iteration scheme (2.42), one can set $\delta B_N = \sum \delta a_i R_i$, where the iteration obviously terminates if $\delta B_N \equiv 0$ and since the only possible expansion of the zero function is all $\delta a_i = 0$, we have $\sum \delta a_i^2 = 0$. By being more precise, one has to add that $\underline{\delta a}$ depends on the iteration step n, such that $\underline{\delta a}_n$ forms a (null) sequence with respect to n and hence a further requirement could be the error $\underline{\delta a}_n$ to decrease *(strictly) monotonically* for all n greater than some n_0. Overall we impose the stopping criterion on $\underline{\delta a}$ at the iteration step n_s as

$$\sum_{i=0}^{N} \delta a_i^2 < \epsilon \quad \text{and} \quad \underline{\delta a}_n \overset{s}{\searrow} \text{ in } \ell^2,\ n_0 \leq n \leq n_s. \tag{2.44}$$

Note that one could alternatively consider the termination criterion in ℓ^1, which can slow down the iteration process, without gaining more accuracy for the solution $\underline{a}$. Setting ϵ in (2.44) at the order of magnitude of the machine precision, typically $\epsilon \approx 10^{-16}$, and using $N > N_0$ yields sufficiently accurate solutions.

Remark 2.19. A serious issue with the procedure (2.42) combined with the criteria (2.44) is the choice of the initial vector $\underline{a}_0$. Since the iteration is in fact a Newton method, the initial guess has to be close (in some sense) to the solution in order for the process to converge.
The method developed in Powell (1970) remedies this disadvantage. Therein it was also shown, that the proposed hybrid method converges (under certain requirements) even if the elements of the Jacobian matrix are calculated approximately. The routine HYBRD1, developed in Moré et al. (2000) and implemented in the *NAG* packages mentioned above, utilizes Powell's method, such that the user only has to provide the equation system in the form of, e.g. (2.40). Due to the convergence proof for the iteration given in Powell (1970), we have an independent test for solutions obtained via (2.42).

In the following we will show the results of some concrete computations for the problems (2.33) and (2.30) using Powell's hybrid method as described in Remark 2.19 and the iteration scheme (2.42). The notions *upper branch* and *lower branch* shall refer to the bifurcation diagram in Figure 4 (for more details on this subject see, e.g. Stewartson et al. (1982) and

Brown & Stewartson (1983) for the two-dimensional case, as well as Duck (1989) and Braun & Kluwick (2002) for the three-dimensional problem).

Say $\Gamma = 2$, $g \equiv 0$ and take $\underline{C}, \underline{K}$ from (2.39). Naturally, one starts with the initial vector $\underline{a}_0 \equiv 0$ to see whether the iteration converges. Table 1 shows the ℓ^2 norm of $\underline{\delta a}_n$ for various N (meeting the criteria (2.44)). The (almost perfect) quadratic convergence of the scheme (2.42) is striking (as expected, since it is essentially a Newton method). In addition, this means that $\underline{a} \equiv 0$ is already close to the solution.

n / N	10	20	40	160
1	2.81×10^{-1}	2.82×10^{-1}	2.83×10^{-1}	2.83×10^{-1}
2	2.42×10^{-2}	2.41×10^{-2}	2.41×10^{-2}	2.41×10^{-2}
3	2.22×10^{-4}	2.24×10^{-4}	2.24×10^{-4}	2.24×10^{-4}
4	1.82×10^{-8}	1.87×10^{-8}	1.87×10^{-8}	1.87×10^{-8}
5	1.22×10^{-16}	1.27×10^{-16}	1.27×10^{-16}	1.27×10^{-16}

Table 1: Iteration error $\|\underline{\delta a}_n\|_{\ell^2}^2$ for various N with initial guess $\underline{a}_0 \equiv 0$.

One can also infer from Table 1 that the iteration not only converges per se (in terms of $\exists n_s \ll \infty$) but also with $N \nearrow$ (considering the lines in the table as sequences in terms of N). Thus we have established the convergence of the iteration.

As for the solution $\underline{a}_{n_s}$ we have the criteria of convergence (as a Cauchy sequence cf. (2.43)) in the ℓ^1 and ℓ^2 norm, see Table 2, and the uniqueness of the coefficients themselves, which are given in Table 3.

N / *norm*	$\|\cdot\|_{\ell^2}^2$	$\|\cdot\|_{\ell^1}$
10	0.48170	1.46535
20	0.48280	1.50366
40	0.48353	1.53942
80	0.48372	1.55689
160	0.48377	1.56593
320	0.48378	1.57039

Table 2: ℓ norms of the solution $\underline{a}_{n_s}$ for various N.

With the difference $\|\underline{a}_{2N} - \underline{a}_N\|_\ell$ (as elements of a Cauchy sequence in N) from the results in Table 2 we have

norm / N	10	20	40	80	160
$\|\cdot\|_{\ell^2}^2$	1.1×10^{-3}	7.3×10^{-4}	1.9×10^{-4}	5×10^{-5}	1×10^{-5}
$\|\cdot\|_{\ell^1}$	3.8×10^{-2}	3.6×10^{-2}	1.7×10^{-2}	9×10^{-3}	4.5×10^{-3}

,

where one can see the slower convergence in the ℓ^1 norm.

N	a_0	a_1	a_2	a_3	a_4	a_{10}	a_{20}
10	−0.5930	−0.1893	0.2123	0.1077	0.1630	0.0065	–
20	−0.5902	−0.1879	0.2161	0.1094	0.1680	0.0113	0.0009
40	−0.5896	−0.1878	0.2173	0.1094	0.1693	0.0128	0.0043
80	−0.5894	−0.1878	0.2177	0.1095	0.1696	0.0132	0.0047
160	−0.5894	−0.1878	0.2177	0.1095	0.1697	0.0133	0.0048
320	−0.5894	−0.1878	0.2178	0.1095	0.1697	0.0133	0.0048

Table 3: Some coefficients of the solution $\underline{a}_{n_s}$ for various N.

For small i the coefficients of the solution (a_i) do not necessarily form a monotone decreasing sequence (as can be seen in Table 3), as one would infer from Theorem 3.15. The results therein are of asymptotic kind, meaning that they only hold for $i \gg 1$. Say $i \geq 20$, cf. Table 3, then the coefficients satisfy $|a_i| \leq 10^{-2}$ and therefore they do not contribute to the characteristics of the solution B_N, but the assertion in Theorem 3.15 becomes applicable. Furthermore, we have that the graphs of $A = A(x)$ plotted in Figure 3 using $N = 40$ polynomials are practically indistinguishable from the graphs using $N = 320$ polynomials, which also becomes obvious when considering the change in the leading coefficients from Table 3. (Even using only 10 polynomials yields an acceptable solution). For the sake of completeness we mention that the Powell method yields the exact same coefficients as the Newton iteration.

Previous works, e.g. Stewartson et al. (1982), suggest at least one other solution for certain values of Γ in (2.33). To obtain such solutions we make an educated guess for the starting vector $\underline{a}_0$ for the iteration scheme (2.42), say $a_0 = -1$, $a_1 = -1$ and $a_i = 0$ otherwise. As mentioned in Remark 2.19 the Newton algorithm is very sensitive with respect to the initial guess and therefore imposing this $\underline{a}_0$ did not lead to convergence (independently of N). The Powell algorithm, on the other hand, did stop at an acceptable solution using $N = 40$. Then, taking this solution as the starting vector for $N = 10, \ldots, 320$, we are again able to show convergence of the Newton iteration, see Table 4.

N / *norm*	$\|\cdot\|_{\ell^2}^2$	$\|\cdot\|_{\ell^1}$
10	5.31191	5.80099
20	5.33291	6.30836
40	5.34614	6.35881
80	5.34667	6.37136
160	5.34681	6.37918
320	5.34684	6.38336

Table 4: ℓ norms of the (lower branch) solution $\underline{a}_{n_s}$ for various N.

By calculating again the difference $\|\underline{a}_{2N} - \underline{a}_N\|_\ell$, as done above, we have

$norm\ /\ N$	10	20	40	80	160
$\|\cdot\|_{\ell^2}^2$	2.1×10^{-2}	1.3×10^{-2}	5.3×10^{-4}	1.4×10^{-4}	3×10^{-5}
$\|\cdot\|_{\ell^1}$	5.1×10^{-1}	5×10^{-2}	1.3×10^{-2}	7.8×10^{-3}	4.2×10^{-3}

.

Thus, convergence for the second solution is also (heuristically) shown and we shall provide again the leading coefficients, see Table 5.

$N\ /\ a_i$	a_0	a_1	a_2	a_3	a_4	a_{10}	a_{20}
10	−1.5818	−1.0007	0.6846	0.6986	0.7324	−0.1461	−
20	−1.5726	−1.0308	0.6244	0.6655	0.7285	−0.1150	0.0337
40	−1.5742	−1.0324	0.6276	0.6675	0.7299	−0.1108	0.0288
80	−1.5741	−1.0323	0.6279	0.6675	0.7302	−0.1104	0.0291
160	−1.5740	−1.0323	0.6280	0.6675	0.7303	−0.1103	0.0293
320	−1.5740	−1.0323	0.6280	0.6675	0.7303	−0.1103	0.0293

Table 5: Some coefficients of the (lower branch) solution $\underline{a}_{n_s}$ for various N.

Comparing Tables 3 and 5 it is clear that the lower branch solutions need more polynomials to achieve a certain accuracy. This is also obvious from the shape of the graphs in Figure 3. Considering the leading coefficients with respect to the above mentioned Cauchy criterion, Table 5 also shows that for a given ϵ the appropriate N_0 is definitely higher in case of lower branch than for upper branch solutions.

To show that the Newton method is in fact able to obtain lower branch solutions, we use a_0 to a_4 from Table 5 and $a_i = 0$ otherwise as the staring vector and run the iteration for the usual N. Interestingly, the convergence depends strongly on these $a_0, \ldots, a_4$ in the initial guess, rather than on N, meaning that if we use only a_0, a_1 from Table 5, the iteration does not stop, independently of N. Table 6 depicts the difference in number of iteration steps and ℓ^2 norms of $\underline{\delta a}_n$ between taking a_0 through a_4 from Table 5 at $N = 40$ and taking a_0 through a_3 at $N = 80$. This again emphasizes the sensitivity of the Newton iteration.

Figure 3 shows upper and lower branch solutions of problem (2.33) for $\Gamma = 2$, found by using $N = 40$ polynomials for B_N in (2.38), where the coefficients are taken from Tables 3 and 5, and finally A from (2.35).

The existence of multiple solutions for various Γ has been studied for example in Stewartson et al. (1982) and Brown & Stewartson (1983) using finite difference and trapezoidal quadrature methods. The obvious part of the bifurcation (sometimes called fundamental curve) is plotted in Figure 4, which is best obtained by incrementally increasing Γ, taking the previous solution as the initial guess for $\Gamma \pm \Delta\Gamma$, starting somewhere on the upper and lower branch, respectively.

n / N	40	80
1	8.82×10^{-1}	1.26×10^{0}
2	1.37×10^{-1}	2.42×10^{-1}
3	7.24×10^{-3}	2.59×10^{-2}
4	4.54×10^{-5}	5.71×10^{-4}
5	4.06×10^{-10}	6.17×10^{-8}
6	6.83×10^{-20}	1.28×10^{-15}
7	–	9.79×10^{-31}

Table 6: Iteration error $\|\underline{\delta a}_n\|_{\ell^2}^2$ for $N = 40$ and $N = 80$ with different initial guesses.

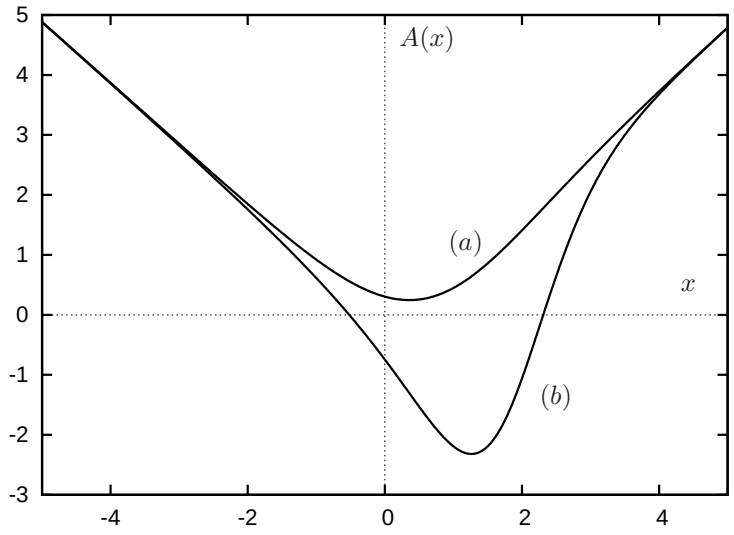

Figure 3: Solutions $A(x)$ of (2.33) for the upper (a) and lower (b) branch at $\Gamma = 2$.

From the fundamental curve one can claim that there exists a certain value Γ_c, above which no (real) solutions can be found. Interestingly, with the approximations used (i.e. polynomial collocation), the Newton and the Powell iterations do not *diverge* for Γ greater than this maximum value. The error $\underline{\delta a}$, on the other hand, does not satisfy (2.44) and therefore convergence can also not be observed. The MINPACK scheme provided in Moré et al. (2000), employing the Powell algorithm, does stop somehow, suggesting the existence of possible local minima. As described in Powell (1970), procedures relaxing the iteration steps can be prone to stop at such minima. But eventually, from the non-decreasing error of the Newton method, we assert the non-existence of solutions.

Remark 2.20. Note the negative values of A in Figure 3(b) indicate the existence of a separation bubble, whereas the upper branch solution at $\Gamma = 2$ is fully attached. This already shows the sensitivity of the marginal separation steady states regarding potential bubble burst, since it is not clear which of the multiple solutions is physically realized when setting Γ at a certain value. We will show, on the other hand, in Section 2.3.1 that the upper

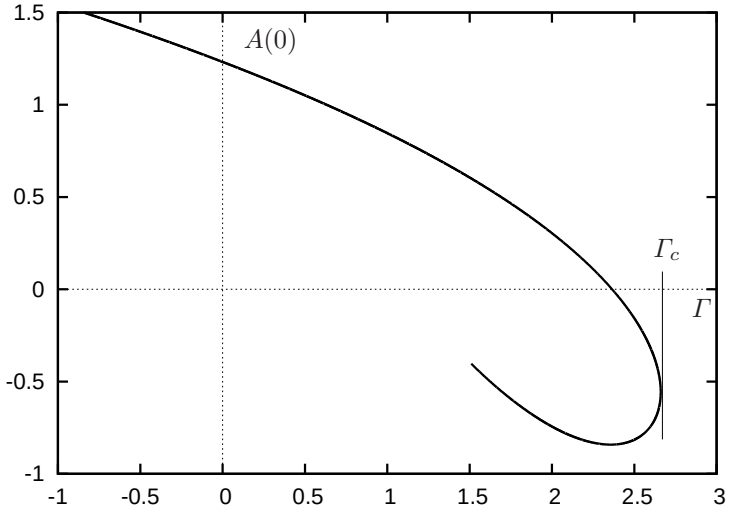

Figure 4: Part of the bifurcation occurring in (2.33), shown as $A(0)$ versus Γ

branch is stable in some sense, rendering the lower branch unstable. Moreover, from Figure 4 one can claim situations with separation bubbles of a certain size to be prone to burst.

Overall, we have presented enough evidence to claim convergence for the iteration procedure and the approximation using the polynomial expansion (2.38) for $N_z = 0$. Since the only difference when computing solutions to (2.41) lies in the matrices $\underline{C}$ and $\underline{K}$, it is now sufficient to obtain similar tables as Table 2 and Table 3 when solving (2.41) to claim the convergence of the approximation (2.38) in the three-dimensional case. In addition, this shows the applicability of the algorithms from Section 3.3.2 yielding $\underline{K}$.

For comparison reasons (i.e. with results presented in Braun & Kluwick (2002)), we show solutions of equation (2.41) for $\Gamma = 1$, including a forcing term of the form

$$g(x,z) = -\frac{\lambda}{2\pi}\,\mathcal{J}_{-\infty}^{1/2}\,\mathcal{R}^1\,[\partial_{\xi_1}^3 + \partial_{\xi_1}\partial_{\xi_2}^2]h(x,z), \tag{2.45}$$

with h representing the hump at the surface (see Section 2.1) given as

$$h(x,z) = 4(1-x^2)^3\,\mathbb{1}_{[-1,1]}(x)\,\exp(-z^2).$$

The formula for g is due to the original argument $A - h$ in the fundamental equation (2.28), where such a local alteration of the surface has to be smooth enough, or say, has to satisfy a certain boundedness regarding its curvature, such that it leaves the boundary layer equations invariant. To calculate $g(x_j, z_l)$, i.e. evaluating the forcing at the collocation points, we say

$$g(x,z) \approx \sum_{i,j=0}^{N} b_{ij} R_i(x) R_j(z),$$

and compute b_{ij} using Lemma 3.19. From Theorems 3.23 and 3.24, the smoothness of h and its rapid decay, we assert spectral convergence of the approximation above (cf. Remark 3.16), meaning very view polynomials are sufficient for the necessary smallness of the discretization error of the forcing term to be negligible. To obtain the forcing matrix $\underline{g}$ from (2.45), one can now use the same algorithm yielding $\underline{K}$ (as mentioned above).

In virtue of Section 2.3, Remark 2.25, the unknown B shall be approximated by a weighted Chebyshev series

$$B_N(x,z) = \frac{1}{\sqrt{1+x^2}} \sum_{i=0}^{N_x} \sum_{k=0}^{N_z} a_{ik} R_i(x) R_k(z), \tag{2.46}$$

where, due to the z-symmetry of the perturbation g, we only need to consider *even* k in B_N. Table 7 depicts the summability of the coefficients (a_{ik}) in (2.46) and Figure 5 shows the according solution of (2.41) including the forcing term (2.45).

N / *norm*	$\|\cdot\|_{\ell^2}^2$	$\|\cdot\|_{\ell^1}$
40/8	1.86254	5.09231
60/20	1.83234	5.26795
60/30	1.82945	5.32358
80/40	1.87524	5.38463
100/24	1.84692	5.65930
120/40	1.86450	6.11253

Table 7: ℓ norms of the solution (a_{ik}) for various $N = (N_x, N_z)$.

The somewhat irregular convergence behavior in Table 7, in contrast to the two-dimensional case, cf. Table 2, stems from the varying resolution of the perturbation g (i.e. the distribution of the collocation points), and the fact that the gradients and curvatures of the solution with respect to the x-coordinate change stronger than in z direction, see Figure 5. For this reason we compared the graphs of $A(x,z)$, using $(N_x, N_z) = (80, 40)$, $(100, 24)$ and $(120, 40)$, with the result presented in Braun & Kluwick (2002), which all turn out to be virtually indistinguishable.

Another way to claim summability, is to consider the convergence and uniqueness of the leading coefficients (see Table 8). Due to the weighted expansion (2.46) and the far field behavior of B, cf. (2.35), one cannot expect a fast decrease of $|a_{ik}|$ as $i, k \to \infty$ (as given in Theorem 3.15). Hence, to assert absolute summability (i.e. to have uniform convergence), much more polynomials are needed than given in Table 7. Square summability, on the other hand, can be easily seen from Figures 6, 7 and 8.

Since the computational costs in the three-dimensional case grow quadratically compared to its two-dimensional equivalent, we consider the connection between these two to show the validity of the approximation algorithm from Section 3.3.2. As mentioned earlier, setting $N_z = 0$ in the expansion (2.38) yields its two-dimensional analogue, such that solving the three-dimensional case (equation (2.41) with its according scheme) for $N_z = 0$ must give the

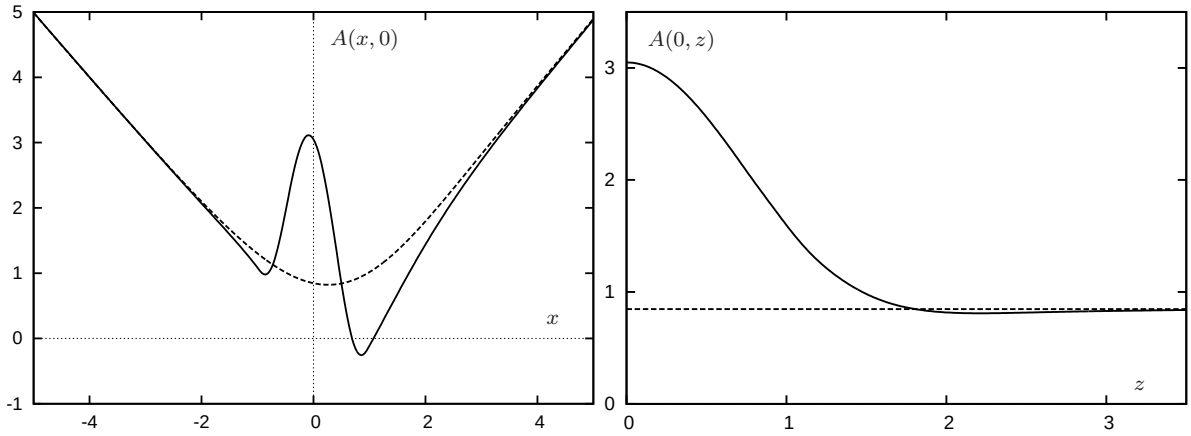

Figure 5: Solid: solution of (2.30) with $(N_x, N_z) = (120, 40)$, dashed: solution of (2.33) with $N = 40$.

N / a_{ik}	a_{00}	a_{02}	a_{10}	a_{12}	a_{20}	a_{22}	a_{30}
40/8	−0.37009	−0.12296	−0.44428	0.23478	−0.74872	0.54053	0.17932
60/20	−0.34859	−0.12598	−0.44340	0.23391	−0.74536	0.53794	0.17579
60/30	−0.34588	−0.12480	−0.44346	0.23359	−0.74583	0.53628	0.17540
80/40	−0.37393	−0.10902	−0.44265	0.23619	−0.75279	0.53761	0.18176
100/24	−0.35751	−0.12292	−0.44483	0.23509	−0.74742	0.54022	0.17536
120/40	−0.36665	−0.11541	−0.44694	0.23562	−0.75327	0.53473	0.17551

Table 8: Some coefficients of the solution (a_{ik}) for various $N = (N_x, N_z)$.

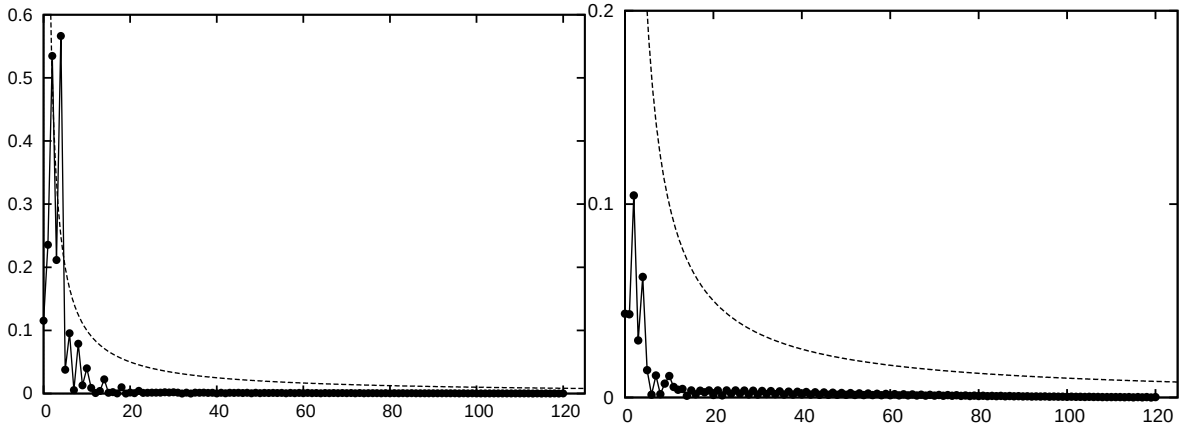

Figure 6: Solution $|a_{ik}|$ at $k = const$, $i = 0, \ldots, 120$ for $(N_x, N_z) = (120, 40)$. Left: $k = 0$, right: $k = 2$, dashed line: $1/i$.

exact same solution as when solving (2.39), cf. Remark 2.14. Table 9 depicts the ℓ norms for (a_{i0}), $i = 0, \ldots, N_x$ as the solution of (2.41) at $\Gamma = 2$ with $g \equiv 0$.

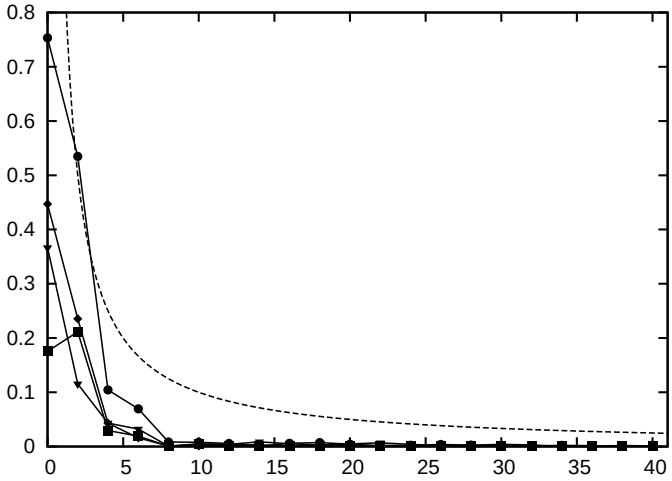

Figure 7: Solution $|a_{ik}|$ at $i = const$, $k = 0, \ldots, 40$ $(even)$, for $(N_x, N_z) = (120, 40)$. Triangles: $i = 0$, diamonds: $i = 1$, circles: $i = 2$, squares: $i = 3$, dashed line: $1/i$.

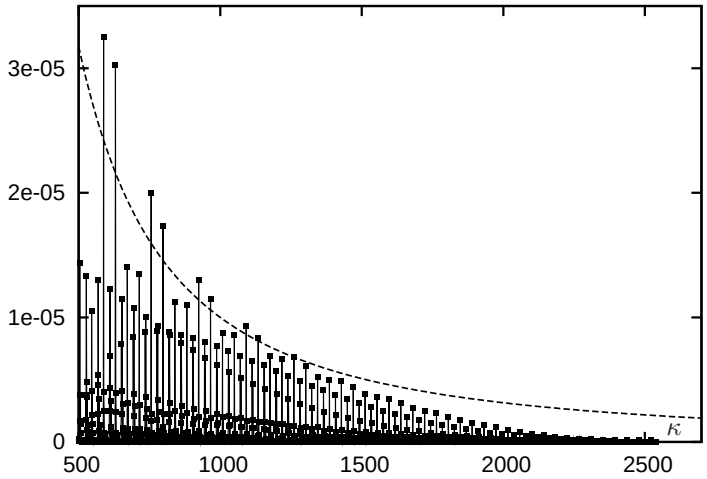

Figure 8: Solution $|a_{ik}|^2$ vs. $\kappa = i\,N_z + k$, $k = 0, \ldots, 20$, $i = 0, \ldots, 120$, for $(N_x, N_z) = (120, 40)$, dashed line: $1/\kappa$.

N_x / *norm*	$\|\cdot\|_{\ell^2}^2$	$\|\cdot\|_{\ell^1}$
20	1.41689	2.22812
40	1.43700	2.26451
80	1.40001	2.27665
160	1.41831	2.40480

Table 9: ℓ norms of the solution (a_{i0}) for various N_x and $N_z = 0$.

For convergence being not as obvious here as, for example, in Table 2, we further compare the leading coefficients for various N_x to the according solution of (2.39) using $N = 160$ polynomials, see Table 10 and Figure 9.

Caveat: Although the solution $A = A(x)$ here is exactly the same as in Figure 3, the coefficients are different due to the weight used in (2.46).

N_x / a_{ik}	a_{00}	a_{10}	a_{20}	a_{30}	a_{50}	a_{60}
20	−1.05158	−0.38061	−0.37543	0.06666	0.11112	0.04507
40	−1.06006	−0.38033	−0.37566	0.07123	0.11676	0.04679
80	−1.04417	−0.37623	−0.37642	0.06875	0.11470	0.04443
100	−0.99730	−0.37386	−0.38242	0.05415	0.10167	0.03698
150	−1.04909	−0.38316	−0.38214	0.06203	0.10826	0.04002
160	−1.04993	−0.37814	−0.37720	0.06969	0.11593	0.04536

Table 10: Convergence of some coefficients of the solution (a_{i0}) for various N_x with $N_z = 0$ in comparison to the solution from the two-dimensional algorithm at $N = 160$ (last line).

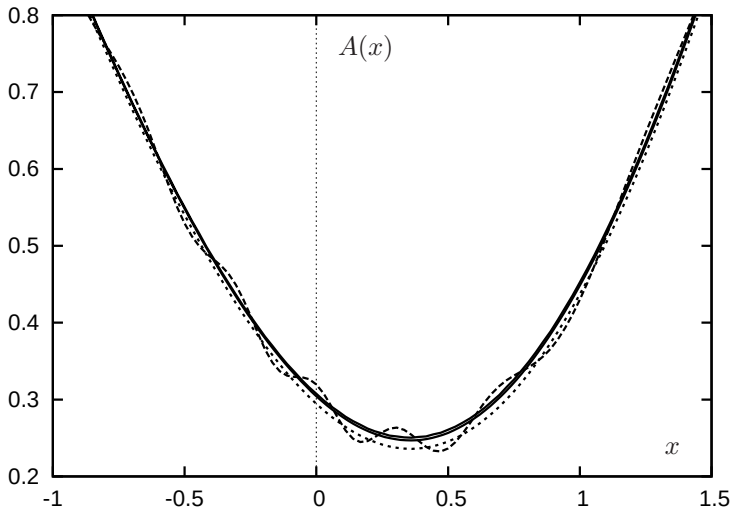

Figure 9: Solution $A(x)$ from three-dimensional scheme, with $N_z = 0$ and $N_x = 20$ (dashed), $N_x = 40$ (dotted), $N_x = 150$ (solid, almost indistinguishable from $A(x)$ (solid) as in Figure 3(a) gained with $N = 40$).

Finally, by having established all necessary convergence properties of the Chebyshev expansions, we shall emphasize the small equation systems (when compared to the deployed schemes in previous works) needed to obtain the required accuracy of the solutions. This then becomes more important in Section 2.3, where the above algorithms will be used for the spatial discretizations of time dependent problems, hence making it possible to obtain *fast* explicit and implicit Euler procedures.

Remark 2.21. In Fromme & Golberg (1979) it was mentioned that from a converging Galerkin scheme one can assert the general existence and uniqueness of a solution (in terms of function spaces). Since collocation methods can be viewed as Galerkin methods, with the inner product integrals approximated by quadrature schemes, cf. Remark 3.40, we can formally claim existence of solutions to problems (2.33) and (2.30) in a subspace of L^2_u (with respect to the differentiability requirements).

Remark 2.22. An interesting test case is to set $g \equiv 0$ in (2.41) and to start the iteration procedure with an initial guess independent of z. According to Remark 2.14 for such functions the three-dimensional problem reduces to its two-dimensional equivalent. This means, if the approximation is done correctly and accurately, the iteration, although performed for (2.41), should yield the same solution as if done for (2.39). This has been tested with $N_x = 100$ and various N_z and initial guesses for the Newton scheme, yielding satisfying results, i.e. steady (upper branch) solutions, as shown in Figure 3. If Γ is chosen near its critical value Γ_c (cf. Figure 4) it is highly likely that the two-dimensional solutions will not be obtained by the three-dimensional scheme as the calculations in Braun & Kluwick (2002) demonstrate.

2.3 Cauchy Problems

In this section we deal with the main objective of the present treatise, namely Cauchy problems, their well-posedness and the finite time blow-up of certain solutions. Mentioned in Remark 2.8, initial conditions imposed on these Cauchy problems are yet to be determined and, as we will demonstrate in Section 2.3.1, may be subject to some restrictions, additionally to the physical meaning they shall have, as will be seen in Section 2.3.3.

The initial value problem per se, including a far field condition, is given via (2.28) and (2.29). In complete accordance to Section 2.2 the interaction pressure can be substituted into (2.28), thus yielding one equation governing time dependent solutions to the steady states shown in the previous section. Given the polar coordinates $(x, z) \to (r, \phi)$, $A = A(x, z, t)$ shall satisfy following problem in $\mathbb{R}^2 \times [0, T]$

$$A^2 - x^2 + \Gamma = \frac{\lambda}{2\pi} \int_{-\infty}^{x} \frac{1}{(x-s)^{1/2}} \int_{\mathbb{R}^2} \frac{\partial^3_{\xi_1} + \partial_{\xi_1}\partial^2_{\xi_2}}{|(s,z) - (\xi_1, \xi_2)|} A(\xi_1, \xi_2, t) d\xi_1 d\xi_2 \, ds - \\ -\gamma \int_{-\infty}^{x} \frac{1}{(x-s)^{1/4}} \partial_t A(s, z, t) \, ds + g(x, z, t) \tag{2.47}$$

$$A(x, z, 0) = A_0(x, z), \text{ in } \mathbb{R}^2$$
$$A(x, z, t) \sim c(\phi) r \quad \text{as} \quad r \to \infty, \text{ in } [0, T].$$

The far field condition expressed in Cartesian coordinates can obviously be taken from the steady problem, i.e. (2.31) and g again summarizes the forcing contributions from the proposed hump and/or suction/blowing devices (see Section 2.1).

As done in Section 2.2 we rewrite the equation in (2.47) by identifying the Abel operators and the Riesz potential as

$$A^2 - x^2 + \Gamma = \frac{\lambda}{2\pi}\,\mathcal{J}_{-\infty}^{1/2}\,\mathcal{R}^1\left([\partial_x^3 + \partial_x\partial_z^2]A\right)(x,z,t) - \gamma\,\mathcal{J}_{-\infty}^{3/4}\big(\partial_t A\big)(x,z,t) + g(x,z,t). \quad (2.48)$$

The planar problem in $\mathbb{R}\times[0,T]$ has originally been derived in Ruban (1982) and Smith (1982) (without including the perturbation g) and is governed by

$$A^2 - x^2 + \Gamma = \lambda\,\mathcal{J}_{\infty}^{1/2}\big(\partial_x^2 A\big)(x,t) - \gamma\,\mathcal{J}_{-\infty}^{3/4}\big(\partial_t A\big)(x,t) + g(x,t), \quad (2.49)$$

subject to $A(x,0) = A_0(x)$ and $A(x,t) = O(|x|)$ as $|x|\to\infty$.

Remark 2.23. Similarly to the steady case, problem (2.47) can be transformed into (2.49) by assuming a solution to be independent of z. Also, as described in Remark 2.11 the global, (z-symmetric) three-dimensional set-up studied in Duck (1990) contains the same operators, such that the difference lies only in the explicit z dependence of the left hand side.

Remark 2.24. To utilize what has been presented in Remarks 2.12 and 2.13 in terms of function spaces for solutions, we consider (as is often done in the theory of evolution equations) the problem (2.48) as an ordinary differential equation with values in some Banach space X, i.e. $A : [0,T]\to X$ is continuously differentiable and the right hand side operators map their domain onto X as well. Obviously, as mentioned in Remark 2.12, X is the space of bounded, continuous functions with an existing limit at infinity. Normally, one also requires the initial condition A_0 to lie in the domain of the operators, but under certain conditions such assumptions can be weakened (cf. the definition of a classical solution of abstract Cauchy problems (2.62)). Here we will confine the set of initial conditions to the domain of the integro-differential operators, i.e. three times continuously differentiable functions with at most linear growth at infinity. (In Section 2.3.3 we will be even stricter and only allow for physically meaningful initial values.)

The properties of the operators in (2.48) on the given function spaces (cf. Remark 2.13) can only provide a quite general insight into the structure of the problem and possible solutions. But since we are also interested in certain local qualitative and quantitative characteristics of solutions (relating to specific initial conditions), a numerical treatment of (2.47) is needed.

As done in Section 2.2 we want to use a polynomial collocation method in spatial coordinates, which is usually set up in, e.g. $(C_l(\mathbb{R}^2), \|\cdot\|_\infty)$, for all $t>0$ and hence needs an at least bounded unknown function, cf. Remark 2.15. As shown in (2.35) it is sufficient to subtract the linear growth $\sqrt{1+x^2}$, obtaining the new unknown $B = B(x,z,t)$. Thus, by expanding B into a Chebyshev series with respect to (x,z) we inherit all the convergence properties presented in the study of the steady problem in Section 2.2.

Reformulating the Cauchy problem in terms of B gives

$$\begin{aligned} B^2 + 2\sqrt{1+x^2}B + \Gamma + 1 =& \frac{\lambda}{2\pi}\,\mathcal{J}_{-\infty}^{1/2}\,\mathcal{R}^1\left([\partial_x^3 + \partial_x\partial_z^2]B\right)(x,z,t) - \\ & - \gamma\,\mathcal{J}_{-\infty}^{3/4}\big(\partial_t B\big)(x,z,t) + f(x) + g(x,z,t) \\ B(x,z,0) = B_0(x,z) < \infty \;\; & \text{in } \mathbb{R}^2 \\ \left.\begin{array}{ll} B(x,z,t) < \infty & \text{as } (x^2+z^2) \to \infty \\ B(x,\cdot) = O(|x|^{-1}) & \text{as } |x| \to \infty \end{array}\right\} \;\; & \text{in } [0,T], \end{aligned} \tag{2.50}$$

with f taken from (2.37), such that

$$B(x,z,t) \approx B_N(x,z,t) = \frac{1}{\sqrt{1+x^2}} \sum_{i=0}^{N_x} \sum_{k=0}^{N_z} a_{ik}(t) R_i(x) R_k(z), \tag{2.51}$$

cf. (2.46), where the coefficients $a_{ik} = a_{ik}(t)$ are now functions of time. Considering z-symmetric disturbances g, we only need to sum over even polynomials in z, i.e. k is even. By setting $N_z = 0$ we obtain the expansion for the unknown B in its two-dimensional version.

Remark 2.25. The weight function $(1+x^2)^{-1/2}$ stems from the operator $\mathcal{J}_{-\infty}^{3/4}$ acting on the time derivative and thus it is only necessary in the unsteady problem. As shown in Section 3.3.2, Remark 3.51, a weight function is needed in order to make the Abel integral acting on rational Chebyshev polynomials exist. Furthermore, if the function to be expanded satisfies $B \sim |x|^a$ as $|x| \to \infty$, a being less than the negative exponent of the operator, then the weight is given as $(1+x^2)^{-b}$, $2b + a \leq 0$, $2b$ being greater than the exponent of the operator. Here this means, for $B(x,\cdot) = O(|x|^{-1})$, $a = -1 < -3/4$, such that $3/8 < b \leq 1/2$. Hence, $b = 1/2$ is an appropriate choice, although one has to expect slower convergence rates of the sums in (2.51), since they now have to assume a non-zero constant at infinity, cf. e.g. Theorems 3.16 and 3.17.

Substituting the expansion (2.51) into problem (2.50) yields (using the same notation as in (2.39) through (2.41))

$$\big(\underline{Cw}\,\underline{a}(t)\big)^2 + 2\underline{C}\,\underline{a}(t) + \Gamma + 1 = \underline{K}\,\underline{a}(t) - \underline{D}\,\frac{d}{dt}\underline{a}(t) + \underline{f} + \underline{g}, \tag{2.52}$$

where $\underline{Cw}$ denotes the matrix gained from (2.51) with $R_i(x)$ replaced by $(1+x^2)^{-1/2}R_i(x)$. The same has been done for $\underline{K}$, as explained in detail in Section 3.3.2, i.e. equation (3.114) and what follows. The set-up of the matrix $\underline{D} = \gamma R_k(z_l)\mathcal{J}_{-\infty}^{3/4}\big((1+\xi^2)^{-1/2}R_i(\xi)\big)(x_j)$ is also shown in Section 3.3.2, i.e. equation (3.106) and Remarks 3.52 through 3.55.

As for the two-dimensional equivalent of (2.52) with $N_z = 0$, it is obvious that the structure in principle remains the same, and the only relevant changes occur in $\underline{K}$, which is built as described in (3.113) and Remark 3.56, and in $\underline{D}$, where $R_k(z_l)$ becomes $R_0 = 1$.

To fully discretize (2.52) one still needs to employ an approximation for the time derivative $\frac{d}{dt}$, but before doing so, we have to consider the aspect of *well-posedness* of (2.47), which is done in the following section.

2.3.1 Ill-Posedness and Regularized Dynamics

For (nonlinear) Cauchy problems (involving partial differential equations) the question of *existence and uniqueness* is governed by the *Cauchy-Kowalewsky theorem*, if the Cauchy data, given on an analytic, noncharacteristic hypersurface, and the coefficients are analytic.

Another, maybe even more important question, was raised in Hadamard (1923). He argued that approximating non-analytic data by analytic ones is not so much about how little the data is altered, but whether the solutions would change much. Hadamard presented his famous example of the Cauchy problem for the Laplace equation in Zürich 1917, which demonstrates that a third requirement (besides existence and uniqueness) should be imposed on Cauchy problems - *continuity with respect to given functions*. Later this was paraphrased as *continuous dependence on the data*, which might include coefficients of the differential operators, and is now commonly known as *well-posedness*.

Furthermore, due to the connection of partial differential equations and the according initial-boundary-value problems to physics and mechanics, Hadamard (1923) subsequently assumed, as a rule, every treated Cauchy problem to be "correctly set", i.e. well-posed. As a definition, *ill-posed* problems are regarded as those, violating at least one of the three requirements stated above.

Petrowsky (1937) defines the Cauchy problem for u, with initial data ϕ, to be correctly set if (formally speaking)

i) for some initial data $\bar{\phi}$, which differs only by an ϵ from ϕ, there exists only one solution $\bar{u}$ and

ii) for every η there exists a δ, such that the difference of $\bar{u}$ and u is less than η, if $\bar{\phi}$ is as close as δ to ϕ.

He then proved for general, quasilinear Cauchy problems, by means of Fourier expansions (without terming them as such), that the Fourier coefficients $a_i(k)$ of the solution satisfy

$$\sum_{i,k} |a_i(k)(t)|^2 \leq e^{ct}\big(c_1 \sum_{i,k} |a_i(k)(0)|^2 + c_2 \int_0^t \sum_{i,k} |f_i(k)(\tau)|^2 d\tau\big), \tag{2.53}$$

where $f_i(k)$ are the Fourier coefficients of a (if present) inhomogeneity, and c, c_1, c_2 are positive constants. Using Parseval's identity and the L^2 norm (which Petrowsky stated as the "growth rate of an integral") yields the well-known result for well-posed systems of evolution problems

(in Banach spaces)

$$\sum_i \|u_i(t)\|^2 \leq e^{ct}\big(c_1 \sum_i \|u_i(0)\|^2 + c_2 \int_0^t \sum_i \|f_i(\tau)\|^2 d\tau\big).$$

Item *ii*) in Petrowsky's requirement for well-posedness, with the set of initial data $u_i(0)$, is now obvious from such an inequality.

Next we apply this fact to a homogeneous partial differential initial value problem to derive a very simple necessary (and in some cases sufficient) condition for well-posedness. Given a function f and variables $k, x \in \mathbb{R}^n$, with the usual inner product denoted by $\langle \cdot, \cdot \rangle$, the Fourier transform shall be defined as

$$\mathcal{F}(f)(k) := \int_{\mathbb{R}^n} f(x) e^{-i\langle k,x\rangle} dx. \tag{2.54}$$

Consider a linear, partial differential evolution equation of the form

$$\partial_t u = \mathcal{P}(\partial_x) u \tag{2.55}$$

subject to some initial condition u_0, where $\mathcal{P}$ denotes a polynomial (of arbitrary, finite degree) with real, constant coefficients. As it is well-known for smooth functions f (with sufficient decay)

$$\mathcal{F}(\partial_x^m f) = (ik)^m \mathcal{F} f, \tag{2.56}$$

where for the multi-index $m = (m_1, \ldots, m_n) \in \mathbb{N}^n$, $\partial_x^m = \prod_j \partial_{x_j}^{m_j}$ and $(ik)^m = \prod_j (ik_j)^{m_j}$. Formally, applying the Fourier transform (with respect to x) to (2.55) yields (denoting $\mathcal{F}u = \hat{u}$)

$$\partial_t \hat{u} = \mathcal{P}(ik)\hat{u} \quad \Rightarrow \quad \hat{u}(k,t) = \hat{u_0}(k) e^{\mathcal{P}(ik)t}.$$

Hence, a condition for well-posedness can be readily deduced from (2.53), i.e.

$$\sum_k |\hat{u}(k,t)|^2 = \sum_k |\hat{u_0}(k)|^2 |e^{\mathcal{P}(ik)t}|^2 \leq c_1^2 e^{2c_2 t} \sum_k |\hat{u_0}(k)|^2, \quad \forall t$$

if and only if

$$|e^{\mathcal{P}(ik)t}| \leq c_1 e^{c_2 t} \quad \forall k, t, \tag{2.57}$$

or, necessarily,

$$\Re\, \mathcal{P}(ik) \leq const., \quad \forall k \in \mathbb{R}^n, \tag{2.58}$$

meaning that the real part of the complex polynomial has to be *bounded from above*.

Remark 2.26. For the sake of simplicity consider the polynomial in one dimension. Writing it explicitly as

$$\mathcal{P}(ik) = a_0 + a_1(ik) + a_2(ik)^2 + \cdots + a_m(ik)^m$$

shows that for the boundedness condition (2.58) it is sufficient to check only the boundedness for the highest derivative (of which the real part is non-zero).

Replacing f by $e^{i\langle k,x\rangle}$ in (2.56), without using the Fourier transform, shows that changing from differentiation to multiplication also holds for these exponential functions with the exact same *Fourier multipliers*, i.e.

$$\partial_x^m e^{i\langle k,x\rangle} = (ik)^m e^{i\langle k,x\rangle}. \tag{2.59}$$

Originated from physical problems the notion of *dispersion relations* is often used to consider the behavior of disturbances in evolution equations, i.e. introducing the perturbation $\tilde{u}(x,t) = e^{\lambda t}e^{i\langle k,x\rangle}$, $\lambda \in \mathbb{C}$, substitution into (2.55) yields

$$e^{i\langle k,x\rangle}\,\partial_t e^{\lambda t} = e^{\lambda t}\,\mathcal{P}(\partial_x)e^{i\langle k,x\rangle}$$

and hence,

$$\lambda e^{\lambda t}e^{i\langle k,x\rangle} = \mathcal{P}(ik)e^{\lambda t}e^{i\langle k,x\rangle} \quad \Rightarrow \quad \lambda = \mathcal{P}(ik).$$

As a necessary condition for well-posedness $\tilde{u}$ obviously has to remain bounded (or even decay) for all times, meaning that the real part of λ has to be bounded from above for all k, i.e.

$$\Re\,\lambda = \Re\,\mathcal{P}(ik) \leq const., \quad \forall k \in \mathbb{R}^n, \tag{2.60}$$

which is the exact same condition as (2.58), derived from the Fourier transformed equation.

Remark 2.27. Some authors use the ansatz function $e^{\pm i\lambda t}e^{i\langle k,x\rangle}$ to derive the dispersion relation, which changes the left hand side in (2.60), but not the actual condition. Say $\lambda = \lambda_r + i\lambda_i$, thus $\pm i\lambda = \pm i\lambda_r \mp \lambda_i$, such that the imaginary part of λ is responsible for the boundedness of $\tilde{u}$. Therefore, for the exponent $\pm i\lambda$ the condition reads

$$\mp\Im\,\lambda = \Re\,\mathcal{P}(ik) \leq const.$$

Thus, from expanding a possible solution into a Fourier series and continuing with its coefficients (i.e. applying a Fourier transform to the whole equation) one obtains the exact same conditions as if substituting a perturbation ansatz of the form shown above. The advantage of the dispersion relation can be best seen in the case of nonlinear problems, where a linearization around some steady state has to be performed.

Let a (weakly nonlinear) Cauchy problem be given as

$$\partial_t u = \mathcal{P}(\partial_x)u + u^2$$

and $u(x,t) = u_0(x) + \tilde{u}(x,t)$, then $\tilde{u}$ satisfies

$$\partial_t \tilde{u} = \mathcal{P}(\partial_x)\tilde{u} + 2\tilde{u}u_0,$$

where a Fourier transform would introduce the problem of dealing with the term $\mathcal{F}(\tilde{u}u_0)$, which, in general, leads to $\mathcal{F}(\tilde{u}) * \mathcal{F}(u_0)$, and one cannot always assume u_0 to be Fourier transformable. On the other hand, assuming $\tilde{u}(x,t) = e^{\lambda t}e^{i\langle k,x\rangle}$, as above, one has

$$\lambda\, e^{\lambda t}e^{i\langle k,x\rangle} = \mathcal{P}(ik)\, e^{\lambda t}e^{i\langle k,x\rangle} + 2u_0\, e^{\lambda t}e^{i\langle k,x\rangle}, \tag{2.61}$$

with the dispersion relation yielding $\Re\,\lambda = \Re\,\mathcal{P}(ik) + 2u_0(x)$. Although this means that the behavior of the disturbance depends on u_0, for well-posedness one still needs the upper bound for $\Re\,\mathcal{P}(ik)$. Hence, using the dispersion relation ansatz easily shows that nonlinearities do not change the requirements for well-posedness, just maybe enhance or delay the (temporal) growth of perturbations.

Excursus I: Abstract Cauchy Problems

In modern mathematics, where evolution equations are often formulated as ordinary differential equations in Banach spaces, the concept of well-posedness has to be generalized in a way, such that the above mentioned partial differential equations and the according conditions can be treated as special cases.

We will now present, in a short and formal manner, the main aspects of these generalizations, where more details of the theory (which is beyond the scope of this treatise) can be found e.g. in Engel & Nagel (2000).
Let us define an *abstract Cauchy problem* as the initial value problem given by

$$\begin{aligned} &\partial_t u(t) = \mathcal{A}u(t), \quad t \geq 0, \\ &u(0) = u_0. \end{aligned} \tag{2.62}$$

With X being some Banach space, we further assume the initial value $u_0 \in X$, $\mathcal{A} : D(\mathcal{A}) \to X$ to be a linear operator and call $u : \mathbb{R}^+ \to X$ a *classical solution* of the Cauchy problem, if it is continuously differentiable with respect to X, $u(t) \in X$, $\forall t \geq 0$ and (2.62) holds. The usual definition of a *strongly continuous semigroup* $\big(T(t)\big)_{t\geq 0}$ then gives $u(t) = T(t)u_0$.

Remark 2.28. Note that this definition does coincide with the requirement for classical solutions of (2.47), cf. Remark 2.24, where we assumed the solution to map $[0,T]$ continuously differentiable onto X. Here, since $T(t)$ is (strongly) continuous in t, it is sufficient to require u to lie in X at all times.

As it is well-known, if $\mathcal{A}$ is a closed operator, which generates a strongly continuous semigroup, then there exists a unique solution of (2.62) for every $u_0 \in X$. For the continuous dependence on the data, Engel & Nagel (2000) proved the equivalence of the generation of the semigroup with the assertion that $\mathcal{A}$ has a dense domain and for every sequence of initial values $(v_n) \in X$, with $\lim_{n\to\infty} v_n = 0$, it follows that $\lim_{n\to\infty} u(t, v_n) = 0$, uniformly in compact intervals $[0,T]$. Using this as a definition of well-posedness Engel & Nagel (2000) essentially proved

Lemma 2.1. *For a closed operator $\mathcal{A}$ the associated abstract Cauchy problem is well-posed, if and only if $\mathcal{A}$ generates a strongly continuous semigroup on X.*

Let us denote such a semigroup by $\big(T(t)\big)_{t\geq 0}$, then (under certain conditions on $\mathcal{A}$) the semigroup can be expressed as an exponential $T(t) := e^{\mathcal{A}t}$ with a solution given by $u(t) = T(t)u_0$ (cf. matrix exponentials in systems of ordinary differential equations and the Hille-Yosida theorem for contraction semigroups). Furthermore, there exist constants $\omega \in \mathbb{R}$ and $M \geq 1$ such that

$$\|T(t)\| \leq Me^{\omega t}, \quad \forall t \geq 0. \tag{2.63}$$

Then the infimum ω_0 of the set

$$\{\omega \in \mathbb{R} : \ \exists M_\omega \geq 1, \text{ such that (2.63) holds for } M_\omega\} \tag{2.64}$$

is called *(exponential) growth bound*. Combining this with the fact that for $\lambda \in \mathbb{C}$, $\Re\,\lambda > \omega$, we have $\lambda \in \rho(\mathcal{A})$, i.e. the resolvent set, and defining the spectral bound

$$s(\mathcal{A}) := \sup\{\Re\,\lambda : \ \lambda \in \sigma(\mathcal{A})\}, \quad \sigma = \mathbb{C}\backslash\rho,$$

one readily obtains $-\infty \leq s(\mathcal{A}) \leq \omega_0 < \infty$.

Eventually, by relating $\mathcal{P}(ik)$ to the spectrum of the differential operator $\mathcal{P}(\partial_x)$, we have found the more or less exact same conditions for well-posedness in the abstract Cauchy problem setting as for classical partial differential initial value problems, cf. (2.57) and (2.58). This can be shown to also hold for a broader class of operators, see Lemma 2.4.

Moreover, as proved in Engel & Nagel (2000), replacing the partial differential operator ∂_x by (ik), denoting $a(k) := \mathcal{P}(ik)$ (also allowing for complex coefficients) and defining $\mathcal{A}$ via these multipliers (cf. (2.56)), the equivalence

$$\mathcal{A} \text{ generates a strongly continuous semigroup} \quad \Leftrightarrow \quad \sup_{k\in\mathbb{R}^n} \Re\, a(k) < \infty \tag{2.65}$$

holds for all such $\mathcal{A}$ acting on $L^2(\mathbb{R}^n)$.

Remark 2.29. The equation for the Cauchy problem stated in (2.47) or (2.48) is inhomogeneous and therefore one has to make the connection to according inhomogeneous abstract problems by assuming the homogeneous problem to be well-posed, where a possible solution is given by the variation of parameters formula, i.e.

$$u(t) = T(t)u_0 + \int_0^t T(t-s)g(s)ds, \quad \forall t \geq 0,$$

with g containing all inhomogeneous terms. Such a description is often called *mild solution*, where further details can again be found in Engel & Nagel (2000).

End of Excursus I

Motivated by simple partial differential equations and their Cauchy problems we have established necessary and sufficient conditions for well-posedness via the boundedness of strongly continuous semigroups, which relates to the boundedness of associated Fourier multipliers. To extend this to the operators involved in the problems stated in (2.48), i.e. combinations of singular integral and classical differential operators, we need certain properties of such integrals.

One of the earliest works dealing with Fourier transforms and multipliers of (singular integral) operators is that of Mikhlin (1936). He introduced the notion of a *symbol* of an operator and first mentioned that sums and products of (singular) integral operators correspond to the sums and products of their according symbols.

This has been considered further, providing more details, in the monograph of Mikhlin (1965), where the important connection to the Fourier transform has been made, resulting in the fact that *the symbol of a singular integral operator coincides with the Fourier transform of its kernel*. That is, let $\mathcal{K}$ be a singular integral operator with kernel K then the symbol $sb(\mathcal{K})$ can be given as (see Mikhlin (1965))

$$sb(\mathcal{K}) = \mathcal{F}K.$$

Thus, for f in a suitable function space, another form to represent $\mathcal{K}$ would be

$$\mathcal{K}f = \mathcal{F}^{-1} sb(\mathcal{K}) \mathcal{F}f. \tag{2.66}$$

N.b.: This is actually a very general and often used form to view operators (and combinations thereof), which possess a symbol.
Furthermore, it is now completely obvious how sums and combinations of operators are connected to their symbols, e.g.

$$\mathcal{F}\big((\mathcal{K}_1 + \mathcal{K}_2)f\big) = \big(\mathcal{F}\mathcal{K}_1 + \mathcal{F}\mathcal{K}_2\big)f = \big(sb(\mathcal{K}_1) + sb(\mathcal{K}_2)\big)\mathcal{F}f,$$

$$\mathcal{F}\big(\mathcal{K}_1(\mathcal{K}_2 f)\big) = sb(\mathcal{K}_1)\mathcal{F}(\mathcal{K}_2 f) = sb(\mathcal{K}_1)sb(\mathcal{K}_2)\mathcal{F}f.$$

With the use of the Fourier transform it becomes clear why Mikhlin (1965) only considered actual singular integrals, meaning that the kernel is given as

$$K(x-y) = \frac{u(\theta)}{r^n}, \qquad r = |x-y|, \quad \theta = (x-y)/r, \tag{2.67}$$

where u is called the *characteristic*, which does not depend on the pole (cf. the Riesz transforms defined in Remark 3.62, as well as Theorem 3.40 and the treatise of Calderon & Zygmund (1952)). In this case Mikhlin (1965) showed, that the combination of two such singular operators (and the multiplication of their symbols) commute. For weakly singular kernels, e.g. the one defining the Riesz potential operator (3.98), the Fourier transform and hence the symbol have to be considered in a distributional sense, see Remark 2.31 below.

Remark 2.30. Although using the Fourier transform to gain symbols of (singular integral) operators is a very well established and practicable technique, associating symbols with operators can be understood in a much broader sense. Mikhlin & Prößdorf (1980) introduced the symbol as an element of a *symbol ring* $\mathfrak{r}$, which is the image of a homomorphism $\mathfrak{h}$ from the ring of linear operators $\mathfrak{R}$. One then consequently has for $a \in \mathfrak{r},\ A \in \mathfrak{R}$ that there exists exactly one a, such that $\mathfrak{h}(A) = a$ (the image of A) and at least one A being the pre-image of a. Also, if $A, B \in \mathfrak{R},\ a, b \in \mathfrak{r}$ and $\mathfrak{h}(A) = a,\ \mathfrak{h}(B) = b$ then $\mathfrak{h}(A + B) = a + b$ and $\mathfrak{h}(AB) = ab$. We shall refrain from paraphrasing more details presented in Mikhlin & Prößdorf (1980) and just finally mention that the symbol ring $\mathfrak{r}$ is by far not unique. If the ring $\mathfrak{r}_1$ is the homomorphic image of $\mathfrak{r}$, then $\mathfrak{r}_1$ can also be taken as a symbol ring for $\mathfrak{R}$.

Next we apply the results above to the operators appearing in (2.32) and (2.48) (as well as their two-dimensional equivalents). As shown in Gorenflo & Vessella (1991) the Fourier transform of the Abel kernel, $K_A(x) = H(x)x^{\alpha-1}$, see (3.96) for the definition, is

$$\mathcal{F}(K_A) = \Gamma(\alpha)(ik)^{-\alpha} = sb(\mathcal{J}^{\alpha}_{-\infty}) \tag{2.68}$$

and analogously $sb(\mathcal{J}^{\alpha}_{\infty}) = \Gamma(\alpha)(-ik)^{-\alpha}$. Here, for definiteness, we set (cf. Gorenflo & Vessella (1991))

$$(\pm ik)^{-\alpha} = |k|^{-\alpha}\, \exp(\mp i\alpha\frac{\pi}{2}\mathrm{sgn}(k)).$$

Similarly the Fourier transform of the Riesz potential kernel, $K_R(x) = |x|^{\alpha-n}$, can be calculated (distributionally) to be

$$\mathcal{F}(K_R) = \gamma(\alpha)|k|^{-\alpha} = sb(\mathcal{R}^{\alpha}), \tag{2.69}$$

see (3.98) for the definition.

As demonstrated in equations (2.59) through (2.60) the ansatz function $e^{\lambda t}e^{i\langle k,x\rangle}$ and the resulting dispersion relations yield the boundedness condition for the multipliers in the same way as the Fourier transform does. Therefore, we want that approach to be applicable to integral operators as well. The following lemma states the straight forward extension of the multiplier property in (2.59).

Lemma 2.2. *Let $\mathcal{K}$ be an integral operator with a convolution kernel K, of which the Fourier transform exists (in some sense). Then*

$$\mathcal{K}e^{i\langle k,x\rangle} = \mathcal{F}(K)e^{i\langle k,x\rangle}$$

holds for all $x, k \in \mathbb{R}^n$.

Proof. To see this, we simply modify

$$\mathcal{K}e^{i\langle k,x\rangle} = \int\limits_{\mathbb{R}^n} K(x-y)e^{i\langle k,y\rangle}dy = \int\limits_{\mathbb{R}^n} K(y)e^{i\langle k,x-y\rangle}dy =$$

$$= \int\limits_{\mathbb{R}^n} K(y) e^{i\langle k,x\rangle} e^{-i\langle k,y\rangle} dy = e^{i\langle k,x\rangle} \int\limits_{\mathbb{R}^n} K(y) e^{-i\langle k,y\rangle} dy = e^{i\langle k,x\rangle} \mathcal{F}(K),$$

such that substitution of (2.68) and (2.69) shows the validity of Fourier multipliers for integral operators. □

Remark 2.31. Equation (2.69) actually represents a formal application of the Fourier transform, where Stein (1970) proved the precise meaning of it in a distributional sense. We shall paraphrase the main idea of this result. Let ϕ lie in the Schwartz space of rapidly decaying functions, then the assertion of $\gamma(\alpha)|x|^{-\alpha}$ being the Fourier transform of $|x|^{\alpha-n}$ is understood as

$$|x|^{\alpha-n} = \mathcal{F}^{-1}(\mathcal{F}|x|^{\alpha-n}) \overset{formally}{=} \mathcal{F}^{-1}(\gamma(\alpha)|k|^{-\alpha}),$$

with the second equality actually meaning

$$\begin{aligned}\int |x|^{\alpha-n}\phi(x)dx &= \int \gamma(\alpha)\mathcal{F}^{-1}(|k|^{-\alpha})\phi(x)dx = \iint \gamma(\alpha)|k|^{-\alpha}e^{i\langle k,x\rangle}dk\,\phi(x)dx = \\ &= \int \gamma(\alpha)|k|^{-\alpha}\int \phi(x)e^{i\langle k,x\rangle}dx\,dk = \int \gamma(\alpha)|k|^{-\alpha}\overline{\mathcal{F}(\phi)}dk,\end{aligned}$$

such that $|x|^{\alpha-n}$ is the inverse Fourier transform of $\gamma(\alpha)|x|^{-\alpha}$ in the sense of distributions.

Remark 2.32. For the Abel kernel there is another, constructive way, to prove Lemma 2.2. Consider

$$\begin{aligned}&\int (x-y)^n e^{iky}dy = \\ &= \int (x-y)^n e^{ikx}e^{-ik(x-y)}dy = e^{ikx}(ik)^{-(n+1)}\int (x-y)^n (ik)^{n+1}e^{-ik(x-y)}dy = \\ &= e^{ikx}(ik)^{-(n+1)}\int ((x-y)ik)^n (ik)e^{-ik(x-y)}dy,\end{aligned}$$

where the coordinate transform $t = ik(x-y)$, $dy = -(ik)^{-1}dt$ then yields

$$e^{ikx}(ik)^{-(n+1)}\int (-1)t^n e^{-t}dt = e^{ikx}(ik)^{-(n+1)}\Gamma(n+1,t),$$

with $\Gamma(n,t)$ being the incomplete gamma function, such that $\Gamma(n,0) = \Gamma(n)$. In case of the Abel operator $n = \alpha - 1$, $0 < \alpha < 1$, and hence

$$\mathcal{J}^{\alpha}_{-\infty}(e^{ikx}) = e^{ikx}(ik)^{-\alpha}\big(\Gamma(\alpha,0) - \Gamma(\alpha,i\infty)\big) = \Gamma(\alpha)(ik)^{-\alpha}e^{ikx},$$

since $\Gamma(n,i\infty) = 0$ if $n < 1$ (and analogously for $\mathcal{J}^{\alpha}_{\infty}$). □

Remark 2.33. The connection, or even say equivalence, between the symbol of a (singular integral) operator and the Fourier multiplier property, as well as the abstract *multiplication semigroup*, lead to the common agreement of also calling $(ik)^m$ in (2.56) the symbol of the derivative operator ∂_x^m.

Denoting any type of operator possessing a symbol by $\mathcal{A}$, we have ($\sum\prod$ symbolizes all appearing sums and combinations of the operators)

$$\mathcal{F}\big(\sum\prod \mathcal{A}_j f\big) = \big(\sum\prod sb(\mathcal{A}_j)\big)\mathcal{F}(f),$$

for f in a suitable function space and at least for integro-differential operators

$$\big(\sum\prod \mathcal{A}_j\big)e^{i\langle k,x\rangle} = \big(\sum\prod sb(\mathcal{A}_j)\big)e^{i\langle k,x\rangle} \tag{2.70}$$

holds.

Finally we restate a necessary condition for well-posedness of evolution problems involving all kinds of combinations of $\mathcal{A}_j$ as

$$\Re\sum\prod sb(\mathcal{A}_j)(k) \leq const. \quad \forall k \in \mathbb{R}^n. \tag{2.71}$$

Remark 2.34. The study of singular integrals and their symbols also lead to the development of the theory of *pseudo-differential* operators. For the sake of completeness we shall briefly demonstrate the relation of such operators to the operators in equations (2.49) and (2.48).
If m is a real number then $S^m = S^m(\mathbb{R}^n \times \mathbb{R}^n)$ is the set of all $a \in C^\infty(\mathbb{R}^n \times \mathbb{R}^n)$ such that for all multi-indices γ, δ there exists a positive constant C

$$\left|\partial_x^\gamma \partial_\xi^\delta a(x,\xi)\right| \leq C(1+|\xi|)^{m-|\delta|}, \quad \forall x, \xi \in \mathbb{R}^n, \tag{2.72}$$

where C might depend on γ, δ. S^m is then called the space of symbols a of order m.
Let $a \in S^m$ and f be in the Schwartz space then (with $\hat{f}$ denoting the Fourier transform of f)

$$\mathcal{A}f(x) := (2\pi)^{-n} \int\limits_{\mathbb{R}^n} e^{i\langle x,\xi\rangle} a(x,\xi)\hat{f}(\xi)d\xi \tag{2.73}$$

defines a pseudo-differential operator of order m, with $\mathcal{A}f$ being again in the Schwartz space. The above definitions are taken from the renowned monograph of Hörmander (1985).
Since $\mathcal{A}$ maps the Schwartz space into itself, by ignoring the factor $(2\pi)^{-n}$, which is due to different definitions of the Fourier transform, one can rewrite (2.73) in the form

$$\mathcal{F}(\mathcal{A}f) = a(x,k)(\mathcal{F}f)(k),$$

which shows the relation to singular integrals and classical derivatives, cf. (2.56) and (2.66). Allowing the symbol a to also depend on x is in accordance to non-constant coefficient derivative type operators and singular integrals, where the characteristic depends on x, cf. (2.67).

Considering the symbols $sb(\mathcal{J}^\alpha_{\pm\infty}) \propto (\pm ik)^{-\alpha}$ as well as $sb(\mathcal{R}^\alpha) \propto |k|^{-\alpha}$ in virtue of the boundedness condition (2.72) readily shows, that *singular integrals are not pseudo-differential operators*. On the other hand, it is straight forward to show, that every classical derivative

operator $\mathcal{P}(\partial_x^m)$ (allowing also for non-constant coefficients) is in fact a pseudo-differential operator.

Furthermore, combinations of derivatives with singular integrals, as those appearing in the equations considered here, cannot be regarded as pseudo-differential operators (at least not in the sense of Hörmander), since higher derivatives of terms like $|k|^\alpha$, $\alpha \notin \mathbb{N}$, will eventually lead to a singularity of the symbol at $k = 0$ and consequently violate (2.72).
However, some authors still view operators associated with non-smooth symbols as pseudo-differential operators, if the exponent α is at least greater than one, i.e. if at least one order of differentiation appears.

Remark 2.35. A probably more appropriate way to view operators with symbols of the form $|k|^\alpha$, $\alpha \in \mathbb{R}^+$, is via the notion of *fractional derivatives*. In this context a combination such as $\partial_x^n \mathcal{J}_{-\infty}^\alpha$, is sometimes called *Weyl fractional derivative* of order $n - \alpha$, in contrast to *Riemann-Liouville fractional derivatives*, where the bounds of the integral are finite.
Another important type of derivatives are that of *Caputo*, as given, for example, in Podlubny (1999),

$$ {}_{-\infty}\mathbf{D}_t^\alpha f(t) = \frac{1}{\Gamma(n-\alpha)} \int_{-\infty}^{t} \frac{f^{(n)}(\tau)}{(t-\tau)^{\alpha+1-n}} d\tau, \quad n-1 < \alpha < n, $$

such that one can easily show, for $\alpha \to n$, ${}_{-\infty}\mathbf{D}_t^\alpha \to d^n/dt^n$.
Thus, for example, ${}_{\infty}\mathbf{D}_t^{3/2}$ (omitting the constant $1/\Gamma(1/2)$) yields the same operator as $\mathcal{J}_\infty^{1/2}\partial_x^2$, which appears in (2.34). Further properties of fractional calculus can also be found in Podlubny (1999). It is worth mentioning that he devoted a chapter to what is known as *spectral relationships*. That is certain fractional integrals applied to special types of orthogonal polynomials yield again sets of orthogonal polynomials. Interestingly such relationships only exist for operators on bounded intervals, so that they cannot be applied to the cases in Section 3.3.2.

We will now show how all the above mentioned applies to the Cauchy problem (2.47). Let us start with the two-dimensional analogue (2.49). First formally inverting $\mathcal{J}_{-\infty}^{1/4}$ and writing the time derivative term on the left hand side, to gain a typical evolution equation, yields

$$ \partial_t A = \tfrac{\lambda}{\gamma}\big[\mathcal{J}_{-\infty}^{3/4}\big]^{-1} \mathcal{J}_\infty^{1/2}\big(\partial_x^2 A\big)(x,t) + F(A), \tag{2.74} $$

where $F(A)$ contains the nonlinearity and the inhomogeneity, cf. Ruban (1982).

Remark 2.36. The formal inversion of an Abel integral operator has been demonstrated in Gorenflo & Vessella (1991), by an actual inversion of the integral. A much easier way to do this is with respect to its symbol (and since it is only a formal inversion we need not to worry about the applicability of the Fourier transform). Thus, for $0 < \alpha < 1$,

$$ sb\Big(\big[\tfrac{1}{\Gamma(\alpha)}\mathcal{J}_{-\infty}^\alpha\big]^{-1}\Big) = sb(\tfrac{1}{\Gamma(\alpha)}\mathcal{J}_{-\infty}^\alpha)^{-1} = (ik)^\alpha = (ik)(ik)^{\alpha-1} = $$

$$= \tfrac{1}{\Gamma(1-\alpha)}(ik)\underbrace{\Gamma(1-\alpha)(ik)^{\alpha-1}}_{=sb(\mathcal{J}_{-\infty}^{1-\alpha})},$$

where we finally obtain

$$\left[\tfrac{1}{\Gamma(\alpha)}\mathcal{J}_{-\infty}^{\alpha}\right]^{-1} = \tfrac{1}{\Gamma(1-\alpha)}\,\partial_x\mathcal{J}_{-\infty}^{1-\alpha}. \tag{2.75}$$

Note that we have taken into account the constants involving the gamma function, such that the formula coincides with the result given in Gorenflo & Vessella (1991). □

As for the commutativity of classical derivatives with fractional integrals we refer to Podlubny (1999). Additionally, Prößdorf (1974) relates the inversion to the symbol by proving, e.g. for Wiener-Hopf equations on $\mathbb{R}^+$, that for bounded operators with absolute integrable kernels the (one-sided) inverse is bounded if and only if the symbol does not have any zeros on $[-\infty,\infty]$. This does in fact also hold for the Abel operator above (as shown in Gorenflo & Vessella (1991) via the ill-posedness of equations of the first kind on L^2).

It has been demonstrated above, cf. (2.61), when considering the conditions for well-posedness we do not need to take into account the additional term F in (2.74). Thus, in virtue of abstract Cauchy problems and Fourier symbols we say (omitting all positive constants)

$$\begin{aligned}&\mathcal{A}_{2D} := \left[\mathcal{J}_{-\infty}^{3/4}\right]^{-1}\mathcal{J}_{\infty}^{1/2}\partial_x^2 \;\Rightarrow\\ &sb(\mathcal{A}_{2D}) = (ik)^{3/4}(-ik)^{-1/2}(ik)^2 = (ik)^{3/4}(-ik)^{3/2}\end{aligned} \tag{2.76}$$

and as established in (2.65) in connection with (2.71), we calculate

$$\Re\, sb(\mathcal{A}_{2D}) = \Re\,|k|^{9/4}e^{i\,\mathrm{sgn}(k)\frac{3\pi}{8}} = |k|^{9/4}\underbrace{\cos\left(\tfrac{3\pi}{8}\right)}_{>0}, \tag{2.77}$$

which proves that no upper bound for the real part of the symbol can be found $\forall k \in \mathbb{R}$. Thus, the Cauchy problem associated with (2.49) is *ill-posed* in general.

Performing the same modifications in (2.48) yields (relating (x,z) with (k,l) in the Fourier transform and omitting positive constants)

$$\begin{aligned}&\mathcal{A}_{3D} := \left[\mathcal{J}_{-\infty}^{3/4}\right]^{-1}\mathcal{J}_{-\infty}^{1/2}\,\mathcal{R}^1\,(\partial_x^3+\partial_x\partial_z^2) \;\Rightarrow\\ &sb(\mathcal{A}_{3D}) = (ik)^{3/4}(ik)^{-1/2}(k^2+l^2)^{-1/2}((ik)^3 - ikl^2) = -(ik)^{5/4}(k^2+l^2)^{1/2},\end{aligned} \tag{2.78}$$

and by again considering

$$\Re\, sb(\mathcal{A}_{3D}) = -\Re\,|k|^{5/4}e^{i\,\mathrm{sgn}(k)\frac{5\pi}{8}}(k^2+l^2)^{1/2} = -|k|^{5/4}(k^2+l^2)^{1/2}\underbrace{\cos\left(\tfrac{5\pi}{8}\right)}_{<0}, \tag{2.79}$$

general *ill-posedness* can be inferred as well.

Remark 2.37. Say A_{st} is a solution of the steady problem (2.33) and $A(x,t) = A_{st}(x) + \tilde{A}(x,t)$, assuming the perturbation to be small, $|\tilde{A}| \ll 1$, then $\tilde{A}$ has to satisfy

$$\partial_t \tilde{A} = \tfrac{\lambda}{\gamma}\big[\mathcal{J}_{-\infty}^{3/4}\big]^{-1}\,\mathcal{J}_{\infty}^{1/2}\big(\partial_x^2 \tilde{A}\big)(x,t) - 2/\gamma\big[\mathcal{J}_{-\infty}^{3/4}\big]^{-1}\,\tilde{A}A_{st}.$$

As mentioned above, to derive the dispersion relation one substitutes $\tilde{A}(x,t) = e^{\omega t + ikx}$, such that (omitting all positive constants)

$$\omega = sb(\mathcal{A}_{2D}) - 2(ik)^{3/4}A_{st}, \tag{2.80}$$

where we have assumed a plane-parallel, unperturbed flow, such that $A_{st} = const.$, as given in Ruban (1982), which is valid for large, negative Γ. Therein k was further assumed to be real and positive. In Ryzhov & Smith (1984) a similar relation has been calculated, where A_{st} was replaced by a linear approximation. Both works mention the occurrence of instabilities for a certain range of k, whereas only Ryzhov & Smith (1984) put this in the context of incorrectly posed Cauchy problems.

For the three-dimensional problem (2.47), a relation of the form (2.79) has never been derived so far, although numerical solutions and further studies for the steady and unsteady case have been presented, e.g. in Duck (1989), Duck (1990), Braun & Kluwick (2002) and Braun & Kluwick (2004), where the question rather is and remains, in which sense (if any) do solutions have to be understood.

We will delay the theoretical investigation of ill-posed problems and first present some illustrations by numerically solving (2.49) and (2.47), i.e. by fully discretizing (2.52). Say

$$t_m = m\Delta t, \quad \underline{a}(t_m) = \underline{a}_m, \quad \frac{d}{dt}\underline{a}(t)\bigg|_{t_m} \approx \frac{\underline{a}_{m+1} - \underline{a}_m}{\Delta t}$$

then the explicit Euler forward scheme for (2.52) reads

$$\underline{a}_{m+1} = \underline{a}_m + \Delta t\,\underline{D}^{-1}[\underline{K}\,\underline{a}_m + \underline{f} + \underline{g} - \big(\underline{Cw}\,\underline{a}_m\big)^2 - 2\underline{C}\,\underline{a}_m - \Gamma - 1], \tag{2.81}$$

which applies to both (2.49) and (2.47), as explained in the previous section. The matrix $\underline{D}$ can very well be expected to be non-singular for all (N_x, N_z), although one shall carefully consider the issues mentioned in Gorenflo & Vessella (1991), since this represents a formal inversion of an Abel operator, cf. (2.75).

From the analysis presented in, e.g. Gottlieb & Orszag (1977) and Hesthaven et al. (2007) one can reasonably assume the Chebyshev collocation method combined with the Euler forward time integration to be conditionally stable for well-posed problems, i.e. as long as a certain restriction on the time step Δt is satisfied (cf. CFL conditions). In one space dimension such a condition might read $\Delta t \leq cN^{-p}$, p depending on the spatial operator (for example, $p = 2$ for a simple advection operator and $p = 4$ for linear diffusion problems, see e.g. Canuto et al. (2006) for more details).

The scheme for the fully discretized two dimensional problem (2.49), with results depicted in Figure 10 using the parameters $N = 50$ (left), $N = 100$ (right) and $\Delta t = 10^{-8}$, can be expected to lie within the stability region of the method, provided the according problem is well-posed. Thus, we conclude that the oscillations shown are not due to numerical instabilities, but come exclusively from the ill-posedness.

What can be further inferred from Figure 10 is that when such problems are numerically solved with a direct method (i.e. the explicit scheme in (2.81)), the better the resolution (i.e. the higher the spatial accuracy) the worse the instabilities, meaning here, that they grow faster (cf. $t = 1.2 \times 10^{-2}$ on the left versus $t = 4 \times 10^{-3}$ on the right). Gorenflo & Vessella (1991) made a similar assertion for solving Abel equations of the first kind.

Another aspect relates directly to the dispersion relation and the symbol of the (fractional) derivative operator. Due to the parabolic-type shape of the real part, cf. (2.77), one can see that larger k result in larger ω in (2.80), i.e. the disturbances grow faster. Since k can be (formally) associated with the coefficient of the polynomial of degree k (for Chebyshev expansions are directly linked to Fourier series, as mentioned, e.g. in Mason & Handscomb (2003)), one can readily explain the different behavior for $N = 50$ and $N = 100$ in Figure 10. N.b.: The connection to Fourier analysis, with the conclusion of instabilities getting worse with higher spatial resolution, is also the reason for the stability condition (for well-posed problems) to become stricter if more polynomials are used.

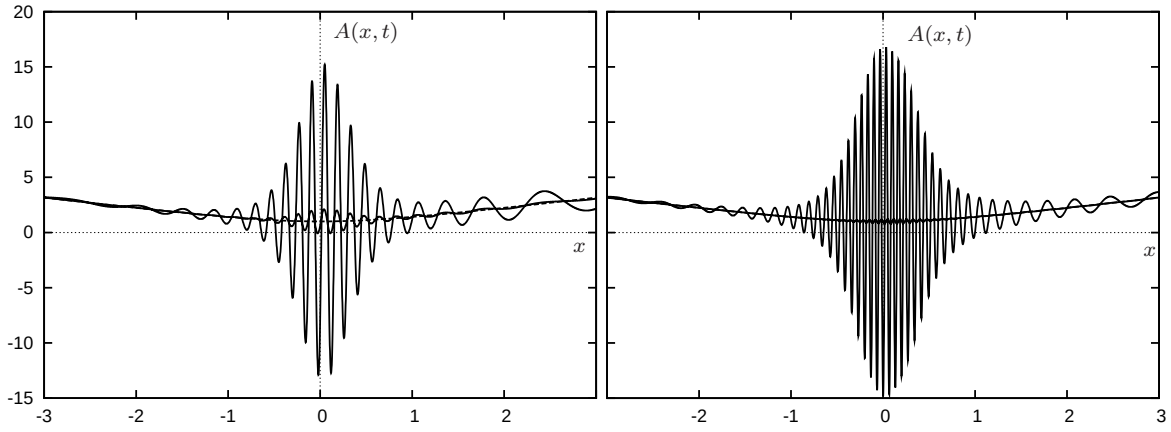

Figure 10: Solution of (2.49) using an explicit Euler scheme, $\Delta t = 10^{-8}$, $\Gamma = 2$, $g \equiv 0$, $A_0(x) = \sqrt{1 + x^2}$ (dashed). Left: $N = 50$, $t = 10^{-2}$ (small oscillations), $t = 1.2 \times 10^{-2}$ (large oscillations). Right: $N = 100$, $t = 3 \times 10^{-3}$ (small oscillations), $t = 4 \times 10^{-3}$ (large oscillations).

