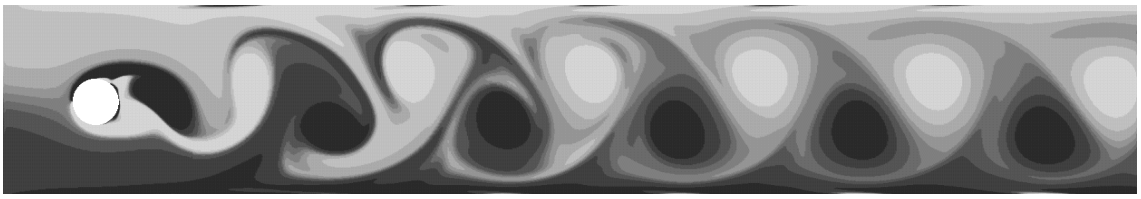
Serge Kräutle

A Navier-Stokes Solver Based on CGBI and the Method of Characteristics

Dissertationsschrift



Den Naturwissenschaftlichen Fakultäten
der Friedrich-Alexander-Universität Erlangen-Nürnberg

vorgelegt 2001

dedicated to Uwe

Front page: Vorticity field of a 2d-flow past a cylinder in a channel at Reynolds number 500, computed in parallel by CGBI.

A Navier-Stokes Solver Based on CGBI and the Method of Characteristics

Den Naturwissenschaftlichen Fakultäten
der Friedrich-Alexander-Universität Erlangen-Nürnberg
zur
Erlangung des Doktorgrades

vorgelegt von

Serge Kräutle

aus Bad Driburg

Als Dissertation genehmigt von den Naturwissenschaftlichen Fakultäten der Universität Erlangen-Nürnberg

Tag der mündlichen Prüfung:	19.02.2002
Vorsitzender der Promotionskommision:	Prof. Dr. A. Magerl
Erstberichterstatter:	Prof. Dr. W. Borchers
Zweitberichterstatter:	Prof. Dr. R. Rautmann, Prof. Dr. H. Sohr

Contents

Chapter 1

Introduction

The main subject of this thesis is the description and investigation of the *Conjugate Gradient Boundary Iteration* (CGBI) method. CGBI is a parallelization method for symmetric elliptic boundary value problems based on non-overlapping domain decomposition. It was proposed by Borchers [5]. This thesis gives a detailed overview on the theory of CGBI, the development of preconditioners on different meshes, test runs and the application of CGBI to Navier-Stokes flow problems.

Beside the possibility to make use of the full computational power of parallel computers, domain decomposition methods like CGBI facilitate the combination of different local solvers on different parts of the computational domain, e.g. highly accurate *spectral solvers* on simply-formed parts of the domain with the highly flexible *finite element method* (FE, FEM) on more complicated parts.

Thus, as an application for CGBI, a parallel solver for the incompressible Navier-Stokes equations

$$\begin{aligned} \vec{u}_t + \vec{u}\,\nabla\vec{u} - \nu\,\Delta\vec{u} + \nabla p &= \vec{f}, \\ \operatorname{div}\vec{u} &= 0. \end{aligned} \tag{1.1}$$

is constructed in this paper. Using a *pressure correction scheme* (the pressure correction methods, also called *fractional step methods*, were introduced by Chorin & Temam [51]) our solver splits each Navier-Stokes timestep into one hyperbolic and two elliptic problems (see Chapter 5 and also [5]). The hyperbolic problem is solved with the method of characteristics. The elliptic problems are solved with CGBI. Chapter 2 describes CGBI with its theoretical background and Chapter 3 is devoted to the construction of efficient preconditioners for CGBI. Both chapters also present numerical tests (Sections 2.8, 3.1.5, 3.3, 3.4.3).

In Chapter 4 our scheme of characteristics is investigated. Error estimates and and a theoretical investigation of the stability on different meshes are given.

In Chapter 5 the full Navier-Stokes solver based on CGBI and the method of characteristics is presented. Test runs for the flow past a backward facing step and the flow past a cylinder (both modeled in 2d) are given.

Beside the approximation of the Navier-Stokes equations, both the CGBI method and the characteristics scheme may have lots of applications. Therefore it makes sense to investigate highly accurate versions of these methods although the presently used Navier-Stokes time splitting scheme may reduce this accuracy. In accordance with this philosophy, the Chapters 2 and 3 on the one hand and Chapter 4 on the other hand can be read independently from each other.

In this introduction I will give a brief summary of the CGBI solver and of the characteristics solver. I will also remark on the mathematical methods which are used in this paper.

1. The CGBI solver. The CGBI solver is a domain decomposition method for the solution of symmetric elliptic problems in parallel. Obviously it is easy to find the solution of the global partial differential equation in parallel as soon as the corresponding boundary conditions on the subdomain interfaces are known. CGBI searchs these boundary conditions of *Neumann* type by a *conjugate gradient* (CG) iteration. The name CGBI comes from the fact that the global conjugate gradient method *and its preconditioner* are running on the interfaces ('boundaries') of the subdomains. During each CG iteration step local elliptic problems on the subdomains are to be solved. The CGBI method enables us to couple different local solvers to get a high performance. For example, on rectangular subdomains spectral methods provide highly efficient solvers. On non-rectangular subdomains, finite element methods (FE, FEM) should be used because of their flexibility.

In Chapter 2 the theory of the CGBI method is expounded. In the context of our Navier-Stokes solver, the CGBI method is used

- on resolvent type equations (5.4) for the velocity involving *Dirichlet* boundary conditions on the 'physical' walls and
- on a Poisson equation (5.5) for the pressure which uses *Neumann* boundary conditions on the physical walls.

When *Neumann* boundary conditions are used for the Poisson equation, the question occurs how to avoid ill-posed local problems: The *global* boundary value problem may be well-posed, e.g. because of Dirichlet conditions on the outflow part of the domain, but some subdomains (the so-called 'floating subdomains') may only have Neumann boundaries. Thus, beside the Dirichlet case (Sections 2.2, 2.3) special emphasis is layed on the application of CGBI in the Neumann case (Section 2.4).

The relation of CGBI to domain decomposition methods like *FETI* (finite element tearing and interconnecting method) developed in the early 1990s for problems from structural mechanics and the *Schur* method are discussed in Sec. 2.9. In fact, CGBI is very similar to FETI; the main advantage is the fact that CGBI uses preconditioners acting only on the interfaces (see 2.).

Many publications on domain decomposition methods ([3] [19] [21] [22] [40]) focus on the discrete ('matrix') level for the description of the algorithm and for the investigation of preconditioners. In this thesis we concentrate on the investigation of the underlying continuous problems. This approach directly leads to our preconditioner (see end of Sec. 2.2) as a discretization of the square root of the negative of the Laplacian operator acting on the subdomain interfaces.

2. The CGBI interface preconditioner. The theoretical investigations of Chapter 2 already show how to construct a suitable preconditioner for CGBI. This preconditioner yields a condition number (i.e. a CGBI convergence rate) *independent* of the number of mesh points and the number of subdomains. In contrast to other approaches ([45] Sec. 3.3.2 and the authors cited there and (very recent) [28]), our interface preconditioner does *not* require the solution of any time-consuming subdomain-based problems. Here we have to mention the work of Dryja [15] and Bjørstad & Widlund [4] who proposed interface-based preconditioners in the context of the *Schur* method with local FE/FD solvers in the 1980.

In Chapter 3 some efficient discretizations of our preconditioning operator are developed and tested for the case of rectangular subdomains. These discretizations are easy to find on an *equidistant* mesh. However, our Chebyshev spectral solvers are working on a Chebyshev-Gauss-Lobatto mesh. It turns out that on such a mesh, it is much more difficult to find a discretization of the preconditioner. The mathematical method to construct a Gauss-Lobatto mesh based preconditioner is the interpolation theory of weighted Sobolev spaces. The resulting preconditioner yields for Dirichlet problems a CGBI convergence rate independent of the number of grid points, the number of the subdomains, the size of the channel and the resolvent equation parameter σ, both on equidistant and on Gauss-Lobatto boundary meshes. Only for the combination of a Gauss-Lobatto boundary mesh and *Neumann* boundary conditions on the physical channel walls, a very moderate dependence of the condition number on the discretization parameter may occur.

A disadvantage of our preconditioner is, that the condition number increases for bad aspect ratios ($\lesssim 0.1$) of the subdomains. For this case and Neumann boundary conditions a different preconditioner is proposed in Sec. 3.3.

The discretization of our preconditioners proposed in Sec. 3.1 is based on the spectral decomposition (fast Fourier transform, FFT) of a function given on the interfaces. In Sec. 3.4, discretizations based on sparse matrices are presented as an alternative.

3. The characteristics solver. For the solution of the transport equation we are using a method of characteristics. The characteristics method reflects the 'local' character of this equation. Obviously, this local character is a big advantage for the parallelization. Compared to the *semi-Lagrangian* method

and the *semi-implicit* method [42], our collocation characteristics method allows larger time steps due to better stability properties, whereas the semi-Lagrangian method has the advantage of avoiding spatial interpolations which are costly in the context of highly accurate spectral approximations.

In Chapter 4 we develop our characteristics scheme for the nonlinear equation

$$u_t + a(t, x, u) \cdot \nabla u = 0,$$

which is slightly more general than needed for our Navier-Stokes solver. However, until now, our theoretical investigation is restricted to the case of *one* space dimension. In our approach we are calculating a characteristic starting at a grid point x_k at time t_{n+1} backward in time until $t = t_n$ (see Chapters 4.2, 5 for the details). In the course of this the discrete data of the velocity field are interpolated using polynomials in time and piecewise polynomial ansatz functions in space. In Chapter 4 detailed convergence and stability investigations are performed. It turns out that for linear spatial ansatz functions the stability of the scheme is trivial. For higher order ansatz functions, however, some restrictions appear: For equidistant or quasi-uniform meshes, only the Courant number has to be bounded which is a very weak condition. On a Chebyshev-Gauss-Lobatto mesh the stability proof for higher order interpolation requires some stricter conditions. However, numerical tests in Section 4.4.6 and in Chapter 5 reveal better stability properties than expected due to Chapter 4.4.4-4.4.5.

The basis of our stability investigations are the stability theorems published by Lax [32] [33]. To make these theorems applicable for our purposes, some modifications and generalizations (Theorems 4.6, 4.8) are necessary. The application of these theorems requires the investigation of certain properties (including a Lipschitz condition) of the coefficient functions of the time stepping scheme.

4. Cooperations, note of thanks. The research presented in this thesis was supported by the *Deutsche Forschungsgemeinschaft* (DFG) and by the *Centre National de la Recherche Scientifique* (CNRS). The programming was done in collaboration with the University of Paderborn (Kerstin Wielage, Dr. Nicole Roß, Prof. Rautmann) and the University of Nice, France (Prof. Peyret, Dr. Pasquetti & co-workers). I would like to express my thanks for cooperation and support to all of them; especially to Prof. Borchers (University Erlangen-Nuremberg), who is in charge of this thesis, for his support and advice.

5. How to attain this paper. This paper is available on the World Wide Web at

`http://www.am.uni-erlangen.de/am1/publications/dipl_phd_thesis/PhD_Kraeutle.ps.gz`

Chapter 2

CGBI

CGBI stands for *Conjugate Gradient Boundary Iteration*. It is a domain decomposition method to parallelize the computation of symmetric elliptic boundary value problems. We concentrate on the solution of the Poisson equation and the Helmholtz resolvent equation

$$Lw := -\Delta w + \sigma w = f, \quad \sigma \geq 0, \tag{2.1}$$

on a bounded domain $\Omega \subset \mathbb{R}^n$, $n \geq 2$, with a Lipschitz continuous[1] boundary $\partial\Omega$ and appropriate boundary conditions on $\partial\Omega$. In this paper, for the sake of simplicity $n = 2$ is taken, however, the method is also applicable in higher dimensions without much modification.

Best suited for the domain decomposition method CGBI are domains with a large extension in *one* direction; for practical applications of CGBI in the context of a Navier-Stokes solver we may think of a rectangular channel with (or without) an obstacle or systems of pipes (see Figs. 2.1-2.4). Test runs for the geometry of Fig. 2.1 with a circular obstacle M and of Figs. 2.2, 2.4 (right hand part) will be presented in Chapter 5. Because of its simplicity, the geometry of Fig. 2.2 is used in Chapter 3 for theoretical investigations on the condition number; however, this geometry is also of practical interest, e.g. for the computation of a jet or a Bickley flow [24] or certain heat driven flow problems.

The boundary on the left and on the right side of the channel are called Γ^I rsp. Γ^O. This nomenclature points to the fact that the test runs in Chapter 5 have inflow boundary conditions on the left and outflow boundary conditions on the right part of $\partial\Omega$. We set $\Gamma^W := \partial\Omega \setminus (\Gamma^I \cup \Gamma^O)$, where the 'W' stands for the physical wall of the channel.

We pose (homogeneous or inhomogeneous) Neumann and/or Dirichlet boundary conditions on $\partial\Omega$:

[1] See [1] A 6.2 for the definition of a Lipschitz boundary.

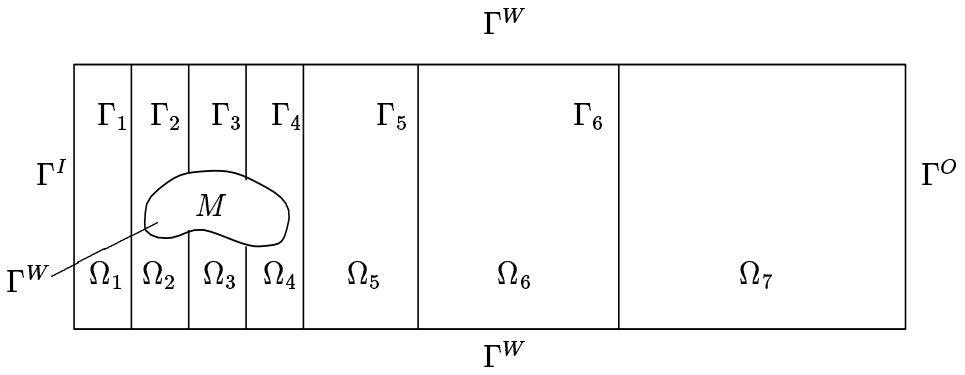


Figure 2.1: Channel with obstacle.

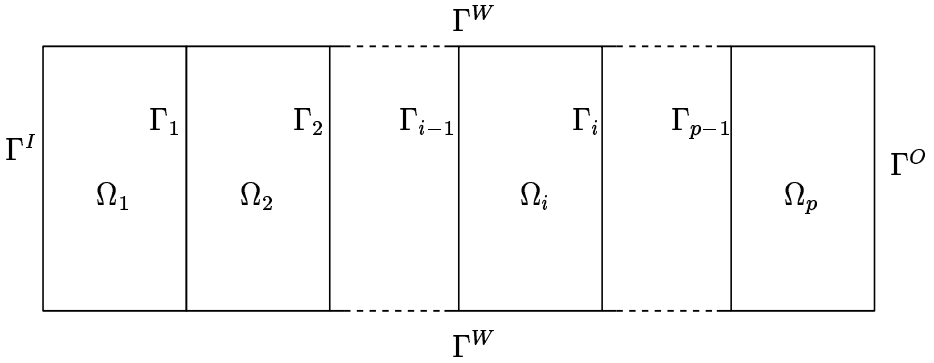


Figure 2.2: Channel without obstacle.

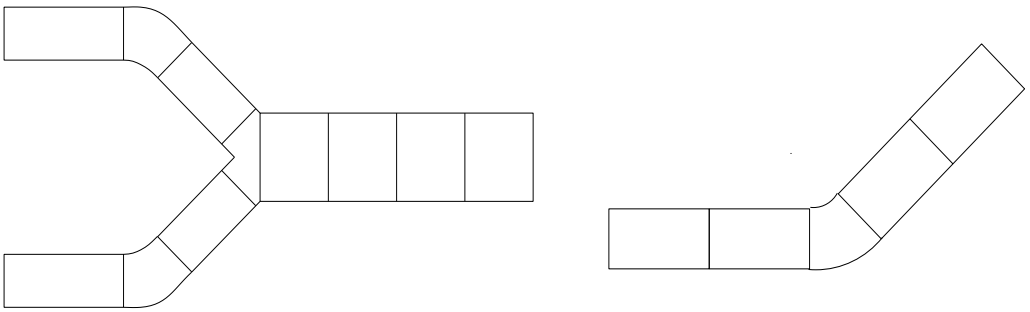


Figure 2.3: (Systems of) Pipes.

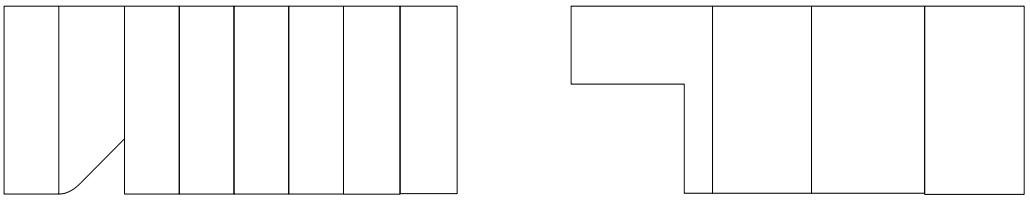


Figure 2.4: Flow behind a step.

$$
\begin{aligned}
w &= g^{Dir} \quad \text{on} \quad \Gamma^{Dir} \\
\frac{\partial w}{\partial n} &= g^{Nm} \quad \text{on} \quad \Gamma^{Nm}
\end{aligned}
\tag{2.2}
$$

where $\Gamma^{Dir} \cup \Gamma^{Nm} = \partial\Omega$, $\Gamma^{Dir} \cap \Gamma^{Nm} = \emptyset$.

The domain Ω is decomposed into the subdomains Ω_1,...,Ω_p where $p \geq 2$ is the number of available processors. The name 'boundary iteration' comes from the fact that the unknowns are distributed only on the artificial boundaries ('interfaces') Γ_i, $i = 1, ..., p-1$ between the subdomains. We set

$$\Gamma := \bigcup_{i=1}^{p-1} \Gamma_i.$$

In this paper we will assume that

$$\bar{\Gamma}_i \cap \bar{\Gamma}_j = \emptyset$$

for $i \neq j$, i.e. there are *no interior crosspoints* of the subdomain boundaries. In our future work this restriction will be dropped. Such a generalization will make the application of CGBI on non-channel-like domains easier.

Processor number i solves an equation of type (2.1) locally on its subdomain Ω_i. The correct *Neumann* boundary conditions for the local problems on the interfaces are unknown. As in Section 2.2 explained, these boundary values are the solution of a certain minimization principle. This minimization principle is solved iteratively by the CG method (see Section 2.3).

The relation between CGBI and other domain decomposition methods is explained in Sec. 2.9.

2.1 Some interpolation spaces and properties

Before we can consider the algorithm and the minimization principle related to CGBI we have to define some symbols and some function spaces, expecially interpolation spaces.

We will use the writing $\|\cdot\|_1 \sim \|\cdot\|_2$ to express that two norms $\|\cdot\|_1$, $\|\cdot\|_2$ on a common vector space are equivalent. To keep the notation simple, we will also use the writing $\|\cdot\|_1^2 \sim \|\cdot\|_2^2$ if $\|\cdot\|_1 \sim \|\cdot\|_2$.

We suppose that Ω and each subdomain Ω_i are bounded and have Lipschitz boundaries. Let $C_0^p(\Omega_i)$, $0 \leq p \leq \infty$, be the vector space of $C^p(\Omega_i)$-functions with a support which is a compact subset of Ω_i. Let $H^1(\Omega_i)$, $H_0^1(\Omega_i)$ be the well known Sobolev spaces

$$\begin{aligned} H^1(\Omega_i) &= closure\{C^\infty(\bar{\Omega}_i)\} \\ H_0^1(\Omega_i) &= closure\{C_0^\infty(\Omega_i)\} \end{aligned}$$

where the closure is taken with respect to the norm

$$\|u\|_{H^1(\Omega_i)} := \left[\int_{\Omega_i} (|\nabla u|^2 + |u|^2)\, dx\right]^{1/2}. \tag{2.3}$$

Furthermore we will need the spaces

$$\begin{aligned} L^2_{mv}(\Omega_i) &:= \left\{ u \in L^2(\Omega_i) \mid \int_{\Omega_i} u\ dx = 0 \right\}, \\ H^1_{mv}(\Omega_i) &:= \left\{ u \in H^1(\Omega_i) \mid \int_{\Omega_i} u\ dx = 0 \right\}. \end{aligned} \tag{2.4}$$

For all the three spaces $H^1(\Omega_i)$, $H_0^1(\Omega_i)$, $H^1_{mv}(\Omega_i)$, (2.3) is a norm; however, using the well known Poincaré inequality[2], $H_0^1(\Omega_i)$ and $H^1_{mv}(\Omega_i)$ can be equipped with the norm

$$\|u\|_{H_0^1(\Omega_i)} := \|u\|_{H^1_{mv}(\Omega_i)} := \|\nabla u\|_{L^2(\Omega_i)}. \tag{2.5}$$

Following the notation of [34], Chapter 1, Theorem 11.7, we define $H_{00}^{1/2}(\Gamma_i)$ as the interpolation space between $L^2(\Gamma_i)$ and $H_0^1(\Gamma_i)$ for the index $1/2$:

$$H_{00}^{1/2}(\Gamma_i) := [L^2(\Gamma_i), H_0^1(\Gamma_i)]_{\frac{1}{2}} \tag{2.6}$$

[2] A rather general version of the Poincaré inequality covering the case of $H^1_{mv}(\Omega)$ is given in Lemma 2.8. The application of this lemma is demonstrated in the proof of 2.9.

Furthermore we set

$$H^{1/2}(\Gamma_i) := [L^2(\Gamma_i), H^1(\Gamma_i)]_{\frac{1}{2}}, \tag{2.7}$$

$$H^{1/2}_{mv}(\Gamma_i) := [L^2_{mv}(\Gamma_i), H^1_{mv}(\Gamma_i)]_{\frac{1}{2}}. \tag{2.8}$$

The spaces (2.6)-(2.8) are equipped with the norms

$$\|u\|^2_{H^{1/2}_{00}(\Gamma_i)^*} := \|u\|^2_{L^2(\Gamma_i)} + \|(-\Delta_0)^{1/4}u\|_{L^2(\Gamma_i)} \tag{2.9}$$

$$\|u\|^2_{H^{1/2}_{mv}(\Gamma_i)^*} := \|u\|^2_{L^2(\Gamma_i)} + \|(-\Delta_{Nm})^{1/4}u\|_{L^2(\Gamma_i)} \tag{2.10}$$

$$\|u\|^2_{H^{1/2}(\Gamma_i)} := \|u\|^2_{L^2(\Gamma_i)} + \|(id-\Delta_{Nm})^{1/4}u\|_{L^2(\Gamma_i)} \tag{2.11}$$

Here, Δ_0 is the Laplacian defined on $H^2(\Gamma_i)\cap H^1_0(\Gamma_i)$ and Δ_{Nm} is the Laplacian defined on the subspace of $H^2(\Gamma_i)$ consisting of all the functions with homogeneous Neumann boundary values.[3] For the explanation of these norms and fractional powers of self-adjoint positive operators see e.g. [34] Chapter 1 Section 2.1 and the texts cited there. Definitions and properties of interpolation spaces are given in Chapter 1 in [34]. Let us summarize the properties of the interpolation spaces which are meaningful for this chapter:

$H^{1/2}(\Gamma_i)$ is equal to the space $H^{1/2}_0(\Gamma_i)$ being the closure of $C^\infty_0(\Gamma_i)$ in the $\|\cdot\|_{H^{1/2}(\Gamma_i)}$-norm (see [34] Chapter 1 Theorem 11.1).

An important characterization of $H^{1/2}_{00}(\Gamma_i)$ is given in [34], Chapter 1, Theorem 11.7: It is

$$H^{1/2}_{00}(\Gamma_i) = \{u\in H^{1/2}(\Gamma_i)\,|\,\rho^{-1/2}u\in L^2(\Gamma_i)\}, \tag{2.12}$$

and the equivalence

$$\|u\|_{H^{1/2}_{00}(\Gamma_i)} \sim \left(\|u\|^2_{H^{1/2}(\Gamma_i)} + \|\rho^{-1/2}u\|^2_{L^2(\Gamma_i)}\right)^{1/2} \tag{2.13}$$

holds in the space $H^{1/2}_{00}(\Gamma_i)$. Herein, ρ is a $C^\infty(\Gamma_i)$-function with $\lim\limits_{x\to x_0}\frac{\rho(x)}{dist(x,\partial\Gamma_i)} = d\neq 0$ for all $x_0\in\partial\Gamma_i$, $x\in\Gamma_i$. (The distance between $\partial\Gamma_i$ and x is taken along Γ_i.)

An alternative characterization of $H^{1/2}_{mv}$ is

$$H^{1/2}_{mv}(\Gamma_i) = \left\{\varphi\in H^{1/2}(\Gamma_i)\,|\,\int\limits_{\Gamma_i}\varphi\,do = 0\right\}. \tag{2.14}$$

This can be proved by regarding L^2_{mv}, H^1_{mv} in (2.8) as quotient spaces and using Theorem 13.2 in [34].

[3] $-\Delta_0$ on $H^2(\Gamma_i)\cap H^1_0(\Gamma_i)$, $id-\Delta_{Nm}$ on $H^2(\Omega)\cap\{u\,|\,\partial u/\partial\nu = 0\}$, $-\Delta_{Nm}$ on $H^2(\Omega)\cap H^1_{mv}(\Omega)\cap\{u\,|\,\partial u/\partial\nu = 0\}$ are self-adjoint and (strictly) positive on the domain of their definition, therefore the fractional powers can be defined. The existence of Neumann boundary values for functions in $H^2(\Gamma_i)$ is a consequence of the Trace Theorem (Satz 8.7 in [57], p. 130).

From (2.13) we derive $H_{00}^{1/2}(\Gamma_i) \subset H^{1/2}(\Gamma_i)$; no '=' holds, as the constant function $1 \in H^1(\Gamma_i) \subset H^{1/2}(\Gamma_i)$, but $\|\rho^{-1/2}\|_{L^2(\Gamma_i)} = \infty$.

Let us consider (2.9)-(2.10). As $-\Delta_0$ and $-\Delta_{Nm}$ are positive definite operators (see Poincaré Lemma 2.8), the same holds for their powers and we get

$$\|(-\Delta_0)^{1/4} u\|_{L^2(\Gamma_i)} \geq c \, \|u\|_{L^2(\Gamma_i)}, \quad \|(-\Delta_{Nm})^{1/4} u\|_{L^2(\Gamma_i)} \geq c \, \|u\|_{L^2(\Gamma_i)}.$$

Therefore dropping the term $\|u\|^2_{L^2(\Gamma_i)}$ in (2.9)-(2.10) leads to the norms

$$\|u\|_{H_{00}^{1/2}(\Gamma_i)} := \|(-\Delta_0)^{1/4} u\|_{L^2(\Gamma_i)}, \tag{2.15}$$

$$\|u\|_{H_{mv}^{1/2}(\Gamma_i)} := \|(-\Delta_{Nm})^{1/4} u\|_{L^2(\Gamma_i)} \tag{2.16}$$

which are equivalent to the norms (2.9), (2.10), respectively.

Function spaces C_0^∞, $H_{00}^{1/2}$, $H^{1/2}$, $H_0^{1/2}$ defined on the interfaces Γ_i are leading to function spaces on Γ in a natural way, e.g.

$$\begin{aligned} C_0^\infty(\Gamma) &:= \{\varphi : \Gamma \to \mathbb{R} \mid \varphi|_{\Gamma_i} \in C_0^\infty(\Gamma_i) \; \forall \, i = 1, ..., p-1\}, \\ H_{00}^{1/2}(\Gamma) &:= \{\varphi : \Gamma \to \mathbb{R} \mid \varphi|_{\Gamma_i} \in H_{00}^{1/2}(\Gamma_i) \; \forall \, i = 1, ..., p-1\}. \end{aligned}$$

The norm associated with $H_{00}^{1/2}(\Gamma)$ will be

$$\|\varphi\|^2_{H_{00}^{1/2}(\Gamma)} := \sum_{i=1}^{p-1} \|\varphi|_{\Gamma_i}\|^2_{H_{00}^{1/2}(\Gamma_i)},$$

and so on.

Let $H^{-1/2}(\Gamma_i)$ be the dual space to $H_{00}^{1/2}(\Gamma_i)$. The application of an element of $H^{-1/2}$ onto an element of $H^{1/2}$ will be denoted as follows:

$$\begin{aligned} \langle \cdot, \cdot \rangle_{\Gamma_i} &:= \langle \cdot, \cdot \rangle_{H^{-1/2}(\Gamma_i), H_{00}^{1/2}(\Gamma_i)}, \\ \langle \cdot, \cdot \rangle_{\Gamma} &:= \langle \cdot, \cdot \rangle_{H^{-1/2}(\Gamma), H_{00}^{1/2}(\Gamma)} \end{aligned}$$

We will use the same notation $\langle \cdot, \cdot \rangle_{\Gamma_i}$ for the duality between $H_{mv}^{-1/2}(\Gamma_i)$ and $H_{mv}^{1/2}(\Gamma_i)$ and between $(H^{1/2}(\Gamma_i))^*$ and $H^{1/2}(\Gamma_i)$.

There is a relation between the elements of $H^{-1/2}(\Gamma)$ and $H^{-1/2}(\Gamma_i)$: For $\varphi_i \in H^{-1/2}(\Gamma_i)$, φ defined by $\langle \varphi, \psi \rangle_\Gamma := \sum_{i=1}^{p-1} \langle \varphi_i, \psi|_{\Gamma_i} \rangle_{\Gamma_i}$ is in $H^{-1/2}(\Gamma)$; we write $\varphi = (\varphi_1, ..., \varphi_{p-1})$. For $\varphi \in H^{-1/2}(\Gamma)$ the restriction $\varphi|_{\Gamma_i} \in H^{-1/2}(\Gamma_i)$ can be defined by $\langle \varphi|_{\Gamma_i}, \psi \rangle_{\Gamma_i} := \langle \varphi, \bar{\psi} \rangle_\Gamma$ where $\bar{\psi} \in H_{00}^{1/2}(\Gamma)$ is the extension by zero of a $\psi \in H_{00}^{1/2}(\Gamma_i)$. For the norms,

$$\|\varphi\|^2_{H^{-1/2}(\Gamma)} = \sum_{i=1}^{p-1} \|\varphi|_{\Gamma_i}\|^2_{H^{-1/2}(\Gamma_i)}$$

holds.

By means of the canonical Riesz identification $L^2(\Gamma) = (L^2(\Gamma))^*$ we can identify $H_{00}^{1/2}(\Gamma)$ with a subspace of $H^{-1/2}(\Gamma)$:

$$H_{00}^{1/2}(\Gamma) \subset L^2(\Gamma) = (L^2(\Gamma))^* \subset H^{-1/2}(\Gamma) \tag{2.17}$$

The same holds for each Γ_i instead of Γ.

Now we will summarize some well known results. The following Trace Theorem is essential for the rest of this section:

Theorem 2.1 (Trace Theorem) *Assume that $\Gamma \subset \bar{\Omega}$ is a Lipschitz curve and ν a unit vector field normal to Γ. Then, the mapping*

$$u \longmapsto u|_\Gamma, \quad C^\infty(\bar{\Omega}) \longrightarrow C^\infty(\Gamma)$$

extends by continuity to a continuous linear mapping

$$u \longmapsto \gamma_0 u, \quad H^1(\Omega) \longrightarrow H^{1/2}(\Gamma). \tag{2.18}$$

This mapping is surjective and there exists a continuous linear right-inverse[4] *$P : H^{1/2}(\Gamma) \to H^1(\Omega)$.*

Proof. See Satz 8.7 and Satz 8.8 in [57].[5] ∎

Further we will make use of the following generalized Stokes formula:

Theorem 2.2 *Let Ω be a domain with Lipschitz boundary,*

$$H_{div}(\Omega) := \{\vec{u} \in (L^2(\Omega))^n \mid div\ \vec{u} \in L^2(\Omega)\}.$$

Then there is a continuous linear operator $g : H_{div}(\Omega) \to (H^{1/2}(\partial\Omega))^*$ with $g(\vec{u}) = \vec{u} \cdot \nu$ for all $\vec{u} \in (C^\infty(\bar{\Omega}))^n$. For $\vec{u} \in H_{div}(\Omega)$

$$\int_\Omega (\vec{u} \cdot \nabla w + w\ div\ \vec{u})\ dx = \langle g(\vec{u}), \gamma_0 w \rangle_{\partial\Omega} \tag{2.19}$$

holds for all $w \in H^1(\Omega)$.

Proof. See Theorem 1.2 and Remark 1.3 in [52], Ch. 1, §1, 1.3. ∎

Corollary 2.3 *Let Ω be a domain with Lipschitz boundary,*

$$H^1_\Delta(\Omega) := \{u \in H^1(\Omega) \mid \Delta u \in L^2(\Omega)\}.$$

Then there exists a continuous linear operator

$$\gamma_1 : H^1_\Delta(\Omega) \longrightarrow (H^{1/2}(\partial\Omega))^*$$

with $\gamma_1 u = \frac{\partial u}{\partial \nu}$ for all $u \in C^\infty(\bar{\Omega})$ and

$$\int_\Omega (\nabla u \cdot \nabla w + w\,\Delta u)\ dx = \langle \gamma_1 u, \gamma_0 w \rangle_{\partial\Omega} \tag{2.20}$$

holds for all $u \in H^1_\Delta(\Omega)$, $w \in H^1(\Omega)$.

[4] i.e. $\gamma_0 P = id_{H^{1/2}(\Gamma)}$

[5] For more regular $\partial\Omega$ we find another proof in [34], Chapter 1, Theorem 8.3.

Proof. Use Theorem 2.2 with $\vec{u}$ replaced by ∇u. ∎

Remark. As $H_{00}^{1/2}(\partial\Omega) \subset H^{1/2}(\partial\Omega)$, $(H^{1/2}(\partial\Omega))^* \subset H^{-1/2}(\partial\Omega)$.

Lemma 2.4 *Let Γ, Γ_0 be two Lipschitz curves, $\Gamma_0 \subset \Gamma$, $\Gamma_0 \neq \Gamma$, $\varphi : \Gamma \longrightarrow \mathbb{R}$ with $u|_{\Gamma\setminus\Gamma_0} = 0$.*

Then $u|_{\Gamma_0} \in H_{00}^{1/2}(\Gamma_0)$ if and only if $u \in H^{1/2}(\Gamma)$.[6] In this case,

$$c\,\|u|_{\Gamma_0}\|_{H_{00}^{1/2}(\Gamma_0)} \leq \|u\|_{H^{1/2}(\Gamma)} \leq c\,\|u|_{\Gamma_0}\|_{H_{00}^{1/2}(\Gamma_0)}. \tag{2.21}$$

Proof. The assertion follows directly from the proof of Theorem 11.4 in [34] Chapter 1 (p. 60-61) for $s = 1/2$.[7] Let us just point out that the validity of (11.34) in [34] is a consequence of the representation (2.12)-(2.13) of $H_{00}^{1/2}(\Gamma_0)$. ∎

Lemma 2.5 *Let B be a subspace of $H^1(\Omega)$ in which a Poincaré inequality*

$$\|u\|_{L^2(\Omega)} \leq c_P\,\|\nabla u\|_{L^2(\Omega)} \quad \forall\, u \in B$$

holds. For

$$\|u\|^2_{H^1_\sigma(\Omega)} := \|\nabla u\|^2_{L^2(\Omega)} + \sigma\|u\|^2_{L^2(\Omega)}, \quad \sigma \geq 0,$$

the estimate

$$\underline{c}\,\|u\|_{H^1(\Omega)} \leq \|u\|_{H^1_\sigma(\Omega)} \leq \overline{c}\,\|u\|_{H^1(\Omega)} \tag{2.22}$$

holds with

$$\underline{c} = \min\left\{1, \left(\frac{1+\sigma\,c_P^2}{1+c_P^2}\right)^{1/2}\right\}, \quad \overline{c} = \max\left\{1, \left(\frac{1+\sigma\,c_P^2}{1+c_P^2}\right)^{1/2}\right\} \tag{2.23}$$

Proof. For $\sigma \leq 1$, $\overline{c}=1$ matches (2.22), obviously. Let $\sigma > 1$. With the approach $c_1+c_2=1$, $0 \leq c_1, c_2 \leq 1$ and the Poincaré inequality we get

$$\|u\|^2_{H^1_\sigma(\Omega)} \leq (1 + c_1\,\sigma\,c_P^2)\,\|\nabla u\|^2_{L^2(\Omega)} + c_2\,\sigma\,\|u\|^2_{L^2(\Omega)}.$$

The best choice of c_1, c_2 is to postulate $1 + c_1\,\sigma\,c_P^2 = c_2\,\sigma$. Then we get

$$\|u\|^2_{H^1_\sigma(\Omega)} \leq \frac{1+\sigma\,c_P^2}{1+c_P^2}\,\|u\|^2_{H^1(\Omega)}$$

and the right part of inequality (2.23) follows.

[6] Replacing $H_{00}^{1/2}(\Gamma_0)$ by $H^{1/2}(\Gamma_0)$ this statement would become false (see Theorem 11.4 in [34] Chapter 1).

[7] That proof makes use of the representation of the $H^{1/2}$-norm by double integrals as in (3.165).

For $\sigma \geq 1$, $\underline{c} = 1$ matches (2.22). Now let $\sigma < 1$. With the same approach as above and $\|\nabla u\|_{L^2(\Omega)} \geq \frac{1}{c_P} \|u\|_{L^2(\Omega)}$ we get

$$\|u\|^2_{H^1_\sigma(\Omega)} \geq c_1 \|\nabla u\|^2_{L^2(\Omega)} + \left(\frac{c_2}{c_P^2} + \sigma \right) \|u\|^2_{L^2(\Omega)}.$$

Selecting c_1, c_2 with $c_1 = \frac{c_2}{c_P^2} + \sigma$ we find

$$\|u\|^2_{H^1_\sigma(\Omega)} \geq \frac{1 + \sigma\, c_P^2}{1 + c_P^2} \|u\|^2_{H^1(\Omega)}$$

and the left part of inequality (2.23) follows. ∎

2.2 The minimization principle for the Dirichlet problem

Now let us consider the Dirichlet case ($\Gamma^{Dir} = \partial\Omega,\ \Gamma^{Nm} = \emptyset$) of the resolvent equation (2.1)-(2.2) at first:

$$\begin{aligned} Lw = -\Delta w + \sigma\, w &= f \quad \text{on } \Omega, \quad \sigma \geq 0, \\ w &= g^{Dir} \quad \text{on } \partial\Omega. \end{aligned} \tag{2.24}$$

To guarantee the existence of the solution of (2.24), we consider the corresponding *weak* problem

$$\begin{aligned} &\int_\Omega \nabla w \cdot \nabla \omega + \sigma\, w\, \omega \; dx = \langle f, \omega \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} \quad \forall\, \omega \in H_0^1(\Omega), \\ &w \in \{ w \in H^1(\Omega) \mid \gamma_0 \omega|_{\partial\Omega} = g^{Dir} \} \end{aligned} \tag{2.25}$$

rather than (2.24). In (2.25), we assume that $f \in H^{-1}(\Omega)$, $g^{Dir} \in H^{1/2}(\partial\Omega)$.

Now we can give a derivation of the CGBI method as it was published in [5] and [7]:

The main idea is to represent the solution w of the global problem (2.24) as the sum

$$w = v_i + u_i, \quad i = 1, ..., p, \tag{2.26}$$

on each Ω_i, where v_i is the solution of the following pre-step:

$$\begin{aligned} Lv_i &= f \text{ in } \Omega_i \\ v_i &= g^{Dir} \text{ on } \partial\Omega_i \cap \partial\Omega \\ \frac{\partial v_i}{\partial \nu_i} &= 0 \text{ on } \partial\Omega_i \setminus \partial\Omega \end{aligned} \tag{2.27}$$

ν_i is the outward normal direction on $\partial\Omega_i$.

Then, obviously u_i has to fulfil $Lu_i = 0$ on Ω_i. Denoting by φ_i the unknown normal derivative of u_i on Γ_i and requiring its continuity we get the following problem for determining u_i:

$$\begin{aligned} Lu_i &= 0 \text{ in } \Omega_i \\ u_i &= 0 \text{ on } \partial\Omega_i \cap \partial\Omega \\ \frac{\partial u_i}{\partial \nu_i} &= -\varphi_{i-1} \text{ on } \Gamma_{i-1} \text{ if } i > 0 \\ \frac{\partial u_i}{\partial \nu_i} &= \varphi_i \text{ on } \Gamma_i \text{ if } i < p. \end{aligned} \tag{2.28}$$

with the unknown boundary value function $\varphi = (\varphi_1, ..., \varphi_{p-1})$. By means of (2.28) we may regard $u = (u_1, ..., u_p) = u(\varphi)$ as a function of φ. We have to find the

φ such that $v+u(\varphi)$ is in $H^1(\Omega)$, i.e. that the jumps of $v+u(\varphi)$ vanish at the interfaces.