A similar situation can be observed in the three-dimensional case, i.e. the explicit Euler scheme shows fast growing disturbances. In Figure 11 we plotted a solution of (2.47), with an initial condition of the form

$$A_0(x, z) = \sqrt{1 + x^2} + \frac{z^2 - 1}{\sqrt{1 + x^2}(z^2 + 1)}, \tag{2.82}$$

which comes from setting $a_{02} = 1$, $a_{ik} = 0$ otherwise, in (2.51). Up to $t \approx 5 \times 10^{-3}$ the solution $A = A(x, z, t)$ remains almost equal to A_0 (dashed lines in Figure 11) and then, in full agreement with the dispersion relation (cf. the symbol in (2.79)), disturbances start to grow.

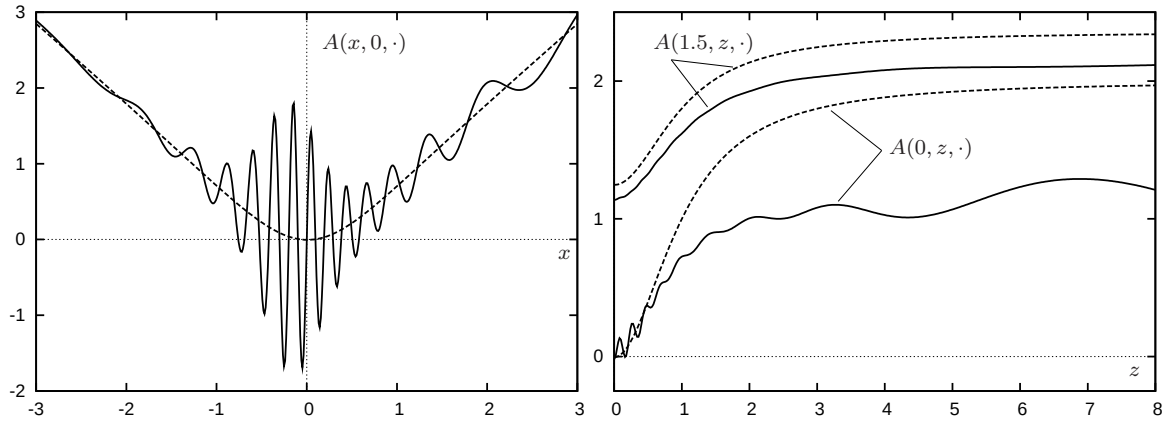

Figure 11: Solution of (2.47) using an explicit Euler scheme, $(N_x, N_z) = (35, 40)$, $\Delta t = 10^{-5}$, $\Gamma = 2$, $g \equiv 0$, A_0 from (2.82), at $t = 5 \times 10^{-3}$ (dashed) and $t = 1.5 \times 10^{-2}$ (solid)

Remarkable about the situations illustrated in Figures 10 and 11 (considered as solutions of (2.50)) is that the data, i.e. the initial condition, is analytical in the two-dimensional case (all $a_i = 0$, hence $B_0 \equiv 0$) and smooth in the three-dimensional case, i.e. $B_0 = (1 + x^2)^{-1/2} R_2(z)$ (cf. (2.82)). Also, as the unknowns are the coefficients, the initial data are prescribed with exact values. We therefore conclude that with the discretization per se, meaning solutions for $t > 0$ are only approximately represented, sufficient perturbations are introduced.

Remark 2.38. As shown in Remark 2.14 the operator $\mathcal{J}_{-\infty}^{1/2} \mathcal{R}^1 (\partial_x^3 + \partial_x \partial_z^2)$ is equivalent (modulo some constants) to $\mathcal{J}_{\infty}^{1/2} \partial_x^2$ when acting on functions independent of z. In the unsteady problem this means we solve problem (2.50), $g \equiv 0$, for solutions in $\mathbb{R}^2$ with $B_0 = B_0(x)$. Despite the growing instabilities (using e.g. $(N_x, N_z) = (50, 20)$) the solution mimics a behavior of $A(x, z, t) = A_{2D}(x, t) + \tilde{A}(x, z, t)$, with $|\tilde{A}| \ll 1$ and A_{2D} being almost equal to the solution in the left graph in Figure 10 (cf. also Remark 2.22). Then, before the whole solution A may develop visible z-dependence, it gets destroyed by the heavy oscillations in x. Comparing this observation with Figure 11 we assert that the approximation of $\mathcal{J}_{-\infty}^{1/2} \mathcal{R}^1 (\partial_x^3 + \partial_x \partial_z^2)$ is accurate enough not to introduce too much artificial z-dependence to purely x-dependent solutions, even though, the approximated operator itself initiates disturbances in z. Furthermore, since the real part of the symbol (2.79) is linear in $|l|$, but "a little more than" quadratic in $|k|$, the ill-posedness is much more severe with respect to x.

After Hadamard presented his seminal example for the Cauchy problem of the Laplace equation, it became common practice (see e.g. Hadamard (1923) and Petrowsky (1955)) to

first prove continuous dependence on the data and hence, to only treat well-posed problems by arguing that everything else would be unphysical.

It was not until the second half of the twentieth century when a theoretical basis for inverse problems started to develop. The two most prominent examples are *integral equations of the first kind*, see e.g. Gorenflo & Vessella (1991), where the operator is compact between the considered function spaces, and the Cauchy problem of the *backward heat equation*, sometimes more generally also called *final value problem*.

The monographical work by Tikhonov & Arsenin (1977) was the first summarizing the, at that time, established knowledge on improperly posed problems. Therein it was stated that the notion of ill-posedness shall always be connected to the function spaces the problem is considered in. That is, speaking in the most general way, given a mapping $\phi : X \to Y$, with (X, Y) being metric spaces, such that the problem of finding x when given y, i.e. $\phi(x) = y$ is ill-posed in some subspaces (X_1, Y_1), might very well be well-posed on other subspaces (X_2, Y_2) or in different metrics. To work on metric spaces is necessary, since one needs a precise meaning of the distance between various given data and their according solutions (cf. conditions *i*) and *ii*) from Petrowsky (1937) stated above).

Remark 2.39. For the class of abstract Cauchy problems such as (2.62) (which also contains problem (2.47)), the inverse of the mapping ϕ is the strongly continuous semigroup $\big(T(t)\big)_{t\geq 0}$, where we have shown above, via the symbols (2.77) and (2.79) and the equivalence (2.65), that the Cauchy problems dealt with here are ill-posed on $L^2(\mathbb{R}^2)$, i.e. $T(t)$ is unbounded (and thus discontinuous) on every bounded time interval, cf. (2.63). As mentioned in Remarks 2.12 and 2.15 a classical solution is actually expected to lie in a subspace (due to the differentiability requirements), where it is reasonable to assume the existence of smaller subspaces for solutions (and initial data) in which the problem (2.47) is well-posed. On the other hand, the question of the meaning of such restrictions with respect to the (physical) interpretation of the solutions remains.

Finding subspaces in which the considered problems are well-posed has been termed *selection method* in Tikhonov & Arsenin (1977), where the following lemma was proved, providing a very basic assertion.

Lemma 2.3. *Suppose that a compact subset X of a metric space X_0 is mapped onto a subset Y of a metric space Y_0. If the mapping $X \to Y$ is continuous and one-to-one then the inverse mapping $Y \to X$ is also continuous.*

Proof. see Tikhonov & Arsenin (1977) □

In other words, if we allow only for solutions lying in a compact subset, we obtain an admissible set of initial data, on which the mapping is continuous, i.e. the solutions depend continuously on the data. Such a result is obviously only the theoretical basis for selecting appropriate spaces and metrics and does not give a concrete strategy.

Remark 2.40. An attempt to impose or derive some restrictions on solutions of (2.49) has been made in Ryzhov & Smith (1984) by providing, as an example, a concrete formula for a

possible disturbance $\tilde{A}$ (cf. Remark 2.37), such that $\mathcal{F}\tilde{A}$ has compact support. As mentioned further therein, a compact support of the Fourier transformed perturbation is sufficient, but not necessary, for it to remain bounded (on finite time intervals). Braun & Kluwick (2004), on the other hand, remarked that solutions, when being in L^1 with respect to x, decay to zero in its Fourier transformed representation, which stands as a necessary but not sufficient condition, since nothing can be said about the rate of decay, which has to be faster than $|k|^{9/4}$, according to (2.79). What remains is the slighter, but still unbounded growth of disturbances with respect to $|l|$.

Remark 2.41. Concluding from the previous remark, we claim that for most problems (especially in physics and engineering) Lemma 2.3 is more or less inapplicable (as was also stated in Tikhonov & Arsenin (1977)), either due to the involved operators or due to the set and structure of the initial data. As said, because of the real parts of the symbols (2.77) and (2.79), regarding the Fourier transform of possible solutions of (2.48) and (2.49) the requirement of either *compact support* in $\mathbb{R}$ and $\mathbb{R}^2$, respectively, or fast enough decay, is sufficient.

For the sake of completeness it is worth mentioning that another remedy would be to bound derivatives of solutions (up to some order and in a certain way) on bounded time intervals. Thus, oscillations of the form shown in Figure 10, for example, are not permissible to occur in solutions. This can be associated to the decay of the Fourier transformed representation via the concept of *Gevrey classes* and *ultradistributions*, which are often found in connection with pseudo-differential operators. Since this subject is beyond the scope of this treatise, we shall only mention (see, e.g. Rodino (1993)) that if the Fourier transform of a function f belongs to the dual of the Schwartz space, satisfying for C and $\epsilon > 0$

$$|\mathcal{F}(f)(k)| \leq Ce^{-\epsilon|k|^{1/s}},$$

then f lies in the Gevrey class $\mathcal{G}^s(\mathbb{R}^n)$, i.e. for every compact subset $K \subset \mathbb{R}^n$ there exists a positive constant c, such that $\forall x \in K$ and multi-indices m

$$|\partial_x^m f(x)| \leq c^{|m|+1}(m!)^s,$$

consequently providing bounds on the derivatives and/or sufficient decay for the Fourier transformed solutions.

Lemma 2.3 and Remarks 2.40 and 2.41 are of rather theoretical nature, in the sense that there exist initial data and according solutions for which the Cauchy problem (2.47) and its two-dimensional analogue are well-posed, but they do not include any instruction on how to choose appropriate initial data in concrete, let alone possible schemes to depict solutions approximately.

Moreover, if one would impose some restrictions on the solution, such as compact support in its Fourier transformed or choosing a certain Gevrey class, an argument has to be found on how to relate this to the physics the solution shall describe. Another question remains

on whether the impositions are too strict, such that one can only obtain a small part of the information the problem actually contains.

For this reason one shall always take into consideration the fact that investigating ill-posed problems in applications is two-fold. On the one side, a theoretical study has to be done on how or in which sense the problem is improperly posed or unstable, as has been presented so far in this section, on the other side, appropriate methods have to be found or developed to actually find qualitative and/or quantitative descriptions of initial data and solutions, without being too restrictive. Such strategies are now commonly referred to as *regularization methods*.

It was stated in Tikhonov & Arsenin (1977) that in most applied problems, as the one studied here, the set, in which possible solutions are expected to lie, is not compact. Thus, as argued before, considering the problem only on compact subsets might leave out just the information one wants to obtain. Such problems were termed *genuinely ill-posed*, for which the concept of *regularizing operators* was introduced.

Given the equation $\mathcal{A}x = y$, such that the exact solution x^* corresponds to the datum y^* and metrics ρ_X and ρ_Y, then an operator $\mathcal{B}(y, \alpha)$ is said to be *regularizing* in a neighborhood of x^* if

i) there exist numbers α_0, δ_0, such that $\mathcal{B}$ is defined for any α and y with $0 < \alpha < \alpha_0$ and $\rho_Y(y, y^*) < \delta_0$;

ii) there exists an $\alpha(\delta)$, such that for every $\epsilon > 0$ there is a number $\delta(\epsilon) < \delta_0$, where for $\rho_Y(y, y^*) < \delta(\epsilon)$ and an $x_\alpha \in X$, $\rho_X(x_\alpha, x^*) < \epsilon$ holds.

We then call $x_\alpha = \mathcal{B}(y, \alpha(\delta))$ the *regularized solution* and α the *regularization parameter*.

Remark 2.42. As mentioned in Tikhonov & Arsenin (1977) a regularizing operator is not uniquely associated to a given equation and vice versa. Also, one can choose $\alpha(\delta)$, such that for $\delta \to 0$ (i.e. in the limit of exactly given data), the regularized solution approaches the exact one, i.e. $\rho_X(x_\alpha, x^*) \to 0$. It was thus proved in Tikhonov & Arsenin (1977) that if for an operator $\mathcal{B}$

$$\lim_{\alpha \to 0} \mathcal{B}(\mathcal{A}x, \alpha) = x$$

holds for every $x \in X$, $\mathcal{B}$ is a regularizing operator for $\mathcal{A}x = y$.

Tikhonov & Arsenin (1977) further develop methods of constructing regularizing operators and finding the optimal value of the parameter α. We shall not go into such details and continue with how to regularize Cauchy problems of evolutionary equations.

In virtue of the definition above Lavrent'ev et al. (1986) demonstrated for the abstract, homogeneous Cauchy problem (2.62) with $\mathcal{A}$ positive, self-adjoint and unbounded (e.g. the negative Laplacian) and thus rendering the problem ill-posed on $[0, T]$, the operator

$$\mathcal{B} = e^{\mathcal{A}t}\left(I + \mu(t)e^{\mathcal{A}T}\right)^{-t/T}, \quad \mu(t) = \frac{\delta}{M}(1 - t/T)^{-2+t/T} \tag{2.83}$$

to be bounded and regularizing in the sense that for initial data $\|u_{0,\delta} - u_0\| \leq \delta$, $\|\mathcal{B}u_{0,\delta} - u(t)\| \to 0$ on $[0,T)$, as $\delta \to 0$. Here I is the identity and M is an upper bound for solutions u on $[0,T]$. Since $\mathcal{B}(t)$ is a strongly continuous semigroup, not quite practicable in this representation, Lavrent'ev et al. (1986) further approximated $\mathcal{B}$ by a polynomial in $e^{\mathcal{A}T}$ and t, and together with the spectral description obtained a more applicable form.

Other works on regularizing abstract Cauchy problems, e.g. Showalter (1974), use the technique of *quasireversibility* (a notion owing to the works of R. Lattès and J.-L. Lions). Roughly speaking, this method adds or subtracts terms multiplied by a (small) parameter to the equation, such that the thus resulting operator generates a semigroup on the desired function spaces.
Obviously, the extension to inhomogeneous problems can be made, see e.g. Campbell Hetrick & Hughes (2007), who regularized the initial value problem (with $-\mathcal{A}$ generating a semigroup)

$$\partial_t u(t) = \mathcal{A}u(t) + h(t), \quad u(0) = u_0 \quad \rightsquigarrow \quad \partial_t v(t) = f(\mathcal{A})v(t) + h(t), \quad v(0) = u_0, \tag{2.84}$$

with $f(\mathcal{A})$ generating a semigroup (e.g. $f(\mathcal{A}) = \mathcal{A} - \alpha\mathcal{A}^2$), such that the problem for v is well posed and proving that v is close to u in some norm on $[0,T)$. This has then been successfully applied to the backward heat equation, i.e. $-\Delta$ is replaced by $-\Delta - \alpha\Delta^2$.
It is worth mentioning that all these methods fall into the category of regularizing operators, demonstrating again the multiple possibilities to regularize an ill-posed problem.

Although, with using regularizing operators, one is able to circumvent the difficulties in ill-posed problems, by obtaining stable (in some sense) approximate solutions, close to the sought ones, they do not necessarily result in a practicable scheme to actually calculate solutions (in some sense).

We will hence distinguish two strategies. One is to regularize the problem via the analysis above and then to solve the approximate problem with established discretization techniques. The other is to first discretize (or approximate) the ill-posed problem and then stabilize the finite-dimensional version. Eventually, in an appropriate limit, both approaches should yield the same, or at least very close, solutions.

The first method has been successfully applied to the backward solution of parabolic problems in e.g. Eldén (1982), where (2.83) was used on $t \in [0,1]$ to describe the well-posed approximation, which was then further discretized, in the spectral representation of the semigroup, by Padé approximation in time (of which the backward Euler procedure is a special case). There are, of course, other possibilities to first approximate and then discretize as shown, for example, in Jonas & Louis (2000), where the concept of mollifiers was used. Such an approach falls into the class of *filter methods*. The monograph of Louis (1989) provides a good introduction (among other techniques) to such types of regularization.

Remark 2.43. Restrictions to solutions with compact support (or fast decay) in their Fourier transform (cf. Remarks 2.40 and 2.41) can be viewed as a filter method. In fact, instead of only allowing for solutions with compact support, one could introduce a "cut-off"-function

(i.e. a filter), such that all solutions obtained by some method using this filter automatically have this compact support. The general idea, as presented in Louis (1989), is to consider the so called *singular system* of the *generalized inverse* (i.e. minimizing functional) and then cut off the growing part of the singular values.

We have now gathered sufficient mathematical information to assert that the Cauchy problems (2.48) and (2.49) can be regularized by *adding or subtracting an appropriate higher derivative term* on the right hand side, cf. (2.84). The idea obviously is to replace the equation in (2.50), such that the Cauchy problem becomes well-posed for the same set of initial conditions. This is, of course, everything but a unique way, and hence we need some constraints. By generalizing the equivalence (2.65) we get

Lemma 2.4. *Let the operator $T(t)$ be defined via $T(t)u := e^{\mathcal{A}t}u$, $u \in L^2(\mathbb{R}^n)$. If $\mathcal{A}$ possesses a symbol $sb(\mathcal{A})$, then $\mathcal{A}$ generates a strongly continuous semigroup on $L^2(\mathbb{R}^n)$, if and only if the real part of $sb(\mathcal{A})$ is bounded from above. Thus $\big(T(t)\big)_{t\geq 0}$ is continuous on L^2 for every finite time $t \leq T$. Additionally, the spectra of $\mathcal{A}$ and $sb(\mathcal{A})$ are equal.*

Proof. This has essentially been proved in Engel & Nagel (2000). Although the assertion was shown for the case of classical derivative operators the argumentation is the same. For the sake of completeness we shall summarize the main idea. The symbol $sb(\mathcal{A})$ regarded as a (continuous) function on $\mathbb{R}^n$ generates a strongly continuous semigroup $M(t)$ (a multiplication semigroup, to be precise), if and only if $\sup(\Re\, sb(\mathcal{A})) < \infty$ over all $k \in \mathbb{R}^n$.
Since $\mathcal{F}$ is an isomorphism on L^2, we have that the strongly continuous semigroup $T(t) = \mathcal{F}^{-1}M(t)\mathcal{F}$ is given by the generator $\mathcal{B} = \mathcal{F}^{-1}sb(\mathcal{A})\mathcal{F}$. By using the alternative description of operators from (2.66) we also have $\mathcal{A} = \mathcal{F}^{-1}sb(\mathcal{A})\mathcal{F}$. The equality of the spectra follows immediately. □

From Lemma 2.4 we infer that by altering the right hand side derivative type operator in (2.50), such that its symbol satisfies the requirements, we obtain a well-posed problem, which we then numerically solve by the usual combination of spectral collocation in spatial coordinates and a time marching scheme. But this imposes a constraint on finding the appropriate higher derivative terms, that is, the order of differentiation should be as small as possible, otherwise the restrictions on the discretization parameters become too strong.

An educated guess thus yields $-\alpha\partial_x^3 B$ for the two-dimensional and $\alpha(-\partial_x^3 + \partial_z^2)B$ for the three-dimensional problem, with α being the (usually small) regularization parameter. That these derivatives are of higher order than the original ones is straight forward and with Remark 2.26 it should suffice to consider the real parts of their symbols. But due to other aspects, discussed later, we shall give the full formulae for the new overall right hand side operators.

Caveat: Neither the signs nor the order of differentiation in those higher derivative terms have any special meaning so far. They were just chosen in such a way that one can find the required upper bounds and to work as "mollifying" as possible for the numerical calculations.

For the new symbols, we define, in accordance to (2.76) and (2.78),

$$\begin{aligned}\mathcal{A}^*_{2D} &:= \left[\mathcal{J}^{3/4}_{-\infty}\right]^{-1}\left(\mathcal{J}^{1/2}_{\infty}\partial_x^2 - \alpha\partial_x^3\right)\\ \mathcal{A}^*_{3D} &:= \left[\mathcal{J}^{3/4}_{-\infty}\right]^{-1}\left(\mathcal{J}^{1/2}_{-\infty}\mathcal{R}^1\left(\partial_x^3 + \partial_x\partial_z^2\right) + \alpha(-\partial_x^3 + \partial_z^2)\right),\end{aligned} \tag{2.85}$$

such that (omitting the usual constants)

$$\begin{aligned}sb(\mathcal{A}^*_{2D}) =& (ik)^{3/4}(-ik)^{3/2} - \alpha(ik)^{15/4} \;\Rightarrow\; \Re\, sb(\mathcal{A}^*_{2D}) = |k|^{9/4}(c_1 - c_2|k|^{3/2})\\ sb(\mathcal{A}^*_{3D}) =& -(ik)^{5/4}(k^2+l^2)^{1/2} - \alpha(ik)^{15/4} - \alpha(ik)^{3/4}l^2 \;\Rightarrow\\ &\Re\, sb(\mathcal{A}^*_{3D}) = c_1|k|^{5/4}(k^2+l^2)^{1/2} - c_2|k|^{15/4} - c_3|k|^{3/4}l^2,\end{aligned}$$

with $c_1, c_2 = c_2(\alpha), c_3 = c_3(\alpha)$ being positive.

Remark 2.44. The real part of $sb(\mathcal{A}^*_{2D})$ is obviously bounded from above, since

$$c_1 - c_2|k|^{3/2} \geq 0 \quad\Leftrightarrow\quad |k| \leq (c_1/c_2)^{2/3},$$

but, we shall mention this for later discussions, there exist intervals where the real part of the symbol is positive. As for $sb(\mathcal{A}^*_{3D})$, we consider its real part in polar coordinates $(k,l) \mapsto (r,\phi)$, obtaining

$$\Re\, sb(\mathcal{A}^*_{3D}) = r^{9/4}|\cos(\phi)|^{3/4}\big(c_1|\cos(\phi)|^{1/2} - c_2 r^{3/2}|\cos(\phi)|^3 - c_3 r^{1/2}\sin(\phi)^2\big),$$

where it is easy to see that some finite $R(\phi)$ exists, such that

$$c_1|\cos(\phi)|^{1/2} - c_2 r^{3/2}|\cos(\phi)|^3 - c_3 r^{1/2}\sin(\phi)^2 \geq 0 \quad\Leftrightarrow\quad r \leq R(\phi).$$

Thus we have again an upper bound for the real part and some regions, where the real part is positive.
Overall, the regularizing operators generate strongly continuous semigroups on L^2 and hence, from Lemma 2.1 the associated homogeneous Cauchy problem is well-posed. □

Lemma 2.4 cannot be extended in a straight forward manner to hold on $L^\infty(\mathbb{R}^n)$, since it needs the isomorphism of the Fourier transform. To show the boundedness of the semigroup $T(t)$ in the L^∞ norm, we employ a different strategy. Application of the Fourier transform to (2.62) obviously yields $T(t)u = \mathcal{F}^{-1}(e^{sb(\mathcal{A})t}) * u$. Consider

$$|T(t)u| = |\mathcal{F}^{-1}(e^{sb(\mathcal{A})t}) * u| \leq \|u\|_\infty \int |\mathcal{F}^{-1}e^{sb(\mathcal{A})t}(x)|dx = \|u\|_\infty \|\mathcal{F}^{-1}e^{sb(\mathcal{A})t}\|_{L^1}, \tag{2.86}$$

such that the inverse Fourier transform of the exponential of the symbol needs to be absolute integrable for all $t < T$. For symbols of the form $-|k|^\lambda$, $1 < \lambda \leq 2$ this has been proved in Droniou et al. (2002) (on the real line), where they additionally showed that the L^1 norm

is equal to one $\forall t > 0$, i.e. the semigroup remains bounded in the long time limit and furthermore, a maximum principle holds in such cases.

To relate this to the present problem we state

Lemma 2.5. *The regularizing operator $\mathcal{A}_{2D}^*$ defined in* (2.85) *generates a strongly continuous semigroup on $L^\infty(\mathbb{R})$ for $t < T$.*

Proof. One needs to show condition (2.86) to be satisfied by the symbol of the operator. First, boundedness follows from

$$\left| \int e^{sb(\mathcal{A}_{2D}^*)t} e^{ikx} dk \right| \leq \int \left| e^{sb(\mathcal{A}_{2D}^*)t} e^{ikx} \right| dk = \int \left| e^{sb(\mathcal{A}_{2D}^*)t} \right| dk,$$

since $|e^{sb(\mathcal{A}_{2D}^*)t}| = e^{\Re\, sb(\mathcal{A}_{2D}^*)(k)t} < \infty$, $\forall k$ and decays faster than $|k|^{-1}$ at infinity (see Remark 2.44). In other words, $e^{sb(\mathcal{A}_{2D}^*)t} \in L^1(\mathbb{R})$, which is necessary for the inverse Fourier transform to exist. As for the decay in terms of x, we use integration by parts twice, where the boundary terms vanish due to the strong decay of the exponential, yielding

$$\left| \int_{\mathbb{R}} e^{sb(\mathcal{A}_{2D}^*)t} e^{ikx} dk \right| = \frac{1}{x^2} \left| \int_{\mathbb{R}} \partial_k^2 e^{sb(\mathcal{A}_{2D}^*)t} e^{ikx} dk \right| \leq \frac{1}{x^2} \| \partial_k^2 e^{sb(\mathcal{A}_{2D}^*)t} \|_{L^1}.$$

Calculating $|\partial_k^2 sb(\mathcal{A}_{2D}^*)|$ one can readily deduce the existence of the L^1 norm. The estimate of quadratic decay with respect to x of the term on the very left hand side finishes the proof. □

Remark 2.45. In contrast to the symbols considered in Droniou et al. (2002), the semigroup generated by $\mathcal{A}_{2D}^*$ does not remain bounded as $t \to \infty$ because of the positive parts of the symbol (cf. Remark 2.44). Droniou et al. (2002) also mentioned for results in more than one dimension, e.g. for the semigroup generated by $\mathcal{A}_{3D}^*$, a proof as done above becomes heavily involved. Although having $e^{sb(\mathcal{A}_{3D}^*)t} \in L^1(\mathbb{R}^2)$ is again straight forward, using integration by parts to obtain asymptotic estimates for the decay with respect to x is not recommendable. Still, we can expect the inverse Fourier transform of $e^{sb(\mathcal{A}_{3D}^*)t}$ to decay sufficiently fast in x and hence has finite L^1 norm.

Remark 2.46. To show well-posedness on $[0, T)$ for the whole regularized problem (i.e. using the operators in (2.85) in (2.48)) one needs to consider the mild formulation of the solution

$$B(t) = e^{\mathcal{A}^* t} B_0 + \int_0^t e^{\mathcal{A}^*(t-s)} F(B) ds, \tag{2.87}$$

where F includes all other terms appearing in (2.50) (and its two-dimensional analogue). From the (local) Lipschitz continuity of the nonlinearity in B and the boundedness of the inhomogeneity we formally claim, based on the findings in Achleitner et al. (2011), the boundedness (in the L^∞ norm) of the second term.

We will now turn to the numerical solutions of the regularized problem. Since well-posedness is sufficiently established, we can apply the explicit Euler scheme described in (2.81). The term $\underline{D}^{-1}\underline{K}$ is replaced by the collocation approximation of $\mathcal{A}^*$ from (2.85). An advantage of the polynomial approach is having an exact description of the regularizing operator (since derivatives of polynomials can be given in closed form, cf. Section 3.1, Remark 3.2), such that the regularization does not introduce additional discretization errors.

As described in the paragraph following (2.81), for stability certain restrictions on the time step have to be satisfied. The parameters used to obtain Figures 10 and 11 are assumed to meet these conditions and hence we use the same setting for numerically solving the regularized equation.

As expected, the explicit Euler scheme is stable and yields some plausible time evolution on, e.g. $t \in [0,1]$, which is shown in Figure 12 (left), starting again from $A_0(x) = \sqrt{1+x^2}$.

Remark 2.47. As it is known from classical textbooks such as Forsythe & Wasow (1960), if an initial value problem is well-posed, an implicit scheme (e.g. backwards Euler or Crank-Nicholson) is unconditionally stable (under certain conditions) and therefore allows for larger time steps. The discrete problem (2.81) thus reads

$$\underline{a}_{m+1} = \left[\underline{D} - \Delta t\underline{K} + 2\Delta t\,\underline{C}\right]^{-1}\left(\underline{D}\,\underline{a}_m + \Delta t(\underline{f} + \underline{g} - \left(\underline{Cw}\,\underline{a}_m\right)^2 - \Gamma - 1)\right), \tag{2.88}$$

which is still first order accurate in time, cf. (2.81). As usual we can reasonably expect the matrix on the right hand side to be non-singular. Notice that this should rather be called a *semi-implicit* scheme, since the quadratic term in $\underline{a}$ is evaluated in an explicit sense, i.e. at the previous time step. As proposed in Scheichl et al. (2008), a linearization of the nonlinear term at the respective time step $m+1$, with respect to previous ones $m, m-1, \ldots$, can yield a higher accuracy for the right hand side evaluations at the respective time step.

In Figure 12 (right), we used the backward or implicit Euler scheme, aiming to depict some long time dynamics of the problem. Since the results for $t > 0$ gained from the regularized, explicit Euler approach can be regarded as sufficiently accurate, for the time step is very small, we have to compare them with results from the implicit scheme using different Δt. Such comparisons can then be utilized to find an upper bound for the time step in the implicit scheme, regarding the growth of discretization errors as $\Delta t \nearrow$. In the case studied with approximately $\Delta t \leq 1/100$ the solutions from the explicit and implicit method are virtually indistinguishable. Hence, results involving longer time intervals are best obtained with the implicit method (allowing for faster, but equally accurate, computations).

Remark 2.48. The most striking observation, which can be made from Figure 12, is the fact that at $t = 20$ the solution visually coincides with the steady state solution obtained from (2.33) (in terms of B, as usual) with the additional term $-\alpha\partial_x^3 B$ on the right hand side. Although, the movement toward this steady state is considerably decelerating (cf. $t = 10$ in Figure 12 (right)).

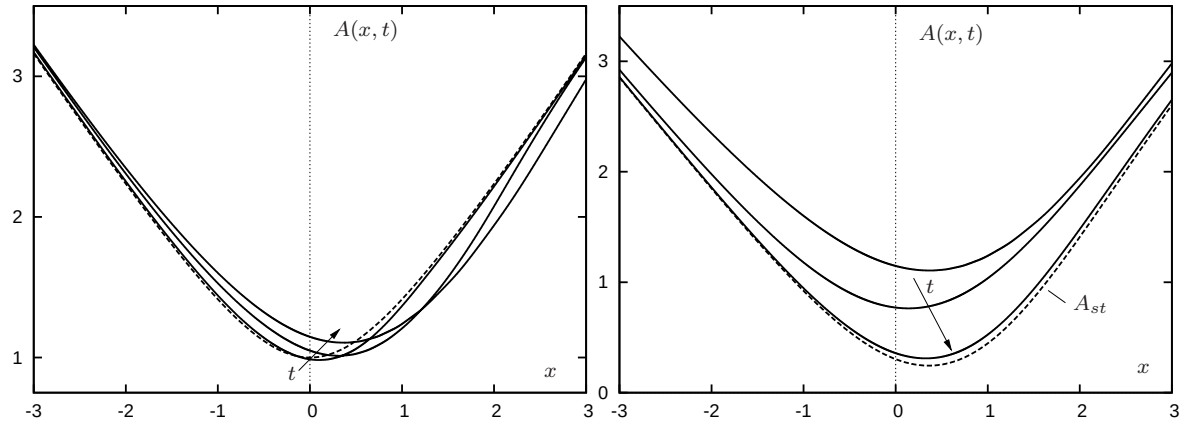

Figure 12: Regularized solution of (2.49) using $N = 50$, $\Gamma = 2$, $g \equiv 0$, $\alpha = 1/100$, $A_0 = \sqrt{1+x^2}$. Left: explicit Euler, $\Delta t = 10^{-8}$ at $t = 0$ (dashed), $t \in \{0.1, 0.5, 1\}$. Right: implicit Euler, $\Delta t = 10^{-2}$ at $t \in \{1, 5, 10\}$, A_{st} (dashed) is the steady state including the regularization.

One might thus be led to claim some sort of stability or attractiveness of such stationary solutions in terms of dynamical systems. Heuristically, we state that if a solution is close to the steady state it remains in a certain neighborhood of it and for $t \gg 1$ it is identical to the equilibrium (cf. the definitions of *Lyapunov* and *asymptotic stability*). In virtue of the bifurcation diagram in Figure 4 we conclude further the upper branch to be stable and the lower branch to be unstable. This is confirmed by some additional calculations performed within this type of regularization by taking as an initial condition a slightly perturbed lower branch solution (cf. graph (*b*) in Figure 3). It seems that the lower branch is repelling, either toward the upper branch or to some unsteady behavior (e.g. continuous growth of the absolute values).

Caveat: All these results and findings are only to be regarded in the sense of regularized dynamics, otherwise the notion of a steady state itself has no meaning. Additionally, as shown in Remark 2.44, certain parts of the spectrum of $\mathcal{A}^*$ have positive real parts, such that, not necessarily but likely, some destabilization occurs in long time asymptotics. As mentioned in Ruban (1982), with the formal stability analysis performed therein, claims about stability by ignoring the ill-posedness are only valid for very large negative values of the bifurcation parameter Γ, where the flow is far from separation.

To still connect such conclusions to the actual Cauchy problem we compare the above mentioned (regularized) steady state to stationary solutions calculated in Section 2.2, cf. Figure 3 (upper branch). By comparing the leading order coefficients of the respective expansions, see Table 11, with the coefficients for the original solution taken from Table 3 and plotting the according graphs shows that they are more or less identical.

Therefore, we formally assert that the *regularized dynamics* sufficiently describe, at least qualitatively, the time evolution of solutions to the original Cauchy problem, as long as α is small enough, such that regularized solutions remain close to the sought ones.

a_i	original	regularized
a_0	-0.5894	-0.5904
a_1	-0.1878	-0.1887
a_2	0.2178	0.2190
a_3	0.1095	0.1111
a_4	0.1697	0.1699
a_{10}	0.0133	0.0133

Table 11: Leading coefficients for the stationary solutions of the original and regularized steady problem, $\alpha = 0.01$.

Remark 2.49. Although from a theoretical viewpoint, the Cauchy problems remain well-posed for all $\alpha > 0$ (where maybe T changes), this does not hold for the actual computations. In Figure 13 we depict the results for $\alpha = 10^{-3}$ at $N = 50$ and $N = 100$ (left) and $\alpha = 1$ (right). The conclusion drawn from this is, on the one hand, that all parameters are connected through some functional relation in order to act regularizing, and on the other hand, that the parameter α reveals a very typical behavior of regularized problems - it has to be neither too small nor too large. A neat example for this can be found in Louis (1989) for the case of discrete differentiation. Simply put, if α gets smaller, it acts destabilizing (and we would have to use more polynomials), and if it increases, the solution deviates more and more from the sought result.

As for the overall, practical motivation of the present study, i.e. airfoil theory, the subsection on the streamline curvature, Section 2.3.2, will provide some physical meaning for α by relating it to the Reynolds number via $\alpha = O(Re^{-6/20})$. In normal flight conditions the Reynolds number is of the order of magnitude of 10^4, meaning $\alpha \approx 1/100$ (as chosen for most of the present computations) is reasonable.

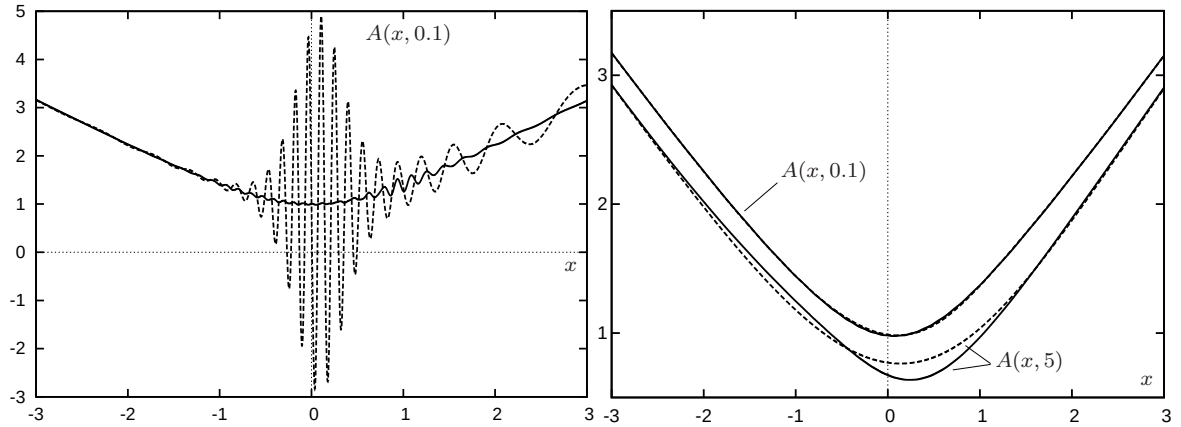

Figure 13: Left: solution of (2.49) at $t = 0.1$ for $\alpha = 10^{-3}$ with $N = 50$ (dashed) and $N = 100$ (solid). Right: comparison of solutions at $t = 0.1$ and $t = 5$ for $\alpha = 0.01$ (dashed) and $\alpha = 1$ (solid).

Not surprisingly, all the results and conclusions established above hold in the same way for the three-dimensional problem. Setting here again $\alpha = 0.01$, the explicit scheme, with sufficiently small Δt, yields the same solutions as the implicit scheme, which allows for larger time steps to study long time behaviors with reasonable computing effort. While the functional relation between Δt, α and N contains the number of polynomials in z as well, their influence on whether the regularization and/or stability holds is negligible. The maximum degree N_z should therefore be chosen with respect to the shape of the function to be approximated (see Example 2.1 and Figure 23 in Section 2.3.3).

Using now the same initial condition (Equation (2.82)) as for the calculations in Figure 11, solutions of the regularized Cauchy problem (2.50) in the discretized form of (2.88), with $\underline{K}$ being the collocation matrix for $\mathcal{A}^*_{3D}$ in (2.85), are shown in Figure 14.

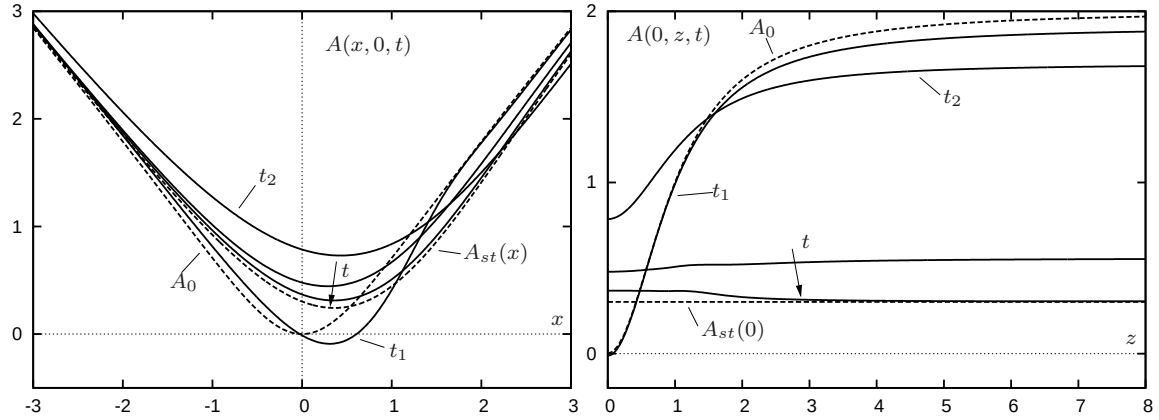

Figure 14: Regularized solution of (2.47) at $t_1 = 0.2$, $t_2 = 2.5$ and as indicated at $t = 7.5$ to $t = 30$ with initial condition A_0 from (2.82) (dashed) and the stationary solution A_{st} (dashed) from Figure 12, parameters $\alpha = 1/100$, $\Delta t = 5 \times 10^{-3}$, $(N_x, N_z) = (50, 20)$. Left: A at $z = 0$. Right: A at $x = 0$.

Remark 2.50. The dynamics in Figure 14 show the indication of convergence of an initially z dependent solution to an equilibrium similar to a steady solution of the two-dimensional case (upper branch, $\Gamma = 2$, cf. Figure 4). This is reasonable, on the one hand, since the local three-dimensionality stems from z dependent perturbations of a planar outer flow (see Section 2.1) and on the other hand, as explained in Remarks 2.14 and 2.22, a solution of (2.33) does satisfy (2.30) and hence the planar (upper branch) steady states can also be attractive equilibria in the three-dimensional time evolution.

So far we have shown by regularizing the Cauchy problem by adding operators, the problem itself was altered and consequently a different problem was solved. Theoretically, as said in Remark 2.42, we would have to prove that solutions of the Cauchy problem containing $\mathcal{A}^*$ can be made arbitrarily close to solutions of the original problem as $\alpha \to 0$. The proof presented in Lavrent'ev et al. (1986) for the homogeneous Cauchy problem, cf. (2.83), cannot be applied here directly, due to the nonlinearity and inhomogeneity, i.e. the mild formulation of the regularized solution (cf. (2.87)). Nevertheless, there are techniques to show the required

convergence for such cases. Since our approach is primarily via approximated solutions and numerical computations, we heuristically state *the more polynomials used, the smaller α can be chosen to act regularizing* and thus, by the overall convergence of the scheme the requirement is met empirically. Table 12 compares the leading coefficients of solutions gained with $N = 300$, $\Delta t = 5 \times 10^{-3}$ at $t = 1$ for different α, which shows the changes in the solution to be negligible, even when decreasing α tenfold.

a_i / α	1×10^{-3}	5×10^{-4}	1×10^{-4}
a_0	0.00577	0.00578	0.00578
a_1	−0.31594	−0.31592	−0.31590
a_2	−0.16895	−0.16895	−0.16895
a_5	0.11644	0.11641	0.11636
a_7	0.09029	0.09033	0.09043

Table 12: Leading coefficients for different α at $t = 1$ with initial condition $a_i \equiv 0$.

Additionally, Table 12 implies that, as $\alpha \to 0$, for certain initial conditions the original (upper branch) steady states, studied in Section 2.2, are approached by the time evolution. And this is the only way how these equilibria should be understood - as *stationary solutions of regularized (well-posed) Cauchy problems in the limit of vanishing regularization.*

Overall, when considering characteristics of the time evolution, one obviously would like to have α as small as possible, to also be quantitatively close to actual solutions. The unsatisfying fact thereby is that then overproportionally high numbers of polynomials have to be used, i.e. while steady states can be calculated to sufficient accuracy with $N = 50$ (see Table 3), the time evolution should be run with $N \geq 200$ (even more in the three-dimensional case).

To avoid adding additional terms to the equation or altering certain operators, we consider again the possibility of filters, as mentioned in Remark 2.43. The parabolic shape of the real parts of the symbols, as shown in (2.77) and (2.79), suggests to proportionally dampen the fast growing parts of the spectrum. The following excursus shows how this can be achieved in an easy, but carefully to use, manner.

Caveat: It is most important to state here that filtering certain parts of a Fourier or singular value decomposition shall be done with care, in order not to cancel out important information on the solution. This holds especially for the choice of how to connect the regularization parameter to the filter and to the necessary convergence - hence, as a general rule, one shall not use implicit schemes for ill-posed problems without having additional information on the solutions.

On the other hand, such a strategy might be closer to what Lemma 2.3 proposes, in the sense that the semigroup $T(t)$ might be continuous on the set of solutions with mollified Fourier decompositions.

Excursus II: Explicit versus Implicit Time Integration

To study the difference between explicit and implicit (or forward and backward) schemes we consider again the usual abstract homogeneous Cauchy problem on some Banach space

$$\partial_t u(t) = \mathcal{A}u(t), \quad u(0) = u_0$$

and assume the solution can be expanded into a Fourier series, i.e.

$$u(x,t) = \sum_{k\in\mathbb{Z}} \hat{u}_k(t) e^{i\langle k,x\rangle}.$$

Let the operator $\mathcal{A}$ possess a symbol, then substitution of the ansatz yields, using Lemma 2.2 or equation (2.71),

$$\sum_{k\in\mathbb{Z}} \partial_t \hat{u}_k(t) e^{i\langle k,x\rangle} = \sum_{k\in\mathbb{Z}} sb(\mathcal{A})\, \hat{u}_k(t) e^{i\langle k,x\rangle}.$$

In other words, the Fourier coefficients shall satisfy

$$\partial_t \hat{u}_k(t) = sb(\mathcal{A})\, \hat{u}_k(t), \quad \forall k \in \mathbb{Z}.$$

Next we will do the exact opposite of what is often called *method of lines* and thus utilize the advantage of Fourier multipliers, we only discretize in time. Applying forward differences gives

$$\partial_t \hat{u}_k(t_m) \approx \frac{\hat{u}_k^{m+1} - \hat{u}_k^m}{\Delta t} = sb(\mathcal{A})\hat{u}_k^m \quad \Rightarrow \quad \hat{u}_k^{m+1} = \big(1 + \Delta t\, sb(\mathcal{A})\big)\, \hat{u}_k^m,$$

whereas the backward differences yield

$$\partial_t \hat{u}_k(t_m) \approx \frac{\hat{u}_k^{m+1} - \hat{u}_k^m}{\Delta t} = sb(\mathcal{A})\hat{u}_k^{m+1} \quad \Rightarrow \quad \hat{u}_k^{m+1} = \big(1 - \Delta t\, sb(\mathcal{A})\big)^{-1}\, \hat{u}_k^m.$$

Note that replacing 1 with the identity and the symbol with its according operator transfers the above relations back to the same result as if applying finite differences in time directly to the Cauchy problem.

As usual, we want to study the behavior of the absolute values of the Fourier coefficients over time regarding their summability (cf. (2.53)). Thus, denoting the multipliers $\tilde{q}_e = 1 + \Delta t\, sb(\mathcal{A})$ and $\tilde{q}_i = (1 - \Delta t\, sb(\mathcal{A}))^{-1}$ for the explicit and implicit scheme, respectively, yields

$$|\hat{u}_k^{m+1}| = \Re\, \tilde{q}\, |\hat{u}_k^m| \quad \text{or} \quad |\hat{u}_k^{m+1}| = q^m\, |\hat{u}_k^0|, \quad \Re\, \tilde{q} =: q, \quad \forall k \in \mathbb{Z}.$$

Obviously, we are dealing with a geometrical sequence, where one has for

$$\left.\begin{array}{lll} q=1 & |\hat{u}_k^m| = |\hat{u}_k^0|, & \dots \text{constant} \\ q=-1 & |\hat{u}_k^m| = (-1)^m |\hat{u}_k^0|, & \dots \text{alternating} \\ |q|<1 & |\hat{u}_k^{m+1}| < |\hat{u}_k^m| & \dots \text{decay} \\ |q|>1 & |\hat{u}_k^{m+1}| > |\hat{u}_k^m| & \dots \text{growth} \end{array}\right\} \quad \forall m. \tag{2.89}$$

For the sake of presentability we only consider symbols in the form of $\Re\, sb(\mathcal{A}) = c|k|^a$, $k \in \mathbb{R}$. As stated in Lemma 2.4 for well-posedness of the Cauchy problem it is necessary and sufficient for the real part of the symbol to be bounded from above. In the example here this reduces to the sign of c.

Without loss of generality say $c = 1$ and consider the operator $\Re\, \mathcal{A}_{2D} \propto |k|^{9/4}$ from (2.77), hence

$$q_e = 1 + \Delta t\, |k|^{9/4}, \qquad q_i = \frac{1}{1 - \Delta t\, |k|^{9/4}}. \tag{2.90}$$

By formally substituting $c = -1$ and thus obtaining a well-posed problem, we obtain according multipliers, which shall be denoted by q_e^* and q_i^*. These multipliers are depicted in Figure 15.

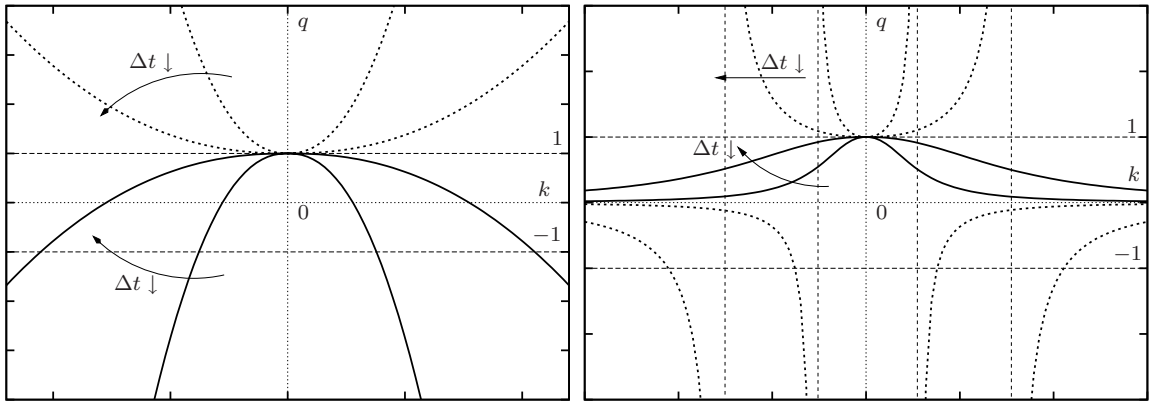

Figure 15: The multipliers from (2.90) as functions of k for various Δt. Left: q_e^* (solid) and q_e (dashed), right: q_i^* (solid) and q_i (dashed).

It becomes absolutely clear from Figure 15 (left) why the explicit scheme used in Figure 10 cannot work at all, independently of how small Δt is chosen. Since every value of k defines a Fourier coefficient $\hat{u}$ and a multiplier $q(k)$ it is also easy to see that the more polynomials are used for a truncated Fourier series (or in the case here, Chebyshev series), the faster the absolute value of the unknown function u grows.

As for the well-posed situation, i.e. the solid lines in Figure 15 corresponding to $q_e^* = 1 - \Delta t|k|^{9/4}$, the smaller the time step gets, the more coefficients lie within the strip ± 1 and thus their value decreases over time, cf. (2.89).

The more important implications for the case studied can be drawn from the graphs on the right of Figure 15, i.e. the implicit scheme. Here, the well-known unconditional stability of implicit schemes for well-posed problems is depicted by $q_i^*(k) = (1 + \Delta t|k|^{9/4})^{-1}$ (solid lines), which remain, independently of Δt, within the strip ± 1 and thus the coefficients decay for all times.

Also, we have thus proved as to why the implicit scheme *can work regularizing* for ill-posed problems. Considering the dashed lines, it is obvious that the coefficients $\hat{u}_k$, for $k \gg 1$ decay for all values of Δt, but this does not imply unconditional stability, since the smaller Δt, the less multipliers q_i have absolute value less than 1. On the other hand, for those k where $q_i > 1$, decreasing Δt means slowing down the growth, by having more and more $q_i(k)$ just slightly greater than 1 (for example say $q_i(k) = 1.001^{100} \approx 1.105$, such that the according $\hat{u}_k$ has grown only 10% at $m = 100$). When choosing instead Δt too large, although damping more $\hat{u}_k$, those with associated $q_i(k) > 1$ are much more amplified. In general, we thus state that

the implicit time integration filters the fast growing parts of a Fourier (or other types of orthogonal) decomposition of the solution, provided the time step is neither too small nor too large.

For the fully discretized system (2.88) the interval Δt can be taken of, depends, of course, on the number of polynomials appearing in the spatial expansion. As a rule (implying some type of convergence), one can claim that the more polynomials used, the larger the allowed interval for the time step gets. Ergo, implicit schemes for ill-posed problems are not unconditionally stable.

Remark 2.51. Needless to say, the above describes exactly the situation of the forward and backward (one-dimensional) heat equation, where all coefficients $\hat{u}_k \to 0$, $\forall k \neq 0$ in the well-posed case, for their multipliers lie all within the strip ± 1 (with sufficiently small Δt for q_e^*) and $\hat{u}_0(t) = \hat{u}_0(0)$, i.e. indicating that with Dirichlet boundary conditions (equal on both boundaries) the solution tends to the constant, namely $\hat{u}_0$.

Finally, the multipliers applied to the regularizing operators (2.85) are given as

$$q_e = 1 + \Delta t\,|k|^{9/4}(1 - \alpha|k|^{3/2}), \qquad q_i = \frac{1}{1 - \Delta t\,|k|^{9/4}(1 - \alpha|k|^{3/2})}. \tag{2.91}$$

Remark 2.52. As mentioned in Remark 2.44, in principle there are always regions of the regularized spectra which have positive real parts. For the multipliers in discretized time integration here this means that there are k, such that $q(k) > 1$, i.e. growing coefficients. As indicated in Figure 16, this can be alleviated by decreasing the time step, although one shall not be mislead by the graphs, since $\alpha = 1$ and hence the denominator in q_i has no real zeros. Decreasing α in q_e results in more k where $q_e(k) > 0$ and in larger absolute values of these $q_e(k)$, but a remedy again is lowering Δt.

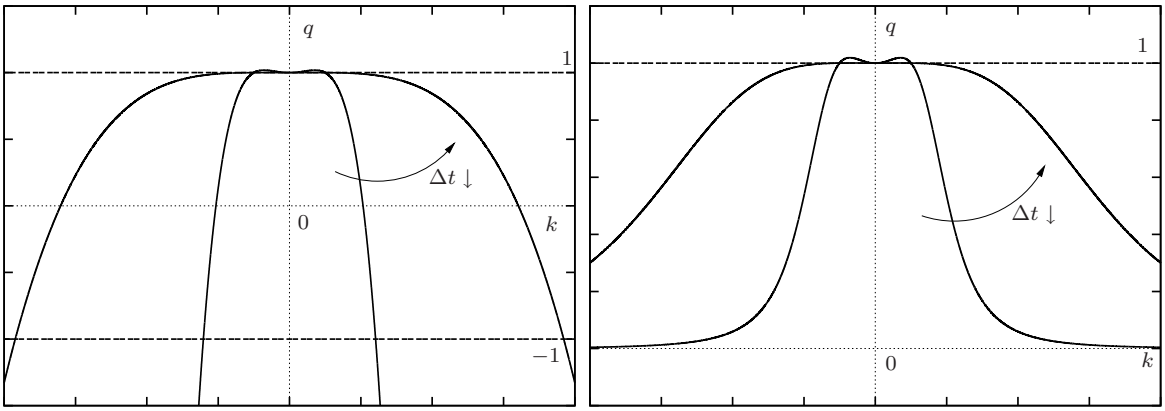

Figure 16: The multipliers for the regularized operators as given in (2.91) with $\alpha = 1$, left: q_e, right: q_i.

As for the implicit case, the interplay between α and Δt can lead to singularities in q_i, similar to the non-regularized situation (cf. Figure 15). In general, decreasing α needs a decrease in Δt as well, to have a non-negative q_i, see Figure 17. All conclusions made above also hold for problems in more than one dimension in the exact same way.

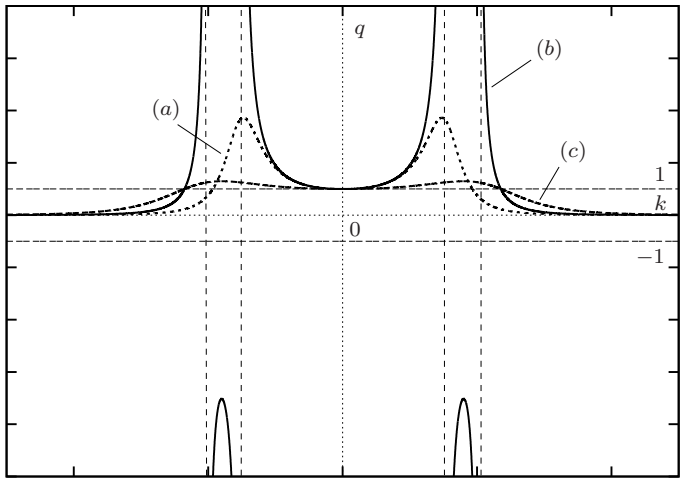

Figure 17: The multiplier $q_i = q_i(\alpha, \Delta t)$ from (2.91). (a): $\alpha = \frac{1}{10}$, $\Delta t = \frac{1}{2}$, (b): $\alpha = \frac{5}{100}$, $\Delta t = \frac{1}{2}$, (c): $\alpha = \frac{5}{100}$, $\Delta t = \frac{1}{10}$

Caveat: Comparing the multipliers in Figures 16 (right) and 17 (right) shows the difference in how Fourier coefficients are damped between directly solving an ill-posed problem with an implicit method and the regularized approach. Thus, one needs additional information to provide some meaning to solutions resulting from the direct scheme. To put this in another

way, one can only be sure in the case of well-posedness to filter the right coefficients when using an implicit method.

Since we have shown that solutions of the Cauchy problems (2.47) and (2.49) have to be understood as limits of the according regularization, a comparison to the direct implicit approach connects the regularizing operator to the filter technique. As done in several occasions above we consider the leading coefficients of the Chebyshev expansion of the solution at time $t = 1$ for both methods (for the two- and three-dimensional problem), see Table 13.

a_i	regularized	direct
a_0	0.00577	0.00578
a_1	−0.31594	−0.31590
a_2	−0.16895	−0.16896
a_5	0.11644	0.11633
a_7	0.09029	0.09045

a_{ik}	regularized	direct
a_{00}	−0.05247	−0.05059
a_{02}	0.93001	0.92473
a_{10}	−0.32438	−0.30444
a_{12}	0.28195	0.27024
a_{20}	−0.16327	−0.16213

Table 13: Leading coefficients at $t = 1$. Left: initial condition $a_i \equiv 0$, values for the regularized (from Table 12, $\alpha = 10^{-3}$) and direct method using the same numerical parameters. Right: initial condition $a_{ik} = 0$, $a_{02} = 1$, $(N_x, N_z) = (50, 20)$, for the regularized (using $\Delta t = 5 \times 10^{-3}$, $\alpha = 2/100$) and the direct method (with $\Delta t = 0.1$).

As mentioned earlier for our general findings on the direct implicit method, the time step has to be chosen out of a certain interval, which depends strongly on the number of polynomials used. This becomes quite apparent in Table 13 (right), where $\Delta t = 0.1$ was set for the direct approach, which was more or less the lowest time step possible for $(N_x, N_z) = (50, 20)$. Taking into account that this is almost three orders of magnitude larger than the time step in the regularized case clarifies, why the coefficients differ already in the third significant digit. In other words, the difference in the time step contributes stronger to different values of the coefficients than the methods applied.

For the sake of completeness, Figure 18 shows the time evolution of the two-dimensional problem gained with the direct implicit scheme starting from $A_0(x) = \sqrt{1 + x^2}$ and approaching the stationary solution as depicted in Figure 3(a). This behavior is qualitatively and even quantitatively indistinguishable from the results obtained with the regularizing operator and shown in Figure 12.

Concluding the excursus, by referring to Remark 2.42 (i.e. regularizing operators are not unique), it is worth mentioning that we found two different methods for solving the ill-posed Cauchy problems at hand, where the according limits of the schemes result in the same solutions on bounded time internals. Hence we sufficiently established in which sense these solutions are to be understood.

End of Excursus II

Considering the regularized time evolution, the steady states given by (2.33) and (2.30) admit some kind of stability. In classical dynamical systems theory linearizing at some equilibrium

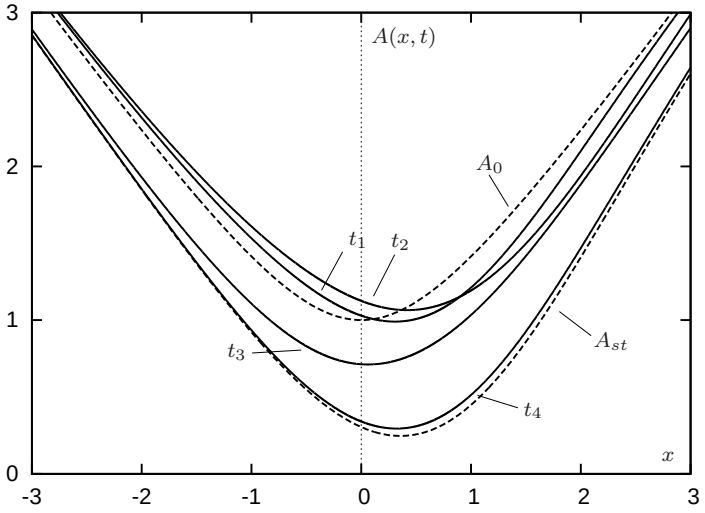

Figure 18: The solution $A(x,t)$ from the direct implicit method using $\Delta t = 1/100$, $N = 100$, $A_0(x) = \sqrt{1+x^2}$ (dashed), at $(t_1, z_2, t_3, t_4) = (0.5,\ 1,\ 5,\ 10)$ (solid lines) and the steady state $A_{st}(x)$ (dashed)

yields a Jacobian matrix, where it is well-known that if all real parts of its eigenvalues are less or equal to zero the equilibrium is stable. Thus, Remark 2.44 would stand in contrast to the numerical findings. This is due to neglecting the nonlinear term when calculating the upper bounds of the symbols. In fact, considering the dispersion relation, including the linearization, as done in Remark 2.37, and regularization one has

$$\omega(k,\alpha) = c_1|k|^{9/4} - c_2(\alpha)|k|^{15/4} - c_3 A_{st}|k|^{3/4}, \tag{2.92}$$

where $\omega(k,0)$ is not bounded from above ($c_2(0) = 0$) but admits negative values for small $|k|$, which strongly depend on A_{st}. On the other hand, one can find combinations of α and A_{st}, such that $\omega(k,\alpha) \leq 0$ for all k.
Now say $\big(S(t)\big)_{t\geq 0}$ is the semigroup generated by the linearized operator, i.e. $\omega(k,\alpha)$, and u^* is a steady state, i.e. $S(t)u^* = u^*$, $\forall t$. Given some u close to u^*, such that u lies in the domain of $S(t)$, then

$$\begin{aligned}\|S(t)u - u^*\| &= \|S(t)u - S(t)u^* + S(t)u^* - u^*\| \leq \\ &\leq \|S(t)u - S(t)u^*\| + \underbrace{\|S(t)u^* - u^*\|}_{=0} \leq \|S(t)\|\,\|u - u^*\|.\end{aligned}$$

Thus, $\|S(t)\| \to 0$ as $t \to \infty$ (i.e. exponentially and strongly stable semigroups, see Engel & Nagel (2000)) is sufficient for u^* to be asymptotically stable, whereas for Lyapunov stability one only needs $\|S(t)\| \leq \infty$ (i.e. continuity of $S(t)$). Both conditions are certainly satisfied if the growth bound, cf. (2.64), is less than zero, which implies that the spectral bound (here the upper bound of $\omega(k,\alpha)$) is also less than zero. For Lyapunov stability the growth bound

can also be zero. But since the assumption of $A_{st} = const.$ is highly artificial (except for some limiting cases), one does not gain any substantial insight into the subject of stability from such a relation. Still, the filter technique revealed the same type of stability as the regularized operators, although (cf. Figure 15) some Fourier coefficients are amplified. Interestingly, with appropriate numerical parameters (i.e. a stable scheme) we were not able to see any onset of destabilization from the steady state as depicted in Figure 12 (even for, say $t \geq 10000$). Overall, for regularizations are normally viewed on finite time intervals, we will not go into further details on the (long time) stability of the stationary solutions.

Remark 2.53. By reversing the time in the original Cauchy problems one can also obtain an upper bound for the resulting real parts of the symbols. This is straight forwardly done by just changing the sign in (2.77) and (2.79) and also holds for the full dispersion relation, cf. $-\omega(k,0) \leq const.$, $\forall k$ in (2.92). This is very well supported by using the numerical methods described above, explicit as well as implicit, with negative time steps. Despite starting from arbitrary initial conditions, such as $A_0(x) = \sqrt{1+x^2}$, the according time evolution does not show any significant dynamics, the solutions do not experience any oscillations. In virtue of $-\omega(k,0)$ the according multipliers q_e and q_i as defined in Excursus II might be greater than one for some (small) k and hence we have to carefully choose the time step.

The most interesting consideration here is to go backward in time from some previously computed regularized solution at some $0 < t < T$. Take, for example, the solution $A(x,5)$ depicted in Figure 18 at $t = 5$. By now running the explicit Euler scheme (with small enough Δt to be stable), backward in time without any regularization and using as an initial condition the solution $A_0(x) = A(x,5)$, one can observe that at all times $5 - t$ the solution passes through every regularized solution at that time and eventually approaches the original initial condition $A_0 = \sqrt{1+x^2}$. Therefore we have found another way to provide some meaning to the filter automatically applied by implicit schemes. That is, one can allow to dampen decomposition values of solutions, if these solutions correspond to initial conditions with certain regularities.

The so far presented time evolution results did not include the perturbation g. In Section 2.2 for certain three-dimensional steady states a hump, cf. (2.45), was taken into account. The fundamental equation (2.28) also includes suction or blowing devices of the form

$$g(x,z,t) = -\gamma \mathcal{J}_{-\infty}^{3/4} v_w(x,z,t),$$

where v_w represents the suction or blowing velocity at the wall (and perpendicular to it), see Section 2.1. Scheichl et al. (2008), for example, used $v_w(x,t) = V(x)G(t)$ with

$$V(x) = \mathbb{1}_{[x_c - l/2, x_c + l/2]}(x), \quad G(t) = \mathbb{1}_{[0,T^*]}(t),$$

x_c representing the center of the blowing slot and l its length. The steps in the function G were mollified to yield a continuous "switch on – switch off" behavior. As for the three-dimensional problem we multiply v_w by $p = a\,p(z)$, $a \in \mathbb{R}$, which has to decay to zero, e.g.

$p(z) = (1+z^2)^{-1}$. Overall, this yields a perturbation of the form

$$g(x,z,t) = -\gamma a\, G(t)\, p(z)\, \mathcal{J}_{-\infty}^{3/4} V(x), \tag{2.93}$$

where the Abel operator applied to the characteristic function can be easily given by a closed formula.

A physical (maybe even experimental) idea to study the time evolution is to create a laminar, steady boundary layer flow and then impose certain disturbances. It is thus interesting to also consider the case where the initial condition is set to be the steady state solution at the upper branch, $A_0 = A_{st}$, cf. Figure 3. From the evidence on stability mentioned on several occasions above, one can claim, that $A(x,t) = A_{st}(x)$, $\forall t$, with $g \equiv 0$ (and the numerical findings confirm this). So by having a non-zero disturbance in the sense that

$$\int_0^{T^*} \|g(t)\|_{L^1}\, dt = const. \neq 0, \tag{2.94}$$

the solution moves away from the equilibrium. Taking T^* less than the overall time for which the regularization holds, reveals additional insight into the dynamics of the system. First, if the constant in (2.94) is too small this means $A(\cdot, T^*)$ is still within the *basin of attraction* of A_{st} and thus the solution has to reapproach its equilibrium. Again, this is sufficiently confirmed by numerical computations, yielding additional evidence for the stability of the upper branch steady states (take, for example, $a = 1$ in (2.93) and $T^* = 2$). Second, strong enough perturbations pushing the solution outside the basin of attraction result in a *finite time blow-up scenario*, which we will study in Section 2.3.3.

2.3.2 Regularization and Higher Order Asymptotic Expansions

We have demonstrated in the above the independence of regularized solutions regarding the regularizing techniques, i.e. higher derivatives versus direct implicit time integration, and thus the following shall provide some additional and, more importantly, physical meaning to the ill-posedness and its regularization.

As pointed out by A. I. Ruban (private communications), when considering the regularizing operators (2.85), the third derivative term, with A being the displacement function as before, relates to the streamline curvature in the main deck. Of course, within the interaction region, due to the scalings of the variables and the considered order of approximation, the streamline curvature is neglected, but since the arising instabilities have much shorter length scales, curvature effects might not be negligible anymore, such that a y gradient of pressure perturbations is induced.

We will investigate such contributions by presenting some heurisitcal aspects and arguments to see how one can find the proposed pressure terms and in what way they relate to the fundamental problem in terms of A, see Section 2.1. Then the method of matched asymptotic

expansions is utilized to embed the heuristical findings into a more strict and substantiated setting.

Let us, for the moment, consider the situation of planar flows along some surface, which admits a non-zero curvature $\kappa = \kappa(x)$, then it is common to write the Navier-Stokes equations (2.2) in curvilinear coordinates (cf. the introduction in Section 2.1). By applying classical, second order boundary layer theory, one obtains

$$\partial_y p \propto \pm \kappa\, u^2,$$

where the sign depends on the orientation of the surface. Since we still have no relevant surface curvature within the interaction region (note that the hump, if included, has to be chosen according to restrictions on its curvature), we shall expect such a κ to result from strongly bent streamlines within the boundary layer.

Basically, from the definition of a streamline, its slope is proportional to v/u and consequently $\kappa \propto |\partial_x(v/u)|$. Assuming further u to be independent of x, which holds at leading order for the main deck expansion (cf. $u_{20} = U_0(y_2)$ in (2.10)), one obtains $\kappa \propto |1/u\, \partial_x v|$, relating to the pressure via

$$\partial_y p \propto |u\, \partial_x v|. \tag{2.95}$$

From the expansions and the coordinate scalings for the main deck given in (2.10), we have established that leading order terms represent a steady and planar flow. Hence, the arguments above do apply here. Heuristically, one would now conclude, in virtue of (2.13), the pressure perturbations are of equal order of magnitude as the (so far y_2 independent) interaction pressure, leading to

$$\partial_{y_2} p \propto |\partial_x^2 A|.$$

Introducing this to the fundamental equation (2.28) would mean that we can include higher derivatives of A in the problem, by taking the streamline curvature effects into account. Certain issues remain at this point, as to how the y gradient of some pressure perturbation can be introduced in the procedure of deriving the solvability condition, at which order (with respect to the Reynolds number) this occurs and, most importantly, which sign additional derivatives of A finally have (compared to the time derivative term) to actually act regularizing.

As said above, incorporating the streamline curvature has to originate in the main deck, hence we shall make a formal extension to the existing expansions (2.10). Say $n_1, \ldots, n_4 \in \mathbb{N}$ (and obviously $n_1 > 4$, $n_2, n_3, n_4 > 10$) we then expand the flow field as

$$\begin{aligned} u_2 &\sim U_0 + \epsilon^4 u_{21} + \cdots + \epsilon^{n_1} u_n \\ v_2 &\sim \epsilon^{10} v_{21} + \cdots + \epsilon^{n_2} v_n \\ w_2 &\sim \epsilon^{10} w_{21} + \cdots + \epsilon^{n_3} w_n \\ p_2 &\sim \tfrac{1-U_{00}^2}{2} + \epsilon^4 p_{00} x + \epsilon^{10} p_i + \cdots + \epsilon^{n_4} p_n \end{aligned} \tag{2.96}$$

with p_i denoting the interaction pressure as found in Section 2.1.

Remark 2.54. The dots in the expansions (2.96) shall symbolize that, depending on the values of $n_1, \dots, n_4$, one might still have non-zero expansion terms at certain orders lying inbetween. Also, from the definition of $\epsilon = Re^{-1/20}$ it is reasonable to assume all orders, i.e. exponents of ϵ, to be natural numbers. Of course, as it is always the case with such formal asymptotic techniques, there is no guarantee that integer powers of the given ϵ are the only possibilities to define expansion orders. Nevertheless, the plausibility of the resulting equations and matching rules provides sufficient argumentation for such an assumption.

As usual, we substitute the expansions (2.96) into the Navier-Stokes equations (2.2). Then, as has already been shown for the main deck in Section 2.1, one obtains from the momentum and mass balance at order ϵ^0

$$\begin{aligned} &U_0 \partial_x u_{21} + v_{21} U_0' = U_0'' - p_{00} \\ &\partial_{y_2} p_i = 0 \\ &\operatorname{div}(u_{21}, v_{21}) = 0, \end{aligned} \tag{2.97}$$

which can be integrated to obtain u_{21} and v_{21} in the form known from Section 2.1, Equation (2.13), as

$$\begin{aligned} v_{21} &= -U_0 \left(\frac{\partial_x A}{p_{00}} + \int_0^{y_2} \frac{U_0''(s) - p_{00}}{U_0^2(s)} ds \right) \\ u_{21} &= U_0' \left(\frac{A}{p_{00}} + x \int_0^{y_2} \frac{U_0''(s) - p_{00}}{U_0^2(s)} ds \right) + x \frac{U_0'' - p_{00}}{U_0}. \end{aligned} \tag{2.98}$$

The next order of interest here is ϵ^6, as it contains the term $U_0 \partial_x v_{21}$ in the y momentum equation on the left hand side. In virtue of (2.95) this shall be proportional to the y gradient of the pressure. Since so far the n_i are arbitrary, the momentum equations contain terms, such as $\epsilon^{n_1 - 1} \partial_t u_n$ or $\epsilon^{n_2} v_n \partial_{y_2} v_{21}$ or $-\epsilon^{n_4 - 4} \partial_x p_n$ or $\epsilon^{n_3} [(\epsilon^{12} \partial_x^2 + \partial_{y_2}^2 + \epsilon^{12} \partial_z^2) w_{22}]$, to name but a few. The crucial term obviously appears in the y momentum balance in the form of

$$-\epsilon^{n_4 - 10} \partial_{y_2} p_n,$$

where we can now choose $n_4 = 16$, such that at order ϵ^6 we have

$$U_0 \partial_x v_{21} = -\partial_{y_2} p_n. \tag{2.99}$$

Remark 2.55. Considering *all* possible combinations of $n_1, \dots, n_4$, together with their constraints $n_1 > 4$ and $n_2, n_3, n_4 > 10$, shows that only the y gradient of the pressure at order $n_4 = 16$ can enter the equation at order ϵ^6. In other words, the y gradient of the pressure p_n is uniquely determined by (2.99), independently of the values of $n_1, \dots, n_3$, and $n_4 = 16$ is the only possibility to have such a gradient proportional to $U_0 \partial_x v_{21}$. Additionally, as required from the principles of matched asymptotic expansions, the lower order equations, e.g. (2.97), determining all lower order terms, remain unchanged.

Remark 2.56. Observe, that the equation yielding the new pressure term p_n only involves lower order terms, which are already known. With $n_4 = 16$, the pressure expansion now reads

$$p_2 \sim \tfrac{1-U_{00}^2}{2} + \epsilon^4 p_{00} x + \epsilon^{10} p_i + \cdots + \epsilon^{16} p_n,$$

where we obviously have several (integer) orders between the interaction pressure and the new term. It is yet to be determined whether there are (relevant) pressure contributions between the expansion orders 10 and 16. Nevertheless, (2.99) can be viewed as an ordinary differential equation with respect to y_2 and hence substituting v_{21} from (2.98) and integration yields

$$\partial_{y_2} p_n = \frac{U_0^2(y_2)}{p_{00}} \partial_x^2 A \quad \Rightarrow \quad p_n(x, y_2, z, t) = p_n(x, 0, z, t) + \frac{1}{p_{00}} \partial_x^2 A \int_0^{y_2} U_0^2(y) dy.$$

Thus we found a higher order pressure contribution containing higher derivatives of the function A. The matching rule then demands in principle

$$p_1 = p_2 \quad \text{as} \quad y_2 \to \infty,\ y_1 \to 0,$$

which can be written as (we shall demonstrate this later in more detail)

$$p_n(x, 0, z, t) - \frac{c}{p_{00}} \partial_x^2 A = p_{1n}(x, 0, z, t),$$

where p_{1n} would be the according higher order term in the upper deck and c some positive constant. Since here we do not consider or change the original upper deck expansions, $p_{1n} = 0$. Therefore, we get

$$p_n(x, 0, z, t) = \frac{c}{p_{00}} \partial_x^2 A.$$

In Section 2.1 we defined the interaction pressure p_i to be p_{32}, which, for being independent of y_3, can also be defined to be p_{32} as $y_3 \to \infty$ (which matches with the main deck pressure at $y_2 = 0$). This then means, by writing the pressure expansion in the lower deck as

$$p_3 \sim \tfrac{1-U_{00}^2}{2} + \epsilon^4 p_{00} x + \epsilon^{10}(p_{32} + \epsilon^6 p_{3n}),$$

such that we can redefine $p_{in} := p_{32} + \epsilon^6 p_{3n} = p_{32} + \epsilon^6 p_n|_{y_2=0}$ and view this as an asymptotic expansion of the interaction pressure, we arrive at

$$p_{in}(x, z, t) = p_i(x, z, t) + \alpha \partial_x^2 A(x, z, t), \quad \alpha = Re^{-3/10} \frac{c}{p_{00}} \to 0.$$

By substituting p_{in}, using (2.29) for p_i, into (2.28) one can derive a new fundamental problem for A, including an additional (regularizing) operator with some arbitrary but small and positive parameter α.

Remark 2.57. As asymptotic expansions are not unique, one can not claim that there are actual "right" or "wrong" expansions and since convergence in any sense is not an issue, every ansatz is justified, as long as the principles are applied correctly. Hence, choosing an expansion can rather be a matter of physical interpretation and relevance to the original equations. Here, for example, p_n has been shown to relate to the streamline curvature. Still, ignoring the dots in (2.96) and the expansions in the other decks and assuming the next relevant order (for the pressure) is 16 would be a very crude application of the principles of asymptotic expansions, for which we neither have a mathematical nor physical argument.

With the above remark in mind, we shall now take into account all three decks and, by being as thorough as possible, expand the flow fields in consecutive integer orders. To make the analysis more accessible, the expansions, resulting equations and (partial) solutions governing the higher order terms are presented first for all decks, such that a subsequent application of the according matching rules finally leads to the sought results.

The upper deck. To abridge the calculations, in virtue of the lower deck expansions as given in Braun et al. (2012) for the two dimensional flow case, we claim the next relevant pressure term to be at order 13. Thus, we write the expansions for the *upper* deck as

$$\begin{aligned}
u_1 &\sim U_{00} - \epsilon^4 U_{01} x + \epsilon^{10} u_{12} + \epsilon^{11} u_{13} + \epsilon^{12} u_{14} + \epsilon^{13} u_{15} + \epsilon^{14} u_{16} + \epsilon^{15} u_{17} + \epsilon^{16} u_{18} \\
v_1 &\sim \epsilon^4 U_{01} y_1 + \epsilon^{10} v_{12} + \epsilon^{11} v_{13} + \epsilon^{12} v_{14} + \epsilon^{13} v_{15} + \epsilon^{14} v_{16} + \epsilon^{15} v_{17} + \epsilon^{16} v_{18} \\
w_1 &\sim \epsilon^{10} w_{12} + \epsilon^{11} w_{13} + \epsilon^{12} w_{14} + \epsilon^{13} w_{15} + \epsilon^{14} w_{16} + \epsilon^{15} w_{17} + \epsilon^{16} w_{18} \\
p_1 &\sim \frac{1 - U_{00}^2}{2} + \epsilon^4 p_{00} x + \epsilon^{10} p_i + \epsilon^{13} p_{13} + \epsilon^{14} p_{14} + \epsilon^{15} p_{15} + \epsilon^{16} p_{16}.
\end{aligned}$$

Substitution into the conservation of mass shows

$$\operatorname{div}(u_{1j}, v_{1j}, w_{1j}) = 0, \quad j \geq 2, \tag{2.100}$$

whereas the momentum equations reveal at orders ϵ^7 and ϵ^8

$$\begin{aligned}
U_{00}\partial_x u_{13} &= 0 \qquad & U_{00}\partial_x u_{14} &= 0 \\
U_{00}\partial_x v_{13} &= 0\ , \qquad & U_{00}\partial_x v_{14} &= 0 \\
U_{00}\partial_x w_{13} &= 0 \qquad & U_{00}\partial_x w_{14} &= 0.
\end{aligned}$$

Obviously, for both velocity terms, indexed 13 and 14, the trivial solution is possible, as we do not know any matching condition at $y_1 = 0$ (yet). Nevertheless, we will (for now) assume them non-zero, just to demonstrate their possible contributions. For the orders ϵ^9 through ϵ^{12} the procedure of differentiating the three momentum equations with respect to x, y_1 and z, respectively, adding the resulting three equations and using (2.100), shall be applied. In the following, if not stated otherwise, all equations are assumed to hold for all $t \in [0, T]$.