Strictly speaking, we are of course solving the corresponding *weak* problems

$$\int_{\Omega_i} \nabla v_i \cdot \nabla \omega + \sigma\, v_i\, \omega \; dx = \langle f, \omega \rangle_{H^{-1}(\Omega_i), H^1(\Omega_i)} \qquad \forall\, \omega \in H^1_{Dir,Nm}(\Omega_i) := \{\omega \in H^1(\Omega_i) \mid \gamma_0 \omega|_{\partial\Omega_i \cap \partial\Omega} = 0\},$$
$$v_i \in \{u \in H^1(\Omega_i) \mid \gamma_0 u|_{\partial\Omega_i \cap \partial\Omega} = g^{Dir}\} \tag{2.29}$$

instead of (2.27) and

$$\int_{\Omega_i} \nabla u_i \cdot \nabla \omega + \sigma\, u_i\, \omega \; dx = \langle \varphi_i, \gamma_0 \omega \rangle_{\Gamma_i} - \langle \varphi_{i-1}, \gamma_0 \omega \rangle_{\Gamma_{i-1}} \qquad \forall\, \omega \in H^1_{Dir,Nm}(\Omega_i)$$
$$u_i \;\in\; H^1_{Dir,Nm}(\Omega_i) \tag{2.30}$$

instead of (2.28).

The method is now to determine φ on the interfaces such that the jump of $v+u = v+u(\varphi)$ on the interfaces is continuous in the sense of traces:

$$[v + u(\varphi)] = 0 \quad \text{on } \Gamma \tag{2.31}$$

where we have put

$$[w] := \gamma_0 w_i - \gamma_0 w_{i+1} \quad \text{on } \Gamma_i,\ i = 1, ..., p-1.$$

The idea of CGBI is to find this φ by minimizing the quadratic functional

$$J(\varphi) := \sum_{i=1}^{p-1} \left\langle \varphi_i, \gamma_0 v_i - \gamma_0 v_{i+1} + \frac{1}{2} \left(\gamma_0 u_i(\varphi) - \gamma_0 u_{i+1}(\varphi)\right) \right\rangle_{\Gamma_i} \longrightarrow \text{Min.}, \tag{2.32}$$

with u_i depending linearly on φ via (2.30). The equivalence of problem (2.31) to the minimization problem (2.32) (both with v, $u(\varphi)$ defined by (2.29), (2.30)) is demonstrated in Theorem 2.6. Using the more compact notation, (2.32) becomes

$$J(\varphi) = \langle \varphi, [v + \frac{1}{2}\, u(\varphi)] \rangle_\Gamma. \tag{2.33}$$

Introducing the associated bilinear form $b : H^{-1/2}(\Gamma) \times H^{-1/2}(\Gamma) \to \mathbb{R}$

$$b(\psi, \varphi) := \langle \psi, [u(\varphi)] \rangle_\Gamma \tag{2.34}$$

and the linear form $l : H^{-1/2}(\Gamma) \to \mathbb{R}$

$$l(\varphi) := \langle \varphi, [v] \rangle_\Gamma$$

(2.32) becomes

$$J(\varphi) = \frac{1}{2}\, b(\varphi, \varphi) + l(\varphi) \longrightarrow \text{Min. in } H^{-1/2}(\Gamma) \tag{2.35}$$

Theorem 2.6 *The bilinear form b has the following properties:*

a) b is symmetric.

b) b is bounded in the sense

$$b(\varphi,\varphi) \le c \, \|\varphi\|^2_{H^{-1/2}(\Gamma)} \quad \forall \varphi \in H^{-1/2}(\Gamma). \tag{2.36}$$

c) b is coercive:

$$b(\varphi,\varphi) \ge c \, \|\varphi\|^2_{H^{-1/2}(\Gamma)} \quad \forall \varphi \in H^{-1/2}(\Gamma).$$

d) The problem (2.35) has a unique solution $\varphi^0 \in H^{-1/2}(\Gamma)$ which coincides with the boundary value φ in (2.29)-(2.31).

The ratio of the constants in b) and c) only depends on the shape of the subdomains Ω_i, but not on the number p of subdomains or the length of the channel or σ.

Remark. These properties and their meaning for the construction of a preconditioner (see next remark) were stated by Borchers [5], but a proof was not given in that paper.

Proof. For $\varphi \in H^{-1/2}(\Gamma)$, the Trace Theorem 2.1 and Lemma 2.4 guarantee that $u_i(\varphi)|_{\Gamma_i}, u_{i+1}(\varphi)|_{\Gamma_i} \in H^{1/2}_{00}(\Gamma_i)$, thus $[u(\varphi)] \in H^{1/2}_{00}(\Gamma)$. That means that b is well defined.
ad a). From the definition of $u(\psi)$ in (2.28) we get

$$\left.\frac{\partial u_i(\psi)}{\partial \nu_i}\right|_{\Gamma_i} = - \left.\frac{\partial u_{i+1}(\psi)}{\partial \nu_{i+1}}\right|_{\Gamma_i}, \quad \text{i.e.} \quad \gamma_1 u_i(\psi) = -\gamma_1 u_{i+1}(\psi) \ \text{ on } \Gamma_i \tag{2.37}$$

for $i = 1, \ldots, p-1$, $\psi \in H^{-1/2}(\Gamma)$. With help of (2.37) we conclude

$$\begin{aligned}
b(\psi,\varphi) &= \sum_{i=1}^{p-1} \langle \psi_i, \gamma_0 u_i(\varphi) - \gamma_0 u_{i+1}(\varphi) \rangle_{\Gamma_i} \\
&= \sum_{i=1}^{p-1} \langle \gamma_1 u_i(\psi), \gamma_0 u_i(\varphi) - \gamma_0 u_{i+1}(\varphi) \rangle_{\Gamma_i} \\
&= \sum_{i=1}^{p-1} \langle \gamma_1 u_i(\psi), \gamma_0 u_i(\varphi) \rangle_{\Gamma_i} + \langle \gamma_1 u_{i+1}(\psi), \gamma_0 u_{i+1}(\varphi) \rangle_{\Gamma_i} \\
&= \sum_{i=1}^{p-1} \langle \gamma_1 u_i(\psi), \gamma_0 u_i(\varphi) \rangle_{\Gamma_i} + \sum_{i=2}^{p} \langle \gamma_1 u_i(\psi), \gamma_0 u_i(\varphi) \rangle_{\Gamma_{i-1}} \\
&= \sum_{i=1}^{p} \langle \gamma_1 u_i(\psi), \gamma_0 u_i(\varphi) \rangle_{\partial\Omega_i}
\end{aligned} \tag{2.38}$$

where we have used the homogenity of the boundary conditions for $u(\varphi)$ on $\partial\Omega$ in (2.28) in the last step.

As $\Delta u_i(\varphi) = \sigma\, u_i(\varphi) \in L^2(\Omega_i)$, we can apply the Stokes formula (2.20) and get

$$\begin{aligned} b(\psi,\varphi) &= \sum_{i=1}^{p} \int_{\Omega_i} (u_i(\varphi)\ \Delta u_i(\psi) + \nabla u_i(\varphi)\cdot\nabla u_i(\psi))\ dx \\ &= \sum_{i=1}^{p} \int_{\Omega_i} (\sigma\ u_i(\varphi)\ u_i(\psi) + \nabla u_i(\varphi)\cdot\nabla u_i(\psi))\ dx \end{aligned} \tag{2.39}$$

ad b). Let us apply Lemma 2.5 to all subdomains $\Omega_1,...,\Omega_p$ and let $\underline{C}$ be the minimum off all the constants $\underline{c}$. Let c_T be the maximum of the norms of all trace operators $H^1(\Omega_i) \to H^{1/2}(\partial\Omega_i)$, $i=1,...,p$. Let us apply Lemma 2.4 to all curves $\Gamma_i \subset \partial\Omega_i$, $\Gamma_i \subset \partial\Omega_{i+1}$ and let c_F be the reciprocal of the minimum of the constants c on the left hand side of (2.21). Using Lemma 2.5, the representations (2.39), (2.38), Lemma 2.4, the Trace Theorem 2.1 and the Cauchy-Schwarz inequality we get

$$\underline{C}^2 \sum_{i=1}^{p} \|u_i(\varphi)\|^2_{H^1(\Omega_i)} \le b(\varphi,\varphi) = \sum_{i=1}^{p-1} \langle \varphi_i,\, \gamma_0 u_i(\varphi) - \gamma_0 u_{i+1}(\varphi)\rangle_{\Gamma_i} \tag{2.40}$$

$$\begin{aligned} &\le \sum_{i=1}^{p-1} \|\varphi_i\|_{H^{-1/2}(\Gamma_i)}\ (\|\gamma_0 u_i(\varphi)\|_{H^{1/2}_{00}(\Gamma_i)} + \|\gamma_0 u_{i+1}(\varphi)\|_{H^{1/2}_{00}(\Gamma_i)}) \\ &\le c_F \sum_{i=1}^{p-1} \|\varphi_i\|_{H^{-1/2}(\Gamma_i)}\ \left(\|\gamma_0 u_i(\varphi)\|_{H^{1/2}(\partial\Omega_i)} + \|\gamma_0 u_{i+1}(\varphi)\|_{H^{1/2}(\partial\Omega_{i+1})}\right) \\ &\le c_F\, c_T \sum_{i=1}^{p-1} \|\varphi_i\|_{H^{-1/2}(\Gamma_i)}\ \left(\|u_i(\varphi)\|_{H^1(\Omega_i)} + \|u_{i+1}(\varphi)\|_{H^1(\Omega_{i+1})}\right) \\ &\le 2\, c_F\, c_T\ \|\varphi\|_{H^{-1/2}(\Gamma)}\ \left(\sum_{i=1}^{p} \|u_i(\varphi)\|^2_{H^1(\Omega_i)}\right)^{1/2} \end{aligned}$$

Thus,

$$\left(\sum_{i=1}^{p} \|u_i(\varphi)\|^2_{H^1(\Omega_i)}\right)^{1/2} \le \frac{2\, c_F\, c_T}{\underline{C}^2}\ \|\varphi\|_{H^{-1/2}(\Gamma)}$$

and

$$b(\varphi,\varphi) \le \frac{4\, c_F^2\, c_T^2}{\underline{C}^2}\ \|\varphi\|^2_{H^{-1/2}(\Gamma)}.$$

$\underline{C}$, c_T, c_F are defined as maximums/minimums of constants which only depend on the *local* geometry Ω_i, Γ_i. c_T and c_F are independent of σ, and $\underline{c}$ in (2.23) is bounded independently of $\sigma \in [0,\infty)$. Hence, the constant in (2.36) is *independent* of p, σ and the length of the channel Ω.

ad c). Let us apply Lemma 2.5 to all subdomains Ω_1,...,Ω_p and let $\overline{C}$ be the maximum off all the constants $\bar{c}$. It holds

$$\|\varphi\|^2_{H^{-1/2}(\Gamma)} = \sum_{i=1}^{p-1} \|\varphi_i\|^2_{H^{-1/2}(\Gamma_i)} = \left(\sum_{i=1}^{p-1} \sup_{\psi_i \in H^{1/2}_{00}(\Gamma_i)} \frac{\langle \varphi_i, \psi_i \rangle_{\Gamma_i}}{\|\psi_i\|_{H^{1/2}_{00}(\Gamma_i)}} \right)^2 . \qquad (2.41)$$

Every $\psi_i \in H^{1/2}_{00}(\Gamma_i)$ can be extended to a $\bar{\psi}_i \in H^{1/2}(\partial\Omega_i)$ by $\bar{\psi}_i|_{\partial\Omega_i \setminus \Gamma_i} :\equiv 0$; $\|\bar{\psi}_i\|_{H^{1/2}(\partial\Omega_i)} \le C \|\psi_i\|_{H^{1/2}_{00}(\Gamma_i)}$ holds (Lemma 2.4). Hence,

$$\begin{aligned} \langle \varphi_i, \psi_i \rangle_{\Gamma_i} &= \langle \gamma_1 u_i(\varphi), \psi_i \rangle_{\Gamma_i} = \langle \gamma_1 u_i(\varphi), \bar{\psi}_i \rangle_{\partial\Omega_i} \\ &= \langle \gamma_1 u_i(\varphi), \gamma_0 P_i \bar{\psi}_i \rangle_{\partial\Omega_i} = \int_{\Omega_i} \sigma\, u_i(\varphi)\, P_i \bar{\psi}_i + \nabla u_i(\varphi) \cdot \nabla P_i \bar{\psi}_i \, dx \\ &\le \bar{C} \|u_i(\varphi)\|_{H^1_\sigma(\Omega_i)} \|P_i \bar{\psi}_i\|_{H^1(\Omega_i)} \end{aligned}$$

where we have used Theorem 2.1, (2.38)/(2.39) and Lemma 2.5. Furthermore,

$$\|P_i \bar{\psi}_i\|_{H^1(\Omega_i)} \le c_{P_i} C \|\psi_i\|_{H^{1/2}_{00}(\Gamma_i)},$$

where c_{P_i} is the norm of the prolongation operator $P_i : H^{1/2}(\partial\Omega_i) \to H^1(\Omega_i)$ (Theorem 2.1). Hence, in (2.41),

$$\|\varphi\|^2_{H^{-1/2}(\Gamma)} \le \bar{C}^2 c_P^2 C^2 \sum_{i=1}^{p-1} \|u_i(\varphi)\|^2_{H^1_\sigma(\Omega_i)} = \bar{C}^2 c_P^2 C^2 b(\varphi, \varphi),$$

where c_P is the maximum of the c_{P_i}.

ad d). The resolvent problem (2.25) has a unique solution $w \in H^1(\Omega)$. With help of Corollary 2.3, $\varphi^0 := (\frac{\partial w}{\partial \nu_1}|_{\Gamma_1}, ..., \frac{\partial w}{\partial \nu_{p-1}}|_{\Gamma_{p-1}})$ lies in $H^{-1/2}(\Gamma)$. Then $\varphi^0 \in H^{-1/2}(\Gamma)$ is a solution of (2.29)-(2.31), i.e. $v + u(\varphi^0) = w$. Using $[v] = -[u(\varphi^0)]$ and the symmetry of b we get for arbitrary $\varphi \in H^{-1/2}(\Gamma)$

$$\begin{aligned} J(\varphi) - J(\varphi^0) &= \langle \varphi - \varphi^0, [v] \rangle_\Gamma + \frac{1}{2} \langle \varphi, [u(\varphi)] \rangle_\Gamma - \frac{1}{2} \langle \varphi^0, [u(\varphi^0)] \rangle_\Gamma \\ &= \frac{1}{2} \langle \varphi^0, [u(\varphi^0)] \rangle_\Gamma - \langle \varphi, [u(\varphi^0)] \rangle_\Gamma + \frac{1}{2} \langle \varphi, [u(\varphi)] \rangle_\Gamma \\ &= b(\varphi - \varphi^0, \varphi - \varphi^0) \end{aligned}$$

which is zero for $\varphi = \varphi^0$ and, due to c), strictly positive otherwise. Thus, J has a unique minimum at $\varphi = \varphi^0$. ∎

Corollary 2.7 *For each isomorphism $C : H^{1/2}_{00}(\Gamma) \to H^{-1/2}(\Gamma)$ that generates a positive definite bilinear form $(\xi, \eta) \mapsto \langle C\xi, \eta \rangle_\Gamma$ on $H^{1/2}_{00}(\Gamma) \times H^{1/2}_{00}(\Gamma)$ there are constants $c_1, c_2 > 0$ such that*

$$c_1 \langle \varphi, C^{-1}\varphi \rangle_\Gamma \le b(\varphi, \varphi) \le c_2 \langle \varphi, C^{-1}\varphi \rangle_\Gamma \qquad (2.42)$$

holds for all $\varphi \in H^{-1/2}(\Gamma)$.

Proof. Using at first Theorem 2.6 and then the continuity of C and C^{-1}, we get the following chain of equivalences on $H^{-1/2}(\Gamma)$:

$$b(\varphi,\varphi) \sim \|\varphi\|^2_{H^{-1/2}(\Gamma)} \sim \|C^{-1}\varphi\|^2_{H^{1/2}_{00}(\Gamma)} \tag{2.43}$$

The positivity

$$\langle C\xi,\xi\rangle_\Gamma \geq c\,\|\xi\|^2_{H^{1/2}_{00}(\Gamma)}$$

and the continuity

$$\langle C\xi,\xi\rangle_\Gamma \leq \|C\xi\|_{H^{-1/2}(\Gamma)}\,\|\xi\|_{H^{1/2}_{00}(\Gamma)} \leq c\,\|\xi\|^2_{H^{1/2}_{00}(\Gamma)},$$

both applied to $\xi := C^{-1}\varphi$ yield the equivalence

$$\langle\varphi, C^{-1}\varphi\rangle_\Gamma \sim \|C^{-1}\varphi\|^2_{H^{1/2}_{00}(\Gamma)}. \tag{2.44}$$

■

Remarks on Corollary 2.7 and on preconditioning. We are going to find the solution of the minimization principle (2.35) by the conjugate gradient method. Therefore we are interested in a mapping $C : H^{1/2}_{00}(\Gamma) \to H^{-1/2}(\Gamma)$ for which (2.42) holds because these mappings (rsp. its discretizations) are suitable preconditioners: They are expected to cause a condition number κ *independent* of the discretization parameter.

Obviously, the *Riesz isometry* from $H^{1/2}_{00}(\Gamma)$ to $H^{-1/2}(\Gamma)$ which is defined by

$$\langle C\xi,\eta\rangle_\Gamma = (\xi,\eta)_{H^{1/2}_{00}(\Gamma)} \quad \forall\,\xi,\eta \in H^{1/2}_{00}(\Gamma)$$

meets the requirements in Corollary 2.7. (In the proof, last equivalence in (2.43) and equivalence (2.44) can be replaced by '=', if C is the Riesz isometry.) Furthermore, if we replace the norm in $H^{1/2}_{00}(\Gamma)$ by an equivalent norm (this induces a new norm on $H^{-1/2}(\Gamma)$ and a new Riesz isometry), the assertion also holds for the new Riesz isometry. Let us focus on a special ('canonical') Riesz isometry:

$(-\Delta_0)^{1/2} : H^1_0(\Gamma) \to L^2(\Gamma)$ is a Riesz isometry between the spaces $H^1_0(\Gamma)$ with the norm (2.5) and $L^2(\Gamma)$. By continuity, this mapping can be extended to a Riesz isometry

$$(-\Delta_0)^{1/2} : H^{1/2}_{00}(\Gamma) \to H^{-1/2}(\Gamma). \tag{2.45}$$

Locally on each interface we get

$$(-\Delta_0)^{1/2} : H^{1/2}_{00}(\Gamma_i) \to H^{-1/2}(\Gamma_i).$$

If we identify Γ_i with the interval $(0,\pi)$, this operator has obviously the eigenfunctions $\sin kx$, $k \in \mathbb{N}$. $(-\Delta_{Nm})^{1/2}$ on $H^{1/2}_{mv}(\Gamma_i)$ has the eigenfunctions $\cos kx$,

$k \in \mathbb{N}$. The accompanying eigenvalues are equal to k both for $(-\Delta_0)^{1/2}$ and $(-\Delta_{Nm})^{1/2}$ and $\sqrt{1+k^2}$ for $(id-\Delta_{Nm})^{1/2}$.

A natural way for the choice of C is this Riesz isometry (2.45). For this choice of C, the proofs of Theorem 2.6 b), c) and of Corollary 2.7 imply estimate (2.42) with constants independent of p and the channel length. So the condition number κ can be expected to be independent of the discretization parameter, the number p of subdomains and the length of the channel, but it may depend on the shape of the subdomains. Of course κ will also depend on the accuracy of the discrete realization of C.

All these effects can be observed in Chapter 3 were preconditioning operators C and its discretizations are developed and tested.

Another, less sophisticated choice of the transition operator $C : H_{00}^{1/2}(\Gamma) \to H^{-1/2}(\Gamma)$ would be the Riesz isometry between $L^2(\Gamma)$ and its dual $(L^2(\Gamma))^*$, restricted to $H_{00}^{1/2}(\Gamma)$. This operator violates the spectral equivalence of the previous corollary. In the discretized version of CGBI, this transition operator would be represented by the identity matrix on a finite dimensional space of discrete solutions φ^h.

Both transition operators are subjected to numerical tests and comparisons in Chapter 3.

2.3 The algorithm for the Dirichlet problem

The method described in Section 2.2 leads to the following algorithm. Step 1 corresponds to the pre-step (2.27), in step 2 the boundary condition φ which fulfils the minimization principle (2.32) is calculated, and in step 3 the corresponding $w_i = v_i + u_i(\varphi)$ is calculated.

For the sake of clarity, the obvious modifications in the boundary conditions for the first and the last subdomain are omitted.

Procedure CGBI(f):
(Done in a parallel mode by processors $i = 1, ..., p$.)

step 1: Solve (in a parallel mode)

$$\begin{aligned} &Lv = f \text{ on } \Omega_i, \\ &\tfrac{\partial v}{\partial \nu_i} = 0 \text{ on } \Gamma_i,\ \Gamma_{i-1} \\ &v = g^{Dir} \text{ on } \partial\Omega_i \setminus (\Gamma_i \cup \Gamma_{i-1}) \\ &\text{with FEM, FDM or Chebyshev method.} \end{aligned} \tag{2.46}$$

Let $g := [v]$ be the jump of v on the artificial boundaries; g is defined on $\bigcup_{i=1}^{p-1} \Gamma_i$.

step 2: $\varphi := CG(g)$; *(see below)*
φ is defined on $\bigcup_{i=1}^{p-1} \Gamma_i$.

step 3: Solve (in parallel)

$$\begin{aligned} &Lu = 0 \text{ on } \Omega_i, \\ &\tfrac{\partial u}{\partial \nu_i} = +\varphi \text{ on } \Gamma_i,\ \tfrac{\partial u}{\partial \nu_i} = -\varphi \text{ on } \Gamma_{i-1} \\ &u = 0 \text{ on } \partial\Omega_i \setminus (\Gamma_i \cup \Gamma_{i-1}) \\ &\text{with FEM, FDM or Chebyshev method.} \end{aligned} \tag{2.47}$$

Return $w := u + v$ which is the solution of
$Lw = f$ on Ω, $w = g^{Dir}$ on $\partial\Omega$.

Procedure $CG(g_i)$**:**
(Done in parallel; each processor $i = 1...p$ calculates on one artificial boundary Γ_i and on one subdomain Ω_i.)
$\varphi_i := 0$
$d_i := -Cg_i$
$\delta_0 := -\sum_{i=1}^{p-1} \langle g_i, d_i \rangle \quad (\rightarrow communication)$
get d_{i-1} from process $i-1 \quad (\rightarrow communication)$

repeat
{ solve

$$\begin{aligned} &Lu = 0 \text{ on } \Omega_i, \\ &\tfrac{\partial u}{\partial \nu_i} = +d_i \text{ on } \Gamma_i, \quad \tfrac{\partial u}{\partial \nu_i} = -d_{i-1} \text{ on } \Gamma_{i-1} \\ &u = 0 \text{ on } \partial\Omega_i \setminus (\Gamma_i \cup \Gamma_{i-1}) \\ &\text{with FEM, FDM or Chebyshev method} \end{aligned} \tag{2.48}$$

$\gamma_i := u_{i+1} - u_i$ on Γ_i ('jump') $\quad (\rightarrow communication)$
$\tau := \frac{\delta_0}{\sum_{i=1}^{p-1} \langle d_i, \gamma_i \rangle} \quad (\rightarrow communication)$
$\varphi_i := \varphi_i + \tau d_i$
$g_i := g_i + \tau \gamma_i, \quad h_i := Cg_i$
$\delta_1 := \sum_{i=1}^{p-1} \langle g_i, h_i \rangle \quad (\rightarrow communication), \quad \beta := \frac{\delta_1}{\delta_0}, \quad \delta_0 := \delta_1$
$d_i := -h_i + \beta d_i$
get d_{i-1} from process $i-1 \quad (\rightarrow communication)$
} until $\delta_1 < \epsilon$
return φ_i

The above algorithm is given in a non-discretized version. It computes the boundary condition $\varphi \in H^{-1/2}(\Gamma)$ and the solution $w \in H^1(\Omega)$ of (2.1). For details on the discretization of this algorithm see Sections 2.6, 2.7. The local solvers in (2.46), (2.47), (2.48) are described in Section 2.5.

For more details on the transition operator C in the procedure CG see Chapter 3 and the end of Section 2.2.

2.4 The Neumann problem

As long as $\sigma > 0$, the CGBI algorithm of Section 2.3 can be applied to the Neumann case $\Gamma^{Nm} \neq \emptyset$ (see (2.2)) as well. Just the boundary condition (2.2) has to be taken into account on the outer boundaries $\Gamma^{Nm} \cap \partial\Omega_i$ for the local problems (2.27) and the corresponding *homogeneous* boundary condition on $\Gamma^{Nm} \cap \partial\Omega_i$ for the local problems (2.28). Concerning theory (Sec. 2.2), the spaces $H_{00}^{1/2}(\Gamma)$ and $H^{-1/2}(\Gamma)$ have to be replaces by new ones if the interfaces hit Γ^{Nm}.

If $\sigma = 0$, then a more important difficulty arises: The well-posedness of the local problems gets lost if $\Gamma^{Dir} \cap \partial\Omega_i = \emptyset$, even if the global problem (2.1) *is* well-posed (e.g. $\Gamma^{Dir} = \Gamma^O$). These subdomains where the local problem has no Dirichlet boundary are frequently called *floating subdomains*. Having in mind the pressure problem of the Navier-Stokes solver (Chapter 5), we will focus on the case $\sigma = 0$, $\Gamma^{Dir} = \Gamma^O$. For this case, we will discuss the modification for the CGBI algorithm and the related function spaces for the boundary functions φ and the jumps $[u(\varphi)]$.

2.4.1 General remarks on the Neumann problem

In this section some well known properties of the Poisson problem with pure Neumann boundary conditions are recapitulated. We regard the following three formulations:

(i) **Strong formulation.** Find a $H^2(\Omega)$-function u such that

$$-\Delta u = f \text{ on } \Omega, \quad \frac{\partial u}{\partial \nu} = g \text{ on } \partial\Omega. \tag{2.49}$$

To assure the uniqueness of a solution, we may additionaly postulate e.g.

$$u \in H^1_{mv}(\Omega) \tag{2.50}$$

(see def. (2.4)).

(ii) **A variational formulation.** For $f \in H^{-1}(\Omega)$, $g \in H^{-1/2}(\partial\Omega)$, find $u \in H^1_{mv}(\Omega)$ such that

$$\int_\Omega \nabla u \cdot \nabla w = \langle f, w \rangle_{H^{-1}(\Omega), H^1(\Omega)} + \langle g, \gamma_0 w \rangle_{H^{-1/2}(\partial\Omega), H^{1/2}(\partial\Omega)} \quad \forall\, w \in H^1(\Omega).$$

(iii) **Another variational formulation.** For $f \in H^{-1}(\Omega)$, $g \in H^{-1/2}(\partial\Omega)$, find $u \in H^1_{mv}(\Omega)$ such that

$$\int_\Omega \nabla u \cdot \nabla w = \langle f, w \rangle_{H^{-1}(\Omega), H^1(\Omega)} + \langle g, \gamma_0 w \rangle_{H^{-1/2}(\partial\Omega), H^{1/2}(\partial\Omega)} \quad \forall\, w \in H^1_{mv}(\Omega). \tag{2.51}$$

The Divergence Theorem shows that (i) can only have a solution if the compatibility condition

$$\int_\Omega f\ dx + \int_{\partial\Omega} g\ do = 0 \tag{2.52}$$

holds. On the other hand, the Lax-Milgram theorem implies[8] the existence of a unique solution of problem (iii) *without any compatibility condition on f and g*. By setting $w :\equiv 1$ we see that in (ii),

$$\langle f, 1\rangle_{H^{-1}(\Omega),H^1(\Omega)} + \langle g, 1\rangle_{H^{-1/2}(\partial\Omega),H^{1/2}(\partial\Omega)} = 0 \tag{2.53}$$

(which is, in fact, a 'weak' formulation of (2.52)) is a necessary compatibility condition for the existence of a solution. So (iii) is the only formulation among (i)-(iii) without any compatibility condition for f and g.

Obviously, the solution u of (iii) does not change if a constant is added to f. (In the spirit of (2.17), $\langle f + c, w\rangle = \langle f, w\rangle + c\int_\Omega w\ dx$.) So we may substitute f in (iii) by $\tilde f = f + c$ such that (2.53) holds for $\tilde f$. Then, (ii) and (iii) have the same solution. Assuming that this solution has $H^2(\Omega)$-regularity, it also solves (i).

It should be emphasized that *arbitrary* $H^2(\Omega)$-solutions of (iii) (i.e. without validity of (2.53)) do not meet (i), but only

$$-\Delta u = f - c \text{ on } \Omega, \quad \frac{\partial u}{\partial \nu} = g \text{ on } \partial\Omega \tag{2.54}$$

For a constant $c \in \mathbb{R}$.[9]

2.4.2 Getting rid of ill-posed local problems

Let us assume the conditions made in the beginning of Section 2.4: $\sigma = 0$, $\Gamma^{Dir} = \Gamma^O$. Comparing Section 2.4.1, especially (2.53)/(2.52), with the problems (2.27), (2.28) where the Dirichlet conditions on $\partial\Omega$ are just replaced by Neumann conditions, we see that these problems would be ill-posed. We apply the following modifications to ensure solvability and uniqueness:

Solvability. For both the pre-step and the main step, there seem to be two different ways to ensure the solvability: To modify the source term of the partial differential equations or to modify the boundary conditions on the interfaces. As we want the problem of the main step to be homogeneous, we leave the source terms unchanged. So we modify the boundary conditions.

[8] Here we are using the Poincaré inequality on $H^1_{mv}(\Omega)$ (Lemma 2.9).

[9] To prove (2.54), one decomposes arbitrary $w \in H^1(\Omega)$ into $w = w_0 + c_w$, $c_w := |\Omega|^{-1}\int_\Omega w\,dx$, $w_0 \in H^1_{mv}(\Omega)$. With help of (2.51), $\int_\Omega \nabla u \cdot \nabla w\,dx = \int_\Omega \nabla u \cdot \nabla w_0\,dx = \int_\Omega f\,w_0\,dx + \int_{\partial\Omega} g\,\gamma_0 w_0\,do = \int_\Omega f(w - c_w)\,dx + \int_{\partial\Omega} g(\gamma_0 w - c_w)\,do = \int_\Omega (f - |\Omega|^{-1}\int_\Omega f - |\Omega|^{-1}\int_{\partial\Omega} g)\,w\,dx + \int_{\partial\Omega} g\,\gamma_0 w\,do$ from which (2.54) follows. Let us remark that our spectral solver (Sec. 2.5) happens to solve (2.54) if confronted with an incompatible problem (2.49).

Let us consider the main step at first. From the compatibility condition (2.53), the homogenity of the partial differential equation and the homogenity of the boundary condition on $\partial\Omega$ we get the condition

$$\left\langle \frac{\partial u_i(\varphi)}{\partial \nu_i}, 1 \right\rangle_{\partial\Omega_i \setminus \partial\Omega} = 0, \quad i = 1, ..., p-1. \tag{2.55}$$

Successive application for $i=1, ..., p-1$ leads to

$$\langle \varphi_i, 1 \rangle_{\Gamma_i} = 0 \quad \forall\, i = 1, ..., p-1.$$

That means that we are searching the boundary condition φ in the space

$$H_{mv}^{-1/2}(\Gamma) := \{\varphi \in (H^{1/2}(\Gamma))^* \mid \langle \varphi_i, 1 \rangle_{\Gamma_i} = 0 \; \forall\, i=1, ..., p-1\}$$

(instead of $H^{-1/2}(\Gamma)$).

Now to the pre-step. Here, (2.53) leads to the conditions

$$\begin{aligned} \langle f_1, 1 \rangle_{H^{-1}(\Omega_1), H^1(\Omega_1)} + \langle g^{Nm}, 1 \rangle_{\Gamma^W \cap \partial\Omega_1} + \langle \varphi_1, 1 \rangle_{\Gamma_1} \quad &= 0 \\ \langle f_i, 1 \rangle_{H^{-1}(\Omega_i), H^1(\Omega_i)} + \langle g^{Nm}, 1 \rangle_{\Gamma^W \cap \partial\Omega_i} + \langle \varphi_i, 1 \rangle_{\Gamma_i} - \langle \varphi_{i-1}, 1 \rangle_{\Gamma_{i-1}} &= 0 \; \forall\, i=2, ..., p-1. \end{aligned} \tag{2.56}$$

Let us notice that there is no compatibility condition on Ω_p due to the Dirichlet boundary condition on Γ^O. The easiest way to fulfil (2.56) is to define each φ_i, $i = 1, ..., p-1$, as a constant c_i on Γ_i. This leads to the following recurrency equation for the constants c_i:

$$\begin{aligned} \langle f_1, 1 \rangle_{H^{-1}(\Omega_1), H^1(\Omega_1)} + \langle g^{Nm}, 1 \rangle_{\Gamma^W \cap \partial\Omega_1} + c_1 \, |\Gamma_1| \quad &= 0, \\ \langle f_i, 1 \rangle_{H^{-1}(\Omega_i), H^1(\Omega_i)} + \langle g^{Nm}, 1 \rangle_{\Gamma^W \cap \partial\Omega_i} + c_i \, |\Gamma_i| - c_{i-1} \, |\Gamma_{i-1}| &= 0, \quad i = 2, ..., p-1. \end{aligned} \tag{2.57}$$

Uniqueness. The local solutions v_i of the pre-step and $u_i(\varphi)$ in $H^1(\Omega_i)$ ($i = 1, ..., p-1$) are only defined up to a constant. Therefore we choose the following additional condition:

$$\int_{\Gamma_i} [v] \, do = 0, \quad \int_{\Gamma_i} [u(\varphi)] \, do = 0 \quad \forall\, i = 1, ..., p-1 \tag{2.58}$$

As v_p and $u_p(\varphi)$ are already well defined due to the Dirichlet condition on Γ^O, (2.58) serves as a recurrency equality to determine the $v_{p-1}, v_{p-2}, ...$, $u_{p-1}(\varphi), u_{p-2}(\varphi), ...$ by

$$\int_{\Gamma_i} \gamma_0 v_i \, do = \int_{\Gamma_i} \gamma_0 v_{i+1} \, do, \qquad \int_{\Gamma_i} \gamma_0 u_i(\varphi) \, do = \int_{\Gamma_i} \gamma_0 u_{i+1}(\varphi) \, do, \qquad i = 1, ..., p-1. \tag{2.59}$$

So we can state that the jumps $[v]$, $[u(\varphi)]$ are situated in the space $H^{1/2}_{mv}(\Gamma)$ defined in (2.14), (2.8) and the solution $u(\varphi)$ itself is situated in the space

$$H^1_{mv}(\Omega) := \{u \in L^2(\Omega) \mid u|_{\Omega_i} \in H^1(\Omega_i),\ u|_{\Gamma^{Dir}} = 0, \int_{\Gamma_i} [u]\ do = 0\ \forall i = 1, ..., p-1\} \tag{2.60}$$

and v in

$$\{u \in L^2(\Omega) \mid u|_{\Omega_i} \in H^1(\Omega_i),\ u|_{\Gamma^{Dir}} = g^{Dir},\ \int_{\Gamma_i} [u]\ do = 0\ \forall i = 1, ..., p-1\}.$$

Obviously, the solution $w_i = v_i + u_i(\varphi)$ fulfils the partial differential equation on each subdomain Ω_i, the boundary conditions on $\partial\Omega \cap \partial\Omega_i$ and the continuity conditions $\gamma_0 u_i = \gamma_0 u_{i+1},\ \gamma_1 u_i = -\gamma_1 u_{i+1}$ on the interfaces Γ_i. Thus, w is the solution of the global problem.

Embedding of the function spaces. In the Neumann case we replace (2.17) by the embedding

$$H^{1/2}_{mv}(\Gamma) \subset L^2_{mv}(\Gamma) = (L^2_{mv}(\Gamma))^* \subset H^{-1/2}_{mv}(\Gamma) \tag{2.61}$$

with

$$\begin{aligned} L^2_{mv}(\Gamma) &:= \{\varphi \in L^2(\Gamma) \mid \int_{\Gamma_i} \varphi = 0\ \forall\ i = 1, ..., p-1\} \\ (L^2_{mv}(\Gamma))^* &= \{\varphi \in (L^2(\Gamma))^* \mid \langle \varphi, 1 \rangle_{(L^2(\Gamma_i))^*, L^2(\Gamma_i)} = 0\ \forall\ i = 1, ..., p-1\} \end{aligned}$$

and the Riesz identification

$$j : L^2_{mv}(\Gamma) \longrightarrow (L_{mv}(\Gamma))^*, \quad \varphi \longmapsto j(\varphi) := \bar{\varphi}, \quad \bar{\varphi}(\psi) = \int_{\Gamma} \varphi\ \psi\ dx.$$

Remark concerning (2.58). Beside (2.58), there are other possibilities to make $u(\varphi)$ unique. The definition of the function space of the $u(\varphi)$ determines the function space of the traces on Γ and vice versa. One might propose

$$[u_i(x_i)] = 0, \quad \forall\ i = 1, ..., p-1, \tag{2.62}$$

where $x_i \in \Gamma_i$ are fixed points, or

$$\int_{\Omega_i} u\ dx = 0 \quad \forall\ i = 1, ..., p-1, \tag{2.63}$$

to substitute (2.58). The bilinear form b would be unaffected: (2.62) or (2.63) instead of (2.58) would only shift the u_i by additive constants; the application of a $\psi \in H^{-1/2}_{mv}$ to a constant is zero.

However, (2.62) is only meaningful in the context of *discretized* u. So the difficulty would arise how to define the function space replacing $H^{1/2}_{mv}(\Gamma)$ and the corresponding imbedding (2.61) properly.

If we would replace (2.58) by (2.63), the explicit characterization of the space of the traces $\gamma_0 u$ as the space $H^{1/2}_{mv}(\Gamma)$ would get lost; the space of traces would be a certain subspace of $H^{1/2}(\Gamma)$. This would be a bit less convenient for the construction of preconditioners which have to be isomorphisms from this space to $H^{-1/2}_{mv}(\Gamma)$.

2.4.3 The algorithm in the Neumann case

Section 2.4.2 showed that we can define the bilinear form b on the space $H^{-1/2}_{mv}(\Gamma) \times H^{-1/2}_{mv}(\Gamma)$ by

$$b(\varphi,\psi) = \langle \varphi, [u(\psi)] \rangle_\Gamma, \tag{2.64}$$

where $u(\varphi)$ is the solution of the problem

$$\begin{aligned} u &\in H^1_{mv}(\Omega), \\ \int_{\Omega_i} \nabla u_i \cdot \nabla \omega \, dx &= \langle \varphi_i, \gamma_0 \omega \rangle_{\Gamma_i} - \langle \varphi_{i-1}, \gamma_0 \omega \rangle_{\Gamma_{i-1}} \\ &\quad \forall\, \omega \in \{\omega \in H^1(\Omega_i) \mid \omega|_{\Gamma^{Dir}} = 0\}, \end{aligned} \tag{2.65}$$

and that v is the solution of the folling pre-step:

$$\begin{aligned} v_i &\in H^1(\Omega_i), \quad v_p|_{\Gamma^O} = g^{Dir}, \\ \int_\Gamma [v] \, do &= 0, \\ \int_{\Omega_i} \nabla v_i \cdot \nabla \omega &= \langle f, \omega \rangle_{\Omega_i} + \langle g^{Nm}, \gamma_0 \omega \rangle_{\partial\Omega_i \cap \Gamma^{Nm}} \\ &\quad + c_i \int_{\Gamma_i} \gamma_0 \omega \, do - c_{i-1} \int_{\Gamma_{i-1}} \gamma_0 \omega \, do \\ &\quad \forall\, \omega \in \{\omega \in H^1(\Omega_i) \mid \omega|_{\Gamma^{Dir}} = 0\} \end{aligned} \tag{2.66}$$

with the c_i from (2.57), determined such that the compatibility condition (2.53) rsp. (2.55) is satisfied.

The corresponding strong formulation of (2.66) is

$$\begin{aligned} -\Delta v_i &= f \text{ in } \Omega_i \\ \frac{\partial v_i}{\partial \nu_i} &= g^{Nm} \text{ on } \partial\Omega_i \cap \Gamma^{Nm} \\ v_i &= g^{Dir} \text{ on } \Gamma^{Dir} = \Gamma^O \text{ if } i = p \\ \frac{\partial v_i}{\partial \nu_i} &= -c_{i-1} \text{ on } \Gamma_{i-1} \text{ if } i > 1 \end{aligned} \tag{2.67}$$

$$\begin{aligned} \frac{\partial v_i}{\partial \nu_i} &= c_i \text{ on } \Gamma_i \text{ if } i<p \\ \int_{\Gamma_i} [v]\, do &= 0 \quad \forall\, i=1,...,p-1 \end{aligned} \tag{2.68}$$

(compare to the Dirichlet case (2.27)) and

$$\begin{aligned} -\Delta u_i &= 0 \text{ in } \Omega_i \\ \frac{\partial u_i}{\partial \nu_i} &= 0 \text{ on } \partial\Omega_i \cap \Gamma^{Nm} \\ u_i &= 0 \text{ on } \Gamma^{Dir}=\Gamma^O \text{ if } i=p \\ \frac{\partial u_i}{\partial \nu_i} &= -\varphi_{i-1} \text{ on } \Gamma_{i-1} \text{ if } i>1 \\ \frac{\partial u_i}{\partial \nu_i} &= \varphi_i \text{ on } \Gamma_i \text{ if } i<p \\ \int_{\Gamma} [u]\, do &= 0 \end{aligned} \tag{2.69}$$

for the main problem (2.65) (compare to the Dirichlet case (2.28)); (2.57) corresponds to the strong formulation

$$\int_{\Omega_1} f_1\, dx + \int_{\Gamma^W \cap \partial\Omega_1} g^{Nm}\, dx + c_1\, |\Gamma_1| = 0,$$

$$\int_{\Omega_i} f_i\, dx + \int_{\Gamma^W \cap \partial\Omega_i} g^{Nm}\, dx + c_i\, |\Gamma_i| - c_{i-1}\, |\Gamma_{i-1}| = 0, \quad i = 2,...,p-1.$$

The last line of (2.67) and of (2.69) are ensured by adding, for $i = p-1$, $p-2$,...,1 a constant to each v_i rsp. u_i. To determine these constants, the recurrency equalities (2.59) are used.

In the following let us check that all the properties of Theorem 2.6 stay true for the bilinear form b defined by (2.64)-(2.65) on the space $H^{-1/2}_{mv}(\Gamma) \times H^{-1/2}_{mv}(\Gamma)$. Also all the remarks on the preconditioning at the end of Section 2.2 stay valid for the Neumann case if the spaces $H^{1/2}_{00}$, $H^{-1/2}$, L^2, $(L^2)^*$ are replaced by $H^{1/2}_{mv}$, $H^{-1/2}_{mv}$, L^2_{mv}, $(L^2_{mv})^*$ and Δ_0 by Δ_{Nm}.

At first let us check the validity of a Poincaré inequality in the space $H^1_{mv}(\Omega_i)$:

Lemma 2.8 (a general Poincaré inequality) *Let $\Omega \subset \mathbb{R}^n$, $n \geq 1$, be a bounded domain, $m \in \mathbb{N}$. Let $\eta \in (H^m(\Omega))^*$ with the property that $\eta(P) \neq 0$ for all nonzero polynomials $P : \Omega \to \mathbb{R}$ of degree $deg(P) \leq m-1$. Then there is a constant $c = c(\eta, \Omega)$ such that*

$$\|u\|_{H^{m-1}(\Omega)} \leq c \left(\sum_{|\alpha|=m} \|D^\alpha u\|_{L^2(\Omega)} + |\eta(u)| \right) \quad \forall\, u \in H^m(\Omega) \tag{2.70}$$

where $\|u\|^2_{H^i(\Omega)} := \sum_{|\alpha|\le i} \|D^\alpha u\|^2_{L^2(\Omega)}$.

Proof. $|\eta(\cdot)|$ is a norm in $\{P \mid P : \Omega \to \mathbb{R} \text{ is a polynomial with } deg(P) \le m-1\}$. So we can apply Satz 2.17 in [55]. ∎

From this lemma we can derive the Poincaré inequality for the space $H^1_{mv}(\Omega_i)$:

Lemma 2.9 (Poincaré inequality in $H^1_{mv}(\Omega_i)$) *There is a constant $c = c(\Omega_i)$ such that*

$$\|u\|_{L^2(\Omega_i)} \le c \, \|\nabla u\|_{L^2(\Omega_i)}$$

for all $u \in H^1_{mv}(\Omega_i)$.

Proof. $\eta(u) := \int_{\Omega_i} u \, dx$ defines a functional $\eta_i \in (H^1(\Omega_i))^*$. Let us apply Lemma 2.8 with $m=1$. We get the estimate

$$\|u\|_{L^2(\Omega_i)} \le c_i \left(\|\nabla u\|_{L^2(\Omega_i)} + |\int_{\Omega_i} u \, dx| \right). \tag{2.71}$$

Due to the definition of $H^1_{mv}(\Omega_i)$, the last integral vanishes. ∎

Remark. It is also possible to derive a 'global' Poincaré inequality in $H^1_{mv}(\Omega)$ (2.60). It reads

$$\|u\|_{L^2(\Omega)} \le c \sum_{i=1}^{p} \|\nabla u\|_{L^2(\Omega_i)} \quad \forall \, u \in H^1_{mv}(\Omega). \tag{2.72}$$

To prove this we replace η in the previous proof by $\eta_i(u) := \int_{\Gamma_i} \gamma_0 u \, do$ and get

$$\|u_i\|_{L^2(\Omega_i)} \le c_i \left(\|\nabla u_i\|_{L^2(\Omega_i)} + |\int_{\Gamma_i} \gamma_0 u_{i+1} \, do| \right)$$

which evaluates to

$$\|u_p\|_{L^2(\Omega_p)} \le c_p \, \|\nabla u_p\|_{L^2(\Omega_p)}$$

on the *last* subdomain and

$$\|u_i\|_{L^2(\Omega_i)} \le c_i \left(\|\nabla u_i\|_{L^2(\Omega_i)} + |\int_{\Gamma_i} \gamma_0 u_{i+1} \, do| \right)$$

on all the other subdomains $i = 1, ..., p-1$ (we have used (2.59)). With help of the Trace Theorem, the last integral can be estimated by the $H^1(\Omega_{i+1})$-norm of u_{i+1}. Successive application of the gained estimate for $i = p-1, ..., 1$ leads to (2.72).

Conclusion. Theorem 2.6 is valid in the Neumann case with the bilinear form b defined by (2.64)-(2.65) on the space $H^{-1/2}_{mv}(\Gamma) \times H^{-1/2}_{mv}(\Gamma)$.

Proof. a) and d) are analogous to the Dirichlet case.

ad b) As $\sigma = 0$, we have $b(\varphi,\varphi) = \sum_{i=1}^{p} \|\nabla u_i(\varphi)\|^2_{L^2(\Omega_i)}$. Using the 'global' Poincaré inequality (2.72) for the first step in (2.40), we can prove (2.36) with a constant c depending on the Poincaré constant in (2.72). This constant, again, may depend on the length of the domain Ω! To avoid this dependence on *global* parameters, we proceed like this, using the 'local' Poincaré inequality on $H^1_{mv}(\Omega_i)$:

Let X_i be the projection

$$X_i : L^2(\Omega_i) \longrightarrow L^2_{mv}(\Omega_i), \quad u \longmapsto u - \frac{1}{|\Omega_i|}\int_{\Omega_i} u\, dx.$$

Following (2.40) we get

$$\begin{aligned}\sum_{i=1}^{p} \|\nabla u_i(\varphi)\|^2_{L^2(\Omega_i)} &= b(\varphi,\varphi)\\ &\le \sum_{i=1}^{p-1} \|\varphi_i\|_{H^{-1/2}_{mv}(\Gamma_i)}\, \|\gamma_0 u_i(\varphi) - \gamma_0 u_{i+1}(\varphi)\|_{H^{1/2}_{mv}(\Gamma_i)}. \end{aligned} \tag{2.73}$$

Using the definition of the norm $\|\cdot\|_{H^{1/2}_{mv}(\Gamma_i)}$ (2.16) and the fact that constant functions are in the kernel of $(-\Delta_{Nm})^{1/4}$, then (2.11), the Trace Theorem for $H^1(\Omega_i) \to H^{1/2}(\Gamma_i)$ and the 'local' Poincaré inequality in $H^1_{mv}(\Omega_i)$ we get

$$\begin{aligned}\|\gamma_0 u_i(\varphi) - \gamma_0 u_{i+1}(\varphi)\|_{H^{1/2}_{mv}(\Gamma_i)} &= \|(-\Delta_{Nm})^{1/4}\,(\gamma_0 u_i(\varphi) - \gamma_0 u_{i+1}(\varphi))\|_{L^2(\Gamma_i)}\\
&= \|(-\Delta_{Nm})^{1/4}\,(\gamma_0 X_i u_i(\varphi) - \gamma_0 X_{i+1} u_{i+1}(\varphi))\|_{L^2(\Gamma_i)}\\
&\le c\,\|\gamma_0 X_i u_i(\varphi) - \gamma_0 X_{i+1} u_{i+1}(\varphi)\|_{H^{1/2}(\Gamma_i)}\\
&\le c\,\|\gamma_0 X_i u_i(\varphi)\|_{H^{1/2}(\Gamma_i)} + c\,\|\gamma_0 X_{i+1} u_{i+1}(\varphi)\|_{H^{1/2}(\Gamma_i)}\\
&\le c\,\|X_i u_i(\varphi)\|_{H^1(\Omega_i)} + c\,\|X_{i+1} u_{i+1}(\varphi)\|_{H^1(\Omega_{i+1})}\\
&\le c\,\|\nabla X_i u_i(\varphi)\|_{L^2(\Omega_i)} + c\,\|\nabla X_{i+1} u_{i+1}(\varphi)\|_{L^2(\Omega_{i+1})}\\
&= c\,\|\nabla u_i(\varphi)\|_{L^2(\Omega_i)} + c\,\|\nabla u_{i+1}(\varphi)\|_{L^2(\Omega_{i+1})}\end{aligned}$$

where the c are generic. We get (2.40), (2.36) now independent of the global shape of Ω.

ad c) We only have to check that the prolongation of $\psi \in H^{1/2}_{mv}(\Gamma_i)$ to a $\bar\psi \in H^{1/2}_{mv}(\partial\Omega_i)$ with $\bar\psi|_{\Gamma_{i-1}} \equiv 0$ in the proof of Theorem 2.6 part c) is possible.

Let us choose a parametrization of $\partial\Omega_i$ which maps the interval $[0, 2\pi)$ onto $\partial\Omega_i$, $(0,\pi)$ onto Γ_i, $(2/3\,\pi, 5/6\,\pi)$ onto Γ_{i-1}. Then by setting $\psi(x-\pi) := \psi(x)$, ψ is prolongated to a $H^{1/2}_{mv}(\partial\Omega_i)$-function. (The $H^{1/2}$-property can be checked easily with the representation of $H^{1/2}$ by double integrals (3.165). Then, ψ is multiplied by a $C^1(\partial\Omega_i)$-function which is 1 on Γ_i and 0 on Γ_{i-1} such that the product has mean value zero. This operation preserves the $H^{1/2}$-property, as can be seen e.g. by using the double integral representation again. ∎

2.5 The local solvers

Within each CGBI iteration step (2.48) and in the pre-step (2.46) and post-step (2.47) a local partial differential equation has to be solved on each subdomain. The CGBI parallelization method enables the use and the *coupling* of arbitrary local solvers on the subdomains Ω_i. On *rectangular* subdomains the use of *spectral* solvers is sensible. The high accuracy of spectral solvers enables the use of a very low discretization parameter on those subdomains. On non-rectangular subdomains (as they occur when the flow around an obstacle is simulated), finite element solvers may be used. On channel-like domains, the possibility of coupling different solvers should make CGBI an interesting alternative to, for example, parallel FE/multigrid solvers.

Our program includes three different local solvers:

1. A finite element solver.
   The FE solver uses a triangular mesh and linear shape functions. On each FE subdomain, the resulting set of equations is solved by a sequential CG method. (However, a sequential multigrid method might be preferable).

2. A Chebyshev collocation spectral solver.
   This solver uses the collocation method with respect to the so-called Gauss-Lobatto points, which are for the two-dimensional unit square $[-1,1] \times [-1,1]$

$$\{(x_i, y_j)^T \mid i{=}0, ..., N_X,\ j{=}0, ..., N_Y\}, \quad x_i = cos\frac{\pi i}{N_X},\ \ y_j = cos\frac{\pi j}{N_Y}.$$

   The Chebyshev collocation method uses the explicit knowledge of the relation between the values of a function of the (polynomial) ansatz space at the Gauss-Lobatto points and the values of its (first, second order) derivative at the same points. As the dimensions decouple, these relations are given by $(N_X{+}1) \times (N_X{+}1)$- rsp. $(N_Y{+}1) \times (N_Y{+}1)$-matrices for $\partial/\partial x$, $\partial^2/\partial x^2$ rsp. $\partial/\partial y$, $\partial^2/\partial y^2$. In the case of the first order derivative, this matrix can be found in [12], p. 69 (and in [16]). The second order derivative matrix was found by Peyret 1986. It is given in [16], p. 7.

   After eleminating the boundary grid points by means of the boundary conditions, a diagonalization[10] is performed on the resulting system matrix. For time-dependent problems (flow problems) this time-consuming diagonalization only takes place *once*. If the diagonalization is performed, the application of the Chebyshev spectral solver takes $O(N^3)$ $(N = N_X = N_Y)$ operations.

[10] i.e. a computation of the eigenvalues and eigenfunctions

A short summary of the method can be found in [7]. A more detailed description is given in [16], [17].

3. A finite difference (FD-) solver.
It uses the ordinary 5-point-stencil approximation

$$-\Delta u(x_i, y_j) \approx \tag{2.74}$$
$$\frac{1}{h^2}\left(4\,u(x_i, y_j) - u(x_{i+1}, y_j) - u(x_{i-1}, y_j) - u(x_i, y_{j+1}) - u(x_i, y_{j-1})\right)$$

on an equidistant mesh $x_i = y_i = ih$, $h = N^{-1}$, $i = 0, ..., N$ on the unit sqare $[0,1]\times[0,1]$.

The system of equations is solved efficiently by a *multigrid* method. We use V-cycles with two relaxed Gauss-Seidel steps before and two after the descent.

2.6 The discretization on the interfaces

A quality of CGBI is its flexibility: Almost all questions of discretization are left to the local solvers which facilitates the use of arbitrary local solving modules. Beside an ordinary CG iteration in which the local solvers are called only the following items are left to CGBI:

1) the calculation of the jump at the interfaces

2) the application of the transition operator $C : H_{00}^{1/2}(\Gamma) \to H^{-1/2}(\Gamma)$ rsp. $C : H_{mv}^{1/2}(\Gamma) \to H_{mv}^{-1/2}(\Gamma)$

The only requirement to the local solving modules is that they provide a method to evaluate a local solution at some grid points on the interfaces. This is sufficient to enable CGBI to perform 1) and 2). Let us concentrate on an interface Γ_i, and let us denote $G_-^i \subset \Gamma_i$ the set of points on which the solver on Ω_{i-1} provides the solution and $G_+^i \subset \Gamma_i$ the set of points on which the solver on Ω_i provides its solution.
ad 1). If G_-^i and G_+^i *coincide* then CGBI uses this boundary mesh $G_-^I = G_+^i$, and the calculation of the jump on Γ_i is trivial. Otherwise some interpolation has to be applied: If Γ_i is surrounded by a FD subdomain and a Chebyshev subdomain our program uses interpolation onto an equidistant grid by piecewise linear polynomials. If Γ_i is surrounded by two Gauss-Lobatto domains we perform the CG iteration on the Gauss-Lobatto grid; finite order interpolation onto another grid would destroy the spectral accuracy.

If FE-subdomains are involved, there are two possibilities:

a) Construct the FE-meshes such that its restriction onto the interfaces coincides with the restriction of the mesh of the adjoint subdomain (equidistant or Gauss-Lobatto). In this case the FE-subdomain can be treated as a FD- rsp. a Chebyshev subdomain.

b) If it is not suitable to construct such a FE-mesh, interpolation onto (e.g.) the boundary mesh induced by the adjoint subdomain is necessary.

ad 2). Chapter 3 deals with the transition operator C. It turns out that a discrete version of C can be found easily if we can access the *equidistant* grid values of the function being subject to C. As already mentioned in 1), the application of CG with respect to an equidistant boundary mesh would diminish the accuracy if more than one spectral subdomain is involved. One might try to use the equidistant mesh *only* for the preconditioner, i.e. to perform CG on the Gauss-Lobatto mesh and to interpolate before and after the application of C:

$$C_{h,GL} := I_{eq \to GL} \circ C_{h,equ} \circ I_{GL \to eq}$$

This method seems to be problematic (see Section 3.2). So the construction of discrete preconditioners on *non-equidistant* grids in Chaper 3 seems to be indispensable.

2.7 The discrete scalar product

Interpretation and evaluation of the duality couple $\langle\cdot,\cdot\rangle_{\Gamma_i}$. In the CGBI algorithm in Section 2.3 the application of a $\varphi \in H^{-1/2}(\Gamma_i)$ onto a $\psi \in H_{00}^{1/2}(\Gamma_i)$ (rsp. $\varphi \in H_{mv}^{-1/2}(\Gamma_i)$, $\psi \in H_{mv}^{1/2}(\Gamma_i)$ in the Neumann case) has to be calculated. How has this to be treated in the discretized version of CGBI?

Discrete solution spaces are usually subspaces of L^2 (or they can be identified with subspaces of L^2); by means of the embedding (2.17) rsp. (2.61) we can interprete the expressions $\langle\varphi_h,\psi_h\rangle_{\Gamma_i}$ as scalar products in $L^2(\Gamma_i)$. But it should be emphasized that CGBI also works if the exact solution φ^0 is *not* situated in $L^2(\Gamma)$ but only in $H^{-1/2}(\Gamma)$ ($H_{mv}^{-1/2}(\Gamma)$)! In this case, the discrete approximations of φ^0 lying in $L^2(\Gamma)$ converge to φ^0 in $H^{-1/2}(\Gamma)$ ($H_{mv}^{-1/2}(\Gamma)$) if the discretization parameter tends to zero.

However, we may interprete the expressions $\langle\varphi_h,\psi_h\rangle_{\Gamma_i}$ as scalar products $(\varphi_h,\psi_h)_{L^2(\Gamma_i)}$. Furthermore, let us assume that the discrete solution spaces enable the pointwise evaluation of functions (for FE and for spectral methods, this is obviously true). If there is an *equidistant* mesh on Γ_i we can approximate the $L^2(\Gamma_i)$ scalar product with help of the trapezoid rule

$$(\varphi_h,\psi_h)_h := \frac{|\Gamma_i|}{N}\left(\frac{\varphi_h(x_0)\,\psi_h(x_0)}{2} + \sum_{i=1}^{N-1}\varphi_h(x_i)\,\psi_h(x_i) + \frac{\varphi_h(x_N)\,\psi_h(x_N)}{2}\right) \quad (2.75)$$

where the x_i are the equidistant grid points.

In case of a *Gauss-Lobatto* mesh on Γ_i we may use the transformation[11]

$$\int_{-1}^{1}\varphi\,\psi\;dx = \int_{0}^{\pi}\varphi\circ\cos\;\psi\circ\cos\;\sin\xi\;d\xi.$$

The discretization of the right hand side by the trapezoid rule[12] with respect to an equidistant grid leads to the discrete scalar product

$$(\varphi_h,\psi_h)_{h,GL} := \frac{\pi}{N}\sum_{i=1}^{N-1}\varphi(\bar{x}_i)\,\psi(\bar{x}_i)\,\sin\frac{i\pi}{N} \quad (2.76)$$

where the $\bar{x}_i = \cos(i\pi/N)$ are the Chebyshev-Gauss-Lobatto grid points.

[11] For the sake of simplicity let us assume $\Gamma_i = (-1,1)$ here.

[12] Let us note that the trapezoid rule is highly accurate here, as the integrand is periodic ([47] Section 8.2.1).

The symmetry of the system matrix. The classical CG method requires the symmetry of the system matrix. But collocation methods like the Chebyshev method mentioned in Section 2.5 are known for destroying this property, i.e. for a symmetric differential operator the system matrix may be non-symmetric.

Therefore it seems reasonable to investigate the (non-)symmetry of the system matrix A_h for our operator

$$A : \varphi \longrightarrow [u(\varphi)] \tag{2.77}$$

corresponding to the bilinear form b and the minimization principle J. Let us mention that the entries of the system matrix are not actually present in the CGBI program code, as the application of A_h is performed 'implicitely' by solving boundary value problems and calculating the 'jump'. As a measure for the (non-)symmetry of the discrete operator A_h with respect to a scalar product $(\cdot,\cdot)_h$ we may choose a basis $\varphi^1, ..., \varphi^{N-1}$ of the N–1-dimensional discrete solution space and calculate

$$\begin{aligned} s(A_h) &= s(A_h; \varphi^1, ..., \varphi^{N-1}) \\ &:= \frac{\left(\sum_{i,j=1}^{N-1} |(\varphi^i, A_h\varphi^j) - (A_h\varphi^i, \varphi^j)|^2 \right)^{1/2}}{2 \left(\sum_{i,j=1}^{N-1} |(\varphi^i, A_h\varphi^j)|^2 \right)^{1/2}} \in [0,1]. \end{aligned} \tag{2.78}$$

If A_h is symmetric with respect to $(\cdot,\cdot)$, $s(A_h)=0$, if A_h is antisymmetric, $s(A_h)=1$.

If the FD (or the FE) solver is used we should expect A_h being symmetric with respect to the scalar product (2.75). And in fact, due to numerical round off errors we achieve for $s(A_h)$ values between 10^{-8} and 10^{-7}.

Fig. 2.5, however, shows $s(A_h)$ in the case of 2 subdomains with Chebyshev spectral solver and a Chebyshev-Gauss-Lobatto mesh on the interface. On the left of Fig. 2.5, Dirichlet boundary conditions are imposed, and on the right Neumann boundary conditions. On the horizontal axis, the discretization parameter $N = h^{-1}$ is given. For the *full* lines, the 'correct' scalar product (2.76) was used[13]. For the *broken* lines, the 'wrong' scalar product

$$\frac{1}{N} \left(\frac{1}{2}\,\varphi(\bar{x}_0)\,\psi(\bar{x}_0) + \sum_{i=1}^{N-1} \varphi(\bar{x}_i)\,\psi(\bar{x}_i) + \frac{1}{2}\,\varphi(\bar{x}_N)\,\psi(\bar{x}_N) \right) \tag{2.79}$$

was used. The figure shows that with the 'correct' scalar product the symmetry of A_h is much better than with the other. Test runs confirmed that the classical CG method is applicable, if the correct scalar product (full lines) is used. For the

[13] Correct in the sense that it approximates the $H_{00}^{1/2}$-$H^{-1/2}$-duality (rsp. the $H_{mv}^{1/2}$-$H_{mv}^{-1/2}$-duality in the Neumann case) in which the exact operator A is symmetric.

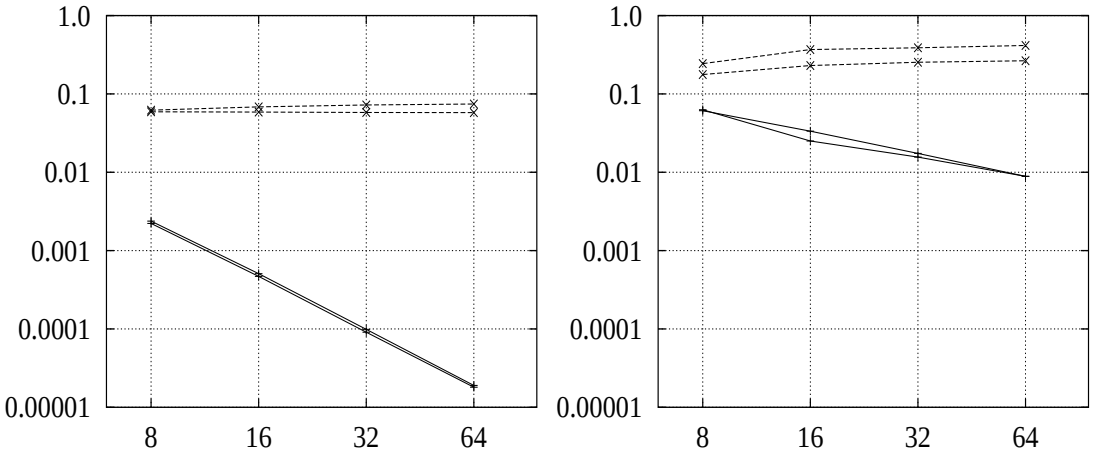


Figure 2.5: Asymmetry $s(A_h)$ in the Dirichlet case (left figure) and in the Neumann case (right figure). Discretization parameter N on the horizontal axis. The full lines represent (2.78) with the scalar product (2.76), the broken lines (2.78) with (2.79). Two different bases $(\varphi^i)_{i=1,\ldots,N-1}$ where used; therefore each line appears twice.

other scalar product (broken lines), the classical CG algorithm only converges for the Dirichlet case, but not for the Neumann case. This can be explained by the larger asymmetry in the Neumann case (compare broken lines left↔right). This larger asymmetry, in turn, can be explained by the fact that the two scalar products (2.76) and (2.79) differ by a weight which is singular at the boundary of Γ_i, and for functions being zero at the boundary (=Dirichlet case) this weight function loses importance.

Our tests (Sec. 2.8 and 3.1.5) showed that our CGBI method using the classical CG algorithm for symmetric matrices handles the slight asymmetry (Fig. 2.5, full lines) of the discrete operator without any problem. The implementation of a CG version for *asymmetric* matrices turned out to be unnecessary.

2.8 Test runs

In this section some test runs with FDM and with Chebyshev subsolvers, with Dirichlet and with Neumann boundary conditions are made. Diagrams showing the errors of the numerical solution of the Poisson equation with respect to the discretization parameter are given.

In Section 3.1.5 (also 3.3, 3.4), far more tests are made including the coupling of the different local solvers and the variation of several parameters. There, special emphasis is laid on the examination of the *speed* of convergence. Here in Sec. 2.8 we focus on the final CGBI error and its dependence on the discretization parameter.

We suppose that we have a rectangular domain Ω (Fig. 2.2).

For all the tests, the exact solution is given so that the error can be calculated. These are the tested exact solutions:

Test function 1:

$$u(x,y) = \frac{108}{L}\, x^2(x-L)\, y^2\, (1-y)^2 \tag{2.80}$$

(see Fig. 2.6) has $\max\limits_{(x,y)\in\Omega} |u(x,y)| = L^2$, $\Omega = (0,L)\times(0,1)$. This function has homogeneous Dirichlet boundary values on $\partial\Omega$ and also homogeneous Neumann boundary values on $\Gamma^W \cup \Gamma^I$.

Test function 2:

$$u(x,y) = e^4\, e^{-1/y}\, e^{-1/(2-y)}\, e^{-L/x}\, e^{-L/(2L-x)} \tag{2.81}$$

(see Fig. 2.7) has $\sup\limits_{(x,y)\in\Omega} |u(x,y)| = 1$ and homogeneous Neumann boundary conditions on $\partial\Omega$. The higher order derivatives of (2.81) take rather large values.

Test function 3:

$$u(x,y) = x\,(L-x)\, y\,(1-y)\, \psi(y-x/L) \tag{2.82}$$

with

$$\psi(d) = \begin{cases} 1, & d \le 0 \\ 1 - d^2/4, & d > 0 \end{cases}$$

The second order derivatives of u are discontinuous at the diagonal $y = x/L$. The two factors $x\,(L-x)\,y\,(1-y)$ and $\psi(y-x/L)$ are displayed in Fig. 2.9; u itself is given in Fig. 2.8. (2.82) lies in $C^1(\Omega)$ and in $H^2(\Omega)$, but not in $C^2(\Omega)$ or in $H^3(\Omega)$. The correct interface condition $\varphi = \frac{\partial u}{\partial x}|_\Gamma$ for (2.82) is only a continuous function, but it is not in $C^1(\Gamma)$. $\frac{\partial u}{\partial x}$ is displayed in Fig. 2.10. Fig. 2.12 shows the decay of the Chebyshev series for functions of different regularity.

Test function 4:

$$u(x,y) = \frac{25}{L^2}\, e^{-y}\, y\,(1-y)\, \cos(Wx)\, x\,(L-x) \tag{2.83}$$

is a function with homogeneous Dirichlet boundary values on $\partial\Omega$. In the test runs and in Fig. 2.11 we take $W=5$. In that case, $\max_{(x,y)\in\Omega} |u(x,y)| \approx 1$.

This function is used for larger channel lengths L. For integer L and W, the cosine oscillation is not commensurable with the length of the domain.

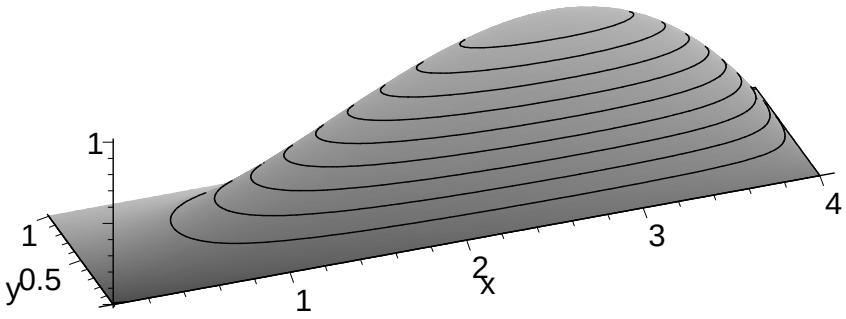


Figure 2.6: Test function 1.

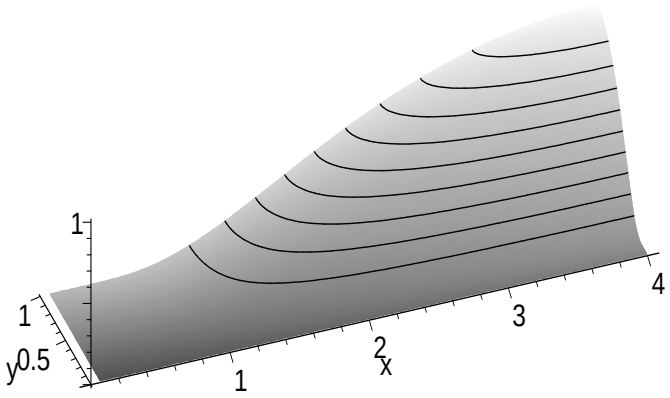


Figure 2.7: Test function 2.

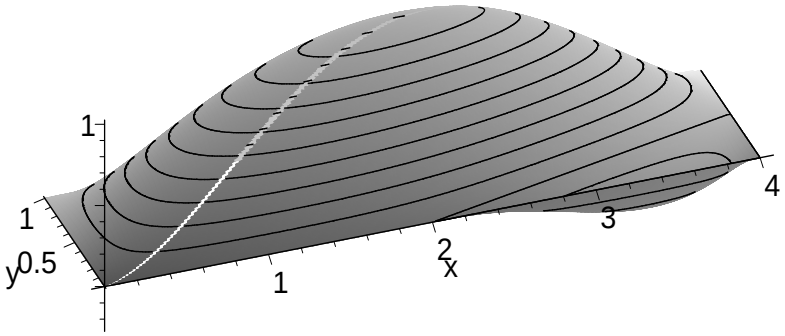


Figure 2.8: Test function 3.

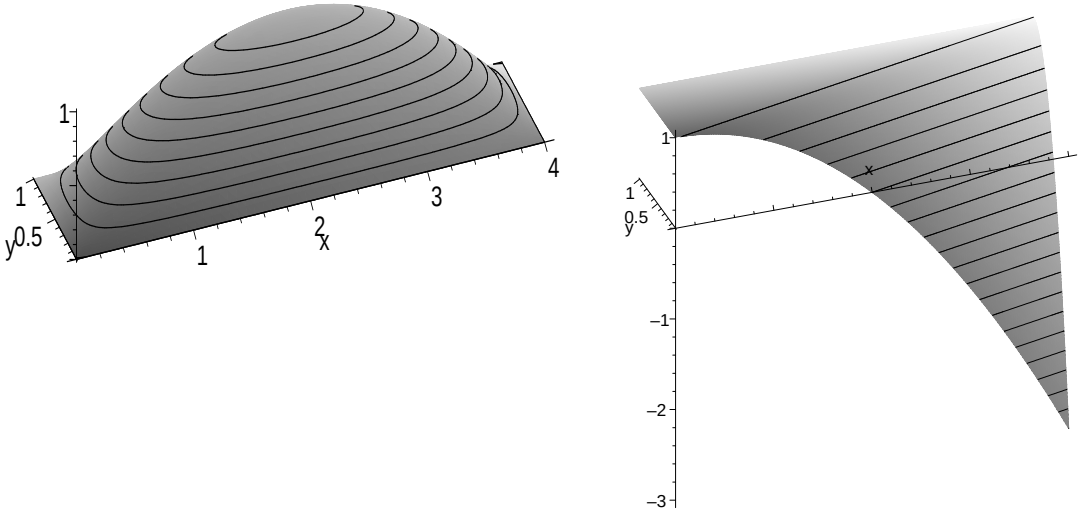


Figure 2.9: The smooth and the non-smooth factor of test function 3.

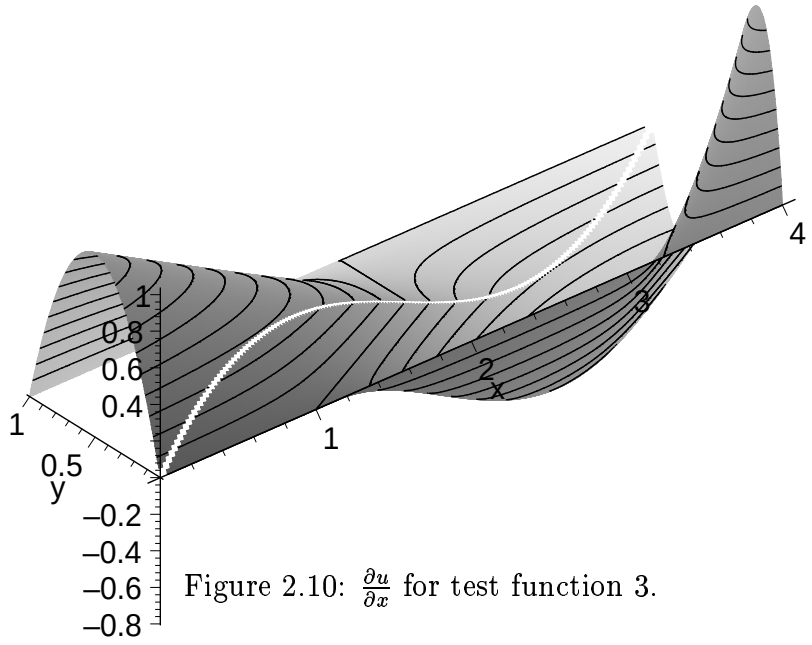


Figure 2.10: $\frac{\partial u}{\partial x}$ for test function 3.

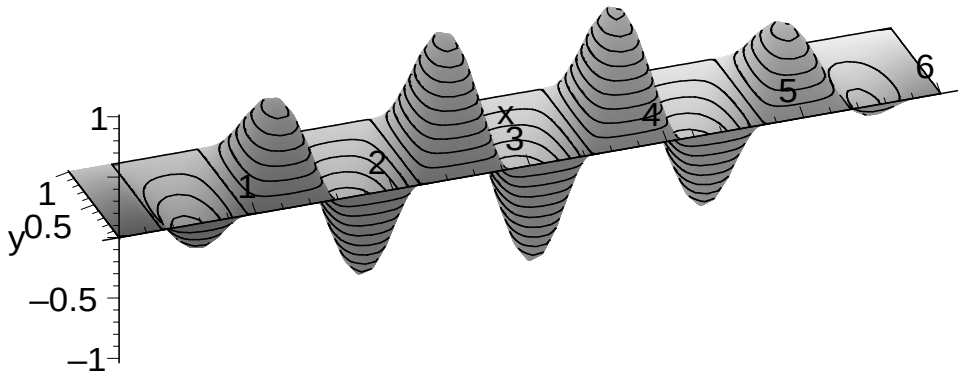


Figure 2.11: Test function 4.

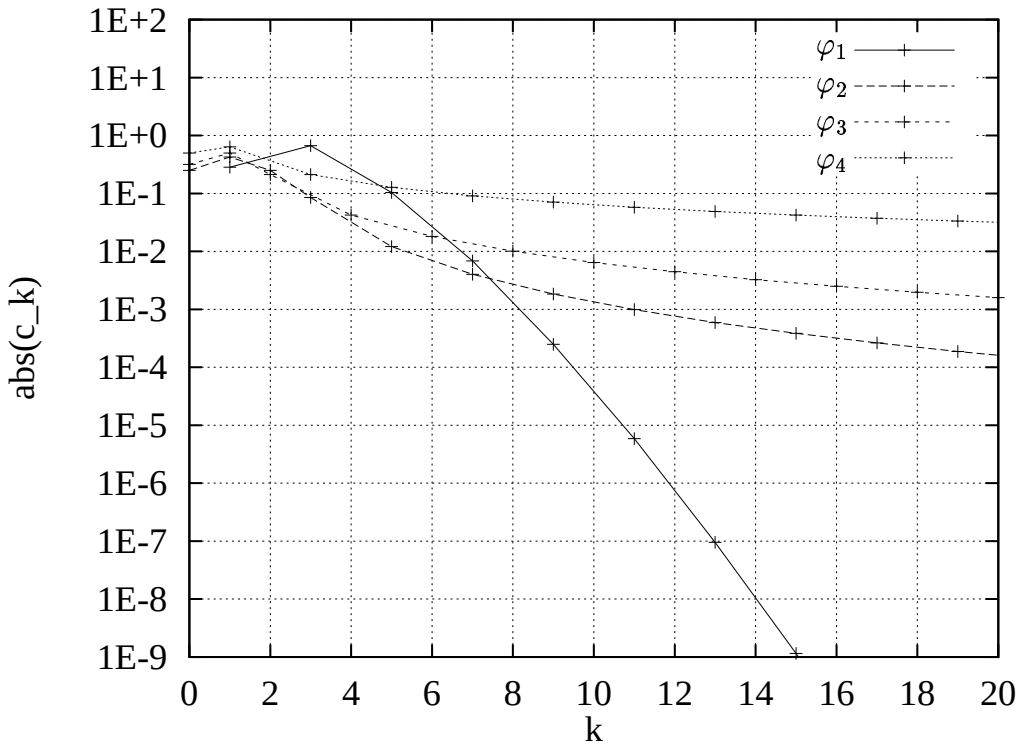


Figure 2.12: Absolute value of the Chebyshev coefficients c_k for functions $f(x) = \sum_{k=0}^{\infty} c_k T_k(x)$, $T_k(x) = \cos(n \arccos x)$, of different regularity on $[-1, 1]$. Only the non-vanishing coefficients are displayed. $\varphi_1(x) = \sin \pi x$, $\varphi_2(x) = x^2 \chi(x \geq 0)$, $\varphi_3(x) = x \chi(x \geq 0)$, $\varphi_4(x) = \chi(x \geq 0)$. $\chi(M)$ is the characteristic function of a set $M \subset [-1, 1]$. The higher the regularity of φ, the faster the decay of the Chebyshev series. Concerning regularity, φ_3 corresponds to the boundary condition $\partial u/\partial x|_{\Gamma_2} = \varphi|_{\Gamma_2}$ for the test function 3 for $p = 4$, $L = 4$. For the other test functions 1, 2, 4, $\partial u/\partial x|_{\Gamma_i}$ is in $C^\infty(\Gamma_i)$ which corresponds to φ_1 in the figure.

The test runs. Figs. 2.13 and 2.14 show the error of the numerical solution in the $L^\infty(\Omega)$-norm with respect to the discretization parameter N. 4 square subdomains with $(N+1) \times (N+1)$ grid points each were used. The error after the CGBI becomes stationary is displayed. The test function (t.f.) 2 uses Neumann boundary conditions. For all the other cases, Dirichlet conditions are applied.

If FD solvers are used (Fig. 2.13) the second order FD discretization error leads to a global error of second order in N^{-1}. Only for t.f. 3, as expected, the lack of regularity causes a lower order.