Thus, we obtain at order ϵ^9

$$\begin{array}{ll} U_{00}\partial_x u_{15} &= -\partial_x p_{13} \\ U_{00}\partial_x v_{15} &= -\partial_{y_1} p_{13} \\ U_{00}\partial_x w_{15} &= -\partial_z p_{13} \end{array} \quad \Rightarrow \quad \begin{array}{ll} \Delta p_{13} = 0 & \text{on } \mathbb{R}\times\mathbb{R}^+\times\mathbb{R} \\ \partial_{y_1} p_{13} = -U_{00}\partial_x v_{15} & \text{at } y_1 = 0. \end{array} \tag{2.101}$$

This Neumann problem posed for p_{13} is exactly the same as the one for the interaction pressure p_i, cf. (2.8), with the same general solution, v_{15} replacing v_{12}.

The next higher order pressure term p_{14} has to satisfy a Neumann problem for the Poisson equation, since at order ϵ^{10} we find

$$\begin{array}{lll} U_{01}x\ \partial_x u_{12} - U_{01}y_1\ \partial_{y_1} u_{12} + U_{01}u_{12} + & U_{00}\partial_x u_{16} = -\partial_x p_{14} \\ U_{01}x\ \partial_x v_{12} - U_{01}y_1\ \partial_{y_1} v_{12} - U_{01}v_{12} + & U_{00}\partial_x v_{16} = -\partial_y p_{14} \\ U_{01}x\ \partial_x w_{12} - U_{01}y_1\ \partial_{y_1} w_{12} + & U_{00}\partial_x w_{16} = -\partial_z p_{14} \end{array} \tag{2.102}$$

$$\Rightarrow \quad \begin{array}{ll} \Delta p_{14} = 2U_{01}(\partial_{y_1} v_{12} - \partial_x u_{12}) =: F & \text{on } \mathbb{R}\times\mathbb{R}^+\times\mathbb{R} \\ \partial_{y_1} p_{14} = -U_{01}x\ \partial_x v_{12} + U_{01}v_{12} - U_{00}\partial_x v_{16} =: f & \text{at } y_1 = 0. \end{array}$$

Assuming the solvability condition for the Poisson problem (i.e. the integral over the inhomogeneity equals the integral over the boundary condition) to be satisfied (otherwise the Neumann problem does not even have a solution), we can write the general solution formula using the Neumann Green's function from (2.8). We will not go into any further details of this problem at the moment and proceed with the next order.

At order ϵ^{11} we have

$$\begin{array}{lll} \partial_t u_{12} + U_{01}x\ \partial_x u_{13} - U_{01}y_1\ \partial_{y_1} u_{13} + U_{01}u_{13} + & U_{00}\partial_x u_{17} = -\partial_x p_{15} \\ \partial_t v_{12} + U_{01}x\ \partial_x v_{13} - U_{01}y_1\ \partial_{y_1} v_{13} - U_{01}v_{13} + & U_{00}\partial_x v_{17} = -\partial_{y_1} p_{15} \\ \partial_t w_{12} + U_{01}x\ \partial_x w_{13} - U_{01}y_1\ \partial_{y_1} w_{13} + & U_{00}\partial_x w_{17} = -\partial_z p_{15} \end{array} \tag{2.103}$$

$$\Rightarrow \quad \begin{array}{ll} \Delta p_{15} = 2U_{01}(\partial_{y_1} v_{13} - \partial_x u_{13}) & \text{on } \mathbb{R}\times\mathbb{R}^+\times\mathbb{R} \\ \partial_{y_1} p_{15} = -\partial_t v_{12} - U_{01}x\ \partial_x v_{13} + U_{01}v_{13} - U_{00}\partial_x v_{17} & \text{at } y_1 = 0. \end{array}$$

Again, this represents the Neumann problem for the Poisson equation and the same arguments made above hold. Similarly, the problem at order ϵ^{12} for p_{16} reads

$$\begin{array}{lll} \partial_t u_{13} + U_{01}x\ \partial_x u_{14} - U_{01}y_1\ \partial_{y_1} u_{14} + U_{01}u_{14} + & U_{00}\partial_x u_{18} = -\partial_x p_{16} \\ \partial_t v_{13} + U_{01}x\ \partial_x v_{14} - U_{01}y_1\ \partial_{y_1} v_{14} - U_{01}v_{14} + & U_{00}\partial_x v_{18} = -\partial_{y_1} p_{16} \\ \partial_t w_{13} + U_{01}x\ \partial_x w_{14} - U_{01}y_1\ \partial_{y_1} w_{14} + & U_{00}\partial_x w_{18} = -\partial_z p_{16} \end{array} \tag{2.104}$$

$$\Rightarrow \quad \begin{array}{ll} \Delta p_{16} = 2U_{01}(\partial_{y_1} v_{14} - \partial_x u_{14}) & \text{on } \mathbb{R}\times\mathbb{R}^+\times\mathbb{R} \\ \partial_{y_1} p_{16} = -\partial_t v_{13} - U_{01}x\ \partial_x v_{14} + U_{01}v_{14} - U_{00}\partial_x v_{18} & \text{at } y_1 = 0. \end{array}$$

As we are rather interested here in determining whether pressure terms between the orders 10 and 16 contain derivatives of the displacement function A, we shall make the assumption of (u_{13}, v_{13}, w_{13}) and (u_{14}, v_{14}, w_{14}) to be zero, since these are possible solutions satisfying the necessary decay at infinity. Thus, problems (2.103) and (2.104) can be rewritten as

$$\begin{aligned} \Delta p_{15} &= 0 && \text{on } \mathbb{R}\times\mathbb{R}^+\times\mathbb{R} \\ \partial_{y_1} p_{15} &= -\partial_t v_{12} - U_{00}\partial_x v_{17} && \text{at } y_1 = 0. \end{aligned}$$

and

$$\begin{aligned} \Delta p_{16} &= 0 && \text{on } \mathbb{R}\times\mathbb{R}^+\times\mathbb{R} \\ \partial_{y_1} p_{16} &= -U_{00}\partial_x v_{18} && \text{at } y_1 = 0. \end{aligned}$$

The problem for p_{14}, Equation (2.102), contains u_{12} and v_{12} on the right hand side F and in the boundary condition f. Denoting the according Green function by G (cf. the calculations following (2.8)) we formally write a solution as

$$p_{14} = -\int\limits_{\mathbb{R}\times\mathbb{R}^+\times\mathbb{R}} GF d\xi_1 d\xi_3 d\xi_2 + \int\limits_{\mathbb{R}^2} Gf d\xi_1 d\xi_2,$$

where, in general, one also has an undetermined integration constant proportional to one over the volume of the domain, which evaluates to zero in the present case.

We will return to the pressure terms above later (when deriving the matching rules) and shall just make two more remarks. As it is commonly known, especially in the two dimensional steady flow case, a general solution of Laplace's equation with a zero far field condition (here as $y_1 \to \infty$) has an exponential behavior, such that one may write the interaction pressure and the vertical velocity component as

$$p_i(x, y_1) = p_i(x, 0)e^{-cy_1}, \quad v_{12}(x, y_1) = v_{12}(x, 0)e^{-cy_1}, \tag{2.105}$$

with c being some positive constant. The same can be easily shown to hold in the three dimensional case by applying the method of separation of variables.

Finally, note that p_{16} is at the order of the heuristically derived p_n above, which means, in virtue of the presence of p_{13}, p_{14} and p_{15}, if we would have continued with the crude expansions (2.96), where we also left the upper deck unchanged and p_n followed p_i as the next relevant term, we would have obtained only a fraction of the actually contained information.

The main deck. We write the according expansions as

$$\begin{aligned} u_2 &\sim U_0(y_2) + \epsilon^4 u_{21} + \epsilon^5 u_{22} + \epsilon^6 u_{23} + \epsilon^7 u_{24} + \epsilon^8 u_{25} + \epsilon^9 u_{26} + \epsilon^{10} u_{27} \\ v_2 &\sim \epsilon^{10} v_{21} + \epsilon^{11} v_{22} + \epsilon^{12} v_{23} + \epsilon^{13} v_{24} + \epsilon^{14} v_{25} + \epsilon^{15} v_{26} + \epsilon^{16} v_{27} \\ w_2 &\sim \epsilon^{10} w_{21} + \epsilon^{11} w_{22} + \epsilon^{12} w_{23} + \epsilon^{13} w_{24} + \epsilon^{14} w_{25} + \epsilon^{15} w_{26} + \epsilon^{16} w_{27} \\ p_2 &\sim \frac{1 - U_{00}^2}{2} + \epsilon^4 p_{00} x + \epsilon^{10} p_i + \epsilon^{13} p_{23} + \epsilon^{14} p_{24} + \epsilon^{15} p_{25} + \epsilon^{16} p_{26}, \end{aligned}$$

with the conservation of mass now reading

$$\begin{aligned}
&\mathrm{div}(u_{2j}, v_{2j}) = 0, \quad j = 1, \ldots, 6 \\
&\mathrm{div}(u_{27}, v_{27}, w_{21}) = 0, \\
&\partial_z w_{2j} = 0, \quad j > 1.
\end{aligned} \tag{2.106}$$

The behavior of U_0 is given in (2.11) and formulae for u_{21} and v_{21} are given in (2.98). At orders ϵ^1 through ϵ^3 the momentum and mass balances combined yield simple differential equations, which can be integrated in general to read

$$U_0' v_{22} + U_0 \partial_x u_{22} = U_0' v_{22} - U_0 \partial_{y_2} v_{22} = 0 \quad \Rightarrow \quad \left\{ \begin{array}{l} v_{22} = -U_0(y_2) \partial_x B_2(x, z, t) \\ u_{22} = U_0'(y_2) B_2(x, z, t) \end{array} \right. \tag{2.107}$$

and

$$U_0' v_{23} + U_0 \partial_x u_{23} = U_0' v_{23} - U_0 \partial_{y_2} v_{23} = 0 \quad \Rightarrow \quad \left\{ \begin{array}{l} v_{23} = -U_0(y_2) \partial_x B_3(x, z, t) \\ u_{23} = U_0'(y_2) B_3(x, z, t) \end{array} \right.$$

and

$$U_0' v_{24} + U_0 \partial_x u_{24} = U_0' v_{24} - U_0 \partial_{y_2} v_{24} = 0 \quad \Rightarrow \quad \left\{ \begin{array}{l} v_{24} = -U_0(y_2) \partial_x B_4(x, z, t) \\ u_{24} = U_0'(y_2) B_4(x, z, t) \end{array} \right. \tag{2.108}$$

$$\partial_y p_{23} = 0,$$

where B_2, B_3 and B_4 are so far undetermined functions representing displacement effects similar to A.

At the order ϵ^4 lower order terms enter as inhomogeneities to give

$$\begin{aligned}
&U_0' v_{25} + v_{21} \partial_{y_2} u_{21} + u_{21} \partial_x u_{21} + U_0 \partial_x u_{25} = \partial_{y_2}^2 u_{21} \\
&\partial_y p_{24} = 0,
\end{aligned}$$

with the general solution

$$\begin{aligned}
v_{25} &= -U_0(y_2) \left(\partial_x B_5(x, z, t) + \int_0^{y_2} \frac{h(x, s, z, t)}{U_0^2(s)} ds \right) \\
h &= \partial_{y_2}^2 u_{21} - v_{21} \partial_{y_2} u_{21} - u_{21} \partial_x u_{21},
\end{aligned} \tag{2.109}$$

where B_5 is in analogy to the other B_i's.

At order ϵ^5 we find

$$\begin{aligned}
&U_0' v_{26} + v_{22} \partial_{y_2} u_{21} + v_{21} \partial_{y_2} u_{22} + \partial_t u_{21} + u_{22} \partial_x u_{21} + u_{21} \partial_x u_{22} + U_0 \partial_x u_{26} = \partial_{y_2}^2 u_{22} \\
&\partial_{y_2} p_{25} = 0,
\end{aligned}$$

with the general solution

$$
\begin{aligned}
v_{26} &= -U_0(y_2)\left(\partial_x B_6(x,z,t) + \int_0^{y_2} \frac{h(x,s,z,t)}{U_0^2(s)} ds\right) \\
h &= \partial_{y_2}^2 u_{22} - (v_{22}\partial_{y_2} u_{21} + v_{21}\partial_{y_2} u_{22} + \partial_t u_{21} + u_{22}\partial_x u_{21} + u_{21}\partial_x u_{22}).
\end{aligned}
\tag{2.110}
$$

And finally at order ϵ^6 we have

$$
\begin{aligned}
&U_0' v_{27} + v_{23}\partial_{y_2} u_{21} + v_{22}\partial_{y_2} u_{22} + v_{21}\partial_{y_2} u_{23} + \partial_t u_{22} + u_{23}\partial_x u_{21} + \\
&\qquad + u_{22}\partial_x u_{22} + u_{21}\partial_x u_{23} + U_0\partial_x u_{27} = \partial_{y_2}^2 u_{23} - \partial_x p_i \\
&U_0\partial_x v_{21} = -\partial_{y_2} p_{26} \quad \Rightarrow \quad \partial_{y_2} p_{26} = \frac{U_0^2}{p_{00}}\partial_x^2 A \\
&U_0\partial_x w_{21} = -\partial_z p_i \quad \Rightarrow \quad w_{21} = -\frac{1}{U_0}\int_{-\infty}^{x} \partial_z p_i(s,z,t) ds,
\end{aligned}
\tag{2.111}
$$

such that p_{26} is equal to p_n in (2.96), with the general solution again reading

$$
\begin{aligned}
p_{26}(x,y_2,z,t) &= p_{26}(x,0,z,t) + \frac{1}{p_{00}}\partial_x^2 A \int_0^{y_2} U_0^2(y) dy = \\
&= p_{26}(x,0,z,t) + \frac{U_{00}^2}{p_{00}} y_2 \partial_x^2 A - \frac{U_{00}^2}{p_{00}}\partial_x^2 A \underbrace{\int_0^{y_2}\left(1 - \frac{U_0^2(y)}{U_{00}^2}\right) dy}_{=:c_p},
\end{aligned}
\tag{2.112}
$$

where the integral in the second line exists as $y_2 \to \infty$ and $c_p > 0$ in this limit.

We shall now reconsider the solution formulae for the expansion terms of v_2, as we need them for establishing a working matching rule. Note that v_{22}, v_{23} and v_{24}, Equations (2.107) and (2.108) essentially have the same structure, only differing in the integration "constants" B_i. Since $U_0 \to U_{00}$ for large y_2, these velocity terms evaluate to a function independent of y_2 at the boundary to the upper deck. By the usual matching procedure they consequently have to match with the upper deck expansion terms of v_1 at the same order. But since orders ϵ^{11} and ϵ^{12} are not present in the upper deck, we may also set B_2 and B_3 equal to zero.

The term v_{24} appears at order ϵ^{13} and by the same argument as before has to match with the velocity contribution at the same order in the upper deck, cf. v_{15}. Hence (2.108) remains, and so does (2.109).

Setting $v_{22} = 0$ in (2.110) leaves only $\partial_t u_{21}$ in the function h, such that by using (2.98) we obtain

$$\begin{aligned} v_{26} &= -U_0(y_2)\bigg(\partial_x B_6(x,z,t) - \int_0^{y_2} \frac{\partial_t u_{21}(x,s,z,t)}{U_0^2(s)} ds\bigg) = \\ &= -U_0(y_2)\bigg(\partial_x B_6(x,z,t) - \frac{\partial_t A}{p_{00}} \int_0^{y_2} \frac{U_0'(s)}{U_0^2(s)} ds\bigg). \end{aligned} \tag{2.113}$$

Analogously, in (2.111) substituting $v_{22} = v_{23} = 0$, and combining the x and z momentum equations using (2.106), we obtain

$$\begin{aligned} &U_0' v_{27} - U_0 \partial_y v_{27} = -\partial_x p_i - \int_{-\infty}^{x} \partial_z^2 p_i(s,z,t) ds \\ &\Rightarrow v_{27} = -U_0(y_2)\bigg(\partial_x B_7(x,z,t) - \bigg(\partial_x p_i + \int_{-\infty}^{x} \partial_z^2 p_i ds\bigg) \int_0^{y_2} \frac{1}{U_0^2(s)} ds\bigg). \end{aligned} \tag{2.114}$$

Matching procedures. We start by matching the upper and main deck pressure and vertical velocity v in the general form of

$$\left.\begin{aligned} &\stackrel{!}{=} \begin{cases} \epsilon^{10} p_i + \epsilon^{13} p_{13} + \epsilon^{14} p_{14} + \epsilon^{15} p_{15} + \epsilon^{16} p_{16} \\ \epsilon^{10} p_i + \epsilon^{13} p_{23} + \epsilon^{14} p_{24} + \epsilon^{15} p_{25} + \epsilon^{16} p_{26} \end{cases} \\ &\stackrel{!}{=} \begin{cases} \epsilon^4 U_{01} y_1 + \epsilon^{10} v_{12} + \epsilon^{13} v_{15} + \epsilon^{14} v_{16} + \epsilon^{15} v_{17} + \epsilon^{16} v_{18} \\ \epsilon^{10} v_{21} + \epsilon^{13} v_{24} + \epsilon^{14} v_{25} + \epsilon^{15} v_{26} + \epsilon^{16} v_{27} \end{cases} \end{aligned}\right\} \quad \begin{aligned} &\text{as} \quad y_1 \to 0 \\ &\qquad\ \ y_2 \to \infty. \end{aligned} \tag{2.115}$$

Since Equations (2.98), (2.108), (2.109), (2.113) and (2.114) contain solution formulae in closed form for the individual expansion terms v_2, we evaluate these with respect to large y_2, i.e.

$$\begin{aligned} \epsilon^{10}: &\quad v_{21} \sim -U_{00}\bigg(\frac{\partial_x A}{p_{00}} + c_{21}\bigg) \\ \epsilon^{13}: &\quad v_{24} \sim -U_{00}\partial_x B_4 \\ \epsilon^{14}: &\quad v_{25} \sim -U_{00}\bigg(\partial_x B_5 + c_{25}\bigg) \\ \epsilon^{15}: &\quad v_{26} \sim -U_{00}\bigg(\partial_x B_6 - \frac{\partial_t A}{p_{00}} c_{26}\bigg) \\ \epsilon^{16}: &\quad v_{27} \sim -U_{00}\bigg(\partial_x B_7 - \big(\partial_x p_i + \int_{-\infty}^{x} \partial_z^2 p_i ds\big)\, c_{27}\bigg), \end{aligned} \tag{2.116}$$

where

$$
\begin{aligned}
c_{21} &= \int_0^{y_2} \frac{U_0''(s) - p_{00}}{U_0^2(s)} ds && \sim \begin{cases} y_2 & y_2 \to \infty \\ const. & y_2 \to 0 \end{cases} \\
c_{25} &= \int_0^{y_2} \frac{\partial_s^2 u_{21} - v_{21}\partial_s u_{21} - u_{21}\partial_x u_{21}}{U_0^2(s)} ds && \sim \begin{cases} y_2 & y_2 \to \infty \\ y_2^{-1} & y_2 \to 0 \end{cases} \\
c_{26} &= \int_0^{y_2} \frac{U_0'(s)}{U_0^2(s)} ds && \sim \begin{cases} const. & y_2 \to \infty \\ y_2^{-2} & y_2 \to 0 \end{cases} \\
c_{27} &= \int_0^{y_2} \frac{1}{U_0^2(s)} ds && \sim \begin{cases} y_2 & y_2 \to \infty \\ y_2^{-3} & y_2 \to 0. \end{cases}
\end{aligned}
\tag{2.117}
$$

For the matching procedure to work one has to split the individual singularities in the integrals in (2.117) to render them bounded, see Appendix A. These singularities must then be matched to the according upper and lower deck expansion terms. We will not go into any further details here, as it will be argued later that the contributions from the vertical velocity v_2 do not necessarily play an important role with respect to our main purpose of regularization.

Next we write the individual solutions of the Laplace and Poisson problems, Equations (2.101) through (2.104), of the upper deck pressure terms in their Green's function formulae and consider them for $y_1 \to 0$, i.e.

$$
\begin{aligned}
p_{13}(x,0,z,t) &= \frac{U_{00}}{2\pi} \int_{\mathbb{R}^2} \frac{1}{|(x,z)-(\xi_1,\xi_2)|} \partial_{\xi_1} v_{15}(\xi_1,0,\xi_2,t)\, d\xi_1 d\xi_2 \\
p_{14}(x,0,z,t) &= \int_{\mathbb{R}^2} G[-U_{01}\xi_1\ \partial_{\xi_1} v_{12} + U_{01} v_{12} - U_{00}\partial_{\xi_1} v_{16}](\xi_1,0,\xi_2,t)\, d\xi_1 d\xi_2 - \\
&\quad - \int_{\mathbb{R}\times\mathbb{R}^+\times\mathbb{R}} G[\partial_{\xi_3} v_{12} - \partial_{\xi_1} u_{12}] d\xi_1 d\xi_3 d\xi_2 \\
p_{15}(x,0,z,t) &= \frac{1}{2\pi} \int_{\mathbb{R}^2} \frac{1}{|(x,z)-(\xi_1,\xi_2)|} [\partial_t v_{12} + U_{00}\partial_{\xi_1} v_{17}](\xi_1,0,\xi_2,t)\, d\xi_1 d\xi_2 \\
p_{16}(x,0,z,t) &= \frac{U_{00}}{2\pi} \int_{\mathbb{R}^2} \frac{1}{|(x,z)-(\xi_1,\xi_2)|} \partial_{\xi_1} v_{18}(\xi_1,0,\xi_2,t)\, d\xi_1 d\xi_2,
\end{aligned}
\tag{2.118}
$$

where G denotes Green's function, as mentioned earlier, and p_{14} is given modulo some positive constants.

To apply the matching rule between the upper and main deck we observe that the main deck velocity terms (2.116) either evaluate to a function independent of y_2 or grow linearly as $y_2 \to \infty$. According to the scaling of y_1 and y_2 given in (2.6) the growth in the main deck has to match with terms in the upper deck (6 orders below). For example at order ϵ^{10} in the

main deck we have

$$v_{21} \sim -U_{00}\frac{\partial_x A}{p_{00}} - U_{00}c_{21} \quad \text{as} \quad y_2 \to \infty,$$

where the first term is independent of y_2 and hence matches to v_{12} and the second term, which grows with y_2, cf. Appendix A, matches to $U_{01}y_1$ (at order ϵ^4), which decays linearly as $y_1 \to 0$. Therefore this matching procedure is complete. In consequence the remaining vertical velocities in the limit $y_2 \to \infty$ shall be given as

$$\begin{aligned}
v_{24} &\sim -U_{00}\partial_x B_4(x,z,t)\\
v_{25} &\sim -U_{00}\Big(\partial_x B_5(x,z,t) + c^*_{25} - \frac{p_{00}^2 x}{U_{00}^4}y_2\Big)\\
v_{26} &\sim -U_{00}\Big(\partial_x B_6(x,z,t) + \frac{\partial_t A}{p_{00}}c^*_{26}\Big)\\
v_{27} &\sim -U_{00}\Big(\partial_x B_7(x,z,t) - \big(\partial_x p_i + \int_{-\infty}^{x}\partial_z^2 p_i ds\big)(c^*_{27} + y_2/U_{00}^2)\Big),
\end{aligned}$$

with the modified integral terms c^*_{25}, c^*_{26} and c^*_{27} as derived in Appendix A and where the growth for $y_2 \to \infty$ matches to the according terms in the upper deck. Furthermore, we can now match the derivatives with respect to x of the velocity terms in order to obtain a full description of the pressure terms in (2.118). Therefore we have

$$\begin{aligned}
v_{24} \;\text{ match to }\; v_{15} \quad &\Rightarrow \quad \partial_x v_{15}(x,0,z,t) = -U_{00}\partial_x^2 B_4(x,z,t)\\
v_{25} \;\text{ match to }\; v_{16} \quad &\Rightarrow \quad \partial_x v_{16}(x,0,z,t) = -U_{00}\Big(\partial_x^2 B_5(x,z,t) + \partial_x c^*_{25}\Big)\\
v_{26} \;\text{ match to }\; v_{17} \quad &\Rightarrow \quad \partial_x v_{17}(x,0,z,t) = -U_{00}\Big(\partial_x^2 B_6(x,z,t) + \frac{\partial^2_{xt} A}{p_{00}}c^*_{26}\Big)\\
v_{27} \;\text{ match to }\; v_{18} \quad &\Rightarrow \quad \partial_x v_{18}(x,0,z,t) = -U_{00}\Big(\partial_x B_7(x,z,t) - \big(\partial_x^2 p_i + \partial_z^2 p_i\big)c^*_{27}\Big),
\end{aligned}$$

yielding

$$\begin{aligned}
p_{13}(x,0,z,t) &= -\frac{U_{00}^2}{2\pi}\int_{\mathbb{R}^2}\frac{1}{|(x,z)-(\xi_1,\xi_2)|}\partial_\xi^2 B_4\, d\xi_1 d\xi_2\\
p_{14}(x,0,z,t) &= \int_{\mathbb{R}^2} G\big[-U_{01}\xi_1\ \partial_{\xi_1}v_{12} + U_{01}v_{12} + U_{00}^2\big(\partial_x^2 B_5 + \partial_x c^*_{25}\big)\big](\xi_1,0,\xi_2,t)\, d\xi_1 d\xi_2 -\\
&\quad - \int_{\mathbb{R}\times\mathbb{R}^+\times\mathbb{R}} G[\partial_{\xi_3}v_{12} - \partial_{\xi_1}u_{12}]d\xi_1 d\xi_3 d\xi_2\\
p_{15}(x,0,z,t) &= \frac{1}{2\pi}\int_{\mathbb{R}^2}\frac{1}{|(x,z)-(\xi_1,\xi_2)|}\big[-\frac{U_{00}}{p_{00}}\partial^2_{xt}A - U_{00}^2\big(\partial_x^2 B_6 + \frac{\partial^2_{xt}A}{p_{00}}c^*_{26}\big)\big]\, d\xi_1 d\xi_2
\end{aligned}$$

$$p_{16}(x,0,z,t) = -\frac{U_{00}^2}{2\pi}\int\limits_{\mathbb{R}^2} \frac{1}{|(x,z)-(\xi_1,\xi_2)|}\left[\partial_x^2 B_7 - (\partial_x^2 p_i + \partial_z^2 p_i)c_{27}^*\right] d\xi_1 d\xi_2.$$

Regarding the pressure p_2 in (2.115) we have no formulae for the terms p_{23}, p_{24} and p_{25} but know that they do not vary with y_2. The linear growth of p_{26} (at order ϵ^{16}), cf. (2.112), matches to the linear decay of p_{12} (at order ϵ^{10}), stemming from the exponential, see (2.105). Therefore the matching rule yields

$$p_{26} \quad \text{match to} \quad p_{16} \quad \Rightarrow \quad p_{16}(x,0,z,t) = p_{26}(x,0,z,t) - \frac{c_p}{p_{00}}\partial_x^2 A.$$

Eventually, we want the higher order pressure terms to relate to the original interaction pressure p_i in the form of representing an expansion of p_i. Hence, we shall evaluate these pressure terms in the lower deck, where p_i was defined as p_{32}, cf. Section 2.1, Equation (2.28). Consequently, with p_3 being independent of y_3 and p_2 having only p_{26} varying with y_2, we can define a new interaction pressure p_i^* at $y_2 = 0$ as

$$p_i^*(x,z,t) = p_i(x,z,t) + \epsilon^3 p_{13}(x,0,z,t) + \epsilon^4 p_{14}(x,0,z,t) + \epsilon^5 p_{15}(x,0,z,t) + \epsilon^6 p_{26}(x,0,z,t).$$

Remark 2.58. One can now argue, as all the new terms are asymptotically small against the original p_i, they shall not appear in the fundamental equation for A. Nevertheless, as the length scale of a disturbance or as parts of the solution (in terms of its Fourier decomposition) gets smaller, spatial derivatives of the unknown A do not remain (asymptotically) small and hence are not negligible. Furthermore, all terms in p_i^*, not containing A, can be viewed, with respect to the fundamental problem (2.28), as asymptotically small inhomogeneities or contributions, neither stabilizing or destabilizing the solution.

Thus, recalling the main goal of this section, i.e. finding regularizing operators in higher order expansions, we will modify p_i^* for the sake of numerical computations and comparison with the regularization derived in the previous section.

Obviously, all B_i can be collected to form some inhomogeneity and are hence left out, e.g. p_{13} on the whole. In virtue of the original p_i, where $\mathcal{R}^1\partial_x^2 A$ represents a differential operator acting on A, we will compare all other appearing types of derivatives of A to this term, to decide whether we have to take them into account. Considering (2.98) it can be easily seen that p_{14} will only comprise of the same order of differentiation of A as p_i (at the maximum). It additionally contains nonlinearities such as $A\partial_x A$, which also do not contribute to any regularizing effects. Hence, p_{14} can be canceled from p_i^*. The pressure term p_{15} contains the time derivative of A and shall thus be taken into account, as this could change the structure of the Cauchy problem per se. The most important term here, as we have demonstrated in the heuristic motivation, is p_{26} at $y_2 = 0$, as this term accounts for the streamline curvature.

Hence, we eventually obtain the new interaction pressure to be

$$\begin{aligned}
p_i^*(x,z,t) &= p_i(x,z,t) + \epsilon^5 \frac{1}{2\pi} \int\limits_{\mathbb{R}^2} \frac{1}{|(x,z)-(\xi_1,\xi_2)|} \big[-\frac{U_{00}}{p_{00}} \partial_{xt}^2 A - U_{00}^2 \frac{\partial_{xt}^2 A}{p_{00}} c_{26}^* \big] \, d\xi_1 d\xi_2 + \\
&+ \epsilon^6 (-\frac{U_{00}^2}{2\pi} \int\limits_{\mathbb{R}^2} \frac{1}{|(x,z)-(\xi_1,\xi_2)|} \big[-(\partial_x^2 p_i + \partial_z^2 p_i) c_{27}^* \big] \, d\xi_1 d\xi_2 + \frac{c_p}{p_{00}} \partial_x^2 A) = \\
&= p_i(x,z,t) - \epsilon^5 c_1 \int\limits_{\mathbb{R}^2} \frac{1}{|(x,z)-(\xi_1,\xi_2)|} \partial_{xt}^2 A \, d\xi_1 d\xi_2 + \\
&+ \epsilon^6 c_2 \int\limits_{\mathbb{R}^2} \frac{1}{|(x,z)-(\xi_1,\xi_2)|} (\partial_x^2 p_i + \partial_z^2 p_i) \, d\xi_1 d\xi_2 + \epsilon^6 c_3 \partial_x^2 A,
\end{aligned}$$

where c_1, c_2, c_3 are positive constants. This formula for the interaction pressure is quite involved and becomes even more so when included in the fundamental problem for A. Therefore we will further analyze the contributions of the new terms in virtue of the dispersion relation, which we dealt with in the previous section.

First, consider the potential integral over the derivatives of p_i. When substituting (2.29) for p_i and using the operator notation for the potential integral we obtain

$$\int\limits_{\mathbb{R}^2} \frac{1}{|(x,z)-(\xi_1,\xi_2)|} (\partial_x^2 p_i + \partial_z^2 p_i) \, d\xi_1 d\xi_2 = -\mathcal{R}^1\big(\mathcal{R}^1([\partial_x^2 + \partial_z^2]\partial_x^2 A)\big),$$

although in this version not much insight into the meaning of this term has been gained. But by using the operator symbols and their characteristics, as derived in the previous section, see e.g. (2.69) and (2.70), one can easily see

$$-\mathcal{F}\Big(\mathcal{R}^1\big(\mathcal{R}^1([\partial_x^2 + \partial_z^2]\partial_x^2 A)\big)\Big) = -(k^2+l^2)^{-1}((ik)^2 + (il^2))(ik)^2 \mathcal{F}A = -k^2 \mathcal{F}A,$$

which is exactly the same Fourier symbol as for $\partial_x^2 A$ in p_i^*. From a practical numerical computation viewpoint, the potential integral is either not sufficiently resolved or yields too big matrices when satisfying a certain accuracy requirement. Hence, by the equality to the classical second derivative, with respect to the regularization, of course, we can safely claim the information coming from the double application of the potential integral is already contained in the derivative term.

Such an argumentation cannot be done in this straight forward manner for the mixed (time and spatial) derivative and we will thus carry it along when substituting p_i^* for p_{32} in (2.28), where consequently two new terms appear on the right hand side, reading (modulo some constants)

$$\int\limits_{-\infty}^{x} \frac{1}{(x-s)^{1/2}} \int\limits_{\mathbb{R}^2} \frac{\partial_{\xi_1}^2 + \partial_{\xi_2}^2}{|(s,z)-(\xi_1,\xi_2)|} \, \partial_t A \, d\xi_1 d\xi_2 \, ds - \int\limits_{-\infty}^{x} \frac{1}{(x-\xi)^{1/2}} (\partial_\xi^3 A + \partial_\xi \partial_z^2 A) d\xi,$$

where the Fourier symbol notation yields

$$-(ik)^{-1/2}(k^2+l^2)^{1/2}\mathcal{F}\partial_t A - (ik)^{-1/2}((ik)^3-(ik)l^2)\mathcal{F}A.$$

Note that we did not say anything about the existence of the Fourier transform of A and one might therefore replace A with the ansatz $e^{\omega t}e^{i(kx+lz)}$, made for the dispersion relation in the previous section. Linearizing then the fundamental problem around some constant steady state (as mentioned in Remark 2.37) and only taking into account derivative terms yields

$$\begin{aligned}\left[\mathcal{J}_{-\infty}^{3/4} - \mathcal{J}_{-\infty}^{1/2}\mathcal{R}^1(\partial_x^2+\partial_z^2)\right]\partial_t e^{\omega t}e^{i(kx+lz)} = \\ = \mathcal{J}_{-\infty}^{1/2}\mathcal{R}^1(\partial_x^3+\partial_x\partial_z^2)\,e^{\omega t}e^{i(kx+lz)} - \mathcal{J}_{-\infty}^{1/2}(\partial_x^3+\partial_x\partial_z^2)e^{\omega t}e^{i(kx+lz)},\end{aligned}$$

such that the simplified dispersion relation is given as

$$\begin{aligned}\omega &= [(ik)^{-3/4}+(ik)^{-1/2}(k^2+l^2)^{1/2}]^{-1}\big((ik)^{1/2}[-(k^2+l^2)^{1/2}+(k^2+l^2)]\big) = \\ &= [1+(ik)^{1/4}(k^2+l^2)^{1/2}]^{-1}\big((ik)^{5/4}[-(k^2+l^2)^{1/2}+(k^2+l^2)]\big).\end{aligned} \tag{2.119}$$

Comparing this to the original right hand side symbols (2.78) one can see the influence of the mixed derivative (i.e. the inverse term in square brackets) and the second derivative, where for the latter we have the physical connection to the streamline curvature. With respect to a possible regularization, it can be easily seen, on the one hand, that the contribution from the mixed derivative in its Fourier symbol representation remains bounded and even tends to zero as $|(k,l)|\to\infty$. On the other hand, the purely spatial derivatives are obviously of higher order (in some sense) and thus, in virtue of adding a regularizing operator, we only consider the streamline curvature term.

This finally results in the interaction pressure to be

$$p_i^*(x,z,t) = -\frac{1}{2\pi}\int\limits_{\mathbb{R}^2}\frac{1}{|(x,z)-(\xi_1,\xi_2)|}\partial_{\xi_1}^2 A\,d\xi_1 d\xi_2 + \alpha\,\partial_x^2 A(x,z,t),$$

with $\alpha=\epsilon^6 c_p>0$ (cf. (2.112)), where we also have applied the affine transform mentioned in Section 2.1, and in the additional term on the right hand side in (2.28) of the form

$$-\alpha\int\limits_{-\infty}^{x}(x-\xi)^{-1/2}(\partial_\xi^3 A+\partial_\xi\partial_z^2 A)d\xi = -\alpha\mathcal{J}_{-\infty}^{1/2}(\partial_x^3 A+\partial_x\partial_z^2 A)(x,z,t). \tag{2.120}$$

Reconsidering the dispersion relation (2.119) without the mixed derivative term yields

$$\omega = e^{i\,\mathrm{sgn}(k)5\pi/8}(-|k|^{5/4}(k^2+l^2)^{1/2}+|k|^{5/4}(k^2+l^2)),$$

with the real part of the right hand side being (in polar coordinates (r, ϕ))

$$\underbrace{\cos(5\pi/8)}_{<0}(-|k|^{5/4}(k^2+l^2)^{1/2}+|k|^{5/4}(k^2+l^2)) = |\cos(\phi)|^{5/4}(r^{9/4}-r^{13/4}),$$

where it is now easy to see that the right hand side is bounded from above for all r and ϕ. Additionally, this shows that the regularizing derivatives proposed in (2.85) already represent a good guess and numerical solutions gained with this term are almost indistinguishable from solutions including (2.120). In virtue of such an observation, due to the rather involved numerical treatment of (2.120), this term shall be seen as a physical justification of including higher derivatives to regularize the Cauchy problem (2.47).

Remark 2.59. The planar equivalent of the regularization term (2.120) can also be found performing the same steps as above, which results in the same term as if taking A independent of z in (2.120).

For comparison reasons we computed a solution of the planar problem (2.49) including (2.120) up to time $t = 1$ (with the usual initial conditions of $\underline{a} \equiv 0$ and excluding any perturbation). Table 14 shows the leading coefficients for different N in comparison to the coefficients obtained from using the direct implicit time integration. One can clearly see the need of overproportionally high polynomial degrees when using (2.120) to regularize the time evolution. This, of course, also depends strongly on the parameter α.

a_i	direct	$N = 100$	$N = 300$
a_0	0.00578	−0.03601	0.00623
a_1	−0.31590	−0.32431	−0.31609
a_2	−0.16896	−0.18280	−0.17048
a_5	0.11633	0.13254	0.11647
a_7	0.09045	0.09506	0.09072

Table 14: Comparison of the leading coefficients for the solution at $t = 1$ between the direct implicit method (cf. Table 13 (left)) and the regularization with curvature term at $\alpha = 1/100$, $\Delta t = 1/100$ and various N.

The influence of the value of the regularization parameter α can be further seen from the steady state solutions including (2.120), see Figure 19. This again shows a typical characteristic of regularizing operators, namely to find the balance between altering the problem (and its solutions) and a good working regularization, when choosing the value of the regularization parameter.

2.3.3 Self-Similar Finite Time Blow-up

In the following the Cauchy problems and the presented solutions are always to be understood in the regularized sense established in the previous sections. Also, numerical results are gained with highest resolution necessary to depict the sought characteristics, using the filter

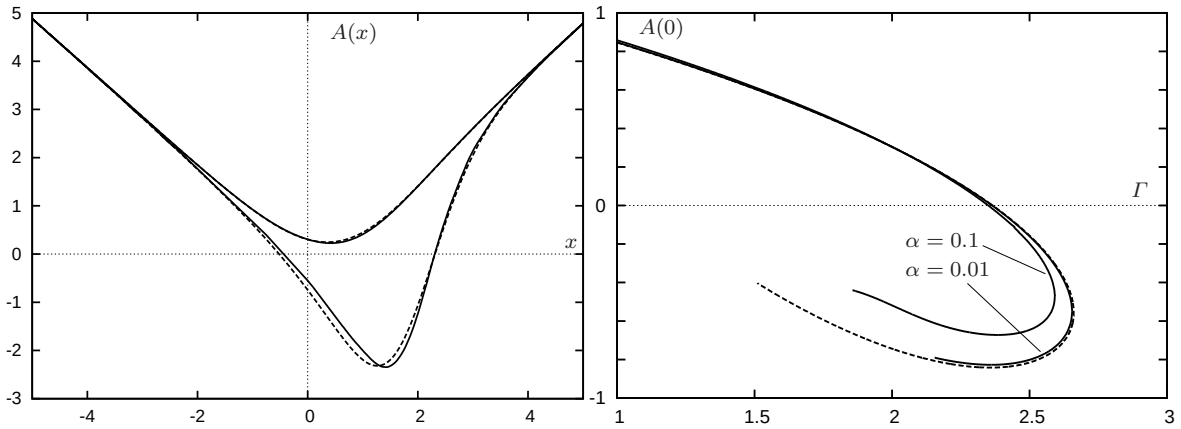

Figure 19: Steady states including curvature term. Left: at $\Gamma = 2$ for $\alpha = 0$ (dashed) and $\alpha = 0.1$ (solid). Right: change in fundamental curve, $\alpha = 0$ (dashed).

(i.e. direct implicit) method. Hence, we will refrain from studying convergence properties and accuracy as well as regularization issues, for which we refer to the previous sections.

The whole present section deals with one particular set-up, where for the sake of more physical meaning we will not be as general as in Section 2.3.1. Consider the stationary two-dimensional upper branch solution at a certain Γ, cf. Figure 3(a) as an initial condition. By introducing a three-dimensional blowing slot, i.e. the forcing function given in (2.93), the flow, so far independent of z, will experience disturbances in z direction and is thus governed by the Cauchy problem (2.47).

As mentioned in Section 2.3.1, in the paragraph following (2.93), the intensity of the perturbation, i.e. (2.94), has to exceed a certain threshold, to push the solution outside the basin of attraction of the steady state. In such situations the occurrence of finite time singularities are well established for the planar problem (2.49) (see Smith (1982) and for more details Scheichl et al. (2008)), whereas for (2.47), to the authors knowledge, only the study in Duck (1990) (for a globally three-dimensional setting) provides a first glance at the finite time blow-up. Supporting its existence in principle can be done by viewing the Cauchy problems as

$$\partial_t A(t) = \mathcal{A}^* A(t) + F(A), \quad F(A) = \left[\mathcal{J}_{-\infty}^{3/4}\right]^{-1}(-A^2 + x^2 - \Gamma + g)$$

with $\mathcal{A}^*$ as defined in (2.85) (maybe also for $\alpha = 0$) and $F(A)$ as a nonlocal nonlinearity.

Remark 2.60. Ball (1977) provides an existence and uniqueness theorem for (maximally defined mild) solutions on $[0, T)$ of such problems, where $\mathcal{A}^*$ has to be the generator of a strongly continuous semigroup and F has to be locally Lipschitz. He shows further, that if $T < \infty$ the solution becomes unbounded in the given norm. A textbook example for this would be the heat equation with quadratic nonlinearity or more general *reaction-diffusion*

equations as mentioned in, e.g. Galaktionov & Vázquez (2002). For an operator theoretic approach we refer to Payne (1975) and references therein.

One common result of these considerations is that blow-up has to be always connected to some norm, i.e. while a classical solution (considered in the L^∞ norm) can experience a singularity, this might not be seen in a weaker norm (e.g. weak or mild solutions). Another common aspect is the space of allowable initial conditions to contain a subset, on which the evolution blows up in finite time.

In the problem here, parts of a subset of initial conditions leading to the singularity can be found empirically by varying the integral criterion (2.94) for a given perturbation and thus single out solutions at time T^*, which when used as an initial condition admit the finite time blow-up. For the case at hand, the blowing slot centered at $x_c = -2$ with length $l = 1$ (cf. Scheichl et al. (2008)), $p(z) = \frac{1}{1+z^2}$ and the choice of parameters $a = 3$ and $T^* = 2$ proved to be sufficient.

Apart from the principle existence of the blow-up scenario, Galaktionov & Vázquez (2002) studied various further aspects of occurring singularities by raising (and answering) some basic questions concerning *when, where and how does the singularity occur* and can one compute it approximately? We will tackle the latter of them in more detail and consequently answer the others.

N.b.: Since this treatise deals with boundary layers (although in a special way) it is worth mentioning that even the according Prandtl equations blow up under certain conditions, see E & Engquist (1997).

Given the perturbation g as defined above and set $A_0(x, z) = A_{st}(x)$, Figure 20 shows $A(x, z, t)$ for $0 \leq t \leq T^* = 2$, i.e. the evolution with the perturbation present. The resulting

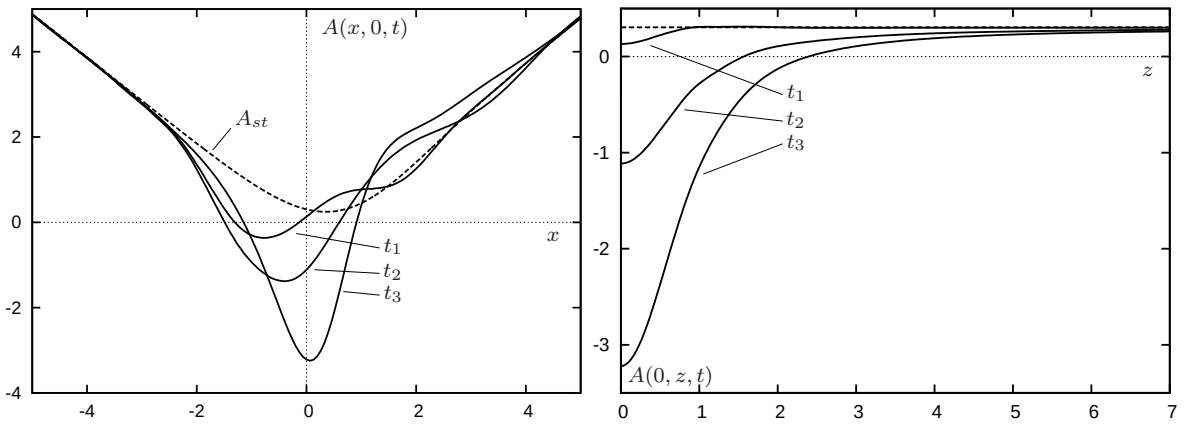

Figure 20: The solution $A(x, z, t)$ at $(t_1, t_2, t_3) = (1, 1.5, 2)$, left: at $z = 0$, right: at $x = 0$.

solution $A(x, z, T^*)$, now taken as an initial condition for the unperturbed Cauchy problem, has deviated sufficiently from the steady state as to not reapproach it for $t > T^*$. Instead the minimum formed in these first time steps will become more and more pronounced with larger

negative values, where eventually the blow-up occurs in a single point (x_s, z_s) and from the z-symmetry of the solution obviously $z_s = 0$, see Figure 21.

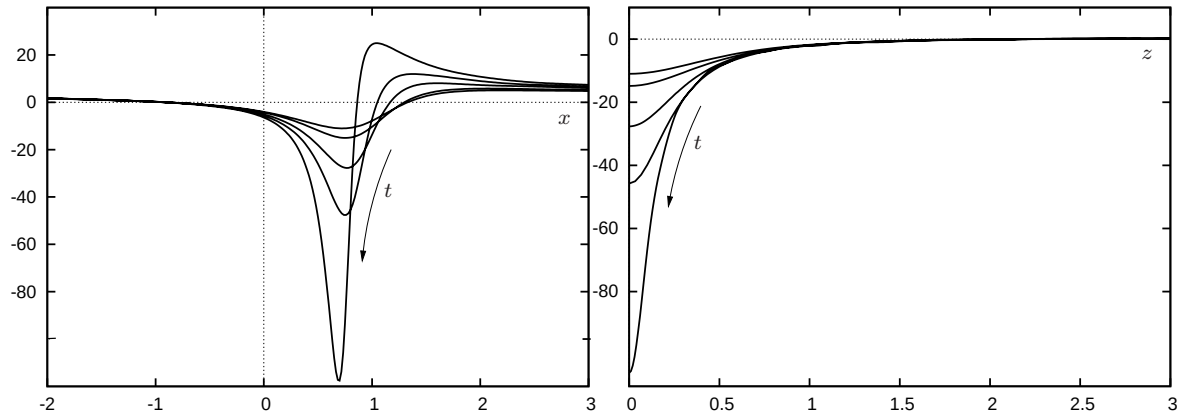

Figure 21: The solution $A(x, z, t)$ of the unperturbed problem continuing Figure 20 at $t \in \{2.61, 2.67, 2.73, 2.76, 2.77\}$, left: at $z = 0$, right: at $x = 0$.

Remark 2.61. In approaching the singularity one cannot use a fixed time step in the numerical computations (if it was not already chosen unreasonably small), since the blow-up time t_s is not known a priori. Scheichl et al. (2008) demonstrated that adapting Δt with respect to the relative change in the minimum of the solution is most appropriate to be able to obtain $A(\cdot, t)$ as close as possible to t_s. For the computation presented in Figure 21 we used

$$m_i = \big|\min_{x\in\mathbb{R}}\big(A(x, 0, t_i)\big)\big|, \quad \text{such that} \quad \Delta t_i = \Delta t_{i-1}\frac{m_{i-1}}{m_i} \quad \text{if} \quad \frac{m_i}{m_{i-1}} > 1.05.$$

Plotting the minimum of A versus the time, the existence of the finite time singularity becomes even more apparent, see Figure 22.

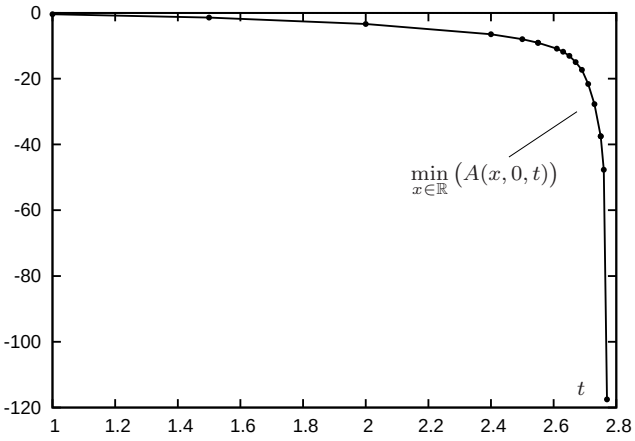

Figure 22: The evolution of the minimum of $A(x, z, t)$, with the results taken from Figures 20 and 21, revealing a blow-up time of $t_s \approx 2.8$.

Also, what is often recognized in regularized problems, near the (stable) steady state the dynamics are comparably slow, i.e. with a moderate amplitude of the perturbation the activation time T^* is large in comparison to the difference to the blow-up time. This is in complete accordance to what has been shown in Section 2.3.1 regarding the slow convergence toward the equilibrium, cf. Figure 12.

With resolving the near-singular behavior or shapes of functions, meaning locally steep gradients and strong curvatures, a disadvantage of the polynomial expansion in contrast to finite differences can be observed. Functions (locally) violating a certain regularity regarding their derivatives need *overproportionally many polynomials* in according orthogonal global approximations to reach a certain accuracy. We shall illustrate this by a simple

Example 2.1. Consider the Chebyshev expansion of (see also Example 3.1 in Section 3.1)

$$f(x) = \frac{1}{1+x^2} = \frac{1}{2}R_0(x) - \frac{1}{2}R_2(x).$$

Taking $f^2 = 1/(1+x^2)^2$, which has a more pronounced maximum at $x = 0$, the according exact expansion reads

$$f^2(x) = \left(\frac{1}{2}R_0(x) - \frac{1}{2}R_2(x)\right)^2 = \frac{3}{8}R_0(x) - \frac{1}{2}R_2(x) + \frac{1}{8}R_4(x),$$

where we used Lemma 3.2(v) for the R_2^2-term. Therefore, if one would have used $N = 2$ to expand f and f^2 the latter would be insufficiently "approximated".

Such facts have to be considered when finding the lowest number of polynomials necessary to compute the time evolution near the blow-up. One has to take into account the ill-posedness in form of fast growing oscillations when the time step is too small, as explained in the paragraph following Figure 15 in Section 2.3.1. This then determines the number of polynomials N_x due to the corresponding symbol (2.79). Thus, oscillations confined to the z coordinate might not necessarily stem from growing instabilities, but are due to the lack of local approximation qualities, i.e. larger pointwise errors. On the left of Figure 23 we compared Chebyshev expansions using $N = 20$, to three different functions (dashed lines) mimicking the time evolution with respect to z at $x = 0$ of the solution A as depicted in Figure 21. One can readily observe that the approximation coincides virtually exactly for functions with the minima at -20 and -40 (and are hence left out in the figure), whereas in the third situation oscillations occur in the approximation, especially in the vicinity where the curvature of the original function changes its sign. By using $N = 40$ polynomials these imperfections vanish, with the expansion being indistinguishable from the given function. On the right of Figure 23, to show the similarity, we plotted the solution gained from the time evolution at $t \approx 2.77$ using $N_z = 20$ (solid) and $N_z = 40$ (dashed) polynomials.

Eventually we claim that polynomial expansions lose their advantage against finite differences when trying to resolve (finite time) singularities, independently of the quality of the methods leading to the equation systems in spatial coordinates and the schemes for the

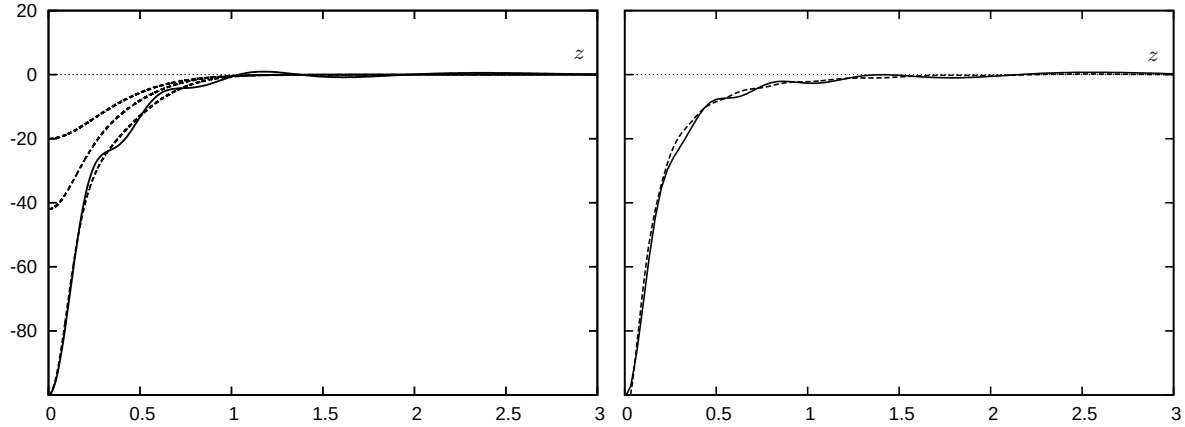

Figure 23: Left: Approximation using $N = 20$ Chebyshev polynomials (solid) of the dashed graphs (the first two are indistinguishable from their approximation). Right: Solution $A(0, z, t)$ at $t \approx 2.77$ with $N_z = 20$ (solid) and $N_z = 40$ (dashed, cf. Figure 21).

discrete time integration. This also holds for finding discontinuities and shocks in time evolution problems, cf. the Gibbs phenomenon described in Section 3.2.2, Example 3.7 and what follows.

Remark 2.62. It is worth noting the singularity to occur only in one singular point x_s in all two-dimensional cases studied. Such an observation can not only be found in the case of unforced problems, where Γ was chosen to be at or above its critical value Γ_c, cf. Figure 4, Smith (1982) and Scheichl et al. (2008), but also for various different perturbations applied to the problems when starting at an upper branch stationary solution. For example, when forcing the solution to form two local minima in x in the first time steps, eventually one of them will dominate the behavior and result in the blow-up. We claim this to be due to the non-symmetric properties of the Abel operators involved.

In the three-dimensional set-up, solutions of the unperturbed problem are always symmetric with respect to z and thus multiple, simultaneously occurring singularities in z (at one common x_s) are possible. We show this by using the previously applied blowing device (2.93), where the function $p = p(z)$ is now assumed to have two (symmetric) maxima, e.g.

$$p(z) = \frac{1}{1 + (z-1)^2} + \frac{1}{1 + (z+1)^2}.$$

As done for the perturbation with one maximum, Figure 24 now shows the solution for the first time steps, when the new forcing is activated, i.e. $t \leq T^*$, using the same numerical parameters as before.

As in the previous computation, a minimum appears in x at $z = 0$, but also with respect to z at $x = 0$, where the latter is smaller, rendering the minimum at $z = 0$ to be only local. Again, for the results at $t > T^*$ the forcing is switched off and the unperturbed problem ends with the finite time singularity, where $z_s \neq 0$ and $t_s \approx 2.45$, see Figure 25.

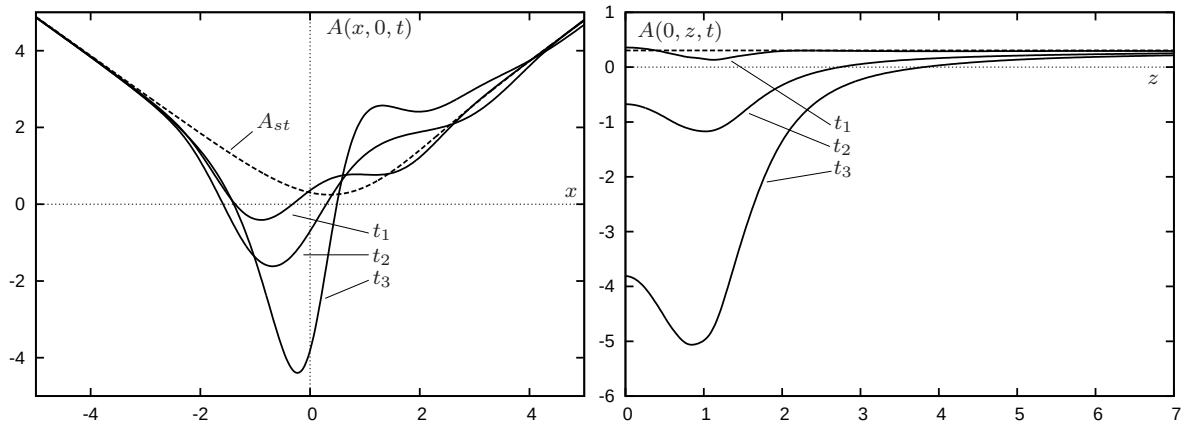

Figure 24: The solution $A(x,z,t)$ at $(t_1,t_2,t_3) = (1,\ 1.5,\ 2)$, left: at $z=0$, right: at $x=0$.

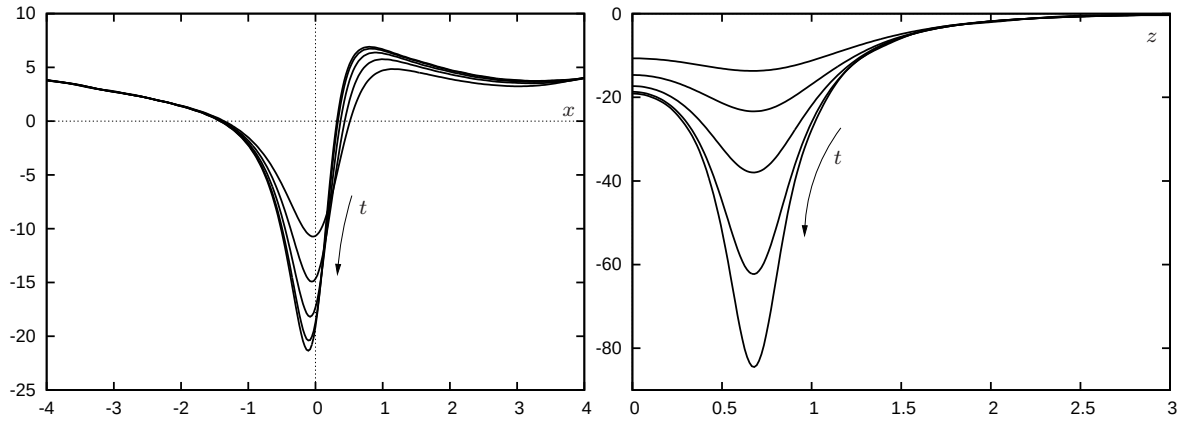

Figure 25: The solution $A(x,z,t)$ of the unperturbed problem continuing Figure 24 at $t \in \{2.306,\ 2.363,\ 2.386,\ 2.4,\ 2.405\}$, left: at $z=0$, right: at $x=0$.

Remark 2.63. Forcing functions with multiple symmetric maxima do not necessarily yield the same number of (symmetrically arranged) singularities in the solution. In the above, we placed the maxima in p at $z_m = \pm 1$, which was approximately the necessary distance to obtain two blow-up points in z. Thus we state that maxima in the perturbation function being too close yield only one common singularity. Without any forcing applied, but by varying Γ, such that it exceeds its critical value Γ_c at some time step, finite time singularities do occur as well. Here, one has to distinguish between Γ depending on z and t (as done in Duck (1990)) and depending only on t, for $\Gamma = \Gamma(z,t)$ can be seen as a type of forcing, yielding similar results as presented in the above, and $\Gamma = \Gamma(t)$ would just mean to start at a steady state, but the significant time evolution is taken for a situation where no steady state exists. Here, without having it studied in all details, we assert that $z_s \neq 0$ appears naturally.

Having demonstrated that blow-up occurs at different points and different times, the usual question is whether there is some generality behind the development of the singularity, i.e. a structure, independent of (x_s, z_s, t_s). As mentioned in the monograph by Barenblatt (1979),

one shall investigate the problem and its solutions regarding the so-called *self-similarity*. This became a prominent technique to gain further insight into blow-up phenomena of time dependent problems, cf. the survey by Eggers & Fontelos (2009).

Following the definition in Barenblatt (1979), let $u = u(x,t)$ be a solution to a given problem. If there exist time dependent scales $\tilde{x}(t)$ and $\tilde{u}(t)$, and an unknown function Φ, such that

$$u(x,t) = \tilde{u}(t)\,\Phi(x/\tilde{x}(t)),$$

then Φ is a *similarity solution* in the *similarity variable* $x/\tilde{x}$. Note that these similarity scaling renders Φ independent of t. Furthermore, if $\tilde{x}$ and $\tilde{u}$ are unique in the sense that the relation above is the only possible way to define Φ, such that u is a solution of the given problem, then we call Φ and the scalings *self-similar*.

For the planar Cauchy problem (2.49), Smith (1982) argued that if t approaches t_s then the left hand side of (2.49) is dominated by A^2, as long as $|x - x_s|$ is small, whereas the first term on the right hand side is of order $A\,|x-x_s|^{-3/2}$ and the second term of $A\,|x-x_s|^{3/4}(t_s-t)^{-1}$. To obtain the similarity variable $x/\tilde{x}(t)$ as in the definition above, both terms on the right hand side have to be of the same order. By then equating their (common) order with the order of A^2 yields the desired function $\tilde{u}(t)$, rescaling the similarity solution. Transferring this idea to the three-dimensional problem (2.47) gives the similarity coordinates and the similarity solution as (cf. Duck (1990))

$$\left.\begin{aligned} &\tau = t_s - t, \quad x - x_s = \tau^{4/9}\hat{x}, \quad z - z_s = \tau^{4/9}\hat{z} \\ &A(x,z,t) = \tau^{-2/3}\hat{A}(\hat{x},\hat{z}) + \mathrm{o}(\tau^{-2/3}) \end{aligned}\right\} \quad \text{as } \tau \to 0. \tag{2.121}$$

One can now utilize these local coordinates further to obtain better estimates for the time t_s of the blow-up and its spatial location.

Define

$$\min_{(x,z)\in\mathbb{R}^2} \big(A(x,z,t_i)\big) =: m_i,$$

it then follows from the second line in (2.121) that for any t_1, t_2 close to t_s

$$\frac{m_1}{m_2} = \Bigg(\underbrace{\frac{t_s - t_1}{t_s - t_2}}_{=:c}\Bigg)^{-2/3} \quad \Rightarrow \quad c = \left(\frac{m_1}{m_2}\right)^{-3/2} \quad \Rightarrow \quad t_s = \frac{t_1 - c\,t_2}{1-c}$$

and also, for (x_i, z_i) being the point where $A(\cdot, t_i)$ attains its minimum m_i, one has

$$\frac{x_1 - x_s}{x_2 - x_s} = c^{4/9} \quad \Rightarrow \quad x_s = \frac{c^{4/9}x_2 - x_1}{c^{4/9} - 1}$$

and analogously for z_s. Applying the above formulae to the results shown in Figure 21 yields

$$x_s \approx 0.66, \quad t_s \approx 2.775,$$

and for the solutions in Figure 25, one obtains

$$x_s \approx -0.1, \quad z_s \approx \pm 0.66, \quad t_s \approx 2.418. \tag{2.122}$$

With knowing the point (x_s, z_s, t_s) one can also compute an estimate for the profile $\hat{A}$ in (2.121) via

$$\hat{A}(\hat{x}, \hat{z}) \approx \frac{A(x, z, t_1) - A(x, z, t_2)}{(t_s - t_1)^{-2/3} - (t_s - t_2)^{-2/3}}, \quad (\hat{x}, \hat{z}) \in \mathbb{R}^2, \tag{2.123}$$

with (x, z) replaced by $(\hat{x}\tau^{4/9} + x_s, \hat{z}\tau^{4/9} + z_s)$. As done in Scheichl et al. (2008) one can use such estimates to graphically show the convergence of the solutions $A = A(\hat{x}, \hat{z}, t)$ (viewed in the similarity variables) to the profile $\hat{A} = \hat{A}(\hat{x}, \hat{z})$ as $t \to t_s$, see Figure 26.

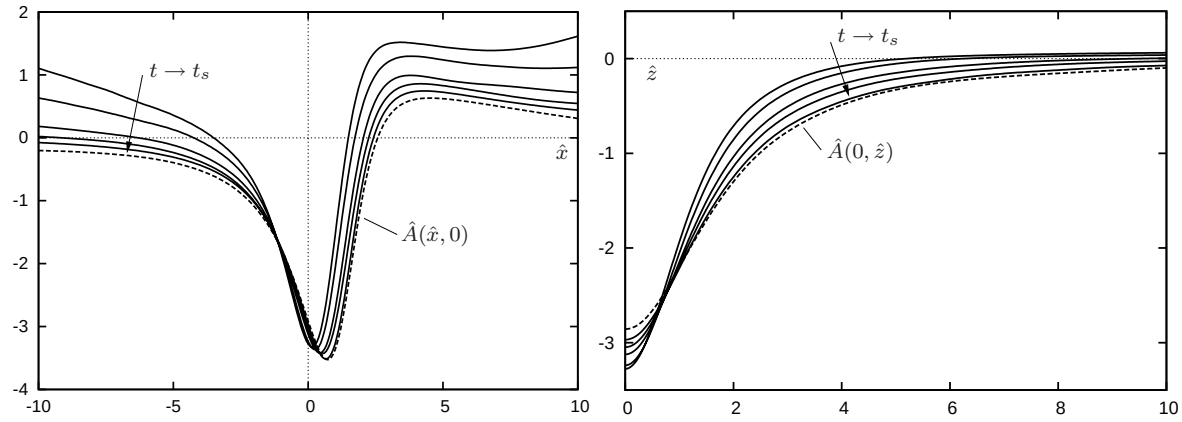

Figure 26: The solution $A = A(x, z, t)$ taken from Figure 21 plotted as $\tau^{2/3} A(x, z, t)$ as a function of $(\hat{x}, \hat{z})$ with (x, z) replaced by $x = \hat{x}\tau^{4/9} + x_s$, $z = \hat{z}\tau^{4/9} + z_s$, with the estimated profile $\hat{A}$ from (2.123) (dashed lines). Left: at $\hat{z} = 0$, right: at $\hat{x} = 0$.

Remark 2.64. A similar convergence behavior as in Figure 26 can be seen when using the solutions in Figure 25 with the according blow-up data (2.122). It is worth noting that the estimated blow-up profiles coincide very well in both cases, suggesting this structure to be an intrinsic property of the singularity in the Cauchy problem, independent of when and where the blow-up occurs. This has been confirmed further by replacing p in the perturbation (2.93) by some exponential function (or similar), always leading to the same limiting structure.

Remark 2.65. The passage of the non-similar behavior to the similarity structure is not only necessary for the similarity transform to be valid, but can also be used in some occasion to determine unknown parameters, as described in Barenblatt (1979). Here we could argue to have used this to also confirm the blow-up data estimates above.

With the convergence of the rescaled solution to the similarity profile and its independence of the initial condition and forcing imposed on the Cauchy problem we can claim, according to the definition above, that the scalings (2.121) are in fact *self-similar*.

The obvious next step, and it is quite usual to do so (cf. Eggers & Fontelos (2009) and Galaktionov (2009)), is to substitute the similarity variables (2.121) into the Cauchy problem to obtain an equation determining the blow-up profile $\hat{A}$. Additionally, information on the far field behavior of $\hat{A}$ can be obtained. Since the blow-up occurs in one singular point only, one can expect the solution to remain bounded away from (x_s, z_s), that is

$$|A(x,z,t_s)| < \infty, \quad \forall (x,z), \text{ such that } \; |x-x_s| = |z-z_s| = O(1), \quad \text{as } \tau \to 0.$$

Furthermore, from the scalings (2.121)

$$|x-x_s| = |z-z_s| = O(1) \quad \Leftrightarrow \quad |(\hat{x},\hat{z})| \to \infty.$$

Introducing the usual polar coordinates as $(x,z) \mapsto (r,\phi)$ (centered at (x_s,z_s)) and $(\hat{x},\hat{z}) \mapsto (\hat{r},\hat{\phi})$ yields

$$\left.\begin{array}{ll} x - x_s = r\,\cos(\phi), & \hat{x} = \hat{r}\,\cos(\hat{\phi}) \\ z - z_s = r\,\sin(\phi), & \hat{z} = \hat{r}\,\sin(\hat{\phi}) \end{array}\right\} \quad \overset{\text{subst. (2.121)}}{\Longrightarrow} \quad \begin{array}{l} r\,\cos(\phi) = \tau^{4/9}\hat{r}\,\cos(\hat{\phi}) \\ r\,\sin(\phi) = \tau^{4/9}\hat{r}\,\sin(\hat{\phi}) \end{array}.$$

Taking the square and adding the equations on the right yields $r^2 = \tau^{8/9}\hat{r}^2$, whereas division gives $\tan(\phi) = \tan(\hat{\phi})$, thus

$$\frac{\hat{r}^2}{r^2} = \tau^{-8/9} \;\Rightarrow\; \left(\frac{\hat{r}^2}{r^2}\right)^{3/4} = \tau^{-2/3} \;\Rightarrow\; A = \left(\frac{\hat{r}^2}{r^2}\right)^{3/4} \hat{A}.$$

We mentioned earlier that for $A = O(1)$ as $r = O(1)$ one has $\hat{r} \to \infty$, such that (in this limit)

$$(\hat{r}^2)^{3/4}\hat{A} \overset{!}{=} O(1) \quad \Rightarrow \quad \hat{A} = O\big((\hat{r}^2)^{-3/4}\big) \;\text{ or }\; \hat{A} \sim c(\hat{\phi})(\hat{r}^2)^{-3/4}, \tag{2.124}$$

which reads in Cartesian coordinates (cf. Duck (1990))

$$\hat{A} \sim c^*(\hat{z}/\hat{x})(\hat{x}^2+\hat{z}^2)^{-3/4} \;\text{ as }\; \hat{x}^2+\hat{z}^2 \to \infty.$$

Remark 2.66. The functions c and c^*, describing the far field decay of $\hat{A}$ in more detail, can be assumed to be bounded and continuous on $[-\pi,\pi]$. From the graphs in Figure 26 (dashed lines) and the far field relation one readily obtains that $\hat{A}$ is not in $L^1(\mathbb{R}^2)$ but in $L^2(\mathbb{R}^2)$ and hence one has to carefully consider the existence of its Fourier transform, analytically (cf. Duck (1990)) as well as in the discrete version, since the decay at infinity might be too weak to yield plausible results.

To finally state the full problem governing the terminal structure we apply a linear coordinate transform, such that $x_s = z_s = t_s = 0$, and substitute the similarity coordinates into (2.47), using

$$\begin{aligned} \partial_t A &= \tfrac{2}{3}(-t)^{-5/3}\big[\hat{A} + \tfrac{2}{3}\big(\hat{x}\,\partial_{\hat{x}} + \hat{z}\,\partial_{\hat{z}}\big)\hat{A}\big] \\ \partial_x^2 A &= (-t)^{-14/9}\partial_{\hat{x}}^2\hat{A}, \quad \partial_z^2 A = (-t)^{-14/9}\partial_{\hat{z}}^2\hat{A} \end{aligned} \tag{2.125}$$

and instead of the Riesz potential $\mathcal{R}^1$, the Riesz transform $\mathcal{R}_1$, i.e. shifting one ∂_x onto the kernel, cf. Remark 3.62, is used. By taking the limit $(-t) \to 0$ this yields

$$\begin{aligned} &\hat{A}^2 = -\tfrac{\lambda}{2\pi}\, \mathcal{J}_{-\infty}^{1/2} \mathcal{R}_1 \Delta \hat{A} - \tfrac{2}{3}\gamma\, \mathcal{J}_{-\infty}^{3/4} \big[\hat{A} + \tfrac{2}{3} \left(\hat{x}\, \partial_{\hat{x}} + \hat{z}\, \partial_{\hat{z}} \right) \hat{A} \big], \quad (\hat{x}, \hat{z}) \in \mathbb{R}^2 \\ &\hat{A}(\hat{r}, \hat{\phi}) \sim c(\hat{\phi})\, \hat{r}^{-3/2} \quad \text{as} \quad \hat{r} \to \infty, \end{aligned} \tag{2.126}$$

where Δ denotes the Laplacian with respect to $\hat{x}, \hat{z}$.

Remark 2.67. It is fairly obvious that the first term on the right hand side in (2.126) is equal to $\frac{\lambda}{2\pi}\, \mathcal{J}_{-\infty}^{1/2} \mathcal{R}^1 \, [\partial_{\hat{x}}^3 + \partial_{\hat{x}} \partial_{\hat{z}}^2] \hat{A}$ and thus one obtains the same operators as for the Cauchy problem, with ∂_t being replaced by $I + \frac{2}{3} \left(\hat{x}\, \partial_{\hat{x}} + \hat{z}\, \partial_{\hat{z}} \right)$, where I denotes the identity. And, in virtue of Remark 2.14, by assuming $\hat{A} = \hat{A}(\hat{x})$ in (2.126) one obtains the according two-dimensional equivalent of the similarity profile equation, cf. Smith (1982) and Scheichl et al. (2008).

Remark 2.68. As mentioned in Remark 2.23, Duck (1990) considered a globally (z-symmetric) three-dimensional Cauchy problem for A, where the exact same equation determining $\hat{A}$ was derived. Due to the approach used therein, the combination of the Riesz transform and the Abel operator, as given in (2.126), is replaced by a type of potential integral with an additional kernel involving a complete elliptic integral. The equivalence of these two descriptions is shown in Section 3.3.2, Remark 3.66. Hence, we can assert the independence of the blow-up profile of whether the z symmetric physical set-up was initially assumed to be (globally) three-dimensional or the z dependence was introduced to the planar problem by forcing functions. This may not apply to problems where no z symmetry was assumed.

Remark 2.69. One of the most important characteristics of (2.126) is that it is a *nonlinear, homogeneous* equation and hence $\hat{A} \equiv 0$ is obviously a solution, which cannot be admissible as a similarity profile in the vicinity of the blow-up. Hence, certain measures have to be taken to find possible non-trivial solutions.

To study the full characteristics of the blow-up profile one has to solve (2.126), which is essentially a numerical task. The main issue, as stated in Remark 2.69 and Scheichl et al. (2008), is the nonlinearity and how to keep the thus needed iteration schemes from (mainly) converging to the trivial solution, which appears to be "highly attractive". All other arising difficulties are inherited from the Cauchy problem, since the operators involved are the same. As has been said in Section 2.2 regarding numerically solving the steady problems, various different approaches were used in the previous works to address the issues concerning the properties of the singular operators, the unbounded domain and the nonlinearity. Interestingly, for the profile equation (2.126) the finite difference method was found to be most successful, especially compared to the polynomial expansions, which stands in contrast to experiences made when solving the Cauchy problems. Hence, the following shall provide the strategy (in principle) for setting up a finite difference scheme for equation (2.126). For the details we refer to Section 3.4.

First, an application of the mapping

$$\tau : [-1,1] \to \mathbb{R}, \quad \tau(u) = c\,\frac{u}{1-u^2} \tag{2.127}$$

yields a compact computation domain $[-1,1]^2$, which is subsequently divided into evenly distributed squares, cf. (3.127). Here the parameter c provides the possibility to adjust the according distribution of the evaluation points on $\mathbb{R}^2$ in order to capture certain local characteristics of the operators or the solution. We then obtain the discrete unknowns

$$\tilde{A}_{ij} := \tilde{A}(u_i, v_j) = \hat{A}(\tau(u_i), \tau(v_j)), \quad i = 0, \ldots, M_i, \quad j = 0, \ldots, M_j.$$

As (2.126) admits only z symmetric solutions, $v \in [0,1]$, thus saving computational costs.

By abbreviating the discretized operators (following the analysis in Section 3.4) as

$$\mathcal{J}_{-\infty}^{1/2}\mathcal{R}_1\Delta \quad \rightsquigarrow \quad \underline{K}, \qquad \mathcal{J}_{-\infty}^{3/4}\big[I + \tfrac{2}{3}\left(\hat{x}\,\partial_{\hat{x}} + \hat{z}\,\partial_{\hat{z}}\right)\big] \quad \rightsquigarrow \quad \underline{J},$$

where, for simplification of the programming effort, one can split $\underline{K}$ and $\underline{J}$ into discretizations of the individual integrals and derivatives (as described in Section 3.4), one obtains the discrete composition by simple matrix multiplication.

Overall, this leads to the nonlinear system in $\big(\tilde{A}_{ij}\big)_{i,j} =: \underline{\tilde{A}}$

$$\underline{\tilde{A}}^{\,2} \;=\; \underline{M}\;\underline{\tilde{A}}, \quad \underline{M} := -\big[\,\tfrac{\lambda}{2\pi}\;\underline{K} \;+\; \tfrac{2}{3}\gamma\;\underline{J}\,\big]. \tag{2.128}$$

To incorporate the boundary conditions we note that $\tilde{A}(\pm 1, v) = 0$ and $\partial_u \tilde{A}(\pm 1, v) = \partial_v \tilde{A}(u, 0) = 0$, due to the far field condition and the symmetry with respect to z.

Remark 2.70. Theorem 3.41 in Section 3.4 states the convergence of the approximation of the matrix-vector (or tensor-matrix, to be precise) description on the right hand side of (2.128) and also provides estimates for the order of the discretization. Some simple tests show the invertibility of $\underline{M}$ in principle, such that we claim that a Newton method applied to (2.128) can be expected to converge, see Section 2.2, Remarks 2.18 and 2.19.

Caveat: It is somehow (from numerical testing) neither necessary nor recommendable to approximate the Abel operators appearing in (2.126) to higher orders (as has proved successful for the two-dimensional problem, cf. Scheichl et al. (2008)) than the potential integral.

A direct application of the Newton-Powell iteration scheme to equation (2.126) does not yield any other result than the trivial solution (see Remark 2.69), or in other words, it seems almost impossible to choose appropriate starting vectors to obtain any other solution. Thus, as proposed in Scheichl et al. (2008), one has to force the unknown vector not to be identical to zero by keeping one entry fixed at a non-zero constant.

Assume $\tilde{A}$ is a non-trivial solution of the mapped version of (2.126), then there exists a constant $a \neq 0$, such that $\tilde{A}(u,v) = a\Phi(u,v)$. Say $\tilde{A}(u^*, v^*) \neq 0$, setting $a := \tilde{A}(u^*, v^*)$ consequently gives $\Phi(u^*, v^*) = 1$. The substitution of $\tilde{A}_{ij} = a\Phi_{ij}$ into (2.128) yields the

equation system

$$\begin{aligned} a\,\underline{\Phi}^2 &= \underline{M}\,\underline{\Phi} \\ \Phi_{i^*j^*} &= 1, \end{aligned} \tag{2.129}$$

where we have gained an additional unknown a and, as an additional equation, the fact that there exists a pair of indices $\{i^*, j^*\}$ at which $\underline{\Phi}$ has to be 1.