If Chebyshev spectral solvers are used (Fig. 2.14) only the round off error of $\approx 10^{-8}$ occurs[14] as soon as the modes of the exact solution are resolved in the local ansatz spaces, i.e. spectral accuracy is reached. T.f. 1 is a polynomial of low order, i.e. it is situated *in* the ansatz spaces. Therefore spectral accuracy is gained even for small N. T.f. 2 requires a very large polynomial ansatz space to resolve its modes. The lack of regularity of t.f. 3 destroys the spectral accuracy in this case. However, it seems to be remarkable that the error is still smaller than in the FD case.

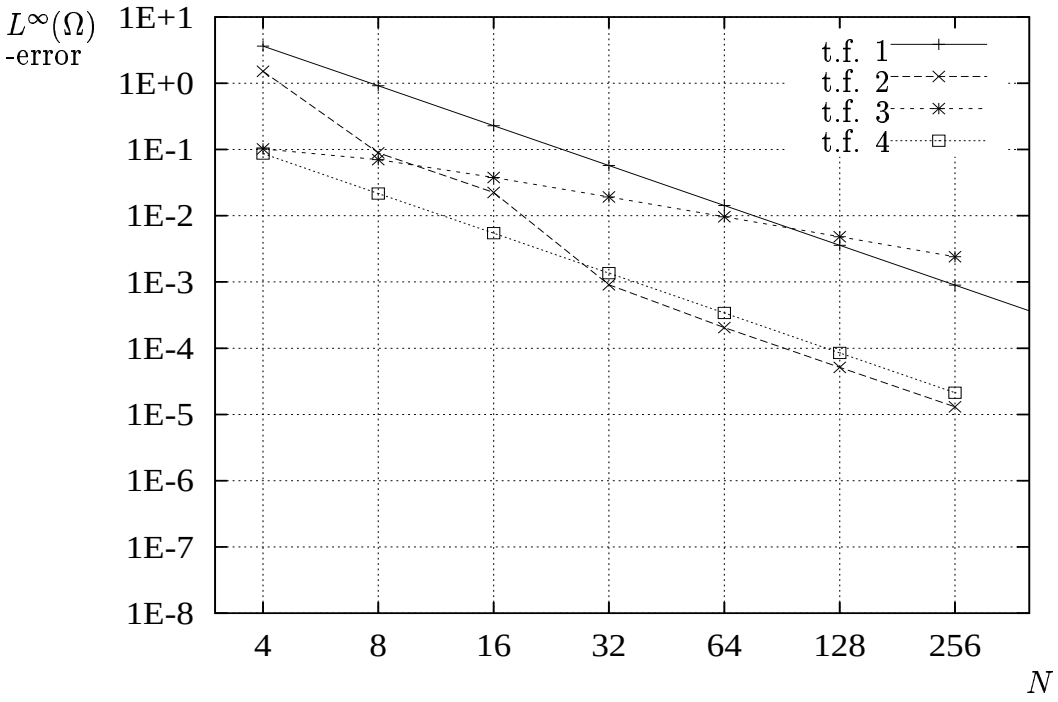


Figure 2.13: CGBI error for 4 FD subdomains.

[14] 10^{-8} with respect to the maximum of the solution.

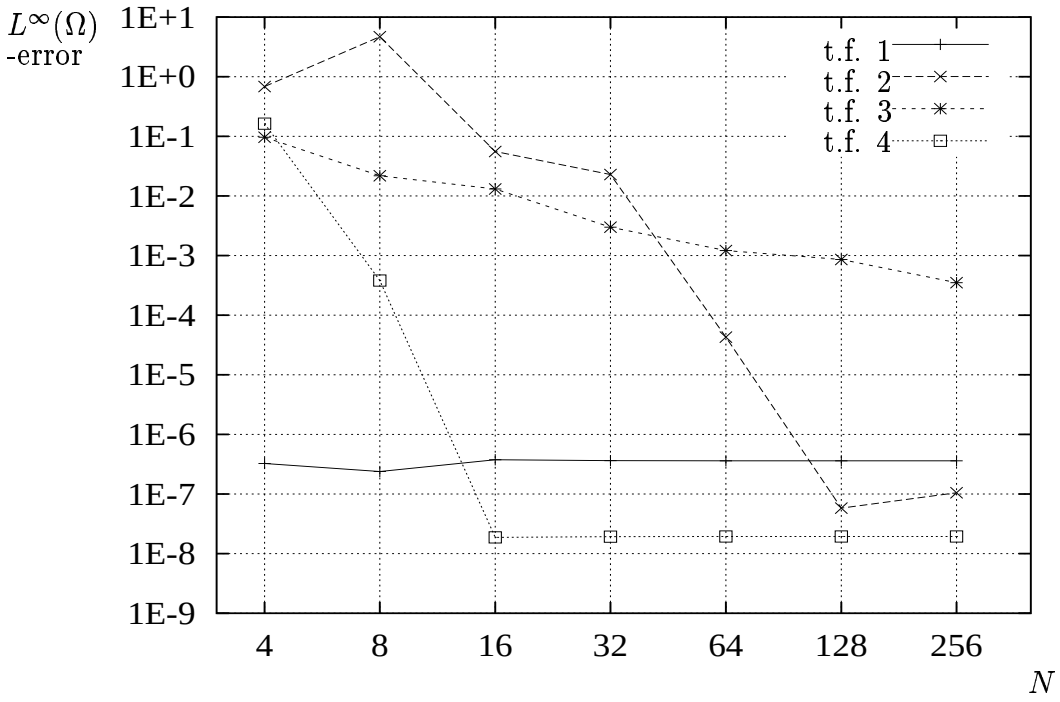


Figure 2.14: CGBI error for 4 spectral subdomains.

2.9 CGBI and other domain decomposition methods

In the last twenty years a large variety of algorithms for parallelization by domain decomposition was presented. In the following we restrict ourselves to approaches which are related to CGBI;[15] we do not discuss e.g. the Schwarz method for overlapping subdomains.

CGBI reduces the global problem (2.1) to the inversion of the operator

$$A : \varphi \longrightarrow [u(\varphi)] \tag{2.84}$$

where $u(\varphi)$ is the function matching (2.1) locally on each Ω_i with *Neumann* boundary data φ on the interfaces Γ (and homogeneous Neumann or Dirichlet data on $\partial\Omega$, see Sections 2.2, 2.4). The continuity condition for the normal derivative $[\partial u/\partial\nu] = 0$ is intrinsically guaranteed by this approach, while the CG iteration process causes $[u] \to 0$.

Closely related to CGBI is the following 'dual' approach: Guarantee $[u] = 0$ and try to minimize the jump of the normal derivatives $[\partial u/\partial\nu] \to 0$. This approach leads to the operator

$$A^* : \psi \longrightarrow \left[\frac{\partial u^*(\psi)}{\partial\nu}\right] \tag{2.85}$$

where $u^*(\psi)$ is the function matching (2.1) on each Ω_i with homogeneous *Dirichlet* boundary data ψ on the interfaces Γ. In fact, the use of operator (2.85) is more classical than (2.84); as a method for parallelization it was investigated from the early 1980s on ([4], [15], [35], [45] p. 4, [53] and papers cited there).

As A (see Theorem 2.6), also A^* is symmetric, coercive and continuous ([45] p. 8-9). That means that a CG iteration is possible to invert (2.85). The condition number of FE discretizations of A^* is estimated in [35] and, more generally, in [11].

The mapping $\varphi \mapsto u(\varphi)$ is usually called the *Poincaré-Steklov operator*. (In [45], instead the mapping A^* has this name.)

Both methods have in common that they 'reduce' the problem given on the domain Ω to a 'smaller' ('dual') problem on the interfaces Γ; both methods use a preliminary step to handle the inhomogenities of the global problem and determine the correct interface boundary condition φ rsp. ψ, then. If we restrict ourselves to the model case of only two subdomains where one is the mirror image of the other, both operators are even inverse to each other up to a constant factor:

$$\frac{\partial}{\partial\nu}u(\varphi) = \varphi, \quad A^*A = 4\,id \tag{2.86}$$

[15] An overview over several methods is given in [45] [53].

If regarded on a discrete ('matrix') level the methods related to A and A^* are called *dual Schur methods* or *Schur complement methods* because the matrices of the discretized problems (2.84), (2.85) are the Schur complements of the discretized global problem (2.1). A very common method which was developed for the *finite element* approximation of problems in *structural mechanics* is the *FETI* method (Finite Element Tearing and Interconnecting) by Farhat & Roux (first publications: [18] [19] [20] [21]; more recent e.g. [22] [40] [3]). FETI and CGBI have in common that the resulting local problems correspond to boundary value problems with *natural* boundary conditions. The FETI method expresses the interface condition by Lagrange multipliers and starts from a saddle-point problem which is discretized. Main differences between FETI and CGBI are the construction of the preconditioner and the handling of the so-called *floating subdomains* (see Sec. 2.4). The problem of the floating subdomains is more grave in the context of structural mechanics. This leads to a *projected* CG algorithm for FETI. For flow problems, CGBI is able to handle the floating subdomains by the simple fact that its preconditioner C defines a proper isomorphism of the function spaces $R(A)$ and $D(A)$ (see Sec. 3.1.1); $C\varphi$ is automatically in the correct space, no projection is needed.

A lot of authors have developed preconditioners for the two operators. Most have in common that they lead to a condition number

$$\kappa \approx c\,(1 + \log H/h)^{\gamma}$$

where h is the mesh size of the discretization, H is the diameter of the subdomains and γ is usually equal to 2 or 3 ([45] Sec. 3.3.2 and the authors cited there and (very recent) [28]). So the condition number depends on the discretization parameter which is not the case for our preconditioner developed in the following chapter. Furthermore, these preconditioners require the solution of additional local problems on the subdomains which is much more time consuming than the preconditioner of Chapter 3. However, the development of those preconditioners was done in a more general setting concerning the geometry of Ω, its decomposition and the given operator L, whereas we consider geometries given on p. 8. So the challenge remains to investigate the effectivity of our preconditioner in the more general case of interior crosspoints of the interfaces. Let us mention the early work by Dryja [15] and Bjørstad & Widlund [4] in which a preconditioner acting only on the interface and leading to a condition number independent of h was constructed for A^* instead of A. They presented test runs for a model case of two subdomains and a FD discretization on an equidistant Cartesian mesh. Both Dryja and Bjørstad & Widlund used the square root of the *discretized* negative Laplacian while Chapter 3 of this thesis proposes the non-discretized. See remark at the bottom of page 92. In this thesis we extend the application of interface-based preconditioners to Gauss-Lobatto boundary meshes and to Neumann boundaries causing the 'floating' subdomains, and we demonstrate the

application for FE-Chebyshev couplings and for Navier-Stokes flow problems.

Let us emphasize that the construction of our preconditioner for the operator A is also meaningful for the Schur complement method based on A^*: From (2.86) or from [15] it is clear that the *inverse* of our preconditioner is well suited as a preconditioner for the Schur complement method concerning A^*, at least in the case of two subdomains. The application of the method of the proof of Theorem 3.12 to the operator A^* instead of A shows that the inverse of our preconditioner is well suited as a preconditioner for A^* in the case of an *arbitrary* number of subdomains, too. In fact, A^* with preconditioner $(-\Delta)^{-1/2}$ and A with preconditioner $(-\Delta)^{1/2}$ lead to exactly *the same* condition number for the geometry of Sec. 3.1.4 and equidistant FD meshes ($\sigma = 0$).

Finally, let us mention that for the Bramble domain decomposition approach a condition number independent of h was found, but just for geometries without interior crosspoints. See [9] [10], esp. Theorem 1 in [9].

CGBI and FETI handle the problem of applying A^{-1} rsp. $(A^*)^{-1}$ by solving local boundary value problems. A modified approach consists in assembling the matrices related to A, A^* (i.e. the Schur complement matrices) explicitely. This is expensive both concerning memory and time, as the matrices are containing blocks which are dense, but this method may be considered if the number of nodes on the interfaces is not too large, as it happens to be the case with spectral methods. Comparational computations are presently done [43].

Chapter 3

Preconditioning Techniques for CGBI

The theoretical results of the previous chapter (Cor. 2.7 and the following remark) already showed how to construct a preconditioner for CGBI. In *this* chapter we will mainly deal with a simplified geometry of the domain Ω. In fact, we will assume that Ω is a rectangle and that all subdomains Ω_i are rectangles of the same size. In this case it is possible to calculate the eigenvalues and eigenfunctions of the operator (2.77) related to the minimization principle (more or less) explicitely. This knowledge of the eigenvalues and eigenfunctions will give us another (less theoretical) approach to construct preconditioners. Of course, the results of this new approach will be similar to those of Chapter 2. The advantages of the new approach using the simplified geometry is that it allows to

- find explicit bounds for the estimate (2.42) of the preconditioned operator (in (2.42), these bounds depend on the unknown norms of some restriction and prolongation operators)

- investigate the dependence of these constants on σ and on the subdomain aspect ratio r and modify the preconditioner to make the bounds independent of these parameters.

Within this chapter, two main discretizations of the square root of the negative Laplacian (i.e. the preconditioning operator) are given: A 'spectral' one based on FFT (Section 3.1) and an approach based on sparse matrices (Section 3.4). Originally, both approaches require the equidistance of the boundary mesh on Γ. In Section 3.1.3, the interpolation theory of Sobolev spaces is used to adapt the discrete preconditioner to a Chebyshev-Gauss-Lobatto boundary mesh.

Of course, the *discretization techniques* (i.e. the use of FFT (p. 54), the use of sparse matrices (Sec. 3.4), the reduction of the Gauss-Lobatto case to the equidistant case (Sec. 3.1.3)) introduced in this third chapter also apply in the

case of a more general global geometry (Chapter 2). Only the explicit knowledge of the bounds of the eigenvalues gets lost, then.

3.1 Eigenvalues and the spectral preconditioner

3.1.1 The main concept

To construct preconditioners, we will examine the *exact* (non-discrete) operator A related to the bilinear form b (2.34):

$$A : D(A) \longrightarrow R(A), \quad \varphi \longmapsto [u(\varphi)] \tag{3.1}$$

with

- $u = u(\varphi)$ defined by (2.30),
- and $D(A) = H^{-1/2}(\Gamma)$, $R(A) = H^{1/2}_{00}(\Gamma)$

in the case of Dirichlet boundary conditions on Γ^W and

- $u = u(\varphi)$ defined by (2.65),
- $D(A) = H^{-1/2}_{mv}(\Gamma)$, $R(A) = H^{1/2}_{mv}(\Gamma)$ in the Neumann case.

As pointed out in Chapter 2.2, A (rsp. b) is symmetric and positive definite.

Concerning the geometry and the boundary conditions we are making the following assumptions in this chapter:

- The domain Ω is a rectangle (i.e. there is no obstacle in the channel) and all the subdomains Ω_i are rectangles of the same size.
- The type of the boundary condition (Dirichlet or Neumann) does not change within Γ^W, Γ^I, Γ^O, i.e. for example $\Gamma^{Dir} = \Gamma^W$ or $\Gamma^{Dir} = \Gamma^I \cup \Gamma^O$ or $\Gamma^{Dir} = \partial\Omega$.

If we know the eigenvalues $\lambda_k > 0$ and the eigenfunctions φ_k of A we can decompose each $\varphi \in D(A)$:

$$\varphi = \sum_{k=1}^{\infty} \alpha_k \, \varphi_k \tag{3.2}$$

Thus, we may write

$$A : D(A) \longrightarrow R(A), \quad \varphi = \sum_{k=1}^{\infty} \alpha_k \, \varphi_k \longmapsto A\varphi = \sum_{k=1}^{\infty} \alpha_k \lambda_k \, \varphi_k. \tag{3.3}$$

The preconditioner C to be constructed should be a good approximation of A^{-1}. It would be optimal if we had

$$C : R(A) \longrightarrow D(A), \quad \varphi = \sum_{k=1}^{\infty} \alpha_k \varphi_k \longmapsto C\varphi = \sum_{k=1}^{\infty} \alpha_k \lambda_k^{-1} \varphi_k. \tag{3.4}$$

In this case C would be inverse to A, and the condition number κ of a discretized version of the preconditioned operator AC could be expected to be very close to 1.

The construction of the preconditioners consists of the following steps:

1. Try to determine (analytically) the eigenvalues and eigenfunctions of the exact operator A.

2. Decompose the boundary conditions φ which occur during the CG-iteration into the eigenfunctions according to (3.2).

3. Multiply each coefficient with the reciprocal value of the related eigenvalue and sum up the terms (see (3.4)).

Fortunately, the restrictions of the eigenfunctions to the interfaces $\varphi_k|_{\Gamma_i}$ are simple trigonometric functions (Lemma 3.1, Theorem 3.12). Therefore we can use the fast Fourier transform (FFT) for the decomposition in 2 and the recomposition in 3.

This approach through the investigation of the eigenvalues and eigenfunctions is completely different from the approach of Chapter 2, but the results are very similar.

Throughout this chapter we are using expansions of the kind

$$\varphi(y) = \sum_{k=1}^{\infty} \alpha_k \sin \pi k y$$

for $\varphi \in H^{-1/2}(\Gamma_i)$, $\Gamma_i = [0,1]$. These series do not converge in $L^2(\Gamma_i)$, but only in $H^{-1/2}(\Gamma_i)$; $\alpha_k \to 0$ is not necessarily true. For $\psi \in H_{00}^{1/2}(\Gamma_i)$, $\psi = \sum_{k=1}^{\infty} \beta_k \sin \pi k y$, we have using (2.15)

$$\begin{aligned} \|\psi\|^2_{H_{00}^{1/2}(\Gamma_i)} &= \sum_{k=1}^{\infty} \beta_k^2 \, \|(-\Delta_0)^{1/4}(\sin \pi k y)\|^2_{L^2(\Gamma_i)} \\ &= \frac{1}{2} \sum_{k=1}^{\infty} \beta_k^2 \pi k. \end{aligned} \tag{3.5}$$

Hence, any $\varphi = \sum_k \alpha_k \sin \pi k y$ with $\sum_{k=1}^{\infty} \frac{1}{k} \alpha_k^2 < \infty$ defines (in the spirit of (2.17)) an element of $H^{-1/2}(\Gamma_i)$ by

$$\begin{aligned} \langle \varphi, \psi \rangle_{\Gamma_i} &= \sum_{k=1}^{\infty} \alpha_k \beta_k \langle \sin \pi k y, \sin \pi k y \rangle_{\Gamma_i} \\ &= \sum_{k=1}^{\infty} \alpha_k \beta_k \| \sin \pi k x \|^2_{L^2(\Gamma_i)} = \frac{1}{2} \sum_{i=1}^{\infty} \alpha_k \beta_k \\ &\leq \frac{1}{2} \left(\sum_{k=1}^{\infty} \frac{1}{k} \alpha_k^2 \right)^{1/2} \left(\sum_{k=1}^{\infty} k \beta_k^2 \right)^{1/2} < \infty. \end{aligned}$$

A side effect of (3.5) is that such a calculation for $\| \cdot \|_{H_{00}^{1/2}(\Gamma_i)^*}$ (see def. (2.9)) shows easily that the norms $\| \cdot \|_{H_{00}^{1/2}(\Gamma_i)}$ and $\| \cdot \|_{H_{00}^{1/2}(\Gamma_i)^*}$ are equivalent. Obviously, the similar considerations hold for $H_{mv}^{1/2}(\Gamma_i)$ and $H_{mv}^{1/2}(\Gamma_i)^*$.

In Chapter 3 we will develop a numerical realization of 2.–3. with computational costs that are neglectable compared to the computational costs of the local solvers. Furthermore we will see that step 2 is very easy if there is an equidistant grid on all the interfaces but rather tricky in the case of a Chebyshev-Gauss-Lobatto grid. In fact, the main difficulty of the construction of a preconditioner is the transference of the equidistant grid preconditioners to the Gauss-Lobatto grid case.

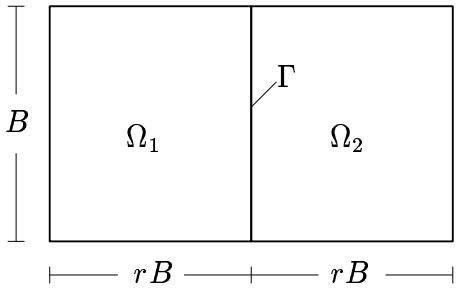
The preconditioners are constructed by examination of the non-discrete operator A. We should be aware that a discretized operator A_h (the kind of discretization depends on the local grids and the local solvers) may have completely different eigenvalues/eigenfunctions.[1] Nevertheless, as A_h is an approximation of A, the use of the eigenvalues/eigenfunctions of the non-discrete operator is justified. The efficiency of the preconditioners will be demonstrated by many numerical test runs.

Even in the case of local FEM solvers it is not compelling to consider the case of *arbitrary* grids on the interfaces. For many flow problems the finite element mesh on the subdomains can be constructed so that its restriction on the interfaces is equidistant, e.g. if the obstacle M is not hit by (or is not very close to) any interface. However, the more general case of nonequidistant meshes on the interfaces is considered in Section 3.5.

[1] Numerical tests showed that the quotient of the largest and the smallest eigenvalue of the discretization of A using $N+1$ grid points on each interface is $O(N^2)$ for the spectral solver whereas it is $O(N)$ for the FD solver. This result corresponds to the behaviour of the spectral solver applied to the Laplacian equation (see [12], Table 4.1, p. 100) having a condition number of $O(N^4)$: We already know that the square root $(-\Delta_0)^{1/2}$, $(-\Delta_{Nm})^{1/2}$ of the related operator is an approximation of A.

3.1.2 The case of $p=2$ subdomains and equidistant boundary mesh

In this section we will show a way to find out the eigenvalues and eigenfunctions of the operator (3.1) in the simple case of only 2 subdomains.



Of course, this is just an exemplary case. In Section 3.1.4 the more relevant (but more difficult) case of arbitrary numbers of subdomains is handled.

At first we consider the Poisson equation (i.e. $\sigma = 0$ in (2.1)) with pure Dirichlet boundary conditions.

Due to the symmetry of the domain and the boundary conditions we have

$$[u(\varphi)] = 2\, u^1(\varphi)|_\Gamma$$

where u^1 is the restriction of u on the left subdomain Ω_1 and $u^1|_\Gamma$ is the trace of u^1 on the interface Γ.

Starting with a product approach $u^1(x,y) = v(x)\, w(y)$ and taking into account the boundary conditions on the left ($x=0$), the upper ($y=B$) and the lower ($y=0$) boundary of Ω_1, we get solutions

$$u_k(x,y) = \sinh\frac{\pi k x}{B}\, \sin\frac{\pi k y}{B}, \quad k \in \mathbb{N}, \tag{3.6}$$

for u^1. We observe that

$$\frac{\partial u_k}{\partial x}|_\Gamma(y) = \frac{\pi k}{B}\, \cosh \pi k r\, \sin\frac{\pi k y}{B} \tag{3.7}$$

and

$$[u_k] = 2\, u_k^1|_\Gamma(y) = 2\, \sinh \pi k r\, \sin\frac{\pi k y}{B} \tag{3.8}$$

where r is the aspect ratio of the subdomains. A comparison of (3.7) and (3.8) shows that $\sin\frac{\pi k y}{B}$ are eigenfunctions of (3.1), and it also gives the eigenvalues as the quotient of the coefficients of the sine functions:

$$\lambda_k = \frac{2B}{\pi k} \tanh \pi k r \tag{3.9}$$

If we consider the more general case $\sigma \geq 0$ in (2.1) we get

$$\begin{aligned} u_k(x,y) &= \sinh x\sqrt{\sigma + \frac{\pi^2 k^2}{B^2}}\ \sin\frac{\pi k y}{B}, \\ \frac{\partial u_k}{\partial x}|_\Gamma(y) &= \frac{c_k}{B}\cosh c_k r\ \sin\frac{\pi k y}{B}, \\ u_k|_\Gamma(y) &= \sinh c_k r\ \sin\frac{\pi k y}{B}, \\ \lambda_k &= \frac{2B}{c_k}\ \tanh r c_k \end{aligned}$$

with

$$c_k := \sqrt{\sigma B^2 + \pi^2 k^2}$$

instead of (3.6) - (3.9) which means that the eigenfunctions are the same, but the eigenvalues depend on σ now.

Similar calculations for different boundary conditions result in the following lemma:

Lemma 3.1 *For $p = 2$ subdomains, the eigenfunctions of (3.1) are*

$$\varphi_k(y) = \sin\frac{\pi k y}{B}, \quad k = 1, ..., \infty \tag{3.10}$$

in the case of Dirichlet boundary conditions on Γ^W and

$$\varphi_k(y) = \cos\frac{\pi k y}{B}, \quad k = 1, ..., \infty \tag{3.11}$$

in the case of Neumann boundary conditions on Γ^W, both independent of the boundary condition at Γ^I and Γ^O. The related eigenvalues are

$$\lambda_k = \frac{2\,B\ \tanh r\sqrt{\sigma B^2 + \pi^2 k^2}}{\sqrt{\sigma B^2 + \pi^2 k^2}} \tag{3.12}$$

for Dirichlet boundary conditions on $\Gamma^I \cup \Gamma^O$ and

$$\lambda_k = \frac{2\,B\ \coth r\sqrt{\sigma B^2 + \pi^2 k^2}}{\sqrt{\sigma B^2 + \pi^2 k^2}} \tag{3.13}$$

for Neumann boundary conditions on $\Gamma^I \cup \Gamma^O$, both independent of the boundary condition on Γ^W.

Proof. By calculations similar to the above. ∎

Corollary 3.2 *For the preconditioner (3.4) with φ_k and λ_k from the previous lemma, $C = A^{-1}$ holds.*

Following Section 3.1.1, the preconditioning operator (3.4) can be implemented as follows:

- To get the sine rsp. cosine coefficients of φ, use the Fast Fourier Transform (FFT) on the equidistant grid values of φ after odd rsp. even prolongation of the record to the double length.
- Multiply each coefficient with λ_k^{-1} from Lemma 3.1.
- Use FFT^{-1} on these coefficients to calculate the equidistant grid values of $C\varphi$.

In lots of applications of the CG method, the costs of the preconditioner is proportional to the costs of the rest of the CG step. This is also true for the widespread FETI and dual Schur preconditioners (see Sec. 2.9 and the papers cited there). In case of CGBI, the amount of work for the preconditioning is only $O(N \log N)$ which is neglectable compared to the $O(N^2)$ operations for a FDM local solver on a subdomain with $N \times N$ grid points.

In the next section, the preconditioner is applied to the Gauss-Lobatto grid. As a preparation, we will derive a simplification of the preconditioner (3.4):

Lemma 3.3 *Let $(\varphi_k)_{k\in\mathbb{N}}$ be a complete orthonormal system of a subspace H of $L^2(\Gamma)$. Let $A, B : H \to H^*$ be two operators defined by*

$$A : \varphi = \sum_{k=1}^{\infty} \alpha_k \varphi_k \quad \longmapsto \quad A\varphi = \sum_{k=1}^{\infty} a_k \alpha_k \varphi_k,$$
$$B : \varphi = \sum_{k=1}^{\infty} \alpha_k \varphi_k \quad \longmapsto \quad B\varphi = \sum_{k=1}^{\infty} b_k \alpha_k \varphi_k,$$

$a_k, b_k > 0$. If there are constants $c_1, c_2 > 0$ with

$$c_1 \le \frac{a_k}{b_k} \le c_2$$

for all $k \in \mathbb{N}$, then the norms $\langle A\varphi, \varphi\rangle_{H^*,H}$ and $\langle B\varphi, \varphi\rangle_{H^*,H}$ are equivalent on H.

Proof. From the assumptions we get easily

$$\langle A\varphi, \varphi\rangle_{H^*,H} = \sum_{k=1}^{\infty} \alpha_k^2 a_k,$$
$$\langle B\varphi, \varphi\rangle_{H^*,H} = \sum_{k=1}^{\infty} \alpha_k^2 b_k,$$

and therefore

$$\langle A\varphi, \varphi\rangle_{H^*,H} \le c_2 \langle B\varphi, \varphi\rangle_{H^*,H},$$
$$\langle A\varphi, \varphi\rangle_{H^*,H} \ge c_1 \langle B\varphi, \varphi\rangle_{H^*,H}.$$

■

The lemma leads to the following idea: If we replace $1/\lambda_k$ by a (simpler) expression $g_k > 0$ in the definition of the preconditioner (3.4) with

$$c_1 \leq \lambda_k \, g_k \leq c_2, \quad , k = 1, ..., \infty, \tag{3.14}$$

c_1, c_2 independent of k, then the new preconditioner

$$\tilde{C} : R(A) \longrightarrow D(A), \quad \varphi = \sum_{k=1}^{\infty} \alpha_k \, \varphi_k \longmapsto \tilde{C}\varphi = \sum_{k=1}^{\infty} \alpha_k \, g_k \, \varphi_k \tag{3.15}$$

generates a norm $\langle \tilde{C}\cdot, \cdot \rangle_\Gamma$ which is *equivalent* to the norm $\langle C\cdot, \cdot \rangle_\Gamma$. This method can be used to approximate the preconditioner C by a simpler one: A linear combination of $(-\Delta_0)^{1/2}$ rsp. $(-\Delta_{Nm})^{1/2}$ and the identity operator, because for large k, λ_k^{-1} in (3.12), (3.13) behaves like a linear function in k. If we find a linear function $g_k = \alpha k + \beta$ fulfilling (3.14), then we can approximate C by

$$\tilde{C} = \alpha \frac{B}{\pi} (-\Delta_0)^{1/2} + \beta \, id$$

without losing the property of a condition number bounded independent of N.

To find suitable α and β, we regard $\lambda_k = \lambda(k)$ as a function of $\mathbb{R}^+$ and postulate that

$$\lim_{k\to\infty} \frac{g(k)}{k} = \lim_{k\to\infty} \frac{(\lambda(k))^{-1}}{k} \text{ and } \lim_{k\to 0} g(k) = \lim_{k\to 0} (\lambda(k))^{-1}, \tag{3.16}$$

i.e.

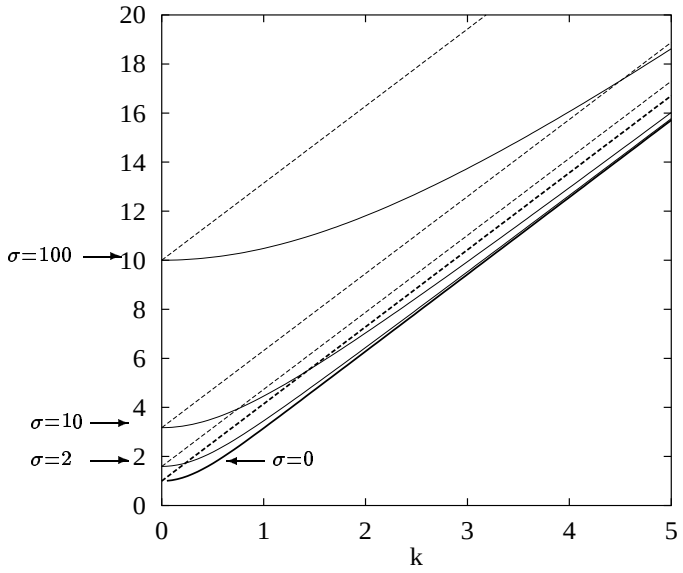


Figure 3.1: $1/\lambda(k)$ (full lines) and its approximation $g(k)$ (dotted lines) in the Dirichlet case for $B=1$, $r=1$ and $\sigma = 0, 2, 10, 100$.

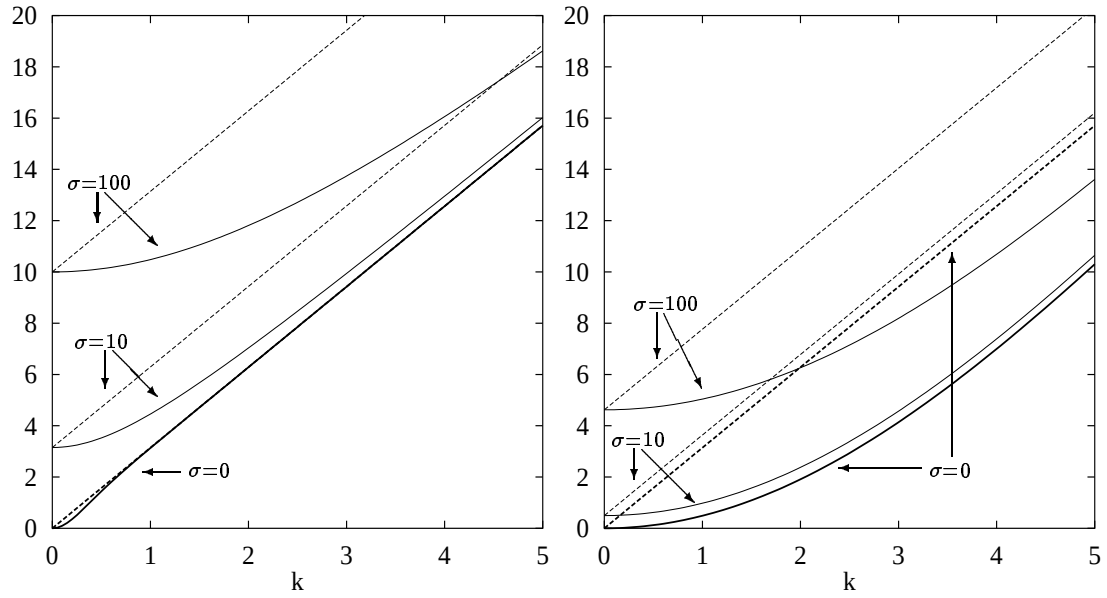


Figure 3.2: $1/\lambda(k)$ (full lines) and its approximation $g(k)$ (dotted lines) in the Neumann case for $B=1$ and $\sigma = 0, 10, 100$. On the left, $r=1$, on the right, $r=0.05$. If $\sigma=0$, the approximation becomes worse for small values of r because $1/\lambda(k)$ behaves like a parabola which is approximated by a straight line.

$$\alpha := \lim_{k\to\infty} \frac{1}{k\,\lambda(k)}, \quad \beta := \lim_{k\to 0} \frac{1}{\lambda(k)}.$$

For Dirichlet boundary conditions on $\Gamma^I \cup \Gamma^O$ we get from (3.12)

$$\begin{aligned} \alpha &= \frac{\pi}{2B}, \\ \beta &= \frac{\sqrt{\sigma}}{2\,\tanh rB\sqrt{\sigma}} \quad \text{for } \sigma > 0, \\ \beta &= \frac{1}{2rB} \quad \text{for } \sigma = 0 \end{aligned}$$

(which coincides with the limit of the result for $\sigma > 0$, $\sigma \to 0$). So we get

$$g_k := \frac{\pi}{2B}\,k + \frac{\sqrt{\sigma}}{2\,\tanh rB\sqrt{\sigma}} \tag{3.17}$$

with obvious modifications for $\sigma = 0$. For Neumann boundary conditions on $\Gamma^I \cup \Gamma^O$ we get from (3.13)

$$\begin{aligned} \alpha &= \frac{\pi}{2B}, \\ \beta &= \frac{\sqrt{\sigma}}{2}\,\tanh rB\sqrt{\sigma}. \end{aligned}$$

So we get

$$g_k := \frac{\pi}{2B}\, k + \frac{\sqrt{\sigma}}{2} \tanh rB\sqrt{\sigma}. \tag{3.18}$$

So we arrive at the following lemma which additionaly estimates the quality of the approximation by giving an estimate for the ratio c_2/c_1 in (2.42):

Lemma 3.4 *Let us set $(-\Delta)^{1/2} := (-\Delta_0)^{1/2}$ in the case of Dirichlet boundary conditions on Γ^W and $(-\Delta)^{1/2} := (-\Delta_{Nm})^{1/2}$ in the case of Dirichlet boundary conditions on Γ^W.*

The norms generated by $C = A^{-1}$ (see (3.4)) and by $\tilde{C}$,

$$\begin{aligned} \tilde{C} &:= (-\Delta)^{1/2} + \frac{\sqrt{\sigma}}{\tanh rB\sqrt{\sigma}}\, id, \quad \text{if } \sigma > 0, \\ \tilde{C} &:= (-\Delta)^{1/2} + \frac{1}{rB}\, id, \quad \text{if } \sigma = 0, \end{aligned}$$

in the case of Dirichlet boundary conditions on $\Gamma^I \cup \Gamma^O$ (='Dirichlet case') and

$$\tilde{C} := (-\Delta)^{1/2} + \sqrt{\sigma}\, \tanh rB\sqrt{\sigma}\, id$$

in the case of Neumann boundary conditions on $\Gamma^I \cup \Gamma^O$ (='Neumann case') are equivalent, i.e.

$$c_1 \langle C\varphi, \varphi\rangle_\Gamma \le \langle \tilde{C}\varphi, \varphi\rangle_\Gamma \le c_2 \langle C\varphi, \varphi\rangle_\Gamma$$

holds. $c_2/c_1 = 28/9$ can be found independent of $\sigma \ge 0$, $r > 0$, $\varphi \in H_{00}^{1/2}(\Gamma)$ in the Dirichlet case. In the Neumann case, c_2/c_1 can be found independent of $\varphi \in H_{mv}^{1/2}(\Gamma)$, but not independent of $\sigma, r \to 0$.

Remark. The preconditioner $\tilde{C}$ of Lemma 3.4 is very similar to the preconditioner $(-\Delta)^{1/2}$ of Chapter 2. In fact, the norms generated by them are equivalent.

Proof.
(a) The Dirichlet case.
(i) The upper bound. From (3.12), (3.17) we conclude that $\frac{1}{2}\tilde{C}A$ multiplies each eigenfunction (3.10) rsp. (3.11) by

$$g(k)\,\lambda(k) = \left(\pi k + \frac{B\sqrt{\sigma}}{\tanh rB\sqrt{\sigma}}\right) \frac{\tanh r\sqrt{\sigma B^2 + \pi^2 k^2}}{\sqrt{\sigma B^2 + \pi^2 k^2}}$$

(with obvious modifications for $\sigma = 0$) for which we have to find bounds. For the hyperbolic tangent we will use the estimates

$$\frac{3}{4} \le \tanh x \le 1 \quad \text{for } x \ge 1, \tag{3.19}$$

$$\frac{3}{4}\, x \le \tanh x \le x \quad \text{for } |x| \le 1, \tag{3.20}$$

$$\tanh x \le \min\{1, x\} \quad \text{for } x \ge 0. \tag{3.21}$$

If $rB\sqrt{\sigma} \geq 1$, we get using (3.19)

$$g(k)\ \lambda(k) \leq \frac{\pi k + \frac{4}{3} B\sqrt{\sigma}}{\sqrt{\sigma B^2 + \pi^2 k^2}}.$$

Using the fact that for $\alpha \geq 0$, $\beta > 0$,

$$\sup_{k \in \mathbb{R}^+} \frac{\alpha + \pi k}{\sqrt{\beta^2 + \pi^2 k^2}} = \sqrt{1 + \frac{\alpha^2}{\beta^2}}, \tag{3.22}$$

we arrive at

$$g(k)\ \lambda(k) \leq \frac{5}{3}.$$

If $rB\sqrt{\sigma} \leq 1$, the application of (3.20) and (3.21) on $g(k)\ \lambda(k)$ gives

$$\begin{aligned} g(k)\ \lambda(k) &\leq \left(\pi k + \frac{4}{3r}\right) \frac{\tanh r\sqrt{\sigma B^2 + \pi^2 k^2}}{\sqrt{\sigma B^2 + \pi^2 k^2}} \\ &\leq \left(\pi k + \frac{4}{3r}\right) \min\left\{r, \frac{1}{\sqrt{\sigma B^2 + \pi^2 k^2}}\right\} \\ &= \min\left\{\pi k r + \frac{4}{3}, \frac{\pi k + \frac{4}{3r}}{\sqrt{\sigma B^2 + \pi^2 k^2}}\right\}. \end{aligned}$$

If $\pi k r \leq 1$, we get

$$g(k)\ \lambda(k) \leq \frac{7}{3}.$$

Otherwise, we get

$$g(k)\ \lambda(k) \leq \frac{\pi k\ \left(1 + \frac{4}{3}\right)}{\sqrt{\sigma B^2 + \pi^2 k^2}} \leq \frac{7}{3}.$$

(ii) The lower bound. For $r\sqrt{\sigma B^2 + \pi^2 k^2} \geq 1$ we have $\tanh r\sqrt{\sigma B^2 + \pi^2 k^2} \geq \frac{3}{4}$. Therefore

$$g(k)\ \lambda(k) \geq \frac{3}{4} \cdot \frac{(\pi k + B\sqrt{\sigma})}{\sqrt{\sigma B^2 + \pi^2 k^2}}.$$

Using

$$\inf_{k \in \mathbb{R}^+} \frac{\alpha + \pi k}{\sqrt{\beta^2 + \pi^2 k^2}} = \min\left\{1, \frac{\alpha}{\beta}\right\}, \tag{3.23}$$

we get

$$g(k)\ \lambda(k) \geq 3/4.$$

For $r\sqrt{\sigma B^2+\pi^2k^2} \le 1$ we have $\tanh r\sqrt{\sigma B^2+\pi^2k^2} \ge \frac{3}{4}\ r\sqrt{\sigma B^2+\pi^2k^2}$ and $\tanh rB\sqrt{\sigma} \le rB\sqrt{\sigma}$. Therefore

$$g(k)\ \lambda(k) \ge \left(\pi k + \frac{1}{r}\right)\ \frac{3}{4}\, r \ge \frac{3}{4}.$$

So we have $c_2/c_1 = 28/9$ independent of r, B, $\sigma > 0$.

Due to continuity, these bounds also hold for $\sigma = 0$.

(b) The Neumann case, $\sigma = 0$.

In this case we get from (3.13), (3.18)

$$g(k)\ \lambda(k) = \frac{1}{\tanh r\pi k}$$

and therefore

$$\inf_{k\in\mathbb{N}} g(k)\ \lambda(k) = 1, \quad \sup_{k\in\mathbb{N}} g(k)\ \lambda(k) = \frac{1}{\tanh \pi r}$$

which is unbounded for $r \to 0$.

(c) The Neumann case, $\sigma > 0$.

(i) The upper bound. From (3.13) and (3.18) we get

$$g(k)\ \lambda(k) = \frac{\pi k + B\sqrt{\sigma}\tanh rB\sqrt{\sigma}}{\sqrt{\sigma B^2+\pi^2k^2}\ \tanh r\sqrt{\sigma B^2+\pi^2k^2}}.$$

If $r\sqrt{\sigma B^2+\pi^2k^2} \ge 1$, using estimate (3.19) on the hyperbolic tangent and (3.22) we get

$$g(k)\ \lambda(k) \le \frac{\pi k + B\sqrt{\sigma}}{\sqrt{\sigma B^2+\pi^2k^2}} \cdot \frac{4}{3} \le \frac{4}{3}\sqrt{2}.$$

If $r\sqrt{\sigma B^2+\pi^2k^2} \le 1$, using the estimate (3.20) for the hyperbolic tangent and the fact that

$$\sup_{k\in\mathbb{R}^+} \frac{\pi k+\alpha}{\pi^2k^2+\beta^2} = \frac{\sqrt{\alpha^2+\beta^2}+\alpha}{2\,\beta^2} \tag{3.24}$$

we get

$$\begin{aligned} g(k)\ \lambda(k) &\le \frac{\pi k + r\sigma B^2}{r(\sigma B^2+\pi^2k^2)} \cdot \frac{4}{3} \le \frac{2}{3}\ \frac{\sqrt{r^2\sigma^2B^4+\sigma B^2}+r}{r\sigma B^2} \\ &= \frac{2}{3}\ \left(\sqrt{1+\frac{1}{r^2\sigma B^2}}+\frac{1}{\sigma B^2}\right) \end{aligned}$$

Remembering the result for $\sigma = 0$, we could not expect this bound being independent of σ.

(ii) The lower bound. If $rB\sqrt{\sigma} \geq 1$ we have, using (3.19), (3.23),

$$g(k)\,\lambda(k) \geq \frac{\pi k + \frac{3}{4}B\sqrt{\sigma}}{\sqrt{\sigma B^2 + \pi^2 k^2}} \geq \min\{1, \frac{3}{4}\} = \frac{3}{4}.$$

In the same manner we proceed in the case $rB\sqrt{\sigma} \leq 1$

$$g(k)\,\lambda(k) \geq \frac{\pi k + \frac{3}{4}r\sigma B^2}{\sqrt{\sigma B^2 + \pi^2 k^2}} \geq \min\left\{1, \frac{3}{4}rB\sqrt{\sigma}\right\} = \frac{3}{4}rB\sqrt{\sigma}.$$ ∎

3.1.3 The case of $p = 2$ subdomains and Gauss-Lobatto grid, application of the interpolation theory of Hilbert spaces

In Section 3.1.2 a preconditioner C for an equidistant grid on the interface was developed. Unfortunately, this preconditioner cannot be applied directly in the case of a non-equidistant grid because the discrete Fourier transform requires the *equidistant* function values of φ to get the sine rsp. cosine coefficients. (FFT applied to the Gauss-Lobatto grid values of φ gives the *Chebyshev* coefficients of a polynomial φ instead of the cosine coefficients.)

One might think that *interpolation* from the Gauss-Lobatto grid to an equidistant grid can solve the problem. In Section 3.2 this idea is investigated and the problem that arises is discussed.

Beside interpolation, there is a more elegant way of solving the problem, and this is one of the main results of this paper: Whereas the decomposition of φ given on a Chebyshev-Gauss-Lobatto mesh into a trigonometric series seems to be difficult, the decomposition of $\varphi \circ \cos$ into such a series is easy, as the Gauss-Lobatto mesh points of φ *coincide* with the equidistant mesh points of $\varphi \circ \cos$![2] That means that the numerical implementation of $(-\Delta)^{1/2}(\varphi \circ \cos)$ and $\tilde{C}(\varphi \circ \cos)$ is easy. So we have to find a relation between $(-\Delta)^{1/2}\varphi$ and $(-\Delta)^{1/2}(\varphi \circ \cos)$. To be more precise: We have to express the norm $\langle (-\Delta)^{1/2}\varphi, \varphi\rangle_\Gamma^{1/2}$ in terms of an *equivalent* norm depending on $\varphi \circ \cos$.

In the case of Dirichlet boundary conditions on Γ^W we succeed by proving that the norms generated by $(-\Delta_0)^{1/2}(\cdot)$ and by $(-\Delta_0)^{1/2}(\cdot \circ \cos)$ are equivalent (see Section 3.1.3.2). The Neumann case, however, seems to be more complicated (see Section 3.1.3.3). In both cases we use the interpolation theory of weighted Sobolev spaces.

[2] For the moment let us consider $\Gamma = (-1, 1)$. For $\Gamma = (0, B)$ we replace cos by cõs from Theorem 3.9.

3.1.3.1 Some results of the interpolation theory of weighted Sobolev spaces

Definitions. Let $\Omega \subset \mathbb{R}^n$ be a C^∞-domain. Let $s \in \mathbb{R}^{\geq 0}$. Let w_1, w_2 be two positive continuous functions on Ω. If $s=0$, let us assume that $w_1 = w_2$. Let us decompose $s = \lfloor s \rfloor + \{s\}$ where $\lfloor s \rfloor \in \mathbb{N}_0$, $\{s\} \in [0,1)$.

Let us define the norm

$$\|\varphi\|^2_{H^s_{w_1,w_2}(\Omega)} := \sum_{|\alpha|=s} \int_\Omega w_1 \, |D^\alpha \varphi|^2 \, dx + \int_\Omega w_2 \, |\varphi|^2 \, dx \tag{3.25}$$

if $s \in \mathbb{N}_0$ and

$$\begin{aligned} \|\varphi\|^2_{H^s_{w_1,w_2}(\Omega)} &:= \sum_{|\alpha|=\lfloor s \rfloor} \int_\Omega \int_\Omega \frac{|w_1(x)^{1/2} D^\alpha \varphi(x) - w_1(y)^{1/2} D^\alpha \varphi(y)|^2}{|x-y|^{n+2\{s\}}} \, dx \, dy \\ &\quad + \|\varphi\|^2_{H^{\lfloor s \rfloor}_{w_1,w_2}} \end{aligned} \tag{3.26}$$

for non-integer s.

Let

$$\begin{aligned} H^s_{w_1,w_2}(\Omega) &:= \text{closure}\{C^\infty(\Omega)\}, \\ \overset{\circ}{H}{}^s{}_{w_1,w_2}(\Omega) &:= \text{closure}\{C_0^\infty(\Omega)\} \end{aligned} \tag{3.27}$$

with respect to the norm $\|\cdot\|_{H^s_{w_1,w_2}(\Omega)}$ in the space $H^s_{loc}(\Omega)$. We are going to write $H^s_{w_1,w_2}$ for $H^s_{w_1,w_2}(\Omega)$ if no confusion arises. Additionally, we set $L^2_w := H^2_{w,w}$. For $\varphi \in \overset{\circ}{H}{}^s{}_{w^1,w^2}(\Omega)$ we may use the notation $\|\varphi\|_{\overset{\circ}{H}{}^s{}_{w^1,w^2}(\Omega)} := \|\varphi\|_{H^s_{w^1,w^2}(\Omega)}$.

Remark. Obviously, $H^1_{1,1} = H^1$, $\overset{\circ}{H}{}^1{}_{1,1} = H^1_0$, $H^0_{1,1} = \overset{\circ}{H}{}^0{}_{1,1} = L^2$ holds.

Also for $s = 1/2$, $w_1 = w_2 = 1$, the definition (3.27) is compatible with the previous definition of $H^{1/2}$ and $H^{1/2}_{00}$ in Chapter 2:

Lemma 3.5 *a)* $H^{1/2}_{1,1} = H^{1/2} = \overset{\circ}{H}{}^{1/2}{}_{1,1}$,

b) $H^{1/2}_{00}(0,\pi) = \overset{\circ}{H}{}^{1/2}{}_{1,w}(0,\pi)$ *where* $w = 1/\sin$.

Proof. **a)** For $H^{1/2}_{1,1} = H^{1/2}$ and $\overset{\circ}{H}{}^{1/2}{}_{1,1} = H^{1/2}_0$ see [34] Chapter 1 Theorem 10.2 and Remark 10.5 (p. 52). Then use $H^{1/2}_0 = H^{1/2}$ from Theorem 11.1 in [34].
b) See (2.13). ∎

Lemma 3.6 *Let $A_1 \subset A_0$, $B_1 \subset B_0$ be two couples of Hilbert spaces with continuous injections. Let $\|\cdot\|_{\mathcal{L}(A_i,B_i)}$ be the operator norm in the space of linear continuous mappings from A_i to B_i.*

a) Let $I : A_0 \to B_0$ be a continuous injection of A_0 into B_0 and of A_1 into B_1, as well. Then, for $0<\Theta<1$, $A_\Theta := [A_0, A_1]_\Theta$ is a subspace of $B_\Theta := [B_0, B_1]_\Theta$ with continuous injection I and

$$\|I\|_{\mathcal{L}(A_\Theta, B_\Theta)} \le \|I\|^{1-\Theta}_{\mathcal{L}(A_0,B_0)} \|I\|^{\Theta}_{\mathcal{L}(A_1,B_1)}. \tag{3.28}$$

b) If $A_0 = B_0$ and $A_1 = B_1$ in the sense of equivalent norms, then

$$[A_0, A_1]_\Theta = [B_0, B_1]_\Theta$$

holds (in the sense of equivalent norms).

Proof. ad a). We apply the Calderon-Lions interpolation theorem (see e.g. [26] Theorem IX.20). Therefore, we have to construct an analytic, uniformly bounded, continuous $\mathcal{L}(A_0, B_0)$-valued function T on the strip $[0,1] \times i\,\mathbb{R} \subset \mathbb{C}$. We choose $T(z)(\varphi) := I(\varphi)$ which is constant in z. In order to apply the Calderon-Lions theorem, we have to check that

$$M_0 := \sup_{y\in\mathbb{R}} \|T(iy)\|_{\mathcal{L}(A_0,B_0)} < \infty, \quad M_1 := \sup_{y\in\mathbb{R}} \|T(1+iy)\|_{\mathcal{L}(A_1,B_1)} < \infty :$$

Obviously, $M_0 = \|I\|_{\mathcal{L}(A_0,B_0)} < \infty$, $M_1 = \|I\|_{\mathcal{L}(A_1,B_1)} < \infty$. Now the theorem yields

$$\|T(\Theta)\|_{\mathcal{L}(A_\Theta,B_\Theta)} \le \|T(\Theta)\|^{1-\Theta}_{\mathcal{L}(A_0,B_0)} \|T(\Theta)\|^{\Theta}_{\mathcal{L}(A_1,B_1)},$$

i.e. (3.28) holds.
ad b). Twice repeated application of a) yields b).
Another possibility to prove b) is given in [34] (see Chapter 1, Remark 2.3) ∎

Lemma 3.7 *Let us consider a positive weight function $w \in C^\infty(\Gamma_i)$ with*

$$c_1\ dist(\partial\Gamma_i, x) \le (w(x))^{-1} \le c_2\ dist(\partial\Gamma_i, x), \quad c_1, c_2 > 0. \tag{3.29}$$

Let $s_1, s_2 \ge 0$, $s_1 \ne s_2$, $\nu_i \ge \mu_i + 2s_i$, $i=1,2$ such that

$$(\mu_1 - \nu_1)\, s_2 = (\mu_2 - \nu_2)\, s_1.$$

Let

$$s := (1-\Theta)\, s_1 + \Theta s_2, \quad \nu := (1-\Theta)\, \nu_1 + \Theta\nu_2.$$

If s is an integer then let us assume that s_1, s_2 are integers, too. For $i=1,2$ with $s_i > 0$, let μ be defined by

$$\frac{\mu-\nu}{s} = \frac{\mu_i - \nu_i}{s_i}.$$

Then

$$[\overset{\circ}{H}{}^{s_1}{}_{w^{\mu_1},w^{\nu_1}}, \overset{\circ}{H}{}^{s_2}{}_{w^{\mu_2},w^{\nu_2}}]_\Theta = \overset{\circ}{H}{}^{s}{}_{w^{\mu},w^{\nu}} \tag{3.30}$$

holds in the sense of equivalent norms.

Proof. The assumption follows directly from Theorem 3.4.2, (a) and (f) on page 275f. in [54]. (We restrict ourselves to the Hilbert space case $p_1 = p_2 = 2$.) The Definition 3.2.3.2 [54] (p. 251) of the involved spaces is different from ours ((3.25)-(3.27)). However, due to Theorem 3.2.4.2 (p. 254 in [54]) and Theorem 3.2.4.1 (p. 253 in [54]) and (11) on p. 252, both definitions lead to equivalent norms and the same spaces. ∎

In the last lemma, the restriction $\nu_i \geq \mu_i + 2s_i$ was essential. The following lemma deals with smaller ν_i:

Lemma 3.8 *Let us consider a positive weight function $w \in C^\infty(\Gamma_i)$ of type (3.29). Let $s_1, s_2 \geq 0$, $s_1 \neq s_2$, $\nu_i < \mu_i + 2s_i$, $i = 1, 2$. In the case $s_i = 0$, $\mu_i = \nu_i$ may be allowed. Let $\{s_i\} \neq 1/2$ and*

$$\mu_i \neq 1 - 2\,s_i + 2\,k_i \quad \textit{where } k_i = 0, ..., [s_i] - 1,$$

for $i = 1, 2$. Let

$$s := (1-\Theta)\,s_1 + \Theta s_2, \quad \mu := (1-\Theta)\,\mu_1 + \Theta\mu_2.$$

If s is an integer then let us assume that s_1, s_2 are integers, too. Then

$$[\overset{\circ}{H}{}^{s_1}{}_{w^{\mu_1}, w^{\nu_1}}, \overset{\circ}{H}{}^{s_2}{}_{w^{\mu_2}, w^{\nu_2}}]_\Theta = \overset{\circ}{H}{}^{s}{}_{w^{\mu}, w^{\mu+2s}} \tag{3.31}$$

holds in the sense of equivalent norms.

Proof. The assumption follows from Theorem 3.4.3, on page 277 in [54]. (Again, we restrict ourselves to the Hilbert space case $p_1 = p_2 = 2$.) If s is an integer we apply (b)/(5) in [54], otherwise (a).

The case $s_i = 0$ (i.e. $\mu_i = \nu_i$ due to Def. 3.2.6 (p. 262/263 [54])) is not explicitely mentioned in Theorem 3.4.3 [54]. However, the proof of Theorem 3.4.3 still holds in this case, as equation (6) in [54] is still true.[3] ∎

3.1.3.2 The Dirichlet case

Now we are able to state the main result of this section (one of the main results of this paper):

Theorem 3.9 *a) Defining the transformation*

$$\tilde{\cos} : [0, \pi] \longrightarrow [0, B], \quad y \longmapsto \frac{B}{2}\,(1 + \cos y),$$

[3] In our case, equation (6) just means that the closure of $C^\infty(\bar{\Gamma}_i)$ and of $C_0^\infty(\Gamma_i)$ with respect to the weighted L^2-norm $|\varphi|^2 := \int_{\Gamma_i} \varphi^2\, w^{\nu_i}\, dx$ are identical.

the equality

$$\{\varphi : [0,B] \to \mathbb{R} \mid \varphi \circ \tilde{\cos} \in H_{00}^{1/2}(0,\pi)\} = H_{00}^{1/2}(0,B) \tag{3.32}$$

holds with equivalent norms

$$\|\varphi \circ \tilde{\cos}\|_{H_{00}^{1/2}(0,\pi)} \sim \|\varphi\|_{H_{00}^{1/2}(0,B)} \tag{3.33}$$

b) Let us define the weight function

$$w : [0,B] \longrightarrow \mathbb{R} \cup \{\infty\}, \quad y \longmapsto \left[1 - \left(\frac{2\,y}{B} - 1\right)^2\right]^{-1}. \tag{3.34}$$

The norms generated by $(-\Delta_0)^{1/2}$ *and by* C_{GL}*,*

$$C_{GL}\varphi := w^{1/2}\,(-\Delta_0)^{1/2}(\varphi \circ \tilde{\cos}) \circ \tilde{\cos}^{-1}, \tag{3.35}$$

generate equivalent norms on $H_{00}^{1/2}(0,B)$*, i.e. there are* c_1*,* $c_2 > 0$ *such that*

$$c_1\,\langle(-\Delta_0)^{1/2}\varphi,\varphi\rangle_\Gamma \le \langle C_{GL}\varphi,\varphi\rangle_\Gamma \le c_2\,\langle(-\Delta_0)^{1/2}\varphi,\varphi\rangle_\Gamma.$$

In short form:

$$\langle C_{GL}\varphi,\varphi\rangle_\Gamma \sim \langle(-\Delta_0)^{1/2}\varphi,\varphi\rangle_\Gamma$$

or

$$C_{GL} \sim (-\Delta_0)^{1/2} \quad \text{on } H_{00}^{1/2}(\Gamma).$$

Remark. If we would normalize the width of the channel onto the intervall $[-1,1]$ instead of $[0,B]$ the transformation $\tilde{\cos}$ would simplify to cos and the weight function $w^{1/2}$ in (3.35) to the well known Chebyshev weight function $(1-y^2)^{-1/2}$. But the usage of the intervall $[-1,1]$ would be less convenient for the representation of the eigenfunctions of A. However, I will use $[0,B]$ throughout this chapter.

Proof of the theorem.

ad a). Let us consider smooth functions φ taken from the space $C_0^\infty(\Gamma)$ which is dense in $H_{00}^{1/2}(\Gamma)$ (see p. 11). We express the norms generated by *id* and $-\Delta_0$ in terms of the function

$$\tilde{\varphi} := \varphi \circ \tilde{\cos}$$

by integral transformation:

$$\begin{aligned}
\|\varphi\|^2_{L^2(0,B)} &= \int_0^B \varphi^2(y)\,dy = \frac{B}{2}\int_0^\pi (\varphi(\tilde{\cos}\, y))^2 \sin y\, dy \\
&= \frac{B}{2}\,\|\tilde{\varphi}\|^2_{L^2_{\tilde{w}^{-1}}(0,\pi)} \\
(-\Delta_0\varphi,\varphi)_{L^2(0,B)} &= \|\nabla\varphi\|^2_{L^2(0,B)} = \int_0^B (\varphi'(y))^2\,dy = \frac{B}{2}\int_0^\pi (\varphi'(\tilde{\cos}\, y))^2 \sin y\, dy \\
&= \frac{2}{B}\int_0^\pi \left(\frac{d}{dy}(\varphi(\tilde{\cos}\, y))\right)^2 \frac{1}{\sin y}\,dy = \frac{2}{B}\,\|\tilde{\varphi}\|^2_{L^2_{\tilde{w}}(0,\pi)} \qquad (3.36)
\end{aligned}$$

with the weight function

$$\tilde{w}(y) := \frac{1}{\sin y}.$$

Due to (3.36) and a density argument, the mapping $\varphi \mapsto \tilde{\varphi} = \varphi \circ \cos$ is an isomorphism from $L^2(0,B)$ to $L^2_{w^{-1}}(0,\pi)$ on the one hand and from $H^1_0(0,B) = \overset{\circ}{H}{}^1{}_{1,1}(0,B)$ to $\overset{\circ}{H}{}^1{}_{\tilde{w},\tilde{w}^{-1}}(0,\pi)$ on the other hand. Applying Lemma 3.6 b) with interpolation index $1/2$ we get

$$\|\varphi\|_{[L^2(0,B),H^1_0(0,B)]_{1/2}} \sim \|\varphi\circ\tilde{\cos}\|_{[L^2_{\tilde{w}^{-1}}(0,\pi),\overset{\circ}{H}{}^1{}_{\tilde{w},\tilde{w}^{-1}}(0,\pi)]_{1/2}}.$$

Using the definition (2.6) of $H^{1/2}_{00}$ on the left hand side and Lemma 3.8 on the right hand side we get

$$\|\varphi\|_{H^{1/2}_{00}(0,B)} \sim \|\varphi\circ\tilde{\cos}\|_{\overset{\circ}{H}{}^1{}_{1,\tilde{w}}(0,\pi)}$$

The application of (2.12) yields a).

ad b). b) is a consequence of a):

Let us restrict ourselves to $\varphi \in C^\infty_0(0,B)$ at first. Using a), the definition of the norm (2.15) and an integral transformation we get

$$\begin{aligned}
\|\varphi\|_{H^{1/2}_{00}(0,B)} &\sim \|\varphi\circ\tilde{\cos}\|^2_{H^{1/2}_{00}(0,\pi)} \\
&= ((-\Delta_0)^{1/2}(\varphi\circ\tilde{\cos}),\varphi\circ\tilde{\cos})_{L^2(\Gamma)} \\
&= \int_0^\pi (-\Delta_0)^{1/2}(\varphi\circ\tilde{\cos})(y)\cdot\varphi(\tilde{\cos}\, y)\,dy \\
&= \frac{2}{B}\int_0^B w(x)^{1/2}\cdot(-\Delta_0)^{1/2}(\varphi\circ\tilde{\cos})(\tilde{\cos}^{-1}x)\cdot\varphi(x)\,dx \\
&= \frac{2}{B}\,(C_{GL}\varphi,\varphi)_{L^2(0,B)}. \qquad (3.37)
\end{aligned}$$

(3.37) and the density of C_0^∞ in $H_{00}^{1/2}$ (see Section 2.1) yield the assertion. ∎

Alternative proof of a). Using the equivalence $H_{00}^{1/2} = \overset{\circ}{H}{}^{1/2}{}_{1,w}$ (Lemma 3.5, w from (3.34)) we can apply Lemma 3.7 and get

$$H_{00}^{1/2}(0,B) = [\overset{\circ}{L}{}^2{}_w(0,B), \overset{\circ}{H}{}^1{}_{w^{-1},w}(0,B)]_{1/2}.$$

As in (3.36), an integral transformation shows easily that

$$\begin{aligned} \|\varphi\|_{L^2_{w^\alpha}(0,B)} &= c\,\|\tilde{\varphi}\|_{L^2_{\tilde{w}^{2\alpha-1}}(0,\pi)}, \\ \|\varphi'\|_{L^2_{w^\beta}(0,B)} &= c\,\|\tilde{\varphi}'\|_{L^2_{\tilde{w}^{2\beta+1}}(0,\pi)}, \end{aligned} \tag{3.38}$$

$\alpha, \beta \in \mathbb{Z}$. As in the previous version of the proof of a), we apply Lemma 3.6 b) with $\Theta = 1/2$ onto the norms (3.38) with $\alpha = 1$, $\beta = -1$ and get

$$\begin{aligned} \varphi &\in H_{00}^{1/2}(0,B) = [\overset{\circ}{L}{}^2{}_w(0,B), \overset{\circ}{H}{}^1{}_{w^{-1},w}(0,B)]_{1/2} \\ \Longleftrightarrow \varphi \circ \tilde{\cos} &\in [L^2_{\tilde{w}}(0,\pi), \overset{\circ}{H}{}^1{}_{\tilde{w}^{-1},\tilde{w}}(0,\pi)]_{1/2} \end{aligned}$$

with equivalent norms

$$\|\varphi\|_{H_{00}^{1/2}(0,B)} \sim \|\varphi \circ \tilde{\cos}\|_{H^{1/2}_{[L^2_{\tilde{w}}(0,\pi), H^1_{\tilde{w}^{-1},\tilde{w}}(0,\pi)]_{1/2}}}.$$

Application of Lemma 3.7 onto $[L^2_{\tilde{w}}(0,\pi), H^1_{\tilde{w}^{-1},\tilde{w}}(0,\pi)]_{1/2}$ yields

$$\varphi \in H_{00}^{1/2}(0,B) \Longleftrightarrow \varphi \circ \tilde{\cos} \in \overset{\circ}{H}{}^{1/2}{}_{1,\tilde{w}}(0,\pi).$$

∎

Remark. From Lemma 3.4 and the proof of Theorem 3.9 it is clear that the equivalence constants of the three norms

$$C \;\sim\; \tilde{C} = (-\Delta_0)^{1/2} + \frac{\sqrt{\sigma}}{\tanh rB\sqrt{\sigma}}\, id \;\sim\; C_{GL} + \frac{\sqrt{\sigma}}{\tanh rB\sqrt{\sigma}}\, id \tag{3.39}$$

do *not* depend on r and σ. So the right hand term in (3.39) is a suitable preconditioner for the Gauss-Lobatto case.

Numerical realization. To apply the preconditioner $C_{GL} + c\,id$ to a function φ known at the Gauss-Lobatto points on Γ, we just have to apply FFT onto the set of function values (after odd prolongation to a data set of double length). This gives the sine coefficients of $\varphi \circ \tilde{\cos}$. Then, we multiply each coefficient α_k by its index k (i.e. we apply $(-\Delta_0)^{1/2}$) and add c. Then we apply FFT^{-1}. At the end, the multiplication with the weight function w is to perform. At the end points of Γ, w is singular. At these points, we use the boundary condition $C_{GL}\varphi(x) = 0$ to gain the function values.

For test runs, see Section 3.1.5.

3.1.3.3 The Neumann case

In the case of Neumann conditions on Γ^W we have to investigate the space $\{\varphi \circ \tilde{\cos} \,|\, \varphi \in H^{1/2}_{mv}(0,B)\}$. We may reduce this problem to the question of identifying the space

$$\mathcal{H}_{Nm} := \{\varphi \circ \tilde{\cos} \,|\, \varphi \in H^{1/2}(0,B) = \overset{\circ}{H}{}^{1/2}{}_{1,1}\,(0,B)\}$$

with the norm

$$\|\varphi \circ \tilde{\cos}\| := \|\varphi\|_{H^{1/2}(0,B)}.$$

This case is more difficult to handle than the Dirichlet case because:

- The Lemmas 3.7 and 3.8 are dealing with interpolation spaces $\overset{\circ}{H}{}^{1/2}{}_{w^\mu,w^\nu}$ with $\nu \geq \mu+1$ rsp. $\nu = \mu+1$. We, however, are interested in the case $\nu = \mu = 0$.
- Leaving the spaces which are closures of C_0^∞, we may use the equivalence

$$H^{1/2} = [L^2, H^1]_{1/2}. \tag{3.40}$$

  Then we may apply the equivalences

$$\varphi \in L^2(0,B) \iff \tilde{\varphi} \in L^2_{\tilde{w}^{-1}}(0,\pi) \tag{3.41}$$
$$\varphi \in H^1(0,B) \iff \tilde{\varphi} \in H^1_{\tilde{w},\tilde{w}^{-1}}(0,\pi) \tag{3.42}$$

  (see (3.36)). But we cannot continue by applying Lemma 3.7, 3.8 as $H^1_{\tilde{w},\tilde{w}^{-1}} \neq \overset{\circ}{H}{}^1{}_{\tilde{w},\tilde{w}^{-1}}$. (The last inequality is a result of the transformation $\varphi \to \varphi \circ \tilde{\cos}$ and $H^1 \neq H^1_0$.)

- Theorems dealing with $H^{s_i}_{w_1,w_2}$- (instead of $\overset{\circ}{H}{}^{s_i}{}_{w_1,w_2}$)-spaces seem to be restricted to non-singular weight functions (Teorema 3.3 in [23]) or to cases where the derivative order of both spaces coincides ($s_1 = s_2$) (Théorème 5.4 in [27] and Teorema 3.2 in [23]).

However, the following two lemmas give an estimation of the kind

$$\|\varphi\|_{H^{1/2}(0,B)} \leq \overline{c}\, \|\varphi \circ \tilde{\cos}\|_{H^{1/2}_{\tilde{w}^0,\tilde{w}^{-1}}} \tag{3.43}$$
$$\|\varphi\|_{H^{1/2}(0,B)} \geq \underline{c}_\epsilon\, \|\varphi \circ \tilde{\cos}\|_{H^{1/2}_{\tilde{w}^{-1-\epsilon},\tilde{w}^{-\epsilon}}} \tag{3.44}$$

i.e. the inclusions

$$H^{1/2}_{1,\tilde{w}^{-1}} \subset \mathcal{H}_{Nm} \subset H^{1/2}_{\tilde{w}^{-1-\epsilon},\tilde{w}^{-\epsilon}} \tag{3.45}$$

hold with continuous injections, $\epsilon > 0$.

Lemma 3.10 *For $0<\epsilon<1$, $\mathcal{H}_{Nm} \subset H^{1/2}_{\tilde{w}^{-1-\epsilon},\tilde{w}^{-\epsilon}}$ holds with continuous injection.*

Proof. Let us follow the idea of the second item of the enumeration at the beginning of this section. For $\varphi \in L^2(0,B)$, let us define[4]

$$\hat{\varphi}(x) := \varphi(\tilde{\cos} x)\, \sin x = \tilde{\varphi}(x)\, (\tilde{w}(x))^{-1}.$$

(i) Obviously,

$$\varphi \in L^2(0,B) \quad \Longleftrightarrow \quad \tilde{\varphi} = \varphi \circ \tilde{\cos} \in L^2_{\tilde{w}^{-1}}(0,\pi) \quad \Longleftrightarrow \quad \hat{\varphi} \in L^2_{\tilde{w}^1}(0,\pi).$$

(ii) Now let $\varphi \in H^1(0,B)$, i.e.

$$\tilde{\varphi} \in H^1_{\tilde{w},\tilde{w}^{-1}}(0,\pi) \tag{3.46}$$

(see (3.42)). As a well known fact, $\varphi \in C^0([0,B])$ and $\sup|\varphi(x)| \le c\,\|\varphi\|_{H^1(0,B)}$. Then, of course, $\sup|\tilde{\varphi}(x)| \le c\,\|\varphi\|_{H^1(0,B)}$. Therefore

$$\int_0^\pi \tilde{\varphi}^2\, \tilde{w}^{1-2\epsilon}\, dx \le \sup|\tilde{\varphi}(x)|^2 \int_0^\pi \tilde{w}^{1-2\epsilon}\, dx \le c\,\|\varphi\|_{H^1(0,B)}, \tag{3.47}$$

with $c=c(\epsilon)$ independent of φ. (3.46) and (3.47) yield

$$\tilde{\varphi}' \in L^2_{\tilde{w}}(0,\pi), \qquad \tilde{\varphi} \in L^2_{\tilde{w}^{1-2\epsilon}}(0,\pi).$$

Obviously, $\hat{\varphi} \in L^2_{\tilde{w}^{3-2\epsilon}}(0,\pi)$ follows, and by using the representation $\hat{\varphi}' = \tilde{\varphi}' \sin + \tilde{\varphi} \cos$ we get $\hat{\varphi}' \in L^2_{\tilde{w}^{1-2\epsilon}}(0,\pi)$, hence $\hat{\varphi} \in H^1_{\tilde{w}^{1-2\epsilon},\tilde{w}^{3-2\epsilon}}(0,\pi)$. Remark 3.2.6.6 in [54] (p. 265) reads $H^1_{\tilde{w}^{1-2\epsilon},\tilde{w}^{3-2\epsilon}}(0,\pi) = \overset{\circ}{H}{}^1_{\tilde{w}^{1-2\epsilon},\tilde{w}^{3-3\epsilon}}(0,\pi)$.

(iii) All in all we can state that the mapping $\varphi \mapsto \hat{\varphi}$ is a continuous injection from $L^2(0,B)$ into $L^2_{\tilde{w}^1}(0,\pi)$, as well as a continuous injection from $H^1(0,B)$ into $\overset{\circ}{H}{}^1_{\tilde{w}^{1-2\epsilon},\tilde{w}^{3-3\epsilon}}(0,\pi)$. Using part a) of Lemma 3.6 we derive that $\varphi \mapsto \hat{\varphi}$ is a continuous injection from $H^{1/2}(0,B)$ into $[L^2_{\tilde{w}^1}(0,\pi), \overset{\circ}{H}{}^1_{\tilde{w}^{1-2\epsilon},\tilde{w}^{3-3\epsilon}}(0,\pi)]_{1/2}$. Applying Lemma 3.8, the last interpolation space is equivalent to $\overset{\circ}{H}{}^{1/2}_{\tilde{w}^{1-\epsilon},\tilde{w}^{2-\epsilon}}(0,\pi)$. Finally, the characterisation (3.26) of the weighted $H^{1/2}$-norms shows easily that

$$\|\hat{\varphi}\|_{\overset{\circ}{H}{}^{1/2}_{\tilde{w}^{1-\epsilon},\tilde{w}^{2-\epsilon}}(0,\pi)} = \|\tilde{\varphi}\|_{\overset{\circ}{H}{}^{1/2}_{\tilde{w}^{-1-\epsilon},\tilde{w}^{-\epsilon}}(0,\pi)},$$

so the mapping $\varphi \to \tilde{\varphi}$ is a continuous injection of $H^{1/2}(0,B)$ into $\overset{\circ}{H}{}^{1/2}_{\tilde{w}^{-1-\epsilon},\tilde{w}^{-\epsilon}}(0,\pi)$. ∎

[4] As explained in the beginning of this section, the interpolation lemmas cannot be applied on $\tilde{\varphi} = \varphi \circ \tilde{\cos}$ because $\tilde{\varphi}$ does not have zero boundary values for $\varphi \in H^1$. Instead, we will apply the interpolation on $\hat{\varphi}$ which decays at the ends of the interval $(0,B)$.

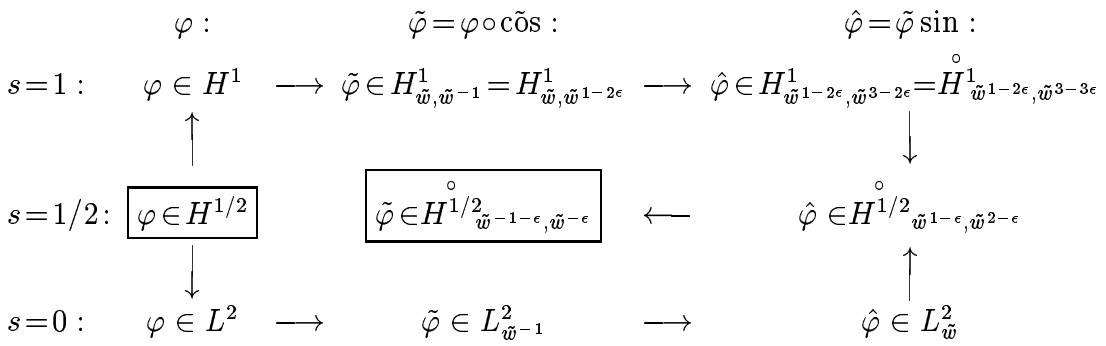


Figure 3.3: Visualisation of the proof 3.10. Vertical arrows symbolize interpolation. Horizontal arrows symbolize e.g. integral transformations like (3.36).

Lemma 3.11 *$H^{1/2}_{1,\tilde{w}^{-1}} \subset \mathcal{H}_{Nm}$ holds with continuous injection.*

Proof. We have to show that there is a constant $\overline{c}>0$ such that

$$\|\varphi\|_{H^{1/2}(0,B)} \le \overline{c}\,\|\tilde{\varphi}\|_{H^{1/2}_{1,\tilde{w}^{-1}}(0,\pi)}.$$

As $\|\varphi\|_{L^2(0,B)} = B/2\,\|\tilde{\varphi}\|_{L^2_{\tilde{w}^{-1}}(0,\pi)}$ it remains to show that

$$\int_0^B\int_0^B \left|\frac{\varphi(\xi)-\varphi(\eta)}{\xi-\eta}\right|^2 d\xi\,d\eta \le \overline{c}\int_0^\pi\int_0^\pi \left|\frac{\tilde{\varphi}(x)-\tilde{\varphi}(y)}{x-y}\right|^2 dx\,dy \tag{3.48}$$

for $\overline{c}>0$ independent of φ. Substituting $\xi=\tilde{\cos}\,x$, $\eta=\tilde{\cos}\,y$, the left hand side of (3.48) is equal to

$$\int_0^\pi\int_0^\pi \left|\frac{\tilde{\varphi}(x)-\tilde{\varphi}(y)}{\cos x-\cos y}\right|^2 \sin x\,\sin y\,dx\,dy. \tag{3.49}$$

A comparison of (3.49) with the right hand side of (3.48) shows that it is sufficient to prove that

$$F(x,y) := \left|\frac{x-y}{\cos x-\cos y}\right|^2 \sin x\,\sin y \le c \tag{3.50}$$

on $Q\setminus\bar{D}$, where $Q := [0,\pi]\times[0,\pi]$, $D := \{(x,x)\,|\,x\in(0,\pi)\}$. As F is continuous on $Q\setminus\bar{D}$, it is sufficient to show that for any sequence (x_n,y_n) converging to $(x_0,y_0)\in\bar{D}$, $F(x_n,y_n)$ is bounded.

Let $(x_0,y_0)\in D$ at first. The mean value theorem applied to the denominator in (3.50) shows that $\lim F(x_n,y_n) = 1$, then. For checking the case $(x_n,y_n)\in \bar{D}\setminus D$, it is sufficient to restrict ourselves to the case $x_0=y_0=0$. Furthermore, we

are allowed to restrict ourselves to the case $\delta_n := y_n - x_n > 0$ for all n. Applying the Taylor expansions

$$\begin{aligned}\cos(x_n+\delta_n) &= \cos x_n - \delta_n \sin x_n - \frac{\delta_n^2}{2}\cos\zeta_n \le \cos x_n - \delta_n \sin x_n \\ &\quad (\text{for } x_n+\delta_n \le \pi/2), \\ \cos(y_n-\delta_n) &= \cos y_n + \delta_n \sin y_n - \frac{\delta_n^2}{2}\cos\zeta_n^* \ge \cos y_n + \delta_n \sin y_n - \frac{\delta_n^2}{2}\end{aligned}$$

to the expression

$$F(x_n, y_n) = \frac{\delta_n^2 \sin x_n \sin y_n}{(\cos x_n - \cos(x_n+\delta_n))(\cos(y_n-\delta_n) - \cos y_n)}$$

we get

$$F(x_n, y_n) \le \frac{1}{\left(1 - \frac{\delta_n}{2\sin y_n}\right)}.$$

Due to $\delta_n < y_n$, $\limsup F(x_n, y_n) \le 2$ follows. ∎

Remark. For the proof of Lemma 3.10 it seems to be more straightforward to use (as for the proof of Lemma 3.11) the representation (3.26) of the weighted $H^{1/2}$-norm instead of the application of interpolation theorems on $\varphi \circ \tilde{\cos}\,\sin$. Indeed, it is possible to derive Lemma 3.10 using the integral representation (3.26). As before, this method is able to prove the assumption for all $\epsilon > 0$, and it fails for $\epsilon = 0$:

Alternative proof of Lemma 3.10. It has to be shown that

$$\int_0^\pi \int_0^\pi \frac{(\tilde\varphi(x)\sin^{(1+\epsilon)/2} x - \tilde\varphi(y)\sin^{(1+\epsilon)/2} y)^2}{(x-y)^2}\, dx\, dy \tag{3.51}$$

can be estimated by (3.49) times a constant plus the L^2-norm of φ times a constant. Because of the symmetry we can restrict ourselves to the integral

$$\iint_{Q_1 \cup Q_2 \cup Q_3} \frac{(\tilde\varphi(x)\sin^{(1+\epsilon)/2} x - \tilde\varphi(y)\sin^{(1+\epsilon)/2} y)^2}{(x-y)^2}\, dx\, dy \tag{3.52}$$

instead of (3.51). See Fig. 3.4 for Q_1, Q_2, Q_3.