Remark 2.71. It is fairly obvious that *if* an iteration scheme applied to (2.129) converges, it can only yield a solution vector $\underline{\Phi}$ not identical to zero. We do not have to be concerned about the pathological case of obtaining $\Phi_{ij} = 0, \forall i, j \neq i^*, j^*$ as a possible solution, since $\hat{A}$, as a solution of (2.126), is at least twice continuously differentiable with respect to $\hat{x}$ and $\hat{z}$. Hence, if $\hat{A}$ is non-zero at some point, it has to be non-zero in a neighborhood with positive measure around this point. Ergo, if the mesh is sufficiently dense, this pathology will not occur. In other words, a convergent iteration algorithm for the system (2.129) provides strong evidence for the existence of a non-trivial solution of (2.126).

For the final set-up of the numerical scheme a few comments are in order. The parameter c in (2.127) shall be used separately in x and z and seems to have an optimal value at $c \approx 3$, meaning for the mesh in $\mathbb{R}^2$ that the points are distributed away from zero. Apparently this is needed in the finite difference scheme to depict the far field behavior more accurately, resulting in a better convergence behavior of the iteration. A good choice for the initial guess was found to be $(1 + \hat{x}^2 + \hat{z}^2)^{-3/4}$ and $a = 1$, where this function already satisfies (up to a constant) the far field condition (2.124).

As pointed out in Scheichl et al. (2008), the blow-up profile for the two-dimensional case is unique, in the sense that numerically one is not able to find another non-trivial solution. Consequently, tests regarding possible uniqueness of the profile at hand have been performed (by variation of the numerical parameters and the initial guesses), where, not surprisingly, only one non-trivial solution could be found as well.

Also, the convergence of the iteration in principle is highly sensitive, especially with respect to the mapping τ and the according distribution of the points. The reason for this might lie in the additional equation in (2.129), which can be seen as a very strict constraint for possible solutions found by the iteration procedure. Note that if the points are distributed in a way, such that the function in the neighborhood of u^*, v^*, in virtue of the differentiability requirement, cannot be resolved to some order of accuracy, the iteration might not converge.

Support for the claimed uniqueness is found in the estimated profiles from the time evolution, where different initial conditions (i.e. forcing functions) result in different blow-up times and points but yield the same intrinsic structure in the similarity variables.

Concluding, we assert the similarity profile $\hat{A}$ given via (2.121), to be in fact *self-similar*, since the numerical findings show that (2.121) represents the only possible way to introduce similarity variables, such that $A = A(x, z, t)$ seen as the rescaled solution in time from $\hat{A} = \hat{A}(\hat{x}, \hat{z})$ is the unique solution of the original Cauchy problem (2.47). Figure 27 finally shows the good agreement between the estimated blow-up profile from the time evolution and the non-trivial solution satisfying (2.126). Combining the (numerically found) uniqueness of

the blow-up structure and the independence of the physical set-up, cf. Remark 2.68, we thus found a generic intrinsic structure for the finite time singularity of the Cauchy problems (2.49) and (2.47).

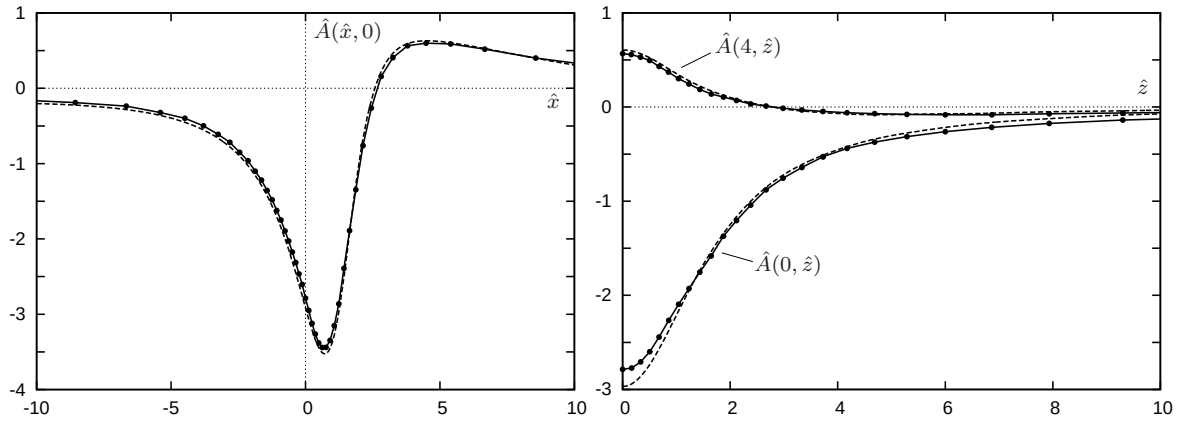

Figure 27: The blow-up profile gained from the finite difference scheme (solid lines) with equidistant grid $(M_i, M_j) = (50, 30)$ on $[-1, 1] \times [0, 1]$ versus the estimated profile from the time evolution, cf. Figure 26 (dashed lines).

Remark 2.72. The uniqueness regarding the two-dimensional version can be further supported by solutions found applying Chebyshev polynomial expansions. Without any constraints to avoid the trivial solutions we were able to obtain the exact same profile as presented in Scheichl et al. (2008) using between $N = 30$ and $N = 80$ polynomials. Interestingly the numerical experiments performed showed that the non-trivial solution, although being unique, is not attractive or stable in the sense that the Newton algorithm immediately converges to the trivial solution, if the initial guess, or some intermediate iterated solution, is not already close to the sought solution.

A *homotopy approach*, such as taking the steady state problem, where upper branch solutions are easily found (see Section 2.2) can be written as

$$A^2 + (1-s)(2A\sqrt{1+x^2} + \Gamma + 1 - f) = \lambda \mathcal{J}_\infty^{1/2} \partial_x^2 A - s\,\tfrac{2}{3}\gamma\, \mathcal{J}_{-\infty}^{3/4}[I + \tfrac{2}{3} x \partial_x] A, \quad s \in [0,1],$$

where $s = 0$ represents the steady problem and $s = 1$ the blow-up profile equation. Although no proof for this to obtain the non-trivial solution is available, with the consistency properties of the operators, see Section 3.3.1, and the convergence of the iteration algorithm and the orthogonal projection, enough evidence for the uniqueness of the blow-up profile for the two-dimensional problem has been presented. As for the profile shown in Figure 27, the polynomial approach did not yield satisfying results, which might be due to the difficulty finding appropriate functions to start the iteration scheme.

Remark 2.73. The evidence that there might be one and only one non-trivial blow-up profile seems sufficient, but this fact per se is highly non-trivial. Galaktionov (2009) investigated

the blow-up and its structure for a fourth-order reaction diffusion type problem for different values of the power of the nonlinearity. It was found therein that at least two non-trivial solutions exist, even for less than quadratic nonlinearities. Surprisingly, some such solution turned out to be strictly positive and hence sign changes and zeros of the similarity profiles are important properties. The results obtained from the numerical experiments performed with respect to various grids and initial guesses regarding equation (2.129) show no possibility for non-trivial solutions of (2.126) to be non-positive or non-negative, which leads even more to the conclusion of uniqueness of the blow-up profile depicted in Figure 27.

Remark 2.74. If one formally assumes the Cauchy problem and according solutions to have some meaning for $t > t_s$, then a reconsideration of the similarity variables (2.121) for $\tau < 0$ would mean the blow-up is approached from the "future" and the question thus raised is whether the intrinsic structure is different. Say, for $\tau > 0$

$$x - x_s = (-\tau)^{4/9}\hat{x} = (-1)^{4/9}\tau^{4/9}\hat{x},$$

where the value of $(-1)^{4/9}$ shall be defined as δ^4, where δ is chosen from all roots of $\delta^9 = -1$, which has nine symmetric solutions on the unit circle, i.e. if δ_i is a solution, so is the complex conjugate. Since nine is an odd number, one δ_i is always found to be on the real line, i.e. ± 1. Thus, $(-1)^{4/9} = 1$, as well as $(-1)^{-2/3} = 1$ is always a possibility and we hence assert a formal *time symmetry* for the blow-up structure, meaning that if a solution does exist in some sense beyond the blow-up time, the structure of the singularity does not change if approached backward in time.

In the above we established the existence and uniqueness of self-similar finite time blow-up solutions of the Cauchy problem (2.47), regularized using the direct implicit numerical technique, which corresponds to filtering or cutting off higher order Fourier coefficients of the solution. But, as demonstrated in Section 2.3.1, this is not the only possibility to gain well-posedness of the time evolution, adding higher derivative operators, as we have shown, works equally well. Moreover, these do also have a physical correlation to streamline curvature contributions (cf. the streamline curvature term (2.120)). Including such terms actually alters the Cauchy problem and yields similar results to the filtering technique only if the regularization parameter is small enough. Nevertheless, with respect to the problem posed in this section, i.e. starting from the upper branch steady state and applying the blowing device, we have to consider the time evolution with these regularizing operators present. Recalculating the first time steps, cf. Figure 20, using the term from (2.85) with the parameter $\alpha = 0.01$, shows almost negligible differences in the structure of the solutions (as expected). When computing further in time, the mollifying characteristics of the regularization become more apparent in the sense that the minimum indicating the singularity is shifted and the blow-up per se seems to be delayed. This, of course, raises the question whether the finite time singularity, when including the higher derivatives, actually occurs. This is reasonable to ask, since if the absolute value of the solution A becomes infinite at some point, all

terms in the equation including A are equally relevant. From a more detailed numerical investigation we still claim the existence of the finite time blow-up, even in the presence of the regularizing operators, as long as the parameter α is sufficiently small. In fact, one might claim the singularity to result from the presence of the interaction term (i.e. the potential integral in Equation (2.48)), as this term, in principle, acts magnifying to the absolute value of a solution over time. Only when starting near an upper branch steady state solutions remain bounded for finite radii $|(x, z)|$. Interestingly, as shown in Smith & Elliott (1985), when considering the non-interactive boundary layer equations and deriving the according solvability condition, one also arrives at problem (2.47), but without the interaction term. Then, as further demonstrated therein, the time evolution of A does *not* terminate in a singularity, but develops a shock-like structure (locally in some bounded region away from the far field). This we were able to confirm numerically with the methods described in the previous sections. Such findings then support the assertion that the interaction term is mainly responsible for the blow-up scenario (under the conditions established above). Although in the non-interactive case no (classical) derivative of A is present in the equation, the issue of ill-posedness, or instability of certain disturbances, still remains in some sense (cf. the arguments made by Smith & Elliott (1985)). Nevertheless, including higher order derivatives or using the implicit time integration does not show any fast growing oscillations in the time evolution but a more or less mollified shock, depending on the regularizing parameter.

Remark 2.75. As said in the preceding paragraph the finite time blow-up still occurs with the regularizing operators included. Applying then the arguments leading to the self-similar coordinate scalings (2.121) one obtain the balances

$$A^2 \;\sim\; |x - x_s|^{-3/2} A \;\sim\; |x - x_s|^{3/4} (t_s - t)^{-1} A \;\sim\; |x - x_s|^{-5/2} A,$$

where it is obviously not possible to take into account all four terms for the similarity transform. So the argumentation would now be that the square on the left hand side and the time derivative term have to be included, but from the spatial derivative type operators only one can be present to determine the blow-up profile. Since the Cauchy problem does not experience the singularity without the interaction term (but instead shows the above mentioned shock), we assert that in the vicinity of the blow-up the regularization does not contribute to the intrinsic structure. In other words, the blow-up scenario has to be seen as the regularized solution of the Cauchy problem subject to certain initial conditions in the limit of vanishing regularization.

2.4 Concluding Remarks

In the preceding sections, apart from providing accuracy and convergence aspects of the numerical schemes used, we studied the two main characteristics of the present description of marginally separated flows in a locally three-dimensional set-up. That is, the ill-posed Cauchy problem and the finite time singularity. It is of high importance at this point to

emphasize the difference between them, as one might associate both with the *breakdown* of the asymptotic triple-deck structure. We shall therefore give some further explanations and clarifications on these (general) subjects.

Let us first note that a finite time blow-up is *not* necessarily connected to the ill-posedness per se. Consider, for example, the reaction-diffusion problem $\partial_t u = \Delta u + F(u)$, which is certainly well-posed on some Sobolev space W (assuming F to be (locally) Lipschitz). Nonetheless, for some initial conditions one can show $\lim_{t\to T} \|u\| \to \infty$, depending on F, of course. In other words, in this case one has existence, uniqueness and continuous dependence of solutions on the data on $t \in [0, T)$. Replacing Δ with $-\Delta$ renders the problem ill-posed on W (in general), such that one cannot even speak of solutions for any $t > 0$, and hence certainly not of a finite time blow-up.

Additionally, as has been sufficiently demonstrated in Section 2.3.1, the ill-posedness is an intrinsic property of the Cauchy problem itself and is not introduced by linearization or discretization. In fact, in some situations (as it is the case here) discretization can be a type of regularization, yielding a well-posed finite-dimensional approximation. Furthermore, ill-posed problems shall always be considered in connection with the function spaces solutions and initial conditions lie in. Thus, heuristically, we might just say that the problem is ill-posed because of the lack of additional a priori information on the sought solutions. Eventually, having found an appropriate regularization, one can focus on properties of the actual time evolution.

We shall make another, more physical, remark on the ill-posedness here. As it is often the case when applying asymptotic expansions, due to their non-uniqueness, choosing different coordinate scalings and different expansion orders might lead to completely different problems. The slow time scale given in (2.6) is introduced in a way to enter the solvability condition in the lower deck, cf. Equation (2.28). Note the fact that no time derivative appears even in the investigated equations for the higher order terms, just a variation in time of the inhomogeneity (cf. Equation (2.20)). Hence, using a different time scale yields different problems for the expansion terms in the three decks and might also give a well-posed Cauchy problem for A.

Concluding further, one can assert that the ill-posedness is not necessarily "un-physical", but might just lack the information to resolve appearing short-scaled parts of possible solutions. We may thus justify the argumentation and numerical results given in Section 2.3.1 by not allowing functions containing higher order Fourier terms to be solutions of the posed problem. The proposed regularizing operators in form of higher derivatives (stemming from the streamline curvature) can be seen as the resulting mathematical formulation of such a physical argument.

In doing so, we have given some meaning to the steady states computed in Section 2.2 by (numerically) investigating the long time behavior in the vicinity of these equilibria (cf. Remark 2.48 and what follows) and by applying control (or forcing) functions (see the paragraph after Remark 2.53). With the relation of the solution A to the wall shear stress τ (see

Section 2.1), one can draw some physical conclusions regarding the separation of the boundary layer. Take, for example, the upper branch solution in Figure 3, where $A > 0$, $\forall x$, i.e. the flow is fully attached. Starting the time evolution with a small separation region already present (e.g. by slightly altering the fully attached steady state accordingly), the numerical results then show the attraction of the upper branch steady state. Thus, without any further disturbances, the separation bubble disappears over time, neither destroying nor significantly changing the main part of the boundary layer. One can hence claim the triple-deck structure and its asymptotic expansions of the flow field to be uniformly valid in such situations.

The crucial situation (being the other main characteristic of the Cauchy problem) appears for a sufficiently strong perturbation or if the initial condition lies outside a certain neighborhood of the upper branch steady state (at some Γ) – that is the finite time blow-up scenario. Note again that this singularity occurs despite the well-posedness of the regularized problem. Nevertheless, solutions blow-up only for certain initial settings, whereas otherwise long time existence and uniqueness have been (heuristically) established. Yet another way is to say that within the subspace of admissible solutions and initial conditions regarding the well-posedness of the original Cauchy problem, there exists a subspace of initial conditions, for which finite time singularities can be found.

The blow-up has been argued to be connected to the emergence of vortices within transitional separation bubbles. Since these bubbles are prone to burst (under the influence of small disturbances), such vortical structures can be considered unstable in this sense. Considering the velocity component v_{31} in the lower deck, as given in (2.15), shows its singular absolute value at the blow-up point, meaning that the lower deck significantly changes the behavior in the boundary layer and the outer flow region. Furthermore, it indicates the breakdown of the triple-deck structure of marginal separation (due to the non-uniform validity of the according asymptotic expansions for all times).

This raises the question of what happens at and beyond this singularity or whether the singularity terminates the whole evolution problem. Here one has to take into account several aspects.

(*i*) At the singularity, independently of whether the time evolution can be continued, we have shown in Section 2.3.3 the existence of a self-similar structure as $t \to t_s$, containing a generic and unique blow-up profile. Also, demonstrated by numerical results, there appear to be only singular blow-up points, or say, blow-up regions with (Lebesgue) measure zero. Nevertheless, the possibility of multiple blow-up points (cf. Figure 25 and Remark 2.63), although z-symmetric, exists. The most important implication then is that independently of when and where the singularity occurs, the time evolution always terminates in the same manner, revealing the unique self-similar profile.

(*ii*) Galaktionov & Vázquez (2002) studied blow-up scenarios in virtue of the issues of *when*, *where* and *how* it occurs and what happens *beyond*. The former have been answered for the problem at hand in Section 2.3.3. From the explanations in the previous item we infer that, regarding the possibility of continuation, none of the first questions has immediate

influence. It is rather the spatial structure of the singularity which determines the behavior beyond blow-up (and whether there is even one). Thus, Galaktionov & Vázquez (2002) distinguish three cases. First, the *complete* blow-up, for which solutions cannot be continued, since they then would have to be infinite everywhere. Second, the *incomplete* blow-up, where the solutions remain bounded in some regions for $t > t_s$ (and infinite in the complement). And at last, the *transient* blow-up, resulting in an everywhere bounded solution immediately after the occurrence of the singularity. We may claim, in the case investigated here, an incomplete scenario to be present, since the formal time symmetry argument (cf. Remark 2.74) suggests the same self-similar structure to appear when formally approaching the blow-up time from beyond. In fact, a condition for the existence of the self-similar variables is the boundedness of the solution away from the (spatial) blow-up point.

So the question rather is whether the singularity is (spatially) integrable in some sense. In other words, whether one can find an integral over the solution for which the time derivative is zero for *all* times. This leads to the subject of finding weaker norms and according function spaces for the solution to lie in, such that the problem does not experience a finite time blow-up in this sense. Ergo, for weak and mild solutions a long time behavior might exist. If this is the case, one can treat the solution as an almost everywhere bounded function with (maybe moving) regions of unbounded absolute value. In Braun & Kluwick (2004) this has been investigated with respect to the Fisher equation, where two moving singularities appear beyond blow-up. Note that the conservation of some integral (norm) of the solution is still an open issue in this case.

(*iii*) A completely different approach is the consideration of the fact that (in general) any kind of singularity or unboundedness appearing at some point in space-time in some expansion term at some order of an asymptotic series shows the breakdown of the proposed expansion (cf. Goldstein's singularity in classical steady boundary layers). As mentioned above, such an expansion is then not *uniformly* valid on the defined domain. This is exactly what happens at the blow-up time, i.e. the function A becomes singular and, for being included in the expansion terms of the velocity field in every deck (see Section 2.1), we infer the uniform validity of the triple-deck expansions to be violated. Already mentioned in Section 2.1 and as shown in Stewartson (1970), the Goldstein singularity cannot be resolved using triple-deck arguments. Interestingly in the present case we have two situations. On the one hand, one can find settings (i.e. different initial conditions), for which the expansions in all decks remain (heuristically) uniformly valid for all times. On the other hand (and in contrast to the Goldstein singularity), the finite time blow-up and the form of the breakdown of the expansions consistently define shorter spatio-temporal scales, which reveal a new triple-deck type problem.

A formal connection between the Cauchy problem (2.47) in terms of A and a reaction-diffusion type equation (here Fisher's equation) has been shown in Braun & Kluwick (2002), Braun & Kluwick (2003) and Braun & Kluwick (2004) to hold in the vicinity of the critical value of the parameter Γ, cf. Figure 4. The idea, in principle, is to say $\Gamma_c - \Gamma =: \epsilon^4$ and

$A(x,z,t) = A_{2d}(x) + \epsilon^2 a_1(x,\epsilon z,\epsilon^2 t) + \ldots$ and to then determine a_1, whereas A_{2d} is the solution of the two-dimensional stationary problem (2.33) at Γ_c. In Braun & Kluwick (2004) further investigations then revealed that $a_1 = b(x)c(\epsilon z, \epsilon^2 t)$ and (for the sake of readability say $\epsilon z \to z$, $\epsilon^2 t \to t$) the behavior of the function c to be governed by

$$\partial_t c = \partial_z^2 c + c - c^2.$$

If a_1 in fact does approximate the dynamics of A near Γ_c, then, what is most striking here is the fact that the Fisher equation is *well-posed*. One might be lead to infer the well-posedness from the independence of c from x, cf. the Fourier symbol (2.78). Although, such an argument can hold in some cases, here we rather claim the well-posedness to be related to the time scaling $t \to \epsilon^2 t$, which results in different combinations of time and spatial derivatives (with respect to z). Nevertheless, as remarked at the beginning of this section, finite time blow-up can occur in such nonlinear well-posed evolution problems. Using classical PDE analysis, one can show for locally Lipschitz continuous nonlinearities (as it is clearly the case here), in the usual L^2 set-up, that there exists a t_{max}, until which a unique weak solution exists and if $t_{max} < \infty$, then the L^2 norm of the solution becomes unbounded at t_{max}. Note that if one would want to argue a possible continuation after blow-up, other L^p norms would have to be used, which might not be found. In conclusion, even though the time scale of the Fisher equation is slower than the one for A, near the bifurcation point Γ_c the upper and lower branch steady states and the possible singularity found for A are mirrored here for c.

As argued at the end of the previous section, we can leave out any type of regularizing operators near the blow-up point and hence continue with the original problem. Following the ideas in Smith (1982) and Elliott & Smith (1987) (for the planar flow case) from the singularity in A we can deduce, via the relationship (2.29), that the term containing the interaction pressure, being initially asymptotically small compared to the imposed pressure gradient (cf. the expansion for p_3 in (2.14)), becomes comparable to the latter when approaching the blow-up time. Consequently, one wants to find the time scale where both terms become of the same order, i.e.

$$O\big(\epsilon^4 p_{00}\big) = O\big(\epsilon^{10}\partial_x p_i\big) \quad \text{as} \quad t \to t_s. \tag{2.130}$$

From (2.121) we have that $x = z = O(\tau^{4/9})$ as $\tau \to 0$ and thus with

$$p_i = \int\limits_{\mathbb{R}^2} \underbrace{\frac{1}{|(x,z)-(\xi_1,\xi_2)|}}_{=O(\tau^{-4/9})} \underbrace{\partial_{\xi_1}^2 A(\xi_1,\xi_2,t)}_{=O(\tau^{-14/9})} \underbrace{d\xi_1 d\xi_2}_{=O(\tau^{8/9})} \quad \Rightarrow \quad O(p_i) = O(\tau^{-10/9}),$$

and consequently $\partial_x p_i = O\big(\tau^{-14/9}\big)$. The fact that p_{00} is an order one quantity in all limits, (2.130) provides a connection to ϵ and hence to the Reynolds number, i.e. $\tau^{14/9} = \epsilon^{-6} = Re^{3/10} \Rightarrow \tau = O\big(Re^{-27/140}\big)$. Since we want to utilize this relationship to obtain rescaled spatio-temporal variables $(\tilde{x},\tilde{y},\tilde{z},\tilde{t})$, such that $\tilde{t} \to -\infty$ correlates to t_s, using (2.6) yields

$|t_s - t| = \epsilon|t_s^* - t^*| = O\big(Re^{-27/140}\big)$ and therefore

$$|t_s^* - t^*| = O\big(Re^{-27/140+1/20}\big) = Re^{-1/7}\tilde{t} \quad \Rightarrow \quad t^* = Re^{1/20}t_s + Re^{-1/7}\tilde{t},$$

with $\tilde{t} = O(1)$ as $Re \to \infty$ (cf. Smith (1982)). Redefining $\epsilon := Re^{-1/7}$ and assuming the blow-up point shifted to the origin, we obtain further the rescaled spatial variables $x^* = O(Re^{-1/5}\tau^{4/9}) = Re^{-2/7}$, thus $x^* = \epsilon^2\tilde{x}$ and analogously $z^* = \epsilon^2\tilde{z}$. As we have defined this new scalings with the individual pressure contributions being of the same order, we can simply say $p^* = O(Re^{-1/5}p_{00}x) = O(Re^{-1/2}p_i) = O(Re^{-1/2}\tau^{-10/9})$ and therefore $p^* = \epsilon^2\tilde{p}$.

Remark 2.76. Being a little more precise regarding the rescaled pressure $\tilde{p}$, or say its gradient with respect to $\tilde{x}$, we shall consider (cf. Braun et al. (2012)) as $\tau = O(Re^{-27/140})$

$$\partial_{x^*}p^* = Re^{-1/5}\partial_{x^*}p_{00}x + Re^{-1/2}\partial_{x^*}p_i = O(1) \quad \Rightarrow \quad \partial_{x^*}p^* - p_{00} = \partial_{\tilde{x}}\tilde{p}$$

and with $\partial_{\tilde{x}}\tilde{p} = O(1)$ one also obtains $p^* = \epsilon^2\tilde{p}$ (as found above).

From the fact that the appearance of the singularity is inherently a time-dependent process related to fast growing pressure perturbations, it is clear that within the Navier-Stokes equations, viewed in this new scales, the time derivative has to be of the same order as the pressure gradient, leading to

$$\partial_t^* \begin{pmatrix} u^* \\ v^* \\ w^* \end{pmatrix} \sim \nabla p^* \quad \Rightarrow \quad \partial_t^* \begin{pmatrix} u^* \\ v^* \\ w^* \end{pmatrix} = \epsilon^{-1}\partial_{\tilde{t}} \begin{pmatrix} u^* \\ v^* \\ w^* \end{pmatrix} \sim \begin{pmatrix} \partial_{x^*} \\ \partial_{y^*} \\ \partial_{z^*} \end{pmatrix} p^* = O(1)$$

$$\Rightarrow \quad \begin{cases} u^* = \epsilon\tilde{u} \\ w^* = \epsilon\tilde{w}, \end{cases}$$

since $O(\partial_{x^*}p^*) = O(\partial_{\tilde{x}}\tilde{p})$ and x^* scales as z^*, whereas the scaling factor for y^* is yet unknown. Next, we observe that the blow-up essentially happens within the viscous sublayer (the lower deck in Section 2.1) and consequently one has to balance the pressure gradient with the viscous terms, yielding

$$\nabla p^* \sim \epsilon^7\Delta \begin{pmatrix} u^* \\ v^* \\ w^* \end{pmatrix} \quad \Rightarrow \quad \begin{cases} \partial_{x^*}p^* \sim \epsilon^7\Delta u^* = \epsilon^7\epsilon^{-4}\epsilon(\partial_{\tilde{x}}^2 + \partial_{\tilde{z}}^2)\tilde{u} + \epsilon^7\epsilon^{-2a}\epsilon\partial_{\tilde{y}}^2\tilde{u} \\ \partial_{y^*}p^* \sim \epsilon^7\Delta v^* \\ \partial_{z^*}p^* \sim \epsilon^7\Delta w^* = \epsilon^7\epsilon^{-4}\epsilon(\partial_{\tilde{x}}^2 + \partial_{\tilde{z}}^2)\tilde{w} + \epsilon^7\epsilon^{-2a}\epsilon\partial_{\tilde{y}}^2\tilde{w} \end{cases}$$

$$\Rightarrow \quad \epsilon^8\epsilon^{-2a} \stackrel{!}{=} 1 \quad \Rightarrow \quad y^* = \epsilon^4\tilde{y},$$

since $O(\partial_{x^*}p^*) = O(\partial_{z^*}p^*) = O(1)$, where the y gradient of the pressure compared to the viscous term in v remains undetermined at this point. Finally, we have to take into account the conservation of mass. Without making any a priori assumptions or restrictions, we have

in general

$$\partial_{x^*}u^* \sim \partial_{y^*}v^* \sim \partial_{z^*}w^* \quad \Rightarrow \epsilon^{-1}\partial_{\tilde{x}}\tilde{u} \sim \epsilon^{-4}\epsilon^{a}\partial_{\tilde{y}}\tilde{v} \sim \epsilon^{-1}\partial_{\tilde{z}}\tilde{w} \quad \Rightarrow \quad v^* = \epsilon^3\tilde{v}.$$

Considering now the nonlinearity in the Navier-Stokes equations and comparing its order to, say the pressure gradient, yields

$$\nabla p^* \sim \begin{pmatrix} u^* \\ v^* \\ w^* \end{pmatrix} \cdot \nabla \begin{pmatrix} u^* \\ v^* \\ w^* \end{pmatrix} \quad \Rightarrow \quad \epsilon^{-1}\partial_{\tilde{x}}\tilde{p} \sim \epsilon^{-1}\tilde{u}\partial_{\tilde{x}}\tilde{u} + \epsilon^{-1}\tilde{v}\partial_{\tilde{y}}\tilde{u} + \epsilon^{-1}\tilde{w}\partial_{\tilde{z}}\tilde{u}$$

and analogously for the y and z gradient, which means that we are dealing with a fully nonlinear problem for the lower deck.

We shall thus summarize the above found scalings (for comparison review the original scalings (2.6) and the according expansions for the three decks, (2.7), (2.10) and (2.14)) by regarding the resulting triple-deck as the next stage within the marginal separation setting, see Figure 28. As argued in Elliott & Smith (1987) this new structure comprises a potential

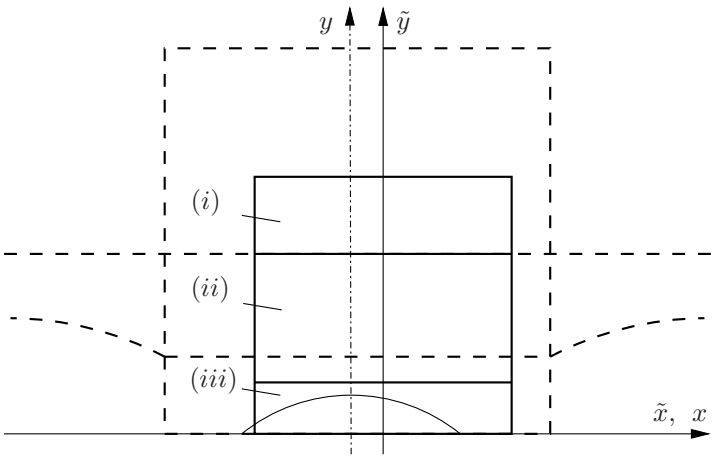

Figure 28: The next stage triple-deck structure, cf. Figure 2 and the dashed lines here, with the rescaled spatial coordinates $(\tilde{x}, \tilde{y}, \tilde{z})$ and the fast time scale $\tilde{t}$. Again the regions (i)–(iii) indicate the potential flow, the main part of the boundary layer and the viscous sublayer, respectively.

flow region, transforming displacements into pressure perturbations, an inviscid, rotational boundary layer and the viscous sublayer derived above. In virtue of the descriptions for the three decks given in Section 2.1 we obtain the following scaled coordinates and expansions for the next stage

$$t^* = \epsilon\,\tilde{t}, \quad x^* = \epsilon^2\,\tilde{x}, \quad z^* = \epsilon^2\,\tilde{z}, \quad y^* = \begin{cases} \epsilon^2\,\tilde{y}_1 \\ \epsilon^{7/2}\,\tilde{y}_2 \\ \epsilon^4\,\tilde{y}_3 \end{cases}, \quad \epsilon = Re^{-1/7} \tag{2.131}$$

upper deck (i)	main deck (ii)	lower deck (iii)
$\tilde{u}_1 \sim U_{00} + \epsilon^2\,\tilde{u}_{11}$	$\tilde{u}_2 \sim U_0(\tilde{y}_2) + \epsilon^{1/2}\,\tilde{u}_{21}$	$\tilde{u}_3 \sim \epsilon\,\tilde{u}_{31}$
$\tilde{v}_1 \sim \epsilon^2\,\tilde{v}_{11}$	$\tilde{v}_2 \sim \epsilon^2\,\tilde{v}_{21}$	$\tilde{v}_3 \sim \epsilon^3\,\tilde{v}_{31}$
$\tilde{w}_1 \sim \epsilon^2\,\tilde{w}_{11}$	$\tilde{w}_2 \sim \epsilon^2\,\tilde{w}_{21}$	$\tilde{w}_3 \sim \epsilon\,\tilde{w}_{31}$
$\tilde{p} \sim p_0 + \epsilon^2\,(p_{00}\tilde{x} + \tilde{p}_i)$	$\tilde{p} \sim p_0 + \epsilon^2\,(p_{00}\tilde{x} + \tilde{p}_i)$	$\tilde{p} \sim p_0 + \epsilon^2\,(p_{00}\tilde{x} + \tilde{p}_i)$.

(2.132)

Substituting these expansions into (2.2) yields (in the same way as in the upper deck in Section 2.1, cf. problem (2.8)) $\tilde{p}_i$ to satisfy the Neumann problem of the Laplace equation in the upper deck. In the main deck, for being a classical boundary layer, conservation of mass and the momentum equations read

$$\left.\begin{aligned} &\partial_{\tilde{x}}\tilde{u}_{21} + \partial_{\tilde{y}_2}\tilde{v}_{21} = 0\\ &\tilde{v}_{21}U_0' + U_0\partial_{\tilde{x}}\tilde{u}_{21} = 0\\ &\partial_{\tilde{y}_2}\tilde{p}_i = 0\\ &U_0\partial_{\tilde{x}}\tilde{w}_{21} = -\partial_{\tilde{z}}\tilde{p}_i \end{aligned}\right\} \quad\Rightarrow\quad \begin{aligned} &\tilde{u}_{21} = A_1(\tilde{x},\tilde{z},\tilde{t})U_0'(\tilde{y}_2)\\ &\tilde{v}_{21} = -U_0(\tilde{y}_2)\partial_{\tilde{x}}A_1(\tilde{x},\tilde{z},\tilde{t})\\ &\tilde{w}_{21} = -\frac{1}{U_0(\tilde{y}_2)}\int\limits_{-\infty}^{\tilde{x}} \partial_{\tilde{z}}\tilde{p}_i(\xi,\tilde{z},\tilde{t})d\xi, \end{aligned}$$

with an yet unknown function A_1, similarly to A from the original marginal separation expansions, representing a shift or correction of the displacement thickness in the lower deck.

The crucial problem for this triple-deck structure lies, as usual (cf. Section 2.1), in the lower deck, which is obviously governed by (omitting the tilde and all indices)

$$\left.\begin{aligned} \partial_t u + (u,v,w)^\intercal\cdot\nabla u &= -(p_{00} + \partial_x p) + \partial_y^2 u\\ \partial_y p &= 0\\ \partial_t w + (u,v,w)^\intercal\cdot\nabla w &= -\partial_z p + \partial_y^2 w\\ \operatorname{div}(u,v,w) &= 0 \end{aligned}\right\} \quad \forall (x,y,z,t)\in\mathbb{R}\times\mathbb{R}^+\times\mathbb{R}\times[0,T]. \tag{2.133}$$

The according boundary and matching conditions are, of course, the no-slip condition at the surface $y=0$, a matching to the boundary layer as $y\to\infty$ and as $|(x,z)|\to\infty$, yielding

$$\begin{aligned} &u = v = w = 0 && \text{at}\ \ y = 0\\ &u \sim p_{00}(y^2/2 + yA_1),\quad v \sim -p_{00}\frac{y^2}{2}\,\partial_x A_1,\quad w \sim \frac{2c(x,z,t)}{p_{00}y^2} && \text{as}\ \ y\to\infty\\ &u \sim p_{00}y^2/2,\quad p, A_1 \to 0 && \text{as}\ \ |(x,z)|\to\infty, \end{aligned} \tag{2.134}$$

where the function c abbreviates the x integral of the z derivative of the interaction pressure (which is independent of y) given in the solution of $\tilde{w}_{21}$. Similar to the marginal separation case, the interaction relation (2.29) from the potential flow region reads

$$p(x,z,t) = -\frac{1}{2\pi}\int\limits_{\mathbb{R}^2} \frac{1}{|(x,z)-(\xi_1,\xi_2)|}\partial_{\xi_1}^2 A_1\,d\xi_1 d\xi_2. \tag{2.135}$$

Remark 2.77. In contrast to the original marginal separation set-up, the triple-deck for the next stage has a nonlinear and unsteady lower deck (cf. (2.20), which contains no time derivative of the velocity field). Furthermore, the u and w component satisfy the same momentum equation and (2.131) and (2.132) yield the y gradient of the pressure to be zero and the Laplace term to be represented only by the y derivative, which indicates the problem to be of boundary layer type.

The procedure used to obtain (2.134) represents the simplest form of a matching rule. Going into more detail here, one can either solve (2.133) for large y (where the viscosity term vanishes) applying the method of characteristics or use additional information, to obtain higher order corrections to those given in (2.134). It can be argued, by viewing the function A_1 as the local displacement of the boundary layer (caused by the sublayer), that the streamlines of the boundary layer are shifted by $y_2 + A_1$ at every $z = const.$, i.e. the separation profile reads $U_0 = U_0(y_2 + A)$ (see e.g. Ryzhov (1980)). Hence, with $U_0(y_2) = p_{00}y_2^2/2$ (as $y_2 \to 0$), one obtains

$$u \sim \frac{p_{00}(y+A_1)^2}{2}, \quad v \sim -\frac{p_{00}(y+A_1)^2}{2}\partial_x A, \quad w \sim \frac{2c}{p_{00}(y+A_1)^2} \quad \text{as} \quad y \to \infty, \tag{2.136}$$

which has been applied to the streamfunction in the planar case in Elliott & Smith (1987). Modifying these far field conditions shows that they do contain (2.134).

In (2.133) the unknown functions are obviously time dependent, with an explicit time derivative term. Therefore initial conditions have to be posed for the problem to be closed. From the deduction above it is straight forward to see that for such a condition to be found one has to consider the match with the previous stage for $\tau \to 0$, i.e. $t \to -\infty$. As done in Elliott & Smith (1987) for the planar case substituting the scalings, cf. (2.121),

$$x = |t|^{4/9}\hat{x}, \quad y = |t|^{1/9}\hat{y}, \quad z = |t|^{4/9}\hat{z}, \quad A_1(x,z,t) = |t|^{-2/3}\hat{A}_1(\hat{x},\hat{z})$$

into (2.134) or (2.136) yields the expansions

$$\left.\begin{aligned} u(x,y,z,t) &\sim \frac{p_{00}}{2}|t|^{2/9}\hat{y}^2 + p_{00}|t|^{-5/9}\hat{A}_1(\hat{x},\hat{z})\hat{y} + |t|^{-12/9}\hat{u}(\hat{x},\hat{y},\hat{z}) \\ v(x,y,z,t) &\sim -\frac{p_{00}}{2}|t|^{-8/9}\partial_{\hat{x}}\hat{A}_1(\hat{x},\hat{z})\hat{y}^2 + |t|^{-15/9}\hat{v}(\hat{x},\hat{y},\hat{z}) \\ w(x,y,z,t) &\sim |t|^{-12/9}\hat{w}(\hat{x},\hat{y},\hat{z}) \\ p(x,z,t) &\sim p_{00}|t|^{4/9}\hat{x} + |t|^{-10/9}\hat{p}(\hat{x},\hat{z}) \end{aligned}\right\} \quad \text{as} \quad t \to -\infty. \tag{2.137}$$

By then substituting these expansions into the momentum equations (2.133) one obtains the unknown higher order flow field $(\hat{u},\hat{v},\hat{w},\hat{p})$ to be governed by

$$\begin{aligned} &\frac{p_{00}}{2}\hat{y}^2\partial_{\hat{x}}\hat{u} + p_{00}\hat{y}\hat{v} - \partial_{\hat{y}}^2\hat{u} = -\partial_{\hat{x}}\hat{p} - \frac{\hat{y}^2}{2}\hat{A}_1\partial_{\hat{x}}\hat{A}_1 - \tfrac{2}{3}\hat{y}\big(\hat{A}_1 + \tfrac{2}{3}(\hat{x}\partial_{\hat{x}}\hat{A}_1 + \hat{z}\partial_{\hat{z}}\hat{A}_1)\big) \\ &\frac{p_{00}}{2}\hat{y}^2\partial_{\hat{x}}\hat{w} - \partial_{\hat{y}}^2\hat{w} = -\partial_{\hat{z}}\hat{p}, \end{aligned} \tag{2.138}$$

where the usual conditions of no-slip at the boundary, at most linear growth in the far field, together with the relationship between $\hat{p}$ and $\hat{A}_1$ via (2.135), are imposed.

Most remarkable here is the similarity of (2.138) to (2.20) from Section 2.1. In fact, the only difference is that $y_3\partial_t A$ in (2.20) is obviously replaced by the last term in the first line on the right hand side in (2.138). Thus, by following the procedure described in Section 2.1 starting with (2.138) instead of (2.20) and using a shift of the unknowns of the form $\hat{u} = \hat{A}_1^2/(2p_{00}) + \tilde{u}$ and $\hat{v} = -\hat{y}\hat{A}_1\partial_{\hat{x}}\hat{A}_1/p_{00} + \tilde{v}$ we find a solvability condition for (2.138) in terms of $\hat{A}_1$ to be given by the equation for the similarity profile at the blow-up, Equation (2.126) in Section 2.3.3. Considering, in addition, the formulae in (2.125) one can easily see where the change of the original time derivative for A to the term in $\hat{A}_1$ stems from.

As a conclusion from the matching as $t \to -\infty$ we find $\hat{A}_1$ in the initial condition (2.137) to be identical to the blow-up profile given as $\hat{A}$ in Section 2.3.3. This is in very well agreement with the findings in Elliott & Smith (1987) for the planar flow case.

Let us recapitulate what has been derived at this point. In the event of the finite time singularity shorter length and time scales (2.131) appear within the original marginal separation set-up, yielding a new triple deck stage, where the lower deck is governed by the (now nonlinear) momentum equations (2.133), subject to the boundary and matching conditions (2.134), as well as the interaction relation (2.135). Since the flow field is unsteady, initial (matching) conditions, i.e. where the new time variable $t \to -\infty$, have to be imposed via the expansions (2.137). The displacement function $\hat{A}_1$ therein has been shown to be equal to the self-similar blow-up profile $\hat{A}$ in (2.126).

Elliott & Smith (1987) continued the investigation of the planar nonlinear triple-deck problem by linearizing the momentum equation for the streamfunction around some steady function comprising the far field behavior. As a consequence, an approximated dispersion relation (cf. Remark 2.37) was then found, showing $\omega \propto k^{5/3}$. Regarding the well-posedness one can easily see the violation of the necessary condition derived in Section 2.3.1, such that the arguments made further in Elliott & Smith (1987) with respect to a possible finite time breakdown have to be seen in the context of *ill-posedness*. Even though the calculations made contain some simplifications one cannot expect the full, nonlinear problem to be well-posed, which is supported by the numerical findings in Elliott & Smith (1987).

In an analogous manner one can perform a similar analysis for the three dimensional equivalent (2.133) presented above, which we claim to lead to the same conclusions. Remaining an open question is whether and how additional terms such as the streamline curvature term found in Section 2.3.2 can enter and possibly regularize this nonlinear triple-deck problem.

With the analysis and remarks above we have provided some possibilities as to how a continuation of the time evolution after the blow-up can be understood. But, as argued, e.g. in Barenblatt (1979), the self-similar structure and let alone scenarios beyond blow-up are only worth considering, if the singularity is stable (in some sense). Otherwise even the similarity coordinates (2.121) do not have any special meaning. A first, simple test for the

stability would be to alter the data at some $t < t_s$ and compare the changes in the similarity profile. If these are uncontrollable, the similarity structure has no meaning. One can easily perform this for the Cauchy problems from Section 2.3 numerically by slightly changing the values of the coefficients in the expansions (2.51) at some t and computing the blow-up point (x_s, z_s, t_s) and its profile. Here we found that small alterations in the data at any time $t < t_s$ lead to reasonable changes in the blow-up characteristics.

Since this represents a rather heuristic approach, Barenblatt (1979) suggested the stability to be studied by saying, if ζ is the self-similar variable connecting (x, t) and a solution $u(x, t)$ is hence represented by $U(\zeta)$, the latter is stable if for a perturbed solution $u(x, t) = U(\zeta) + w(\zeta, t)$, one has $w(\zeta, t) \to 0$ as $t \to \infty$.

Remark 2.78. Obviously, in the present case, $\hat{A} = \hat{A}(\hat{x}, \hat{z})$ is invariant with respect to the one-parameter group of translations, i.e. shifting the blow-up point and time (together with the formal symmetry in (z, t), cf. Remark 2.74). Therefore, one might write in general for the perturbed state above $u(x, t) = U(\zeta + a) + w(\zeta, t)$.

In Eggers & Fontelos (2009) this definition was adopted in the following form. Given the self-similar coordinates (2.121) and say $\rho = -\log(\tau)$ and $A(x, z, t) = \tau^{-2/3}\hat{A}^*(\hat{x}, \hat{z}, \rho)$, then substitution into the Cauchy problem (2.47) yields an initial value problem for $\hat{A}^*$, where steady states correspond to the originally derived blow-up profile $\hat{A} = \hat{A}(\hat{x}, \hat{z})$. Consequently the time derivative term $\partial_t A$ reads in the new independent variables

$$\partial_t A = \tfrac{2}{3}(-t)^{-5/3}\big[\hat{A}^* + \tfrac{2}{3}\big(\hat{x}\,\partial_{\hat{x}} + \hat{z}\,\partial_{\hat{z}}\big)\hat{A}^* + \tfrac{3}{2}\partial_\rho \hat{A}^*\big],$$

where comparison with (2.125) eventually confirms the blow-up profiles to be equilibria of the initial value problem

$$\gamma\,\mathcal{J}^{3/4}_{-\infty}\,\partial_\rho \hat{A}^* = -(\hat{A}^*)^2 - \tfrac{\lambda}{2\pi}\,\mathcal{J}^{1/2}_{-\infty}\mathcal{R}_1\Delta\hat{A}^* - \tfrac{2}{3}\gamma\,\mathcal{J}^{3/4}_{-\infty}\big[\hat{A}^* + \tfrac{2}{3}\big(\hat{x}\,\partial_{\hat{x}} + \hat{z}\,\partial_{\hat{z}}\big)\hat{A}^*\big],$$

subject to the far field decay as in (2.126) and some initial condition. Here, one can continue by applying the standard techniques of substituting $\hat{A}^*(\hat{x}, \hat{z}, \rho) = \hat{A}(\hat{x}, \hat{z}) + \tilde{A}(\hat{x}, \hat{z}, \rho)$, $|\tilde{A}| \ll 1$, and expanding $\tilde{A}$ in eigenfunctions of the according linearized, right hand side operator. For the present problem, as we have demonstrated on several occasions in the previous sections, the stability might still be best tested using numerical experiments.

Remark 2.79. The Cauchy problem for $\hat{A}^*$ in the form above will most likely not yield traveling wave solutions. To check whether these are still possible, one would have to use the transform $A(x, t) = \tau^{-2/3}\hat{\psi}(\hat{x} + v\rho)$ (see Eggers & Fontelos (2009) for some further and general aspects of self-similar equilibria).

It has been suggested and shown in Weideman (2003) that the occurrence of a finite time blow-up in real solutions of certain evolution equations can be interpreted as vanishing imaginary parts of (permanently) existing complex singularities. The reaction diffusion equation

$\partial_t u = \partial_x^2 u + u^2$, for example, has the behavior $u(x,t) \sim \left(t_s - t + \frac{1}{8}x^2 \log|t_s - t|^{-1})\right)^{-1}$, such that (asymptotically) if $x = \pm i\sqrt{8(t_s - t)\log|t_s - t|}$, u has a complex pole, approaching the real axis as $t \to t_s$.

The blow-up in the Cauchy problem (2.47) is asymptotically given by (2.121), i.e. $A(x,z,t) \sim c(x,z,t)(t_s - t)^{-2/3}$, which is nicely reflected by tracking the minimum of A as $t \to t_s$, see Figure 29. With $c(x,z,t)$ converging to the blow-up profile $\hat{A}$ and hence being bounded, no complex singularities can be found in a similar way as in the example above for $t < t_s$.

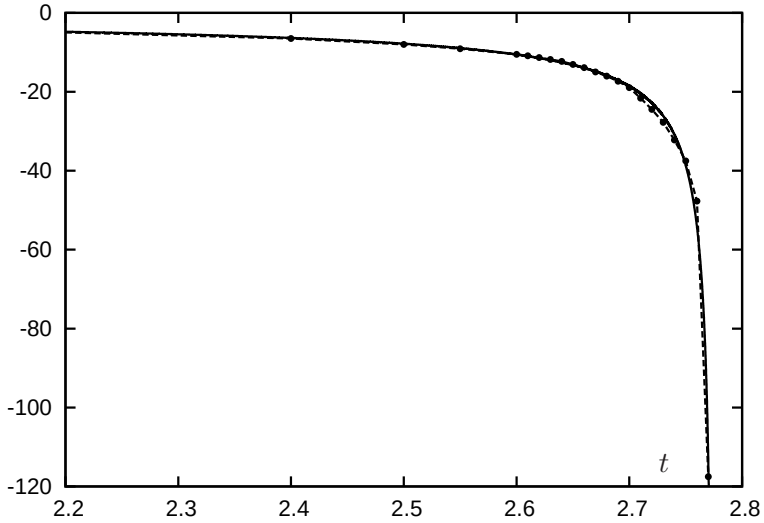

Figure 29: The evolution of the minimum of $A(x,z,t)$ (dashed), cf. Figure 22, versus the algebraic singularity in time suggested by (2.121), e.g. $-3.3\,(2.775 - t)^{-2/3}$ (solid).

As for the Fisher equation, the existence of moving, complex singularities prior to blow-up and moving, real singularities afterwards has been demonstrated in Braun & Kluwick (2004), where this was also connected to a formal time symmetry combined with a change to complex independent variables. When formally replacing t with $-t$ in the Cauchy problem (2.47) it is not straight forward to find according transforms for A and the spatial variables, such that the equations remain invariant. This is due to the explicit appearance of x on the left hand side and the combinations of integral operators on the right. In consequence, although we have shown a formal time symmetry to hold for the blow-up profile, actual continuation beyond blow-up here needs another approach.

We shall make two more remarks on the important issues regarding the uniqueness of the blow-up profile and the continuation of the solution beyond blow-up. As mentioned in the paragraph following Remark 2.72, homotopy analysis methods might be utilized to show the existence and uniqueness of non-trivial blow-up profiles for the two and three dimensional case. The approach used therein, of course, only represents one of many possibilities to introduce the homotopy parameter and perform the numerical calculations. In the theory of homotopy analysis this is known as the zeroth order deformation equation. The whole

technique then is constructive in the sense that one can (numerically) compute solutions for different values of the homotopy parameter and observe the change from the initial to the sought solution. To obtain a classical existence and uniqueness result one might employ *fixed point theorems*, such as Schauder or Banach. The difficulty in the present case certainly lies in the contraction principle needed for using Banach's argument. As for Schauder's theorem "only" continuity (among other characteristics) has to be assumed, which then yields existence but not necessarily uniqueness.

As for the continuation beyond blow-up we stated above that necessarily the conservation of some integral norm has to hold for the function A. One possibility would be to take into account the conservation of mass and momentum in their integral form. Assuming the flow field considered in Section 2.1 satisfies the Navier-Stokes equations then the according integral descriptions are satisfied as well. Substituting the lower deck expansions for the velocity field into the equations yields integral relations for the function A. No actual conservation of A could be found with this approach and hence the integrability and its conservation over time remains an open question.

As an additional perspective for future considerations one might make the assumption (A. I. Ruban, private communications) of a moving surface in or a moving singularity of a solution for the original problem (2.47). First let the surface be moving with a (relative) velocity u_w^*. Given as an additional condition for the original Navier-Stokes equations (2.2), u_w^* enters the boundary conditions for the lower deck expansion term u_{32}, see (2.14), if $u_w^* = Re^{-2/5}u_w$, such that $u_{32} = u_w$ at $y_3 = 0$, cf. (2.21). Following then the analysis of deriving the solvability condition, u_w just changes the boundary conditions for $\widetilde{u}_{32}$ (assuming we have already applied the affine transform) to be $\widetilde{u}_{32} = -A^2 + x^2 - \Gamma + u_w$, cf. (2.24). Eventually, for the steady and unsteady problems (2.30) and (2.47) it is straight forward to deduce that the surface velocity represents a shift in the bifurcation parameter Γ, with the obvious conclusions (for example, it does not alter the self-similar structure).

Another approach would be to assume the spatial location of the singularity (x_s, z_s) to depend on the time $t > t_s$, such that $x_s \to x_s + vt$ and $z_s \to z_s + vt$. Consequently, the self-similar variables (2.121) then read

$$\begin{aligned} x - x_s &= (-\tau)^{4/9}\hat{x} + v(-\tau) \\ z - z_s &= (-\tau)^{4/9}\hat{z} + v(-\tau) \end{aligned} \quad \text{and} \quad A(x,z,t) = (-\tau)^{-2/3}\hat{A}(\hat{x},\hat{z};v,\tau), \quad \tau = t_s - t.$$

It is yet to be established, if such a view has some meaning or yields meaningful solutions in terms of, e.g. traveling waves (cf. Braun & Kluwick (2004) for the Fisher equation).

3 Polynomial Approximation and Numerical Analysis

The subject of describing arbitrary functions (depending on one or more variables) using finitely many simpler functions (of which the behavior is known and well-defined) is named *approximation theory*. This idea is presented in a very general context, for example, in Collatz & Krabs (1973) and goes back at least to *Weierstraß'* theorem, saying that *polynomials are dense in the space of continuous functions*.

Polynomial approximation can then be seen as a special case of linear approximation problems (in contrast to nonlinear problems and/or exponential, trigonometric and rational approximations). In numerical analysis the classical polynomials are almost never the best choice and hence one always turns to *orthogonal* polynomials, such as the well-known *Chebyshev* polynomials.

Since these were originally defined only on $[-1, 1]$ this section deals with extensions to unbounded multi-dimensional domains. First we define how such polynomials can be mapped onto the real line, while remaining smooth and bounded, and prove certain characteristics of these mapped analogues in the context of weighted Lebesgue spaces.

Well-known results in approximation theory applied to spectral methods, such as convergence rates, interpolation properties and the aliasing phenomenon will then be formulated and proved in spaces defined on $\mathbb{R}^n$.

Due to the overall application to problems given in Section 2 we utilize the derived properties of the mapped Chebyshev polynomials, now well-defined in certain function spaces, to solve equations involving singular integral operators using spectral collocation methods. This is done, as usual, in a functional analytic or operator theory setting, by treating the approximations as projections onto finite-dimensional subspaces.

3.1 Rational Chebyshev Polynomials

For a basic and comprehensive introduction to orthogonal polynomials and approximation theory the reader is referred to e.g. Szegö (1939), Timan (1963), Cheney (1966), Guo (1998), Boyd (2001) and Mason & Handscomb (2003). In the following the notion *rational Chebyshev polynomials* is used for classical Chebyshev polynomials T_n mapped onto the whole real axis, i.e.

$$R_n(x) := T_n(\psi(x)) = \cos(n\phi(x)), \quad \forall x \in \mathbb{R}, \tag{3.1}$$

where

$$\begin{aligned} &\psi : \mathbb{R} \to [-1, 1], \quad \psi(x) = \frac{x}{\sqrt{1 + x^2}}, \\ &\phi : \mathbb{R} \to [-\pi, 0], \quad \phi(x) = \arctan(x) - \pi/2, \end{aligned} \tag{3.2}$$

both bijective and diffeomorphic on the according open intervals. Algebraic simplification shows that every mapped polynomial can be written as $R_n(x) = \frac{p(x)}{q(x)}$, where $q(x) \neq 0, \forall x \in \mathbb{R}$. Hence, Boyd (1987) termed them *rational* Chebyshev polynomials, although in the odd case $q(x)$ is not an actual polynomial (due to the square root). Equation (3.3) and Figure 30

depict a few examples.

$$
\begin{aligned}
R_0(x) &= 1, & R_1(x) &= \frac{x}{(1+x^2)^{1/2}},\\
R_2(x) &= \frac{x^2-1}{x^2+1}, & R_3(x) &= \frac{x^3-3x}{(1+x^2)^{3/2}},\\
R_4(x) &= \frac{x^4-6x^2+1}{x^4+2x^2+1}, & R_5(x) &= \frac{x^5-10x^3+5x}{(1+x^2)^{5/2}}.
\end{aligned}
\tag{3.3}
$$

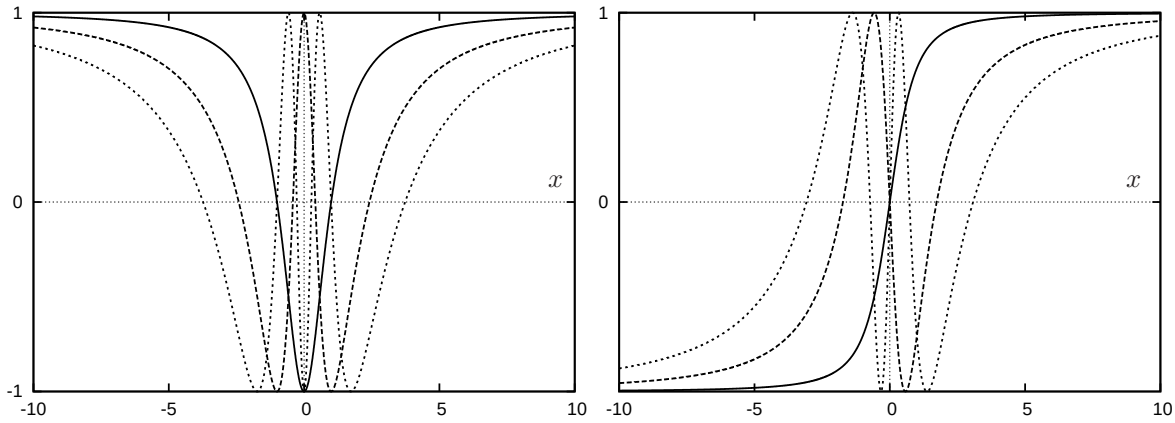

Figure 30: Left: even polynomials R_2 (solid), R_4 (dashed), R_6 (dotted). Right: odd polynomials R_1 (solid), R_3 (dashed), R_5 (dotted).

The following results represent an extension of some aspects of classical Chebyshev polynomials for the above mentioned global analogs (see e.g. Mason & Handscomb (2003) and Guo (1998) for some general ideas and proofs).

Theorem 3.1. *The set* $\{R_n\}$, $n \in \mathbb{N}$, *forms a complete orthogonal set in the space* $L^2_w(\mathbb{R})$ *with the weight function* $w(x) = \frac{1}{1+x^2}$.

Proof. To see this we utilize the fact that $\{T_n\}$ is complete and orthogonal in $L^2_v([-1,1])$, $v(t) = \frac{1}{\sqrt{1-t^2}}$. By applying ψ from (3.2) to the definition of completeness (i.e. the $span\{T_i\}$ is dense in L^2_v), one has for an arbitrary function $f \in L^2_v([-1,1])$ that

$$
\exists a_i \in \mathbb{R} : \|f - \sum_{i=0}^{\infty} a_i T_i\|_v^2 = \int_{-1}^{1} |f(t) - \sum_{i=0}^{\infty} a_i T_i(t)|^2 v(t) dt = 0. \tag{3.4}
$$

A straight forward variable transform $t = \psi(x)$, with $dt = (1+x^2)^{-3/2}dx$, shows that $\{R_n\}$ combined with the weight $w(x) = \frac{1}{1+x^2}$ satisfies (3.4) for (arbitrary) functions $f(t) = f(\psi(x)) =: g(x)$, $x \in \mathbb{R}$, such that $g \in L^2_w(\mathbb{R})$.

Performing the same steps for the orthogonality relation one obtains

$$\langle R_n, R_m\rangle_w := \int_{\mathbb{R}} R_n(x)R_m(x)w(x)dx = \begin{cases} 0 & \text{if } n \neq m \\ \|R_n\|_w^2 & \text{otherwise,} \end{cases} \tag{3.5}$$

with $\|R_0\|_w^2 = \pi$ and $\|R_n\|_w^2 = \pi/2$, $\forall n > 0$. □

Lemma 3.2. *The following holds for all R_n defined by* (3.1).

(i) Recurrence relation

$$R_{n+1} = 2\psi R_n - R_{n-1}$$

(ii) $R_n \in C^\infty(\mathbb{R})$, $R_n(x) \in [-1, 1]$, $\forall x \in \mathbb{R}$, $n \in \mathbb{N}$

(iii) Asymptotic behavior

$$R_n(x) \sim (\pm 1)^n\big(1 - \frac{n^2}{2x^2}\big) \quad \text{as} \quad x \to \pm\infty \tag{3.6}$$

(iv) R_n *is an even or odd function if* n *is even or odd, respectively.*

(v) Product of polynomials

$$R_n R_m = \tfrac{1}{2}(R_{m+n} + R_{|m-n|})$$

(vi) Generating sum in powers of ψ

$$R_n = \sum_{k=0}^{\lfloor n/2\rfloor} c_{k,n}\psi^{n-2k}$$

where $c_{k,n} = (-1)^k\, 2^{n-2k-1}\frac{n}{n-k}\binom{n-k}{k}$ *and* $\lfloor\cdot\rfloor$ *takes the integer part*

(vii) Generating formula

$$\begin{aligned} R_n(x) &= \tfrac{1}{2}\left[\left(\psi(x) + \sqrt{\psi(x)^2 - 1}\right)^n + \left(\psi(x) - \sqrt{\psi(x)^2 - 1}\right)^n\right] = \\ &= \sum_{k \ even}^{n} \binom{n}{k}\psi(x)^{n-k}(\psi(x)^2 - 1)^{k/2} \end{aligned} \tag{3.7}$$

(viii) Derivatives of R_n

$$\frac{dR_n}{dx} = 2n\frac{d\psi}{dx}\sum_{n-i \ odd}^{n-1}{}' R_i,$$

where the dash denotes that the kth term is halved if $k = n/2$ *and* n *is even*

(ix) Zeros x_i *of* R_n *(*$n \geq 1$*)*

$$x_i = \psi^{-1}\big(\cos\left(\frac{(2i-1)\pi}{2n}\right)\big) \qquad i = 1, \ldots, n$$

with (not sharp) bounds $x_i \in [-\frac{2}{\pi}n, \frac{2}{\pi}n] \subset [-\frac{2}{3}n, \frac{2}{3}n]$.

Proof. Most assertions in this lemma can be seen by a straight forward substitution of the mappings ψ and ϕ for the original variables in the proofs given in e.g. Mason & Handscomb (2003). So in what follows only arguments for the non-trivial extensions will be presented.

In (ii) the smoothness and the bounds ± 1 follow directly from the definition via the cosine function in (3.1) as a combination of smooth functions (ϕ is smooth by induction).

(iii) can be shown by setting $x = \pm 1/y$ in ψ (such that $x \to \pm\infty$ as $y \to 0^+$) and expanding

$$R_n(y) = \cos\left(n \arccos\left(\pm \frac{1}{\sqrt{1+y^2}}\right)\right)$$

into a Taylor series around $y = 0$, i.e.

$$\cos\left(n \arccos\left(\pm \frac{1}{\sqrt{1+y^2}}\right)\right) = (\pm 1)^n - (\pm 1)^n \frac{n^2}{2} y^2 + \mathrm{O}(y^4) \quad \text{as} \quad y \to 0^+$$

(cf. Boyd (2001a), where $(\pm 1)^n$ is missing in the second term).

The expression given in (vii) stems from considering the polynomials in the complex plane. Hence it cannot be used in the current form to (numerically) evaluate real-valued polynomials ($\sqrt{\psi(x)^2 - 1} \notin \mathbb{R}$). By saying $a := \psi(x)$ and $b := \sqrt{\psi(x)^2 - 1}$ we obtain (from the binomial formula)

$$R_n = \tfrac{1}{2}((a+b)^n + (a-b)^n) = \sum_{k \ even}^{n} \binom{n}{k} \psi(x)^{n-k} (\psi(x)^2 - 1)^{k/2},$$

where the odd terms can be cancelled out, which leaves only real addends.

The zeros in (ix) are, of course, just the mapped zeros of T_n. To prove the estimate for the bounds, it is sufficient to consider $i = 1$, since a simple evaluation shows $x_n < x_n - 1 < \cdots < x_1$ and $x_n = -x_1$. With $\psi^{-1}(x) = x/\sqrt{1-x^2}$ we have

$$x_1(n) = \frac{\cos(\frac{\pi}{2n})}{\sqrt{1 - \cos(\frac{\pi}{2n})^2}} = \frac{\cos(\frac{\pi}{2n})}{\sin(\frac{\pi}{2n})} = \frac{1}{\tan(\frac{\pi}{2n})}. \tag{3.8}$$

For $n = 1, 2, 3, \ldots$ the argument of the tangent function decreases (to zero) and hence setting $y := \frac{\pi}{2n}$, $y \in [0, \pi/2]$, an expansion of the last term in (3.8) around $y = 0$ should yield the desired estimate. However, a Taylor series of $\frac{1}{\tan(y)}$ around 0 cannot be gained in the usual way (due to the singularity). But one can easily obtain (using L'Hospital's rule for the coefficients) an expression for

$$\frac{1}{\tan(y)} - \frac{1}{y} = \frac{y - \tan(y)}{y \tan(y)} = -\frac{y}{3} + \mathrm{O}(y^3) \quad \text{as} \quad y \to 0.$$

By adding $1/y$ to the right-hand side above and combining this with (3.8) we arrive at

$$x_1(n) = \frac{1}{\tan(\frac{\pi}{2n})} = \frac{2n}{\pi} - \frac{\pi}{6n} + \mathrm{O}(n^{-3}) < \frac{2n}{\pi} < \frac{2}{3}n$$

(if $n > 1$), where the bound becomes sharper as n increases. □

Remark 3.1. One can infer from Lemma 3.2(iii) that the R_n are (in general) *not* integrable over $\mathbb{R}$. On the other hand, using integration by parts and the fact that all derivatives of all R_n vanish at infinity, it can be easily seen that

$$\int_{\mathbb{R}} \tfrac{d^m}{dx^m} R_n(x)dx = \begin{cases} 2 & m = 1 \text{ or } n \text{ odd} \\ 0 & \text{otherwise.} \end{cases}$$

Remark 3.2. There are several ways to express derivatives of Chebyshev polynomials. Considering the definition (3.1) it is obvious that all derivatives can be written in closed form by just differentiating the cosine combined with the mapping ψ. Lemma 3.2(viii) shows that one can reduce derivatives to sums of the original polynomials. Another way is to use Chebyshev polynomials of the second kind, or more generally *Gegenbauer* polynomials (see e.g. Boyd (2001) and Mason & Handscomb (2003)), to describe derivatives in a more or less compact formula. For the sake of completeness it shall be noted that the generating function given in (3.7) can also be differentiated to arbitrary order, which does not necessarily yield a practicable description, especially for numerical usage. In view of such characteristics, the best form to be used depends strongly on the problem to be solved. In case of mainly numerical evaluations closed forms are to be preferred, due to possible (round-off) error accumulations in the sums. Trefethen (2000) provides a mathematically thorough and applicable treatment of various aspects of evaluating derivatives of polynomials (making use of so called *differentiation matrices*).

Remark 3.3. As mentioned in Boyd (2001a) indefinite or definite integrals (over a compact interval, cf. Remark 3.1) of individual R_n's can be calculated in a straight forward manner. Therein a recurrence relation was also given to obtain integrals of R_n's for arbitrary high degrees. Interestingly, a general antiderivative in closed form for all degrees cannot be found directly. This is in strong contrast to the classical polynomials where it is easy to see that

$$\int T_n(x)dx = \begin{cases} T_1(x) & n = 0 \\ \frac{1}{4}T_2(x) & n = 1 \\ \frac{1}{2}\left(\frac{T_{n+1}(x)}{n+1} - \frac{T_{n-1}(x)}{n-1}\right) & n > 1, \end{cases} \tag{3.9}$$

which can be most easily shown via the cosine definition (see e.g. Mason & Handscomb (2003)).

Remark 3.4. Obviously, by applying a coordinate transform, it is only possible to utilize (3.9) to gain an equivalent general description for mapped polynomials, if the derivative of the transform is independent of the integration variable. Thus, it cannot be done for the mapping ψ in (3.2). Alternatively, integrals of Chebyshev polynomials can be given by considering the generating sum or the generating function in Lemma 3.2(vi),(vii). In the classical case, both can be easily integrated and algebraically modified to be equal to (3.9).

Apart from the recurrence relation stated in Boyd (2001a), one can calculate indefinite integrals of rational Chebyshev polynomials for arbitrary degrees (using a computer algebra system such as *Mathematica*) to be

$$\int R_n(x)dx = \sum_{k=0}^{\lfloor n/2 \rfloor} c_{k,n} \frac{x^{1+n-2k}}{1+n-2k} \, {}_2\mathrm{F}_1(\tfrac{n}{2}-k, \tfrac{n+1}{2}-k, \tfrac{n+3}{2}-k; -x^2),$$

where the coefficients $c_{k,n}$ are the same as in Lemma 3.2(vi) and ${}_2\mathrm{F}_1$ is the *Gauss hypergeometric function*. This is a very general description, which might not be practicable for numerical schemes.

When applying affine transforms of the form $y = ax + b$, $x \in [-1, 1]$, such that $y \in [b-a, b+a]$, the polynomials $T_n^* = T_n^*(y)$ are then defined on $[b-a, b+a]$. As mentioned in Mason & Handscomb (2003) such polynomials are not very useful in general, except when one needs to approximate a function only locally around some special point where a Taylor series representation is not the best choice.

Considering (3.9) and the fact that $dy/dx = a$ one can easily show

Lemma 3.3. *Given $y \in [b-a, b+a]$ and $y = ax + b$ with $a, b \in \mathbb{R}$, such that $T_n^*(y) := T_n(\frac{y-b}{a}) = T_n(x)$, then the following holds $\forall n \in \mathbb{N}$*

$$\int T_n^*(y)dy = \begin{cases} T_1^*(y) & n = 0 \\ a\frac{1}{4}T_2^*(y) & n = 1 \\ a\frac{1}{2}\left(\frac{T_{n+1}^*(y)}{n+1} - \frac{T_{n-1}^*(y)}{n-1}\right) & n > 1 \end{cases} \quad \textit{and hence}$$

$$\int_{b-a}^{b+a} T_n^*(y)dy = a\frac{1+(-1)^n}{1-n^2} = \begin{cases} \frac{2a}{1-n^2} & n \textit{ even} \\ 0 & n \textit{ odd.} \end{cases}$$

The second equation above follows directly from the first one, since the polynomials at the integration bounds evaluate to ± 1.

Remark 3.5. In addition to the non-integrability of all R_n over $\mathbb{R}$, the fact that there is no general rule for calculating closed form descriptions of indefinite integrals of R_n for arbitrary degrees presents a non-negligible disadvantage when dealing with algorithms or numerical schemes to approximate integrals of functions, solve differential equations with a Galerkin method or solve integral equations (on unbounded domains).

The following simple example shall illustrate this problem.

Example 3.1. Consider the integral

$$\int_{\mathbb{R}} \left[\tfrac{1}{2}R_0(x) - \tfrac{1}{2}R_2(x)\right]dx = \tfrac{1}{2}\left[\int_{\mathbb{R}} R_0 dx - \int_{\mathbb{R}} R_2 dx\right],$$

a situation possibly arising in the treatment of integral equations or Galerkin schemes (see Section 3.3). As shown above both integrals on the right-hand side do not exist. Although, by being more precise, calculating these integrals as a limit (by definition)

$$\tfrac{1}{2}\Big[\int_{\mathbb{R}} R_0 dx - \int_{\mathbb{R}} R_2 dx\Big] = \lim_{a\to\infty} \tfrac{1}{2}\Big[\int_{-a}^{a} R_0 dx - \int_{-a}^{a} R_2 dx\Big],$$

reveals that one first has to find the antiderivatives of R_0 and R_2, evaluate them at $\pm a$ and then take the limit. Since

$$\int R_0 dx = x \quad \text{and} \quad \int R_2 dx = x - 2\arctan(x)$$

one can immediately see, that the linear growth at infinity will eventually be canceled out by the sum, rendering the limit to exist. In fact, what we have is

$$\int_{\mathbb{R}} \underbrace{\big[\tfrac{1}{2}R_0(x) - \tfrac{1}{2}R_2(x)\big]}_{\frac{1}{1+x^2}} dx = \int_{\mathbb{R}} \frac{1}{1+x^2} dx = \pi.$$

All this is even more striking when products of polynomials occur (e.g. in Galerkin schemes), for example

$$\int_{\mathbb{R}} \big[\tfrac{1}{2}R_0(x) - \tfrac{1}{2}R_2(x)\big] R_n(x) dx = \tfrac{1}{2}\Big[\int_{\mathbb{R}} R_0 R_n dx - \int_{\mathbb{R}} R_2 R_n dx\Big].$$

Again, a numerical algorithm would be terminated at this point and even analytically one might not be able to find antiderivatives of such products in order to calculate the limits (as done above). Lemma 3.2(v) provides at least the opportunity to return to integrals of single polynomials, but most interestingly, applying trigonometric identities to the definition of R_n via the cosine function and due to the fact that $1/(1+x^2) = \frac{d}{dx}\arctan(x)$ it is straight forward to show

$$\int_{\mathbb{R}} \big[\tfrac{1}{2}R_0(x) - \tfrac{1}{2}R_2(x)\big] R_n(x) dx = 0 \quad \forall n > 0.$$

By evaluation of the term in square brackets this can also be viewed as $\langle R_n, 1\rangle_w$, which has to be zero by definition.

Remark 3.6. By substituting $x \to x/L$ in the mappings (3.2) (see Boyd (1982) and Boyd (2001) and references therein) one can introduce a stretching parameter L to modify the polynomials in order to better adapt them to certain problems or to reduce approximation errors (see Section 3.2). With $\psi(x) = x/\sqrt{L^2+x^2}$, the weight now reads $w(x) = L/(L^2+x^2)$ and the zeros are shifted significantly, i.e. $x_i \to Lx_i$ and so is the asymptotic behavior, i.e.

$$R_n(x;L) \sim (\pm 1)^n \big(1 - L^2 \frac{n^2}{2x^2}\big) \quad \text{as} \quad x \to \pm\infty$$

(cf. (3.6) and Figure 31). Especially when using these zeros as points of evaluation (e.g. collocation points, see Section 3.3) there is a non-negligible change in the approximation error (see Example 3.6 in Section 3.2). Although $L = 1$ might not be the best choice for optimal convergence (cf. (3.12)), it can provide possibilities to algebraically simplify certain integrals (with special kernels) and derivatives of R_n. Additionally, most of the modifications in the proofs given in Section 3.2 are algebraically easier to perform and furthermore, the conditions for convergence per se given therein must not depend on the choice of L. In actual numerical calculations, especially with the aim to minimize the number of polynomials, including L as a variable can be of advantage, see Example 3.6.

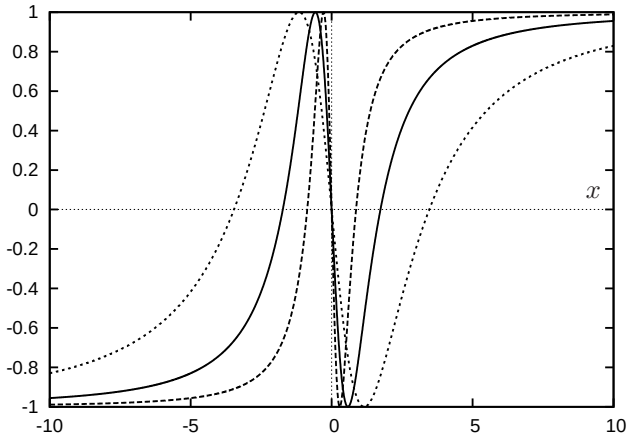

Figure 31: The polynomial R_3 taking $L = 1$ (solid), $L = 1/2$ (dashed) and $L = 2$ (dotted).

Remark 3.7. An extension of classical univariate polynomials T_i, such that they are defined on the hypercube $[-1, 1]^n$, has been done successfully by using a tensor product description, which is essentially just the pointwise product of such polynomials. This can be applied to the rational case without any changes, e.g.

$$R_k(x_1, x_2) = R_{ij}(x_1, x_2) := R_i(x_1)R_j(x_2), \qquad (x_1, x_2) \in \mathbb{R}^2,$$

where k runs through the rows of the array built by the indices i, j (cf. Mason & Handscomb (2003) and references therein). There is no unique way to obtain multivariate polynomials, but from the point of approximation theory (see Section 3.2), the tensor product is the most appropriate one, since it originally stems from formally expanding functions in several variables into a polynomial series, i.e.

$$f(x_1, x_2) = \sum_i a_i(x_2)p_i(x_1), \quad \text{with} \quad a_i(x_2) = \sum_j b_{ij}p_j(x_2) \;\Rightarrow$$
$$f(x_1, x_2) = \sum_{i,j} b_{ij} \underbrace{p_i(x_1)p_j(x_2)}_{p_{ij}(x_1,x_2)}.$$

Using the notion of *multi-indices* $k = (k_1, \ldots, k_n)$, with $|k| = \max_i |k_i|$ and $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, one possible definition (cf. Guo (1998)) then is

$$R_k(x) := \prod_{i=1}^{n} R_{k_i}(x_i). \tag{3.10}$$

Again, cf. (3.4) and (3.5), $\{R_k\}$ forms a complete orthogonal set in $L^2_u(\mathbb{R}^n)$, with the weight $u(x) = \prod_k w(x_k)$ (w given as in Theorem 3.1), and the orthogonality relation

$$\begin{aligned} \langle R_k, R_l \rangle_u &:= \int_{\mathbb{R}^n} R_k(x) R_l(x) u(x) dx = \int_{\mathbb{R}^n} \prod_{i=1}^{n} R_{k_i}(x_i) \prod_{i=1}^{n} R_{l_i}(x_i) \prod_{i=1}^{n} w(x_i) dx = \\ &= \prod_{i=1}^{n} \int_{\mathbb{R}} R_{k_i}(x_i) R_{l_i}(x_i) w(x_i) dx_i = \prod_{i=1}^{n} \langle R_{k_i}, R_{l_i} \rangle_w = \begin{cases} 0 & \text{if } k \neq l \\ \|R_k\|_u^2 & \text{otherwise} \end{cases}, \end{aligned} \tag{3.11}$$

where $k \neq l \Leftrightarrow \exists i : \; k_i \neq l_i, \; i = 1, \ldots, n$ and $\|R_k\|_u^2 := \prod_{i=1}^{n} \|R_{k_i}\|_w^2$, with the one-dimensional inner product and norm given as in (3.5).

Caveat: In the case of multivariate polynomials defined in (3.10) it is possible to write the inner product and hence the norm as the product of the one-dimensional equivalents. In general, the fist equality in (3.11) remains as the definition of the inner product in the space $L^2_u(\mathbb{R}^n)$ with the naturally induced norm $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$. For an illustration in two dimensions see Example 3.2.

We will not derive any other properties of multivariate rational Chebyshev polynomials (cf. Lemma 3.2), as this is a purely algebraic exercise and will not shed any (more) light on the structure of the set $\{R_k\}$.

Remark 3.8. For the sake of completeness it shall be mentioned that the Chebyshev polynomials (among the Legendre and Gegenbauer polynomials) are special cases of the *Jacobi polynomials*. It is hence obvious that these can also be mapped to provide basis functions on the whole real axis. Recently, Narayan & Hesthaven (2011) generalized such considerations including Wiener's basis functions.

In the above only a few properties of orthogonal polynomials were treated regarding their modification for problems on $\mathbb{R}$. A more detailed and classical treatment of orthogonal polynomials has been given by Szegö (1939) and many of the results presented therein can be transformed to the real line in the same way as we have shown in this section.

3.2 Approximation Theory in $\mathbb{R}^n$

This section provides a general treatment of approximating functions using orthogonal polynomials in $\mathbb{R}^n$. This forms the theoretical background for numerically solving differential and integral equations using spectral methods on unbounded domains.

As it has been done in Section 3.1, we use existing definitions and results for orthogonal polynomials provided, for example, in Gottlieb & Orszag (1977), Guo (1998), Mason & Handscomb (2003) and Hesthaven et al. (2007) and extend them to the rational, multivariate case.

Three types of convergence shall be addressed in the following, namely

$$\left.\begin{array}{ll} \textit{convergence in some function space} & \\ & \|f - f_N\|_H \to 0 \\ \textit{pointwise convergence} & \\ & |f(x) - f_N(x)| \to 0 \quad \forall x \\ \textit{uniform convergence} & \\ & \sup_x |f(x) - f_N(x)| \to 0 \end{array}\right\} \quad \text{as } N \to \infty, \tag{3.12}$$

where H denotes some Lebesgue or Sobolev space and the *uniform convergence* can also be seen as *convergence in the L^∞ norm* by taking the essential supremum. Under certain conditions (cf. Lemma 3.14), but not in general, one can infer that

$$L^\infty \textit{ convergence } \Rightarrow \textit{ pointwise convergence } \Rightarrow \textit{ } H \textit{ convergence.}$$

From the completeness of the set $\{R_n\}$ in $L^2_w(\mathbb{R})$ (cf. Theorem 3.1) and the norm induced by the inner product in (3.11) one can readily deduce the following convergence result.

Lemma 3.4. *Given a function $f \in L^2_u(\mathbb{R}^n)$, with $x = (x_1, \ldots, x_n)$ and $u(x) = \prod_{i=1}^{n} 1/(1+x_i^2)$, there exist $a_k \in \mathbb{R}$ and a polynomial $p_N(x) = \sum_{|k|\leq N} a_k R_k(x)$, such that*

$$\|f - p_N\|_u \to 0 \quad \textit{as} \quad N \to \infty.$$

Remark 3.9. The lemma above represents the weakest form of convergence in the sense that it does not say anything about the rate of convergence or how the coefficients a_k can be determined. Also, no additional assumptions on the function f have been made, such as boundedness, differentiability etc.

To measure the quality of an approximation f^* in some normed function space, a typical requirement would be $\|f - f^*\| \leq \epsilon$ (for a given ϵ) and consequently a *best approximation* f^*_B can be defined as $\|f - f^*_B\| \leq \|f - f^*\|$ for all other approximations f^*, see e.g. Mason & Handscomb (2003).

By using the fact that any weighted L^2 space possesses an inner product we can paraphrase a well known result in terms of multivariate polynomials.

Theorem 3.5. *For any $f \in L^2_u(\mathbb{R}^n)$ there exists a unique polynomial p^B_N of maximal degree N, such that p^B_N is the L^2_u best approximation and is given by the (necessary and sufficient)*

condition

$$\langle f - p_N^B, p_N \rangle_u = 0, \qquad \forall \text{ polynomials } p_N \in L_u^2. \tag{3.13}$$

Proof. see e.g. Mason & Handscomb (2003). Therein, apart from some general arguments, only properties of inner products and their induced norms are used and hence the proof does not need to be modified to hold for weighted Hilbert spaces in $\mathbb{R}^n$. □

From the existence of such a polynomial it can be concluded that there must also exist corresponding coefficients a_k, as mentioned in Lemma 3.4. In fact, by exploiting the orthogonality, we have the following

Theorem 3.6. *Let $\{p_k\}$ be a set of orthogonal polynomials on $\mathbb{R}^n$, $f \in L_u^2(\mathbb{R}^n)$ and p_N^B the best approximation found in Theorem 3.5. Then p_N^B can be written as*

$$p_N^B = \sum_{|k| \le N} a_k \, p_k \,, \qquad \text{with} \qquad a_k := \frac{\langle f, p_k \rangle_u}{\|p_k\|_u^2} \,. \tag{3.14}$$

Proof. see e.g. Mason & Handscomb (2003) for the one-dimensional classical case, which can be easily adapted to the multivariate result here. □

Due to the uniqueness of the best approximation polynomial, the coefficients given in (3.14) are thus uniquely defined and, more importantly, do *not* depend on N.
Obviously, the univariate as well as the multivariate rational Chebyshev polynomials defined in (3.1) and (3.10) combined with the inner products (3.5) and (3.11) satisfy the requirements in Theorems 3.5 and 3.6.

Remark 3.10. By formally expanding a function f into a Chebyshev series (in $\mathbb{R}^n$) and taking the (weighted) inner product with an arbitrary Chebyshev polynomial and exploiting the orthogonality, i.e.

$$f = \sum_{i=0}^{\infty} a_i R_i \quad \rightsquigarrow \quad \langle f, R_j \rangle_u = \sum_{i=0}^{\infty} a_i \underbrace{\langle R_i, R_j \rangle_u}_{\neq 0 \Leftrightarrow i = j} \quad \forall j,$$

it is straight forward to see that this yields the same formula for the coefficients as in (3.14). To be more precise, such modifications should be considered in the limit $N \to \infty$ for all the sums over $i = 1, \ldots, N$. Thus, suitable convergence has to be assumed to interchange the limit (and the summation) with the integral from the inner product (cf. the dominated convergence theorem and Lemma 3.12 and Theorem 3.13).

It is nevertheless possible to define an *orthogonal Chebyshev expansion* of a (suitably chosen) function as an infinite series, where every partial sum is the L^2 best approximation. It additionally follows that for a finite (i.e. truncated) Chebyshev series p_N (with the coefficients given via (3.14)), the error $e_N := f - p_N$ not only tends to zero in the L^2 sense (trivial from Lemma 3.4), but is also *minimal* for every N.

3.2.1 Orthogonal Projections

Some types of finite-dimensional approximations can also be considered in terms of *projection methods*, where it is often easier to prove certain consistency, stability and convergence results as well as error estimates. In view of such methods a truncated Chebyshev series expansion operator shall be generally defined as

$$\mathcal{P}_N : H \to V_N, \qquad \mathcal{P}_N f := \sum_{|i|\leq N} a_i R_i, \tag{3.15}$$

with H being some Banach space and V_N the span of all polynomials R_i.
The next result shows under which assumptions such an operator can be seen as an orthogonal projection

Lemma 3.7. *The approximation operator $\mathcal{P}_N : L^2_u(\mathbb{R}^n) \to V_N$, given as*

$$\mathcal{P}_N f = \sum_{|k|\leq N} a_k \, R_k, \tag{3.16}$$

with a_k as in (3.14)*, defines a bounded orthogonal projection.*

Proof. One needs to verify the projection properties. Using the definition of the coefficients and the linearity of the inner product, i.e.

$$\frac{\langle \lambda f + \gamma g, R_i\rangle_u}{\|R_i\|_u^2} = \lambda \frac{\langle f, R_i\rangle_u}{\|R_i\|_u^2} + \gamma \frac{\langle g, R_i\rangle_u}{\|R_i\|_u^2}$$

one can deduce the linearity of $\mathcal{P}_N$.
As for the idempotence ($\mathcal{P}^2 = \mathcal{P}$) say $g := \mathcal{P}_N f$ with coefficients a_i, then

$$\mathcal{P}_N g = \sum_i b_i R_i, \quad b_i = \frac{\langle g, R_i\rangle_u}{\|R_i\|_u^2} = \frac{\langle \sum_j a_j R_j, R_i\rangle_u}{\|R_i\|_u^2} = \sum_j a_j \frac{\langle R_j, R_i\rangle_u}{\|R_i\|_u^2} = a_i \quad \Rightarrow \quad \mathcal{P}_N g = g.$$

Finally, for the boundedness, we use the fact that $\mathcal{P}_N f$ is the best approximation in L^2_u (cf. Theorem 3.5) and deduce $\forall f \in L^2_u$

$$\|f\|_u^2 = \|f - \mathcal{P}_N f + \mathcal{P}_N f\|_u^2 = \|f - \mathcal{P}_N f\|_u^2 + \underbrace{2Re\langle f - \mathcal{P}_N f, \mathcal{P}_N f\rangle_u}_{=0} + \|\mathcal{P}_N f\|_u^2 \geq \|\mathcal{P}_N f\|_u^2$$

and thus $\|\mathcal{P}_N\|_u \leq 1$. On the other hand, from $\mathcal{P}_N^2 = \mathcal{P}_N$ we have $\|\mathcal{P}_N\|_u \leq \|\mathcal{P}_N\|_u^2$, where $\|\mathcal{P}_N\|_u \geq 1$ and consequently $\|\mathcal{P}_N\|_u = 1$. □

Remark 3.11. Guo (1998) defines an orthogonal projection via (3.13) and deduces Theorem 3.6 and Lemma 3.7 from that. Furthermore, extensions to general weighted Sobolev spaces (containing an inner product) are shown therein as well. By reviewing the proof given above, one can easily see that in fact every argument holds for all Hilbert spaces, concluding that *every H best approximation given via orthogonal polynomials is an orthogonal projection from*

H into the span of the polynomials (see Mason & Handscomb (2003) and Hackbusch (1995) for some further and general aspects of projection operators).

The next result, which is often utilized when proving convergence rates, shows that if the projection converges in L^2 the coefficients are *square summable* (also known as *Parseval's identity*, originally derived for Fourier series). Since this shall be done in the most general sense, where a precise understanding of sums over multi-indices is crucial. The following example will illustrate the two-dimensional case.

Example 3.2. (see definition in Remark 3.7)
Say $n = 2$, thus $k = (k_1, k_2)$ and $a_k = a_{(k_1,k_2)} =: a_{k_1k_2}$, which can be also seen as a matrix of coefficients. Now the sum over $|k| \leq N$ or $|k| = 0, 1, 2, \ldots, N$ means summing up all the terms for which

$$\begin{aligned}
&|k| = 0 \quad \rightsquigarrow k = (0,0)\\
&|k| = 1 \quad \rightsquigarrow k = (0,1),\ (1,0),\ (1,1)\\
&|k| = 2 \quad \rightsquigarrow k = (0,2),\ (1,2),\ (2,0),\ (2,1),\ (2,2)\\
&\vdots
\end{aligned}$$

Obviously, all the pairs can be rearranged to yield the matrix mentioned above, where summation runs through every line for every column or vice versa. Ergo,

$$\sum_{|k|\leq N} p_k = \sum_{k_1=0}^{N}\sum_{k_2=0}^{N} p_{k_1k_2}.$$

Overall, the projection operator in $\mathbb{R}^2$ reads

$$(\mathcal{P}_N f)(x_1, x_2) = \sum_{k_1=0}^{N}\sum_{k_2=0}^{N} a_{k_1k_2} R_{k_1}(x_1) R_{k_2}(x_2),$$

with the coefficients

$$a_{k_1k_2} = \frac{\langle f, R_{k_1}R_{k_2}\rangle_u}{\|R_{k_1}\|_w^2 \|R_{k_2}\|_w^2} = \left(\frac{\pi}{2}\right)^2 c_{k_1}c_{k_2} \int_{\mathbb{R}^2} f(x_1, x_2) R_{k_1}(x_1) R_{k_2}(x_2) w(x_1) w(x_2) dx_1 dx_2,$$

where $c_{k_i} = 2$ if $k_i = 0$ and $c_{k_i} = 1$ otherwise (taking into account $\|R_0\|_w^2 = 2\|R_j\|_w^2 = \pi$, see (3.5), cf. definition in Guo (1998) for the coefficients in multi-dimensional Chebyshev transformations). It is now straight forward to see that only in the case of separable functions f the multi-dimensional inner product can be written as the product of its one-dimensional equivalents. One possible way to express multi-dimensional inner products in one-dimensional terms would be by successive application, i.e.

$$\langle f, R_k\rangle_u = \langle \ldots \langle \langle f, R_{k_1}\rangle_w, R_{k_2}\rangle_w \ \ldots, R_{k_n}\rangle_w.$$

The following result states the square summability of the coefficients given in the orthogonal projection.

Theorem 3.8. *Let $f \in L^2_u(\mathbb{R}^n)$ and $\mathcal{P}_N$ be the according orthogonal projection given by* (3.16)*. Then the coefficients satisfy the (modified) Parseval identity*

$$\sum_{|k|=0}^{\infty} c_k a_k^2 = (2/\pi)^n \|f\|_u^2, \tag{3.17}$$

with $c_k = \prod_{i=1}^{n} c_{k_i}$, where $c_{k_i} = 2$ for $k_i = 0$ and $c_{k_i} = 1$ otherwise.

Proof. Measuring the difference of f and its projection in L^2, we get

$$\begin{aligned}
\|f - \mathcal{P}_N f\|_u^2 &= \langle f - \mathcal{P}_N f, f - \mathcal{P}_N f\rangle_u = \langle f, f\rangle_u - 2\langle f, \mathcal{P}_N f\rangle_u + \langle \mathcal{P}_N f, \mathcal{P}_N f\rangle_u = \\
&= \|f\|_u^2 - 2\sum_{|k|\le N} a_k \langle f, R_k\rangle_u + \sum_{|l|,|k|\le N} a_l a_k \underbrace{\langle R_l, R_k\rangle_u}_{=0,\ l\neq k} = \\
&= \|f\|_u^2 - 2\sum_{|k|\le N} a_k^2 \|R_k\|_u^2 + \sum_{|k|\le N} a_k^2 \|R_k\|_u^2 = \\
&= \|f\|_u^2 - \sum_{|k|\le N} a_k^2 \|R_k\|_u^2.
\end{aligned}$$

All equalities above are independent of the dimension n and also hold for the classical polynomials (cf. Mason & Handscomb (2003)).
Lemma 3.4 shows that for $N \to \infty$ the left-hand side of the equation above tends to zero, hence

$$\sum_{|k|=0}^{\infty} a_k^2 \|R_k\|_u^2 = \|f\|_u^2$$

and from $\|R_k\|_u^2 = \prod_{i=1}^{n} \|R_{k_i}\|_w^2$, where $\|R_{k_i}\|_w^2 = c_{k_i}\frac{\pi}{2}$, with the c_{k_i} defined above we arrive at the desired result. □

Note that absolute summability cannot be inferred here. Therefore it is not possible to make any assertion about pointwise or uniform convergence. The only obvious conclusion is that the coefficients must form a null sequence, the necessary condition for convergence of any series.

In two dimensions, for example, the square summability can also be written as

$$\begin{aligned}
&\sum_{k=0}^{\infty}\sum_{l=0}^{\infty} c_{kl} a_{kl}^2 = a_{00}^2 + \sum_{l=0}^{\infty} a_{0l}^2 + \sum_{k=0}^{\infty} a_{k0}^2 + \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} a_{kl}^2 \\
&\Rightarrow \sum_{|k|=0}^{\infty} a_k^2 \le c\|f\|_u^2.
\end{aligned}$$

So far the L^2 convergence for approximations (projections) in terms of (multivariate rational) Chebyshev polynomials is very well established. If the given function is differentiable in some sense, convergence rates can be proved to depend on the approximation parameter N and the order of the derivatives.

Guo (1998), Wang & Guo (2002) and Hesthaven et al. (2007) provide results for the classical types of orthogonal polynomial systems (Fourier, Legendre, Hermite, Chebyshev) with some extensions to multidimensional and unbounded domains.

In the following we will show convergence rates for multivariate rational Chebyshev polynomials in terms of Sobolev type spaces and smooth functions. As usual, for the multi-index k and $x \in \mathbb{R}^n$ say $\partial_x^k := \partial_{x_1}^{k_1}\partial_{x_2}^{k_2}\dots\partial_{x_n}^{k_n}$, $|k|_s = \sum k_i$ and let the space $H_u^r(\mathbb{R}^n)$ be defined as the set

$$H_u^r(\mathbb{R}^n) = \left\{ f \,|\, \partial_x^k f \in L_u^2(\mathbb{R}^n), |k|_s \le r \right\}, \qquad \text{with} \quad \|f\|_r^2 = \sum_{|k|_s \le r} \|\partial_x^k f\|_u^2, \tag{3.18}$$

whereas $H_{u,A}^r$ shall be the set

$$H_{u,A}^r(\mathbb{R}^n) = \left\{ f \,|\, \|f\|_A < \infty \right\} \qquad \text{with} \quad \|f\|_A^2 := \sum_{|k|_s \le r} \Big\| \prod_{j=1}^{n} (1+x_j^2)^{\frac{r/n+k_j}{2}} \partial_{x_j}^{k_j} f \Big\|_u^2. \tag{3.19}$$

Remark 3.12. The different use of absolute values for multi-indices (cf. Remark 3.7) stems from the agreement that mixed partial derivatives are usually regarded as derivatives of order of the sum of its individual derivatives, e.g. the pair $(2,2)$ gives $\partial_x^2\partial_y^2$ ($|(2,2)|_s = 4$), a fourth order differential operator, whereas $|(2,2)| = 2$.

The space $H_{u,A}^r$ is chosen to work with the arguments in the proof for the convergence rate of the projection operator (Wang & Guo (2002) show the general idea in one dimension), cf. Theorem 3.11, where the subscript A will become clear. The following result adds some meaning to this function space.

Lemma 3.9. *The space $H_{u,A}^r$ defined in* (3.19) *is a subspace of H_u^r.*

Proof. One needs to show that for every $f \in H_{u,A}^r(\mathbb{R}^n)$, $\|f\|_r \le c\|f\|_A$ holds, which implies that if $\|f\|_A < \infty \;\Rightarrow\; \|f\|_r < \infty$.
By observing that every addend in both, (3.18) and (3.19) is positive and also the number of terms in both sums is equal, it is sufficient to prove that

$$\|\partial_x^k f\|_u^2 \le c \Big\| \prod_{j=1}^{n} (1+x_j^2)^{\frac{r/n+k_j}{2}} \partial_{x_j}^{k_j} f \Big\|_u^2 \quad \text{for fixed } r, \;\; \forall k. \tag{3.20}$$

Defining $g_r(x) := \prod_{j=1}^{n} (1+x_j^2)^{\frac{r/n+k_j}{2}}$ it follows that $g_r(x) \ge 1$ and unbounded as $|x| \to \infty$. With $f_k^*(x) := \partial_x^k f(x)$, (3.20) reads

$$\|f_k^*\|_u^2 \le c\|g_r f_k^*\|_u^2 \quad \forall k. \tag{3.21}$$

The definition of the L^2_u norm on the left-hand side in (3.21) yields

$$\int\limits_{\mathbb{R}^n} |f_k^*(x)|^2 u(x)dx = \int\limits_{\mathbb{R}^n\setminus S} |f_k^*(x)|^2 u(x)dx + \int\limits_{S} |f_k^*(x)|^2 u(x)dx, \tag{3.22}$$

where S denotes a bounded n-sphere with radius R. Switching to hyperspherical coordinates $x \to (\rho, \phi)$ with $\phi = (\phi_1, \ldots, \phi_{n-1})$, i.e. say $P_n : \mathbb{R}^n \to \mathbb{R}^n$, such that

$$P_2(\rho,\phi) := \begin{pmatrix} \rho\cos(\phi) \\ \rho\sin(\phi) \end{pmatrix} \quad \text{and} \quad P_n(\rho,\phi_1,\ldots,\phi_{n-1}) := \begin{pmatrix} P_{n-1}(\rho,\phi_1,\ldots,\phi_{n-2})\cos(\phi_{n-1}) \\ \rho\sin(\phi_{n-1}) \end{pmatrix},$$

with $\phi \in I := (-\pi,\pi)\times(-\pi/2,\pi/2)^{n-2}$

and the functional determinant $|J| = \rho^{n-1}\prod\limits_{k=2}^{n-1}\cos^{k-1}(\phi_k)$,

and by denoting the functions in these coordinates with an upper bar, (3.22) can be rewritten to be

$$\int\limits_{\mathbb{R}^n} |f_k^*(x)|^2 u(x)dx = \int\limits_{R}^{\infty}\int\limits_{I} |\bar{f}_k^*(\rho,\phi)|^2 \bar{u}(\rho,\phi)|J|\,d\phi\,d\rho + \int\limits_{0}^{R}\int\limits_{I} |\bar{f}_k^*(\rho,\phi)|^2 \bar{u}(\rho,\phi)|J|\,d\phi\,d\rho. \tag{3.23}$$

In S, which represents the second term on the right-hand side in (3.23), $\bar{g}_r = \bar{g}_r(\rho,\phi)$ is a uniformly bounded function with values between $1 \le \bar{g}_r \le const.$ Thus, one obtains

$$\int\limits_{0}^{R}\int\limits_{I} |\bar{f}_k^*(\rho,\phi)|^2 \bar{u}(\rho,\phi)|J|\,d\phi\,d\rho \le c\int\limits_{0}^{R}\int\limits_{I} |\bar{g}_r|^2|\bar{f}_k^*(\rho,\phi)|^2 \bar{u}(\rho,\phi)|J|\,d\phi\,d\rho. \tag{3.24}$$

Considering the far field behavior of the functions

$$\bar{g} \sim \rho^{c_g(\phi)}, \quad \bar{f}_k^* \sim \rho^{c_f(\phi)}, \quad \bar{u} \sim \rho^{c_u(\phi)} \quad \text{as } \rho \to \infty$$

yields for the first integral on the right-hand side of (3.23) (additionally changing the order of integration)

$$\int\limits_{I}\int\limits_{R}^{\infty} |\bar{f}_k^*(\rho,\phi)|^2 \bar{u}(\rho,\phi)|J|\,d\rho\,d\phi \sim \int\limits_{I}\int\limits_{R}^{\infty} \rho^{2c_f+c_u}|J|\,d\rho\,d\phi \le c\int\limits_{I}\int\limits_{R}^{\infty} \rho^{2(c_g+c_f)+c_u}|J|\,d\rho\,d\phi, \tag{3.25}$$

where the inequality holds due to $\bar{g}_r$ being unbounded as $\rho \to \infty$, which means $c_g(\phi) > 0$, $\forall\phi$ and consequently $\rho^{c_g} \ge 1$, for $\rho \ge R$.
In view of (3.23), adding the right-hand sides of (3.24) and (3.25) gives the desired estimate (3.20). □

Remark 3.13. This proof shows that the space $H^r_{u,A}$ collects those functions of H^r_u which satisfy a certain restriction on the decay behavior. Obviously, considering a bounded integra-

tion domain, e.g. the n-sphere in (3.22), if f is $H^r_u(S)$-integrable then f is $H^r_{u,A}(S)$-integrable and vice versa (by a continuity argument there exists a constant c^*, such that the equality in (3.24) holds). But when looking at the far field,

$$\int_I \int_R^\infty \rho^{2(c_g+c_f)+c_u+n-1}\, d\rho\, \cos^{n-2}(\phi_{n-1}) \cos^{n-3}(\phi_{n-2}) \dots \cos(\phi_2)\, d\phi,$$

the integral over ρ can only exist if $2(c_g + c_f) + c_u + n - 1 < -1\ \forall \phi$, or $c_f < (-c_u - n)/2 - c_g$, which is a stronger requirement than $c_f < (-c_u - n)/2$ (for $f \in H^r_u$), since $c_g > 0$.

In one dimension, see Wang & Guo (2002), the difference in the necessary decay can be given explicitly as $f \sim x^{(1-2r)/2}$ (to lie in $H^r_{u,A}(\mathbb{R})$) in contrast to $f \sim x^{1/2}$. Note that in the case of $H^r_u(\mathbb{R})$ the necessary far field condition is independent of r.

As mentioned in Gottlieb & Orszag (1977), the classical Chebyshev polynomial T_l is the eigenfunction to the according Sturm-Liouville operator $A := -\sqrt{1-x^2}\frac{d}{dx}[\sqrt{1-x^2}\frac{d}{dx}]$ with eigenvalues l^2 and $x \in [-1, 1]$. By applying the coordinate transform ψ defined in (3.2), one can derive the analogous operator on $\mathbb{R}$, such that the rational Chebyshev polynomial R_l becomes the eigenfunction to the eigenvalues l^2, i.e.

$$A_x R_l(x) := -(1 + x^2)\partial_x[(1 + x^2)\partial_x]R_l(x) = l^2 R_l(x). \tag{3.26}$$

One can now apply this operator successively m times to a function $f \in C^{2m}(\mathbb{R})$ obtaining

$$A^m_x f(x) = \sum_{k=1}^{2m} (1 + x^2)^{m+\frac{k}{2}} p_k(x) \partial^k_x f(x), \tag{3.27}$$

where p_k are uniformly bounded (rational) functions on $\mathbb{R}$, which can be seen by induction, as proposed in Wang & Guo (2002) (without showing details). In Appendix B the necessary arguments to prove this assertion are provided.

Rewriting the Sturm-Liouville problem (3.26) as $R_l = A_x R_l / l^2$ and substituting this into the definition of the multivariate rational Chebyshev polynomials (3.10) (now $k = (k_1, \dots, k_n)$) then yields

$$\prod_{i=1}^n k_i^2 R_k = \prod_{i=1}^n A_{x_i} R_{k_i}(x_i) = \big(\prod_{i=1}^n A_{x_i}\big) R_k(x) =: A_x R_k(x), \tag{3.28}$$

which can be viewed as a definition for multi-dimensional eigenvalue problems for the Sturm-Liouville operator A_x, $x \in \mathbb{R}^n$.

Remark 3.14. By noting that every component A_{x_i} acts only with respect to x_i one can rearrange the terms in $A_x = A_{x_1} A_{x_2} \dots A_{x_n}$, such that $A_x = (-1)^n \prod(1+x_i^2)\partial_x[\prod(1+x_i^2)\partial_x]$, which is now of a similar type as the original definition in one dimension stated above.

Caveat: In $\mathbb{R}$, $A_x R_0 \equiv 0$ since $R_0 \equiv 1$, which is the only eigenfunction to the eigenvalue 0. In the multivariate case this means that every R_k where at least one $k_i = 0$ is an eigenfunction to the eigenvalue 0, hence one obtains $2^n - 1$ according polynomials. Additionally, if A_x is applied to a function, independent of at least one of the x_i, it follows that $A_x f \equiv 0$.

With this remark and in view of (3.27) it is now straight forward to define the successive application of the multi-dimensional Sturm-Liouville operator to a function $f \in C^{2mn}(\mathbb{R}^n)$

$$A_x^m f(x) := A_{x_1}^m \dots A_{x_n}^m f(x) = \sum_{k_1=1}^{2m} \cdots \sum_{k_n=1}^{2m} \prod_{i=1}^{n} (1+x_i^2)^{m+\frac{k_i}{2}} p_{k_i}(x_i) \partial_{x_i}^{k_i} f(x). \tag{3.29}$$

Lemma 3.9 implies $H_{u,A}^r$ to be a subspace of L^2, thus proposing

Lemma 3.10. *Say $f \in H_{u,A}^r(\mathbb{R}^n)$ and $r = 2mn$, then there exists a positive constant c, such that*

$$\|A_x^m f\|_u \le c \|f\|_A$$

and hence A_x^m continuously maps $H_{u,A}^r$ to L_u^2.

Proof. Again, this was mentioned in Wang & Guo (2002) for the one-dimensional case without showing the details. Since the proof reveals why the space $H_{u,A}^r$ is defined by (3.19), where the subscript A refers to the Sturm-Liouville operator, it shall be given here in detail. Recalling *Cauchy's inequality* for real numbers, i.e. $|\sum_{i=0}^{n} a_i b_i|^2 \le \left(\sum_{i=0}^{n} |a_i|^2\right)\left(\sum_{i=0}^{n} |b_i|^2\right)$, one can write

$$\begin{aligned}
|A_x^m f(x)|^2 &= \Big| \sum_{k_1=1}^{2m} \cdots \sum_{k_n=1}^{2m} \prod_{i=1}^{n} (1+x_i^2)^{m+\frac{k_i}{2}} p_{k_i}(x_i) \partial_{x_i}^{k_i} f(x) \Big|^2 = \\
&= \Big| \sum_{k_1=1}^{2m} p_{k_1}(x_1) \underbrace{\sum_{k_2=1}^{2m} p_{k_2}(x_2) \cdots \sum_{k_n=1}^{2m} p_{k_n}(x_n) \prod_{i=1}^{n} (1+x_i^2)^{m+\frac{k_i}{2}} \partial_{x_i}^{k_i} f(x)}_{=:h_{k_1}} \Big|^2 \le \\
&\le \left(\sum_{k_1=1}^{2m} |p_{k_1}|^2 \right) \sum_{k_1=1}^{2m} |h_{k_1}|^2 \quad \text{plug in } h_{k_1} \text{ and apply the inequality again} \\
&\le \left(\sum_{k_1=1}^{2m} |p_{k_1}|^2 \right) \left(\sum_{k_2=1}^{2m} |p_{k_2}|^2 \right) \sum_{k_1=1}^{2m} \sum_{k_2=0}^{2m} |h_{k_2}|^2 \\
&\vdots \quad \text{with } h_{k_2} \text{ embracing the sums over } k_3 \text{ to } k_n \\
&\vdots \quad \text{repeating this procedure finally yields} \\
&\le \underbrace{\left(\sum_{k_1=1}^{2m} |p_{k_1}|^2 \right) \cdots \left(\sum_{k_n=1}^{2m} |p_{k_n}|^2 \right)}_{=:c} \sum_{k_1=1}^{2m} \cdots \sum_{k_n=1}^{2m} \Big| \prod_{i=1}^{n} (1+x_i^2)^{m+\frac{k_i}{2}} \partial_{x_i}^{k_i} f(x) \Big|^2,
\end{aligned}$$

where the constant c exists due to the boundedness of every p_{k_i}, see Appendix B, and the inequalities hold $\forall x \in \mathbb{R}^n$. Additionally both sides are non-negative functions and hence integration over $\mathbb{R}^n$ (including the positive weight u) will not change the inequality,

$$\int_{\mathbb{R}^n} |A_x^m f(x)|^2 u(x)dx \leq c \int_{\mathbb{R}^n} \sum_{k_1=1}^{2m} \cdots \sum_{k_n=1}^{2m} \Big| \prod_{i=1}^{n} (1+x_i^2)^{m+\frac{k_i}{2}} \partial_{x_i}^{k_i} f(x) \Big|^2 u(x)dx. \tag{3.30}$$

Interchanging summation and integration on the right-hand side and starting the summation from zero then gives

$$\begin{aligned} \|A_x^m f\|_u^2 \leq c \sum_{k_1=0}^{2m} \dots \sum_{k_n=0}^{2m} \int_{\mathbb{R}^n} \Big| \prod_{i=1}^{n} (1+x_i^2)^{m+\frac{k_i}{2}} \partial_{x_i}^{k_i} f(x) \Big|^2 u(x)dx \leq \\ \leq c \sum_{|k|_s \leq r} \Big\| \prod_{j=1}^{n} (1+x_j^2)^{\frac{r/n+k_j}{2}} \partial_{x_j}^{k_j} f \Big\|_u^2 = c\|f\|_A^2, \end{aligned} \tag{3.31}$$

where the last inequality comes from setting $r = 2mn = \max_k |k|_s$ (to include the highest derivative), gaining additional terms (not appearing in the sums in (3.30)), which, for being non-negative, do not change the inequality. Appendix B shows the above result for $r = (2m+1)n$, such that it holds for arbitrary $r \in \mathbb{N}$. □

Next we show a convergence rate for multivariate orthogonal projections in $\mathbb{R}^n$ (in the same way as Wang & Guo (2002) showed the one dimensional version).

Theorem 3.11. *Let $f \in H_{u,A}^r(\mathbb{R}^n)$, then there exists a positive constant c, such that*

$$\|f - \mathcal{P}_N f\|_u \leq cN^{-r/n} \|f\|_A. \tag{3.32}$$

Proof. The strategy of this proof is in complete accordance to Wang & Guo (2002). In general, the application of Sturm-Liouville type operators to show convergence rates for projection operators is very common, see Hesthaven et al. (2007) for the case of ultraspherical polynomials as well as Guo (1998) for other types of orthogonal polynomials.
First we show, how to modify the iteration process to include A_x^m in the description of the expansion coefficients. Let $\mathcal{P}_N f = \sum a_k R_k$ as usual. Then,

$$\begin{aligned} \|R_k\|_u^2 a_k &= \int_{\mathbb{R}^n} f(x) \prod_{i=1}^{n} R_{k_i}(x_i) \tfrac{1}{1+x_i^2} \, dx \overset{(3.28)}{=} \\ &= \prod \tfrac{1}{k_i^2} \int_{\mathbb{R}^n} f(x) \prod A_{x_i} R_{k_i}(x_i) \tfrac{1}{1+x_i^2} \, dx = \\ &= \prod \tfrac{1}{k_i^2} \int_{\mathbb{R}} A_{x_n} R_{k_n}(x_n) \tfrac{1}{1+x_n^2} \int_{\mathbb{R}} \dots \int_{\mathbb{R}} f(x) A_{x_1} R_{k_1}(x_1) \tfrac{1}{1+x_1^2} \, dx = \end{aligned}$$

... substitution of (3.26) cancels the weights

... consecutive integration by parts yields

$\ldots$ (boundary terms vanish due to $f \in H^r_{u,A}$)

$$= \prod \tfrac{1}{k_i^2} \int\limits_{\mathbb{R}} \partial_{x_n} R_{k_n}(x_n) \int\limits_{\mathbb{R}} \ldots \int\limits_{\mathbb{R}} \prod (1+x_i^2) \partial_{x_i} f(x) \partial_{x_1} R_{k_1}(x_1)\, dx = \tag{3.33}$$

$\ldots$ integration by parts and extension with $\prod \frac{1+x_i^2}{1+x_i^2}$ gives

$$= \prod \tfrac{1}{k_i^2} \int\limits_{\mathbb{R}^n} (-1)^n \prod (1+x_i^2) \partial_{x_i} [(1+x_i^2)\partial_{x_i}] f(x) R_{k_i}(x_i)\, dx =$$

$$= \prod \tfrac{1}{k_i^2} \int\limits_{\mathbb{R}^n} \big(A_x f(x)\big) R_k(x) u(x) dx.$$

By repeating this procedure m and $m+1$ times one obtains

$$\|R_k\|_u^2 a_k = \prod \tfrac{1}{k_i^{2m}} \int\limits_{\mathbb{R}^n} \big(A_x^m f(x)\big) R_k(x) u(x) dx \tag{3.34}$$

and

$$\|R_k\|_u^2 a_k = \prod \tfrac{1}{k_i^{2m+2}} \int\limits_{\mathbb{R}^n} \partial_x \big(A_x^m f(x)\big) \partial_x R_k(x) u(x)^{-1} dx. \tag{3.35}$$

As usual (most prominently in the Fourier case) one starts from Parseval's identity (3.17) to get

$$\|f - \mathcal{P}_N f\|_u^2 =$$

$$= \sum_{|k|=N+1}^{\infty} a_k^2 \|R_k\|_u^2 = \begin{cases} \displaystyle\sum_{|k|=N+1}^{\infty} \frac{\Big(\prod \frac{1}{k_i^{2m}} \int\limits_{\mathbb{R}^n} \big(A_x^m f(x)\big) R_k(x) u(x) dx \Big)^2}{\|R_k\|_u^2} \\ \\ \displaystyle\sum_{|k|=N+1}^{\infty} \frac{\Big(\prod \frac{1}{k_i^{2m+1}} \int\limits_{\mathbb{R}^n} \partial_x \big(A_x^m f(x)\big) \partial_x R_k(x) u(x)^{-1} dx \Big)^2}{\|\partial_x R_k\|_{u^{-1}}^2}, \end{cases} \tag{3.36}$$

using (3.34) and (3.35), respectively. For the norm and orthogonality of $\partial_x R_k$, to obtain the second equation, we refer to Wang & Guo (2002).

In Remark 3.14 it was mentioned that whenever $k_i = 0$ the Sturm-Liouville operator cannot be applied as in (3.33). In one dimension this will not occur, where the sum (as in (3.36)) starts at $k = N+1$. In the multivariate case $|k| = N+1$ only means that at least one $k_i = N+1$, which does not exclude other k_i being zero. But, with $R_0 \equiv 1$, every R_{k_i}, $k_i = 0$, can be ignored in the calculations above. Thus, by finding a lower bound for the products $\prod 1/k_i^{2m}$ and $\prod 1/k_i^{2m+2}$ for all $|k| \geq N+1$ and then starting the summation from zero one

obtains from (3.36)

$$\|f - \mathcal{P}_N f\|_u^2 \leq$$
$$\leq \begin{cases} c\frac{1}{N^{4m}} \sum_{|k|=0}^{\infty} \frac{\langle A_x^m f, R_k\rangle_u^2}{\|R_k\|_u^2} \leq cN^{-4m}\|A_x^m f\|_u^2 \leq cN^{-4m}\|f\|_A^2 \\ \\ c\frac{1}{N^{4m+2}} \sum_{|k|=0}^{\infty} \frac{\langle \partial_x(A_x^m f), \partial_x R_k\rangle_{u^{-1}}^2}{\|\partial_x R_k\|_{u^{-1}}^2} \leq c\frac{1}{N^{4m+2}}\|\partial_x(A_x^m f)\|_{u^{-1}}^2 \leq cN^{-4m-2}\|f\|_A^2, \end{cases} \tag{3.37}$$

where Parseval's identity was used for the sums and Lemma 3.10 yields the final inequality (see Appendix B for the second line involving the partial derivative of R_k). Taking the square root and replacing $r = 2mn$ and $r = (2m+1)n$, respectively, finishes the proof. □

Remark 3.15. The fact that the convergence rate depends on the dimension n is connected to the definition of the $H_{u,A}^r$ norm in (3.19). This definition comes naturally in the sense that it represents just a restriction on the decay behavior of the functions compared to H_u^r (cf. Lemma 3.9 and Remark 3.13). Considering the first inequality in (3.31), replacing $r = 2m$ on the right-hand side, and using this as the defined norm in $H_{u,A}^r$ shows Lemma 3.10 to remain valid, but since the summation over $k_i = 0, \ldots, r$ for $i = 1, \ldots, n$ does not include *all* k, such that $|k|_s \leq r$, $H_{u,A}^r$ might in general not be a subspace of H_u^r, although (3.21) still holds. By allowing the multi-dimensional $H_{u,A}^r$ norm to sum only the necessary terms to satisfy the inequality (3.31), replacing $r = 2m$ in (3.37) yields the convergence rate to be N^{-r}. On the other hand, the rate $N^{-r/n}$ correlates to the findings in Theorems 3.16 and 3.17 in the sense that the differentiability requirement therein is also n times the convergence rate. Although Shen & Wang (2009) did not consider rational Chebyshev polynomials per se, it is worthwhile mentioning their findings on the slower convergence rate when using an algebraic mapping (cf. (3.2)) for approximating functions with an oscillating decay behavior at infinity.

Remark 3.16. As mentioned in Guo (1998) and in Wang & Guo (2002) results such as Theorem 3.11 can be proved for all real $r \geq 0$ by what is known as *space interpolation*. This shall not be treated here. More interestingly, $r = 0$ in (3.32) shows the right-hand side to be independent of N, thus one can assert, for a non-differentiable but L_u^2 integrable function, the projection error does tend to zero (cf. Parseval's identity), but the rate of convergence might not be expressible in inverse powers of N. Conversely, if r grows arbitrarily, the definition of $H_{u,A}^r$ then requires more and more derivatives of the function to exist and be integrable in the $H_{u,A}^r$ sense, in addition to a decay, successively more rapid than any inverse power of $|x|$ (cf. *Schwartz* space). On the other hand, Theorem 3.11 reveals (in terms of Lebesgue spaces) a convergence rate for smooth functions with rapid decay faster than any power of N. Theorem 3.16 treats the classical differentiable version, where (sometimes called) *spectral convergence* occurs again for smooth functions.

The Sobolev type results stated above are of a theoretical nature showing convergence and convergence rates in a weak sense, using the least possible requirements for functions (to be approximated). In most applications a usual goal is to plot graphs of solutions to gain insight into the qualitative and also quantitative behavior, cf. Section 2. In these situations a small L^2_w error does not necessarily mean a "good" approximation - locally, the error can be of the same order of magnitude as the function value, or even tend to infinity. The following examples shall demonstrate such occurrences.

Example 3.3. Let $f \in L^2_w(\mathbb{R})$ be given as $f(x) = (1+x^2)^{1/8}$, where $f \to \infty$ as $x \to \pm\infty$. For the orthogonal projection $\mathcal{P}_N f = \sum_{k=0}^{N} a_k R_k$ the coefficients are calculated via

$$a_k = \frac{\langle f, R_k \rangle_w}{\|R_k\|^2_w} = \frac{1}{\|R_k\|^2_w} \int_{\mathbb{R}} R_k(x)(1+x^2)^{-7/8} dx,$$

which possess an analytical expression for every k. Figure 32 depicts the graphs of $\mathcal{P}_N f$ for different N.

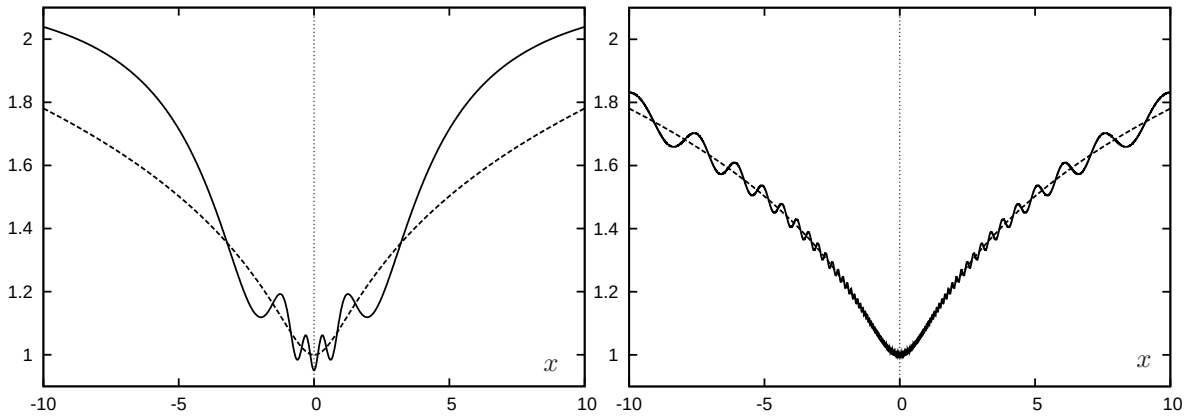

Figure 32: $\mathcal{P}_N f$ (solid), f (dashed). Left: $N = 10$, right: $N = 200$.

Convergence on the interval depicted in Figure 32 is somehow obvious, but very slow and due to the heavy oscillations unsatisfactory from a qualitative point of view. Additionally, as $x \to \pm\infty$, $\mathcal{P}_N f \to const.$, i.e. $\mathcal{P}_N f(\pm\infty) = \sum_{k=0}^{N} a_k$ (cf. asymptotic behavior of R_k in Lemma 3.2(iii)), which leads to an increasing difference between the function and its projection, for all N, as $|x|$ grows.

From Lemma 3.4 one has $\|f - \mathcal{P}_N f\|_w \to 0$ as $N \to \infty$ and with the far field of $f \sim x^{1/4}$ Theorem 3.11 means that the L^2 error decreases with the rate $1/N^r$, whereas $r < 1/4$ (from (3.19)), as shown in Figure 33.

The reason for the good agreement of the L^2 error with the proved convergence rate lies in the fact that f is actually smooth (where every derivative is again an L^2_w function), but due

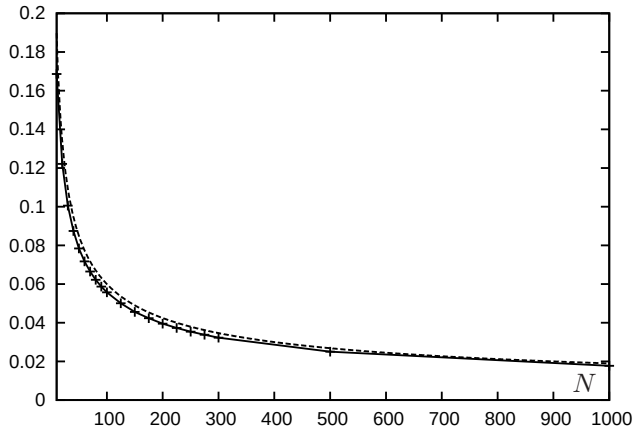

Figure 33: $\|f - \mathcal{P}_N f\|_w^2$ (solid) compared to $c/N^{1/2}$ (dashed)

to the severe restrictions in the $H^r_{w,A}$ norm on the decay behavior one cannot have a faster rate than $r < 1/4$.

Example 3.4. Let $f \in L^2_w(\mathbb{R})$ be given as $f(x) = 1/(x^2)^{1/8}$, where $f \to \infty$ as $x \to 0$. In contrast to the previous example, f now decays at infinity but has an L^2 integrable singularity at zero. Again, the integrals for the projection coefficients can be calculated in an exact manner for every k. Figure 34 shows the projection for different N. Although

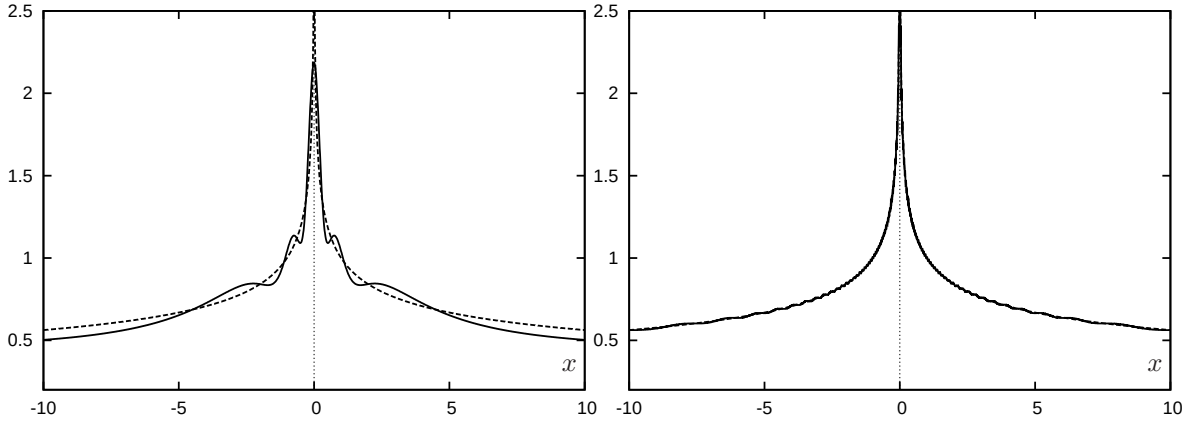

Figure 34: $\mathcal{P}_N f$ (solid), f (dashed). Left: $N = 10$, right: $N = 200$

the approximation appears to be "good" for $N = 200$, again there exists a point, $x = 0$, where the actual difference between the function and its approximation remains infinitely large (independently of N).
From the decay of f at infinity one would calculate $r < 3/4$ as the convergence rate. This cannot be satisfied (see Figure 35), since f has neither a classical nor a weak derivative (in order to apply Theorem 3.11) due to its singularity at zero. By removing the singularity "smoothly", i.e. define $g(x) = 1/(1+x^2)^{1/8}$, where g and f coincide very well away from zero

and thus decay in the same way at infinity, one can see (Figure 35) the predicted convergence rate to be satisfied.

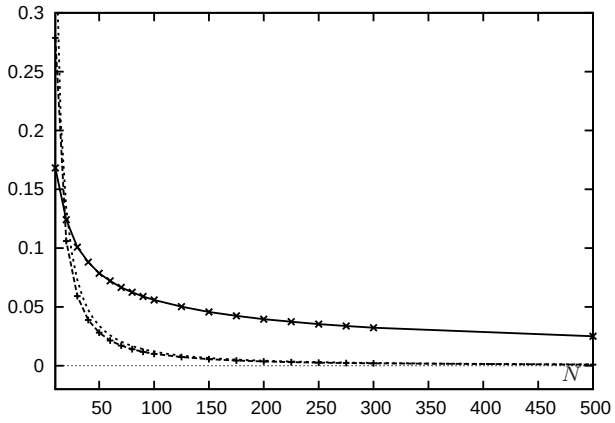

Figure 35: $\|f - \mathcal{P}_N f\|_w^2$ (solid), $\|g - \mathcal{P}_N g\|_w^2$ (dashed) compared to $c/N^{3/2}$ (dotted), where for comparison reasons the absolute values of the dashed and dotted lines have been multiplied by 100.

N.b.: Since every polynomial is smooth in $\mathbb{R}^n$ (Lemma 3.2(ii)) so is the projection $\mathcal{P}_N f$ for every N, and with $|\sum^N a_k R_k| \leq \sum^N |a_k R_k| \leq \sum^N |a_k| = const.$ (i.e. bounded everywhere) it is easy to see that demanding the approximation to be close (in the sense of its actual shape) to the original function needs some additional assumptions.

With such examples in view, the meaning of the infinite sum, i.e. $f(x) = \sum_{k=0}^{\infty} a_k R_k(x)$ (cf. Remark 3.10), has to be dealt with. At the beginning of this section (3.12) shows the other types of convergence used in approximation theory, which are more appropriate in those cases, where the local structure of a function is of interest (and are hence often left out in mainly theoretical works).

As for the following *pointwise convergence* shall be defined as

$$\begin{aligned} &\forall \epsilon > 0 \text{ and } \forall x \in \mathbb{R}^n : \ \exists N, \text{ such that} \\ &\big| f(x) - \sum_{|k| \leq N} a_k R_k(x) \big| \leq \epsilon, \quad (\text{where } N = N(x)). \end{aligned} \tag{3.38}$$

Especially with this definition another way of writing the multivariate projection operator (3.15) might be useful. Given the multi-indices $M = (M_1, \ldots, M_n)$ and analogously k, then the projection $\mathcal{P}_M$ shall be

$$\mathcal{P}_M f := \sum_{k_1=0}^{M_1} \cdots \sum_{k_n=0}^{M_n} a_k R_k \tag{3.39}$$

and by saying $N := |M|$ (which is the maximum value of all M_i) one arrives at the usual form. It is straight forward to see that all results above still hold if $\mathcal{P}_N$ is replaced by $\mathcal{P}_M$, whereas the convergence rate in Theorem 3.11 will then be given by the *minimum* of all M_i (cf. (3.37)).

Remark 3.17. In applications using $\mathcal{P}_M$ is preferable, since it is likely to involve much less terms in the summation and thus yields smaller equation systems to solve. Also, in virtue of the pointwise error, the estimate (3.38) for given ϵ and x might be satisfied with some M_i being less than $|M|$.

Analogous conclusions can be drawn for the *uniform convergence*, given as

$$\begin{aligned} &\forall \epsilon > 0 : \ \exists N, \text{ independent of } x, \text{ such that} \\ &\big| f(x) - \sum_{|k| \leq N} a_k R_k(x) \big| \leq \epsilon, \quad \forall x \in \mathbb{R}^n \end{aligned} \tag{3.40}$$

and, as usual, taking the estimate modulo sets of measure zero yields the L^∞ norm (cf. (3.12)).

As a consequence of Examples 3.3 and 3.4 we prove the following

Lemma 3.12. *Necessary and sufficient conditions for pointwise convergence of $\mathcal{P}_N f$ to f are* continuity *and* boundedness *of f in $\mathbb{R}^n$.*

Proof. It is well known (see e.g. Mason & Handscomb (2003)) that for a continuous function on $[-1, 1]$ the classical Chebyshev series converges pointwise. From assuming f is continuous and bounded on $\mathbb{R}^n$ it follows that f is continuous and bounded on $\mathbb{R}$ in every component x_i. Thus

$$\begin{aligned} &|f(x_1, \ldots, x_i, \ldots, x_n) - \sum_{k_1=0}^{N} \cdots \sum_{k_i=0}^{N} \cdots \sum_{k_n=0}^{N} a_k R_k(x)| = \\ &= |f(x_i) - \sum_{k_i=0}^{N} b_{k_i} R_{k_i}(x_i)|, \quad \text{with} \ \ b_{k_i} := \sum_{k_j, \ j \neq i} a_k \prod_{j \neq i} R_{k_j}(x_j) \end{aligned}$$

for every fixed x_j, $j \neq i$. Applying the mapping $y_i = \psi(x_i)$ from (3.2) renders $f = f(y_i)$ continuous on the compact interval $[-1, 1]$ and $R_{k_i}(x_i) = T_{k_i}(y_i)$, and hence there exists N_i for all x_i from the classical pointwise convergence result. This can then be done $\forall i = 1, \ldots, n$ (thus $\forall x \in \mathbb{R}^n$) and taking the maximum of all N_i yields the desired result.
The necessary condition is best shown by negation. Assuming f is not continuous and bounded means f is discontinuous *or* unbounded in $\mathbb{R}^n$ (in at least one point). In case of a discontinuity it is well known from classical Fourier series approximation (and thus also in the classical Chebyshev case), that the series converges to the average value of the discontinuity step. If f has a singularity at some point or unbounded growth at infinity it follows from the definition of pointwise convergence (3.38) that $\forall \epsilon$ no N can be found, such that $|\infty - \mathcal{P}_N f| \leq \epsilon$. □

Remark 3.18. For discontinuity and unboundedness on $\mathbb{R}^n$ it is sufficient to consider this in at least one component (to utilize the classical results). Continuity in every component x_i, on the other hand, does not mean continuity on $\mathbb{R}^n$, where only the inverse is true.

Formally speaking, given an orthogonal projection operator L^2, convergence can be established for functions having weak singularities and are weakly increasing in the far field.

By then selecting those functions, which are at least bounded and continuous everywhere, pointwise convergence can be achieved. As for the strongest convergence result in addition to continuity, *bounded total variation* has to be required (see Mason & Handscomb (2003) for the classical univariate case). Making use of the bijective mapping ψ (thus ensuring every partition on a compact interval to be a partition on $\mathbb{R}$) the concept of bounded variation can be directly introduced for the whole real line. Hence, the one-dimensional result shown in Mason & Handscomb (2003), i.e. *if a function defined on* $[-1,1]$ *is continuous and of bounded variation, then the Chebyshev series is uniformly convergent*, holds in the exact same way for the projection $\mathcal{P}_N f$ defined on $\mathbb{R}$.

In the situation of multivariate functions on hypercubes $[-1,1]^n$ (and eventually $\mathbb{R}^n$) using total variation does not come straight forward from its one-dimensional equivalent and for this reason Mason (1980) applied what is known as the *Dini-Lipschitz condition* to prove uniform convergence.

To extend this result to functions on $\mathbb{R}^n$ one needs the notion of the *modulus of continuity*, where a general definition can be found e.g. in Timan (1963). In the L^∞ case this reads: Let $f = f(x_1, \ldots, x_n)$ be a real function on the closed bounded domain $G \subset \mathbb{R}^n$, then the modulus of continuity is given as

$$\omega(f; u_1, \ldots, u_n) = \sup_{|x_i - y_i| \leq u_i, \forall i} |f(x_1, \ldots, x_n) - f(y_1, \ldots, y_n)|, \quad \forall x_i, y_i \in G.$$

From this Mason (1980) extracted the partial modulus for each component of f, i.e.

$$\omega_j(f; u_j) := \omega(f; 0, \ldots, u_j, \ldots, 0) = \sup_{|x_j - y_j| \leq u_j} |f(x_1, \ldots, x_j, \ldots, x_n) - f(x_1, \ldots, y_j, \ldots, x_n)|.$$

Timan (1963) then proved for any bounded 2π-periodic function f defined on G, the L^∞ best approximation f^B using trigonometric polynomials, with different degrees of approximation N_j in each component, cf. (3.39), satisfies

$$\|f - f^B\|_\infty \leq C \sum_{j=1}^{n} \omega_j \left(f; \frac{1}{N_j + 1} \right), \tag{3.41}$$

where C is a constant independent of f and N_j. As mentioned later in Handscomb (1966) this holds for any continuous function on a bounded closed region.

Furthermore, as shown in Mason & Handscomb (2003), for every projection operator the inequality

$$\|f - \mathcal{P}_N f\|_\infty \leq (1 + \|\mathcal{P}_N\|_\infty) \|f - f^B\|_\infty \tag{3.42}$$

holds, which was termed *near-best* approximation with a relative distance $\|\mathcal{P}_N\|_\infty$ (and the operator norm taken in the usual sense, cf. Notation Index).

Recalling that for continuous functions the modulus $\omega(f; 0) = 0$, uniform convergence (as $N_j \to \infty$, $\forall j$) of the best approximation f^B is easily seen from (3.41). Although the

projection can be shown (cf. Lemma 3.7) to have a finite L^2 norm, this is not the case in the L^∞ norm with $N \to \infty$ as proved in e.g. Mason & Handscomb (2003). They showed for the classical Fourier and Chebyshev projection $\|\mathcal{P}_N\|_\infty \leq \lambda_N$, with λ_N being the *Lebesgue constant* and $\lambda_N = \mathrm{O}(\log N)$ as N tends to infinity.

Remark 3.19. In the proof of Lemma 3.12 we used the mapping ψ to utilize what is known for convergence on compact intervals. The only condition to add was the function defined on $\mathbb{R}^n$ to be bounded. Or in more detail let $f = f(x)$ be bounded and continuous on $\mathbb{R}^n$ and $(y_1, \ldots, y_n) = (\psi(x_1), \ldots, \psi(x_n))$, such that $R_k(x) = T_k(y)$, then the projection

$$\mathcal{P}_N f(x) = \sum_{|k| \leq N} a_k R_k(x) = \sum_{|k| \leq N} a_k T_k(y) =: \mathcal{G}_N f(y),$$

where $f(y) := f(\psi^{-1}(y_1), \ldots, \psi^{-1}(y_n))$, yields the exact same function values as the classical projection $\mathcal{G}_N f$ for every continuous function f on $[-1, 1]^n$. Consequently, one obtains

$$\sup_{x \in \mathbb{R}^n} |\mathcal{P}_N f(x)| = \sup_{y \in [-1,1]^n} |\mathcal{G}_N f(y)| \quad \Rightarrow \quad \|\mathcal{P}_N\|_\infty = \prod_{j=1}^{n} \lambda_{N_j}, \tag{3.43}$$

as proved in Mason (1980), with the asymptotic behavior of the operator norm being $\mathrm{O}(\prod \log N_j)$. Substituting the order of magnitude estimate for the norm in (3.43) and the estimate for the best approximation (3.41) (setting $\delta_j := 1/(N_j + 1)$) into (3.42) proves the following (cf. Mason (1980))

Theorem 3.13. *If a bounded continuous function f defined on $\mathbb{R}^n$ satisfies the multi-dimensional Dini-Lipschitz condition*

$$\sum_{j=1}^{n} \omega_j(f; \delta_j) \prod_{j=1}^{n} \log \delta_j \to 0 \quad \textit{as} \quad \delta_j \to 0, \ \forall j, \tag{3.44}$$

where the modulus ω is understood as f being mapped onto $[-1, 1]^n$, then the orthogonal projection $\mathcal{P}_N f$ converges uniformly to f.

Remark 3.20. As mentioned in Mason & Handscomb (2003), this Dini-Lipschitz condition is just infinitesimally more than requiring continuity and still much less than differentiability, where $\omega = \mathrm{O}(\delta)$. Also, it is now obvious why this assumption has to be preferred to the concept of bounded variation in higher dimensions (in two dimensions, for example, an additional partial derivative of f needs to be bounded, see e.g. Mason (1967)).

Remark 3.21. The Parseval identity given in Theorem 3.8, which states the square summability of the coefficients of the projection was then used to estimate the according L^2 error,

see (3.36). The same can be done for the L^∞ norm, i.e.

$$\begin{aligned}\sup_{x\in\mathbb{R}^n}|f(x)-\mathcal{P}_Nf(x)| &= \sup_{x\in\mathbb{R}^n}|\sum_{|k|\geq N+1}a_kR_k(x)| \leq \\ &\leq \sup_{x\in\mathbb{R}^n}\sum_{|k|\geq N+1}|a_k|\underbrace{|R_k(x)|}_{\leq 1} \leq \sum_{|k|\geq N+1}|a_k|,\end{aligned} \tag{3.45}$$

such that one can assert that convergence in L^2 or L^∞ depends on how fast the coefficients tend to zero.

The next result shows under which conditions the different types of convergence imply each other.

Lemma 3.14. *Given a function f and a non-negative smooth weight $u = u(x)$, integrable over $\mathbb{R}^n$, then the following assertions for the projection $\mathcal{P}_Nf$ hold as $N \to \infty$:*

$$\begin{aligned}&(i) \quad \|f-\mathcal{P}_Nf\|_\infty \to 0 \quad \Rightarrow \quad |f(x)-\mathcal{P}_Nf(x)| \xrightarrow{ptw.} 0\\ &(ii) \quad \|f-\mathcal{P}_Nf\|_\infty \to 0 \quad \Rightarrow \quad \|f-\mathcal{P}_Nf\|_u \to 0\\ &(iii) \quad |f(x)-\mathcal{P}_Nf(x)| \xrightarrow{ptw.} 0 \quad \Rightarrow \quad \|f-\mathcal{P}_Nf\|_u \to 0\end{aligned}$$

Proof. (i) follows directly from the definitions (3.38) and (3.40).
In (ii) say $\int_{\mathbb{R}^n}u(x)dx =: c$, then for $N \gg 1$

$$\begin{aligned}\int_{\mathbb{R}^n}|f-\mathcal{P}_Nf|^2u\,dx &\leq \sup_{x\in\mathbb{R}^n}|f(x)-\mathcal{P}_Nf(x)|^2\int_{\mathbb{R}^n}u\,dx = \\ &= c\sup_{x\in\mathbb{R}^n}|f(x)-\mathcal{P}_Nf(x)|^2 \leq c\sup_{x\in\mathbb{R}^n}|f(x)-\mathcal{P}_Nf(x)|.\end{aligned}$$

Pointwise convergence implies continuity and boundedness of f (Lemma 3.12) and since the product of an integrable function (here u) and a bounded (measurable) function (here $|f|^2$) is again integrable, implication (iii) holds. □

Remark 3.22. In the proof of Lemma 3.12 it was mentioned that one cannot have pointwise convergence if the function has a singularity or a point of discontinuity. Strictly speaking, convergence cannot be expected at such points, but in every other point (or interval) where the function is bounded and continuous. In other words, if a bounded continuous function has only a finite number of discontinuities and singularities the orthogonal projection converges pointwise on every interval excluding such points. In addition to the approximation converging to the average value of the left- and right-hand side limit at a discontinuity, the height of such a step seems to be increased due to what is known as the *Gibbs phenomenon*, see Figure 41.

Equations (3.34) and (3.35) imply that if a given function f lies in $H^r_{u,A}(\mathbb{R}^n)$, the coefficients of an orthogonal projection satisfy $a_k \propto \prod\frac{1}{k_i^{r/n}}$. The following result shows how a

similar relation can be established in the case of $f \in C^m(\mathbb{R}^n)$ with a certain requirement on the decay behavior.

Theorem 3.15. *Let $f \in C^m(\mathbb{R}^n)$, where $m := sn$, $s \in \mathbb{N}$. If*

(i) $\partial_x^m f \in L^2_u(\mathbb{R}^n)$ and

(ii) $x_1^s \dots x_n^s \partial_x^n f(x) \to 0$, as $x_i \to \pm\infty$, $\forall i$,

then for every coefficient given in (3.14)*, where all $k_i \neq 0$,*

$$|a_k| \propto \prod_{i=1}^{n} \frac{1}{k_i^s} \tag{3.46}$$

holds.

Proof. The following steps proving this result are of rather asymptotic kind, meaning that asymptotic expansions of integrals are used, where a parameter (here $|k|$) is assumed to be large enough.
By substitution of $y_i = \phi(x_i)$ and $dy_i/dx_i = \phi'(x_i) = w(x_i)$, where $y_i \in I := [-\pi, 0]$ and using the cosine description from (3.1) for the polynomials, the coefficients read

$$\|R_k\|^2 a_k = \int_{I^n} \underbrace{f(\phi^{-1}(y_1), \dots, \phi^{-1}(y_n))}_{=:f(y_1,\dots,y_n)} \prod_{i=1}^{n} \cos(k_i y_i)\, dy. \tag{3.47}$$

The strategy will be to use integration by parts consecutively for all components s times. Since the existence of the above written integral is trivially given, the order of integration must not play a role in the argumentation. Thus, by rearranging the integrand and starting the procedure at y_i one obtains

$$a_k \propto \int_{I^{n-1}} \prod_{j\neq i} \cos(k_j y_j) \left[\frac{f(y)\sin(k_i y_i)}{k_i} \Big|_{y_i=-\pi}^{y_i=0} - \frac{1}{k_i} \int_I \partial_{y_i} f(y) \sin(k_i y_i)\, dy_i \right] \prod_{j\neq i} dy_j,$$

where f is, such that the boundary term can be set to zero. Performing this for all the other components one arrives at

$$a_k \propto (-1)^n \prod_{i=1}^{n} \frac{1}{k_i} \int_{I^n} \partial_y^n f(y) \prod_{i=1}^{n} \sin(k_i y_i)\, dy,$$

with the condition $f(y)\sin(k_i y_i) = 0$ at $y_i \in \{-\pi, 0\}$, $\forall i$. Evaluating the boundary term to zero at *both* boundaries is chosen, such that no symmetry has to be required of f at $\pm\infty$ in every component. Obviously, the next step yields

$$a_k \propto (-1)^{2n} \prod_{i=1}^{n} \frac{1}{k_i^2} \int_{I^n} \partial_y^{2n} f(y) \prod_{i=1}^{n} \cos(k_i y_i)\, dy,$$

where $\partial_y^n f(y)\cos(k_i y_i) = 0$ at $y_i \in \{-\pi, 0\}\ \forall i$, such that one can claim an odd number of steps gives the cosine function in the boundary term and the sine function at the even numbers. Clearly, s such modifications finally yield

$$|a_k| \propto \prod_{i=1}^{n} \frac{1}{k_i^s} \left| \int_{I^n} \partial_y^{sn} f(y) \prod_{i=1}^{n} \begin{pmatrix} \cos(k_i y_i) \\ \sin(k_i y_i) \end{pmatrix} dy \right|, \tag{3.48}$$

where the stacked sine and cosine functions symbolize the dependence on whether s is even or odd. Since no restriction shall be made on the parity of s the condition is given as $\partial_y^{(s-1)n} f(y) = 0$ in every component $y_i \in \{-\pi, 0\}$. Appendix C concludes the proof by showing how this condition is derived in detail yielding assumptions (i) and (ii). □

Remark 3.23. If a coefficient a_k contains at least one $k_i = 0$, the cosine function is identical to 1 (the zeroth order polynomial), which means the integration by parts procedure in the proof is just not performed for this component. Eventually, the requirements (i) and (ii) in the theorem above then read

$$\left.\begin{array}{ll} (i) & \prod \partial_{x_i}^s f \in L_u^2(\mathbb{R}^n) \\ (ii) & \prod x_i^s \partial_{x_i} f(x) \to 0, \text{ as } x_i \to \pm\infty \end{array}\right\} \forall i, \text{ such that } k_i \neq 0.$$

The essential assertion (3.46) remains unchanged (provided all $k_i = 0$ are not considered in the product).

As a direct consequence of the asymptotic rate for the coefficients and (3.45) the next result provides the convergence rate in the L^∞ norm, similarly to Theorem 3.11.

Theorem 3.16. *Given a function f satisfying the assumptions in Theorem 3.15 with $s > 1/2$, there exists a constant c independent of N, such that*

$$\|f - \mathcal{P}_N f\|_\infty \leq c N^{1/2-s}. \tag{3.49}$$

Proof. Combining (3.45) and (3.48) and applying Cauchy's inequality yields

$$\begin{aligned} &\|f - \mathcal{P}_N f\|_\infty \leq \\ &\leq \sum_{|k|\geq N+1} |a_k| = \sum_{|k|\geq N+1} \prod_{i=1}^{n} \frac{1}{k_i^s} \left| \int_{I^n} \partial_y^{sn} f(y) \prod_{i=1}^{n} \begin{pmatrix} \cos(k_i y_i) \\ \sin(k_i y_i) \end{pmatrix} dy \right| \|R_k\|_u^{-2} \leq \\ &\leq \left(\sum_{|k|\geq N+1} \prod_{i=1}^{n} \frac{1}{k_i^{2s}} \right)^{1/2} \left(\sum_{|k|\geq N+1} \left| \int_{I^n} \underbrace{\partial_y^{sn} f(y)}_{=:g(y)} \prod_{i=1}^{n} \begin{pmatrix} \cos(k_i y_i) \\ \sin(k_i y_i) \end{pmatrix} dy \right|^2 \|R_k\|_u^{-4} \right)^{1/2}. \end{aligned} \tag{3.50}$$

The coordinate transform $y_i = \phi(x_i)$ used in (3.47) changes the integral on the right-hand side above, such that

$$\int_{I^n} g(y) \prod_{i=1}^{n} \begin{pmatrix} \cos(k_i y_i) \\ \sin(k_i y_i) \end{pmatrix} dy = \int_{\mathbb{R}^n} g(\phi(x)) \underbrace{\prod_{i=1}^{n} \begin{pmatrix} \cos(k_i \phi(x_i)) \\ \sin(k_i \phi(x_i)) \end{pmatrix}}_{=: \binom{R_k(x)}{S_k(x)}} u(x)\, dx = \left\langle g(\phi(x)), \begin{pmatrix} R_k \\ S_k \end{pmatrix} \right\rangle_u,$$

provided $g \circ \phi \in L^2(\mathbb{R}^n)$ (which follows from the assumptions (i) and (ii)), where Appendix D shows the polynomial system $\{S_k\}$ (generated from the sine function) to be a complete orthogonal set in $L^2_w(\mathbb{R}^n)$. One can then define

$$r_k := \frac{\langle g(\phi(x)), R_k \rangle_u}{\|R_k\|_u^2} \quad \text{and} \quad s_k := \frac{\langle g(\phi(x)), S_k \rangle_u}{\|S_k\|_u^2}$$

as the coefficients of an expansion of $g \circ \phi$ in the systems R_k and S_k and since $\|R_k\|_u = \|S_k\|_u$ the second sum on the right-hand side in (3.50) is part of the sum over squared coefficients r_k and s_k, respectively. Hence, in virtue of Parseval's identity this term is bounded by the L^2 norm of $g \circ \phi$.
As for the first sum in (3.50) it shall be noted that for every $|k| = a \in \mathbb{N}$

$$\prod_{i=1}^{n} \frac{1}{k_i^{2s}} \leq \frac{1}{a^{2s}},$$

since every $k_i \geq 1$ and thus the following estimate holds

$$\left(\sum_{|k| \geq N+1} \prod_{i=1}^{n} \frac{1}{k_i^{2s}} \right)^{1/2} \leq \left(\sum_{k \geq N+1} \frac{1}{k^{2s}} \right)^{1/2} \leq c N^{1/2-s}, \tag{3.51}$$

as long as $s > 1/2$ to make sure the sum converges. □

A similar result in terms of the one-dimensional Fourier sum is given in Hesthaven et al. (2007). The Riemann-Lebesgue lemma can be applied to the integrals in (3.50), saying that as $|k| \to \infty$ the integrals tend to zero, concluding that there exists a maximum value c for all k, such that the estimate becomes

$$\|f - \mathcal{P}_N f\|_\infty \leq \sum_{|k| \geq N+1} |a_k| \leq c \sum_{|k| \geq N+1} \prod_{i=1}^{n} \frac{1}{k_i^s} \leq c N^{1-s},$$

provided $s > 1$ for convergence of the sum. This shows that by assuming the convergence of the sum over the absolute values of the coefficients is only due to the multi-index, one loses $1/2$ of the convergence rate. Additionally, comparing this with the L^2 convergence rate established in Theorem 3.11, one can claim that in principle more polynomials are needed to obtain the L^∞ error to be equal to the L^2 error. And as mentioned in Remark 3.16 again

one obtains a convergence rate faster than any negative powers of N for smooth functions, which decay rapidly enough (cf. definition of the *Schwartz space*).

The uniform convergence rate can also be given in terms of the Sobolev type space $H^r_{u,A}$, for applying (3.34) and (3.35) to (3.50) (and thus replacing the integrals therein) yields (in the even case)

$$\|f-\mathcal{P}_N f\|_\infty \leq \sum_{|k|\geq N+1}|a_k| = \sum_{|k|\geq N+1}\prod_{i=1}^{n}\frac{1}{k_i^{\frac{r}{n}}}\left|\langle A_x^{\frac{r}{2n}}f,R_k\rangle_u\right|\|R_k\|_u^{-2} \leq$$
$$\leq \left(\sum_{|k|\geq N+1}\prod_{i=1}^{n}\frac{1}{k_i^{\frac{2r}{n}}}\right)^{1/2}\left(\sum_{|k|\geq N+1}\left|\langle A_x^{\frac{r}{2n}}f,R_k\rangle_u\right|^2\|R_k\|_u^{-4}\right)^{1/2},$$

where the second sum in bounded by Parseval's identity and subsequently by $\|f\|_A$ (as shown in Lemma 3.10) and with the bound for the first sum in (3.51) we have thus proved

Theorem 3.17. *Given $f\in H^r_{u,A}(\mathbb{R}^n)$, $r/n>1/2$ then there exists a constant c independent of N, such that*

$$\|f-\mathcal{P}_N f\|_\infty \leq cN^{1/2-r/n}\|f\|_A. \tag{3.52}$$

Overall, in the above one has all the necessary arguments to why Boyd (2001) recommends an approximation using rational Chebyshev polynomials (only) for differentiable functions, decaying algebraically at infinity.

Example 3.5. Given the function

$$f(x)=\frac{x}{1+x^2},\quad \text{where}\quad x^2f'(x)\to const.\ \text{ and }\ xf'(x)\to 0\ \text{ as }\ |x|\to\infty,$$

such that one can estimate a convergence rate to be N^{-s}, with $\frac{1}{2}<s<\frac{3}{2}$ (or $s\leq\frac{3}{2}$ in $H^r_{u,A}$), as claimed in Theorem 3.16 (i.e. (3.49) combined with assumption (ii) in Theorem 3.15). Figure 36 confirms $s=1$ to be most appropriate, which clearly demonstrates that *spectral convergence* highly depends on the decay rate at infinity, even for smooth functions. Performing the same approximation for $f(x)=\exp(-x^2)$ Table 15 shows the "exponential" dependence of the uniform error on N.

N	error
10	8×10^{-3}
50	1×10^{-8}
100	5×10^{-10}
150	6×10^{-13}
200	1×10^{-15}

Table 15: Uniform error approximating $f(x)=\exp(-x^2)$

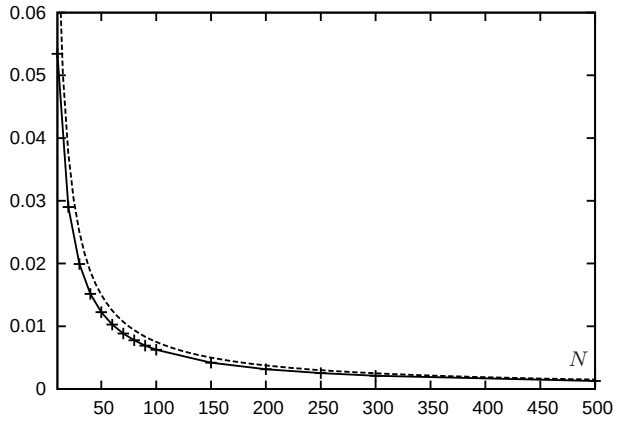

Figure 36: $\|\mathcal{P}_N f - f\|_\infty$ vs. N (solid), c/N (dashed)

Remark 3.6 shows the change in the behavior of the rational Chebyshev polynomials when introducing a stretching factor L in the maps ψ or ϕ. In all the results above $L = 1$ was chosen for the sake of simplicity (of some calculations) and readability, but the general assertions are independent of L. In other words, the convergence per se (for all three types mentioned) is obviously independent of the value of L, whereas for the convergence rate the constants, especially when calculating the lowest one, might depend on L. This is due to the fact that the norm of the error can differ in orders of magnitude at constant N for different L, as the following example will demonstrate.

It shall be noted here, that it is straight forward to see, that the error depends continuously on L and also the minimum of the error cannot be at $L = 0$ or $L = \infty$. So at some finite L for every N there exists a minimum for the error in a given norm. Some aspects of finding such an optimal value in some special cases can be found in Boyd (1982) and Boyd (1987), as well as Boyd (2001a) where it was also mentioned that $L = 1$ is not always the best choice, but still yields good approximations.

Example 3.6. Given $f(x) = (1+x^2)^{-1/2}$, which does not possess a finite Chebyshev expansion for any value of L, Figure 37 depicts approximations calculated using $N = 6$ polynomials for various L. This clearly shows that $L < 1$ might never be a good choice, but experiments confirm L between 1 and approximately 4 to be most appropriate in almost all applications, see Boyd (2001a) for the graph of the error as a function of L.
For the three cases in Figure 37 the error $e_N(L)$ in the L^∞ norm evaluates to $e_N(1) \approx \frac{9}{100}$, $e_N(2) \approx \frac{4}{100}$ and $e_N(\frac{1}{2}) \approx \frac{20}{100}$, whereas the pointwise error attains its maximum at infinity (for all L).

3.2.2 Interpolation and the Aliasing Error

So far the results in this section require functions given in a closed form, in order to calculate the integrals defining the expansion coefficients (cf. (3.14)). One can only expect to obtain

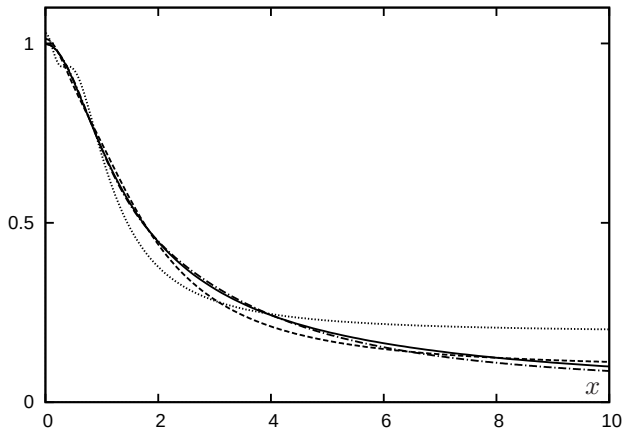

Figure 37: $f(x) = (1 + x^2)^{-1/2}$ (solid), approximation with $L = 1$ (dashed), $L = 2$ (dash-dotted, almost indistinguishable from f) and $L = 1/2$ (dotted).

analytic expressions of such integrals for some special or simple formulae describing these functions, otherwise an approximate evaluation has to be performed. Also, if the function is described only as a set of point values, another form of obtaining the coefficients has to be found.

The most obvious way would be to apply a quadrature scheme as an approximation of the integrals. The problem thereby is that the orthogonality can be restored only approximately. Additionally, higher order integration schemes, especially in higher dimensions, are very expensive in calculational costs.

The best way, as done in all standard textbooks, to circumvent such problems is to define a *discrete inner product* which relates to quadrature schemes using a certain set of points and special weights. This ensures the exact orthogonality and eventually results in unique *interpolation polynomials*.

Given bounded functions f and g on $\mathbb{R}^n$ the *discrete inner product* shall be defined as

$$\begin{aligned} \langle f, g\rangle_N &:= \sum_{j_n=0}^{N}{}'' \dots \sum_{j_1=0}^{N}{}'' f(x_{1_{j_1}}, \dots, x_{n_{j_n}}) g(x_{1_{j_1}}, \dots, x_{n_{j_n}}), \\ \text{with } \; x_{i_{j_i}} &:= \psi^{-1}\Big(\cos\left(\frac{j_i \pi}{N}\right)\Big), \quad j_i = 0, \dots, N \;\text{ and }\; i = 1, \dots, n, \end{aligned} \tag{3.53}$$

ψ taken from (3.2) and the double dash denotes the first and last term to be halved. Furthermore, in the case of rational Chebyshev polynomials using definition (3.1) and the result presented in Mason & Handscomb (2003) this yields

$$\langle R_k, R_l\rangle_N := \sum_{|j|=0}^{N}{}'' R_k(x_j) R_l(x_j) =$$

$$= \sum_{j_n=0}^{N}{}'' \dots \sum_{j_1=0}^{N}{}'' \prod_{i=1}^{n} R_{k_i}(x_{i_{j_i}}) R_{l_i}(x_{i_{j_i}}) =$$

$$= \sum_{j_n=0}^{N}{}'' R_{k_n}(x_{n_{j_n}}) R_{l_n}(x_{n_{j_n}}) \dots \sum_{j_1=0}^{N}{}'' R_{k_1}(x_{1_{j_1}}) R_{l_1}(x_{1_{j_1}}) =$$

$$= \prod_{i=1}^{n} \sum_{j_i=0}^{N}{}'' R_{k_i}(x_{i_{j_i}}) R_{l_i}(x_{i_{j_i}}) = \prod_{i=1}^{n} \langle R_{k_i}, R_{l_i} \rangle_N = \begin{cases} 0 & k \neq l \\ \prod_{i=1}^{n} \|R_{k_i}\|_N^2 & \text{otherwise} \end{cases},$$

with the points $x_{i_{j_i}}$ as in (3.53) and the discrete norm $\|\cdot\|_N$ defined via the square root of the inner product, i.e.

$$\|R_{k_i}\|_N^2 = \begin{cases} N & k_i = 0 \text{ or } k_i = N \\ \frac{N}{2} & \text{otherwise} \end{cases}. \tag{3.54}$$

Obviously, by using ψ^{-1} to gain the points of evaluation, one actually deals with the classical polynomials T_k on $[-1,1]^n$, as seen from the definition $R_k(x) = T_k(\psi(x))$.