(i) On Q_1 we use the estimate $|x-y| \ge \pi/2$ and get

$$\begin{aligned}&\iint_{Q_1} \frac{(\tilde\varphi(x)\sin^{(1+\epsilon)/2} x - \tilde\varphi(y)\sin^{(1+\epsilon)/2} y)^2}{(x-y)^2}\, dx\, dy \\ &\le 2\left(\frac{2}{\pi}\right)^2 \left(\iint_{Q_1} \sin^{1+\epsilon} x\, \tilde\varphi^2(x)\, dx\, dy + \iint_{Q_1} \sin^{1+\epsilon} y\, \tilde\varphi^2(y)\, dx\, dy \right) \\ &\le \frac{2}{\pi}\int_0^\pi \sin^{1+\epsilon} x\, \tilde\varphi^2(x)\, dx \le \frac{4}{\pi B}\, \|\varphi\|^2_{L^2(0,B)}\end{aligned}$$

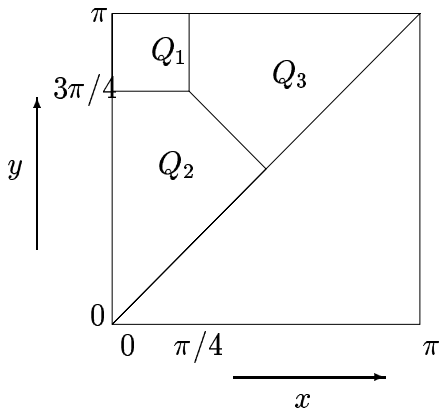


Figure 3.4: Decomposition of the square $[0,\pi]\times[0,\pi]$.

(ii) On Q_2 we have $x\le y$, $\sin x\le\sin y$. We estimate the numer in (3.52) and get

$$\iint\limits_{Q_2}\frac{(\tilde\varphi(x)\sin^{(1+\epsilon)/2}x-\tilde\varphi(y)\sin^{(1+\epsilon)/2}y)^2}{(x-y)^2}\,dx\,dy \tag{3.53}$$

$$\le 2\iint\limits_{Q_2}\frac{|\tilde\varphi(x)-\tilde\varphi(y)|^2\sin^{1+\epsilon}x}{|x-y|^2}\,dx\,dy+2\iint\limits_{Q_2}\tilde\varphi^2(y)\,\frac{|\sin^{(1+\epsilon)/2}x-\sin^{(1+\epsilon)/2}y|^2}{|x-y|^2}.$$

For $0\le y\le\frac{3\pi}{4}$,

$$\frac{\sin\limits_{0\le\xi\le y}\xi}{\sin y}\le\frac{1}{\sin\frac{3\pi}{4}}=\sqrt{2}$$

holds. Therefore, for $(x,y)\in Q_2$,

$$\frac{|\cos x-\cos y|^2}{|x-y|^2}\frac{\sin^\epsilon x}{\sin y}\le\sup_{\xi\in[x,y]}\sin^2\xi\,\frac{\sin^\epsilon x}{\sin y}=2\sin y\sin^\epsilon x\le 2$$

holds. Hence,

$$\frac{\sin^{1+\epsilon}x}{|x-y|^2}\le\frac{2\sin x\sin y}{|\cos x-\cos y|^2}.$$

Thus, the first integral on the right hand side of (3.53) is estimated by

$$2\iint\limits_{Q_2}\frac{|\tilde\varphi(x)-\tilde\varphi(y)|^2}{|\cos x-\cos y|^2}\sin x\sin y\,dx\,dy\le 2\int_0^B\int_0^B\frac{|\varphi(x)-\varphi(y)|^2}{|x-y|^2}\,dx\,dy.$$

For the last integral in (3.53) we proceed like this:

$$\frac{|\sin^{(1+\epsilon)/2}x-\sin^{(1+\epsilon)/2}y|^2}{|x-y|^2}=\frac{|\sin^{1+\epsilon}x-\sin^{1+\epsilon}y|^2}{|x-y|^2(\sin^{(1+\epsilon)/2}x+\sin^{(1+\epsilon)/2}y)^2}$$

$$\begin{aligned} &\leq \frac{(1+\epsilon)^2 \sup_{\xi\in[x,y]} \sin^{2\epsilon}\xi\,|\cos\xi|^2}{(\sin^{(1+\epsilon)/2}x+\sin^{(1+\epsilon)/2}y)^2} \leq \frac{(1+\epsilon)^2\,2\,\sin^{2\epsilon}y}{(\sin^{(1+\epsilon)/2}x+\sin^{(1+\epsilon)/2}y)^2} \\ &\leq 2\,(1+\epsilon)^2\,\frac{\sin^{2\epsilon}y}{x^{1+\epsilon}+y^{1+\epsilon}} \end{aligned}$$

So the last integral on the right hand side of (3.53) is estimated by

$$2\,(1+\epsilon)^2 \iint\limits_{Q_2} \tilde{\varphi}^2(y)\,\frac{\sin^{2\epsilon}y}{x^{1+\epsilon}+y^{1+\epsilon}}\,dx\,dy. \tag{3.54}$$

To compute this integral with respect to x, $0\leq x\leq y$, we use the representation as a geometric series:

$$\begin{aligned} &\int\limits_0^y \frac{1}{x^{1+\epsilon}+y^{1+\epsilon}}\,dx = \lim_{\bar{y}\to y}\int\limits_0^{\bar{y}} \frac{1}{y^{1+\epsilon}}\sum_{n=0}^{\infty}\left(-\left(\frac{x}{y}\right)^{1+\epsilon}\right)^n dx \\ &= \lim_{\bar{y}\to y}\frac{1}{y^{1+\epsilon}}\sum_{n=0}^{\infty}(-1)^n\,\frac{x^{(1+\epsilon)n+1}}{((1+\epsilon)n+1)\,y^{(1+\epsilon)n}}\Bigg|_{x=0}^{x=\bar{y}} \\ &\leq \lim_{\bar{y}\to y}\frac{x}{y^{1+\epsilon}}\sum_{n=0}^{\infty}\left(-\left(\frac{x}{y}\right)^{1+\epsilon}\right)^n\Bigg|_{x=0}^{x=\bar{y}} = \frac{x}{x^{1+\epsilon}+y^{1+\epsilon}}\Bigg|_{x=0}^{x=y} = \frac{1}{2y^{\epsilon}} \end{aligned} \tag{3.55}$$

Thus, (3.54) is smaller or equal

$$(1+\epsilon)^2\int\limits_0^{3\pi/4}\tilde{\varphi}^2(y)\,\sin^{\epsilon}y\,dy \leq (1+\epsilon)^2\,\frac{2}{B}\,\|\varphi\|^2_{L^2_{w^{(1-\epsilon)/2}}(0,B)}. \tag{3.56}$$

It remains to show that the last term exists for $\varphi\in H^{1/2}(0,B)$: Proceeding as in the first version of the proof of Lemma 3.10 part (ii) we know that $H^1=H^1_{1,w^{1-\epsilon}}$ and get by using the Calderon-Lions Lemma (Lemma 3.6) that

$$\varphi\in H^{1/2}=[L^2,H^1]_{1/2}=[L^2,H^1_{1,w^{1-\epsilon}}]_{1/2}\subset[L^2,L^2_{w^{1-\epsilon}}]_{1/2}=L^2_{w^{(1-\epsilon)/2}}. \tag{3.57}$$

The last step in (3.57) is a well known interpolation property which can be deduced with the K-method, for example.

(iii) On Q_3 we use

$$\begin{aligned} &\iint\limits_{Q_3}\frac{(\tilde{\varphi}(x)\sin^{(1+\epsilon)/2}x-\tilde{\varphi}(y)\sin^{(1+\epsilon)/2}y)^2}{(x-y)^2}\,dx\,dy \\ &\leq 2\iint\limits_{Q_3}\frac{|\tilde{\varphi}(x)-\tilde{\varphi}(y)|^2\,\sin^{1+\epsilon}y}{|x-y|^2}\,dx\,dy + 2\iint\limits_{Q_3}\tilde{\varphi}^2(x)\,\frac{|\sin^{(1+\epsilon)/2}x-\sin^{(1+\epsilon)/2}y|^2}{|x-y|^2}. \end{aligned}$$

instead of (3.53). Then, we proceed analogously to (ii). ∎

Let us mention that the estimates (3.51)-(3.54) also hold in the case $\epsilon=0$. However, the left hand side of (3.55) equals $\ln 2$ for $\epsilon=0$. So in (3.56), the $L^2_{w^{1/2}}(0,B)$-norm appears.

Conclusions. We have pointed out that the construction of an 'optimal' preconditioner acting on a Gauss-Lobatto boundary mesh seems to be more difficult in the Neumann case. Proceeding as in (3.37), the bound (3.43) corresponds to the possible preconditioner[5]

$$\overline{C}_{GL,Nm}\varphi := w^{1/2}\,(-\Delta_{Nm})^{1/2}(\varphi\circ\tilde{\cos})\circ\tilde{\cos}^{-1}+\varphi.$$

Numerical tests[6] suggest to orientate towards this bound (3.43) and not to (3.44); i.e. to use $\overline{C}_{GL,Nm}$ or just

$$C_{GL,Nm} := w^{1/2}\,(-\Delta_{Nm})^{1/2}(\varphi\circ\tilde{\cos})\circ\tilde{\cos}^{-1} \tag{3.58}$$

(compare the Dirichlet case (3.35)). Test runs in Section 3.1.5 show that this preconditioner produces satisfactory results. There seems to be only a slight decrease of the CGBI convergence rate if N becomes very large ($N\gg 100$).

Let us mention that there are possibilities to avoid this problem if we are ready to use $O(N^2)$ operations for the preconditioner instead of $O(N\log N)$ for (3.58):

Another possibility to construct a preconditioner is a matrix approach: The Chebyshev collocation spectral solver performs a diagonalization of the negative discrete 1d Laplacian operator $(\varphi(y_0),...,\varphi(y_N))\mapsto -(\varphi''(y_0),...,\varphi''(y_N))$ for a column of grid points of the subdomain, i.e. the eigenvalues and eigenfunctions of this discrete operator are calculated (Sec. 2.5). This information could be used to apply the square root of this operator to the boundary value function. Another starting point for the construction of a matrix-type preconditioner may be the representation of the $H^{1/2}(\Gamma)$-scalar product related to the norm representation (3.26): A coordinate transform as in Sec. 2.6 and the application of the trapezoid rule lead[7] to a matrix preconditioner.

---

[5] The other bound (3.44) would correspond to the operator $\varphi \to w^{(1-\epsilon)/4}\,(id-\Delta_{Nm})^{1/2}(\sin^{(1+\epsilon)/2}\cdot\varphi\circ\tilde{\cos})\circ(\tilde{\cos}^{-1})+w^{(1-\epsilon)/2}\,\varphi$, which could be used if combined with a projection onto the space of functions with mean value zero.

[6] As in the Dirichlet case, the weight $w^{1/2}$ is singular at the ends of Γ. At those points we use the boundary condition $\partial C_{GL,Nm}\varphi/\partial y=0$ to construct the values of $C_{GL,Nm}\varphi$.

[7] As the denominator becomes singular on the diagonal $x_k=y_k$, these values have to be replaced by a limit $(x,y)\to(x_k,y_k)$ before the quadrature rule can be applied.

3.1.4 Eigenvalues, eigenfunctions and preconditioning in the case of more than 2 subdomains

In this chapter the eigenvalues and eigenfunctions of the operator A are calculated for the case of more than two subdomains. In this case the eigenfunctions and eigenvalues are more complicated than in the case $p=2$ (Lemma 3.1). But it is still possible to give them in a (more or less) explicit way.

Then, using the technique of Lemma 3.4, we construct preconditioners of the type

$$\begin{aligned} C_{loc} &:= (-\Delta)^{1/2} + c\,id, \\ C_{glob}\varphi &:= (C_{loc}\varphi|_{\Gamma_1}, ..., C_{loc}\varphi|_{\Gamma_{p-1}}). \end{aligned} \tag{3.59}$$

(I.e. on each interface Γ_i, the boundary value function $\varphi|_{\Gamma_i}$ is treated *separately*. We will see that this preconditioner gives a condition number κ independent of N, but dependent on r with $\kappa \to \infty$ for $r \to 0$. Let us mention that a preconditioner of type (3.59) is less time consuming than the *complete* decomposition into the eigenfunctions proposed for the case $p=2$ (Cor. 3.2). Furthermore, for a preconditioner of type (3.59) it is obvious how to apply the results of Section 3.1.3 in the Gauss-Lobatto case.

The following lemma gives the eigenfunctions and eigenvalues in the case $p>2$. It is convenient to use double indices $\lambda_{k,m}$, $\varphi_{k,m}$. Instead of just verifying the equation

$$A\varphi_{k,m} = \lambda_{k,m}\varphi_{k,m}$$

the proof of the lemma shows a method how to find the $\lambda_{k,m}$, $\varphi_{k,m}$. This method might be applied to cases with different boundary conditions (e.g. Dirichlet on Γ^I and Neumann on Γ^O or vice versa (see Chapter 5) or subdomains of non-equidistant size. In the latter case, numerical methods seem to be necessary to find these values.

Theorem 3.12 *Let $p > 2$. The eigenfunctions of the operator A are:*

(i) Case Neumann b.c. on $\partial\Omega$:

$$\varphi_{k,m}|_{\Gamma_i}(y) = \sin\frac{\pi i m}{p}\,\cos\frac{\pi k y}{B} \tag{3.60}$$

(ii) Case Dirichlet b.c. on Γ^W and Neumann b.c. on $\Gamma^I \cup \Gamma^O$:

$$\varphi_{k,m}|_{\Gamma_i}(y) = \sin\frac{\pi i m}{p}\,\sin\frac{\pi k y}{B} \tag{3.61}$$

(iii) Case Neumann b.c. on Γ^W and Dirichlet b.c. on $\Gamma^I \cup \Gamma^O$:

$$\varphi_{k,m}|_{\Gamma_i}(y) = f(\gamma_{k,m}\,(i-\frac{p}{2}))\,\cos\frac{\pi k y}{B}, \quad f=\sin \text{ or } f=\cos \tag{3.62}$$

(iv) Case Dirichlet b.c. on $\partial\Omega$:

$$\varphi_{k,m}|_{\Gamma_i}(y) = f(\gamma_{k,m}\,(i-\frac{p}{2}))\,\sin\frac{\pi k y}{B}, \quad f=\sin \text{ or } f=\cos \tag{3.63}$$

$m = 1, ..., p-1$, $k = 1, ..., \infty$. The eigenvalues are

$$\lambda_{k,m} = \frac{2\,B}{\sqrt{\sigma B^2+\pi^2k^2}}\,\frac{\cosh r\sqrt{\sigma B^2+\pi^2k^2}-\cos\gamma_{k,m}}{\sinh r\sqrt{\sigma B^2+\pi^2k^2}} \tag{3.64}$$

with

$$\gamma_{k,m} = \frac{\pi m}{p}, \tag{3.65}$$

$m = 1, ..., p-1$, $k = 1, ..., \infty$, in the case of Neumann b.c. on $\Gamma^I \cup \Gamma^O$.

In the case of Dirichlet b.c. on $\Gamma^I \cup \Gamma^O$, the $\gamma_{k,m}$, $m = 1, ..., p-1$, are the $p-1$ roots within the interval $(0, \pi)$ of the equations

$$\cosh r\sqrt{\sigma B^2+\pi^2k^2}\,\sin\frac{\gamma_{k,m}p}{2} = \sin\frac{\gamma_{k,m}\,(p-2)}{2}, \tag{3.66}$$

$$\cosh r\sqrt{\sigma B^2+\pi^2k^2}\,\cos\frac{\gamma_{k,m}p}{2} = \cos\frac{\gamma_{k,m}\,(p-2)}{2}. \tag{3.67}$$

For all $\gamma_{k,m}$ being a root of (3.66), $f=\sin$ has to be taken in (3.62)-(3.63), for all $\gamma_{k,m}$ being a root of (3.67), $f=\cos$ has to be taken.

For all $k = 1, ..., \infty$

$$\min_{m=1,...,p-1}\gamma_{k,m} \le \frac{\pi}{p} \tag{3.68}$$

holds.

It is not very surprising that the results are simpler for Neumann boundary conditions (see (3.65)-(3.67)) because the inner boundary conditions on the interfaces involved in the definition of A are also of Neumann type. (3.68) is necessary to get an estimate for the condition number also in the Dirichlet case.

Proof of Theorem 3.12. Similar to Section 3.1.2 we see that functions $u(x,y) = v(x)\,w(y)$ with

$$v(x) = \sinh\frac{xK}{rB} \quad \text{rsp.} \quad v(x) = \cosh\frac{xK}{rB}$$

and

$$w(y) = \sin \frac{\pi k y}{B} \quad \text{rsp.} \quad w(y) = \cos \frac{\pi k y}{B}$$

are solutions of the partial differential equation (2.1) on each subdomain Ω_i. Here, we have used a local coordinate system with $x \in [0, rB]$, $y \in [0, B]$ and the abbreviation

$$K := r\sqrt{\sigma B^2 + \pi^2 k^2}. \tag{3.69}$$

So we use the approach

$$u(x,y) = \left(d_i \sinh \frac{xK}{rB} + s_i \cosh \frac{xK}{rB} \right) \cos \frac{\pi k y}{B} \tag{3.70}$$

on each subdomain Ω_{i+1} in the cases (i) and (iii) rsp. cos replaced by sin in (3.70) in the cases (ii) and (iv). The continuity of $\frac{\partial u}{\partial x}$ on each Γ_{i+1} gives after division by $\frac{K}{rB}$ and the sine rsp. cosine term the conditions

$$d_{i-1} \cosh K + s_{i-1} \sinh K = d_i, \quad i = 1, ..., p-1. \tag{3.71}$$

The condition that $\varphi = \frac{\partial u}{\partial x}|_\Gamma$ should be an eigenfunction of A means that $[u(\varphi)] = \lambda [\varphi]$ on all Γ_i, i.e.

$$u_i|_{x=0} - u_{i-1}|_{x=rB} = \lambda \left. \frac{\partial u_i}{\partial x} \right|_{x=0}$$

with λ *independent* of i. This yields

$$d_{i-1} \sinh K + s_{i-1} \cosh K - s_i = \lambda \frac{K}{rB} d_i, \quad i = 1, ..., p-1. \tag{3.72}$$

The boundary conditions on $\Gamma^I \cup \Gamma^O$ give

$$d_0 = 0, \quad d_{p-1} \cosh K + s_{p-1} \sinh K = 0 \tag{3.73}$$

in cases (i) and (ii) and

$$s_0 = 0, \quad d_{p-1} \sinh K + s_{p-1} \cosh K = 0 \tag{3.74}$$

in cases (iii) and (iv). Solving (3.71) for s_{i-1} we get

$$s_{i-1} = \frac{d_i - d_{i-1} \cosh K}{\sinh K}, \quad i = 1, ..., p-1. \tag{3.75}$$

Using (3.75) for s_{i-1} and s_i in (3.72) we get

$$-d_{i-1} + 2 \cosh K \, d_i - d_{i+1} = \frac{\lambda K \sinh K}{rB} d_i, \quad i = 1, ..., p-2. \tag{3.76}$$

We use (3.72) on the boundary conditions (3.73) rsp. (3.74) to eleminate s_{p-1}, then (3.75) to eleminate s_{p-2}, and we get

$$d_0 = 0, \quad 2\cosh K\, d_{p-1} - d_{p-2} = \lambda \frac{K \sinh K}{rB}\, d_{p-1} \tag{3.77}$$

in cases (i) and (ii) and

$$\begin{aligned} d_0 &= \frac{1}{\cosh K}\, d_1, \\ \left(2\cosh K - \frac{1}{\cosh K}\right) d_{p-1} - d_{p-2} &= \frac{\lambda K \sinh K}{rB}\, d_{p-1} \end{aligned} \tag{3.78}$$

in cases (iii) and (iv). So we arrive at the discrete eigenvalue problem

$$MD = \frac{\lambda K \sinh K}{rB} D \tag{3.79}$$

with the vector $D = (d_1, ..., d_{p-1})^t$ and the $(p-1) \times (p-1)$-matrix

$$M = \begin{pmatrix} 2\cosh K & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2\cosh K \end{pmatrix} \tag{3.80}$$

in the Neumann cases (i) and (ii) and

$$M = \begin{pmatrix} 2\cosh K - \frac{1}{\cosh K} & -1 & & & \\ -1 & 2\cosh K & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 2\cosh K & -1 \\ & & & -1 & 2\cosh K - \frac{1}{\cosh K} \end{pmatrix} \tag{3.81}$$

in the Dirichlet cases (iii) and (iv).[8]

In case (3.80), the discrete problem (3.79) can be solved explicitly:

$$\begin{aligned} d_i &= \sin\frac{\pi i m}{p}, \\ \frac{\lambda K \sinh K}{rB} &= 2\left(\cosh K - \cos\frac{\pi m}{p}\right), \end{aligned} \tag{3.82}$$

$m = 1, ..., p-1$ arbitrary. As already used for (3.71),

$$\varphi|_{\Gamma_i} = d_i \frac{K}{rB} \cos\frac{\pi k y}{B} \quad \text{rsp.} \quad = d_i \frac{K}{rB} \sin\frac{\pi k y}{B}$$

[8] So matrix (3.80) for the *Neumann* case (i), (ii) corresponds to a 1d boundary value problem with homogeneous *Dirichlet* conditions.

holds. So from (3.82) we get the eigenfunctions $\varphi_{k,m}$ (3.60), (3.61) and the eigenvalues $\lambda_{k,m}$ (3.64) with (3.65).

In the Neumann cases (iii) and (iv) we may use Gershgorim circles to see that M from (3.81) has all eigenvalues in the intervall $[2\cosh K - 2, 2\cosh K + 2]$. So we get

$$\lambda_{k,m} = \frac{2Br}{K} \frac{\cosh K - q_{k,m}}{\sinh K} \quad \text{with some} \quad q_{k,m} \in [-1, 1].$$

Thus, (3.64) holds with $\gamma_{k,m} \in [0, \pi]$. To find the $\gamma_{k,m}$, we use the approach

$$d_i = \sin \gamma \, (i - \frac{p}{2}), \quad i = 1, ..., p-1. \tag{3.83}$$

Obviously, for *any* $\gamma \in \mathbb{R}$ $D = (d_1, ..., d_{p-1})^t$ fulfils the equations no. 2 to $p-2$ of the linear system (3.79) with

$$\frac{\lambda K \sinh K}{rB} = 2\cosh K - 2\cos\gamma. \tag{3.84}$$

The boundary conditions (3.78) lead to the condition

$$\cosh K \sin \frac{\gamma p}{2} = \sin \frac{\gamma (p-2)}{2}. \tag{3.85}$$

With the approach

$$d_i = \cos \gamma \, (i - \frac{p}{2}), \quad i = 1, ..., p-1,$$

instead of (3.83) we get some more solutions of (3.79)/(3.81), if γ is a solution of

$$\cosh K \cos \frac{\gamma p}{2} = \cos \frac{\gamma (p-2)}{2} \tag{3.86}$$

and $\frac{\lambda K \sinh K}{rB} = 2\cosh K - 2\cos\gamma$.

The right hand sides of (3.85) and (3.86) are bounded between -1 and 1. The left hand sides of these equations take the extreme values $\pm \cosh K$ where

$$\cosh K > 1 \tag{3.87}$$

as $K \neq 0$. Between these extreme points there is always one root of the equation (see Figs. 3.5 and 3.6). Counting these extremums we find that (3.85) and (3.86) together must have $p-1$ roots within $(0, \pi)$.

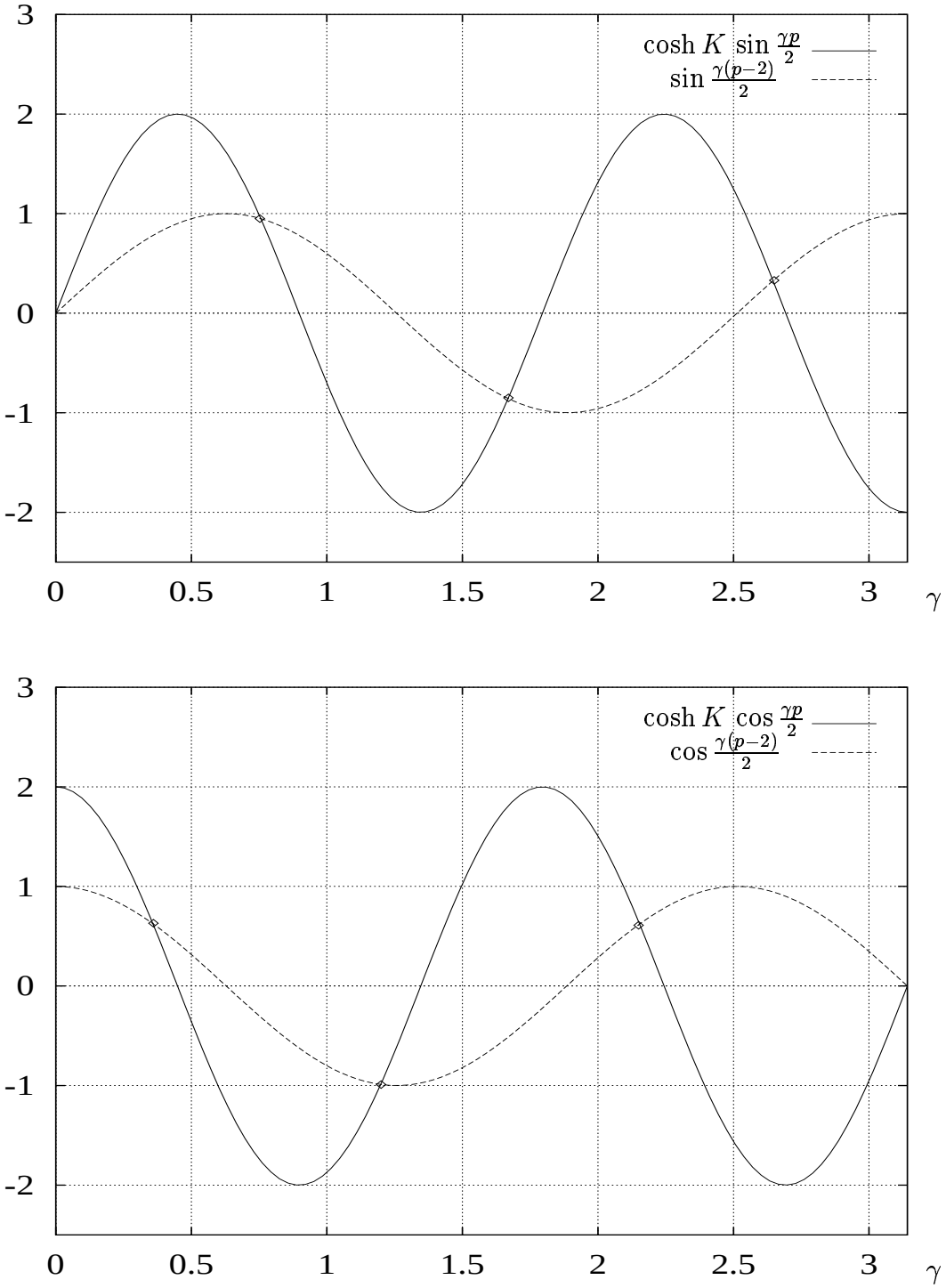


Figure 3.5: Visualization of the equations (3.85) (upper fig.) and (3.86) (lower fig.) and their roots for $p=7$, $\cosh K=2$.

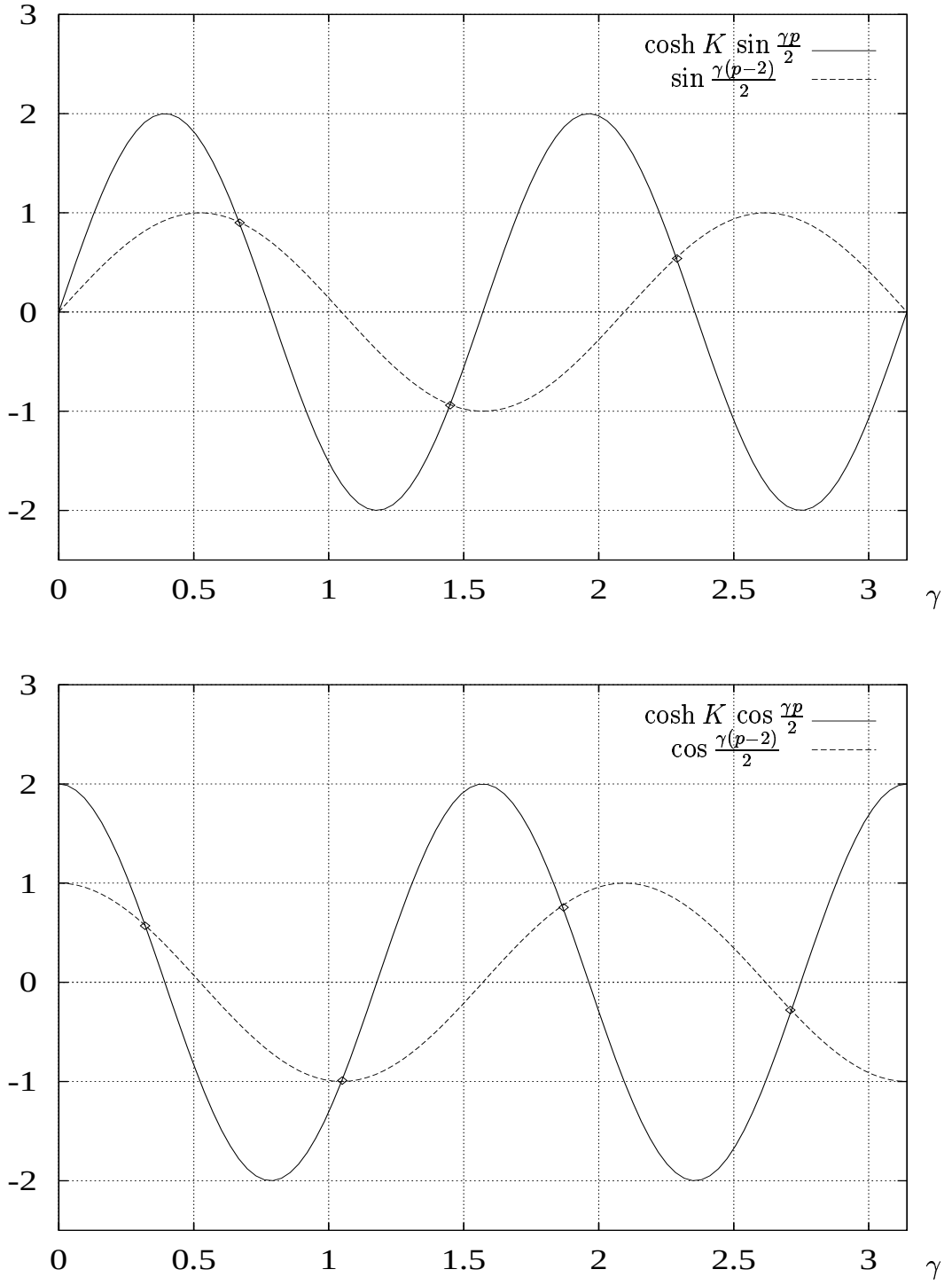


Figure 3.6: Visualization of the equations (3.85) (upper fig.) and (3.86) (lower fig.) and their roots for $p=8$, $\cosh K=2$.

All the $\gamma_{k,m}$ (and therefore all the $\varphi_{k,m}$) are mutually different: Assuming that there is a $\gamma_{k,m}$ fulfilling (3.85) *and* (3.86), the addition of the squared equations gives

$$\cosh^2 K \left(\sin^2 \frac{\gamma_{k,m}\, p}{2} + \cos^2 \frac{\gamma_{k,m}\, p}{2}\right) = \left(\sin^2 \frac{\gamma_{k,m}\,(p-2)}{2} + \cos^2 \frac{\gamma_{k,m}\,(p-2)}{2}\right)$$

which is a contradiction to (3.87).

So all solutions of (3.79) are found. From the fact that the functions $\cos \frac{\pi k}{B}$ rsp. $\sin \frac{\pi k}{B}$ are a complete orthogonal systems in $H_{mv}^{-1/2}(\Gamma_i)$ rsp. $H^{-1/2}(\Gamma_i)$ we can conclude that the $\varphi_{k,m}$, $k \in \mathbb{N}$, $1 \le m \le p-1$, form a complete orthogonal system in $H_{mv}^{-1/2}(\Gamma)$ rsp. $H^{-1/2}(\Gamma)$. Thus, all eigenvalues are found.

Now to (3.68). This estimate follows from the fact that equation (3.86) has a root within $[0, \frac{\pi}{p}]$. ∎

Visualization of the eigenfunctions in the Neumann case. Let $u_{k,m} : \Omega \to \mathbb{R}$ be the function assigned to the eigenfunction $\varphi_{k,m}(y)$ by (2.69). In particular,

$$\left.\frac{\partial u_{k,m}}{\partial x}\right|_{\Gamma} = \varphi_{k,m}, \tag{3.88}$$

then. The following two figures display some of the $u_{k,m}$, $k \in \mathbb{N}$, $m = 1, ..., p-1$, for $p = 5$ quadratic subdomains. In the left column of Fig. 3.7 cuts in x-direction of $u_{k,m}$ for $k = 1$, $m = 1, 2, 3, 4$ are displayed. In the left column of Fig. 3.8 cuts in x-direction for $k = 4$, $m = 1, 2, 3, 4$ are displayed. In the *right* columns, x-cuts of the (continuous) functions $\partial u_{k,m}/\partial x$ are plotted. In these diagrams the values of $\varphi_{k,m}$ occur where the graphs of the functions $\frac{\partial u_{mk,m}}{\partial x}$ meet the interfaces (see (3.88)). The y-cuts of $u_{k,m}$ are just sine rsp. cosine functions with frequency depending on k. In both figures, $\sigma = 0$, $r = 1$, $B = 1$ and Neumann boundary conditions on $\Gamma^I \cup \Gamma^O$ are assumed.

Visualization of the eigenvalues in the Neumann case and development of a preconditioner. As the eigenvalues depend on B, r, σ, p, m and k, it is not easy to get an inpression about their distribution. But this is important in order to construct appropriate preconditioners.

Let us assume $\sigma = 0$ and Neumann conditions on $\Gamma^I \cup \Gamma^O$ at first. In this case, (3.64) becomes

$$\lambda_{k,m} = \frac{2rB}{K} \cdot \frac{\cosh K - \cos \frac{\pi m}{p}}{\sinh K} \tag{3.89}$$

with

$$K = \pi k r, \quad m = 1, ..., p-1, \quad k = 1, ..., \infty. \tag{3.90}$$

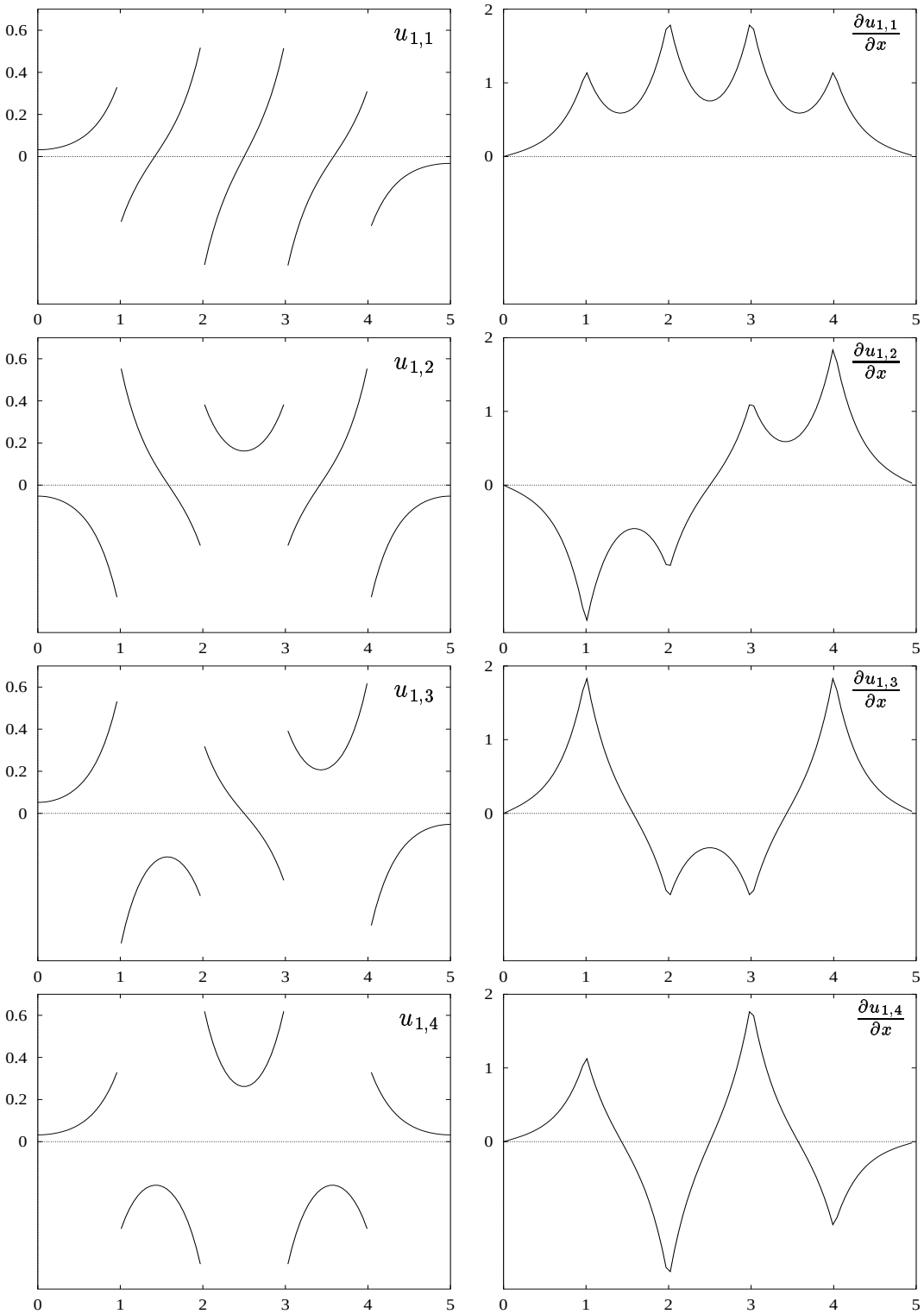


Figure 3.7: Visualization of $u_{k,m}$ (left column) and $\frac{\partial u_{k,m}}{\partial x}$ (right column) for $k=1$, $m=1,...,4$. See explanations in the text (p. 81).