Remark 3.24. Mason & Handscomb (2003) argued that the discrete inner product is not unique and depends on the points of evaluation (here, the extrema of T_N), which themselves are related to the interpolation points of quadrature schemes.

The definition works in the same way when using the zeros of T_{N+1} (resulting in different values for the norm), cf. Lemma 3.2(ix). Comparing the values given in (3.54) with their continuous counterparts in (3.5) shows a difference (due to the dependency on N), although the discrete formula should be exact in the case of polynomials. This is why other authors, e.g. Guo (1998) or Hesthaven et al. (2007) introduce (quadrature-)weights in (3.53).

The following will show that when deriving discrete versions of expansion coefficients such weights will cancel out and will also show how the discrete inner product relates to the *fast Fourier transform*.

Let f be a bounded function on $\mathbb{R}^n$ and $\mathcal{I}_N$ be given as

$$\mathcal{I}_N f(x) := \sum_{|k|=0}^{N} b_k R_k(x), \quad \text{with} \quad b_k = \frac{\langle f, R_k \rangle_N}{\prod_{i=1}^{n} \|R_{k_i}\|_N^2}. \tag{3.55}$$

It is straight forward (cf. Lemma 3.7) to see that this operator is an orthogonal projection with respect to the inner product $\langle \cdot, \cdot \rangle_N$. The next result shows how this projection relates to the given function.

Lemma 3.18. *Let f be bounded and defined on all points in $\mathbb{R}^n$, then the projection $\mathcal{I}_N$ interpolates f at the points $x_{i_{j_i}}$ given in* (3.53).

Proof. The classical one-dimensional result can be found in e.g. Mason & Handscomb (2003) and Hesthaven et al. (2007) (also for the general Jacobi polynomials). To show this in $\mathbb{R}^n$,

evaluation of $\mathcal{I}_N f$ at the given points (with multi-indices l, j) yields

$$\begin{aligned}
\mathcal{I}_N f(x_{1_{l_1}}, \ldots, x_{n_{l_n}}) &= \sum_{|k|=0}^{N} b_k R_k(x_{1_{l_1}}, \ldots, x_{n_{l_n}}) = \sum_{|k|=0}^{N} \frac{\langle f, R_k \rangle_N}{\prod_{i=1}^{n} \|R_{k_i}\|_N^2} R_k(x_{1_{l_1}}, \ldots, x_{n_{l_n}}) = \\
&= \sum_{|k|=0}^{N} \frac{1}{\prod_{i=1}^{n} \|R_{k_i}\|_N^2} \sum_{|j|=0}^{N}{}'' f(x_j) R_k(x_j) R_k(x_{1_{l_1}}, \ldots, x_{n_{l_n}}) = \\
&= \sum_{|j|=0}^{N}{}'' f(x_j) \sum_{|k|=0}^{N} \frac{R_k(x_j) R_k(x_{1_{l_1}}, \ldots, x_{n_{l_n}})}{\prod_{i=1}^{n} \|R_{k_i}\|_N^2} = \qquad (3.56) \\
&= \sum_{|j|=0}^{N}{}'' f(x_j) \underbrace{\sum_{k_n=0}^{N} \cdots \sum_{k_1=0}^{N} \prod_{i=1}^{n} \frac{R_{k_i}(x_{i_{j_i}}) R_{k_i}(x_{i_{l_i}})}{\|R_{k_i}\|_N^2}}_{(*)},
\end{aligned}$$

where it is now sufficient to show (in one dimension) that

$$\frac{1}{N} + \frac{2}{N} \sum_{k=1}^{N-1} R_k(x_i) R_k(x_j) + (-1)^{i+j} \frac{1}{N} = \begin{cases} \delta_{ij} & 1 \leq i \leq N-1 \\ 2\delta_{ij} & i = 0, N \end{cases}, \qquad (3.57)$$

such that $(*)$ in (3.56) is equal to $\prod_i \delta^*_{j_i l_i}$ where δ^* shall abbreviate the right-hand side in (3.57). The factor 2 cancels out with the halved terms in the sums in (3.56) regarding the double dashes.
Considering the fact that $R_k(x_i) = T_k(\cos(i\pi/N)) = \cos(ki\pi/N)$, it is easy to see that (3.57) is the same as the discrete inner product $\langle R_i(x_k), R_j(x_k) \rangle_N$ multiplied by $2/N$ (see Mason & Handscomb (2003)), which concludes the proof. □

Remark 3.25. Since the coefficients in (3.55) are uniquely defined one can assert the uniqueness of the *interpolation polynomial* to a given function from Lemma 3.18. It is fairly obvious that one can use more terms to evaluate b_k in (3.56), whereas the interpolation points might then be shifted, due to (3.57) being defined only for $i, j \leq N$.

Intuitively, since the discrete inner product combined with certain weights is a quadrature formula for the integrals of the continuous inner product, one (reasonably) expects the discrete coefficients to converge to the continuous counterparts and consequently $\mathcal{I}_N$ to tend to $\mathcal{P}_N$. The difference between those two is known as the *aliasing error*, which shall be dealt with later. From the argumentation for (3.57) the connection of the discrete inner product with the FFT can be readily deduced as

Lemma 3.19. *Given the discrete inner product and the points x_j as in* (3.53) *the identity*

$$\left(\frac{2}{N}\right)^{n/2} \big\langle f, R_k \big\rangle_N = multiDCT\big(f(x_j)\big) \qquad (3.58)$$

holds for all bounded functions defined on $\mathbb{R}^n$.

Proof. By *multiDCT* we mean a successively applied classical DCT (the cosine part of the FFT) in all independent variables. There are several definitions of DCTs in numerical libraries and computer algebra systems. The normalization factor $\sqrt{\frac{2}{N}}$ above indicates the use of the version $v = DCT(u)$, where $v = (v_1, \ldots, v_M)$, $u = (u_1, \ldots, u_M)$ and

$$v_s = \sqrt{\frac{2}{M-1}} \left[\frac{u_1}{2} + \sum_{r=2}^{M-1} u_r \cos\left(\frac{\pi}{M-1}(r-1)(s-1) \right) + (-1)^{s-1} \frac{u_M}{2} \right],$$

for a given vector u (see e.g. *Mathematica* documentations). Starting from the inner product and substituting the evaluation points

$$\langle f, R_k \rangle_N = \sum_{j_n=0}^{N}{}'' R_{k_n}(x_{n_{j_n}}) \ldots \sum_{j_1=0}^{N}{}'' f(x_j) R_{k_1}(x_{1_{j_1}}),$$

it can be easily seen that showing the equivalence to the one dimensional DCT is sufficient, i.e.

$$\sum_{j_i=0}^{N}{}'' f(x_j) R_{k_i}(x_{i_{j_i}}) = \sum_{j_i=0}^{N}{}'' f(x_{1_{j_1}}, \ldots, x_{n_{j_n}}) \cos\left(\frac{k_i\, j_i\, \pi}{N} \right) =$$
$$= \frac{f(x_{i_0})}{2} + \sum_{j_i=1}^{N-1} f(x_{i_{j_i}}) \cos\left(\frac{k_i\, j_i\, \pi}{N} \right) + (-1)^{k_i} \frac{f(x_{i_N})}{2}$$

and by shifting the indices $j_i = r - 1$, $k_i = s - 1$, $N = M - 1$ we arrive at the definition of the DCT modulo the normalization factor $\sqrt{\frac{2}{N}}$. Performing this transformation for every independent variable it is straight forward to obtain (3.58). □

Caveat: Considering the evaluation points x_{i_0} and x_{i_N} one can immediately see that the values of the function at infinity are needed. Hence it is of advantage in most numerical schemes to set them a priori.

As mentioned in Remark 3.25, what is known as the *aliasing error* is in fact the difference between the coefficients gained from the discrete and continuous inner products. Hesthaven et al. (2007) and Guo (1998) derived a formula for this difference in the case of Jacobi-type polynomials on a bounded interval. In the following b_k shall always represent the discrete and a_k the continuous coefficients. For one-dimensional rational Chebyshev polynomials this then reads

$$b_k - a_k = \frac{1}{\|R_k\|_N^2} \sum_{l=N+1}^{\infty} a_l \langle R_l, R_k \rangle_N, \tag{3.59}$$

which is completely similar to the classical version, due to the choice of the interpolation points defining the discrete inner product. Using this one can write the aliasing error as

$$\mathcal{A}_N f := (\mathcal{I}_N - \mathcal{P}_N) f = \sum_{l=N+1}^{\infty} (\mathcal{I}_N R_l) a_l, \tag{3.60}$$

indicating that the error relates to the fact that higher order polynomials are not included exactly, but as their interpolation polynomial (cf. Hesthaven et al. (2007)). In other words, higher order terms are given via polynomials with degree less than or equal to N and can thus not be distinguished from corresponding lower order terms (for some examples see Mason & Handscomb (2003)). Equation (3.60) can be deduced via

$$\begin{aligned}\mathcal{A}_N f(x) &= \sum_{k=0}^{N}(b_k - a_k)R_k(x) = \\ &= \sum_{k=0}^{N}\left(\frac{1}{\|R_k\|_N^2}\sum_{l=N+1}^{\infty} a_l \langle R_l, R_k\rangle_N\right) R_k(x) = \\ &= \sum_{l=N+1}^{\infty} a_l \underbrace{\sum_{k=0}^{N} \underbrace{\frac{\langle R_l, R_k\rangle_N}{\|R_k\|_N^2}}_{k\text{th coefficient of interpol. of } R_l} R_k(x)}_{=\mathcal{I}_N R_l}.\end{aligned}$$

For multivariate polynomials the extension is not straight forward. Considering the two-dimensional case shall demonstrate the general strategy. From (3.55) we have that

$$\begin{aligned}b_{k_1k_2} &= \frac{1}{\|R_{k_1}\|_N^2\|R_{k_2}\|_N^2}\sum_{j_1=0}^{N}{}''\sum_{j_2=0}^{N}{}'' f(x_{1_{j_1}}, x_{2_{j_2}})R_{k_1}(x_{1_{j_1}})R_{k_2}(x_{2_{j_2}}) = \\ &= \frac{1}{\|R_{k_1}\|_N^2\|R_{k_2}\|_N^2}\sum_{j_1=0}^{N}{}''\sum_{j_2=0}^{N}{}''\sum_{l_2=0}^{\infty}\sum_{l_1=0}^{\infty} a_{l_1l_2}R_{l_1}(x_{1_{j_1}})R_{l_2}(x_{2_{j_2}})R_{k_1}(x_{1_{j_1}})R_{k_2}(x_{2_{j_2}}) = \\ &= \frac{1}{\|R_{k_1}\|_N^2\|R_{k_2}\|_N^2}\sum_{l_2=0}^{\infty}\sum_{l_1=0}^{\infty} a_{l_1l_2}\underbrace{\sum_{j_1=0}^{N}{}''\sum_{j_2=0}^{N}{}'' R_{l_1}(x_{1_{j_1}})R_{l_2}(x_{2_{j_2}})R_{k_1}(x_{1_{j_1}})R_{k_2}(x_{2_{j_2}})}_{\langle R_{l_1}, R_{k_1}\rangle_N\langle R_{l_2}, R_{k_2}\rangle_N},\end{aligned}$$

then splitting the summation at N to apply the orthogonality gives

$$\begin{aligned}b_{k_1k_2} = \frac{1}{\|R_{k_1}\|_N^2\|R_{k_2}\|_N^2}\Bigg[&\underbrace{\sum_{l_2=0}^{N}\sum_{l_1=0}^{N} a_{l_1l_2}\langle R_{l_1}, R_{k_1}\rangle_N\langle R_{l_2}, R_{k_2}\rangle_N}_{=a_{k_1k_2}\|R_{k_1}\|_N^2\|R_{k_2}\|_N^2} + \\ &+ \underbrace{\sum_{l_2=N+1}^{\infty}\sum_{l_1=0}^{N} a_{l_1l_2}\langle R_{l_1}, R_{k_1}\rangle_N\langle R_{l_2}, R_{k_2}\rangle_N}_{=\sum_{l_2} a_{k_1l_2}\|R_{k_1}\|_N^2\langle R_{l_2}, R_{k_2}\rangle_N} + \end{aligned} \tag{3.61}$$

$$+\underbrace{\sum_{l_2=0}^{N}\sum_{l_1=N+1}^{\infty} a_{l_1 l_2}\langle R_{l_1}, R_{k_1}\rangle_N \langle R_{l_2}, R_{k_2}\rangle_N}_{=\sum_{l_1} a_{l_1 k_2}\|R_{k_2}\|_N^2\langle R_{l_1}, R_{k_1}\rangle_N} +$$
$$+\sum_{l_2=N+1}^{\infty}\sum_{l_1=N+1}^{\infty} a_{l_1 l_2}\langle R_{l_1}, R_{k_1}\rangle_N \langle R_{l_2}, R_{k_2}\rangle_N\Bigg]$$

and finally

$$\begin{aligned} b_{k_1 k_2} - a_{k_1 k_2} = &\frac{1}{\|R_{k_2}\|_N^2}\sum_{l_2=N+1}^{\infty} a_{k_1 l_2}\langle R_{l_2}, R_{k_2}\rangle_N + \\ &+\frac{1}{\|R_{k_1}\|_N^2}\sum_{l_1=N+1}^{\infty} a_{l_1 k_2}\langle R_{l_1}, R_{k_1}\rangle_N + \\ &+\frac{1}{\|R_{k_1}\|_N^2\|R_{k_2}\|_N^2}\sum_{l_2=N+1}^{\infty}\sum_{l_1=N+1}^{\infty} a_{l_1 l_2}\langle R_{l_1}, R_{k_1}\rangle_N \langle R_{l_2}, R_{k_2}\rangle_N. \end{aligned} \tag{3.62}$$

By comparison with (3.59) the above deduction reveals the obvious way to extensions in higher dimensions, i.e. the right-hand side sums up all possible combinations of k_i and l_i in the multi-index of the coefficient a, such that for every l_i a sum runs from $N+1$ to infinity keeping the inner product, whereas for every k_i a norm term can be canceled out in the overall factor $\prod 1/\|R_{k_i}\|_N^2$.

To put this in a more compact form we state

Lemma 3.20. *Say $m := (m_1, \ldots, m_n)$ is a multi-index taking the values $m_i \in \{l_i, k_i\}$. Given the coefficients a_l and b_k ($|k| \leq N$) from the projection operators defined in* (3.14) *and* (3.55) *and a bounded and (piecewise) continuous function f on $\mathbb{R}^n$, b_k can be written as*

$$b_k = \sum_m \Bigg(\prod_{\substack{i=1 \\ m_i \neq k_i}}^{n} \frac{1}{\|R_{m_i=k_i}\|_N^2} \sum_{\substack{i \\ m_i=N+1 \\ m_i=l_i}}^{\infty} a_m \prod_{\substack{i=1 \\ m_i=l_i}}^{n} \langle R_{m_i=l_i}, R_{m_i=k_i}\rangle_N \Bigg). \tag{3.63}$$

Remark 3.26. The first sum in the result above is taken over all possible m and the subscript i in the second sum symbolizes multiple summations depending on how many entries m_i equal l_i. The requirement of (piecewise) continuity can be replaced by assuming smoothness in the sense of $H^1_{u,A}$ (or similar, as mentioned in Hesthaven et al. (2007)) for the sums in (3.61) to be sufficiently convergent.

For the sake of clarity and comparability the three-dimensional case shall be written down in detail. Starting from the multi-index $m = (m_1, m_2, m_3)$ with $m_1 \in \{k_1, l_1\}$, $m_2 \in \{k_2, l_2\}$ and $m_3 \in \{k_3, l_3\}$, such that all possible combinations are

$$m \in \left\{ \begin{array}{llll} (k_1,k_2,k_3), & (l_1,k_2,k_3), & (k_1,l_2,k_3), & (k_1,k_2,l_3), \\ (l_1,l_2,k_3), & (l_1,k_2,l_3), & (k_1,l_2,l_3), & (l_1,l_2,l_3) \end{array} \right\}. \tag{3.64}$$

Obviously, for all multiple dimensions, there is always one combination, such that $a_m = a_k$ and since in that particular case there is no $m_i = l_i$ no sum and no norm factor appears, which means one can write $b_k - a_k$ on the left-hand side (indicating that the resulting right-hand side contains the coefficients of the aliasing error $\mathcal{A}_N f$ given only in terms of a_l).

Overall, the three-dimensional version can be obtained to be

$$
\begin{aligned}
b_{k_1k_2k_3} - a_{k_1k_2k_3} &= \frac{1}{\|R_{k_1}\|_N^2} \sum_{l_1=N+1}^{\infty} a_{l_1k_2k_3} \langle R_{l_1}, R_{k_1} \rangle_N + \\
&+ \frac{1}{\|R_{k_2}\|_N^2} \sum_{l_2=N+1}^{\infty} a_{k_1l_2k_3} \langle R_{l_2}, R_{k_2} \rangle_N + \frac{1}{\|R_{k_3}\|_N^2} \sum_{l_3=N+1}^{\infty} a_{k_1k_2l_3} \langle R_{l_3}, R_{k_3} \rangle_N + \\
&+ \frac{1}{\|R_{k_1}\|_N^2 \|R_{k_2}\|_N^2} \sum_{l_2=N+1}^{\infty} \sum_{l_1=N+1}^{\infty} a_{l_1l_2k_3} \langle R_{l_1}, R_{k_1} \rangle_N \langle R_{l_2}, R_{k_2} \rangle_N + \\
&+ \frac{1}{\|R_{k_1}\|_N^2 \|R_{k_3}\|_N^2} \sum_{l_3=N+1}^{\infty} \sum_{l_1=N+1}^{\infty} a_{l_1k_2l_3} \langle R_{l_1}, R_{k_1} \rangle_N \langle R_{l_3}, R_{k_3} \rangle_N + \\
&+ \frac{1}{\|R_{k_2}\|_N^2 \|R_{k_3}\|_N^2} \sum_{l_3=N+1}^{\infty} \sum_{l_2=N+1}^{\infty} a_{k_1l_2l_3} \langle R_{l_2}, R_{k_2} \rangle_N \langle R_{l_3}, R_{k_3} \rangle_N + \\
&+ \frac{1}{\|R_{k_1}\|_N^2 \|R_{k_2}\|_N^2 \|R_{k_3}\|_N^2} \sum_{l_3=N+1}^{\infty} \sum_{l_2=N+1}^{\infty} \sum_{l_1=N+1}^{\infty} a_{l_1l_2l_3} \langle R_{l_1}, R_{k_1} \rangle_N \langle R_{l_2}, R_{k_2} \rangle_N \langle R_{l_3}, R_{k_3} \rangle_N .
\end{aligned}
$$

Note, most importantly, that for a fixed k the first three terms on the right-hand side above are equal to the one-dimensional version (3.59).

Another possibility to describe $b_k - a_k$ from (3.63) can be obtained combining two results presented in Mason & Handscomb (2003), i.e.

$$
\sum_{j=0}^{N}{}'' \cos(j\theta) = \frac{1}{2} \frac{\sin(N\theta)}{\tan(\theta/2)}, \qquad \cos(l\theta)\cos(k\theta) = \frac{1}{2}[\cos((l+k)\theta) + \cos((l-k)\theta)],
$$

such that

$$
\sum_{j=0}^{N}{}'' R_l(x_j) R_k(x_j) = \sum_{j=0}^{N}{}'' \cos\left(\frac{jl\pi}{N}\right) \cos\left(\frac{jk\pi}{N}\right) = \tfrac{1}{4} \left[\frac{\sin((l+k)\pi)}{\tan\left(\frac{(l+k)\pi}{2N}\right)} + \frac{\sin((l-k)\pi)}{\tan\left(\frac{(l-k)\pi}{2N}\right)} \right]
$$

and with $l > N$ and $k \leq N$ (as given in (3.63)) one finds

$$
\langle R_l, R_k \rangle_N = \sum_{j=0}^{N}{}'' R_l(x_j) R_k(x_j) = \begin{cases} \frac{N}{2} & l + k = 2Np \\ \frac{N}{2} & l - k = 2Np \\ 0 & otherwise \end{cases} \qquad \text{for } p = 0, 1, 2, \ldots \tag{3.65}
$$

resulting in the (one-dimensional) description

$$b_k - a_k = \frac{N}{2\|R_k\|_N^2} \sum_{p=1}^{\infty} (a_{2Np-k} + a_{2Np+k}),$$

see also Hesthaven et al. (2007). Again, for the multivariate analogue, we first state the result in two dimensions, cf. (3.62),

$$\begin{aligned} b_{k_1k_2} - a_{k_1k_2} = & \\ & + \frac{N}{2\|R_{k_2}\|_N^2} \sum_{p_2=1}^{\infty} (a_{k_1,2Np_2-k_2} + a_{k_1,2Np_2+k_2}) + \\ & + \frac{N}{2\|R_{k_1}\|_N^2} \sum_{p_1=1}^{\infty} (a_{2Np_1-k_1,k_2} + a_{2Np_1+k_1,k_2}) + \\ & + \frac{N^2}{4\|R_{k_1}\|_N^2 \|R_{k_2}\|_N^2} \sum_{p_2=1}^{\infty} \sum_{p_1=1}^{\infty} (a_{2Np_1+k_1,2Np_2+k_2} + a_{2Np_1-k_1,2Np_2+k_2} + \\ & \qquad + a_{2Np_1+k_1,2Np_2-k_2} + a_{2Np_1-k_1,2Np_2-k_2}), \end{aligned}$$

where one can easily see that every l_i in (3.62) is switched with $2Np_i \pm k_i$ and the sums are taken over p_i. From the "quasi-orthogonality" in (3.65) every inner product is replaced by the factor $N/2$ and combining that with the actual orthogonality (3.53) gives an overall factor $1/\prod c_k$ with $c_{k_i} = 2$ if $k_i \in \{0, N\}$ and $c_{k_i} = 1$ otherwise, which eventually shows

Lemma 3.21. *Given the same assumptions as in Lemma 3.20 the coefficients of the aliasing error can be written as*

$$b_k - a_k = \sum_{\substack{m \\ m \neq (k_1,\dots,k_n)}} \left(\prod_{\substack{i=1 \\ m_i \neq k_i}}^{n} \frac{1}{c_{m_i=k_i}} \sum_{\substack{p_i=1 \\ m_i=l_i}}^{\infty_i} \sum_{q \in M} a_q \right), \tag{3.66}$$

where in the last sum the set M contains all possible combinations of l_i being replaced by $\{2Np_i - k_i, 2Np_i + k_i\}$ in m.

Using the L^2 norm to measure the aliasing error the following result provides the convergence rate in terms of N.

Theorem 3.22. *Let $f \in H^r_{u,A}(\mathbb{R}^n)$ with $r/n > 1/2$, then there exists a constant c independent of N, such that the aliasing error satisfies*

$$\|\mathcal{A}_N f\|_u \leq c N^{-r/n} \|f\|_A.$$

Proof. Using Parseval's identity with $\mathcal{A}_N f = \sum_k^N \gamma_k R_k$ yields

$$\|\mathcal{A}_N f\|_u^2 = \sum_{|k|=0}^{N} |\gamma_k|^2 \|R_k\|_u^2 \leq const. \sum_{|k|=0}^{N} |\gamma_k|^2,$$

where $|\gamma_k|^2 = |b_k - a_k|^2$. In virtue of (3.64) one can enumerate all possible multi-index values of m by j, such that substituting (3.66) yields

$$|\gamma_k|^2 = \Big| \sum_{\substack{m \\ m\neq(k_1,\dots,k_n)}} \Big(\underbrace{\prod_{\substack{i=1 \\ m_i\neq k_i}}^{n} \frac{1}{c_{m_i=k_i}}}_{=:c_j} \underbrace{\sum_{\substack{p_i=1}}^{\infty}{}_i \sum_{q\in M} a_q}_{\substack{m_i=l_i \\ =:d_j}} \Big) \Big|^2 \leq \Big(\sum_{j=1}^{2^n-1} c_j \, |d_j| \Big)^2, \tag{3.67}$$

where a_m are the coefficients of the expansion of f. Since $f \in H^r_{u,A}$ equation (3.34) holds, by which one can claim $A_x^{r/2n} f$ to possess an expansion into R_l with the coefficients

$$g_l := \frac{\langle A_x^{\frac{r}{2n}} f, R_l \rangle_u}{\|R_l\|_u^2} \quad \Rightarrow \quad a_m = \prod_{i=1}^{n} \frac{1}{m_i^{\frac{r}{n}}} g_m. \tag{3.68}$$

Starting from $m = (k_1, \dots, k_n)$ (which is excluded in the sum) we can subdivide the d_j into groups depending on how many k_i are replaced by l_i (cf. (3.64)) and hence how many sums over p_i are involved. For the sake of readability say $\alpha_i := 2Np_i - k_i$ and $\beta_i := 2Np_i + k_i$, then all d_j with one $k_i \leftrightarrow l_i$ $(i = 1, \dots, n)$ read

$$\begin{aligned} |d_{j,1}| &= \Big| \sum_{p_i=1}^{\infty} (a_{(k_1,\dots,\alpha_i,\dots,k_n)} + a_{(k_1,\dots,\beta_i,\dots,k_n)}) \Big| = \\ &= \Big| \sum_{p_i=1}^{\infty} \big((k_1 \cdots \alpha_i \cdots k_n)^{-\frac{r}{n}} g_{(k_1,\dots,\alpha_i,\dots,k_n)} + (k_1 \cdots \beta_i \cdots k_n)^{-\frac{r}{n}} g_{(k_1,\dots,\beta_i,\dots,k_n)} \big) \Big| \leq \\ &\leq \Big| \sum_{p_i=1}^{\infty} (k_1 \cdots \alpha_i \cdots k_n)^{-\frac{r}{n}} g_{(k_1,\dots,\alpha_i,\dots,k_n)} \Big| + \Big| \sum_{p_i=1}^{\infty} (k_1 \cdots \beta_i \cdots k_n)^{-\frac{r}{n}} g_{(k_1,\dots,\beta_i,\dots,k_n)} \Big|. \end{aligned} \tag{3.69}$$

In the case of two exchanges $k_i \leftrightarrow l_i$ and $k_j \leftrightarrow l_j$ $(i, j = 1, \dots, n)$ one has

$$\begin{aligned} |d_{j,2}| \leq & \Big| \sum_{p_i=1}^{\infty} \sum_{p_j=1}^{\infty} (k_1 \cdots \alpha_i \alpha_j \cdots k_n)^{-\frac{r}{n}} g_{(k_1,\dots,\alpha_i,\dots,\alpha_j,\dots,k_n)} \Big| + \\ & + \Big| \sum_{p_i=1}^{\infty} \sum_{p_j=1}^{\infty} (k_1 \cdots \alpha_i \beta_j \cdots k_n)^{-\frac{r}{n}} g_{(k_1,\dots,\alpha_i,\dots,\beta_j,\dots,k_n)} \Big| + \\ & + \Big| \sum_{p_i=1}^{\infty} \sum_{p_j=1}^{\infty} (k_1 \cdots \beta_i \alpha_j \cdots k_n)^{-\frac{r}{n}} g_{(k_1,\dots,\beta_i,\dots,\alpha_j,\dots,k_n)} \Big| + \\ & + \Big| \sum_{p_i=1}^{\infty} \sum_{p_j=1}^{\infty} (k_1 \cdots \beta_i \beta_j \cdots k_n)^{-\frac{r}{n}} g_{(k_1,\dots,\beta_i,\dots,\beta_j,\dots,k_n)} \Big|. \end{aligned} \tag{3.70}$$

Such formulae immediately become more complicated when replacing more than one k_i and since it is obvious how to proceed, the remaining d_j shall not be written down explicitly.
Next Cauchy's inequality can be applied to the right-hand side terms in (3.69) to modify the

estimate to

$$|d_{j,1}| \leq \Big(\sum_{p_i=1}^{\infty} |k_1 \cdots \alpha_i \cdots k_n|^{-\frac{2r}{n}}\Big)^{\frac{1}{2}} \Big(\sum_{p_i=1}^{\infty} |g_{(k_1,\ldots,\alpha_i,\ldots,k_n)}|^2\Big)^{\frac{1}{2}} + \\ + \Big(\sum_{p_i=1}^{\infty} |k_1 \cdots \beta_i \cdots k_n|^{-\frac{2r}{n}}\Big)^{\frac{1}{2}} \Big(\sum_{p_i=1}^{\infty} |g_{(k_1,\ldots,\beta_i,\ldots,k_n)}|^2\Big)^{\frac{1}{2}},$$

which can be performed in the same way for (3.70) and all other d_j. Since $k_i \leq N$ it follows that α_i, β_i are non-negative and thus $\alpha_i \leq \beta_i$, such that

$$|d_{j,1}| \leq \Big(\sum_{p_i=1}^{\infty} |k_1 \cdots \alpha_i \cdots k_n|^{-\frac{2r}{n}}\Big)^{\frac{1}{2}} \times \\ \times \left[\Big(\sum_{p_i=1}^{\infty} |g_{(k_1,\ldots,\alpha_i,\ldots,k_n)}|^2\Big)^{\frac{1}{2}} + \Big(\sum_{p_i=1}^{\infty} |g_{(k_1,\ldots,\beta_i,\ldots,k_n)}|^2\Big)^{\frac{1}{2}}\right]. \tag{3.71}$$

Assuming one has applied Cauchy's inequality to the four addends in (3.70) it is easy to see that the factor containing $\alpha_i\alpha_j$ is greater or equal to the other three, yielding

$$|d_{j,2}| \leq \Big(\sum_{p_i=1}^{\infty}\sum_{p_j=1}^{\infty} |k_1 \cdots \alpha_i\alpha_j \cdots k_n|^{-\frac{2r}{n}}\Big)^{\frac{1}{2}} \times \\ \times \left[\Big(\sum_{p_i=1}^{\infty}\sum_{p_j=1}^{\infty} |g_{(k_1,\ldots,\alpha_i,\ldots,\alpha_j,\ldots,k_n)}|^2\Big)^{\frac{1}{2}} + \cdots + \Big(\sum_{p_i=1}^{\infty}\sum_{p_j=1}^{\infty} |g_{(k_1,\ldots,\beta_i,\ldots,\beta_j,\ldots,k_n)}|^2\Big)^{\frac{1}{2}}\right] \tag{3.72}$$

and, analogously, estimates for all other d_j have a similar leading factor with all appearing l_i replaced α_i.
Such factors with $k_i \leq N$ allow for a further modification, i.e.

$$\Big(\sum_{p_i=1}^{\infty} |k_1 \cdots \alpha_i \cdots k_n|^{-\frac{2r}{n}}\Big)^{\frac{1}{2}} = \underbrace{\prod_{j\neq i} \frac{1}{k_j^{\frac{r}{n}}}}_{\leq 1} \Big(\sum_{p_i=1}^{\infty} |2Np_i - k_i|^{-\frac{2r}{n}}\Big)^{\frac{1}{2}} \leq \\ \leq \frac{1}{N^{\frac{r}{n}}} \Big(\sum_{p_i=1}^{\infty} \underbrace{|2p_i - k_i/N|^{-\frac{2r}{n}}}_{\leq |2p_i-1|^{-\frac{2r}{n}}}\Big)^{\frac{1}{2}} \leq \frac{1}{N^{\frac{r}{n}}} \Big(\sum_{p_i=1}^{\infty} |2p_i - 1|^{-\frac{2r}{n}}\Big)^{\frac{1}{2}} = const. \frac{1}{N^{\frac{r}{n}}}, \tag{3.73}$$

where the constant in the last term stems from calculating the value of the infinite sum, which converges provided $r/n > 1/2$ (as found in Hesthaven et al. (2007) in the case of one-dimensional Fourier expansions). Considering the factor in (3.72) one can observe that

$$\Big(\sum_{p_i=1}^{\infty}\sum_{p_j=1}^{\infty} |k_1 \cdots \alpha_i\alpha_j \cdots k_n|^{-\frac{2r}{n}}\Big)^{\frac{1}{2}} = \\ = \prod_{l\neq i,j} \frac{1}{k_l^{\frac{r}{n}}} \Big(\sum_{p_i=1}^{\infty} |\alpha_i|^{-\frac{2r}{n}} \sum_{p_j=1}^{\infty} |\alpha_j|^{-\frac{2r}{n}}\Big)^{\frac{1}{2}} \leq const. \left(\frac{1}{N^{\frac{r}{n}}}\right)^2, \tag{3.74}$$

by using (3.73) for the individual sums, which is then done for all other such factors in the remaining d_j.
Collecting now all results from (3.68) to (3.74) and substitute them into (3.67) one obtains

$$\begin{aligned}\Big(\sum_{j=1}^{2^n-1} c_j\,|d_j|\Big)^2 \le const.\Big[&\frac{1}{N^{\frac{r}{n}}}\sum_{\substack{m\\ |\{m_i=l_i\}|=1}}\sum_{l_i\in M}\Big(\sum_{\substack{i\\ p_i=1\\ m_i=l_i}}^{\infty}|g_m|^2\Big)^{\frac12}+\\ &+\Big(\frac{1}{N^{\frac{r}{n}}}\Big)^2\sum_{\substack{m\\ |\{m_i=l_i\}|=2}}\sum_{l_i\in M}\Big(\sum_{\substack{i\\ p_i=1\\ m_i=l_i}}^{\infty}|g_m|^2\Big)^{\frac12}+\\ &\vdots\\ &+\frac{1}{N^r}\sum_{l_i\in M}\Big(\sum_{p_n=1}^{\infty}\cdots\sum_{p_1=1}^{\infty}|g_l|^2\Big)^{\frac12}\Big]^2\end{aligned}\tag{3.75}$$

and with $1/N^{r/n}$ being greater or equal to all other leading factors, the right-hand side can be modified to yield

$$\|\mathcal{A}_N f\|_u^2 \le const.\sum_{|k|=0}^{N}|\gamma_k|^2 \le const.\Big(\frac{1}{N^{\frac{r}{n}}}\Big)^2\sum_{|k|=0}^{N}\Big[\ldots\Big]^2.\tag{3.76}$$

The square brackets, which shall be estimated in the following, contain all the sums from above but without their leading factors. By Parseval's identity we have that

$$\sum_{|l|=0}^{\infty}|g_l|^2 \le const.\|A_x^{r/(2n)}f\|_u^2,\tag{3.77}$$

such that the sum converges and consequently so does every partial sum. Starting again with the sums over p_i given in (3.71) and renaming them to be

$$g_k^{(1)} := \sum_{p_i=1}^{\infty}|g_{(k_1,\ldots,2Np_i-k_i,\ldots,k_n)}|^2, \qquad g_k^{(2)} := \sum_{p_i=1}^{\infty}|g_{(k_1,\ldots,2Np_i+k_i,\ldots,k_n)}|^2$$

(where the superscripts just symbolize an (arbitrarily chosen) enumeration), which can be viewed as partial sums of (3.77), which are therefore convergent. In a similar manner the terms in (3.72) can also be renamed (and viewed again as partial sums). By enumerating all these partial sums of (3.77) in all d_j, estimate (3.76) reads

$$\|\mathcal{A}_N f\|_u^2 \le const.\Big(\frac{1}{N^{\frac{r}{n}}}\Big)^2\sum_{|k|=0}^{N}\Big[\sum_{j=1}^{b}\big(g_k^{(j)}\big)^{1/2}\Big]^2,$$

with $b = |\{m : m_i = k_i \text{ or } m_i = 2Np_i - k_i \text{ or } m_i = 2Np_i + k_i,\ \forall i\}|$. Rewriting the sum on the right-hand side as

$$\begin{aligned}\sum_{|k|=0}^{N}\left[\sum_{j=1}^{b}\left(g_k^{(j)}\right)^{1/2}\right]^2 &= \sum_{|k|=0}^{N}\sum_{j=1}^{b}\sum_{i=1}^{b}\left(g_k^{(i)}g_k^{(j)}\right)^{1/2} \leq \sum_{|k|=0}^{N}\sum_{j=1}^{b}\sum_{i=1}^{b}\left(g_k^{(i)}+g_k^{(j)}\right) = \\ = 2b\sum_{|k|=0}^{N}\sum_{j=1}^{b}g_k^{(j)} &= 2b\sum_{j=1}^{b}\sum_{|k|=0}^{N}g_k^{(j)} \leq const.\sum_{j=1}^{b}\|A_x^{r/2n}f\|_u^2 \leq const.\|f\|_A^2,\end{aligned} \tag{3.78}$$

using Lemma 3.10 for the last inequality, substituting this into the estimate above and taking the square root finishes the proof. □

Remark 3.27. Comparing Theorems 3.11 and 3.22 it has been shown that the L^2 error made by *truncating* a rational Chebyshev series expansion using exact coefficients is of the *same order* as the error made by approximating coefficients using *interpolation*. This proves parts of the assumption made in Boyd (1987) and Boyd (2001) about the truncation error being of the same order of magnitude as the "discretization" error (the difference between the exact and approximated coefficients obtained from solving differential or integral equations), which is similar to the aliasing error.

The overall error made by the discrete projection is bounded in L^2, which stands in complete accordance to Theorem 3.11.

Theorem 3.23. *Let $f \in H^r_{u,A}(\mathbb{R}^n)$, then there exists a positive constant c, such that*

$$\|f - \mathcal{I}_N f\|_u \leq cN^{-r/n}\|f\|_A.$$

Proof. Obviously, by adding and subtracting the continuous projection

$$\|f - \mathcal{I}_N f\|_u = \|f - \mathcal{I}_N f + \mathcal{P}_N f - \mathcal{P}_N f\|_u \leq \|f - \mathcal{P}_N f\|_u + \|\mathcal{P}_N f - \mathcal{I}_N f\|_u, \tag{3.79}$$

where Theorems 3.11 and 3.22 can be applied to arrive at the desired estimate. □

Remark 3.28. Finally, the error made by approximately calculating coefficients of a truncated series (the most common situation when solving operator equations with spectral methods) is at most twice the error made by the continuous projection. The one-dimensional case has been proved in Wang & Guo (2002) using a different approach (i.e. not via the aliasing error).

In Mason (1980) Theorem 3.13 was shown to hold in the same way for the multivariate interpolation operator at (a tensor product of) Chebyshev zeros by making use of Lagrange interpolation polynomials. Again, cf. Remark 3.19, the mapping ψ transforms the rational case into the classical case on compact intervals and thus Theorem 3.13 can also be applied to the discrete projection on $\mathbb{R}^n$, i.e. uniform convergence of $\mathcal{I}_N f$, provided f satisfies the Dini-Lipschitz condition (3.44).

Having a discrete inner product is not the only reason for choosing the interpolation points as in (3.53). The well-known *Runge phenomenon* occurs when using equidistant points. In other words, the problem of interpolating a given function using orthogonal polynomials at equidistant points is ill-posed, as demonstrated e.g. in Hesthaven et al. (2007), meaning that the more points are used, the larger the L^∞ error gets.

This can be seen by calculating the *Lebesgue constant* in the case of equally distributed interpolation points. Whereas in the continuous case and for an interpolation at Chebyshev points this constant grows logarithmically and can hence be bounded by the modulus of continuity (see Mason (1980)), in the case of equidistant interpolation the Lebesgue constant grows exponentially.

In view of the Lebesgue constant and the Runge phenomenon it is worth considering the uniform convergence rate for the interpolation operator. In Theorem 3.17 the convergence rate was presented when using the continuous projection. The next result shows that the same rate holds in the case of the discrete projection.

Theorem 3.24. *Given $f \in H^r_{u,A}(\mathbb{R}^n)$, $r/n > 1/2$, then there exists a constant c independent of N, such that*

$$\|f - \mathcal{I}_N f\|_\infty \leq c N^{1/2-r/n} \|f\|_A.$$

Proof. Since (3.79) holds in every norm, it is sufficient to prove that the aliasing error measured in the L^∞ norm is of the same order as the truncation error in (3.52). Using the same argument as in (3.45) leads to

$$\|\mathcal{A}_N f\|_\infty \leq \sum_{|k|=0}^{N} |\gamma_k| \leq \sum_{|k|=0}^{N} \sum_{j=1}^{2^n-1} c_j \, |d_j|, \tag{3.80}$$

cf. (3.67), whereas here the estimates from (3.69) and (3.70) shall be written as

$$|d_{j,1}| \leq \sum_{p_i=1}^{\infty} \left| a_{(k_1,\ldots,\alpha_i,\ldots,k_n)} \right| + \sum_{p_i=1}^{\infty} \left| a_{(k_1,\ldots,\beta_i,\ldots,k_n)} \right|$$

and analogously for all other d_j.
Changing the order of the sums on the right-hand side in (3.80) yields

$$\begin{aligned} \sum_{j=1}^{2^n-1} \sum_{|k|=0}^{N} c_j \, |d_j| \leq const. \Bigg[& \sum_{|\{m_i = l_i\}|=1}^{m} \left(\sum_{\substack{p_i=1 \\ l_i=\alpha_i}}^{\infty} \sum_{|k|=0}^{N} |a_m| + \sum_{\substack{p_i=1 \\ l_i=\beta_i}}^{\infty} \sum_{|k|=0}^{N} |a_m| \right) + \\ & + \sum_{|\{m_i = l_i\}|=2}^{m} \sum_{l_i \in M} \sum_{\substack{p_i=1 \\ m_i=l_i}}^{\infty}{}_{i} \sum_{|k|=0}^{N} |a_m| + \\ & \vdots \end{aligned} \tag{3.81}$$

$$+\sum_{l_i\in M}\sum_{p_n=1}^{\infty}\cdots\sum_{p_1=1}^{\infty}\sum_{|k|=0}^{N}|a_l|\Big],$$

in contrast to (3.75), where, due to the overall square, the strategy had to be different (with applying Cauchy's inequality earlier) and it was hence not possible to move the sum over k up to the coefficient a_m or g_m. And here lies the main difference in the proof for the infinity norm. By considering $\sum_{p_i}\sum_{|k|}|a_m|$, where we start by changing one m_i to either α_i or β_i, then substitute (3.68) and use Cauchy's inequality we get

$$\begin{aligned}\sum_{p_i=1}^{\infty}\sum_{|k|=0}^{N}|a_m| &= \sum_{p_i=1}^{\infty}\sum_{|k|=0}^{N}\left|(k_1\cdots\alpha_i\cdots k_n)^{-\frac{r}{n}}g_{(k_1,\ldots,\alpha_i,\ldots,k_n)}\right| \leq \\ &\leq \left(\sum_{p_i=1}^{\infty}\sum_{|k|=0}^{N}\left|(k_1\cdots\alpha_i\cdots k_n)^{-\frac{2r}{n}}\right|\right)^{\frac{1}{2}}\left(\sum_{p_i=1}^{\infty}\sum_{|k|=0}^{N}\left|g_{(k_1,\ldots,\alpha_i,\ldots,k_n)}\right|^2\right)^{\frac{1}{2}}.\end{aligned}$$

The second term is bounded by $\|f\|_A$, as seen from (3.78) by taking the square root, whereas for the first term one can derive

$$\begin{aligned}&\sum_{p_i=1}^{\infty}\sum_{|k|=0}^{N}\left|(k_1\cdots\alpha_i\cdots k_n)^{-\frac{2r}{n}}\right| = N^{-\frac{2r}{n}}\sum_{p_i=1}^{\infty}\underbrace{\prod_{j\neq i}\sum_{k_j=1}^{N}\left|\frac{1}{k_j}\right|^{\frac{2r}{n}}}_{\leq const.\ \forall N}\underbrace{\sum_{k_i=0}^{N}\left|2p_i-\frac{k_i}{N}\right|^{-\frac{2r}{n}}}_{\leq N\,|2p_i-1|^{-\frac{2r}{n}}} \leq \\ &\leq const.\,N^{-\frac{2r}{n}+1}\underbrace{\sum_{p_i=1}^{\infty}\left|2p_i-1\right|^{-\frac{2r}{n}}}_{\text{cf. (3.73)}} \leq const.\,N^{-\frac{2r}{n}+1},\end{aligned}$$

where the sums converge if $r/n > 1/2$. For the issue of $k_i = 0$ we refer to the proof of Theorem 3.11 above. Obviously, the same can be done for $m_i = \beta_i$, without any changes, such that the first line in (3.81) is bounded by $cN^{\frac{1}{2}-\frac{r}{n}}\|f\|_A$.
Writing in detail the case of two m_i switched with α_i or β_i shows the second line (excluding the sum over all possible exchanges) in (3.81) to be

$$\begin{aligned}&\sum_{p_i=1}^{\infty}\sum_{p_j=1}^{\infty}\sum_{|k|=0}^{N}\left|a_{(k_1,\ldots,\alpha_i,\ldots,\alpha_j,\ldots,k_n)}\right| + \sum_{p_i=1}^{\infty}\sum_{p_j=1}^{\infty}\sum_{|k|=0}^{N}\left|a_{(k_1,\ldots,\alpha_i,\ldots,\beta_j,\ldots,k_n)}\right| + \\ &+\sum_{p_i=1}^{\infty}\sum_{p_j=1}^{\infty}\sum_{|k|=0}^{N}\left|a_{(k_1,\ldots,\beta_i,\ldots,\alpha_j,\ldots,k_n)}\right| + \sum_{p_i=1}^{\infty}\sum_{p_j=1}^{\infty}\sum_{|k|=0}^{N}\left|a_{(k_1,\ldots,\beta_i,\ldots,\beta_j,\ldots,k_n)}\right|.\end{aligned}$$

Without loss of generality we take the first sum and substitute again (3.68) to get

$$\sum_{p_i=1}^{\infty}\sum_{p_j=1}^{\infty}\sum_{|k|=0}^{N}\left|(k_1\cdots\alpha_i\alpha_j\cdots k_n)^{-\frac{r}{n}}g_{(k_1,\ldots,\alpha_i,\ldots,\alpha_j,\ldots,k_n)}\right| \leq$$
$$\leq \left(\sum_{p_i=1}^{\infty}\sum_{p_j=1}^{\infty}\sum_{|k|=0}^{N}\left|(k_1\cdots\alpha_i\alpha_j\cdots k_n)^{-\frac{2r}{n}}\right|\right)^{\frac{1}{2}}\left(\sum_{p_i=1}^{\infty}\sum_{p_j=1}^{\infty}\sum_{|k|=0}^{N}\left|g_{(k_1,\ldots,\alpha_i,\ldots,\alpha_j,\ldots,k_n)}\right|^2\right)^{\frac{1}{2}}.$$

Again, the second term is bounded by the $H^r_{u,A}$ norm of f and by splitting the first term one obtains

$$\sum_{p_i=1}^{\infty}\sum_{p_j=1}^{\infty}\sum_{|k|=0}^{N}\left|(k_1\cdots\alpha_i\alpha_j\cdots k_n)^{-\frac{2r}{n}}\right| =$$
$$= \left(N^{-\frac{2r}{n}}\right)^2\sum_{p_i=1}^{\infty}\sum_{p_j=1}^{\infty}\underbrace{\prod_{l\neq i,j}\sum_{k_l=1}^{N}\left|\frac{1}{k_l}\right|^{\frac{2r}{n}}}_{\leq const.\ \forall N}\underbrace{\sum_{k_i=0}^{N}\left|2p_i-\frac{k_i}{N}\right|^{-\frac{2r}{n}}}_{\leq N\,|2p_i-1|^{-\frac{2r}{n}}}\underbrace{\sum_{k_j=0}^{N}\left|2p_j-\frac{k_j}{N}\right|^{-\frac{2r}{n}}}_{\leq N\,|2p_j-1|^{-\frac{2r}{n}}} \leq$$
$$\leq const.\left(N^{-\frac{2r}{n}+1}\right)^2\sum_{p_i=1}^{\infty}|2p_i-1|^{-\frac{2r}{n}}\sum_{p_j=1}^{\infty}|2p_j-1|^{-\frac{2r}{n}} \leq const.\left(N^{-\frac{2r}{n}+1}\right)^2.$$

Consequently, by the same argument as before, the second line in (3.81) is bounded by $c\left(N^{\frac{1}{2}-\frac{r}{n}}\right)^2\|f\|_A$, which has a leading factor smaller than or equal to the one in the first line. It is now easy to see how to find bounds for every line with decreasing factors in terms of N. Thus, one can finally conclude

$$\|\mathcal{A}_N f\|_\infty \leq \sum_{j=1}^{2^n-1}\sum_{|k|=0}^{N} c_j\,|d_j| \leq const.\,N^{\frac{1}{2}-\frac{r}{n}}\|f\|_A.$$

□

As mentioned in Remark 3.24 there are other possibilities to define discrete inner products (and according norms) by changing the set of interpolation points. In the above we used the extrema of the last retained polynomial because of the connections to the fast Fourier transform. It can be easily shown that the proved convergence results can be established for other types of inner products as well, for example, when using the set of *zeros of the next higher order polynomial*.

The *Erdös-Turán* theorem shows L^2 convergence of the interpolation polynomial P_N formed by an arbitrary set of orthogonal polynomials $\{p_n\}$ interpolated at the zeros of p_{N+1} to a given continuous functions on a compact interval (see e.g. Cheney (1966) or Szegö (1939) for a proof). In our treatise such interpolation points will be used as *collocation points* (see Section 3.3).

Caveat: Lemma 3.2(ix) shows that the zeros of rational Chebyshev polynomials are by far not distributed over the *whole* real axis (in contrast to the classical case on $[-1,1]$). In

consequence, this means that when interpolating (or evaluating) functions at such points certain characteristics away from these points *cannot* be depicted or included, which might result in a (from the view of pointwise convergence) very poor approximation.

To illustrate the goodness of gaining coefficients by discrete inner products and what happens when changing the function away from the evaluation points, we consider the following

Example 3.7. Given the function $f(x) = (1+x^2)^{-1/2}$ (cf. Example 3.6) one can exactly calculate the continuous inner product to be

$$\|R_k\|_w^2 a_k = \left\langle (1+x^2)^{-\frac{1}{2}}, R_k \right\rangle_w = \frac{2}{1-k^2} \quad (k \text{ even}), \qquad x^2 f'(x) \to \mp 1 \ \text{ as } \ x \to \pm\infty,$$

and hence $|a_k| \propto 1/k^2$ (whereas according to Theorem 3.15 the power of k should be less than 2, but since f is even, the symmetry adds to a stronger decrease). Using the set of extrema of the Nth polynomial, i.e.

$$x_i = \frac{\cos(\frac{i\pi}{N})}{\sqrt{1-\cos^2(\frac{i\pi}{N})}} \quad \text{implies} \quad \left(Nb_0, \frac{N}{2}b_1, \ldots, \frac{N}{2}b_{N-1}, Nb_N \right) = \sqrt{N/2}DCT(f(x_i)),$$

cf. Lemma 3.19. In Table 16 the difference $|a_k - b_k|$ is given for the first 5 even coefficients for various N. Figure 38 depicts this result using $N = 10$, by showing the actual function

N	$\lvert a_0 - b_0\rvert$	$\lvert a_2 - b_2\rvert$	$\lvert a_4 - b_4\rvert$	$\lvert a_6 - b_6\rvert$	$\lvert a_8 - b_8\rvert$	$\lvert a_{10} - b_{10}\rvert$
10	5×10^{-3}	1×10^{-2}	1.1×10^{-2}	1.3×10^{-2}	1.5×10^{-2}	3×10^{-3}
50	2×10^{-4}	4.2×10^{-4}	4.2×10^{-4}	4.2×10^{-4}	4.2×10^{-4}	4.3×10^{-4}
100	5×10^{-5}	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}	1.1×10^{-4}
200	1×10^{-5}	2.6×10^{-5}	2.6×10^{-5}	2.6×10^{-5}	2.6×10^{-5}	2.6×10^{-5}

Table 16: Aliasing error of f, $|a_k - b_k|$ for various k and N

and the approximations on the left and the absolute value of the aliasing error as a function of x on the right. There, one can see the supremum of $\mathcal{A}_N f$ to be at infinity. This is due to Lemma 3.18, from which it follows that $\mathcal{I}_N f(\infty) = f(\infty) = 0$, whereas $\mathcal{P}_N f(\infty) = \sum^N a_k$, i.e. $\mathcal{P}_{10} f(\infty) = \frac{2}{11\pi} \approx 5.8 \times 10^{-2}$. This is in very well accordance to Theorem 3.24, where it was stated that $\|\mathcal{A}_N f\|_\infty \leq \sum^N |a_k - b_k| \leq cN^{1/2-s}$, reading in the case here as

$$\sup |\mathcal{A}_{10} f(x)| = \mathcal{A}_{10} f(\infty) = \sum_{k=0}^{10} = 6 \times 10^{-2} \leq const.\, 10^{1/2-2} = const.\, \frac{\sqrt{10}}{100} \approx const.\, 3 \times 10^{-2}.$$

Changing the given function to

$$\tilde{f}(x) = \begin{cases} (1+x^2)^{-1/2} & x < 5 \\ 0 & x \geq 5 \end{cases},$$

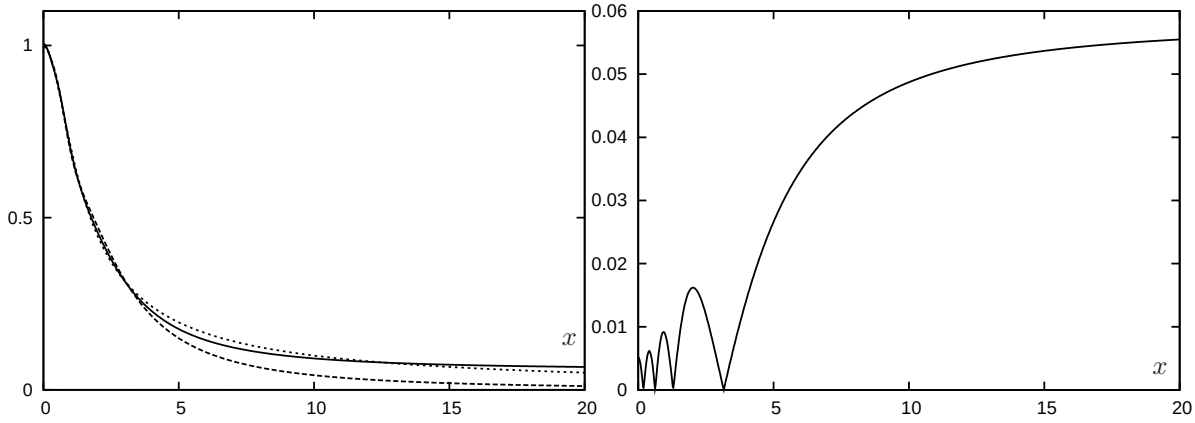

Figure 38: Left: $\mathcal{P}_{10}f$ (solid), $\mathcal{I}_{10}f$ (dashed) vs. f (dotted). Right: aliasing error $|\mathcal{A}_{10}f|$.

which has a discontinuity at $x = 5$, but still decays to zero at infinity, yields the exact same coefficients b_k as for f if $N \leq 15$, since $f \equiv \tilde{f}$ at the interpolation points x_i (discarding the points $|x_0| = |x_N| = \infty$ all other points lie in $(-5, 5)$ for $N \leq 15$).
The coefficients a_k can again be derived analytically for all k, such that subsequently a completely different aliasing error behavior occurs. (Observe, for example, that for $N \leq 15$ $\mathcal{I}_N f$ is still an even function). Figure 39 shows the approximation and the aliasing error for $\tilde{f}$ in the case of $N = 10$. From the aliasing error it is obvious that for $x < 5$ (away from the discontinuity) the approximation behaves analogously to Figure 38. Increasing the number of

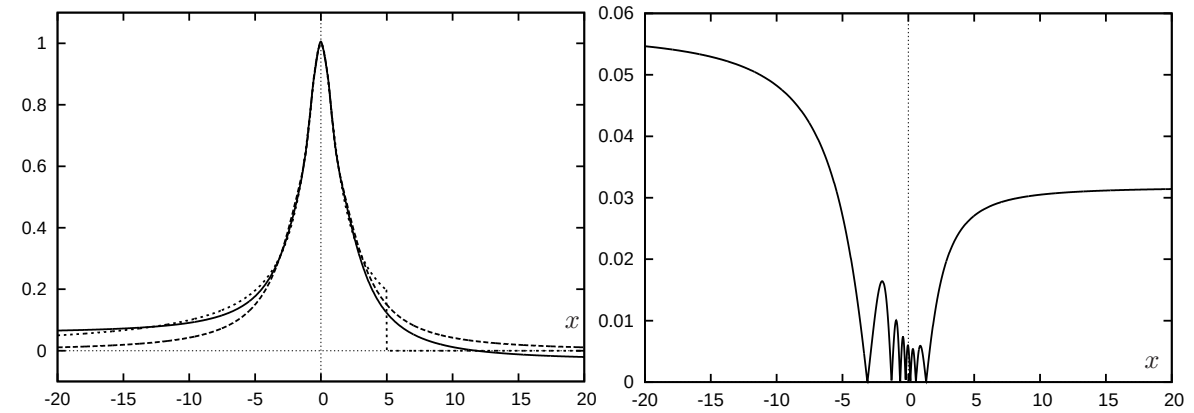

Figure 39: Left: $\mathcal{P}_{10}f$ (solid), $\mathcal{I}_{10}f$ (dashed) vs. $\tilde{f}$ (dotted). Right: aliasing error $|\mathcal{A}_{10}\tilde{f}|$.

polynomials the difference between f and $\tilde{f}$ becomes more apparent, resulting in a complete change of the aliasing error, especially for the point of its supremum. Figure 40 shows the details for $N = 100$, where one can see the aliasing to be maximal in the vicinity around the discontinuity, whereas everywhere else the original function is indistinguishable from its approximations. This gives rise to the treatment of a rather practical, but very important topic of polynomial approximation, the *Gibbs phenomenon*.

If one would continue to approximate $\tilde{f}$ from the example above with an increase in the number of polynomials used, the Gibbs oscillations become more and more visible around

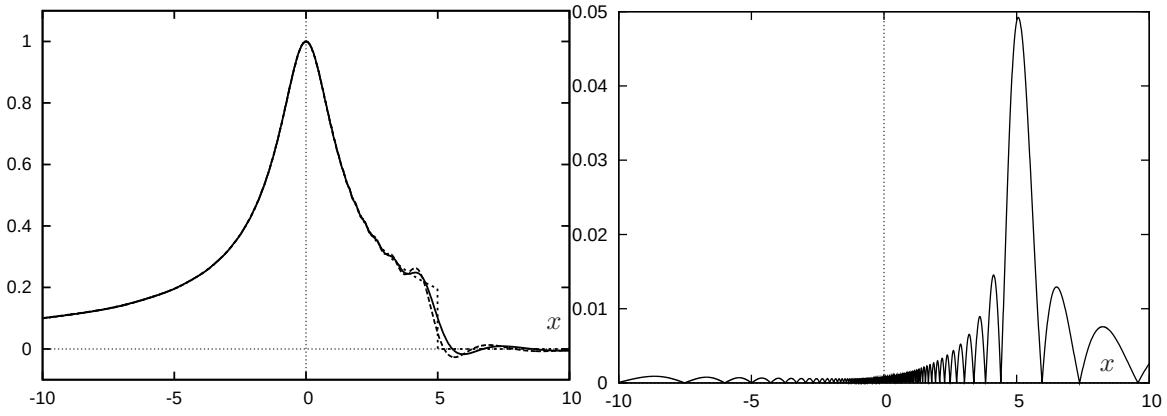

Figure 40: Left: $\mathcal{P}_{100}f$ (solid), $\mathcal{I}_{100}f$ (dashed) vs. $\tilde{f}$ (dotted). Right: aliasing error $|\mathcal{A}_{100}\tilde{f}|$.

the discontinuity step. Also, as mentioned in the proof of Lemma 3.12, the value of the approximation at the discontinuity is equal to the average step height, whereas the oscillations seem to increase the step. Thus, as mentioned in e.g. Mason & Handscomb (2003) the pointwise error at the discontinuity remains an O(1)-quantity as $N \to \infty$, cf. Figure 41. We

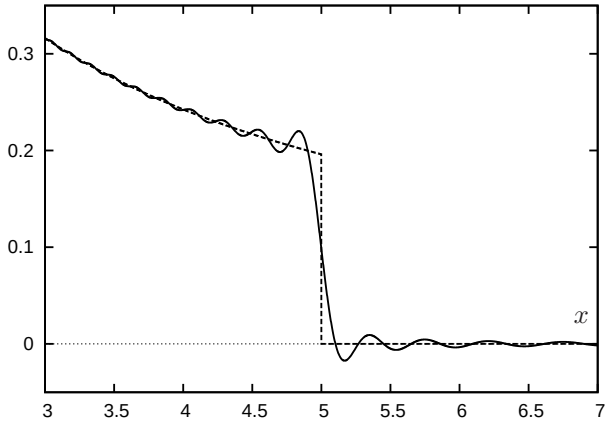

Figure 41: Figure 40 continued using $N = 500$: $\mathcal{P}_{500}f$ (solid) vs. $\tilde{f}$ (dashed)

shall not go further into the details of the phenomenon per se, but consider the aspect of approximating integrals of such discontinuous functions.

Performing numerical calculations, e.g. solving integral equations or applying Galerkin methods, actual values of integrals of polynomials with some kernel functions can be of interest, where the accuracy of the scheme could depend strongly on such values, and furthermore in most cases analytical expressions cannot be obtained and hence approximation techniques are needed. Here, some Sobolev-type convergence might not be sufficient and thus pointwise convergence has to be demanded.

Depicted in Figure 41 and inferred from Lemma 3.12 the approximation in the vicinity of a discontinuity differs strongly from its exact values, even for higher degree projections, meaning that the Gibbs phenomenon not only prevents convergence *at* the discontinuity, but

also slows down convergence near this point. Due to the fact that points of discontinuity have measure zero they do not contribute to the value of integrals and so pointwise and even uniform convergence can be shown to hold in such cases.

Proved in Hesthaven et al. (2007) and Guo (1998) for the case of classical Chebyshev polynomials, $\partial_x \mathcal{P}_N f \neq \mathcal{P}_N \partial_x f$ (but converges in L^2 as $N \to \infty$), the same holds in the case of integrals.

Lemma 3.25. *Given the functions f and g, sufficiently smooth and defined on $[-1,1]$, such that*

$$f(x) = \int g(x)\,dx + c \quad \text{and} \quad f_N(x) := \sum_{k=0}^{N} b_k T_k(x), \quad g_N(x) := \sum_{k=0}^{N} a_k T_k(x)$$

as projections of f and g, respectively, with $\{T_k\}$ being the set of classical Chebyshev polynomials, then

$$b_k = \begin{cases} a_0 - \frac{1}{2}a_2 & k = 1 \\ \dfrac{a_{k-1} - a_{k+1}}{2k} & 1 < k \leq N-1 \end{cases},$$

where b_0 remains undetermined (thus being the integration constant).

Proof. The fact that one can only equate the coefficients of f_N up to $N-1$ shows that $f_N \neq \int g_N$, where equality can only be attained if $N \to \infty$. To see this we start by writing

$$\int g_N(x)\,dx = \sum_{k=0}^{N} a_k \int T_k(x)\,dx = a_0 T_1(x) + \tfrac{1}{4}a_1 T_2(x) + \sum_{k=2}^{N} a_k \tfrac{1}{2}\left(\frac{T_{k+1}(x)}{k+1} - \frac{T_{k-1}(x)}{k-1}\right),$$

where (3.9) is applied to yield the second equality. By splitting the sum and shifting the indices, i.e.

$$\sum_{k=2}^{N} a_k \tfrac{1}{2}\left(\frac{T_{k+1}}{k+1} - \frac{T_{k-1}}{k-1}\right) = \sum_{k=3}^{N+1} a_{k-1}\tfrac{1}{2k}T_k - \sum_{k=1}^{N-1} a_{k+1}\tfrac{1}{2k}T_k,$$

one obtains

$$\int g_N(x)\,dx = (a_0 - \tfrac{1}{2}a_2)T_1(x) + \sum_{k=2}^{N-1} \tfrac{a_{k-1}-a_{k+1}}{2k} T_k(x) + \tfrac{1}{2N}a_{N-1}T_N(x) + \tfrac{1}{2(N+1)}a_N T_{N+1}(x).$$

Comparing coefficients with f_N for every T_k proves the assertion and shows b_0 to be the integration constant if $N \to \infty$ and $b_N \overset{!}{=} \frac{1}{2N}a_{N-1}$, which is highly inaccurate if N is small. Additionally, a (rapidly oscillating) remainder function is given by the last term, since T_{N+1} does not appear in the expansion of f_N. □

The next result shows that for the integral of the simplest discontinuous function, i.e. the *characteristic* or *indicator* function, uniform convergence holds.

Theorem 3.26. *Let $\mathbb{1}_{[-1,c]}$ be the characteristic function with $-1 < c \leq 1$ and define*

$$f(x) := x\mathbb{1}_{[-1,c]}(x) + c\mathbb{1}_{(c,1]}(x), \quad x \in [-1,1],$$

such that f is continuous. The definite integral over $[-1,x]$ of the classical Chebyshev series of $\mathbb{1}_{[-1,c]}$ converges uniformly to $f+1$.

Proof. It is straight forward to see that

$$\int_{-1}^{x} \mathbb{1}_{[-1,c]}(y)\,dy = f(x) + 1 \tag{3.82}$$

and since f is continuous and of bounded variation there exists a uniformly convergent Chebyshev expansion with the coefficients given as

$$b_k = \frac{\langle f, T_k\rangle_v}{\|T_k\|_v^2} = \frac{1}{\|T_k\|_v^2}\int_{-1}^{1} f(x)T_k(x)\tfrac{1}{\sqrt{1-x^2}}\,dx.$$

By applying the usual transform $x = \cos(y)$ the integral above can be calculated in an exact manner (abbreviating $c^* := \arccos(c)$), such that

$$\begin{aligned} b_k = \frac{1}{\|T_k\|_v^2}\left[\int_{c^*}^{\pi} \cos(y)\cos(ky)\,dy + c\int_{0}^{c^*}\cos(ky)dy\right] = \\ = \frac{1}{\pi}\begin{cases} cc^* - \sqrt{1-c^2} & k=0 \\ \pi + c\sqrt{1-c^2} - c^* & k=1 \\ 2\,\dfrac{k\sqrt{1-c^2}\cos(kc^*) - c\sin(kc^*)}{k^3-k} & k>1 \end{cases}. \end{aligned}$$

In the same way, although not being pointwise convergent, one can calculate the coefficients for the expansion of the characteristic function to be

$$a_k = \frac{1}{\|T_k\|_v^2}\int_{-1}^{c} T_k(x)\tfrac{1}{\sqrt{1-x^2}}\,dx = \frac{1}{\|T_k\|_v^2}\int_{c^*}^{\pi}\cos(ky)dy = \frac{1}{\pi}\begin{cases} \pi - c^* & k=0 \\ -2\,\dfrac{\sin(kc^*)}{k} & k>0 \end{cases}. \tag{3.83}$$

Paraphrasing (3.9) for definite integrals over $[-1,x]$ yields

$$\int_{-1}^{x} T_k(y)dy = \begin{cases} T_1(x)+1 & k=0 \\ \frac{1}{4}T_2(x) - \frac{1}{4} & k=1 \\ \frac{1}{2}\left(\frac{T_{k+1}(x)}{k+1} - \frac{T_{k-1}(x)}{k-1}\right) - \frac{(-1)^k}{k^2-1} & k>1 \end{cases}. \tag{3.84}$$

By combining (3.83) and (3.84) one can write the approximate version of the integral in (3.82) in the form

$$\sum_{k=0}^{N} a_k \int_{-1}^{x} T_k(y)\,dy = a_0(T_1(x)+1)+a_1(\tfrac{1}{4}T_2(x)-\tfrac{1}{4})+\sum_{k=2}^{N} a_k \tfrac{1}{2}\left(\tfrac{T_{k+1}(x)}{k+1}-\tfrac{T_{k-1}(x)}{k-1}\right)-\sum_{k=2}^{N} a_k \tfrac{(-1)^k}{k^2-1}$$

and by comparison with the expansion for f (using Lemma 3.25) it is left to show

$$\begin{aligned} &a_0 - \tfrac{1}{4}a_1 - \sum_{k=2}^{N} a_k \tfrac{(-1)^k}{k^2-1} \to b_0 + 1 \quad \text{as} \quad N \to \infty \\ &b_1 = a_0 - \tfrac{1}{2}a_2 \\ &b_k = \tfrac{1}{2k}(a_{k-1} - a_{k+1}) \quad \forall k > 1, \end{aligned} \tag{3.85}$$

such that

$$\int_{-1}^{x} \sum_{k=0}^{N} a_k T_k(y)\,dy \xrightarrow{\text{unif.}} \sum_{k=0}^{N} b_k T_k(x) + 1 \xrightarrow{\text{unif.}} f(x) + 1 \quad \text{as} \quad N \to \infty,$$

which concludes the proof.
To show (3.85) we start with b_1 by modifying (modulo the factor $\frac{1}{\pi}$ for the norm)

$$a_0 - \tfrac{1}{2}a_2 = \pi - c^* + \frac{\sin(2c^*)}{2} = \pi - c^* + \sin(c^*)c = \pi - c^* + c\sqrt{1-c^2} = b_1,$$

using the fact that $\sin(\arccos(x)) = \sqrt{1-x^2}$, and analogously for the other b_k

$$\begin{aligned} \tfrac{1}{2k}(a_{k-1} - a_{k+1}) &= \frac{1}{k}\left(-\frac{\sin((k-1)c^*)}{k-1} + \frac{\sin((k+1)c^*)}{k+1}\right) = \\ &= \frac{-(k+1)\sin((k-1)c^*) + (k-1)\sin((k+1)c^*)}{k^3-k} = \\ &= \frac{2k\sqrt{1-c^2}\cos(kc^*) - 2c\sin(kc^*)}{k^3-k} = b_k, \end{aligned}$$

applying trigonometric identities. As for the integration constant, the first equation in (3.85), the convergence of the sum is obvious, since (a_k) is a null sequence and with the use of some (computer) algebra one then gets

$$\sum_{k=2}^{\infty} a_k \tfrac{(-1)^k}{k^2-1} = \tfrac{2}{\pi}\left[\tfrac{3\sqrt{1-c^2}}{4} + (1+c)\arctan\left(\tfrac{c-1}{\sqrt{1-c^2}}\right)\right] = \tfrac{3\sqrt{1-c^2}}{2\pi} - \frac{2(1+c)\arctan\left(\frac{1-c}{\sqrt{1-c^2}}\right)}{\pi}.$$

With the inverse trigonometric identity $\arccos(x) = 2\arctan(\sqrt{1-x^2}/(1+x))$, $-1 < x \leq 1$ this simplifies to

$$\sum_{k=2}^{\infty} a_k \frac{(-1)^k}{k^2-1} = \frac{3\sqrt{1-c^2}}{2\pi} - \frac{(1+c)\arccos(c)}{\pi},$$

such that overall

$$a_0 - \tfrac{1}{4}a_1 - \sum_{k=2}^{\infty} a_k \tfrac{(-1)^k}{k^2-1} = 1 - \frac{c^*}{\pi} + \frac{\sqrt{1-c^2}}{2\pi} - \frac{3\sqrt{1-c^2}}{2\pi} + \frac{(1+c)\arccos(c)}{\pi} =$$
$$= 1 - \frac{c^*}{\pi} - \frac{\sqrt{1-c^2}}{\pi} + \frac{(1+c)c^*}{\pi} = 1 - \frac{\sqrt{1-c^2}}{\pi} + \frac{cc^*}{\pi} = b_0 + 1,$$

which is the desired equality. □

Remark 3.29. In Theorem 3.26, for readability and simplicity of the proof, the characteristic function was set to contain only one discontinuity step. The general case considering $\mathbb{1}_{[d,c]}$, $-1 < d < c \leq 1$ can be shown in the exact same way as above, with the only difference that the first equation in (3.85) has only b_0 on the right-hand side, since the integration evaluates to zero at -1.

Observe that with this remark integrals of every combination of multiplications of characteristic functions on $[-1, 1]$ have been proved to converge uniformly, due to $\prod_k \mathbb{1}_{A_k} = \mathbb{1}_{\cap A_k} = \mathbb{1}_{[a,b]}$. To utilize this result for arbitrary piecewise continuous functions use can be made of what is known as *simple functions* from measure theory.

A measurable function $s : I \to \mathbb{R}$ is called *simple*, if it can be written as

$$s = \sum_{j=0}^{p} \alpha_j \mathbb{1}_{A_j}, \quad \text{where} \;\; A_j = s^{-1}(\alpha_j) \in \sigma\text{–algebra}. \tag{3.86}$$

In the following the σ-algebra shall be the Borel σ-algebra, such that every continuous function is measurable and since we are dealing with functions defined on real intervals the Borel measure coincides with the Lebesgue measure and consequently all $A_j \in [-1, 1]$ are closed (or open) intervals.

Furthermore, for every positive measurable bounded function f there exists a monotone sequence s_n of simple functions which converges uniformly to f.

Theorem 3.27. *Let f be a piecewise continuous function of bounded variation with finitely many discontinuities, then the integral of the Chebyshev expansion of f converges uniformly.*

Proof. First observe that every piecewise continuous function f can be written as $\sum f_i \mathbb{1}_{A_i}$, where i counts the discontinuities and f_i are continuous functions on A_i. Thus it is sufficient to show the uniform convergence of the integral of $g\mathbb{1}_A$ with g continuous and of bounded variation and $A \in [-1, 1]$.
Without loss of generality one can assume g to be positive (otherwise split g into its positive and negative part, where the argument then holds for each part separately). Hence there exists a monotone sequence of simple functions s_n uniformly convergent to g.
Additionally one can find an integrable function h, such that $|s_n| \leq h$ (to apply the *dominated*

convergence theorem), which yields

$$\int_{-1}^{x} g(y)\mathbb{1}_A(y)\,dy = \lim_{n\to\infty} \int_{-1}^{x} s_n(y)\mathbb{1}_A(y)\,dy = \lim_{n\to\infty} \sum_j \alpha_{j,n} \int_{-1}^{x} \mathbb{1}_{A_j}(y)\mathbb{1}_A(y)\,dy,$$

by substituting $s_n(y) = \sum_j \alpha_{j,n}\mathbb{1}_{A_j}(y)$ as stated in (3.86). With $\mathbb{1}_{A_j}(y)\mathbb{1}_A(y) = \mathbb{1}_{A_j\cap A} =: \mathbb{1}_{[d,c]}$, the approximation of the integral on the right-hand side above converges uniformly, according to Theorem 3.26 and Remark 3.29. This holds for all j and n, which finishes the proof. □

Remark 3.30. The proof above is not constructive in the sense that in applications one would not construct a sequence of simple functions but expand the f_i or g in a Chebyshev series as well, whereas for the uniform convergence the argument still holds for such expansions. Moreover, calculating the product of two series expansions is too expensive and hence finding the Chebyshev series for $g\mathbb{1}_A$ as one function has to be preferred. Not only is producing the product of two series very expensive, but also highly inaccurate compared to an expansion of the product itself. Although one has that if $f_N \to f$ and $g_N \to g$ it follows that $f_N g_N \to fg$ for all x, using Lemma 3.2(v) to rearrange the product of two series to compare coefficients clearly depicts where the inaccuracy stems from and shows the slow rate of convergence.

In contrast to the general topic of this section we treated the Gibbs phenomenon in integrals in the case of classical polynomials. This is due to the fact that, as mentioned in Remark 3.1, integrals of rational Chebyshev polynomials over $\mathbb{R}$ do not exist. But applying the usual mapping to the results stated above shows that the convergence of such integrals holds for the case of mapped polynomials as well, provided one has some integral kernel function, which does not take part in the approximation but takes care of the necessary decay behavior for the individual integrals to exist. In particular Lemma 3.25 cannot be transferred directly to the whole real axis and also needs some modification for arbitrary integral kernels.

But, overall, the alleviation of the Gibbs phenomenon (as well as uniform convergence) using integrals can also be observed in the rational case.

Remark 3.31. It is straight forward to see how Theorems 3.26 and 3.27 can be extended to multivariate functions on $[-1,1]^n$, since $\mathbb{1}_{[a,b]\times[c,d]}(x,y) = \mathbb{1}_{[a,b]}(x)\mathbb{1}_{[c,d]}(y)$ etc. which shows how simple functions are defined in higher dimensions and that every case can be reduced to one-dimensional integrals of characteristic functions.

3.3 Spectral Collocation Methods

This section contains a description of the necessary requirements and tasks for setting up (reasonably fast working and converging) spectral schemes to solve operator equations. There are in general two different approaches, distinguished by the method of minimizing the error made by approximating the exact solution using a finite polynomial expansion, namely

Galerkin and *collocation* methods. The first requires the error to be orthogonal to the finite dimensional subspace (the approximate solution lies in), which is realized using inner products and is hence set up in a Hilbert space. Collocation methods, on the other hand, set the error to zero at a certain set of points.

It shall be noted here that both schemes treat the coefficients of the polynomial expansion as the discrete unknowns and thus fall into the *spectral* category. This correlates with the continuous projection operator from Section 3.2, whereas using interpolation leads to what is called *pseudospectral* schemes. Most authors do not distinguish between the notions *collocation* and *pseudospectral*. The analysis below shall clarify where the difference of those two approaches stems from.

Hackbusch (1995) shows that all the above mentioned discretization methods are in fact connected and it can be proved that under certain circumstances they all even lead to the same equation system.

First, a precise meaning shall be given to how spectral methods are mathematically set up in the usual operator description (see e.g. Hackbusch (1995), Golberg (1979) and Guo (1998)).

Assume $\mathcal{K}$ is a linear operator and $X(\Omega)$, $Y(\Omega)$ are two Banach spaces on the domain Ω, such that $\mathcal{K} : X \to Y$. If $f \in X$ is unknown and $g \in Y$ is given, the general operator equation then reads

$$\mathcal{K}f = g. \tag{3.87}$$

One way to find an approximate solution $f_N \in X_N \subset X$ is by using a projection operator $\mathcal{P}_N : Y \to Y_N$, such that Y_N is a finite dimensional subspace of Y, uniquely defined by $\mathcal{P}_N$, with a basis $\{q_1, \ldots, q_N\}$. Since one cannot expect f_N to be a solution of (3.87), substituting this function generates the error

$$e_N := \mathcal{K}f_N - g, \quad e_N \in Y,$$

which in general does not vanish (everywhere) in Ω, so one requires its projection onto Y_N to be zero, leading to a semi-discrete equation, i.e.

$$\mathcal{P}_N e_N = 0 \quad \Rightarrow \quad \mathcal{P}_N \mathcal{K} f_N = \mathcal{P}_N g.$$

Remark 3.32. From the definition of the Galerkin method (Y being a Hilbert space, e.g. $L^2(\Omega)$ with the inner product $\langle \cdot, \cdot \rangle$)

$$\langle e_N, q_i \rangle = 0 \quad \Leftrightarrow \quad \langle \mathcal{K}f_N, q_i \rangle = \langle g, q_i \rangle, \quad \forall i \leq N, \tag{3.88}$$

it is fairly obvious that this defines an orthogonal projection. As for the collocation method the error is evaluated at the set of points $\{x_1, \ldots, x_N\}$ (commonly termed *collocation points*) in Ω,

$$e_N(x_i) = 0 \quad \Leftrightarrow \quad \mathcal{K}f_N(x_i) = g(x_i), \quad \forall i \leq N, \tag{3.89}$$

where it is not as straight forward how an according (orthogonal) projection can be found.

In Section 3.2 Lemma 3.18 states that the therein defined operator $\mathcal{I}_N$ interpolates a function at a certain set of (interpolation) points. By setting the collocation points equal to the interpolation points one finds that if

$$\mathcal{I}_N \mathcal{K} f_N = \mathcal{I}_N g \quad \Rightarrow \quad \mathcal{K} f_N(x_i) = g(x_i), \ \ \forall i \leq N$$

and hence the collocation can also be treated as a projection method. In contrast to the Galerkin method no inner product is needed for the projection, but due to the definition of the interpolation operator one normally requires Y to contain at least bounded or even continuous functions (e.g. $C(\Omega)$ equipped with $\|\cdot\|_\infty$). The equivalence stated in (3.88) is non-trivial and a proof can be found in e.g. Hackbusch (1995).