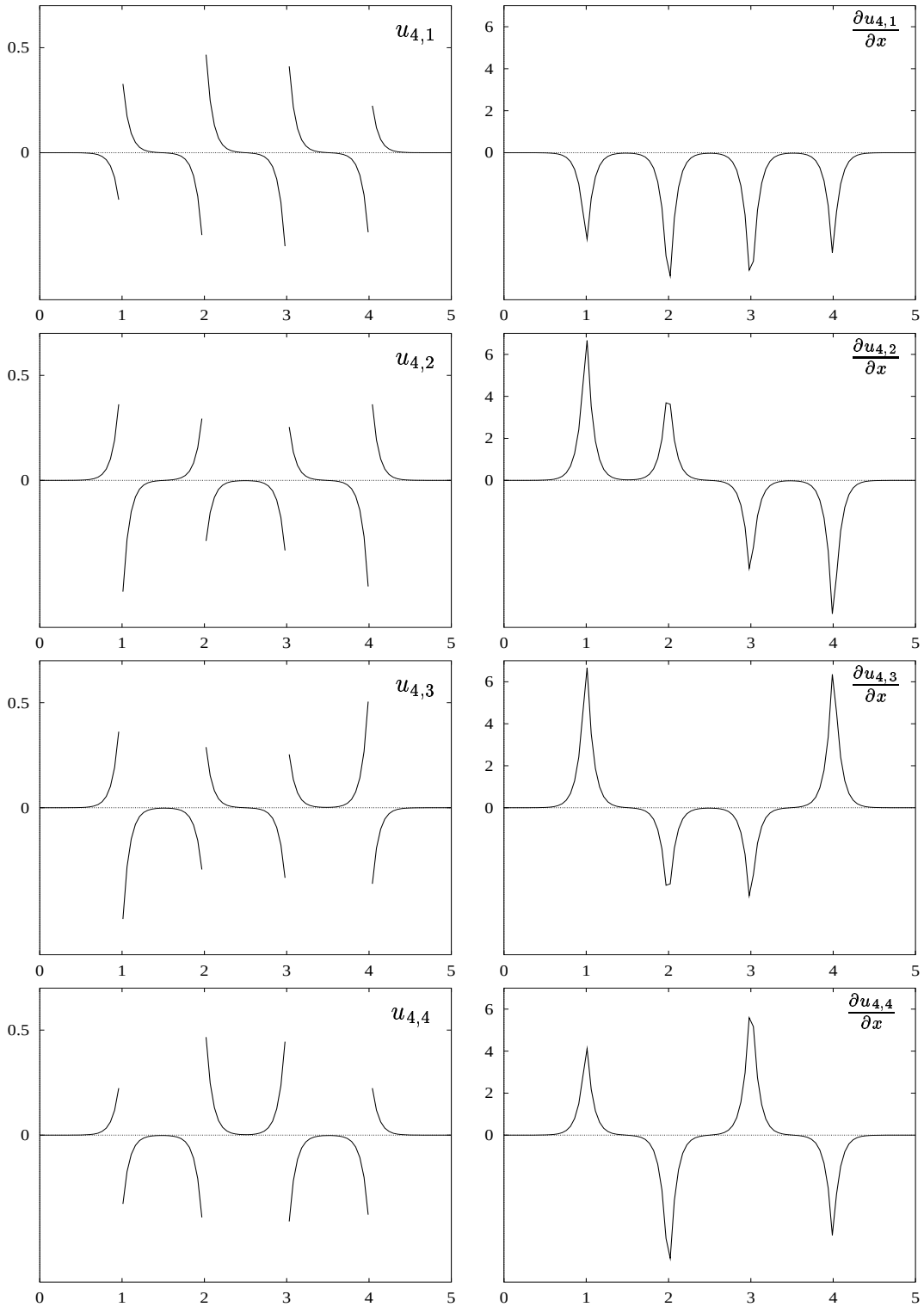


Figure 3.8: Visualization of $u_{k,m}$ (left column) and $\frac{\partial u_{k,m}}{\partial x}$ (right column) for $k=4$, $m=1,...,4$. See explanations in the text (p. 81).

To reduce the dependence of the eigenvalues on k, it seems to be useful to multiply each $\lambda_{k,m}$ by

$$g_{k,m} := k. \tag{3.91}$$

To investigate the condition number of the so preconditioned operator we have to find upper and lower bounds for the expression $k\lambda_{k,m}$ or for

$$(k\lambda_{k,m})^{-1} = \frac{\pi}{2\,B} \cdot \frac{\sinh K}{\cosh K - \cos\frac{\pi m}{p}}.$$

As our interest lies only in the condition number, we may omit the k-independent factor $\frac{\pi}{2B}$. Therefore Figs. 3.9 and 3.10 display the functions

$$f_q(K) := \frac{\sinh K}{\cosh K - q}, \quad q = \cos\frac{\pi m}{p}. \tag{3.92}$$

In Fig. 3.9 the case $p = 4$, $r = 0.7$ is visualized, in Fig. 3.10 the case $p = 8$, $r = 0.2$. Due to this choice of p the functions f_q for $q = 1$, $q = \cos\frac{\pi}{8}$, $q = \cos\frac{2\pi}{8}$,..., $q = -1$ are displayed. We see in both pictures that for $k \to \infty$ (i.e. $K \to \infty$) the preconditioning produces only values $k\lambda_{k,m}$ very close to 1. For small K instead (i.e. $k = 1$, $K = \pi r$), we get rather large *and* rather small eigenvalues. That means that the values $k\lambda_{k,m}$ for $k = 1$ determine the condition number of the preconditioned operator.

If r approaches zero the minimum value of K tends to zero due to (3.90), and the ratio between the largest and the smallest eigenvalue becomes worse (compare the case $r = 0.7$ (Fig 3.9) to $r = 0.2$ (Fig. 3.10)). Watching $f_q(K)$ for q approaching 1 (which has to be taken into account for $p \to \infty$) we see that $\lim_{\substack{q\to 1\\ r\to 0}} \sup_{k\in\mathbb{N}} f_q(\pi k r) = \infty$. That means that for $p \to \infty$, $r \to 0$ the condition number approaches infinity.

So the condition number when multiplying each eigenfunction $\varphi_{k,m}$ by k (i.e. application of $(-\Delta_0)^{1/2}$ rsp. $(-\Delta_{Nm})^{1/2}$ on φ) is to be expected independent of N but approaching infinity for $\sigma = 0$, $r \to 0$, $p \to \infty$.

All these facts are formalized in Theorem 3.13.

Now the case $\sigma > 0$. It is easy to see that for (3.91) the term $g_{k,m}\lambda_{k,m}$ cannot be bounded independent of $\sigma \in [0, \infty)$. Instead, we use

$$g_{k,m} := \frac{\pi}{2\,B}k + \frac{\sqrt{\sigma}}{2}. \tag{3.93}$$

Using K from definition (3.69) in (3.64) we see that

$$(g_{k,m}\lambda_{k,m})^{-1} = \frac{\sqrt{\sigma B^2 + \pi^2 k^2}}{B\sqrt{\sigma} + \pi k} \, \frac{\sinh K}{\cosh K - \cos\gamma_{k,m}} \tag{3.94}$$

Using (3.22), (3.23) we observe that the first of the two fractions is bounded between 1 and $1/\sqrt{2}$ independent of σ, r, B, k. Thus, for the visualization we may

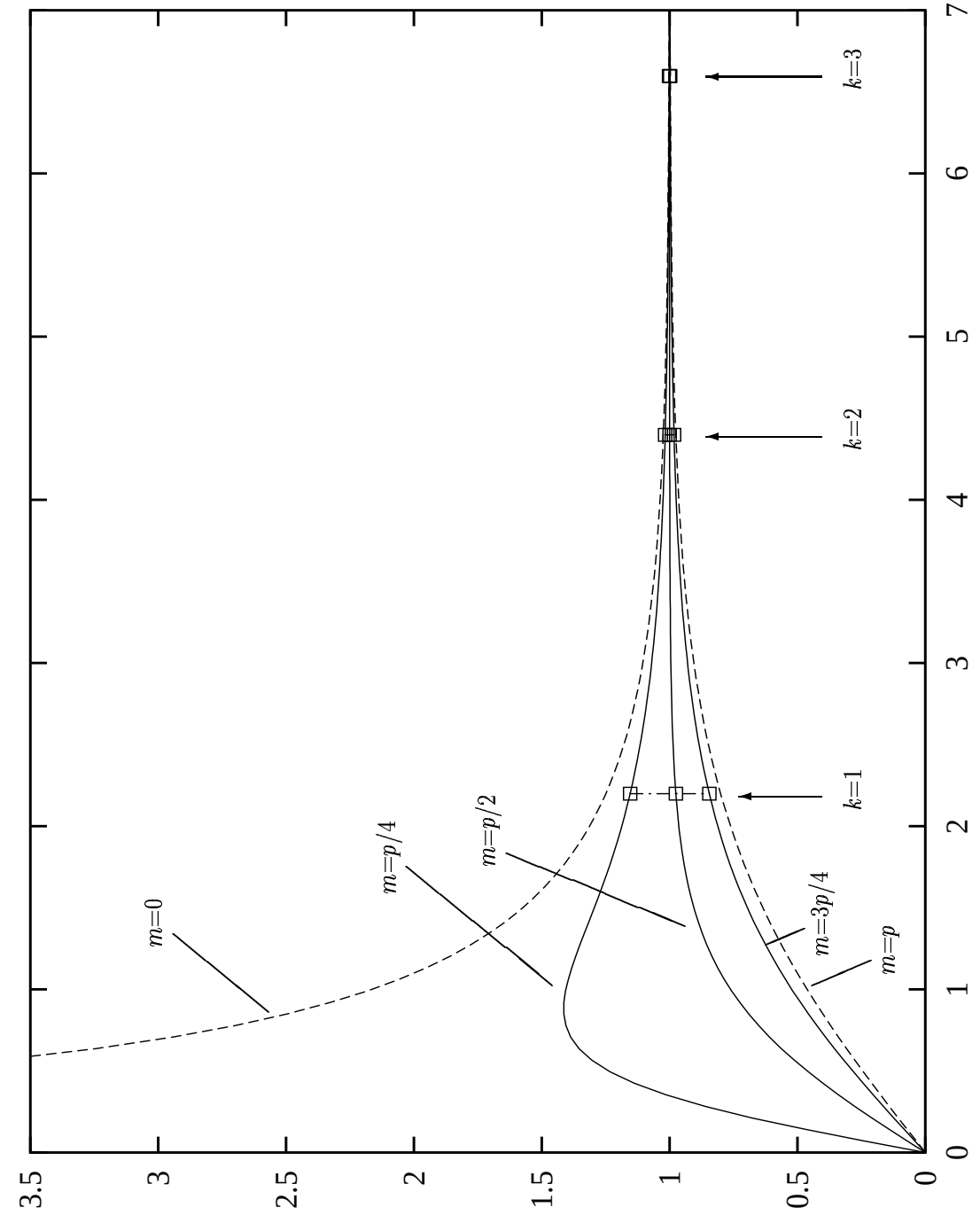


Figure 3.9: Visualization of the eigenvalues $(k\lambda_{k,m})^{-1}$ (3.64)-(3.65) ('dots') for $\sigma = 0$, $r = 0.7$, $p = 4$ and Neumann b.c. on Γ^W. $K := \pi kr$ on the horizontal axis. Explanations in the text (p. 81/84). Compare to Fig. 3.10 where r is smaller and therefore κ bigger. The number of lines depends on p (see Fig. 3.10 where p is larger). For *Dirichlet* b.c. on Γ^W, the lines $m = 1, ..., p-1$ slightly change, but the bounds $m = 0$, $m = p$ are the same.

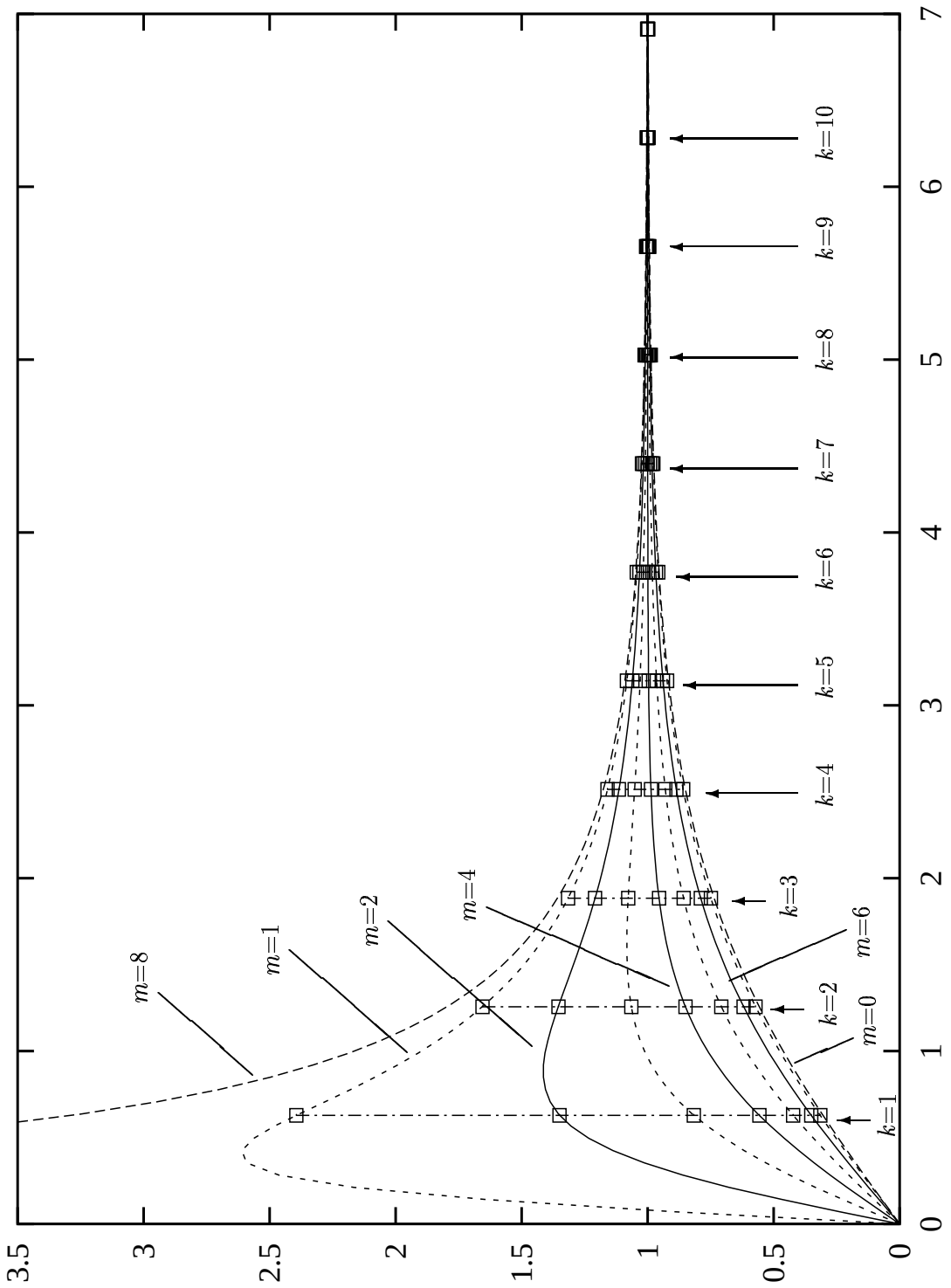


Figure 3.10: Visualization of the eigenvalues $(k\lambda_{k,m})^{-1}$ (3.64)-(3.65) ('dots') for $\sigma = 0$, $r = 0.2$, $p = 8$. $K := \pi kr$ on the horizontal axis. Explanations in the text (p. 81/84).

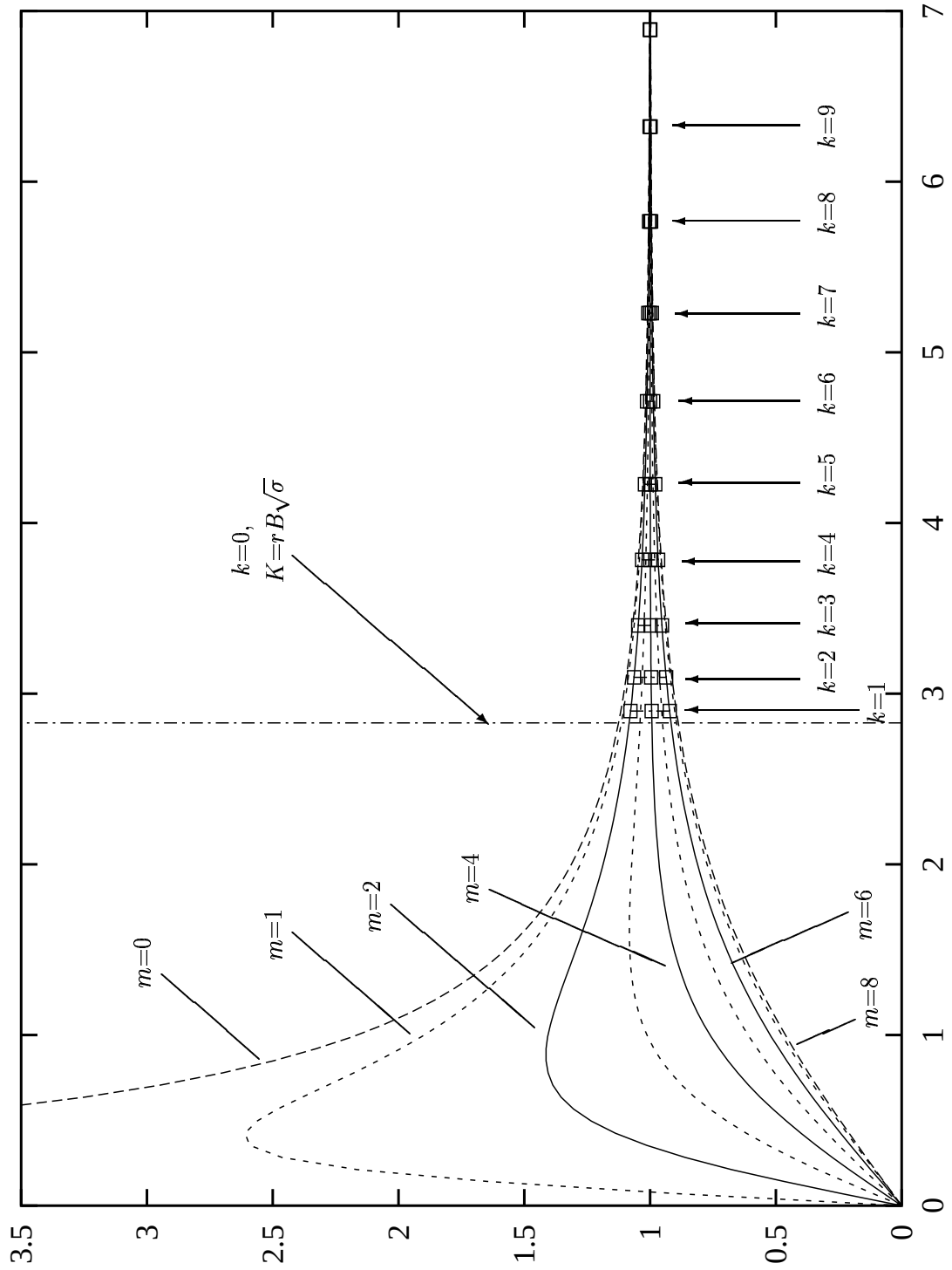


Figure 3.11: Visualization of the eigenvalues $(k\lambda_{k,m})^{-1}$ (3.64)-(3.65) ('dots') for $\sigma = 200,\ r = 0.2,\ p = 4,\ B = 1.$ $K := r\sqrt{\sigma B^2 + \pi^2 k^2}$ on the horizontal axis. Explanations in the text (p. 84/88).

omit this factor. So the visualization of (3.92) in Figs. 3.9-3.11 is still valid for arbitrary $\sigma \geq 0$ (except fot the bounded factor, and taking into account K from (3.69)).

Fig. 3.11 points out the difference for $\sigma > 0$: The lower bound for K (see vertical line in Fig. 3.11) is now $r\sqrt{\sigma B^2 + \pi^2}$ instead of just $r\pi$. So in the case $\sigma > 0$ the situation improves: For large values of σ we can expect that the above preconditioning gives a smaller condition number than for $\sigma = 0$. Again, this relation is formalized in the following theorem.

Theorem 3.13 *Let $(-\Delta)^{1/2}$ be from Lemma 3.4 and let $p > 2$. For all the cases (i)-(iv) of Theorem 3.12 the following three assertions hold:*

(i) For $g_{k,m} = \alpha\, k + \beta$, $\alpha > 0$, $\beta \geq 0$, $\sigma \geq 0$ the eigenvalues $g_{k,m}\lambda_{k,m}$ of the preconditioned operator CA, $C = \frac{\alpha B}{\pi}(-\Delta)^{1/2} + \beta\, id$, are bounded independently of p and $k, m \in \mathbb{N}$:

$$c_1(r, \sigma, B) \leq \lambda_{k,m} \cdot g_{k,m} \leq c_2(r, \sigma, B)$$

(ii) For $\sigma \geq 0$, $r > 0$,

$$g_{k,m} := g_k := \frac{\pi}{2B}\, k + \frac{\sqrt{\sigma}}{2},$$

we can find the c_i in (i) such that

$$\frac{c_2(p, r, \sigma, B)}{c_1(p, r, \sigma, B)} \leq \sqrt{2}\left(1 + \frac{4}{r^2(\sigma B^2 + \pi^2)}\right) \leq \sqrt{2}\left(1 + \frac{4}{\pi^2 r^2}\right), \qquad (3.95)$$

i.e. the c_i are independent of σ and B.

(iii) Under the assumptions $\sigma = 0$, $r \to 0$, $L = pr = const$: For any $g_{k,m} = g_k$ independent of m, there is a constant $c_3 > 0$ so that

$$\frac{\sup\limits_{\substack{m=1,\ldots,p-1 \\ k=1,\ldots,\infty}} g_{k,m}\lambda_{k,m}}{\inf\limits_{\substack{m=1,\ldots,p-1 \\ k=1,\ldots,\infty}} g_{k,m}\lambda_{k,m}} \geq c_3\, r^{-2} \qquad (3.96)$$

holds.

Discussion of Theorem 3.13. Part (i) shows that the use of *any* preconditioner of type

$$C_{loc} = \alpha\,(-\Delta)^{1/2} + \beta\, id, \quad \alpha > 0,\ \beta \geq 0, \qquad (3.97)$$

$$C_{glob}\varphi := (C_{loc}\varphi_1, \ldots, C_{loc}\varphi_{p-1}) \qquad (3.98)$$

(rsp. its discretization) should cause a condition number κ independent of the number of grid points and the number of subdomains. For (3.98) with

$$C_{loc} = (-\Delta)^{1/2} + \sqrt{\sigma}\, id, \tag{3.99}$$

(ii) shows that the condition number is even bounded independent of $\sigma \geq 0$, $B > 0$.

(ii) and (iii) show the behaviour of the condition number if the number of subdomains p is increasing. Let us consider the following two possibilities to increase p:

- If $r = const$ while $p \to \infty$ (i.e. the length $L = pBr$ of the channel tends to infinity), κ remains bounded, as the bounds in (i) do not depend on p, L. This property is very important for the possible use of the CGBI method on very long domains.

- If the length L is fixed while $\sigma = 0$, $p \to \infty$ (i.e. $r \to 0$), κ tends to ∞ as $O(p^2) = O(r^{-2})$ due to (ii). Part (iii) shows that for $\sigma = 0$ this estimate for κ cannot be improved. So the number of CG iteration steps to gain a certain error reduction can be expected to behave like $O(p)$ for very large p. That means that a further increase in parallelism is senseless. This observation is the starting point for some different preconditioners in Section 3.3. The improvement compared to preconditioners of type (3.97) is gained by using factors $g_{k,m}$ which are *not* independent of m. But let us mention that the loss of efficiency of the preconditioner (3.98)-(3.99) is no problem as long as, let us say, $r\sqrt{\sigma B^2 + \pi^2} \gtrsim 0.6$. So only very narrow subdomains of $r \lesssim 0.2$ require different preconditioners.

Numerical implementation of preconditioner (3.98)-(3.99) and the Chebyshev case. To apply the operator (3.98)-(3.99), each $\varphi|_{\Gamma_i}$ is treated separately. No additional communication is necessary. If there is an equidistant grid on the interface Γ_i, C_{loc} can be applied directly via FFT. If there is a Gauss-Lobatto mesh on Γ_i, $(-\Delta)^{1/2}$ has to be replaced by C_{GL} from Theorem 3.9. So we get (3.98) with

$$C_{loc} = C_{GL} + \sqrt{\sigma}\, id, \tag{3.100}$$

then, instead of (3.99). The assertions of Theorem 3.13 are still valid for the Chebyshev case: Theorem 3.9 (see also remark on p. 66) says that there are no additional dependencies occuring in the equivalence constants c_1, c_2 if $(-\Delta)^{1/2}$ is replaced by C_{GL}. C_{GL} then is implemented by FFT again as explained at the end of Section 3.1.3.2. Thus, the application of the preconditioner takes only

$O(N \log N)$ operations for each processor which is again neglectable compared to the local FDM/FEM-Multigrid solver and the local Chebyshev solver.

Proof of Theorem 3.13.
Ad (i). We have to estimate

$$(\alpha k+\beta)\,\lambda_{k,m}.$$

Obviously

$$\frac{2rB}{K}\,\frac{\cosh K-1}{\sinh K} \le \lambda_{k,m} \le \frac{2rB}{K}\,\frac{\cosh K+1}{\sinh K}$$

with $K=K(k)$ defined in (3.69). Both expressions

$$\frac{2rB\,(\alpha k+\beta)}{K}\,\frac{\cosh K \pm 1}{\sinh K}$$

are bounded for $K\in[r\sqrt{\sigma B^2+\pi^2},\infty)$ independendly of k, m, p.
Ad (ii). We have to estimate

$$g_k\lambda_{k,m} = \frac{B\sqrt{\sigma}+\pi k}{\sqrt{\sigma B^2+\pi^2k^2}}\,\frac{\cosh K-\cos\gamma_{k,m}}{\sinh K}. \tag{3.101}$$

As already explained after equation (3.94), the first fraction in (3.101) lies between 1 and $\sqrt{2}$. The other fraction is obviously bounded between $f_1(K) := \frac{\cosh K-1}{\sinh K}$ and $f_2(K) := \frac{\cosh K+1}{\sinh K}$ where $K \ge K_0 := r\sqrt{\sigma B^2+\pi^2}$. Differentiation shows that f_1 is monotoneously increasing and f_2 monotoneously decreasing for positive K (see also Fig. 3.9-3.10 for these functions). Therefore it is sufficient to estimate $f_1(K_0)$ and $f_2(K_0)$. So we arrive at

$$\frac{c_2}{c_1} \le \sqrt{2}\,\frac{f_2(K_0)}{f_1(K_0)} = \sqrt{2}\,\left(1+\frac{2}{\cosh K_0-1}\right) \le \sqrt{2}\,\left(1+\frac{4}{K_0^2}\right)$$

from which (ii) follows.
Ad (iii). Let us consider the two eigenvalues $g_{1,1}\lambda_{1,1}$ and $g_{1,p-1}\lambda_{1,p-1}$. Under the assumptions of (iii), the arguments of the hyperbolic functions and the cosine in (3.64) tend to 0 rsp. to π for these values of k, m. (Here we have used (3.68) for the Dirichlet case.)

Therefore a Taylor expansion of the hyperbolic cosine and estimate (3.143) for the cosine yield for the Neumann case (3.65) *and* for the Dirichlet case (3.68)

$$\begin{aligned}\frac{g_{1,p-1}\lambda_{1,p-1}}{g_{1,1}\lambda_{1,1}} &= \frac{\lambda_{1,p-1}}{\lambda_{1,1}} = \frac{\cosh\pi r-\cos\gamma_{1,p-1}}{\cosh\pi r-\cos\gamma_{1,1}} \ge \frac{\cosh\pi r}{\cosh\pi r-(1-\frac{\pi^2}{2p^2})} \\ &\approx \frac{\frac{\pi^2r^2}{2}+1}{\frac{\pi^2r^2}{2}+\frac{\pi^2}{2p^2}} = \frac{L^2}{L^2+1}\,\left(1+\frac{2}{\pi^2r^2}\right).\end{aligned} \tag{3.102}$$

Thus, (iii) holds. ∎

3.1.5 Numerical results

In this section test runs for the preconditioner(s) (3.98)-(3.100) developed in Section 3.1 are documented. Tests *with* and *without* preconditioning are compared. If not explicitely mentioned the tests were made with $B=1$, $r=1$, $p=4$ and the CG starting vector $\varphi=0$.

In Fig. 3.12, $p=4$ FDM subdomains are used with $\sigma=0$, $r=1$, $N\times N$ grid points per subdomain with $N=16,64,256$, and the exact solution (2.83) ('test function 4'). Three test runs were done with preconditioner (3.98)-(3.99) on equidistant boundary meshes (full lines) and three without preconditioner (broken lines). The test runs with preconditioner needed only half of the iteration steps compared to the other test runs; for large N even less. We can see that the final error when CGBI becomes stationary depends like $O(N^2)$ from the number of grid points (as already in Fig. 2.13), but the final error does *not* depend on the fact if preconditioning was used or not.

In the upper part of Fig. 3.13, the same tests were made with $p=4$ *Chebyshev* subdomains and the exact solution (2.80) ('test function 1').

In Fig. 3.12 and in the upper part of Fig. 3.13 the convergence rate when using the preconditioner does not depend on N. This was predicted by theory (Theorem 3.13 and also Theorem 3.9). But the figures show that the convergence rate when the preconditioner was *not* used also is rather *in*dependent of N. This is not really a contradiction to (3.64) which predicts an unbounded condition number for $N\to\infty$, as we have used *rather smooth* exact solutions (2.80), (2.83). So during CGBI, all the φ have hardly any highly oscillating components. Therefore the bad error reduction property for highly oscillating φ of the unpreconditioned CGBI cannot be observed. To make this effect visible I use two different ideas:

The first is to use *a highly oscillating starting vector* instead of the zero starting vector. The CGBI process has to extinct these oscillations. So in the lower part of Fig. 3.13 a CG starting vector with components which vary (between $\pm 10\cdot\frac{1}{N}\cdot\frac{L}{L+1}$) by random is taken on $\Gamma_{p/2}$. Now we can see (lower part of Fig. 3.13, broken lines) that in the absence of the preconditioner the convergence rate depends on N while it is independent of N when the preconditioner is used (full lines). The number of CG steps is reduced from several dozens to ≈ 8. The error reduction rate in the *preconditioned* case is quite constantly one full power of ten!

In Fig. 3.14 the same test is made for the *mixed* boundary value problem with Neumann boundary conditions on $\Gamma^W\cup\Gamma^L$ and the same exact solution (2.80). In this case of Neumann conditions on Γ^W $C_{GL,Nm}$ (3.58) is used, and the random CGBI starting vector is generated in a way that $\varphi(y)=-\varphi(B-y)$ so that automatically the condition $\int_{\Gamma_i}\varphi\,dy=0$, which was derived in Sec. 2.4.2 from the compatibility condition, holds. The effect of the preconditioner $C_{GL,Nm}$ is similar, but a bit weaker than C_{Gl} (3.35) in the Dirichlet case. The influence of the CG starting vector (upper part $\leftrightarrow$ lower part of Fig. 3.14) is also similar to the Dirichlet case Fig. 3.13.

In 3.15 the same tests are made for the *pure* Neumann problem for the exact solution (2.81). Again, the error reduction rate is worse compared to the Dirichlet case (Fig. 3.13). It was observed that this problem can be made deteriorate by using $C_{GL,Nm} + c\,id$ instead of $C_{GL,Nm}$. The non-connected dots in Fig. 3.15 show the effect for $c = 2$ and $N = 256$. The alternative matrix preconditioners mentioned at the end of Sec. 3.1.3.3 are not tested, yet.

The second possibility to investigate the effect of the preconditioner on boundary value functions φ with highly oscillating parts is to use a solution which is not so smooth. In Fig. 3.16 the solution (2.82) ('test function 3') was used. Its Chebyshev series is decreasing slowlier compared to the more regular functions so that the final error is large (as in Fig. 2.14). In the Chebyshev case (lower diagram) the convergence speed is again one full power of ten per CGBI step. In the FD case (upper diagram), we already reach the final error after only 1-2 CGBI steps!

The following figures deal with varying p, r, σ:

Fig. 3.17 shows that the convergence rate is *independent* of the number of subdomains p if the aspect ratio r is fixed. This observation corresponds to Theorem 3.13 part (ii).

In Fig. 3.18 the dependence of the convergence rate on σ and on r is displayed. If both r and σ are close to zero, the convergence rate drops. Again, this effect could be foreseen by Theorem 3.13 (ii) and Figs. 3.9-3.11. For the case $r \to 0$ in Section 3.3 a better preconditioner is developed.

The case $\sigma = 0$, $r \to 0$ is investigated more closely in Fig. 3.19. The length $L = pr$ of the domain is fixed, but p and r are varied. According to Theorem 3.13 parts (ii), (iii) the convergence rate decreases for $r \to 0$. If we compare the test runs $r = 1/8$ and $r = 1/16$ the estimates (3.95)-(3.96) let us expect that the condition number increases by the factor 4. That means that the number of CG iteration steps should double. This doubling can be observed in Figure 3.19 very clearly.

In Fig. 3.20 the important case of mixing FDM and Chebyshev solvers is investigated. 10 subdomains were used with the FDM solver on the 4^{th} subdomain and Chebyshev local solvers on all the other (full lines). The broken lines represent the case where *only* Chebyshev solvers were used. Of course, on the FDM subdomain the error is higher than on Chebyshev subdomains. But Fig. 3.20 shows that the 'pollution' of the error on Chebyshev subdomains by the FDM error is extremely small: It decreases exponentially along the channel. For $\sigma > 0$ (lower diagram) the decay is faster than for $\sigma = 0$ (upper diagram). All these results comply with theory.

Remark on discretization of $(-\Delta)^{1/2}$. The construction of our preconditioner was based on the analysis of the *non-discretized* operator A. As a result we found

that the preconditioner should be $C = (-\Delta)^{1/2}$, [9] which was discretized, then. Another possibility would be to use the square root of the *discretization* of $-\Delta$; an approach which was followed in [15] and [4] for the 'original' Schur method based on Dirichlet interface conditions and on the operator (2.85) and only for the model case of two subdomains and an equidistant mesh.

This modified approach leads to the preconditioning

$$\sum_{k=1}^{N} \alpha_k \sin \frac{\pi k x}{N} \longmapsto \sum_{k=1}^{N} g_k \, \alpha_k \sin \frac{\pi k x}{N}$$

with $g_k = \sin \frac{k\pi}{2N}$ instead of $g_k = k$. Since $\frac{1}{N} \leq \frac{1}{k} \sin \frac{k\pi}{2N} \leq \frac{\pi}{2N}$, both preconditioning operators are spectrally equivalent, and the condition numbers of the two preconditioned operators differ by a factor of $\pi/2$, only.

In fact, a comparison of the two preconditioners by numerical tests revealed no relevant difference. This was true both for equidistant and for Gauss-Lobatto boundary mesh[10], and also for tests with strongly oszillating CGBI initial vector φ.

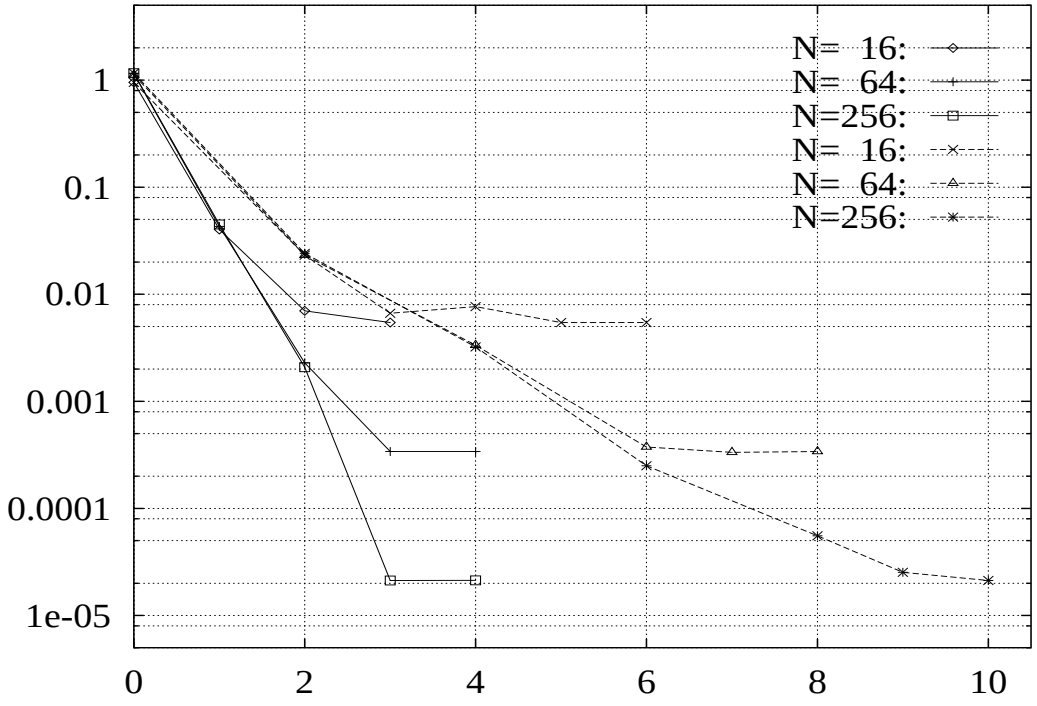


Figure 3.12: The Dirichlet problem on 4 FDM subdomains for solution (2.83). Full lines: using the preconditioner $(-\Delta_0)^{1/2}$. Broken lines: using no preconditioner. Further explanations to all the figures in the text.

[9] Let us assume $\sigma = 0$ for the moment.

[10] using our technique of Sec. 3.1.3 for the Gauss-Lobatto case; numerical realization on p. 66

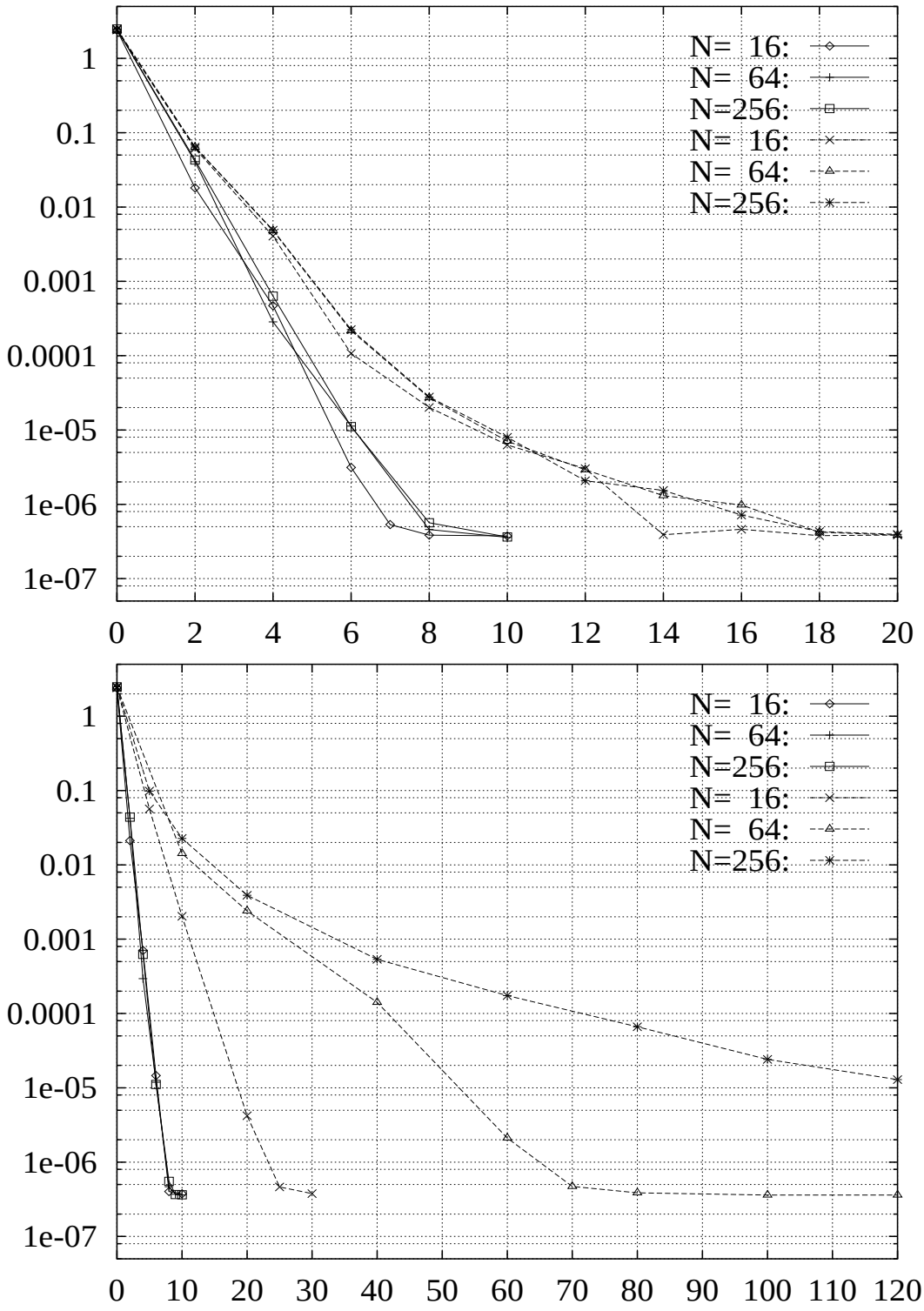


Figure 3.13: The Dirichlet problem on 4 *Chebyshev* subdomains for solution (2.80). Upper fig.: CG starting vector $\varphi = 0$. Lower fig.: highly oscillating CG starting vector. Full lines: using the preconditioner C_{GL}. Broken lines: using no preconditioner.