Remark 3.33. For some problems one may further assume $X = Y$, such that $\mathcal{K} : X \to X$, where then boundedness and compactness properties can be easily utilized, as often done when dealing with integral equations of the second kind, cf. e.g. Golberg (1979), Golberg (1990), Sloan (1990) and Hackbusch (1995), where the latter showed for Volterra and Fredholm integral operators with certain types of kernels that Y is continuously (compactly) embedded in X, thus gaining some regularity for the solution of the according equations.

A very general treatment of consistency, stability and convergence for projection methods, without making any additional assumptions on the involved spaces and their finite dimensional subspaces, can be found in Guo (1998).

To analyze such methods, Hackbusch (1995) defines (with X being a Banach space and $\{\mathcal{P}_N\}$ a set of bounded projections from X to $X_N \subset X$)

$$\begin{aligned}
&(i) \quad \text{if} \ \ \mathcal{P}_N x \to x, \ \ \forall x \in X, \ \text{then the set is called } \textit{convergent}, \\
&(ii) \quad \text{if} \ \ \mathcal{P}_N x \to x, \ \ \forall x \in M, \ M \subset X \ \textit{dense}, \ \text{then the set is called } \textit{consistent}, \\
&(iii) \quad \text{if} \ \ \sup_N \{\|\mathcal{P}_N\|_X\} < \infty, \ \text{then the set is called } \textit{stable}.
\end{aligned} \tag{3.90}$$

One immediate consequence from this definition mentioned further in Hackbusch (1995) is

Lemma 3.28. *A set of projections $\{\mathcal{P}_N\}$ is convergent, if and only if it is consistent and stable.*

Proof. This follows directly from the theorem of *Banach-Steinhaus*, i.e. uniform boundedness principle, which states that given a set F of linear operators, with $T \in F : X \to Y$, if $\sup_{T \in F} \|Tx\|_X < \infty$, $\forall x \in X$, it follows that $\sup_{T \in F} \|T\|_{X \to Y} < \infty$. □

Remark 3.34. From Lemma 3.28 and Theorems 3.23 and 3.24 given in Section 3.2 one can infer that the interpolation operators defined in Section 3.2 are convergent and hence stable and consistent in the according function spaces. But as mentioned in Remark 3.28 the stability strongly depends on the interpolation points, since the uniform boundedness might be violated. For example in the case of equidistant points it has been shown (cf. e.g.

Hesthaven et al. (2007)) that the Lebesgue constant, measuring the norm of the projection, is asymptotically $\lambda_N \sim 2^{N+1}/(N \log N)$, whereas $\|\mathcal{I}_N\|_\infty = \mathrm{O}(\log N)$ using Chebyshev points. In the L^2 case Golberg (1990a) mentioned, as a consequence of the Erdös-Turán theorem, that $\mathcal{I}_N : C([a,b]) \to L^2_w([a,b])$ satisfies $\|\mathcal{I}_N\|^2_{C \to L^2_w} = \int_{[a,b]} w(t)dt$. Another result presented in Golberg (1979) shows that the stability of collocation methods depends not only on the choice of collocation points but also on the basis functions spanning the subspaces Y_N.

Remark 3.35. It shall be noted here that *consistency* relates purely to the involved operators and hence can be treated independently of any equation. Stability and convergence, on the other hand, are always connected to a given equation and this is why for example Guo (1998) (proving general results for equations of the first kind) only needs the inverse of $\mathcal{K}$ to exists to prove a convergence result, whereas Hackbusch (1995) and e.g Sloan (1990) require $\mathcal{K}$ to be compact, since they emphasize on equations of the second kind, i.e. $\lambda f - \mathcal{K}f = g$ (λ being a regular value of $\mathcal{K}$). The reason for this lies in the fact that by saying $\mathcal{K}_N := \mathcal{P}_N K$ the condition

$$\|\mathcal{K}_N - \mathcal{K}\|_X \to 0 \quad \text{as } N \to \infty \tag{3.91}$$

is always assumed to hold to establish convergence. Here, the pointwise convergence defined in (3.90)(i) is sufficient for convergence in the operator norm (i.e. uniform convergence) if $\mathcal{K}$ is *compact*, because compact operators map bounded sets onto precompact sets, for which pointwise implies uniform convergence (cf. Hackbusch (1995)).

Interestingly, as mentioned in Sloan (1990), the pointwise convergence, although being sufficient for compact $\mathcal{K}$, is not necessary for (3.91) to hold. Paraphrasing this one might also state

Lemma 3.29. *Given a bounded operator $\mathcal{K}$ and the set of bounded operators $\{\mathcal{P}_N\}$ then $\|\mathcal{P}_N\mathcal{K} - \mathcal{K}\|_X \to 0$, if $\mathcal{P}_N$ converges uniformly to the identity.*

Proof. Obviously, one has $\|\mathcal{P}_N\mathcal{K} - \mathcal{K}\| = \|(\mathcal{P}_N - I)\mathcal{K}\| \leq \|\mathcal{P}_N - I\| \underbrace{\|\mathcal{K}\|}_{<\infty}$. □

This implies that compactness for $\mathcal{K}$ is not necessary for convergence, if one can find a projection method which converges uniformly, but, for example, neither the Galerkin nor the collocation approach satisfy this in general.

Stability for equations of the second kind is defined via the norm of the solution operator $\|(\lambda I - \mathcal{K}_N)^{-1}\|_X \leq const$, $\forall N \geq N_0$, where (3.91) ensures the uniform boundedness. Conversely, if one considers such operator equations of the first kind as (3.87) with $\mathcal{K}$ being compact, its inverse is in general *unbounded*, which might lead to instability (cf. for example the ill-posedness of Abel integral equations as treated in Gorenflo & Vessella (1991) and Kress (1999), including regularization techniques).

From the definition of consistency given as

$$\lim_{N \to \infty} \|\mathcal{P}_N\mathcal{K}v - \mathcal{K}v\|_Y = 0, \quad \forall v \in X, \tag{3.92}$$

where $\mathcal{K} : X \to Y$, it is sufficient that (3.90)(ii) holds in Y, without any additional requirement on $\mathcal{K}$. Furthermore, Golberg (1990) mentioned that if the projection onto Y_N is uniformly bounded and $\bigcup Y_N$ is dense in Y, then $\|\mathcal{P}_N h - h\|_Y$, $h := \mathcal{K}v$, tends to zero.

Remark 3.36. Summarizing the considerations above, compactness or boundedness of given operators can be sufficient, but need not be necessary, for spectral methods to be consistent, stable and convergent (although it definitely simplifies the analysis and proofs). Also, when using collocation, which is essentially an interpolation, the right choice of collocation points has to be made, otherwise it might not be possible to establish stability or even consistency for certain types of operators.

So far we only dealt with the method of how to minimize the error (for example using interpolation) and possible issues arising when approximating the solution f with $f_N \in X_N$. This then leads to sometimes termed semi-discrete equations, since nothing has been said about how to actually obtain f_N from such a scheme. As has been done for Y_N, the subspace X_N shall again be defined by a projection $\mathcal{Q}_N : X \to X_N$, such that $f_N = \mathcal{Q}_N f$.

Consequently, a discretization of (3.87) then reads

$$\mathcal{P}_N \mathcal{K} \mathcal{Q}_N f = \mathcal{P}_N g.$$

In most of the works mentioned above, the convergence analysis is done without this additional step of projecting f onto X_N. Guo (1998), in contrast, derives consistency, stability and convergence conditions exclusively for a "discretized" (invertible) operator $\mathcal{K}_N : X_N \to Y_N$ (which is *not* equal to $\mathcal{K}_N = \mathcal{P}_N \mathcal{K}$ defined above) and an inverse mapping $\mathcal{S}_N : Y_N \to Y$, such that the approximate equation is

$$\mathcal{K}_N f_N = \mathcal{P}_N g,$$

where the approximation error can then be given as $R_N(v) = \mathcal{K}v - \mathcal{S}_N \mathcal{K}_N \mathcal{Q}_N v$, $\forall v \in X$. Furthermore, the consistency condition is thus said to be

$$\lim_{N\to\infty} \|R_N(v)\|_Y = 0, \quad \forall v \in D \ldots \text{subset of solutions } f.$$

This is in complete accordance to (3.92), which in full discretization yields the requirement

$$\lim_{N\to\infty} \|\mathcal{P}_N \mathcal{K} \mathcal{Q}_N v - \mathcal{K}v\|_Y = 0, \quad \forall v \in X. \tag{3.93}$$

The following results show under which assumptions this can be satisfied.

Lemma 3.30. *If the two (bounded) projections $\mathcal{P}_N$ and $\mathcal{Q}_N$ are (pointwise) convergent, $\mathcal{K}$ being bounded is sufficient for* (3.93) *to hold.*

Proof. Starting from the definition of consistency, some modification yields the inequality

$$\|\mathcal{P}_N \mathcal{K} \mathcal{Q}_N v - \mathcal{K}v\|_Y = \|\mathcal{P}_N \mathcal{K} \mathcal{Q}_N v - \mathcal{K}v + \mathcal{P}_N \mathcal{K}v - \mathcal{P}_N \mathcal{K}v\|_Y \leq$$

$$\leq \| \underbrace{\mathcal{P}_N \mathcal{K} v - \mathcal{K} v}_{\text{set } h := \mathcal{K} v} \|_Y + \| \underbrace{\mathcal{P}_N \mathcal{K} \mathcal{Q}_N v - \mathcal{P}_N \mathcal{K} v}_{= \mathcal{P}_N \mathcal{K} (\mathcal{Q}_N v - v)} \|_Y \leq \tag{3.94}$$

$$\leq \underbrace{\|\mathcal{P}_N h - h\|_Y}_{\to 0} + \underbrace{\|\mathcal{P}_N\|_Y \|\mathcal{K}\|_{X \to Y}}_{< \infty} \underbrace{\|\mathcal{Q}_N v - v\|_X}_{\to 0} \to 0, \quad \forall v \in X.$$

□

Remark 3.37. The boundedness of the projections in Lemma 3.30 is sufficient for the argument in the last line in (3.94). For collocation methods, where $\mathcal{P}_N \equiv \mathcal{I}_N$, $\|\mathcal{P}_N\|_\infty \to \infty$ as $N \to \infty$ (see Remark 3.34), even when using appropriate collocations points. But since the rate of divergence of $\|\mathcal{P}_N\|_\infty$ is slow, Lemma 3.30 still holds, provided $\|\mathcal{Q}_N v - v\|_X$ tends to zero fast enough (as it is usually the case).

Although being sufficient, the boundedness of $\mathcal{K}$ is not necessary. The next result shows that for differential operators, which are unbounded in general, consistency can be established in the appropriate Sobolev spaces.

Theorem 3.31. *Let $\mathcal{D}^m := \sum_{|k|_s \leq m} a_k \partial_x^k$, with $a_k \in \mathbb{R}$, be a classical derivative operator and $v \in H^r(\Omega)$, then $\forall r, q \in \mathbb{R}$, with $m + q < r$, provided $\mathcal{P}_N$ converges (pointwise) in H^{m+q}*

$$\lim_{N \to \infty} \|\mathcal{D}^m v - \mathcal{D}^m \mathcal{P}_N v\|_{H^q} = 0$$

holds.

Proof. This is a generalization of a result presented in Hesthaven et al. (2007), where it was proved for approximations via orthogonal polynomial expansions. The proof here is essentially the same, with the main arguments being

$$\|\mathcal{D}^m v - \mathcal{D}^m \mathcal{P}_N v\|_{H^q} = \| \sum_{|k|_s \leq m} a_k \partial_x^k (v - \mathcal{P}_N v)\|_{H^q} \leq$$

$$\leq \max_k a_k \sum_{|k|_s \leq m} \|\partial_x^k (v - \mathcal{P}_N v)\|_{H^q} = \max_k a_k \sum_{|k|_s \leq m} \underbrace{\left(\sum_{|j|_s \leq q} \|\partial_x^j [\partial_x^k (v - \mathcal{P}_N v)]\|_w^2 \right)^{1/2}}_{\leq \|(v - \mathcal{P}_N v)\|_{H^{|k|_s + q}}} \leq$$

$$\leq c \, \|(v - \mathcal{P}_N v)\|_{H^{m+q}} \to 0.$$

□

Remark 3.38. The assumption of pointwise convergence of the projection $\mathcal{P}_N$ in some H^q has been proved in Guo (1998) in form of v being in H^r and the error measured in H^q, $\forall q \leq r$ (even including the convergence rate), which holds for a variety of orthogonal projections involving trigonometric, Hermite, Laguerre and Chebyshev polynomials.

As an immediate consequence from Theorem 3.31 one can infer consistency of the collocation approach involving general differential operators, provided the function satisfies some dif-

ferentiability requirements. Furthermore, consider a combination of bounded and unbounded operators, i.e. $\mathcal{K} : Y \to Y$ and $\mathcal{D} : X \to Y$, within a projection scheme, then

$$\|\mathcal{P}_N\mathcal{K}\mathcal{D}\mathcal{Q}_N v - \mathcal{K}\mathcal{D}v\|_Y \overset{cf.\ (3.94)}{\leq} \underbrace{\|\mathcal{P}_N\|_Y\|\mathcal{K}\|_Y}_{<\infty}\|\mathcal{D}\mathcal{Q}_N v - \mathcal{D}v\|_Y,$$

which proves that if the approximation of $\mathcal{D}$, using the projection $\mathcal{Q}_N$, is pointwise convergent, arbitrary combinations of bounded operators with those, where the application to projections converges pointwise, lead to a consistent projection method. This result directly applies to the integro-differential operators in Section 2, equations (2.34), (2.32), (2.49), (2.48) and (2.126).

Remark 3.39. As shown in Sloan (1990) if $\mathcal{K}$ is compact, pointwise convergence of the *orthogonal* projections in a Hilbert space is sufficient for consistency. The main arguments in the proof given therein are that by using the fact that orthogonal projections are *self-adjoint*, one has $\|\mathcal{K}\mathcal{P}_N - \mathcal{K}\| = \|\mathcal{P}_N\mathcal{K}^* - \mathcal{K}^*\|$, and since the adjoint $\mathcal{K}^*$ is again compact, such that uniform convergence holds (as mentioned in Remark 3.35). Interestingly, when considering (3.94), one could alternatively say

$$\begin{aligned}\|\mathcal{P}_N\mathcal{K}\mathcal{Q}_N v - \mathcal{K}v\|_Y &= \|\mathcal{P}_N\mathcal{K}\mathcal{Q}_N v - \mathcal{K}v + \mathcal{K}\mathcal{Q}_N v - \mathcal{K}\mathcal{Q}_N v\|_Y \leq \\ &\leq \|\mathcal{K}(\mathcal{Q}_N v - v)\|_Y + \|(\mathcal{P}_N\mathcal{K} - \mathcal{K})\mathcal{Q}_N v\|_Y \leq \\ &\leq \|\mathcal{K}\|_{X\to Y}\|\mathcal{Q}_N v - v\|_X + \underbrace{\|\mathcal{K} - \mathcal{P}_N\mathcal{K}\|_{X\to Y}}_{\mathcal{K}\overset{comp.}{\longrightarrow}0}\|\mathcal{Q}_N v\| \quad \forall v \in X,\end{aligned}$$

where it becomes obvious again that for compact operators, pointwise convergence of the projections is sufficient for consistency. On the other hand, as argued in Sloan (1990), if a compact operator is defined to act on $C([0,1])$ with $\mathcal{P}_N$ being an arbitrary interpolatory projection (not necessarily orthogonal, where self-adjointness cannot be assumed), $\|\mathcal{K}\mathcal{P}_N - \mathcal{K}\| \geq \|\mathcal{K}\|$, which can be found by assuming the functions v satisfy $\|v\| = 1$ and $v(t_i)$ vanish at the collocation points t_i. Thus, $\|(\mathcal{K}\mathcal{P}_N - \mathcal{K})v\|_\infty = \|\mathcal{K}v\|_\infty$. This is in full agreement with Remark 3.34.

There is still another way to look at consistency for compact operators. A classical functional analysis result is

$$\mathcal{K} \text{ compact} \quad \Leftrightarrow \quad v_N \rightharpoonup v \;\Rightarrow\; \mathcal{K}v_N \to \mathcal{K}v,$$

with v_N and v being in the domain of $\mathcal{K}$ and the weak convergence of v_N is defined via an appropriate inner product in the according Hilbert space. Thus one can further infer that for compact operators weak convergence of the projection is sufficient for consistency.

Remark 3.40. Since this section deals with collocation methods we shall not go further into the details of the Galerkin approach. However, a few comments are in order. Hackbusch (1995) and Fromme & Golberg (1979) showed that in the usual setting of (weighted) L^2

Hilbert spaces with the inner product given as an integral, the collocation scheme with the collocation points being the zeros of the next higher order polynomial is equivalent to the Galerkin scheme when approximating the inner product by some quadrature rule (i.e. Jacobi-Gauss or trapezoidal). This becomes also obvious by taking into account that collocation at the zeros of orthogonal polynomials is an interpolation projection, which then can be seen as an approximation of a continuous projection using discrete inner products (cf. (3.53), (3.55) and Lemma 3.18).

What follows here is that with such a connection some stability and convergence results transfer directly from Galerkin to collocation (where the former is in general more stable), especially when operators are involved which do not allow for an exact evaluation of the inner product. The biggest advantage of collocation, most notably in the case of multi-dimensional problems, is the straight forward set-up as a simple pointwise evaluation, while, as said above, inner products become heavily involved in higher dimensions.

3.3.1 Properties of Abel Operators and Riesz Potentials

In virtue of the consistency and convergence results and in connection with the equations given in Section 2 (e.g. (2.49), (2.48) or (2.126)) we state some theorems on boundedness and compactness of certain integral operators.

In accordance to Gorenflo & Vessella (1991) an *Abel integral operator* shall be defined as (in an appropriate function space over $\mathbb{R}$)

$$\mathcal{J}^{\alpha}_{-\infty}(f)(x) := \int_{-\infty}^{x} (x-\xi)^{\alpha-1} f(\xi)\, d\xi \quad \text{and} \quad \mathcal{J}^{\alpha}_{\infty}(f)(x) := \int_{x}^{\infty} (\xi-x)^{\alpha-1} f(\xi)\, d\xi, \tag{3.95}$$

with $0 < \alpha < 1$ and $-\infty < x < \infty$ (where the factor $1/\Gamma(\alpha)$ is omitted).

Gorenflo & Vessella (1991) proved boundedness and compactness results for Abel operators on bounded intervals. Since the extension to unbounded intervals is not straight forward, we prove the boundedness of $\mathcal{J}^{\alpha}_{\pm\infty}$ between weighted L^2 spaces and the space of continuous functions (in virtue of the convergence results for the projection operators from Section 3.2). The integrability of the weak singularity of the kernel in (3.95) requires a necessary decay behavior of the argument function for the integral to exist, thus one has to work in weighted spaces.

Defining the step function as

$$H(x) = \begin{cases} 1 & x \geq 0 \\ 0 & otherwise \end{cases} \quad \Rightarrow \quad \mathcal{J}^{\alpha}_{-\infty}(f)(x) = \int_{\mathbb{R}} \frac{H(x-\xi)}{(x-\xi)^{1-\alpha}} f(\xi)\, d\xi \tag{3.96}$$

and analogously for $\mathcal{J}^{\alpha}_{\infty}$.

Theorem 3.32. *Given the weight function $w(x) = (1+x^2)^{-\lambda}$, $\lambda > \alpha/2$, $x \in \mathbb{R}$ and $f \in L^2_{1/w}(\mathbb{R})$, there exists a positive constant c, such that*

$$\|\mathcal{J}^{\alpha}_{-\infty}(f)\|_w \leq c\|f\|_{1/w}$$

holds (and analogously for $\mathcal{J}^{\alpha}_{\infty}$) and hence $\mathcal{J}^{\alpha}_{\pm\infty}$ is bounded from $L^2_{1/w} \to L^2_w$.

Proof. With (3.96) one has (using Hölder's inequality)

$$\begin{aligned} g(x) := \mathcal{J}^{\alpha}_{-\infty}(f)(x) &= \int\limits_{\mathbb{R}} \frac{H(x-\xi)}{(x-\xi)^{1-\alpha}} f(\xi)\, d\xi \leq \\ &\leq \int\limits_{\mathbb{R}} \Big| \frac{H^{1/2}(x-\xi)}{(x-\xi)^{(1-\alpha)/2}} \frac{H^{1/2}(x-\xi)}{(x-\xi)^{(1-\alpha)/2}} (1+\xi^2)^{\lambda/2} (1+\xi^2)^{-\lambda/2} f(\xi) \Big|\, d\xi \leq \\ &\leq \underbrace{\left(\int\limits_{\mathbb{R}} \frac{H(x-\xi)}{(x-\xi)^{1-\alpha}} (1+\xi^2)^{-\lambda} d\xi \right)^{1/2}}_{=:\sqrt{k},\ as\ \sup_x} \left(\int\limits_{\mathbb{R}} \frac{H(x-\xi)}{(x-\xi)^{1-\alpha}} (1+\xi^2)^{\lambda} f^2(\xi) d\xi \right)^{1/2}, \end{aligned}$$

where the first integral exists $\forall x$ since the singularity is integrable and $\lambda > \alpha/2$ provides the necessary decay. Thus, we obtain further

$$\begin{aligned} \|g\|^2_w = \int\limits_{\mathbb{R}} g^2(x) w(x) dx &\leq k \int\limits_{\mathbb{R}} \int\limits_{\mathbb{R}} \frac{H(x-\xi)}{(x-\xi)^{1-\alpha}} (1+\xi^2)^{\lambda} w(x) f^2(\xi) d\xi\, dx = \\ &= k \int\limits_{\mathbb{R}} f^2(\xi)(1+\xi^2)^{\lambda} \underbrace{\int\limits_{\mathbb{R}} \frac{H(x-\xi)}{(x-\xi)^{1-\alpha}} w(x) dx}_{<\infty\ \forall \xi,\ \lambda > \alpha/2}\, d\xi \leq c\|f\|^2_{1/w}. \end{aligned}$$

□

Remark 3.41. The boundedness of the Abel operator between (weighted) spaces of continuous functions on $\mathbb{R}$ is much more straight forward. Say $f \in (C(\mathbb{R}), \|\cdot\|_\infty)_{1/w}$, meaning that $\|f/w\|_\infty$ is bounded, with the weight given as in Theorem 3.32, then

$$\begin{aligned} |g(x)| := \left| \int_{-\infty}^{x} (x-\xi)^{\alpha-1} f(\xi) d\xi \right| &= \left| \int_{-\infty}^{x} (x-\xi)^{\alpha-1} w(\xi) \frac{f(\xi)}{w(\xi)} d\xi \right| \leq \\ &\leq \int_{-\infty}^{x} (x-\xi)^{\alpha-1} w(\xi) \Big| \frac{f(\xi)}{w(\xi)} \Big| d\xi \leq \sup \Big| \frac{f}{w} \Big| \underbrace{\int_{-\infty}^{x} (x-\xi)^{\alpha-1} w(\xi) d\xi}_{<\infty,\ \forall x} \leq c \Big\| \frac{f}{w} \Big\|_\infty, \end{aligned}$$

where g is obviously continuous, which proves the assertion. □

Sloan (1981) proved that under certain conditions on the kernel, integral operators defined on $[0,\infty)$ are compact from $C_l([0,\infty))$ to itself, where the index l means that every function has a limit at infinity. The conditions derived for the kernel k therein can be easily extended

to the whole real line, reading as

$$\begin{aligned} &(i) \quad \int_{\mathbb{R}} |k(x,s)|ds < \infty \\ &(ii) \quad \lim_{y\to x} \int_{\mathbb{R}} |k(y,s) - k(x,s)|ds = 0 \\ &(iii) \quad \lim_{|x|\to\infty} \sup_{|y|\geq|x|} \int_{\mathbb{R}} |k(y,s) - k(x,s)|ds = 0. \end{aligned} \tag{3.97}$$

From Remark 3.41 it is obvious that the Abel operator on $C_l(\mathbb{R})$ does not exist and hence for the compactness result we again work on the subspace of functions with the necessary decay at infinity.

Theorem 3.33. *The Abel integral operator defined in* (3.95) *is compact from* $(C(\mathbb{R}), \|\cdot\|_\infty)_{1/w}$ *to* $C_l(\mathbb{R})$*, with* $w(x) = (1+x^2)^{-\lambda}$*,* $\lambda > \alpha/2$*.*

Proof. Without loss of generality we prove the assertion for $\mathcal{J}^\alpha_{-\infty}$. Sloan (1981) showed that the conditions (3.97) are necessary and sufficient for an integral operator to be compact on C_l. To utilize this, we use a weighted kernel to define $\mathcal{J}_w$ on C_l as

$$k(x,\xi) := H(x-\xi)(x-\xi)^{\alpha-1}w(\xi), \qquad \mathcal{J}_w(f)(x) := \int_{\mathbb{R}} k(x,\xi)f(\xi)d\xi.$$

It is now easily seen that

$$g(x) := \int_{\mathbb{R}} |k(x,\xi)|d\xi$$

is continuous and bounded on $\mathbb{R}$, with $g(x) \to 0$ (strictly monotonically) as $|x| \to \infty$. Thus condition (i) is satisfied.
Since the kernel is non-negative we have for $y \leq x$, $k(x,\xi) \leq k(y,\xi)$, $\forall \xi \leq y$. Hence (without loss of generality say $y \leq x$), by using the fact that $k(x,\xi) = k(y,\xi) = 0$, $\forall \xi > x$,

$$\begin{aligned} \lim_{y\to x} \int_{\mathbb{R}} |k(y,\xi) - k(x,\xi)|d\xi &= \lim_{y\to x} \left[\int_{-\infty}^{y} |k(y,\xi) - k(x,\xi)|d\xi + \int_{y}^{x} |k(y,\xi) - k(x,\xi)|d\xi \right] = \\ &= \lim_{y\to x} \left[\int_{-\infty}^{y} (k(y,\xi) - k(x,\xi))d\xi + \int_{y}^{x} k(x,\xi)d\xi \right] = \\ &= \lim_{y\to x} \left[\int_{-\infty}^{y} k(y,\xi)d\xi - \int_{-\infty}^{y} k(x,\xi))d\xi + \int_{y}^{x} k(x,\xi)d\xi \right], \end{aligned}$$

and performing the limit one can easily see that the last term on the right-hand side is zero, the first term yields $\lim_{y\to x} g(y) = g(x)$ (since g is continuous) and the second term is just the definition of the improper integral for the weakly singular kernel, which again evaluates to

$g(x)$. As for (iii) we observe that $g(y) \leq g(x)$ if $|y| \geq |x| \gg 1$, such that

$$\sup_{|y|\geq|x|} \int_{\mathbb{R}} |k(y,\xi) - k(x,\xi)| d\xi \leq \sup_{|y|\geq|x|} \int_{\mathbb{R}} (k(y,\xi) + k(x,\xi)) d\xi \leq 2 \int_{\mathbb{R}} k(x,\xi) d\xi = 2g(x)$$

and taking the limit $|x| \to \infty$ where then $g(x) \to 0$ shows condition (iii) holds.
Thus, $\mathcal{J}_w$ is compact on C_l. This relates directly to the compactness of $\mathcal{J}^{\alpha}_{-\infty}$ by saying for every $f^* \in C_l$ we find an $f \in C(\mathbb{R})$, decaying to zero, such that $f^* = f/w$ and substitution into $\mathcal{J}_w$ finishes the proof. □

Remark 3.42. One should be aware that operators with convolution kernels, e.g. of Wiener-Hopf type, cannot satisfy the third condition, as stated in Sloan (1981). Due to the (necessary) weight function in the theorem above, the kernel here does not fall into this category.

Next we define the *Riesz potentials* (cf. e.g. Stein (1970) and Samko (1976)) for $x = (x_1, \ldots, x_n)$ and $\xi = (\xi_1, \ldots, \xi_n)$ as (in an appropriate function space over $\mathbb{R}^n$)

$$\mathcal{R}^{\alpha}(f)(x) := \int_{\mathbb{R}^n} \frac{1}{|x-\xi|^{n-\alpha}} f(\xi)\, d\xi, \quad 0 < \alpha < n, \tag{3.98}$$

omitting the usual constant $1/\gamma(\alpha)$, $\gamma(\alpha) = \pi^{n/2} 2^{\alpha} \Gamma(\alpha/2)/\Gamma(n/2 - \alpha/2)$.

The following theorem shows the existence and boundedness properties of such operators.

Theorem 3.34. *Let $0 < \alpha < n$, $1 \leq p < q < \infty$ and $1/q = 1/p - \alpha/n$.*
(i) If $f \in L^p(\mathbb{R}^n)$, then the integral $\mathcal{R}^{\alpha}(f)$ converges absolutely for almost every $x \in \mathbb{R}^n$.
(ii) If, in addition, $1 < p$, then

$$\|\mathcal{R}^{\alpha}(f)\|_{L^q} \leq c \|f\|_{L^p}$$

holds (where the constant c might depend on p, q, n).

Proof. see Stein (1970). □

Remark 3.43. The result above has also been proved by Sobolev (1938) using a different approach. As an immediate consequence from (i) (since $0 < q < \infty$) one obviously has $\mathcal{R}^{\alpha}(L^p(\mathbb{R}^n))$ is well defined, if $1 \leq p < n/\alpha$. This has been extended in Samko (1999) to $\mathcal{R}^{\alpha}(L^p(\mathbb{R}^n)) \subset L^p_{loc}(\mathbb{R}^n)$ in a distributional sense.

Setting $n = 2$ and $\alpha = 1$ in (3.98), one gains the operator appearing in (2.32) and (2.48),

$$\mathcal{R}^1(f)(x_1, x_2) = \int_{\mathbb{R}^2} \frac{1}{|(x_1, x_2) - (\xi_1, \xi_2)|} f(\xi_1, \xi_2)\, d\xi_1 d\xi_2, \tag{3.99}$$

such that, from Theorem 3.34, boundedness from L^p into L^q, for $1/p - 1/q = 1/2$, as well as the existence (almost everywhere) for $f \in L^p(\mathbb{R}^2)$, $1 \leq p < 2$, follows immediately and hence, the most important function spaces when using spectral methods, L^2 and $(C, \|\cdot\|_{\infty})$, are excluded from the classical boundedness result above in the case of $\mathcal{R}^1$ on $\mathbb{R}^2$. Thus, in order to work in the desired setting, certain additional requirements have to be met.

Remark 3.44. On bounded domains $\Omega \subset \mathbb{R}^n$, *essential boundedness* of the function, i.e. $f \in L^\infty(\Omega)$ is sufficient for L^p integrability. By introducing (positive) weights in the form $w(x) = (1+|x|)^{-a}$, $x \in \mathbb{R}^n$ and $a > n$, essential boundedness implies $f \in L^p_w(\mathbb{R}^n)$. On the other hand, requiring $f \in L^p_{1/w}(\mathbb{R}^n)$ yields a decay behavior for f faster than $|x|^{-(n+a)/p}$. Thus, given a, one can impose a decay rate on f. If f is not essentially bounded, then applying such weight functions will not change the demand on possible singularities to be weaker than $|x|^{-n/p}$ (as $|x| \to 0$) for $f \in L^p_w(\mathbb{R}^n)$.

With this remark we can formulate the following

Theorem 3.35. *Let $f \in L^2(\mathbb{R}^2)$ and $w(x) = (1+|x|)^{-\lambda}$, $\lambda > 2$, $x \in \mathbb{R}^2$, then there exists a constant c, such that*

$$\|\mathcal{R}^1(f)\|_{L^2_w} \leq c\|f\|_{L^2},$$

i.e. $\mathcal{R}^1 : L^2(\mathbb{R}^2) \to L^2_w(\mathbb{R}^2)$ is bounded.

Proof. Say $r = |x-\xi|$ (cf. (3.99)), then with $\epsilon > 0$ but small

$$\begin{aligned} g(x) := \mathcal{R}^1(f)(x) = \int\limits_{\mathbb{R}^2} r^{-1} f(\xi)\, d\xi \leq \\ \leq \int\limits_{\mathbb{R}^2} \left| r^{-1/2}(1+r)^{-(1+\epsilon)/2}\, r^{-1/2}(1+r)^{(1+\epsilon)/2} f(\xi) \right| d\xi, \end{aligned}$$

where one can apply Hölder's inequality to obtain

$$g(x) \leq \left(\int\limits_{\mathbb{R}^2} r^{-1}(1+r)^{-1-\epsilon}\, d\xi \right)^{\frac{1}{2}} \left(\int\limits_{\mathbb{R}^2} r^{-1}(1+r)^{1+\epsilon} f^2(\xi)\, d\xi \right)^{\frac{1}{2}}. \tag{3.100}$$

By using polar coordinates (r, θ) centered around x, i.e.

$$\left.\begin{aligned} x_1 - \xi_1 &= r\, \cos(\theta) \\ x_2 - \xi_2 &= r\, \sin(\theta) \end{aligned}\right\} \quad \Rightarrow \quad \left| \begin{array}{cc} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{array} \right| = r, \tag{3.101}$$

the first integral on the right-hand side in (3.100) can be written as

$$\int\limits_{-\pi}^{\pi} \int\limits_{0}^{\infty} r^{-1}(1+r)^{-1-\epsilon} r\, dr d\theta = 2\pi \int\limits_{0}^{\infty} (1+r)^{-1-\epsilon} dr =: k < \infty,$$

and consequently

$$g^2(x) w(x) \leq k \int\limits_{\mathbb{R}^2} r^{-1}(1+r)^{1+\epsilon} w(x) f^2(\xi)\, d\xi.$$

Furthermore, since both sides above are positive, the integral on the right hand side exists (f decays faster than $|\xi|^{-1}$) and w decays faster than $|x|^{-2}$, such that

$$\int_{\mathbb{R}^2} g^2(x)w(x)dx \leq$$

$$\leq k \int_{|x|=0}^{\infty} \int_{|x-\xi|=0}^{\infty} r^{-1}(1+r)^{1+\epsilon} w(x) f^2(\xi)\, d\xi\, dx = \int_{|\xi|=0}^{\infty} f^2(\xi) \underbrace{\int_{|x-\xi|=0}^{\infty} r^{-1}(1+r)^{1+\epsilon} w(x)\, dx}_{<\infty,\ \forall \xi}\, d\xi$$

$$\Rightarrow \|g\|^2_{L^2_w} \leq c\|f\|^2_{L^2}.$$

□

Remark 3.45. As mentioned in Remark 3.44 one has to take into account singularities and the decay behavior of functions when dealing with L^p spaces on unbounded domains. Boundedness of $\mathcal{R}^1$ in Theorem 3.35 can also be proved between $L^1(\mathbb{R}^2)$ and $L^q_w(\mathbb{R}^2)$, where $1 \leq q < 2$ and the weight has to satisfy $\lambda > 2-q$ (see Samko (1998)). Hence, it is obvious that $\mathcal{R}^1(L^p)$ exists (in the L^1_w sense) for all L^p functions with compact support (or on bounded domains) since $L^p(\Omega) \subset L^1(\Omega)$, $\forall p \geq 1$. A formal consideration of functions $f \in L^p$, $p > 2$, shows the decay to be too weak for $\mathcal{R}^1(L^p)$ to exist on unbounded domains. Thus, weights of the form $1/w$ have to be introduced in the pre-image space to impose the necessary decay, or in other words, to have $L^p_{1/w}(\mathbb{R}^2) \subset L^1(\mathbb{R}^2)$. On the side of the image space this means that the singularities become weaker and can thus be measured again in a (weighted) subspace of L^1_w. Finally, without proof, we formally claim that there exist weights w_1, w_2, such that $\mathcal{R}^1 : L^p_{w_1} \to L^p_{w_2}$ is bounded for all $1 \leq p \leq \infty$. □

Remark 3.46. Consider the space $(C(\mathbb{R}^2), \|\cdot\|_\infty)_{1/w}$ as the pre-image space (with index $1/w$ denoting that f/w is bounded on $\mathbb{R}^2$, cf. Remark 3.41) and say $\bar{f}(r,\theta) := f(\xi_1,\xi_2)$ in (3.99), such that $\bar{f}(0,\theta) = f(x_1,x_2)$, then

$$g(x) := \mathcal{R}^1(f)(x) = \int_{-\pi}^{\pi} \int_{0}^{\infty} \bar{f}(r,\theta) dr d\theta,$$

which exists $\forall x$ if the weight satisfies $\lambda > 1$ (i.e. f decays faster than r^{-1}) and consequently g is continuous and essentially bounded.

To obtain compactness for the Riesz potential on weighted spaces of continuous functions we state the following

Theorem 3.36. *Let X be a normed space and Y be a Banach space. Let the sequence $A_n : X \to Y$ of compact linear operators be norm convergent to a linear operator $A : X \to Y$, i.e., $\|A_n - A\|_{X\to Y} \to 0$, $n \to \infty$. Then A is compact.*

Proof. see e.g. Kress (1999) □

By defining a weakly singular kernel as $|k(x,y)| \leq M|x-y|^{\alpha-n}$, $M>0$, $0<\alpha\leq n$, $x,y\in G\subset\mathbb{R}^n$ bounded, $x\neq y$, Theorem 3.36 was subsequently used to proof

Theorem 3.37. *The integral operator with weakly singular kernel k is compact from $C(G)$ to itself.*

Proof. see e.g. Kress (1999). □

As shown in Remark 3.46 one needs decaying continuous functions, i.e. $\|f/w\|_\infty<\infty$ for $\mathcal{R}^1(f)$ to exist. For every continuous function f, f/w then lies in $C_l(\mathbb{R}^2)$, meaning that every function has a limit as the radius tends to infinity (cf. Theorem 3.33 and Sloan (1981)). With this prerequisites we are now able to prove

Theorem 3.38. *The Riesz potential operator defined in* (3.99) *is compact from* $(C(\mathbb{R}^2),\|\cdot\|_\infty)_{1/w}$ *to* $C_l(\mathbb{R}^2)$*, with* $w(x)=(1+|x|)^{-\lambda}$*,* $\lambda>1$*.*

Proof. For $f\in(C(\mathbb{R}^2),\|\cdot\|_\infty)_{1/w}$ we write $f^*=f/w\in C_l(\mathbb{R}^2)$. Now the weighted potential operator shall be

$$\mathcal{R}_w(f^*)(x):=\int\limits_{\mathbb{R}^2}|x-\xi|^{-1}w(\xi)f^*(\xi)d\xi,\quad k(x,\xi):=|x-\xi|^{-1}w(\xi),$$

which is defined on C_l.
Further say $k_n(x,\xi):=k(x,\xi)\mathbb{1}_{B_n(x)}(\xi)$, such that (by applying polar coordinates), $k(x,\xi)=\frac{w(x;r,\theta)}{r}$ and consequently $k_n(x,\xi)=\frac{w(x;r,\theta)}{r}\mathbb{1}_{[0,n]}(r)$.
Denoting the Riesz potential with kernel k_n as $\mathcal{R}_n$ we can use Theorem 3.37 to claim its compactness from $C_l(\mathbb{R}^2)$ to itself (due to the compact support of k_n).
Next, consider (for $f\in C_l$)

$$|\mathcal{R}_w(f)-\mathcal{R}_n(f)|\leq\int\limits_{\mathbb{R}^2}|k(x,\xi)-k_n(x,\xi)||f(\xi)|d\xi\leq\|f\|_\infty\int\limits_{\mathbb{R}^2}|k(x,\xi)-k_n(x,\xi)|d\xi,$$

and from estimating the last integral

$$\begin{aligned}\int\limits_{\mathbb{R}^2}|k(x,\xi)-k_n(x,\xi)|d\xi&=\int\limits_{-\pi}^{\pi}\int\limits_{0}^{\infty}|1-\mathbb{1}_{[0,n]}(r)|\frac{w(x;r,\theta)}{r}rdrd\theta=\\&=\int\limits_{-\pi}^{\pi}\int\limits_{n}^{\infty}\underbrace{w(x,r,\theta)}_{\sim r^{-\lambda},\ n\gg1}drd\theta\leq cn^{-\lambda+1},\end{aligned}$$

it follows that $\mathcal{R}_n(f)$ converges (uniformly) to $\mathcal{R}_w(f)$ in $C_l(\mathbb{R}^2)$. Furthermore $\|\mathcal{R}_w-\mathcal{R}_n\|_\infty\leq cn^{-\lambda+1}$ and by Theorem 3.36 this implies $\mathcal{R}_w$ is compact on $C_l(\mathbb{R}^2)$. Substituting $f^*=f/w$ from the beginning finishes the proof. □

Remark 3.47. With the above remarks and Theorems 3.35 and 3.38 we have shown, that when dealing with Riesz potentials in integral equations, one can work in the usual L^2 or con-

tinuous functions setting (in Galerkin or collocation schemes) to obtain consistency (or even convergence), as long as the functions decay properly at infinity. But since essential boundedness is a stronger demand than integrable singularities one cannot expect boundedness between L^2 and L^∞ (as shown by a counter example in Stein (1970)).

3.3.2 Collocation Algorithms for Singular Integral Operators

All the results and comments presented in the previous sections show necessary and sufficient conditions for consistency of semi and fully discretized operator equations, without describing how to calculate the unknown approximate solution f_N.

By going back to the collocation equation (3.89) one might further assume f_N to lie in a finite dimensional subspace X_N of X. Such a subspace may then be said to be the span of some basis $\{p_1, \ldots, p_N\}$, such that every element in X_N can be written as

$$f_N = \sum_{j=1}^{N} a_j p_j, \quad a_j \in \mathbb{R}, \quad f_N \in X_N.$$

Being absolutely precise one has to add that the basis functions p_i and the coefficients a_i might depend on the dimension N (cf. Hackbusch (1995)). Obviously, by plugging in the sum for f_N in the collocation equation one obtains

$$\sum_{j=1}^{N} a_j \mathcal{K} p_j(x_i) = g(x_i), \quad \forall x_i, \tag{3.102}$$

where $\mathcal{K}$ was assumed to be linear, such that an equation system can be set up (cf. Hackbusch (1995))

$$\underline{K}\,\underline{a} = \underline{g}, \quad \underline{K} = (K_{ij}) = \mathcal{K} p_j(x_i), \tag{3.103}$$

with the unknowns a_j, which has a solution if and only if $\underline{K}$ is non-singular.

Remark 3.48. Now one advantage of the collocation method compared to Galerkin becomes more accessible, especially for higher dimensions. That is, setting up the matrix $\underline{K}$, which is done by applying the operator of the equation to the basis functions. Evaluating this at certain points gives the collocation matrix, whereas for the Galerkin approach one has to apply a second operator, namely the inner product with another basis function, which, in general, is not a straight forward task.

Remark 3.49. Even in the case of collocation, calculating the matrix entries K_{ij} can be very expensive. Here the difference between integral and differential equations stands out the most. If $\mathcal{K}$ is a combination of differential operators the only task is to differentiate the basis functions, while in the case of integral or integro-differential equations applying the according operators to the basis functions can be a knock out criterion for the whole scheme in terms of calculational costs.

This fact has been neglected in most of the works cited above and hence we shall show some special techniques for the operators in Section 2, equations (2.49) and (2.48), regarding these matrix entries.

The basis for X_N, $\{p_1, \ldots, p_N\}$, has to be chosen wisely in order to obtain a good working numerical scheme. As it is well known, orthogonal basis functions have certain advantages. Also, when finding a set, where its span is dense in X (i.e. a complete set), one can claim the basis functions p_i not to depend on N.

Eventually this might lead to the choice of a complete orthogonal set of polynomials, as has been dealt with in detail in Section 3.2. Thus, if $\{p_i\}$ is a complete orthogonal set in X, an approximate function $f_N \in X_N$ shall be defined as

$$f_N = \mathcal{Q}_N f = \sum_{i=0}^{N} a_i p_i,$$

which can be viewed as an orthogonal projection of f onto X_N.

Note that the subspace X_N is actually $(N+1)$-dimensional, since the sum starts at $i = 0$, which is done to fully relate the practical numerical treatment and algorithm to the results in Section 3.2.

In contrast to choosing an arbitrary basis, by using a set of complete orthogonal polynomials neither all the basis sets $\{p_i\}_{i=0}^{N}$ nor the coefficients of the linear combination forming f_N depend on N (review Section 3.2 for all the necessary details).

Remark 3.50. With the fact that f_N can be found as a projection of f onto f_N we can now provide a precise meaning to some notions appearing when dealing with spectral methods in general.

As derived above what essentially is done when setting up a spectral scheme for a general operator equation is applying projections onto finite dimensional subspaces to arrive at an equation system given by the matrix defined in (3.103), i.e.

$$\mathcal{K}f \;\rightsquigarrow\; \mathcal{P}_N \mathcal{K} \mathcal{Q}_N f$$

Thus one can distinguish the following schemes (replacing $\mathcal{P}_N$, $\mathcal{Q}_N$ above with)

(i) Galerkin: $\mathcal{P}_N \equiv \langle \cdot, q_i \rangle$, $\mathcal{Q}_N \equiv \mathcal{P}_N$ as in Lemma 3.7

(ii) Collocation: $\mathcal{P}_N \equiv \mathcal{I}_N$, $\mathcal{Q}_N \equiv \mathcal{P}_N$ as in Lemma 3.7

(iii) Pseudospectral: $\mathcal{P}_N \equiv \mathcal{I}_N$, $\mathcal{Q}_N \equiv \mathcal{I}_N$

with $\mathcal{I}_N$ as in Lemma 3.18.

From this the main difference between collocation and pseudospectral schemes becomes more obvious. Since $\mathcal{I}_N$ interpolates a function at certain (given) points, one can set up the equation system (3.102) by writing the expansion using Lagrange interpolation polynomials, such that the values of f at the interpolation points are the unknowns (which correlates to the

definition of the discrete inner product). Once those values are found, expansion coefficients can then directly be calculated using the discrete inner product derived in Section 3.2.

One can further claim the collocation to relate to pseudospectral methods via the aliasing error, which is of the same order of magnitude as the approximation error made by truncating the series expansion (see Remark 3.27).

An immediate disadvantage, if one may say so, of the pseudospectral approach is the instability of differentiation, since it cannot be written in terms of derivatives of the basis polynomials, so that one has to use so called *differentiation matrices*. For a theoretical and practical treatment on that subject see Trefethen (2000).

In virtue of equations (2.49) and (2.48), this subsection shall provide a (reasonably) fast working algorithm to obtain the essential system matrix entries, i.e. the combination of operators applied to rational Chebyshev polynomials R_n (in $\mathbb{R}$ and $\mathbb{R}^2$), cf. Section 3.1. Furthermore, the collocation points are chosen to be the zeros of the polynomial R_{N+1} (for an expansion up to R_N, such that the system matrix is square), which are given by Lemma 3.2(ix).

Arguments for taking the zeros of the next higher order polynomial stem from the *Erdös-Turán* theorem, the definition of a variant of the discrete inner product (see Mason & Handscomb (2003) and (3.53)) and its implications (cf. e.g. Theorem 3.24 and Remark 3.34).

To remain as general as possible, we will not present the full scheme, which numerically solves the above mentioned equations, but show how a matrix vector description can be derived for the crucial parts.

The most appropriate function space for collocation methods is $C(\mathbb{R}^n)$ equipped with the $\|\cdot\|_\infty$ norm. Hence, if not otherwise stated, the following calculations are done in this framework (with the additional assumption of the Dini-Lipschitz condition to hold, see Theorem 3.13).

(i) <u>The case $\mathbb{R}$</u>

Let a function f decay fast enough at infinity, such that

$$f \approx f_N(x) = \sum_{i=0}^{N} a_i R_i(x) \quad \text{and} \quad \mathcal{J}^\alpha_{-\infty} f < \infty, \tag{3.104}$$

where $\mathcal{J}^\alpha_{-\infty}$ is the Abel integral operator as given in (3.95). From Theorems 3.32 and 3.33 and Remark 3.41 *consistency* of a collocation projection can be readily proved.

When plugging in f_N, terms such as $\mathcal{J}^\alpha_{-\infty} R_i$ appear in the sum, which do not exist $\forall i$ (see Remark 3.1). Thus, a weight $w_\beta(x) := (1+x^2)^{-\beta}$ has to be found, such that

$$\mathcal{J}^\alpha_{-\infty}(w_\beta R_i) < \infty \quad \text{and} \quad \left\| \frac{f}{w_\beta} - \mathcal{Q}_N \frac{f}{w_\beta} \right\|_\infty \to 0, \tag{3.105}$$

essentially meaning that f/w_β remains at least bounded.

Remark 3.51. Obviously, the requirements in (3.105) impose some constraints on the decay of f, i.e. say $f \sim |x|^a$ as $|x| \to \infty$, then

$$a + 2\beta \leq 0 \quad \text{and} \quad 1 - \alpha + 2\beta > 1 \quad \Rightarrow \quad a < -\alpha, \quad \beta > \alpha/2$$

and since $0 < \alpha < 1$ (by definition), an asymptotic behavior of $f \sim 1/x$ is sufficient (cf. (2.35)). Restrictions on the decay of the given functions and the use of weights are due to the unbounded domain (cf. main theorems in Section 3.2).

In concrete, evaluating the integral in (3.105) at the collocation points yields the matrix entry K_{ij} (as symbolized in (3.103))

$$\begin{aligned} K_{ij} = \mathcal{J}^{\alpha}_{-\infty}(w_\beta R_i)(x_j) &= \int_{-\infty}^{x_j} (x_j - \xi)^{\alpha-1} \frac{1}{(1+\xi^2)^\beta} R_i(\xi)\, d\xi = \\ &= \int_{-\infty}^{x_j} (x_j - \xi)^{\alpha-1} \frac{1}{(1+\xi^2)^\beta} \cos(i\phi(\xi))\, d\xi, \quad \forall\, 0 \leq i,j \leq N \end{aligned} \tag{3.106}$$

with ϕ taken from the definition in (3.2).

Remark 3.52. It is fairly straight forward to see that a closed formula for all K_{ij} above cannot be found in general and also that this matrix does not have any special properties, such as sparseness, triangular shapes or symmetries. It is for these facts that analyzing the equation systems for condition numbers and possible inversions for Abel integral operators (on the real line) combined with spectral methods becomes heavily involved. Hence, a quadrature scheme has to be applied to obtain the system matrix. Additionally, when considering the integral boundaries, the kernel and the integrand, standard numerical integration might not be practicable. This is in sharp contrast to integral operators with simple kernel functions acting on a bounded interval and, of course, differential equations.

Categorizing the whole term in (3.106) one can find three characteristics, which need special treatment

a) unbounded domain

b) (weakly) singular kernel

c) (high) oscillatory integrand,

and since (in general) one has to perform such an integration $(N+1)^2$ times, a reasonably fast and accurate scheme is crucial. There are existing routines, which employ e.g. QUADPACK (see Piessens *et al.* (1983) and the implementations in the *NAG* packages), tackling the characteristics *a*) and *b*). Since the algorithms taking care of the singularity can only be applied to integrals over bounded intervals and the others mapping the infinite integration limit only allow for at least bounded functions, the interval $(-\infty, x_j]$ in (3.106) has to be split into $(-\infty, x_0] \times [x_0, x_j]$.

Remark 3.53. The routine for the infinite range uses algebraic mappings of $(-\infty, x_0]$, $[x_0, \infty)$ or the whole line onto $[0, 1]$, such that the integral can be approximated by a higher order *Gauss-Kronrod* scheme. The necessary decay of the given integrand (for the original integral to exist) then shall cancel out the singular Jacobi determinant of the coordinate transform. Alternatively, for certain values of α, β and x_0 one might be able to evaluate the integral analytically when substituting the asymptotic behavior of the polynomials, derived in Lemma 3.2(iii). As for choosing x_0, it is obvious that $x_0 < x_j$, $\forall j$. With Lemma 3.2(ix) the collocation points can be confounded to a bounded interval for a given N, such that x_0 just has to lie outside this interval. Additionally, this means for all polynomial degrees, indicated by i in the cosine function in (3.106), the integrand does *not* oscillate on $(-\infty, x_0]$. Hence, the QUADPACK routine works fast and accurate in such cases.

Remark 3.54. Weak end-point singularities, such as the one in the general Abel integral operator, are very common among integral kernels and hence, existing quadrature schemes can be easily found in numerical packages. The algorithm used here starts by bisecting the interval $[x_0, x_j]$ and applies a modified *Clenshaw-Curtis* method to the sub-interval containing the singularity and a *Gauss-Kronrod* integration to the remaining part. The singularities have to be provided in the form $(x_j - \xi)^\alpha (\xi - x_0)^\beta$, with $\alpha, \beta > -1$, which is essentially the same as the Abel kernel in (3.106) ($\beta = 0$). Thus, one can expect a highly accurate result for these types of operators.

Remark 3.55. Up to a polynomial degree of $N \approx 200$ the above mentioned routines work perfectly fast and accurate for all K_{ij} needed. If one needs more polynomials the oscillations of the integrand render the algorithm for the singularity unfeasible. Also, it is not suitable to move x_0 closer to x_j, because the scheme dealing with the infinite interval yields unacceptable results. There is a routine in QUADPACK taking care of such situations using *Gauss* 30-points and *Kronrod* 61-points rules, but it slows down the calculations significantly and hence is not recommendable for non-sparse system matrices.

Applying the coordinate transform $y = \phi(\xi) = \arctan(\xi) - \pi/2,\ \phi'(\xi) = (1 + \xi^2)^{-1}$, to (3.106) we obtain

$$
\begin{aligned}
K_{ij} &= \int_{-\pi}^{\phi(x_j)} (x_j - \phi^{-1}(y))^{\alpha-1} \frac{1}{(1 + (\phi^{-1}(y))^2)^\beta} \cos(iy) \frac{dy}{\phi'(\phi^{-1}(y))} = \\
&= \int_{-\pi}^{\phi(x_j)} \underbrace{(x_j - \phi^{-1}(y))^{\alpha-1} (1 + (\phi^{-1}(y))^2)^{1-\beta}}_{=:h(y)} \cos(iy)\, dy = \qquad (3.107) \\
&= \int_{-\pi}^{\phi(x_j)} h(y)\, \cos(iy)\, dy \to 0 \quad \text{as} \quad i \to \infty,
\end{aligned}
$$

which holds due to the Riemann-Lebesgue lemma, provided $h \in L^1([-\pi, \phi(x_j)])$. To see this, observe that $\phi(x_j) < 0$, $\forall x_j$ and hence h is positive on the considered interval, then

$$\|h\|_{L^1} = \int_{-\pi}^{\phi(x_j)} h(y)dy = \int_{-\infty}^{x_j} h(\phi(\xi))\phi'(\xi)\, d\xi = \int_{-\infty}^{x_j} (x_j - \xi)^{\alpha-1} \frac{1}{(1+\xi^2)^\beta}\, d\xi < \infty,$$

obviously yielding the "phase function"-version of the Riemann-Lebesgue lemma, where the phase ϕ has to be differentiable and non-constant within the integration limits.

Mentioned in Remark 3.55, numerical techniques for high oscillatory functions might not be fast and work only for bounded integrands. The result in (3.107) suggests an asymptotic expansion of such integrals for high polynomial degrees. Erdélyi (1956) provides some formulae regarding Fourier integrals and the method of the *stationary phase*. Although, strictly speaking, the integral in (3.106), seen as the real part of a Fourier integral, does not have a stationary phase, i.e. $\phi'(\xi) = 0 \;\Leftrightarrow\; |\xi| \to \infty$, the results are still applicable.

Defining a *stationary point of order m* to be a point x, for which a function ϕ satisfies $\phi'(x) = \cdots = \phi^{(m)}(x) = 0$, $\phi^{(m+1)}(x) \neq 0$, one can state

Lemma 3.39. *Given the interval $[a, b]$ and a differentiable function ϕ, increasing on $[a, b]$, where a, b are either ordinary points or stationary points of some order, such that*

$$\phi'(\xi) = (\xi - a)^{\rho-1}(b - \xi)^{\sigma-1}\phi_1(\xi),$$

where $\rho, \sigma \geq 1$ and $\phi_1 \in C^n([a, b])$ and positive. If $\lambda > 0$ and $\mu \leq 1$ and the function $h \in C^n([a, b])$, then

$$\int_a^b h(\xi)(\xi - a)^{\lambda-1}(b - \xi)^{\mu-1}e^{im\phi(\xi)}d\xi = B(m) - A(m), \tag{3.108}$$

where $A(m) \sim A_n(m)$ and $B(m) \sim B_n(m)$ to n terms as $m \to \infty$, with

$$\begin{aligned} A_n(m) &= -\sum_{k=0}^{n-1} \frac{u^{(k)}(0)}{k!\rho}\Gamma\left(\frac{k+\lambda}{\rho}\right)\exp\left(\frac{i\pi(k+\lambda)}{2\rho}\right) m^{-(k+\lambda)/\rho}\, e^{im\phi(a)} \\ B_n(m) &= -\sum_{k=0}^{n-1} \frac{v^{(k)}(0)}{k!\sigma}\Gamma\left(\frac{k+\mu}{\sigma}\right)\exp\left(\frac{-i\pi(k+\mu)}{2\sigma}\right) m^{-(k+\mu)/\sigma}\, e^{im\phi(b)}, \end{aligned} \tag{3.109}$$

defining the functions u and v via

$$\begin{aligned} \zeta^\rho &:= \phi(\xi) - \phi(a), \quad u(\zeta) = h(\xi)(\xi - a)^{\lambda-1}(b - \xi)^{\mu-1}\zeta^{1-\lambda}\frac{d\xi}{d\zeta} \\ \eta^\sigma &:= \phi(b) - \phi(\xi), \quad v(\eta) = h(\xi)(\xi - a)^{\lambda-1}(b - \xi)^{\mu-1}\eta^{1-\mu}\frac{d\xi}{d\eta}. \end{aligned}$$

Proof. see Erdélyi (1956) □

Comparing (3.106) with (3.108) one obtains $\lambda = 1$, $b = x_j$, $a = x_0$, $\mu = \alpha$, $h(\xi) = 1/(1+\xi^2)^\beta$, from ϕ' it is obvious to get $\rho = \sigma = 1$ and by taking the real part in (3.108),(3.109) the cosine integral remains (keeping m as the polynomial degree instead of i). Then,

$$\begin{aligned}
\zeta = \phi(\xi) - \phi(x_0) = \arctan(\xi) - \arctan(x_0), \quad \xi(\zeta) = \tan(\zeta + \arctan(x_0)) \quad \Rightarrow \\
u(\zeta) = \frac{1}{(1+\xi(\zeta)^2)^\beta}(x_j - \xi(\zeta))^{\alpha-1}\frac{1}{\cos^2(\zeta + \arctan(x_0))} \\
\eta = \phi(x_j) - \phi(\xi) = \arctan(x_j) - \arctan(\xi), \quad \xi(\eta) = \tan(\arctan(x_j) - \eta) \quad \Rightarrow \\
v(\eta) = -\frac{1}{(1+\xi(\eta)^2)^\beta}(x_j - \xi(\eta))^{\alpha-1}\eta^{1-\alpha}\frac{1}{\cos^2(\arctan(x_j) - \eta)},
\end{aligned} \tag{3.110}$$

such that

$$\begin{aligned}
A_n(m) &= -\sum_{k=0}^{n-1} \frac{u^{(k)}(0)}{k!}\Gamma(k+1)\, m^{-(k+1)}\, \cos(m\phi(x_0) + \pi(k+1)/2) \\
B_n(m) &= -\sum_{k=0}^{n-1} \frac{v^{(k)}(0)}{k!}\Gamma(k+\alpha)\, m^{-(k+\alpha)}\, \cos(m\phi(x_j) - \pi(k+\alpha)/2).
\end{aligned} \tag{3.111}$$

The following example shall provide some actual calculations using these formulae for the Abel operator and the weight given in (2.51).

Example 3.8. Consider the Abel operator $\mathcal{J}^\alpha_{-\infty}$, cut off at some x_0, for weighted rational Chebyshev polynomials, with $\alpha = 3/4$ and the weight $1/\sqrt{1+x^2}$, i.e. the term

$$K_{mj} = \int_{x_0}^{x_j} (x_j - \xi)^{-1/4} \frac{1}{(1+\xi^2)^{1/2}}\, \cos(m\phi(\xi))\, d\xi, \tag{3.112}$$

such that (3.110) reads

$$\begin{aligned}
u(\zeta) &= \frac{1}{(1+\tan^2(\zeta + \arctan(x_0)))^{1/2}}(x_j - \tan(\zeta + \arctan(x_0)))^{-1/4}\frac{1}{\cos^2(\zeta + \arctan(x_0))} \\
v(\eta) &= -\frac{1}{(1+\tan^2(\arctan(x_j) - \eta))^{1/2}}(x_j - \tan(\arctan(x_j) - \eta))^{-1/4}\frac{\eta^{1/4}}{\cos^2(\arctan(x_j) - \eta)},
\end{aligned}$$

and with

$$\begin{aligned}
u(0) &= \frac{1}{(1+x_0^2)^{1/2}}(x_j - x_0)^{-1/4}\frac{1}{\cos^2(\arctan(x_0))} = \frac{(1+x_0^2)^{1/2}}{(x_j - x_0)^{1/4}} \\
v(0) &= -(1+x_j^2)^{1/2}\lim_{\eta\to 0}\frac{\eta^{1/4}}{(x_j - \tan(\arctan(x_j) - \eta))^{1/4}} = -(1+x_j^2)^{1/4},
\end{aligned}$$

the first terms ($k = 0$) in the sums in (3.111) can be given as

$$A_1(m) = -\frac{(1+x_0^2)^{1/2}}{(x_j - x_0)^{1/4}}\, m^{-1}\, \cos(m\phi(x_0) + \pi/2)$$
$$B_1(m) = (1+x_j^2)^{1/4}\Gamma(3/4)\, m^{-3/4}\, \cos(m\phi(x_j) - 3\pi/8).$$

The data in Table 17 shall demonstrate how the accuracy of $B_1 - A_1$ depends on x_0 compared to evaluating K_{mj} in (3.112) via a Gauss-Kronrod scheme.
Fix $x_j = 10$, $m = 200$, then the differences $e_1 := |K_{mj} - (B_1 - A_1)|$ and $e_2 := |K_{mj} - (B_1 + B_2 - A_1 - A_2)|$ are calculated. The irregular changes in the difference in Table 17 for

x_0	e_1	e_2
0	1.7×10^{-3}	8.4×10^{-5}
-10	2.2×10^{-3}	1.6×10^{-4}
-50	1.2×10^{-2}	2×10^{-3}
-100	7×10^{-3}	3×10^{-2}
-200	1×10^{-1}	2×10^{-4}

Table 17: Difference approximating K_{mj} in (3.112) via Gauss-Kronrod and asymptotic expansions

$x_0 = -100$ and $x_0 = -200$ stem from the distribution of the zeros of the polynomial R_{200}. Due to Lemma 3.2(ix) the smallest zero lies near $x \approx 130$, such that taking x_0 close to or lower than this value means that there is less direct cancellation in the integral from the oscillations of the integrand. In other words, one is getting close to the stationary point (at infinity), which, apart from the singularity, contributes most to the integral.

In Section 2 another operator term occurs in equations (2.34) and (2.49), given more generally as

$$\mathcal{J}_\infty^\alpha \partial_x^r f = \int_x^\infty (\xi - x)^{\alpha-1}\, \partial_\xi^r f(\xi)\, d\xi < \infty \quad \text{if} \quad f \in L_w^2(\mathbb{R}) \cap C^r(\mathbb{R}),$$

such that, in contrast to (3.104), one does not have to impose restrictions on the function's decay behavior (if $r \geq 2$). Requiring f in $L_w^2(\mathbb{R}) \cap C^r(\mathbb{R})$ is sufficient for the existence of the integral and since $H_{w,A}^r(\mathbb{R}) \subset L_w^2(\mathbb{R})$ (cf. (3.19) and Lemma 3.9) this space would also take care of the differentiability and integrability demands.

As has been done above, system matrix entries

$$K_{ij} = \mathcal{J}_\infty^\alpha(R_i^{(r)})(x_j) = \int_{x_j}^\infty (\xi - x_j)^{\alpha-1} R_i^{(r)}(\xi)\, d\xi, \tag{3.113}$$

have to be approximated with quadrature schemes.

Caveat: Although weight functions need not necessarily be introduced in (3.113), since (cf. Lemma 3.2(iii)) higher derivatives of rational Chebyshev polynomials have sufficient decay at infinity. But, due to the combination of both Abel operators in (2.49), one actually has $K_{ij} = \mathcal{J}_\infty^\alpha[(w_\beta R_i)^{(r)}](x_j)$ (since the expansion of the unknown contains a weight function).

Remark 3.56. Obviously, the issues *a*) - *c*) arise again in (3.113), maybe even in a more severe form, since the terms are (algebraically) more complicated. Also, derivatives of R_i are not bounded by ± 1 and hence steeper gradients, due to oscillations, appear in the integrand. Asymptotic expansions might be applied again, whereas the whole calculation becomes much more involved. With Lemma 3.2(viii) one can circumvent derivatives of the polynomials by using sums over R_i, but as the degree grows, this is not recommendable, since error accumulation might occur.

Overall, with derivatives of $w_\beta R_i$ given as closed formulae, the above mentioned QUADPACK routines work to satisfaction for (3.113) up to a degree of $i = 200$.

(ii) The case $\mathbb{R}^2$

In sharp contrast to the one-dimensional case, routines (provided by numerical libraries) for fast and accurate multi-dimensional quadrature are rare and the ones existing treat almost exclusively bounded integrands over bounded regions. Hence, when using spectral methods involving operators such as the potential integral defined in (3.98), alternatives have to be found (especially for high oscillatory integrands).

We start by saying $f \in C(\mathbb{R}^2)$, decaying fast enough, such that

$$f \approx f_N = \sum_{i=0}^{N} \sum_{k=0}^{N} a_{ik} R_i R_k \quad \text{and} \quad \mathcal{R}^1 f < \infty,$$

with $\mathcal{R}^1$ being the operator defined in (3.99) and where Theorems 3.35 and 3.38 and Remark 3.46 provide the consistency of a collocation approach.

Similar to the Abel operator above, substituting the expansion for f_N in $\mathcal{R}^1 f$ yields integrals over $R_i R_k$, which do not exist in general (see e.g. Theorem 3.34). Again, weights have to be found in a way analogously to (3.105) to obtain the system matrix entries (collocated at the zeros (x_{1_j}, x_{2_l}) of R_{N+1}, hence $x_{1_j} = x_{2_l}$ if $j = l$)

$$K_{ijkl} := \mathcal{R}^1(w_\beta w_\gamma R_i R_k)(x_{1_j}, x_{2_l}) = \int_{\mathbb{R}^2} \frac{w_\beta(\xi_1) w_\gamma(\xi_2)}{|(x_{1_j}, x_{2_l}) - (\xi_1, \xi_2)|} R_i(\xi_1) R_k(\xi_2)\, d\xi, \tag{3.114}$$

with w_β given as in the one-dimensional case and analogously $w_\gamma(x_2) = (1 + x_2^2)^{-\gamma}$.

Obviously, this is not the only choice to introduce a weight, $w(x_1, x_2) = (1 + |(x_1, x_2)|)^{-\lambda}$ has been shown in Theorem 3.35 to be more appropriate. The reason to separate the weight into individual independent variables is due to the connection to the orthogonality relation (3.11). Remark 3.46 showed $\lambda > 1$ is necessary for existence of the integral and hence is $\beta + \gamma > 1/2$.

The following provides a description of a fast working, sufficiently accurate algorithm to calculate the matrix entries, which by categorizing the terms in (3.114) are prone to the same issues mentioned on page 189 for the one-dimensional case (with the additional severity of evaluating the integrals $(N+1)^4$ times).

Defining $B_\epsilon(x) := \{\xi \in \mathbb{R}^2 : |x-\xi| \leq \epsilon\}$ and dividing the integration domain yields ($k(x,\xi) := |x-\xi|^{-1}$)

$$\begin{aligned} K_{ijkl} &= \Bigg[\int\limits_{\mathbb{R}^2 \setminus B_\epsilon(x)} + \int\limits_{B_\epsilon(x)} \Bigg] k(x,\xi) w_\beta(\xi_1) w_\gamma(\xi_2) R_i(\xi_1) R_k(\xi_2)\, d\xi = \\ &= \int\limits_{\mathbb{R}^2} \underbrace{k(x,\xi) \mathbb{1}_{\mathbb{R}^2 \setminus B_\epsilon(x)}(\xi)}_{=:k^*(x,\xi)} [w_\beta w_\gamma R_i R_k](\xi)\, d\xi + \int\limits_{B_\epsilon} k(x,\xi) [w_\beta w_\gamma R_i R_k](\xi)\, d\xi, \end{aligned} \tag{3.115}$$

with $k^* \in L^2_w(\mathbb{R}^2)$ as a new kernel, which is *discontinuous* but *bounded*. Hence, there exists an L^2 convergent expansion of k^* into rational Chebyshev polynomials, but without pointwise convergence due to the Gibbs phenomenon (as has been presented in Section 3.2). On the other hand, those oscillations are damped based on the fact that we integrate this expansion over the whole domain (cf. Theorem 3.27 and Remark 3.31). Thus, say

$$k^*(x,\xi) \approx w_\mu(\xi_1) w_\nu(\xi_2) \sum_{m=0}^{M} \sum_{p=0}^{M} b_{mp}(x) R_m(\xi_1) R_p(\xi_2), \tag{3.116}$$

where the weights are chosen, such that $k^*/(w_\mu w_\nu)$ is bounded.

Remark 3.57. It is possible, by finding a *continuous extension* of k^* over the domain B_ϵ, to avoid Gibbs oscillations. But, in order to have an actual advantage, such an extension would have to be simple enough, to evaluate its integral analytically (otherwise one would artificially introduce another quadrature error). In one-dimension this is always possible by just connecting the boundary points of B_ϵ by a straight line. In the present case, one would have to find a two-dimensional continuous surface, containing ∂B_ϵ. As has been mentioned above, the integral prevents the Gibbs phenomenon from destroying the pointwise convergence of the overall approximation, but since the amplitude of the oscillations grows proportionally to the step height of the discontinuity, the larger ϵ, the higher the accuracy for finite (small) M.

By substituting (3.116) in (3.115), the first integral on the right hand side then reads

$$\int\limits_{\mathbb{R}^2} k^*(x,\xi) [w_\beta w_\gamma R_i R_k](\xi)\, d\xi \approx \sum_{m=0}^{M} \sum_{p=0}^{M} b_{mp}(x) \int\limits_{\mathbb{R}^2} [w_\mu w_\nu R_m R_p\, w_\beta w_\gamma R_i R_k](\xi)\, d\xi.$$

Setting $\mu + \beta = 1$ and $\nu + \gamma = 1$ yields the orthogonality weights (cf. Theorem 3.1 and equation (3.11)) in both variables, such that the right hand side simplifies to

$$\int_{\mathbb{R}^2} k^*(x,\xi)[w_\beta w_\gamma R_i R_k](\xi_1,\xi_2)\,d\xi \approx b_{ik}(x)\|R_i\|_w^2\|R_k\|_w^2. \tag{3.117}$$

The existence of such combinations of weights is straight forward, for example take $\beta = \gamma = 3/4$ and $\mu = \nu = 1/4$, such that $\beta + \gamma > 1/2$ and applying the usual polar coordinates shows $k^* \sim 1/r$, whereas $w_{1/4} \sim 1/\sqrt{r}$.

Remark 3.58. As established in Theorem 3.6 the coefficients in (3.117) above are given via

$$b_{ik} = \|R_i\|_w^{-2}\|R_k\|_w^{-2}\langle \tfrac{k^*}{w_\mu w_\nu}, R_i R_k\rangle_u \tag{3.118}$$

and by substituting this into (3.117), one can immediately see that we actually did not apply any series expansion, but merely identified the integral in (3.115) as a coefficient of a Chebyshev series. For the sake of convergence arguments it is easier to consider it in the way presented above. Much more importantly, one can now apply Lemma 3.19 to (3.118) to see that evaluating the first integral on the right hand side in (3.115) reduces to an application of the FFT (or DCT) algorithm in ξ_1 and ξ_2 consecutively, i.e.

$$b_{ik}(x) \approx \tfrac{N}{2} DCT^2\left(\frac{k^*(x)}{w_\mu w_\nu}(\xi_{1_r},\xi_{2_s})\right), \qquad r,s = 0,\dots,N,$$

with (ξ_{1_r},ξ_{2_s}) as in (3.53). Defining the matrix $(k_{rs})_{r,s=0}^N := (\frac{k^*(x)}{w_\mu w_\nu}(\xi_{1_r},\xi_{2_s}))_{r,s}$, the symbol DCT^2 then means applying the transform two times (cf. proof of Lemma 3.19), i.e.

$$(l_{rs}) := DCT\big((k_{rs})\big) = \big(DCT(k_{r0}),\dots,DCT(k_{rN})\big), \qquad DCT^2\big(k_{rs}\big) = \Big(DCT\big((l_{rs})^T\big)\Big)^T,$$

where T stands for the transposed matrix (DCT shall always act on the either the columns or the rows).

Remark 3.59. With using the FFT, a slight disadvantage occurs. As k^* is a function of ξ and x, the coefficients in (3.116) depend on the latter, which means that the discontinuity moves in $[-\frac{2}{3}(N+1), \frac{2}{3}(N+1)]$, since the collocation points lie within this interval (cf. Lemma 3.2(ix)) and so do the points of evaluation in the FFT algorithm (according to Lemma 3.19). Hence, the approximate coefficients become more inaccurate the farther x moves from $(0,0)$. Taking $M \approx 10N$ (which is almost negligible in terms of calculational costs for the FFT) can alleviate this fact.

To obtain a full discretization of the collocation matrix entries in (3.115), it is left to evaluate the second integral on the right-hand side, containing the singularity. There are, of course, several ways to calculate such integrals, since the kernel is integrable over any

bounded domain. The most obvious one is to use polar coordinates (cf. (3.101)), yielding

$$\int_{B_\epsilon} k(x,\xi)[w_\beta w_\gamma R_i R_k](\xi_1,\xi_2)\,d\xi = \int_0^\epsilon \int_{-\pi}^{\pi} [w_\beta w_\gamma R_i R_k](r,\theta;x) d\theta dr.$$

The whole integrand above is smooth and given by a closed formula, which can be easily used for existing cubature packages. Although, one is advised to use adaptive points of evaluation and cubature weights, due to the different behavior of the weighted polynomials in polar coordinates with respect to x and the degrees i,k. Depending on the package, this can lead to longer computation times.

Remaining in Cartesian coordinates, one can also apply a finite (equidistant) grid on $B_\epsilon(x)$ and approximate the weighted polynomials by locally constant functions on the grid points (where the grid and the distances might be adapted with respect to x). Thus, one obtains

$$\int_{B_\epsilon} k(x,\xi)[w_\beta w_\gamma R_i R_k](\xi_1,\xi_2)\,d\xi \approx \sum_{r=0}^{L}\sum_{s=0}^{L} [w_\beta w_\gamma R_i R_k](\xi_r,\xi_s) \int_{\Omega_r}\int_{\Omega_s} k(x,\xi)d\xi,$$

Ω_r, Ω_s representing the according sections of the grid, see Section 3.4 for more details on such methods. Here, again, one has to be careful with generating the grid, as the weighted polynomials become high oscillatory and thus a piecewise constant approximation can become highly inaccurate (depending on L). The advantage definitely is that all elements (as functions of x) can be calculated using matrix vector multiplication, which is considerably fast (although one generally has a non-sparse matrix of integration weights).

Remark 3.60. Since $w_\gamma R_k, w_\beta R_i \in C^\infty(\mathbb{R})$ one can find a convergent Taylor series expansion around x_1 and x_2 of both functions, such that the B_ϵ integral in (3.115) reduces to the problem of calculating terms of the form

$$\int_{B_\epsilon} k(x,\xi)(x_1-\xi_1)^a(x_2-\xi_2)^b d\xi, \quad a,b\in\mathbb{N},$$

and the according Taylor series coefficients, where it is easily seen that the integral terms are zero if a is even or b is odd. The problem with this strategy is not only the high programming effort (calculating all those terms and coefficients, as well as explicitly programming the resulting Taylor sum), but also the accuracy, or better to say, the large number of expansion terms needed to reach the necessary accuracy, even for small ϵ. Furthermore, this highly depends on the degree of the polynomials and so, overall, a Taylor series approximation is not recommendable.

At last, if we want to utilize the FFT again, we need the integrand to be bounded, such that it can be expanded into a (suitably convergent) Chebyshev series. For this, what is known as *subtraction of the singularity* (often used for Cauchy principal value integrals, cf.

Diethelm (2000)), shall be applied, i.e.

$$\int_{B_\epsilon} k(x,\xi)[w_\beta w_\gamma R_i R_k](\xi_1,\xi_2)\,d\xi =$$
$$= [w_\beta w_\gamma R_i R_k](x) \underbrace{\int_{B_\epsilon} k(x,\xi)d\xi}_{<\infty} + \int_{B_\epsilon} \underbrace{k(x,\xi)\big([w_\beta w_\gamma R_i R_k](\xi_1,\xi_2) - [w_\beta w_\gamma R_i R_k](x_1,x_2)\big)}_{\text{bounded on } B_\epsilon(x)}\,d\xi,$$

where it is straight forward to verify the boundedness of the integrand in the second integral, since the difference of the weighted polynomials tends to zero faster than the kernel tends to infinity as ξ approaches x. Now, the first term on the right-hand side can be given in closed form, whereas the second term can be either calculated using a cubature scheme or by expanding the bounded (but probably discontinuous) integrand into a (classical) Chebyshev series. Thus, say

$$k(x,\xi)\big([w_\beta w_\gamma R_i R_k](\xi_1,\xi_2) - [w_\beta w_\gamma R_i R_k](x_1,x_2)\big) \approx \sum_{r=0}^{L}\sum_{s=0}^{L} c_{rs}(x)T_r(\xi_1)T_s(\xi_2), \qquad (3.119)$$

and by plugging this into the second integral above, we get

$$\sum_{r=0}^{L}\sum_{s=0}^{L} c_{rs}(x)\int_{B_\epsilon} T_r(\xi_1)T_s(\xi_2)d\xi = \sum_{r=0}^{L}\sum_{s=0}^{L} c_{rs}(x)\int_{x_1-\epsilon}^{x_1+\epsilon} T_r(\xi_1)d\xi_1 \int_{x_2-\epsilon}^{x_2+\epsilon} T_s(\xi_2)d\xi_2, \qquad (3.120)$$

where, for the sake of computability, the domain $B_\epsilon(x)$ was taken as a square with center x and side length 2ϵ (which is also advantageous when expanding k^* in (3.115)). Again, as in (3.116), the coefficients c_{rs} can be calculated using the FFT and the integrals are given via the formula in Lemma 3.3.

Overall, we managed to approximate the matrix entries defined in (3.114) by closed formulae and only using the FFT, such that the whole approximation can be easily programmed, is fast working and sufficiently accurate.

Remark 3.61. It is also possible to subtract the singularity without splitting the integral. Consider (3.114) in the following way

$$\int_{\mathbb{R}^2} \frac{w_\beta(\xi_1)w_\gamma(\xi_2)}{|(x_{1_j},x_{2_l})-(\xi_1,\xi_2)|}\left(R_i(\xi_1)R_k(\xi_2) - R_i(x_1)R_k(x_2) + R_i(x_1)R_k(x_2)\right)d\xi_1 d\xi_2 =$$
$$= R_i(x_1)R_k(x_2)\int_{\mathbb{R}^2} \frac{w_\beta(\xi_1)w_\gamma(\xi_2)}{|(x_{1_j},x_{2_l})-(\xi_1,\xi_2)|}d\xi +$$
$$+ \int_{\mathbb{R}^2} w_\beta(\xi_1)w_\gamma(\xi_2)\frac{R_i(\xi_1)R_k(\xi_2) - R_i(x_1)R_k(x_2)}{|(x_{1_j},x_{2_l})-(\xi_1,\xi_2)|}d\xi,$$

where the weights are chosen, such that the first integral on the right hand side exists, and it is easy to verify the boundedness of the integrand in the second integral. Expanding this integrand into a rational Chebyshev series and plugging in this expansion, as done in (3.120), yields integrals over those polynomials, which do not exist (cf. Example 3.1). Thus, a weighted expansion has to be sought, where the weights have to be integrable over $\mathbb{R}$, e.g. $w(\xi_i) = 1/(1+\xi_i^2)^{1/2+\epsilon}$, which would satisfy all conditions if e.g. $\beta = \gamma = 3/4$. In contrast to the algorithm when splitting the integral, one cannot exploit the orthogonality here and also, depending on the used weight functions, a closed formula (as Lemma 3.3 provides for the terms in (3.120)) might not be obtainable for the integrals over weighted rational polynomials.

Next, we present how the above derived algorithm can be adapted when derivatives of the argument function are involved, i.e. terms of the form $\mathcal{R}^1(\partial_x^n f)$. This is, for example, in accordance to equations (2.32) and (2.48), where we have $\mathcal{R}^1([\partial_{x_1}^3 + \partial_{x_1}\partial_{x_2}^2]f)$.

To again be able to exploit the orthogonality and utilize the FFT, we first (formally) shift the derivatives onto the kernel using integration by parts, i.e.

$$\mathcal{R}^1(\partial_{x_1}^n \partial_{x_2}^m f) = \int_{\mathbb{R}^2} k(x,\xi)\partial_{\xi_1}^n \partial_{\xi_2}^m f(\xi) d\xi = (-1)^{m+n} \int_{\mathbb{R}^2} \underbrace{\partial_{\xi_1}^n \partial_{\xi_2}^m k(x,\xi)}_{=:k_1(x,\xi)} f(\xi) d\xi, \tag{3.121}$$

where we then continue from (3.115). When considering the decay behavior of k_1 (with $m, n \geq 1$) it becomes clear that in this case no weights in the expansion for f are needed, because if one expands k_1 similarly to (3.116), one can set $\mu = \nu = 1$ and thus obtains the orthogonality weights.

The existence of the integral in (3.121) is not straight forward due to the stronger singularity in k_1, say r^{-p}, $p > m$ on $\mathbb{R}^m$ when using the usual polar description. Considering the assumptions in Theorem 3.40 below, condition (i) therein then has to be replaced by a *Hölder condition* on f, where the Hölder exponent depends on p and m, which becomes obvious from the proof given in Mikhlin & Prößdorf (1980) (see a similar result mentioned in Diethelm (2000)).

In case of the equations (2.32) and (2.48) the new kernel reads $k_1 = [\partial_{x_1}^3 + \partial_{x_1}\partial_{x_2}^2]k$, where the resulting characteristic defined in Theorem 3.40 does satisfy (3.125). So, trivially, when splitting the integral and extracting the singularity, it is clear that the integral of $k_1 \mathbb{1}_{\mathbb{R}^2 \setminus B_\epsilon} R_i R_k$ exists.

Remark 3.62. If $m = 1$, $n = 0$ or vice versa, one obtains

$$\mathcal{R}^1(\partial_{x_1} f) = -\int_{\mathbb{R}^2} \frac{x_1 - \xi_1}{((x_1-\xi_1)^2 + (x_2-\xi_2)^2)^{3/2}} f(\xi) d\xi, \tag{3.122}$$

which is related to the definition of the *n Riesz transforms* for $f \in L^p(\mathbb{R}^n)$, $1 \leq p < \infty$ (see e.g. Stein (1970))

$$\mathcal{R}_j(f)(x) := \lim_{\epsilon \to 0} \int\limits_{|y| \geq \epsilon} \frac{y_j}{|y|^{n+1}} f(x-y)dy, \quad j = 1, \ldots, n, \tag{3.123}$$

where y_j is the jth component of y. Such integrals have already been dealt with in a seminal paper by Calderon & Zygmund (1952) and also by Mikhlin & Prößdorf (1980), where the following theorem was proved.

Theorem 3.40. *Let $r = |y - x|$, $\theta = (y - x)/r$. Then the singular integral*

$$v(x) = \int\limits_{\mathbb{R}^n} r^{-n} u(x, \theta) f(y) dy \tag{3.124}$$

together with the assumptions

(i) in every ball $B_R := \{y : |y - x| \leq R\}$ the modulus of continuity of f satisfies the Dini condition

$$\int\limits_0^t \tau^{-1} \omega(f, \tau) d\tau < \infty, \qquad \omega(f, t) = \sup_{|y - y_0| \leq t} |f(y) - f(y_0)|, \ y, y_0 \in B_R$$

(ii) for large $|x|$, $f(x) = O(|x|^{-k})$, $k > 0$, holds and

(iii) the characteristic u is bounded and, for fixed x, continuous with respect to θ,

exists, if and only if

$$\int\limits_S u(x, \theta) dS = 0, \tag{3.125}$$

where S denotes the unit sphere and θ varies in S.

Proof. see Mikhlin & Prößdorf (1980) □

Remark 3.63. The singular integral (3.124) in the theorem above can be easily linked to the Riesz transforms (3.123) and further, to (3.122). Therein, when changing to polar coordinates one obtains the kernel

$$\frac{x_1 - \xi_1}{((x_1 - \xi_1)^2 + (x_2 - \xi_2)^2)^{3/2}} = \frac{\cos(\theta)}{r^2} \quad \Rightarrow \quad u(x, \theta) = \cos(\theta) \quad \text{where} \quad \int\limits_{-\pi}^{\pi} \cos(\theta) d\theta = 0.$$

Thus, if f in (3.122) decays to zero at infinity, Theorem 3.40 applies to the whole algorithm in the case of the Riesz potential combined with derivatives (assuming f to be smooth enough, such that every derivative satisfies the Dini condition).

Remark 3.64. The theorem above can be generalized to the *Calderon-Zygmund-inequality*, stating that the integral in (3.124) is bounded from $L^p(\mathbb{R}^n)$ to itself (provided condition (3.125) is satisfied), see e.g. Alt (2002) for a proof.

After shifting the derivatives onto the kernel, splitting the integral, exploiting the orthogonality and using the FFT, one is again left to evaluate the B_ϵ integral, containing now a (possibly) non-integrable singularity. To be able to apply the subtraction of the singularity technique, one has to shift the derivatives back onto the polynomials (gaining boundary terms). For the sake of simplicity say B_ϵ is a square and $k_1 = \partial_{\xi_1} k$, then

$$\begin{aligned}
\int_{B_\epsilon} k_1(x,\xi) R_i(\xi_1) R_k(\xi_2) d\xi &= \int_{x_2-\epsilon}^{x_2+\epsilon} R_k(\xi_2) \int_{x_1-\epsilon}^{x_1+\epsilon} \partial_{\xi_1} k \, R_i(\xi_1) d\xi_1 d\xi_2) = \\
&= \int_{x_2-\epsilon}^{x_2+\epsilon} R_k(\xi_2) \left[k(x,\xi) R_i(\xi_1) \Big|_{\xi_1=x_1-\epsilon}^{x_1+\epsilon} - \int_{x_1-\epsilon}^{x_1+\epsilon} k \partial_{\xi_1} R_i(\xi_1) d\xi_1 \right] d\xi_2 = \\
&= R_i(\xi_1) \int_{x_2-\epsilon}^{x_2+\epsilon} k(x,\xi) R_k(\xi_2) d\xi_2 \Big|_{\xi_1=x_1-\epsilon}^{x_1+\epsilon} - \int_{B_\epsilon} k \partial_{\xi_1} R_i(\xi_1) R_k(\xi_2) d\xi_1 d\xi_2,
\end{aligned}$$

such that the singularity subtraction works for the second integral, whereas the first integral, evaluated at $\xi_1 = x_1 \pm \epsilon$, can be approximately calculated using a classical Chebyshev expansion of the integrand (cf. (3.120), with Lemma 3.3 providing exact formulae for the integral), which is bounded and continuous along the line of integration. Similarly, all other boundary terms are obtained when higher derivatives in both directions appear.

The derivatives of the polynomials occurring in the boundary terms and the resulting B_ϵ integral then should be calculated in closed form (by differentiating the definition (3.1)), for programming reasons.

Remark 3.65. It is fairly straight forward to see that for symmetric functions only even polynomials appear (or need to be considered) in a Chebyshev series expansion (and analogously for odd functions). Furthermore, this means that it is sufficient for determining the coefficients to evaluate the expansion only at the negative (or positive), including zero (if so), collocation points. In (3.114) the matrix entries admit the following symmetry properties. Assume R_i is an even polynomial, i.e. $R_i(x) = R_i(-x)$ and say

$$\begin{aligned}
g(x_1) &= \int_{\mathbb{R}} \frac{w_\beta(\xi_1)}{|(x_1,x_2)-(\xi_1,\xi_2)|} R_i(\xi_1)\, d\xi_1 = \int_{\mathbb{R}} k((x_1-\xi_1)^2) R_i(\xi_1) w_\beta(\xi_1)\, d\xi_1 = \\
&= \int_{\mathbb{R}^+} k((x_1-\xi_1)^2) R_i(\xi_1) w_\beta(\xi_1)\, d\xi_1 + \int_{\mathbb{R}^-} k((x_1-\xi_1)^2) R_i(\xi_1) w_\beta(\xi_1)\, d\xi_1 \overset{R_i \text{ even}}{=} \\
&= \int_{\mathbb{R}^+} k((x_1-\xi_1)^2) R_i(\xi_1) w_\beta(\xi_1)\, d\xi_1 + \int_{\mathbb{R}^+} k((x_1+\xi_1)^2) R_i(\xi_1) w_\beta(\xi_1)\, d\xi_1,
\end{aligned}$$

whereas

$$g(-x_1) = \int_{\mathbb{R}^+} k((-x_1-\xi_1)^2) R_i(\xi_1) w_\beta(\xi_1)\, d\xi_1 + \int_{\mathbb{R}^-} k((-x_1-\xi_1)^2) R_i(\xi_1) w_\beta(\xi_1)\, d\xi_1 =$$
$$= \int_{\mathbb{R}^+} k((x_1+\xi_1)^2) R_i(\xi_1) w_\beta(\xi_1)\, d\xi_1 + \int_{\mathbb{R}^+} k(\underbrace{(-x_1+\xi_1)^2}_{=(x_1-\xi_1)^2}) R_i(\xi_1) w_\beta(\xi_1)\, d\xi_1,$$

thus, g is even if R_i is even and analogously for the odd case and the coordinate x_2.

Remark 3.66. The equations (2.32), (2.48) and (2.126) in Section 2 involve a combination of the operators $\mathcal{J}_{-\infty}^{1/2}$ and $\mathcal{R}^1$, where we have shown in the above how to gain the according matrix entries individually, cf. equations (3.106) and (3.114). The consecutive application of these operators to a function in an appropriate function space on $\mathbb{R}^2$ reads

$$(\mathcal{J}_{-\infty}^{1/2}\mathcal{R}^1) f(x_1,x_2) = \int_{-\infty}^{x_1} (x_1-s)^{-1/2} \int_{\mathbb{R}^2} ((s-\xi_1)^2 + (x_2-\xi_2)^2)^{-1/2} f(\xi) d\xi\, ds =$$
$$= \int_{-\infty}^{x_1} \int_{\mathbb{R}^2} [(x_1-s)((s-\xi_1)^2 + (x_2-\xi_2)^2)]^{-1/2} f(\xi) d\xi\, ds,$$

and with both singularities being integrable, a change of order of integration combined with a coordinate transform $v := x_1 - s$ yields a new operator description for $\mathcal{J}_{-\infty}^{1/2}\mathcal{R}^1$ given as

$$\mathcal{H} f(x_1,x_2) := \int_{\mathbb{R}^2} k_{ell}(x_1-\xi_1, x_2-\xi_2) f(\xi) d\xi, \quad k_{ell}(x,y) = \int_0^\infty [v((x-v)^2+y^2)]^{-1/2} dv,$$

where, using some algebra (or *Mathematica*), k_{ell} can be transformed into

$$k_{ell}(x,y) = \frac{2}{(x^2+y^2)^{\frac{1}{4}}} \mathbb{K}\left[\frac{1}{2}\left(1 + \frac{x}{(x^2+y^2)^{\frac{1}{2}}}\right)\right], \tag{3.126}$$

where $\mathbb{K}$ is the complete elliptic integral of the first kind. $\mathcal{H}$ also appears in the equations derived in Duck (1990) and thus we have shown their equivalence to the equations in Section 2. Several problems arise with this description. It is straight forward to show that k_{ell} becomes essentially unbounded on the *line* ($x_2 = \xi_2, \xi_1 \le x_1$), such that the advantages of the algorithm derived above cannot be utilized, i.e. using the FFT, etc. Also, if one would find some practicable cubature scheme, the argument of $\mathbb{K}$ in (3.126) lies in $[0,1]$, whereas, for example *Mathematica* finds a complex infinity for $\mathbb{K}(1)$, and hence direct numerical evaluation of the kernel k_{ell} has to be avoided in the actual scheme. Thus, in the numerical scheme used to solve the equations in Section 2 the evaluation of the operator $\mathcal{R}^1$ acting on weighted polynomials has been performed with the FFT algorithm above, where the collocation points x_{1_j} (not necessarily zeros of polynomials) actually come from a Nyström approach for $\mathcal{J}_{-\infty}^{1/2}$, see Section 3.4.

Remark 3.67. As mentioned in Remark 3.40, a Galerkin method, where the inner product integrals are approximated by quadrature techniques, is the same as a collocation approach with special collocation points. Another way to view such schemes has been analyzed in Golberg (1990) as *perturbed* projection methods. Given a sequence of linear operators $\mathcal{B}_N : X_N \to Y_N$ and $b_N \in Y_N$, then if v_N satisfies

$$\mathcal{P}_N \mathcal{K} v_N = \mathcal{P}_N g + \mathcal{B}_N v_N + b_N,$$

the sequence $\{v_N\}$ defines a perturbed projection method for solving $\mathcal{K}f = g$. Golberg (1990) then proved a theorem stating that $\mathcal{B}_N$ and b_N converging to zero (in a suitable norm) is sufficient for v_N to converge to f (also showing a convergence rate), provided the operators and projections involved satisfy the usual boundedness (or compactness) conditions.

The algorithms derived in this section obtaining the matrix entries $\mathcal{K}R_i(x_j)$ in the one and two dimensional case, cf. (3.106) and (3.114), can thus be seen as perturbed projection methods. Say $\mathcal{K}_M R_i(x_j)$ represents the matrix entry gained via the mentioned algorithms, where M shall stand for the minimum of all approximation parameters (cf. e.g. (3.116) or (3.119) or the number of quadrature points used in the QUADPACK routines referred to in Remark 3.52), then setting $\mathcal{B}_N = \mathcal{P}_N \mathcal{K}_M - \mathcal{P}_N \mathcal{K}$ yields the perturbed equation. For this Golberg (1990) deduces certain conditions for the perturbed scheme, such that $\mathcal{B}_N$ does converge to zero (e.g. $M \geq N$, the convergence of the quadrature method for continuous functions, etc.). With this we can formally claim the consistency of the presented algorithms with the exact calculation of the collocation matrix.

By replacing inner products with quadrature formulae the Galerkin method is linked to the collocation method and (as partially done in the algorithms above) by approximating all appearing integrals with a (composite) trapezoidal rule, collocation can be seen as a Nyström approach (see Hackbusch (1995)), where Section 3.4 shows how such a scheme can be set up for (weakly) singular integrals.

Remark 3.68. Sloan (1990) proved that *superconvergence* can be observed under certain conditions for the *iterated* Galerkin and collocation method, meaning that the iterated solution converges faster to the exact solution than any (standard) Galerkin or collocation solution. Therein it was concluded further, that if superconvergence occurs in the iterated collocation approach, it so does in the collocation method itself – namely at the collocation points. Similar results have also been shown in Hackbusch (1995).

3.4 Nyström Algorithms for Singular Integral Operators

Numerical quadrature or cubature can be regarded as the most direct method of approximately solving integral equations. The approach is to (locally) discretize the unknown function on some predetermined grid, where the integral operator is replaced by a weighted sum, representing an according quadrature scheme.

The analysis carried out, for example, in Hackbusch (1995) or Kress (1999) then requires this quadrature scheme to be convergent (as a necessary condition, of course). Thus in what follows, we shall present special numerical integration algorithms for the operators in equation (2.126) given in Section 2, whereas for the overall analysis, we refer to the above mentioned textbooks. Also, due to the connections of the Nyström method to the Galerkin or collocation schemes, it is possible to apply certain techniques from the projection approach (e.g. using (locally) piecewise constant basis functions).

The crucial operator to be discretized in equation (2.126) is the combination $\mathcal{J}_{-\infty}^{1/2}\mathcal{R}^1(\partial_{x_1}^3 + \partial_{x_1}\partial_{x_2}^2)$, which shall be calculated in its finite dimensional version as a matrix product of the Abel, the potential and the classical derivative operator.

Since quadrature and cubature methods work best on finite grids, we apply a mapping $\tau : [-1,1] \to \mathbb{R}$ (e.g. the tangent function or rational polynomials such as $\frac{t}{1-t^2}$), which has to be a diffeomorphism on the open intervals. To work on the compact domain $[-1,1]^2$, we generate a mesh and approximate the unknown function piecewise constant.

Hence, given the meshsizes $h_i = 2/M_i$, $i = 1, 2$, say

$$\left.\begin{array}{ll} x_1 = \tau(u), & u_i = -1 + ih_1 \\ x_2 = \tau(v), & v_j = -1 + jh_2 \end{array}\right\} \; f_{ij} = f(\tau(u_i), \tau(v_j)). \tag{3.127}$$

For practical (or programming) reasons, one derivative with respect to x_1 is shifted (via integration by parts) onto the kernel of $\mathcal{R}^1$ (as done in Remark 3.62, cf. Equation (3.122)), such that (with Δ_x denoting the Laplace operator with respect to x)

$$\mathcal{J}_{-\infty}^{1/2}\mathcal{R}^1(\partial_{x_1}^3 + \partial_{x_1}\partial_{x_2}^2)(f)(x) = \int_{-\infty}^{x_1} \frac{1}{\sqrt{x_1 - s}} \int_{\mathbb{R}^2} \frac{s - \xi_1}{((s-\xi_1)^2 + (x_2 - \xi_2)^2)^{3/2}} \Delta_\xi f(\xi) d\xi \, ds =$$

$$= \int_{-1}^{u} \frac{\tau'(\zeta)}{\sqrt{\tau(u) - \tau(\zeta)}} \int_{[-1,1]^2} \frac{(\tau(\zeta) - \tau(w))\tau'(w)\tau'(t)}{((\tau(\zeta) - \tau(w))^2 + (\tau(v) - \tau(t))^2)^{3/2}} \Delta_{\tau(w,t)} f(w,t) dw dt \, d\zeta.$$

Now we write the first integral as

$$\int_{-1}^{u} \frac{\tau'(\zeta)}{\sqrt{\tau(u) - \tau(\zeta)}} \left(\frac{u-\zeta}{u-\zeta}\right)^{\frac{1}{2}} [\cdot] d\zeta = \int_{-1}^{u} \frac{1}{\sqrt{u-\zeta}} \underbrace{\left(\frac{u-\zeta}{\tau(u) - \tau(\zeta)}\right)^{\frac{1}{2}} \tau'(\zeta)[\cdot]}_{=: f^*(u,\zeta)} d\zeta,$$

where f^* can be expected to be differentiable (or at least continuous) on $[-1, u]$, depending on the argument (indicated as $[\cdot]$). Special attention has to be paid to the behavior as $\zeta \to -1$, since the derivative of the mapping τ becomes unbounded. In original coordinates it is sufficient for the argument to decay faster than $s^{-1/2}$, which might not guarantee a decay to zero of f^* as $\zeta \to -1$. On the other hand, if the original integral exists, so does the transformed one, and hence, if f^* has a singularity at ± 1, it must be integrable. It is nevertheless possible to just set $f^*(u, -1) = 0$ when approximating f^* with piecewise

constant functions (since the unknown function contained in f^* tends to zero at $\pm\infty$) and integrating the singular kernel analytically to obtain the quadrature weights. This has been successfully applied in Scheichl et al. (2008) to the equations (2.34) and (2.49) in Section 2. A convergence analysis of such an approach (or other weakly singular integrals) can be found e.g. in Hackbusch (1995), Kress (1999) and Diogo et al. (2006).

Next, we modify the potential integral in a similar manner

$$\int\limits_{[-1,1]^2} \frac{(\tau(\zeta)-\tau(w))\tau'(w)\tau'(t)}{((\tau(\zeta)-\tau(w))^2+(\tau(v)-\tau(t))^2)^{3/2}} \frac{\zeta-w}{\zeta-w} \left(\frac{(\zeta-w)^2+(v-t)^2}{(\zeta-w)^2+(v-t)^2}\right)^{3/2} [\cdot]dwdt =$$

$$= \int\limits_{[-1,1]^2} \frac{\zeta-w}{((\zeta-w)^2+(v-t)^2)^{3/2}} f^*(\zeta,v,w,t)dwdt,$$

with f^* now given as

$$f^*(\zeta,v,w,t) = \frac{(\tau(\zeta)-\tau(w))\big[(\zeta-w)^2+(v-t)^2\big]^{3/2}}{(\zeta-w)\big[(\tau(\zeta)-\tau(w))^2+(\tau(v)-\tau(t))^2\big]^{3/2}}\tau'(w)\tau'(t)[\cdot]. \tag{3.128}$$

Using the usual polar coordinates $(w,t) \to (r,\theta)$ centered around (ζ,v), one can easily see that f^* is bounded, decays to zero as $r \to \infty$ and has a discontinuity at $r = 0$ (where we expect the unknown argument $[\cdot]$ to be continuous and bounded).

Let w,t be discretized on the same grid as u,v (and consequently $\zeta_i = -1+ih_1$), i.e. $w_k = -1+kh_1$ and $t_l = -1+lh_2$ and f^* shall be constant on $I_{kl} := [w_k-\frac{h_1}{2}, w_k+\frac{h_1}{2}]\times[t_l-\frac{h_2}{2}, t_l+\frac{h_2}{2}]$, thus $f^*_{ijkl} = f^*(\zeta_i, v_j, w_k, t_l)$, then

$$\int\limits_{[-1,1]^2} \frac{\zeta-w}{((\zeta-w)^2+(v-t)^2)^{3/2}} f^*(\zeta,v,w,t)dwdt \approx$$

$$\sum_{k=1}^{M_1-1}\sum_{l=1}^{M_2-1} f^*_{ijkl} \underbrace{\int\limits_{I_{kl}} \frac{\zeta_i-w}{((\zeta_i-w)^2+(v_j-t)^2)^{3/2}}dwdt}_{=:q_{ijkl}}, \tag{3.129}$$

such that the integral weights q can be calculated (as a Cauchy principal value integral) to be

$$q_{ijkl} = \log\left(\frac{\big(v_l^- - v_j + r^{--}\big)\big(v_l^+ - v_j + r^{++}\big)}{\big(v_l^- - v_j + r^{+-}\big)\big(v_l^+ - v_j + r^{-+}\big)}\right), \tag{3.130}$$

where the superscripts $+,-$ indicate addition or subtraction of the according $h/2$, similarly $r^{++} := \sqrt{(\zeta_i-\zeta_l^+)^2+(v_k-v_l^+)^2}$ and analogously for r^{+-}, r^{-+}, r^{--}.

Finally, for the sake of completeness, the Laplace operator in x can be transformed into an operator in u,v using

$$\partial^2_{x_1} = \left(\frac{1}{\tau'(u)}\right)^2 \partial^2_u - \frac{\tau''(u)}{\big(\tau'(u)\big)^3}\partial_u,$$

where the derivatives with respect to u, v are then approximated by centered finite differences. For the convergence of the cubature scheme (3.129) we use the following

Theorem 3.41. *Let f, defined on a bounded domain $G \subset \mathbb{R}^2$, satisfy a* Hölder *condition with exponent $\mu > 0$ and $f \equiv 0$ on ∂G. Then, for $(u_i, v_j) \in G$, the following estimate holds*

$$\left| \int\limits_G \frac{u_i - w}{((u_i - w)^2 + (v_j - t)^2)^{3/2}} f(w,t) dw dt - \sum_{k=1}^{M_1 - 1} \sum_{l=1}^{M_2 - 1} f(u_l, v_k) q_{ijkl} \right| \leq O(h^\mu |\log(h)|),$$

where $h = \max(h_1, h_2)$, q given in (3.130) *and the grid is as defined above.*

Proof. see Akimenko (1997) □

Obviously, the integrand f^* in (3.128) is Hölder continuous with exponent $\mu > 0$, except at $r = 0$ (there, $\mu = 0$, i.e. bounded). The argumentation in the proof given in Akimenko (1997) is first done for the domain $G \backslash B_r(u_i, v_j)$, such that one can (continuously) replace f^* on $[0, r]$ by a function, which has a Hölder exponent $\mu > 0$, without altering the value of the discretized integral. Hence the estimate in Theorem 3.41 holds for this modification. Then, in the limit $r \to 0$, the original f^* is recovered, with the replacement function reducing to a point of measure zero, the discontinuity of f^*.

Remark 3.69. Golberg (1979) showed for integral equations of the second kind that (under certain conditions on the inverse of discretized operator) the uniform approximation error between the Nyström solution and the unknown function is dominated by the quadrature error (in the infinity norm), which depends on the smoothness of the unknown function and the kernel, cf. Theorem 3.41. In other words, the order of the quadrature scheme is "a little less" than μ.

APPENDICES

A Higher Order Marginal Separation Expansions

As mentioned in Section 2.3.2 parts of the solution formulae for the main deck vertical velocities contain certain (unbounded) integrals, which need to be examined further in order for a matching rule to work. In doing so we first repeat their description from (2.117), i.e.

$$\left.\begin{aligned} c_{21} &= \int_0^{y_2} \frac{U_0''(s) - p_{00}}{U_0^2(s)} ds, \qquad & c_{25} &= \int_0^{y_2} \frac{\partial_s^2 u_{21} - v_{21}\partial_s u_{21} - u_{21}\partial_x u_{21}}{U_0^2(s)} ds \\ c_{26} &= \int_0^{y_2} \frac{U_0'(s)}{U_0^2(s)} ds, & c_{27} &= \int_0^{y_2} \frac{1}{U_0^2(s)} ds. \end{aligned}\right\} \text{ as } y_2 \to \infty.$$

Obviously, an important function here is the separation profile $U_0 = U_0(y_2)$, where we recall its asymptotic behavior (Equation (2.11), Section 2.1) to be

$$\begin{aligned} U_0(y_2) &\sim \frac{p_{00}}{2} y_2^2 \quad \text{as } y_2 \to 0 \\ U_0(y_2) &\to U_{00} \qquad \text{as } y_2 \to \infty. \end{aligned}$$

In general we will view the above integrals as improper integrals, i.e. in the limit $\eta \to 0$, with η being the lower bound. Let us start with c_{21}. Here, with $U_0'' = p_{00}$ as $y_2 \to 0$, the integrand vanishes identically at the lower bound, whereas at the upper bound it tends to the value $-p_{00}/U_{00}^2$. Hence in the limit $y_2 \to \infty$ we write

$$c_{21} = \int_0^{\infty} \left(\frac{U_0''(s) - p_{00}}{U_0^2(s)} + \frac{p_{00}}{U_{00}^2} \right) ds - \frac{p_{00}}{U_{00}^2} y_2,$$

where the absolute value of the integral is bounded. In the same manner one can analyze c_{25}. The substitution of (2.98) into c_{25} yields a closed description for the integrand, such that one can evaluate its asymptotic behavior at the upper and lower bound. As this is quite involved but essentially a technical task, we refrain from displaying all the calculations here and just state the result as

$$\frac{\partial_{y_2}^2 u_{21} - v_{21}\partial_{y_2} u_{21} - u_{21}\partial_x u_{21}}{U_0^2} \sim \begin{cases} -p_{00}^2 x / U_{00}^4 & y_2 \to \infty \\ -\dfrac{2A\partial_x A}{p_{00}^2 y_2^2} & y_2 \to 0. \end{cases}$$

Thus we shall write, by extracting the growth at the bounds from the integral

$$c_{25} = \lim_{y_2 \to \infty} [c_{25}^* - \frac{p_{00}^2 x}{U_{00}^4} y_2 + g(y_2)] = \lim_{y_2 \to \infty} [c_{25}^* - \frac{p_{00}^2 x}{U_{00}^4} y_2],$$

where c_{25}^* is bounded and $g(y_2) \sim 1/y_2$ as $y_2 \to 0$, but $g(y_2) \to 0$ for large y_2.

Next we consider the integrand in c_{26}. In the limit $y_2 \to \infty$ it tends (algebraically) to zero. As $y_2 \to 0$ we have

$$\frac{U_0'(y_2)}{U_0^2(y_2)} \sim \frac{p_{00}y_2}{p_{00}^2 y_2^4/4} = \frac{4}{p_{00}y_2^3}.$$

Hence we shall write in general

$$c_{26} = \lim_{y_2\to\infty} \left[\int_0^{y_2} \left(\frac{U_0'(s)}{U_0^2(s)} - \frac{f(s)}{p_{00}s^3} \right) ds + g(y_2) \right] =: -c_{26}^*,$$

where $f = f(y_2)$ is bounded and positive everywhere and tends (rapidly) to 4 as $y_2 \to 0$, whereas $g(y_2) \sim y_2^{-2}$ in this limit, but $g \to 0$ as $y_2 \to \infty$. Therefore, by choosing f accordingly, we obtain $c_{26}^* > 0$. Note that $f \equiv 4$ might very well be sufficient.

Analogously we approach c_{27}, where

$$\frac{1}{U_0^2(y_2)} \sim \begin{cases} U_{00}^{-2} & y_2 \to \infty \\ 4/(p_{00}^2 y_2^4) & y_2 \to 0, \end{cases}$$

such that

$$c_{27} = \lim_{y_2\to\infty} \left[\int_0^{y_2} \left(\frac{1}{U_0^2(s)} - \frac{1}{U_{00}^2} - \frac{f(s)}{s^4} \right) ds + \frac{y_2}{U_{00}^2} + g(y_2) \right] =: c_{27}^* + \frac{y_2}{U_{00}^2} \quad \text{as} \quad y_2 \to \infty,$$

with f, g being not the same as in c_{26}, but satisfy similar requirements, such that $c_{27}^* > 0$.

Remark A.1. For all the constants c_i above we used the method of subtracting the singular behavior of the integrand and hence gained bounded integrals plus terms reflecting the unbounded growth of the original integral. By as general as necessary we symbolized the singularity subtraction by the functions f and g. One shall keep in mind that the applied matching procedure will impose some additional constraints on these functions. For our purpose it is not necessary to go into these details. Nevertheless for the matching rules it might be advantageous to write the singularity subtraction in terms of the already existing functions, e.g. in c_{27} the term $f(s)/s^4$ might be written similar to $1/[(U_0')^2 s^2]$.

B Multiple Sturm-Liouville Operators

Induction Argument for A^m

Stated in (3.27), the successive application of the Sturm-Liouville operator can be written as

$$A_x^m f(x) = \sum_{k=1}^{2m} (1+x^2)^{m+\frac{k}{2}} p_k(x) \partial_x^k f(x), \tag{B.1}$$

where we will show the p_k to be uniformly bounded (as mentioned in Wang & Guo (2002)).

Given a function f the Sturm-Liouville operator A_x applied to f reads

$$A_x f(x) = -(x^2+1)\partial_x\big((x^2+1)\partial_x f(x)\big) = -(x^2+1)(2x\partial_x f(x) + (x^2+1)\partial_x^2 f(x)). \quad \text{(B.2)}$$

By saying $m = 1$ in (B.1) one obtains

$$A_x f(x) = (x^2+1)^{3/2} p_1(x)\partial_x f(x) + (x^2+1)^2 p_2(x)\partial_x^2 f(x),$$

where comparison with the right-hand side in (B.2) yields

$$p_1(x) = -\frac{2x}{\sqrt{1+x^2}} \quad \text{and} \quad p_2(x) = -1,$$

which are uniformly bounded on $\mathbb{R}$. Assuming the same has been shown for $A_x^m f$ the next step $A_x^{m+1} f = A_x(A_x^m f)$ is given by

$$A_x^{m+1} f(x) = -\underbrace{(x^2+1)2x\partial_x A_x^m f(x)}_{(*)} - \underbrace{(x^2+1)^2\partial_x^2 A_x^m f(x)}_{(**)}.$$

Calculating the first and second derivatives of $A_x^m f$ the terms on the right-hand side above are given as

$$(*) = \sum_{k=1}^{2m}\big[4\big(\tfrac{k}{2}+m\big)(x^2+1)^{\frac{k}{2}+m}x^2 p_k(x)\partial_x^k f(x) + 2(x^2+1)^{\frac{k}{2}+m+1}x p_k(x)\partial_x^{k+1} f(x) + \\ + 2(x^2+1)^{\frac{k}{2}+m+1}x p_k'(x)\partial_x^k f(x)\big]$$

which, by rearranging terms and shifting indices, can be further modified to give

$$(*) = \sum_{k=1}^{2m}(x^2+1)^{\frac{k}{2}+m+1}\underbrace{\Big[4\alpha_{k,m}\frac{x^2 p_k(x)}{1+x^2} + 2x p_k'(x)\Big]}_{=:a_k(x)}\partial_x^k f(x) + \\ + \sum_{k=2}^{2m+1}(x^2+1)^{\frac{k}{2}+m+1}\underbrace{\frac{2x p_{k-1}(x)}{\sqrt{x^2+1}}}_{=:b_k(x)}\partial_x^k f(x) \quad \text{(B.3)}$$

and analogously

$$(**) = \sum_{k=3}^{2m+2}(x^2+1)^{\frac{k}{2}+m+1}p_{k-2}(x)\partial_x^k f(x) + \\ + 2\sum_{k=2}^{2m+1}(x^2+1)^{\frac{k}{2}+m+1}\underbrace{\Big[2\alpha_{k-1,m}\frac{x p_{k-1}(x)}{\sqrt{x^2+1}} + \sqrt{x^2+1}\,p_{k-1}'(x)\Big]}_{=:c_k(x)}\partial_x^k f(x) + \quad \text{(B.4)}$$

$$+\sum_{k=1}^{2m}(x^2+1)^{\frac{k}{2}+m+1}\underbrace{\left[2\alpha_{k,m}\left(\tfrac{2\alpha_{k-1,m}x^2}{x^2+1}+1\right)p_k(x)+4\alpha_{k,m}xp_k'(x)+(x^2+1)p_k''(x)\right]}_{=:d_k(x)}\partial_x^k f(x)$$

with $\alpha_{k,m}=\frac{k}{2}+m$.
Finally, noting that every sum in (B.3) and (B.4) contains the terms $(x^2+1)^{\frac{k}{2}+m+1}$ and $\partial_x^k f(x)$, one can collect these sums to obtain

$$A_x^{m+1}f(x)=\sum_{k=1}^{2(m+1)}(x^2+1)^{\frac{k}{2}+m+1}p_k^*(x)\partial_x^k f(x),$$

with

$$p_k^*=\begin{cases}a_k+d_k & k=1\\ a_k+d_k+b_k+2c_k & k=2\\ a_k+d_k+b_k+2c_k+p_{k-2} & 3\le k\le 2m\\ b_k+2c_k+p_{k-2} & k=2m+1\\ p_{k-2} & k=2m+2\end{cases},$$

where it is straight forward to see that the p_k^* are uniformly bounded (since the p_k are bounded by assumption). □

With the details presented in the above it is worthwhile mentioning that the p_k depend additionally on m, meaning that, e.g. $p_1|_{m=1}\neq p_1|_{m=2}$ and so forth. In virtue of (3.29) the multi-dimensional case is given by multiplication of the individual p_{k_i}, which are equal to the above derived p_k for all $k_i=k$ and hence are again uniformly bounded on $\mathbb{R}^n$.

Partial Derivative of A^m

As needed in (3.37), where $r=(2m+1)n$, we shall prove

$$\|\partial_x A_x^m f\|_{u^{-1}}^2\le c\|f\|_A^2.$$

It suffices to show $\|\partial_x A_x^m f\|_{u^{-1}}^2=\|B_x^m f\|_u^2$, where the operator B stems from replacing every p_{k_i} in (3.29) by another *bounded* function q_{k_i}, such that the arguments in the proof of Lemma 3.10 still hold.

Let us start with the one-dimensional partial derivative of A^m, which can be written as

$$\begin{aligned}\partial_x A_x^m f&=\partial_x\sum_{k=1}^{2m}(1+x^2)^{m+\frac{k}{2}}p_k(x)\partial_x^k f(x)=\\ &=\sum_{k=1}^{2m}\left[(m+\tfrac{k}{2})(1+x^2)^{m-1+k/2}2xp_k+(1+x^2)^{m+k/2}p_k'\right]\partial_x^k f+\\ &+\sum_{k=1}^{2m}(1+x^2)^{m+k/2}p_k\partial_x^{k+1}v\quad(k\to k-1,\text{second sum})=\end{aligned}$$

$$= \sum_{k=1}^{2m+1} (1+x^2)^{m+k/2-1/2} q_k \partial_x^k f$$

with

$$q_k(x) = (m+\tfrac{k}{2})\frac{2x}{\sqrt{1+x^2}} p_k(x) + \sqrt{1+x^2} p_k'(x) + p_{k-1}(x), \quad p_0 := 0, \quad q_{2m+1} := p_{2m+1}.$$

As shown above the p_k are rational bounded functions, evaluating to a constant at $\pm\infty$. It is easy to see, that derivatives of such fractions have to decay to zero at infinity faster then or equal to x^{-1}. Consequently, the q_k are of the same type as the p_k, especially bounded.

By definition

$$\begin{aligned}\|\partial_x A_x^m f\|_{u^{-1}}^2 &= \int_{\mathbb{R}^n} \Big| \prod_{i=1}^n \partial_{x_i} A_{x_i}^m f \Big|^2 \prod_{i=1}^n (1+x_i^2) dx_1 \dots dx_n = \\ &= \int_{\mathbb{R}^n} \Big| \prod_{i=1}^n \underbrace{(1+x_i^2)\partial_{x_i} A_{x_i}^m f}_{=: B_{x_i}^m f} \Big|^2 u(x) dx_1 \dots dx_n,\end{aligned}$$

where

$$B_{x_i}^m f = \sum_{k_i=1}^{2m+1} (1+x_i^2)^{m+k_i/2+1/2} q_k(x_i) \partial_{x_i}^k f(x)$$

thus yielding

$$\begin{aligned}\|\partial_x A_x^m f\|_{u^{-1}}^2 &= \int_{\mathbb{R}^n} \Big| \prod_{i=1}^n B_{x_i}^m f \Big|^2 u(x) dx = \\ &= \int_{\mathbb{R}^n} \Big| \sum_{k_n=1}^{2m+1} \cdots \sum_{k_1=1}^{2m+1} \prod_{i=1}^n (1+x_i^2)^{\frac{2m+k_i+1}{2}} q_k(x_i) \partial_{x_i}^k f(x) \Big|^2 u(x) dx = \|B_x^m f\|_u^2\end{aligned}$$

with finally replacing $r = (2m+1)n$. □

C Derivatives of $f \circ \phi^{-1}$ and their Far Field Behavior

In the proof of Theorem 3.15 a description of the derivatives (with respect to y) of $(f \circ \phi^{-1})(y)$ is needed in a way to establish the far field behavior in terms of $y = \phi(x)$, with $\phi(x) = \arctan(x) - \pi/2$ being the mapping defined in (3.2).

For the sake of clarity and readability we shall first prove the one-dimensional result, i.e. given a function $f \in C^n(\mathbb{R})$ and ϕ as above, then for

$$x^s \partial_x f(x) \to 0 \ \text{ as } \ x \to \pm\infty \quad \Rightarrow \quad \partial_y^{s-1}(f \circ \phi^{-1})(y) \to 0 \quad \text{as} \ \ y \to \{-\pi, 0\}. \tag{C.1}$$

It shall be noted that $\phi^{-1}(y) = \tan(y + \pi/2)$ and $(\phi^{-1})'(y) = \dfrac{1}{\sin^2(y)} = 1 + x^2$, when substituting $y = \phi(x)$. Consequently, the asymptotic behavior can the be written as

$$\left|\frac{d}{dy^n}\phi^{-1}(y)\right| \sim \frac{1}{y^{n+1}} \quad \text{as} \quad y \to 0 \quad \text{and} \quad \left|\frac{d^n}{d\phi^n}\phi^{-1}(\phi(x))\right| \sim x^{n+1} \quad \text{as} \quad x \to \pm\infty \tag{C.2}$$

and analogously if $y \to -\pi$. The formulae above can be (easily) seen by a Taylor series expansion. Note that the sign is not important here, for it does not matter whether the assumption in (C.1) is satisfied when approaching $0^{\pm}$.

Starting with $s = 1$ one essentially has to show, if $x\partial_x f(x) \to 0 \Rightarrow f(x) \to 0$ as $x \to \pm\infty$. Assuming a non-oscillatory asymptotic behavior of algebraic type (since for exponentially decaying functions the assertion is trivially satisfied and if f evaluates to a constant or tends to infinity in the far field it cannot be satisfied at all) gives $\partial_x f(x) \sim x^a$, such that $x\partial_x f(x) \sim x^{a+1} \to 0$ and from formally integrating the derivative it follows that $f(x) \sim x^{a+1}$ which tends to zero by assumption as $x \to \pm\infty$.

Let us perform the same argumentation for $s = 2$, which means $\partial_y(f \circ \phi^{-1})(y) \to 0$ as (without loss of generality) $y \to 0$ is needed. The chain rule yields

$$\partial_y(f \circ \phi^{-1})(y) = \partial_{\phi^{-1}} f(\phi^{-1}(y))\,(\phi^{-1})'(y) = \partial_x f(x)\,(\phi^{-1})'(\phi(x)) \sim x^2 \partial_x f(x), \tag{C.3}$$

as $x \to \pm\infty$, using (C.2) for the far field. By assuming the right-hand side above (i.e. the asymptotic behavior) to tend to zero as $x \to \infty$ it follows that the very left-hand side, $\partial_y(f \circ \phi^{-1})(y)$, tends to zero as y tends to zero.

To find similar equalities as in (C.3) for arbitrary s a general description of the chain rule for higher derivatives has to be applied. Such a formula was found by *Faa di Bruno*, cf. e.g. Mishkov (2000).

Given $f, g \in C^n(\mathbb{R})$ then

$$\partial_y^n(f \circ g)(y) = \sum c_{k_i}\, \partial_g^k f(g(y)) \prod_{i=1}^{n} (\partial_y^i g(y))^{k_i}, \tag{C.4}$$

where the sum is taken over all non-negative integer solutions of $k_1 + 2k_2 + \cdots + nk_n = n$ with $k = k_1 + \cdots + k_n$. The constants c_{k_i} (also termed structural coefficients) are given via

$$c_{k_i} = \frac{n!}{k_1!k_2!\ldots k_n!(1!)^{k_1}(2!)^{k_2}\ldots(n!)^{k_n}}.$$

A further investigation of the Diophantine equation for the k_i shows that one can rewrite the formula, such that

$$\partial_y^n(f \circ g)(y) = \sum_{k=1}^{n} \partial_g^k f(g(y)) \mathcal{B}_k(g), \tag{C.5}$$

where the terms $\mathcal{B}_k$ can be found in, e.g. Leipnik & Pearce (2007). With this version the appearance of all orders of the derivatives of f becomes more obvious.

Furthermore, the solution $k_1 = k_2 = \cdots = k_{n-1} = 0$, $k_n = 1$ is the only possibility to obtain $k = 1$ (with the coefficient $c = 1$). Hence, for all n, one term in the sum in (C.4) will always read

$$\partial_g f(g(y))\partial_y^n g(y) \quad \rightsquigarrow \quad \partial_x f(x)\partial_y^n \phi^{-1}(y) \sim x^{n+1}\partial_x f(x) \quad \text{as} \quad x \to \infty, \tag{C.6}$$

which is exactly the term needed to prove the assertion in (C.1). Thus, it is left to show that every other term in the sum in (C.4) has the exact same asymptotic behavior as the term with $k_n = 1$, i.e. if $\partial_g f(g(y))\partial_y^n g(y) \to 0$ it follows that the whole sum tends to zero as $y \to 0$.

As an example we will demonstrate this for $n = 3$:

$$\partial_y^3 f(g(y)) = \partial_g^3 f\,(\partial_y g)^3 + 3\,\partial_g^2 f\,\partial_y g\,\partial_y^2 g + \partial_g f\,\partial_y^3 g,$$

where by the same substitution as in (C.6) the third term yields $x^4\partial_x f$, the second $x^2 x^3 \partial_x^2 f$ and the first $(x^2)^3\partial_x^3 f$. With the assumption on the third term to tend to zero, it is obvious that (assuming algebraic far field behavior) both the first and the second term will also vanish. Taking this as the induction basis, the step from $n \to n+1$ is then given as

$$\begin{aligned}
\partial_y(\partial_y^n(f \circ g)) &= \partial_y \sum c_{k_i}\,\partial_g^k f \prod_{i=1}^{n}(\partial_y^i g(y))^{k_i} = \\
&= \sum c_{k_i}\,\partial_g^k f\, k_1(\partial_y g)^{k_1-1}\partial_y^2 g \prod_{i\neq 1}^{n}(\partial_y^i g(y))^{k_i} + \\
&+ \sum c_{k_i}\,\partial_g^k f\, k_2(\partial_y^2 g)^{k_2-1}\partial_y^3 g \prod_{i\neq 2}^{n}(\partial_y^i g(y))^{k_i} + \\
&\vdots \\
&+ \sum c_{k_i}\,\partial_g^k f\, k_n(\partial_y^n g)^{k_n-1}\partial_y^{n+1} g \prod_{i=1}^{n-1}(\partial_y^i g(y))^{k_i} + \\
&+ \sum c_{k_i}\,\partial_g^{k+1} f\,\partial_y g \prod_{i=1}^{n}(\partial_y^i g(y))^{k_i}
\end{aligned}$$

and all the sums are still taken over all solutions $k_1, \ldots, k_n$ of the Diophantine equation.

Evaluating the sum at $y = 0$ (or $x \to \infty$) we can (formally) rewrite the equation above

$$\begin{aligned}
\partial_y(\partial_y^n(f \circ g)) = \frac{\partial_y^2 g}{\partial_y g} \sum c_{k_i} k_1 \, \partial_g^k f \prod_{i=1}^{n} (\partial_y^i g(y))^{k_i} + \\
+ \frac{\partial_y^3 g}{\partial_y^2 g} \sum c_{k_i} k_2 \, \partial_g^k f \prod_{i=1}^{n} (\partial_y^i g(y))^{k_i} + \\
\vdots \\
+ \frac{\partial_y^{n+1} g}{\partial_y^n g} \sum c_{k_i} k_n \, \partial_g^k f \prod_{i=1}^{n} (\partial_y^i g(y))^{k_i} + \\
+ \sum c_{k_i} \, \partial_g^{k+1} f \, \partial_y g \prod_{i=1}^{n} (\partial_y^i g(y))^{k_i},
\end{aligned} \tag{C.7}$$

such that the induction hypothesis can be applied to the first n sums, since in the far field the fraction $\partial_y^{m+1} g / \partial_y^m g \sim x \ \forall m$. In the last line above an additional derivative appears for f (together with the term $\partial_y g$), whereas the combinations of $k_1, \ldots, k_n$ (from the induction hypothesis) still yield the same terms for the product of the derivatives of g. These terms, derived using the solutions $k_1, \ldots, k_n$, only evaluate to the same asymptotic behavior if multiplied with the far field of $\partial_g^k f$. But since $\partial_x f \sim x^{-n-\epsilon}$ (by assumption) $\Rightarrow$ $\partial_g^{k+1} f \sim x^{-n-(k+1)-\epsilon}$ times $\partial_y g \sim x^2$ yields $\partial_g^{k+1} f \, \partial_y g \sim x \, x^{-n-k-\epsilon}$, such that x can be taken out of the summation and the remaining asymptotic behavior equals that of $\partial_g^k f$, which finishes the proof. □

In Mishkov (2000) a version of Faa di Bruno's formula for vector arguments is presented, again using combinatorial aspects to treat summations and in Leipnik & Pearce (2007) the most general case of multivariate higher derivatives of composite functions with multiple arguments can be found. The case considered here, where a multivariate function is combined with a univariate mapping can be viewed as a special case, where no additional formulae and Diophantine equations are needed.

For the sake of readability in the following we will treat the two-dimensional case, where the generalization to the multi-dimensional case then becomes obvious. Considering the first derivative, meaning that the first derivative is taken in every component, i.e.

$$\begin{aligned}
\partial_{y_1 y_2}^2 f(g(y_1), g(y_2)) &= \partial_{y_1} \big(\partial_{y_2} f(g(y_1), g(y_2)) \big) = \\
&= \partial_{y_1} \big(\partial_{g(y_2)} f(\cdot, \cdot) \, g'(y_2) \big) = \partial_{g(y_1) g(y_2)}^2 f(\cdot, \cdot) \, g'(y_1) g'(y_2),
\end{aligned}$$

where the prime indicates the derivative with respect to the according variable and henceforth derivatives with respect to $g(y_i)$ shall be written as ∂_{x_i} (for $x_i = g(y_i)$ when substituting g with ϕ^{-1}). Consequently the second derivatives (in every component) read

$$\begin{aligned}
\partial_{y_1 y_2}^4 f(x_1, x_2) = \partial_{x_1}^2 \partial_{x_2}^2 f(\cdot, \cdot) \, (g'(y_1))^2 (g'(y_2))^2 + \partial_{x_1} \partial_{x_2}^2 f(\cdot, \cdot) \, (g'(y_2))^2 g''(y_1) + \\
+ \partial_{x_1}^2 \partial_{x_2} f(\cdot, \cdot) \, (g'(y_1))^2 g''(y_2) + \partial_{x_1} \partial_{x_2} f(\cdot, \cdot) \, g''(y_1) g''(y_2),
\end{aligned} \tag{C.8}$$

where it is straight forward to see that this can also be written in the form

$$\partial^4_{y_1y_2} f(x_1,x_2) = \big[(g'(y_1))^2\partial^2_{x_1} + g''(y_1)\partial_{x_1}\big]\big[(g'(y_2))^2\partial^2_{x_2} + g''(y_2)\partial_{x_2}\big] f(x_1,x_2),$$

such that the expression in the square brackets is the formula for the second composite derivative in one dimension and thus one can define

$$\mathcal{D}_i f(x_1,\dots,x_n) := \partial_{y_i}(f\circ g)(y_1,\dots,y_n) = \partial_{y_i} f(g(y_1),\dots,g(y_n)), \tag{C.9}$$

such that, e.g. $\partial^4_{y_1y_2} f(x_1,x_2) = [\mathcal{D}_1^2\mathcal{D}_2^2] f(x_1,x_2)$.

As mentioned in the proof of Theorem 3.15 the conditions dealt with in this appendix stem from successive applications of integration by parts to the definition of expansion coefficients in terms of inner products. In two dimensions such integrals read

$$\int_{I^2} f(y_1,y_2)\cos(ky_1)\cos(ly_2)\,dy_1dy_2, \quad k,l\in\mathbb{N},$$

and integration by parts then yields

$$\begin{aligned}
&\int_{I^2} f(y_1,y_2)\cos(ky_1)\cos(ly_2)\,dy_1dy_2 = \\
&= \int \cos(ly_2)\Big[\underbrace{\frac{f(y_1,y_2)\sin(ky_1)}{k}\Big|_{y_1=\partial I}}_{\stackrel{!}{=}0} - \tfrac{1}{k}\int \partial_{y_1} f\,\sin(ky_1)\,dy_1\Big]dy_2 = \\
&= -\tfrac{1}{k}\int \sin(ky_1)\Big[\underbrace{\frac{\partial_{y_1} f(y_1,y_2)\sin(ly_2)}{l}\Big|_{y_2=\partial I}}_{\stackrel{!}{=}0} - \tfrac{1}{l}\int \partial^2_{y_1y_2} f\,\sin(ly_2)\,dy_2\Big]dy_1 = \\
&= \frac{1}{kl}\int_{I^2} \partial^2_{y_1y_2} f(y_1,y_2)\sin(ky_1)\sin(ly_2)\,dy_1dy_2
\end{aligned} \tag{C.10}$$

where the usual integrability condition $\partial^2_{y_1y_2} f \in L^2_w(I^2)$ justifies the change of integration order and the existence of the appearing integrals. The necessary condition for the above equalities to hold are (ignoring the behavior of the sine function at ∂I, see the arguments in (3.47) through (3.48))

$$\begin{aligned}
&(i)\quad f(y_1,y_2)\to 0, \quad \text{as}\ \ y_1\to\partial I\ \ \text{and}\ \ y_2\to\partial I\ \ \text{(separately)}\\
&(ii)\quad \partial_{y_1} f(y_1,y_2)\to 0, \quad \text{as}\ \ y_2\to\partial I\\
&(iii)\quad \partial_{y_2} f(y_1,y_2)\to 0, \quad \text{as}\ \ y_1\to\partial I,
\end{aligned}$$

taking into account that integration by parts has to arrive at the same result when interchanging y_1 and y_2. To show that condition (i) implies (ii) and (iii) (as required) one can

write (ii) as

$$\partial_{y_1} f(y_1, y_2) = \lim_{h \to 0} \frac{1}{h} \underbrace{[f(y_1 + h, y_2) - f(y_1, y_2)]}_{\to 0, \text{ by } (i)} \quad \text{and vice versa for } (iii).$$

In virtue of (C.10) it is obvious how this procedure and the according conditions read for multivariate functions f.

Finally, we will prove the two-dimensional equivalent of (C.1), i.e. assuming $f \in C^n(\mathbb{R}^2)$ and ϕ as usual, then for

$$x_1^s x_2^s \partial^2_{x_1 x_2} \to 0 \quad \text{as} \quad x_i \to \pm\infty \quad \Rightarrow \quad \partial^{(s-1)2}_{y_1 y_2} (f \circ \phi^{-1})(y_1, y_2) \to 0 \quad \text{as} \quad y_i \to \{-\pi, 0\}. \tag{C.11}$$

Using the definition of the composite derivative operator (C.9) the conclusion also reads

$$\partial^{(s-1)2}_{y_1 y_2} (f \circ \phi^{-1})(y_1, y_2) = [\mathcal{D}_1^{s-1} \mathcal{D}_2^{s-1}] f(x_1, x_2) \to 0 \quad \text{as} \quad x_i \to \pm\infty.$$

And from the fact that

$$[\mathcal{D}_1^{s-1} \mathcal{D}_2^{s-1}] f(x_1, x_2) = \mathcal{D}_1^{s-1} [\mathcal{D}_2^{s-1} f(x_1, x_2)] = \mathcal{D}_2^{s-1} [\mathcal{D}_1^{s-1} f(x_1, x_2)]$$

one can readily deduce that for the individual asymptotic behavior, meaning x_1 and x_2 tend to infinity separately, the one-dimensional argumentation, cf. (C.6) - (C.7), applies here. In fact, this is equivalent to the assumption in (C.11) and consequently to condition (ii) given in Theorem 3.15.

Using the standard transformation to polar coordinates, (C.11) can be shown to also hold in the case of $r := |(x_1, x_2)| \to \infty$, which we shall demonstrate in the following for the case of $s = 2$, cf. (C.8).

Say $r(x_1, x_2) = \sqrt{x_1^2 + x_2^2}$ and $\alpha = \arctan(x_2/x_1)$, then the far field of the fourth term in (C.8) can be transformed to give, as $(x_1, x_2) \to \infty$, i.e. $r \to \infty$

$$\begin{aligned}
\partial_{x_1} \partial_{x_2} f(x_1, x_2)\, g''(y_1) g''(y_2) \sim\ & x_1^3 x_2^3 \partial_{x_1} \partial_{x_2} f(x_1, x_2) = \\
& = r^6 \cos^3(\alpha) \sin^3(\alpha) \left[-\frac{\sin(2\alpha)}{2r} \partial_r f(r, \alpha) - \frac{\cos(2\alpha)}{r^2} \partial_\alpha f(r, \alpha) + \right. \\
& \left. + \frac{\cos(2\alpha)}{r} \partial^2_{r\alpha} f(r, \alpha) + \frac{\sin(2\alpha)}{2} \partial_r^2 f(r, \alpha) - \frac{\sin(2\alpha)}{2r^2} \partial_\alpha^2 f(r, \alpha) \right] \sim \\
& \sim r^5 \partial_r f(r, \alpha) + r^4 \partial_\alpha f(r, \alpha) + r^5 \partial^2_{r\alpha} f(r, \alpha) + r^6 \partial_r^2 f(r, \alpha) + r^4 \partial_\alpha^2 f(r, \alpha)
\end{aligned} \tag{C.12}$$

requiring, as done above, the last line to tend to zero (taking the plus sign for all terms to indicate that they all have to tend to zero individually), yielding the asymptotic behavior for all the appearing derivatives of f with respect to r and α.

In the same way one can then calculate the asymptotic behavior of the other terms in (C.8), e.g.

$$\partial_{x_1}^2 \partial_{x_2} f(x_1, x_2) \, (g'(y_1))^2 g''(y_2) \sim \tag{C.13}$$
$$\sim r^5 \partial_r f + r^4 \partial_\alpha f + r^5 \partial_{r\alpha}^2 f + r^6 \partial_r^2 f + r^4 \partial_\alpha^2 f + r^6 \partial_r^2 \partial_\alpha f + r^5 \partial_r \partial_\alpha^2 f + r^7 \partial_r^3 f + r^4 \partial_\alpha^3 f.$$

By now (formal-asymptotically) assuming if $\partial_r^n f \sim r^a \Rightarrow \partial_r^{n+1} f \sim r^{a-1}$ and if $\partial_\alpha^n f \sim r^b \Rightarrow \partial_\alpha^{n+1} f \sim r^b$ one can immediately see, that if the far field in (C.12) tends to zero, so does the far field in (C.13) (if the function f is (strongly) non-separable the derivative with respect to α might also lead to a different asymptotic behavior, i.e. $\partial_\alpha^{n+1} f \sim r^c$, with $c < b$). Performing this analogously for all other terms, with a similar induction argument for the higher derivatives as in (C.7) proves the assertion in two-dimensions.

Due to the product character of the composite derivative operator $\mathcal{D}_i$ the final generalization to multi-dimensions is straight forward (as is the application of multivariate polar coordinates). The integrability of $\partial_x^m f$ in $\mathbb{R}^n$ as a consequence of the integrability of $\partial_y^m (f \circ \phi^{-1})$ (provided the necessary decay derived above) follows from (C.5) and (C.8). □

D The Sine "Polynomial" Version on $\mathbb{R}$

Given a square integrable function $f : [-\pi, 0] \to \mathbb{R}$, extended to be an odd function on $[-\pi, \pi]$, then (under the usual assumptions) the Fourier sine series

$$p_N(y) = \sum_{k=0}^{N} a_k \sin(ky) \tag{D.1}$$

converges to f in $L^2(-\pi, 0)$, i.e.

$$\int_{-\pi}^{0} |f(y) - \sum_{k=0}^{N} a_k \sin(ky)|^2 dy \to 0, \quad \text{as} \quad N \to 0.$$

Since the since functions are orthogonal under the inner product

$$\langle \sin(ky), \sin(ly) \rangle = \int_{-\pi}^{0} \sin(ky) \sin(ly) dy = \frac{\pi}{2} \delta_{kl},$$

the coefficients a_k can be written as (with $a_0 = 0$)

$$a_k = \frac{2}{\pi} \langle f, \sin(ky) \rangle, \quad \forall k > 0.$$

By applying the coordinate transform ϕ from (3.2), which maps $\mathbb{R} \to [-\pi, 0]$, saying $y = \phi(x)$, such that $dy = \frac{1}{1+x^2}dx$, (D.1) now reads

$$\int_{\mathbb{R}} \left| f(\phi(x)) - \sum_{k=0}^{N} a_k \sin(k\phi(x)) \right|^2 \frac{1}{1+x^2} dx \to 0, \quad \text{as} \quad N \to 0.$$

Finally, defining $S_k(x) := \sin(k\phi(x))$ for $x \in \mathbb{R}$, one can assert the set $\{S_k\}$ to be complete and orthogonal in $L^2_w(\mathbb{R})$ with the weight function $w(x) = \frac{1}{1+x^2}$ and equipped with the inner product

$$\langle S_k, S_l \rangle_w := \int_{\mathbb{R}} S_k(x) S_l(x) w(x) dx = \frac{\pi}{2} \delta_{kl},$$

such that the projection onto an N-dimensional subspace can be obtained to be

$$\mathcal{P}_N f = \sum_{k=1}^{N} a_k S_k, \quad \text{where} \quad a_k = \frac{2}{\pi} \langle f, S_k \rangle_w$$

with $S_0 \equiv 0$, which converges in the L^2_w norm to $f \in L^2_w(\mathbb{R})$. □

Obviously, since the weight function is the same as in the rational Chebyshev case, all results hold in the same way for the mapped sine functions (as long as no use was made of specific cosine properties).

Remark D.1. Deriving the first such sine function $S_1(x) = -\frac{1}{\sqrt{1+x^2}}$, shows that the function $f(x) = \frac{1}{\sqrt{1+x^2}}$, which has an infinite rational Chebyshev polynomial expansion, has a finite (i.e. $a_1 = -1$) mapped sine series. Thus, one can infer that the Chebyshev system is not necessarily superior to the mapped sine functions. Nevertheless, as a further investigation reveals two main differences lie in the fact that the mapped sine functions might not all be expressible as polynomials in x, and that the far field behavior shows all S_k to tend to zero. As a consequence, L^2 convergence of the sine expansion is much slower for bounded functions not decaying at infinity, whereas pointwise convergence can never be established in such cases.

So, overall, there are some aspects of inferiority of mapped sine functions, where it might be possible to alleviate them by applying different coordinate transforms. The use of the mapping from (3.2) here shall be argued by the need of the specific weight in order for the proof of Theorem 3.16 to hold.

Notation Index

$\propto$	proportional to
$\rightsquigarrow$	leads to
$\overset{!}{=}$	has to be equal to
$\Re$, $\Im$	real and imaginary part of a complex number
$a \sim b$	asymptotic representation, i.e. $a = b + o(b)$
∂_x^n	derivative operator, i.e. $\frac{\partial^n}{\partial x^n}$
	multi-dimensional partial derivative, if not otherwise indicated: $\partial_x^n = \prod_{i=1}^{m} \partial_{x_i}^{n/m}$
$\mathcal{F}$	Fourier transform, see (2.54)
$sb(\mathcal{K})$	Fourier symbol or multiplier of an operator $\mathcal{K}$
$f * g$	convolution, i.e. $\int f(x)g(y-x)dx$
A_x	(rational) Sturm-Liouville operator, see (3.26)
$\mathcal{J}_{\pm\infty}^{\alpha}$	Abel integral operator, see (3.95)
$\mathcal{R}^{\alpha}$	Riesz potential operator, see (3.98)
$\mathcal{R}_j$	jth Riesz transform, see (3.123)
$\mathcal{P}_N$	projection operator from some Banach space into some vector space of dimension N
$\mathbb{1}_A(x)$	characteristic function with value 1 if $x \in A$ and 0 otherwise
$R_i(x)$	ith rational Chebyshev polynomial for $x \in \mathbb{R}$, see (3.1)
Γ	Gamma function
$\mathbb{K}$	complete elliptic integral defined as $\mathbb{K}(m) = \int_0^{\pi/2} \left(1 - m\sin^2(\theta)\right)^{-1/2} d\theta$
$\mathit{\Gamma}$	real valued control parameter
$\langle \cdot, \cdot \rangle$	general inner product
$\lvert x \rvert$	absolute value for $x \in \mathbb{R}$, Euclidean norm for $x \in \mathbb{R}^n$
$\lvert k \rvert_s$	summed absolute value of a multi-index k, i.e. $\lvert k \rvert_s = \sum k_i$
$C^n(\Omega)$	space of n-times continuously differentiable functions on (the open domain) Ω
$C(\Omega)$	space of continuous functions
$C^{\infty}(\Omega)$	space of smooth functions
$C_l(\mathbb{R}^n)$	space of continuous functions, assuming a limit at infinity

$L^p_w(\Omega)$	space of weighted p-integrable functions on Ω (including $w \equiv 1$)
$\|\cdot\|_w$	norm on L^2_w defined as $\left(\int_\Omega |\cdot|^2 w(x)dx\right)^{1/2}$
$\langle\cdot,\cdot\rangle_w$	weighted inner product on L^2_w defined as $\langle f,g\rangle_w = \int_\Omega f(x)g(x)w(x)dx$
$\|\cdot\|_\infty$	norm on L^∞ defined as $\sup_{x\in\Omega}|\cdot|$, also used in $C^n(\Omega)$
$L^p_{loc}(\Omega)$	space of p-integrable functions on any finite ball in Ω
$H^r(\Omega)$	Sobolev space, $H^r(\Omega) := \{f \mid \|f\|_r < \infty\}$
$\|\cdot\|_r$	norm on H^r_u, see (3.18)
$\|\cdot\|_A$	norm on $H^r_{u,A}$, see (3.19)
$\|\cdot\|_{X\to Y}$	operator norm defined as $\sup_{f\neq 0\in X} \frac{\|\cdot f\|_Y}{\|f\|_X}$, denoted as $\|\cdot\|_X$ if $X = Y$
$\ell^p(\mathbb{N})$	space of p-summable sequences
$\|\cdot\|_{\ell^p}$	norm on ℓ^p, $\|a\|_{\ell^p} = (\sum |a_i|^p)^{1/p}$

References

Achleitner, F., Hittmeir, S. & Schmeiser, C. (2011): *On nonlinear conservation laws with a nonlocal diffusion term*, J. Diff. Equ. **250**, 2177–2196.

Akimenko, V.V. (1997): *On quadrature and cubature formulas for a class of multiple singular integrals*, Ukrainian Math. J. **49**,12, 1891–1898.

Alt, H.W. (2002): *Lineare Funktionalanalysis*, Springer-Verlag, Berlin.

Ball, J.M. (1977): *Remarks on blow-up and nonexistence theorems for nonlinear evolution equations*, Quart. J. Math. **28**, 473–486.

Barenblatt, G.I. (1979): *Similarity, Self-Similarity, and Intermediate Asymptotics*, Consultants Bureau, New York.

Boyd, J.P. (1982): *The Optimization of Convergence for Chebyshev Polynomial Methods in an Unbounded Domain*, Journal of Comp. Physics **45**, 43–79.

Boyd, J.P. (1987): *Spectral Methods Using Rational Basis Functions on an Infinite Interval*, Journal of Comp. Physics **69**, 112–142.

Boyd, J.P. (2001): *Chebyshev and Fourier Spectral Methods*, Dover Publications, Mineola.

Boyd, J.P. (2001a): *Rational Chebyshev Spectral Methods for Unbounded Solutions on an Infinite Interval Using Polynomial-Growth Special Basis Functions*, Computers and Mathematics with App. **41**, 1293–1315.

Braun, S. & Kluwick, A. (2002): *The effect of three-dimensional obstacles on marginally separated laminar boundary layer flows*, J. Fluid Mech. **460**, 57–82.

Braun, S. & Kluwick, A. (2003): *Analysis of a bifurcation problem in marginally separated laminar wall jets and boundary layers*, Acta Mech. **161**, 195–211.

Braun, S. & Kluwick, A. (2004): *Unsteady three-dimensional marginal separation caused by surface-mounted obstacles and/or local suction*, J. Fluid Mech. **514**, 121–152.

Braun, S., Scheichl, S. & Kluwick, A. (2012): *Asymptotic description of incipient separation bubble bursting*, to be published in PAMM.

Brown, S.N. & Stewartson, K. (1983): *On an integral equation of marginal separation*, SIAM J. Appl. Maths **43**, 1119–1126.

Calderon, A.P. & Zygmund, A. (1952): *On the existence of certain singular integrals*, Acta Math. **88**,1-2, 85–139.

Campbell Hetrick, B.M. & Hughes, R.J. (2007): *Continuous dependence results for inhomogeneous ill-posed problems in Banach space*, J. Math. Anal. Appl. **331**, 342–357.

Canuto, C., Hussaini, M.Y., Quarteroni, A. & Zang, T.A. (2006): *Spectral Methods*, Springer-Verlag, Berlin Heidelberg.

Cheney, E.W. (1966): *Introduction to Approximation Theory*, McGraw-Hill, New York.

Collatz, L. & Krabs, W. (1973): *Approximationstheorie*, Teubner, Stuttgart.

Diethelm, K. (2000): *Two general methods for the numerical approximation of multidimensional Cauchy principal value integrals*, ANZIAM J. **42**, E, 1–26.

Diogo, T., Lima, P.M. & Rebelo, M.S. (2006): *Numerical solution of a nonlinear Abel type Volterra integral equation*, Comm. Pure and Appl. Anal. **5**, 277–288.

Droniou, J., Gallouët, T. & Vovelle, J. (2002): *Global solution and smoothing effect for a non-local regularization of a hyperbolic equation*, J. Evol. Equ. **3**, 499–521.

Duck, P.W. (1989): *Three-dimensional marginal separation*, J. Fluid Mech. **202**, 559–575.

Duck, P.W. (1990): *Unsteady three-dimensional marginal separation, including breakdown*, J. Fluid Mech. **220**, 85–98.

E, W. & Engquist, B. (1997): *Blowup of Solutions of the Unsteady Prandtl's Equation*, Comm. Pure Appl. Math. **50**, 1287–1293.

Eckhaus, W. (1973): *Matched Asymptotic Expansions and Singular Perturbations*, North-Holland Publishing, Amsterdam.

Eggers, J. & Fontelos, M.A. (2009): *The role of self-similarity in singularities of partial differential equations*, Nonlinearity **22**, R1–R44.

Eldén, L. (1982): *Time Discretization in the Backward Solution of Parabolic Equations. I*, Math. Comp. **39**,159, 53–68.

Elliott, J.W. & Smith, F.T. (1987): *Dynamic stall due to unsteady marginal separation*, J. Fluid Mech. **179**, 489–512.

Engel, K.-J. & Nagel, R. (2000): *One-Parameter Semigroups for Linear Evolution Equations*, Springer-Verlag, New York.

Erdélyi, A. (1956): *Asymptotic Expansions*, Dover Publications, New York.

Forsythe, G.E. & Wasow, W.R. (1960): *Finite-Difference Methods for Partial Differential Equations*, John Wiley & Sons, New York.

Fromme, J.A. & Golberg, M.A. (1979): *Numerical Solution of a Class of Integral Equations Arising in Two-Dimensional Aerodynamics*, Solution Methods for Integral Equations, M. A. Golberg (ed.), Plenum Press, New York, 109–146.

Galaktionov, V.A. (2009): *Five types of blow-up in a semilinear fourth-order reactiondiffusion equation: an analyticnumerical approach*, Nonlinearity **22**, 1695–1741.

Galaktionov, V.A. & Vázquez, J.L. (2002): *The Problem Of Blow-Up In Nonlinear Parabolic Equations*, DISCR. CONT. DYN. SYST. **8**, 399–433.

Glauert, M.B. (1957): *A boundary layer theorem, with applications to rotating cylinders*, J. Fluid Mech. **2**, 89–99.

Golberg, M.A. (1979): *A Survey of Numerical Methods for Integral Equations*, Solution Methods for Integral Equations, M. A. Golberg (ed.), Plenum Press, New York, 1–58.

Golberg, M.A. (1990): *Perturbed Projection Methods for Various Classes of Operator and Integral Equations*, Numerical Solution of Integral Equations, M. A. Golberg (ed.), Plenum Press, New York, 71–129.

Golberg, M.A. (1990a): *Introduction to the Numerical Solution of Cauchy Singular Integral Equations*, Numerical Solution of Integral Equations, M. A. Golberg (ed.), Plenum Press, New York, 183–308.

Goldstein, S. (1948): *On laminar boundary-layer flow near a position of separation*, Q. J. Mech. Appl. Maths **1**, 43-69.

Gorenflo, R. & Vessella, S. (1991): *Abel Integral Equations*, Lecture Notes in Mathematics, Springer-Verlag, Berlin-Heidelberg.

Gottlieb, D. & Orszag, S.A. (1977): *Numerical Analysis of Spectral Methods*, SIAM, CBMS-NFS, Philadelphia.

Guo, B.-Y. (1998): *Spectral Methods and Their Applications*, World Scientific, Singapore.

Hackbusch, W. (1995): *Integral Equations: Theory and Numerical Treatment*, Birkhäuser, Basel.

Hadamard, J. (1923): *Lectures on Cauchy's Problem in Linear Partial Differential Equations*, Yale University Press, New Haven.

Handscomb, D.C. (1966): *Functions of many variables*, Methods of Numerical Approximation, D. C. Handscomb (ed.), Pergamon, Elmsford, N.Y., 191–194.

Hesthaven, J.S., Gottlieb, S. & Gottlieb, D. (2007): *Spectral Methods for Time-Dependent Problems*, Cambridge University Press, Cambridge.

Hörmander, L. (1985): *The Analysis of Linear Partial Differential Operators III*, Springer-Verlag, Berlin Heidelberg.

Jonas, P. & Louis, A.K. (2000): *Approximate inverse for a one-dimensional inverse heat conduction problem*, Inverse Problems **16**, 175–185.

Kress, R. (1999): *Linear Integral Equations*, Springer-Verlag, New York.

Landau, L.D. & Lifshitz, E.M. (1959): *Fluid Mechanics*, Pergamon Press, London.

Lavrent'ev, M.M., Romanov, V.G. & Shishatskii, S.P. (1999): *Ill-posed Problems of Mathematical Physics and Analysis*, Translations of Mathematical Monographs, Amer. Math. Soc., Providence, RI.

Leipnik, R.B. & Pearce, C.M. (2007): *The multivariate Faa di Bruno formula and multivariate Taylor expansions with explicit integral remainder terms*, ANZIAM J. **48**, 327–341.

Louis, A.K. (1989): *Inverse und schlecht gestellte Probleme*, Teubner, Stuttgart.

Marchioro, C. & Pulvirenti, M. (1994): *Mathematical Theory of Incompressible Nonviscous Fluids*, Springer-Verlag, New York.

Mason, J.C. (1967): *Chebyshev polynomial approximation for the L-membrane eigenvalue problem*, SIAM J. Appl. Math. **15**, 172–186.

Mason, J.C. (1980): *Near-Best Multivariate Approximation by Fourier Series, Chebyshev Series and Chebyshev Interpolation*, J. Approx. Theory **28**, 349–358.

Mason, J.C. & Handscomb, D.C. (2003): *Chebyshev Polynomials*, Chapman & Hall/CRC, Boca Raton.

Mikhlin, S.G. (1936): *Singular integral equations with two independent variables* (in Russian), Mat. Sbor. 1:**4**(43), 535–550.

Mikhlin, S.G. (1965): *Multidimensional singular integrals and integral equations*, Pergamon Press, New York.

Mikhlin, S.G. & Prößdorf, S. (1980): *Singuläre Integraloperatoren*, Akademie-Verlag, Berlin.

Mishkov, R.L. (2000): *Generalization of the formula of Faa di Bruno for a composite function with a vector argument*, Internat. J. Math. & Math. Sci. **24**(7), 481–491.

Moré, J.J., Garbow, B.S. & Hillstrom, K.E. (2000): *User Guide for MINPACK-1*, Technical Report ANL-80-74, Argonne National Laboratory, Argonne, IL, USA.

Narayan, A.C. & Hesthaven, J.S. (2011): *A generalization of the Wiener rational basis functions on infinite intervals. Part I - Derivation and properties*, Math. Comp. **80**(275), 1557–1583.

Ortega, J.M. & Rheinboldt, W.C. (1966): *On Discretization and Differentiation of Operators with Application to Newtons Method*, SIAM J. Numer. Anal. **3**, 143–156.

Payne, L.E. (1975): *Improperly Posed Problems in Partial Differential Equations*, SIAM, Philadelphia, Pa.

Petrowsky, I.G. (1937): *Über das Cauchysche Problem für Systeme von partiellen Differentialgleichungen*, Mat. Sbor. **2**(44), 815–870.

Petrowsky, I.G. (1955): *Vorlesungen über partielle Differentialgleichungen*, Teubner, Leipzig.

Piessens, R., de Doncker-Kapenga, E., Überhuber, C. & Kahaner, D. (1983): *QUADPACK, A Subroutine Package for Automatic Integration*, Springer-Verlag, Berlin-Heidelberg.

Podlubny, I. (1999): *Fractional Differential Equations*, Academic Press, San Diego.

Powell, M.J.D. (1970): *A hybrid method for nonlinear equations*, Numerical Methods for Nonlinear Algebraic Equations, P. Rabinowitz (ed.), Gordon and Breach, London, 87–115.

Prößdorf, S. (1974): *Einige Klassen singulärer Gleichungen*, Birkhäuser Verlag, Basel.

Rodino, L. (1993): *Linear Partial Differential Operators in Gevrey Spaces*, World Scientific, Singapore.

Ruban, A.I. (1981): *Asymptotic theory of short separation regions on the leading edge of a slender airfoil*, Izv. Akad. Nauk SSSR: Mekh. Zhidk. Gaza **1**, 42–51 (Engl. transl. *Fluid Dyn.* **17**, 33–41).

Ruban, A.I. (1982): *Stability of preseparation boundary layer on the leading edge of a thin airfoil*, Izv. Akad. Nauk SSSR: Mekh. Zhidk. Gaza **6**, 55–63, (Engl. transl. *Fluid Dyn.* **17**, 860–867).

Ruban, A.I. (2010): *Asymptotic Theory of Separated Flows*, Asymptotic Methods in Fluid Mechanics: Survey and Recent Advances, CISM vol. 523, H. Steinrück (ed.), Springer-Verlag, Wien New York, 311–408.

Ryzhov, O.S. (1980): *On the Unsteady Three-Dimensional Boundary Layer Freely Interacting with the External Stream*, Prikl. Mathem. Mekhan. **44**, 1035–1052, (Engl. transl. *PMM U.S.S.R.* **44**, 739–750).

Ryzhov, O.S. & Smith, F.T. (1984): *Short-length instabilities, breakdown and initial value problems in dynamic stall*, Mathematika **31**, 163–177.

Samko, S.G. (1976): *On Spaces of Riesz Potentials*, Math. USSR Izvestija **10**,no. 5, 1089–1117.

Samko, S.G. (1998): *A new approach to the inversion of the Riesz potential operator*, Fract. Calc. and Applied Anal. **1**,3, 225–245.

Samko, S.G. (1999): *On Local Summability of Riesz Potentials in the case* $Re\,\alpha > 0$, Analysis Mathematica **25**, 205–210.

Scheichl, S., Braun, S. & Kluwick, A. (2008): *On a similarity solution in the theory of unsteady marginal separation*, Acta Mech. **201**, 153–170.

Sell, G.R. & You, Y. (2002): *Dynamics of Evolutionary Equations*, Springer-Verlag, New York.

Shen, J. & Wang, L.-L. (2009): *Some Recent Advances on Spectral Methods for Unbounded Domains*, Commun. Comput. Phys. **5**, 195–241.

Showalter, R.E. (1974): *The Final Value Problem for Evolution Equations*, J. Math. Anal. Appl. **47**, 563–572.

Sloan, I.H. (1981): *Quadrature Methods for Integral Equations of the Second Kind Over Infinite Intervals*, Mathem. Comp. **36**, 511–523.

Sloan, I.H. (1990): *Superconvergence*, Numerical Solution of Integral Equations, M. A. Golberg (ed.), Plenum Press, New York, 35–70.

Smith, F.T. (1982): *Concerning dynamic stall*, Aeron. Quart. **33**, 331–352.

Smith, F.T. & Elliott, J.W. (1985): *On the abrupt turbulent reattachment downstream of leading-edge laminar separation*, Proc. R. Soc. Lond. A **401**, 1–27.

Sobolev, S.L. (1938): *On a theorem of functional analysis* (in Russian), Mat. Sbor. **4**(46), 471–497.

Stein, E.M. (1970): *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ.

Stewartson, K. (1970): *Is the singularity at separation removable?*. J. Fluid Mech. **44**, 347–364.

Stewartson, K., Smith, F.T. & Kaups, K. (1982): *Marginal separation*. Stud. Appl. Maths **67**, 45–61.

Sychev, V.V., Ruban, A.I., Sychev, Vik.V. & Korolev, G.L. (1998): *Asymptotic Theory of Separated Flows*, Cambridge University Press, Cambridge.

Szegö, G. (1939): *Orthogonal Polynomials*, American Mathematical Society, Providence, RI.

Tikhonov, A.N. & Arsenin, V.Y. (1977): *Solutions of ill-posed problems*, Winston & Sons, Washington D.C..

Timan, A.F. (1963): *Theory of Approximation of Functions of a Real Variable*, Pergamon Press, Oxford.

Trefethen, L.N. (2000): *Spectral Methods in MATLAB*, SIAM, Philadelphia.

Wang, Z.-Q. & Guo, B.-Y. (2002): *A rational approximation and its applications to nonlinear partial differential equations on the whole line*, J. Math. Anal. Appl. **274**, 374–403.

Weideman, J.A.C. (2003): *Computing the dynamics of complex singularities of nonlinear PDEs*, SIAM J. Appl. Dyn. Sys. **2**, 171–186.