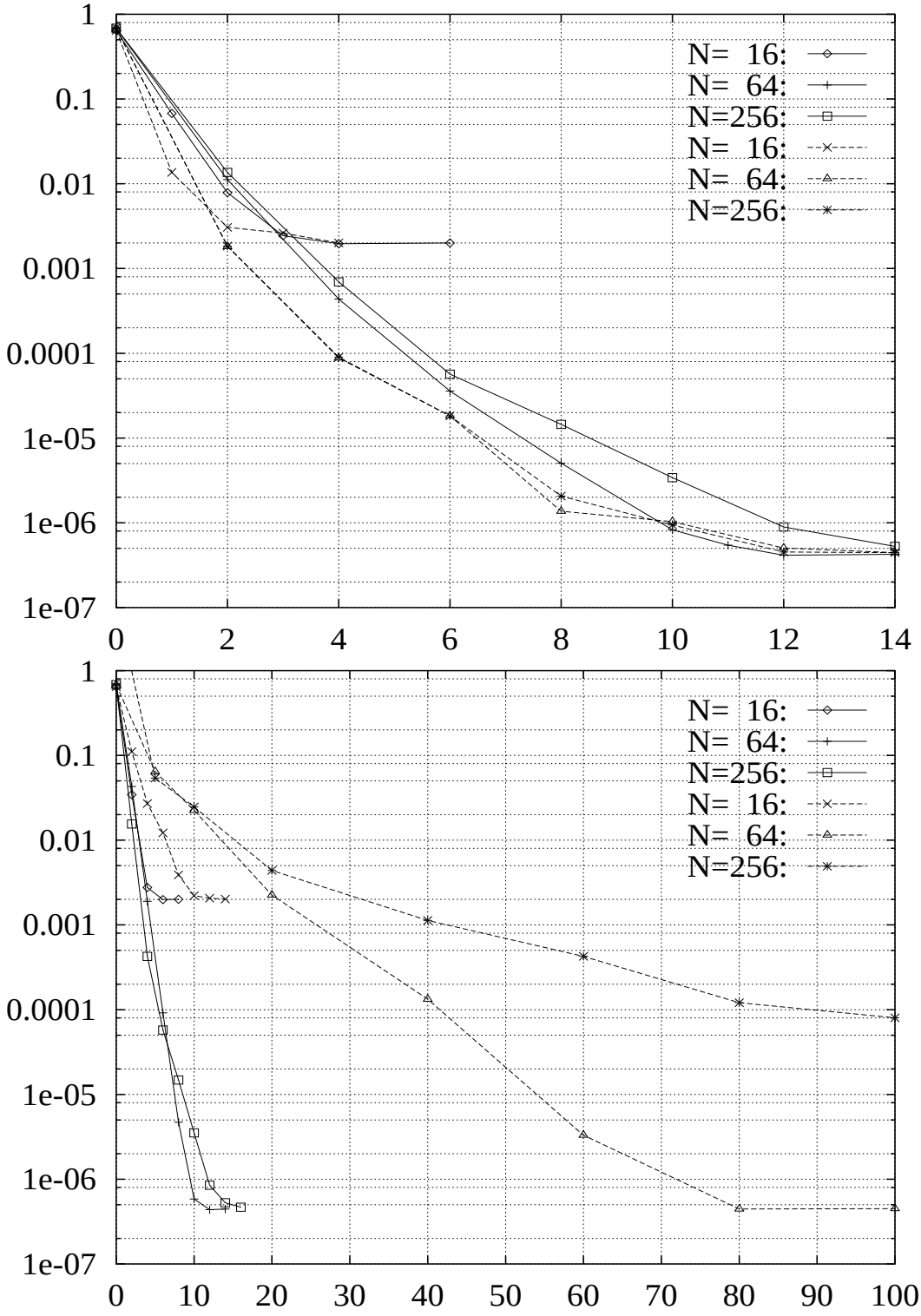


Figure 3.14: The mixed Neumann/Dirichlet problem for solution (2.80) on 4 Chebyshev subdomains. Upper fig.: CG starting vector $\varphi = 0$. Lower fig.: highly oscillating CG starting vector. Full lines: using the preconditioner $C_{GL,Nm}$. Broken lines: using no preconditioner.

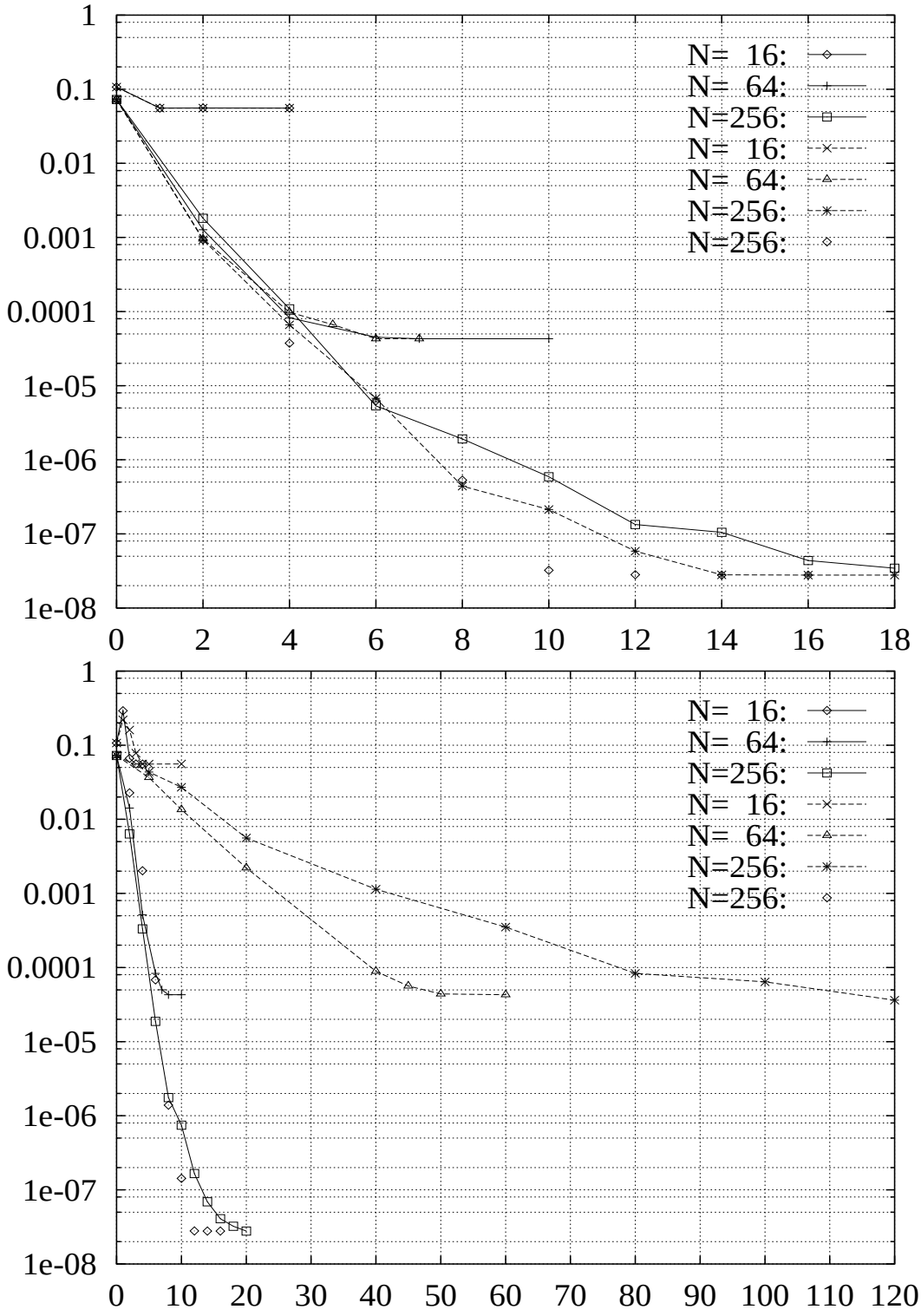


Figure 3.15: The pure Neumann problem for solution (2.81) on 4 Chebyshev subdomains. Upper fig.: CG starting vector $\varphi = 0$. Lower fig.: highly oscillating CG starting vector. Full lines: using the preconditioner $C_{GL,Nm}$. Broken lines: using no preconditioner. No lines: using $C_{GL,Nm} + 2\,id$ instead of $C_{GL,Nm}$.

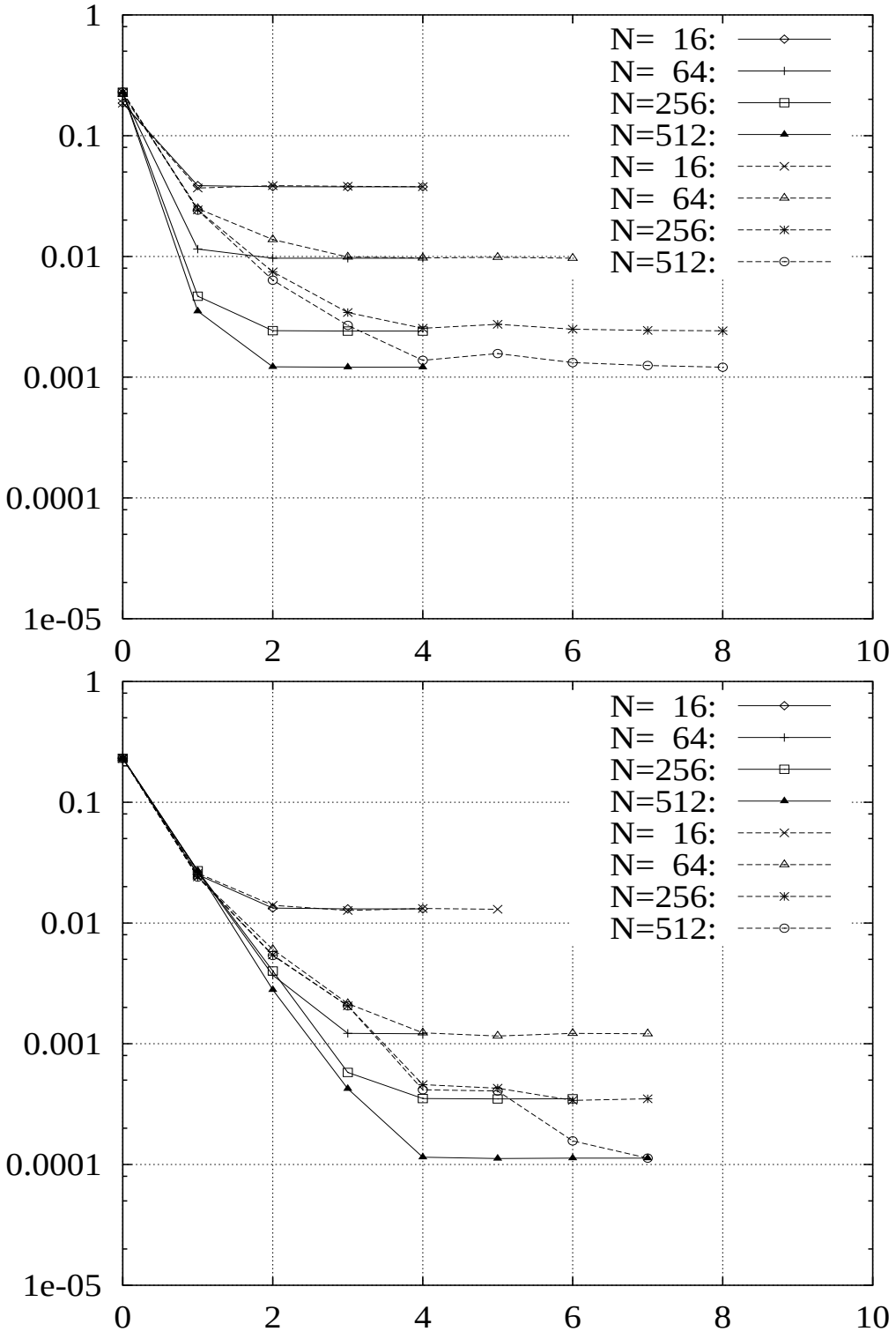


Figure 3.16: Dirichlet problem for the non-smooth solution (2.82). FDM (upper fig.) and Chebyshev (lower fig.) method. CG starting vector $\varphi = 0$. Full lines: using the preconditioner $(-\Delta_0)^{1/2}$ rsp. C_{GL}. Broken lines: using no preconditioner.

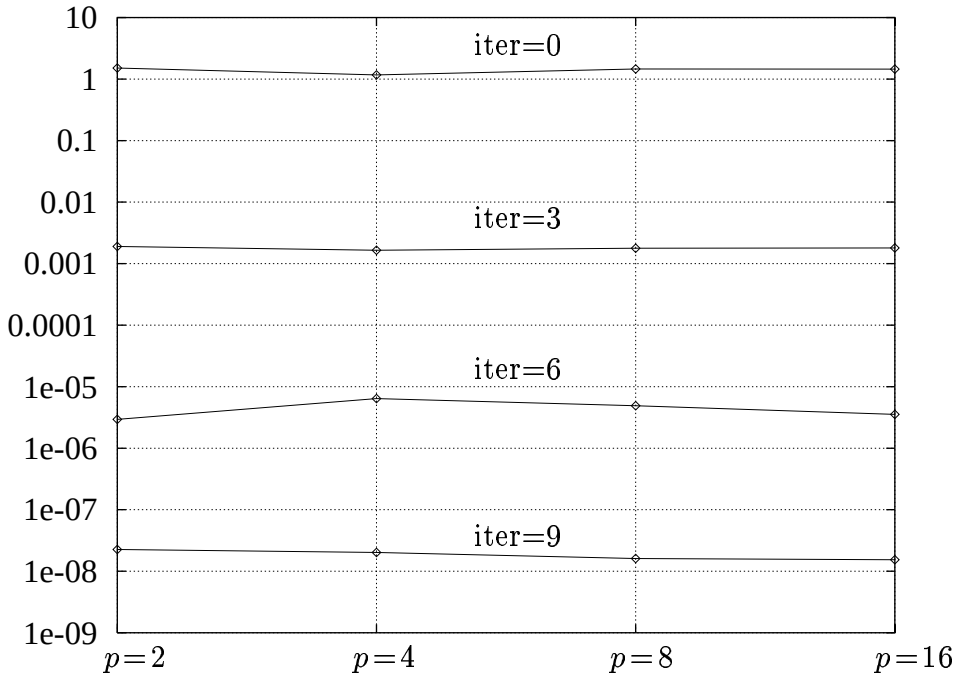


Figure 3.17: Dirichlet problem on p Chebyshev subdomains for the solution (2.83) for $p \to \infty$, $r = 1 = const$, $N = 64$. Global error after 0, 3, 6, 9 preconditioned CGBI steps. CGBI starting vector $\varphi = 0$.

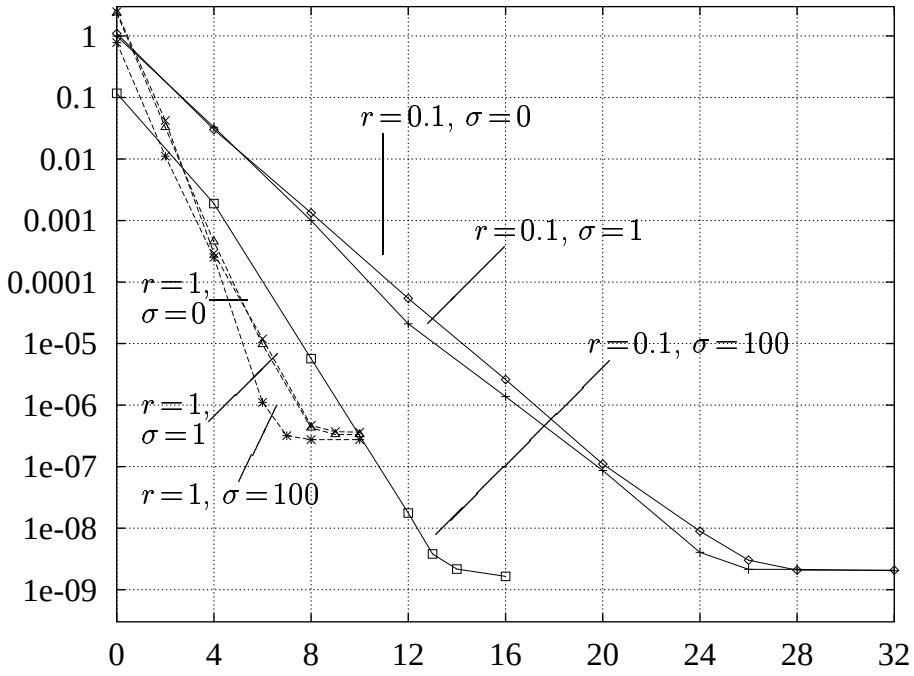


Figure 3.18: Dirichlet problem on 4 Chebyshev subdomains for the solution (2.80) for $N=64$ and different values of r and σ. The combinations $r=0.1$ (full lines), $r=1$ (broken lines), $\sigma=0,1,100$ are displayed. Oscillating CG starting vector $\varphi\neq 0$. If σ and r are close to zero, the convergence rate drops.

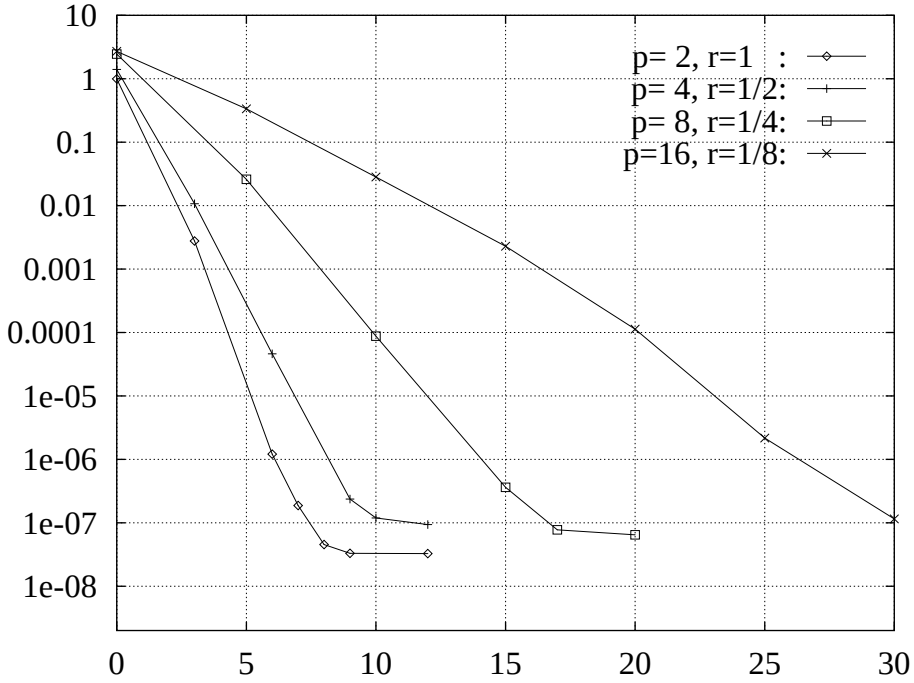


Figure 3.19: Dirichlet problem for the solution (2.80) on a channel of fixed length $L=pr=2$. $N=64$, $\sigma=0$ and an oscillating CG starting vector $\varphi\neq 0$. The smaller r, the slower the convergence is.

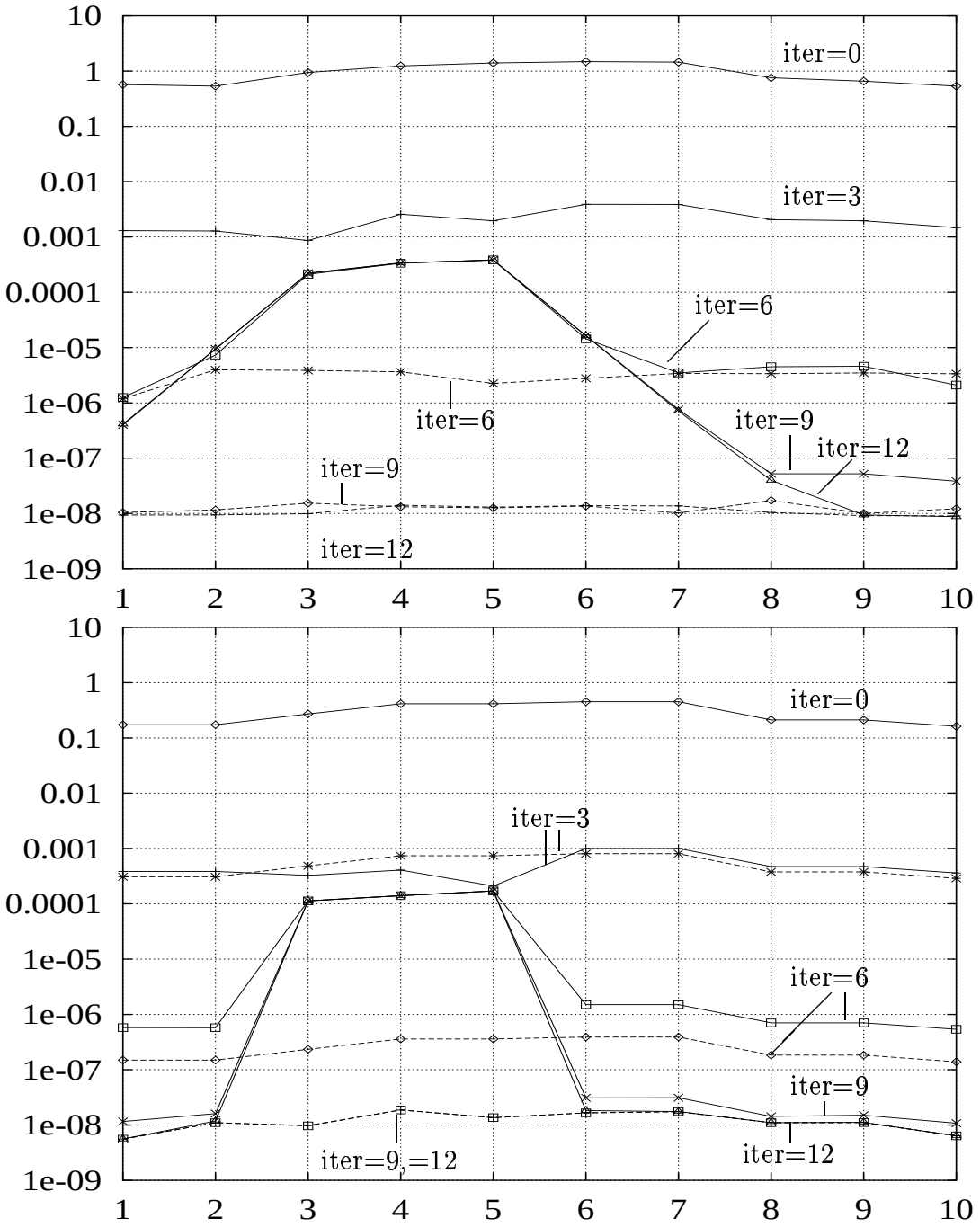


Figure 3.20: Dirichlet problem on 10 subdomains with a combination of FDM and Chebyshev local solvers. On subdomain no. 4 the FD method was used and on all the other the Chebyshev solver (full lines). The local error on each subdomain after 3,6,9,12 CG steps is displayed. The exact solution is (2.83). $\sigma = 0$ in the upper figure and $\sigma = 100$ in the lower figure. $r = 1$ and CG starting vector $\varphi = 0$. For a comparison, the broken lines show the case when Chebyshev solvers are used on *all* subdomains.

3.2 Preconditioning by interpolation from Gauss-Lobatto grid to equidistant grid

On first sight, the easiest possibility to solve to problem of constructing a discrete preconditioner acting on the Gauss-Lobatto points (see Section 3.1.3) seems to be the use of an interpolation between the Gauss-Lobatto mesh values and the equidistant mesh values on the interfaces: To apply the discrete Gauss-Lobatto preconditioner,

1. calculate equidistant boundary mesh values of each φ_i from the Gauss-Lobatto mesh values of φ_i by interpolation
2. apply the discrete equidistant boundary mesh preconditioner C (see end of Section 3.1.2) onto φ_i
3. calculate Gauss-Lobatto boundary mesh values of $C\varphi_i$ from the equidistant mesh values of $C\varphi_i$ by interpolation

Let us refer to this preconditioner as 'interpolation preconditioner' C_{INT}. The question occurs which interpolation should be used.

One possibility is to expand φ_i into a Chebyshev series (by application of FFT). Step 1 can be performed by evaluating the Chebyshev polynomials at the equidistant grid points. Analogously, step 3 can be performed by evaluation trigonometric functions at the Gauss-Lobatto points. Due to the evaluation of the sums, this algorithm takes $O(N^2)$ operations for $N+1$ grid points per interface. Let us call this 'full order' interpolation.

Another possibility to perform the interpolation in step 1 and step 3 is to use a piecewise polynomial interpolation. We checked piecewise 1^{st} and 2^{nd} polynomials as well as cubic b-splines. These kinds of interpolation require $O(N)$ calculation steps, i.e. the whole preconditioner $O(N \log N)$.

It was already mentioned in the beginning of Section 3.1.3 that the preconditioners using interpolation onto an equidistant boundary mesh do not work properly. Let us verify this by numerical tests. Fig. 3.21 shows that CGBI with the interpolation preconditioner C_{INT} becomes stationary earlier than in the case of the Gauss-Lobatto mesh preconditioner C_{GL} (or in the case of *no* preconditioner (not displayed)). For the test runs that use a highly oscillation CGBI starting vector (to simulate a 'worst case' test run, lower fig.) the effect is more grave than in the test runs using the $\varphi = 0$ CGBI starting vector (upper fig.). A closer investigation shows that the preconditioners using interpolation are not able to reduce the high frequency parts of the error. This sounds reasonable, as near the boundary of Γ_i, the oscillations of a Chebyshev polynomial of order N cannot be resolved by a trigonometric sum of the same order.

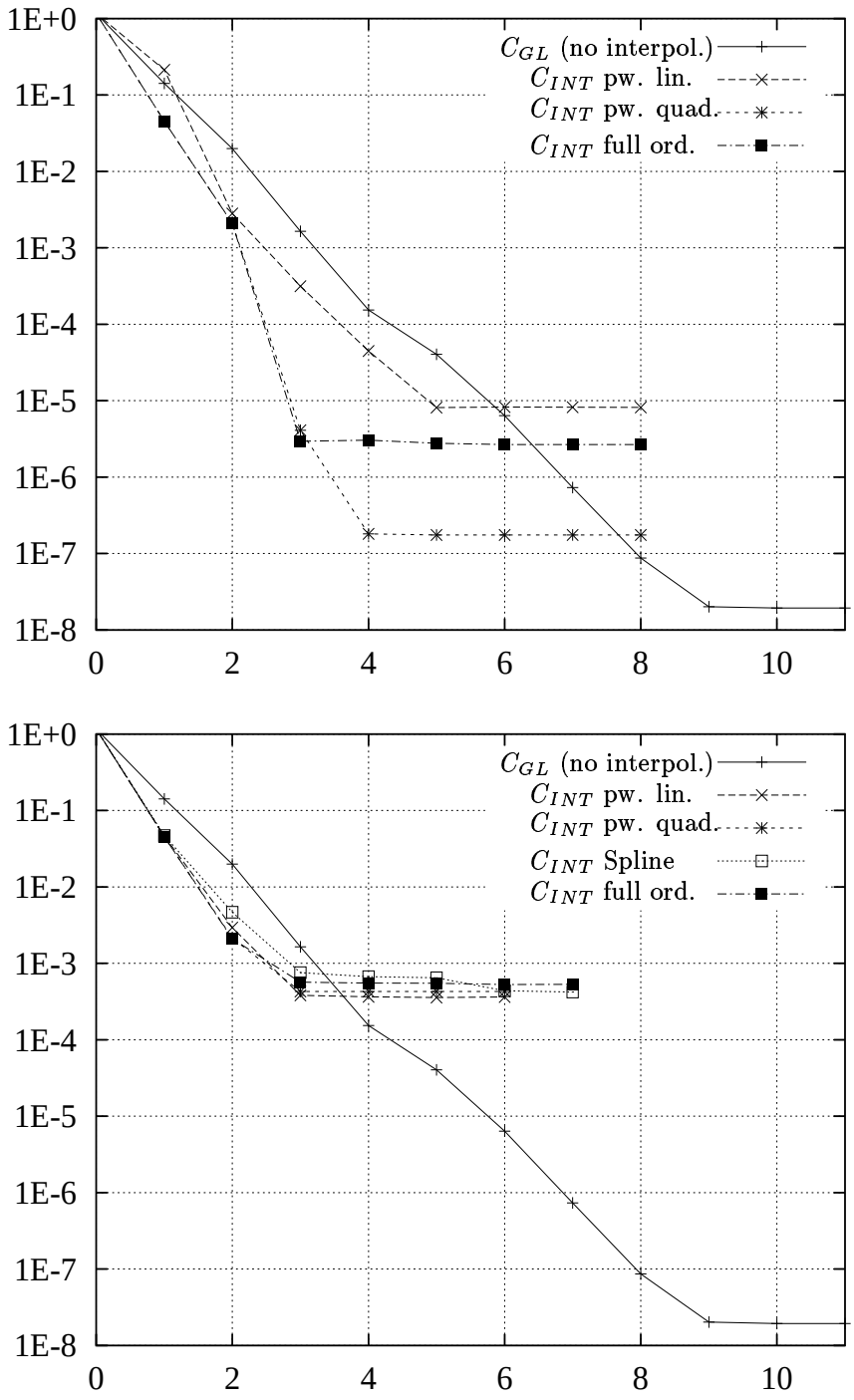


Figure 3.21: Preconditioners on Gauss-Lobatto meshes based on interpolation (piecewise linear, quadratic, cubic spline, full order) and the preconditioner acting on the Gauss-Lobatto (G.L.) mesh. 4 Chebyshev subdomains, $N = 64$, test function 4 (2.83). In the upper figure, CGBI starting vector $\varphi = 0$, in the lower figure, $\varphi \neq 0$ strongly oscillating.

This observation of the inefficiency of the preconditioning method by interpolation is the starting point for the construction of a preconditioner that is acting *directly* on the Gauss-Lobatto mesh (see Sec. 3.1.3).

Furthermore, Fig. 3.21 shows that *in the beginning* of the CGBI process, the interpolation preconditioners have a slightly *better* error reduction property than the Gauss-Lobatto preconditioner.[11] So it may seem reasonable to combine both: In the first CGBI steps, use an interpolation preconditioner; then, if e.g. the residual does not decrease any longer, use the Gauss-Lobatto preconditioner. Such combinations are displayed in Fig. 3.22. Obviously, the saving of CGBI iteration steps is not very large ($\approx$1 iteration step).

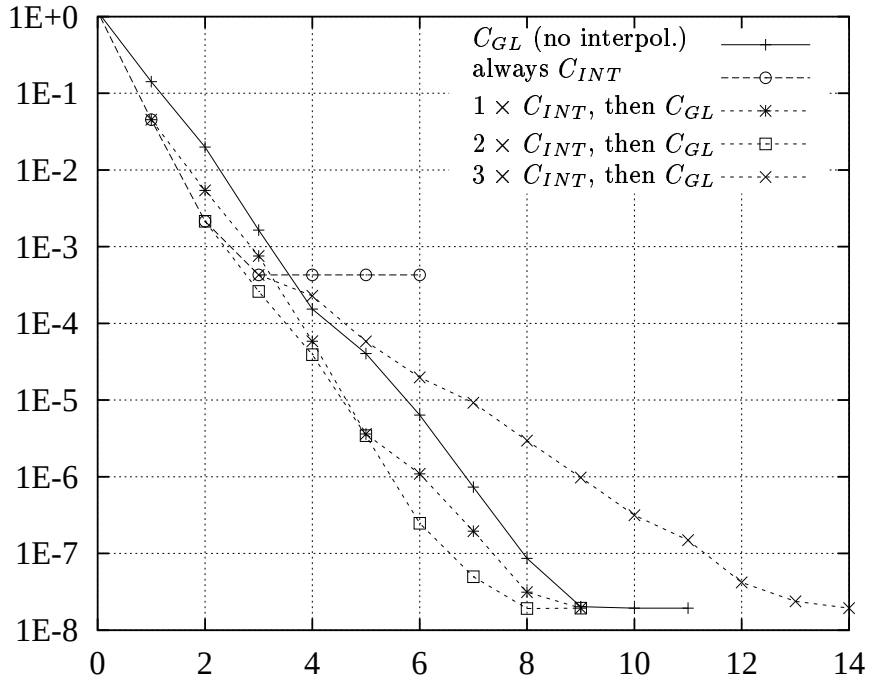


Figure 3.22: A comparison and a combination of an interpolation preconditioner (the one using second order piecewise polynomials) and the preconditioner C_{GL} acting on the Gauss-Lobatto mesh. Oscillating CGBI starting vector. Parameters as in Fig. 3.21. In the first 1, 2, 3 CGBI steps, respectively, C_{INT} is used. After these steps, C_{GL} is used. If C_{INT} is used for the first or for the first two iteration steps, a moderate acceleration of the convergence is reached.

[11] The reason for this is obviously that the preconditioning by interpolation does not make use of the approximation (3.33).

3.3 Preconditioning by convolution

In Section 3.1.4, especially in Theorem 3.13 part (ii) and (iii) it was shown that the method of using a preconditioner *separately* on each $\varphi|_{\Gamma_i}$ becomes inefficient for $r \to 0$. To illustrate this effect we give the following heuristic explanation: The preconditioner C should be an approximation of the operator

$$A^{-1} : \varphi \longrightarrow \frac{\partial}{\partial x}\, \bar{u}(\varphi)|_{\Gamma}$$

where $\bar{u}(\varphi)$ is the solution of the local partial differential equations with the jump φ at the interfaces Γ. $\partial\bar{u}(\varphi)/\partial x|_{\Gamma_i}$ depends on *all* φ_j.[12] The smaller the subdomain aspect ratio r, the closer $\Gamma_{i\pm1}$, $\Gamma_{i\pm2}$,... lie to Γ_i. Thus, the smaller r, the more $\bar{u}|_{\Gamma_i}$ and also $\frac{\partial}{\partial x}\,\bar{u}(\varphi)|_{\Gamma_i}$ should be influenced by $\varphi|_{\Gamma_{i\pm1}}$, $\varphi|_{\Gamma_{i\pm2}}$,... Therefore any approximation C of A^{-1} where $C\varphi$ is independent of $\varphi|_{\Gamma_{i\pm1}}$, $\varphi|_{\Gamma_{i\pm2}}$,... should become bad for $r \to 0$.

Another approach to explain the quality of preconditioning for $r \to 0$ is the following: Beside the computation of the scalar products (i.e. the interprocessor exchange of *scalars*), only next-neighbour communication takes place in CGBI. Therefore $O(p)=O(r^{-1})$ CGBI steps are necessary to spread information over the whole domain Ω. This result of $O(r^{-1})$ necessary CGBI steps corresponds very well to the theoretical result that the condition number is $O(r^{-2})$ (Theorem 3.13). The spreading of information can be accelerated by a preconditioner which uses interprocessor data exchange.

To find such a better preconditioner we will use the following two ideas:

- As in Section 3.1.1 explained (but different from approach (3.59)), we decompose a given φ into the eigenfunctions $\varphi_{k,m}$ (see (3.60)-(3.63)). In the case of equidistant grids we can multiply each coefficient with the exact reciprocal value of the eigenvalue $\lambda_{k,m}$ (see (3.64)-(3.67)) to get an 'ideal' preconditioner $C=A^{-1}$:

$$\begin{aligned} &C : R(A) \longrightarrow D(A), \\ &\varphi = \sum_{k=1}^{\infty}\sum_{m=1}^{p-1} \alpha_{k,m}\,\varphi_{k,m} \longmapsto C\varphi = \sum_{k=1}^{\infty}\sum_{m=1}^{p-1} \alpha_{k,m}\,\lambda_{k,m}^{-1}\,\varphi_{k,m} \end{aligned} \tag{3.103}$$

  If we restrict ourselves to Neumann conditions on $\Gamma^I \cup \Gamma^O$ we can use FFT to perform this decomposition.

  It remains to investigate the Gauss-Lobatto case:

[12] Whereas $(A\varphi)|_{\Gamma_i} = [u(\varphi)]|_{\Gamma_i}$ depends only on φ_{i-1}, φ_i, φ_{i+1}.

- In the case of a Gauss-Lobatto grid, we will decompose a given $\varphi \in H^{1/2}_{mv}(\Gamma)$ into its components $P_1\varphi, ..., P_{p-1}\varphi$ of the $H^{1/2}_{mv}(\Gamma)$-subspaces $S_m := span\{\varphi_{k,m} \mid k \in \mathbb{N}\}$. As we assume Neumann conditions on $\Gamma^I \cup \Gamma^O$, the simplicity of the $\varphi_{k,m}$ (Theorem 3.12) allows us to do this decomposition by FFT.

  Then, *convolution operators* are applied to the $P_m\varphi$. This enables us a finer tuning of the eigenvalues of the resulting preconditioning operator than by only applying $C_{GL,Nm}+c\,id$. The backgrund of this idea is the well known fact that for 2B-periodic functions φ, ψ with the cosine series

$$\varphi(y) = \sum_{k\in\mathbb{N}} \alpha_k \cos\frac{\pi k y}{B}, \quad \psi(y) = \sum_{k\in\mathbb{N}} \beta_k \cos\frac{\pi k y}{B}, \tag{3.104}$$

  the convolution

$$(\varphi * \psi)(y) := \frac{1}{2B}\int_0^{2B} \varphi(y-z)\,\psi(z)\,dz \tag{3.105}$$

  has the cosine series

$$(\varphi * \psi)(y) = \frac{1}{2}\sum_{k\in\mathbb{N}} \alpha_k\beta_k \cos\frac{\pi k y}{B} \tag{3.106}$$

  ((3.106) can be proved by a short straight-forward calculation. A similar relation is given if φ, ψ are represented by sine series.) Let us remark that the *symmetry* of the convolution operator (3.105) is an immediate consequence of (3.106).

In the following we focus on Neumann conditions also on Γ^W and $\sigma=0$. The cases $\sigma>0$ and/or Dirichlet conditions on Γ^W are discussed later on.

By identifying $\Gamma_i=(0,B)$ and 'even extension' $\varphi(B+x)=\varphi(B-x)$ we can assume that every $\varphi_i \in H^{1/2}_{mv}(\Gamma_i)$ is a 2π-periodic function defined on $\mathbb{R}$. The application of the new preconditioner $C_{conv,GL}$ consists of the following steps:

1. Compute the sine-coefficients with respect to m (by FFT) onto the data set of φ given at all the grid points on Γ, i.e. perform the decomposition

$$\varphi(x,y) = \sum_{m=1}^{p-1} P_m\varphi(y)\,\sin\frac{\pi i m}{p}, \quad x = irB, \tag{3.107}$$

   where

$$P_m\varphi := \frac{2}{p}\sum_{i=1}^{p-1} \varphi_i\,\sin\frac{\pi i m}{p}.$$

I will refer to this operation as

$$P\varphi := (P_1\varphi, ..., P_{p-1}\varphi).$$

2. Processor m applies the operator

$$F_m \left(C_{GL,Nm} + \frac{2}{rB}\, id \right) F_m$$

onto $P_m\varphi$ where

$$F_m\varphi := \varphi - \varphi * \psi_m, \tag{3.108}$$

$$\psi_m(y) := 2\sum_{k=1}^{\infty} \exp\left(-\frac{\pi kr}{\sqrt{2(1-\cos\gamma_m)}} \right) \cos\frac{\pi ky}{B}, \tag{3.109}$$

$\gamma_m = \gamma_{k,m}$ from (3.65), the convolution '$*$' from (3.105) and $C_{GL,Nm}$ defined in (3.58). Let us define

$$F(\varphi_1, ..., \varphi_{p-1}) := (F_1\varphi_1, ..., F_{p-1}\varphi_{p-1}). \tag{3.110}$$

3. Apply the reverse sine transformation (by FFT^{-1}) with respect to m onto the data set of $F_m \left(C_{GL,Nm} + \frac{2}{rB}\, id\right) F_m P_m\varphi$, $m = 1, ..., p-1$, i.e. calculate the sums

$$\begin{aligned} C_{conv,GL}\varphi &:= \sum_{m=1}^{p-1} \sin\frac{\pi im}{p} \cdot F_m \left(C_{GL,Nm} + \frac{2}{rB}\, id \right) F_m\, P_m\varphi \\ &= P^{-1}\, F \left(C_{GL,Nm} + \frac{2}{rB}\, id \right) F\, P\, \varphi. \end{aligned} \tag{3.111}$$

Explanation to this preconditioner. Let us consider

$$\begin{aligned} C_{conv,equ}\varphi &:= \sum_{m=1}^{p-1} \sin\frac{\pi im}{p} \cdot F_m \left((-\Delta_{Nm})^{1/2} + \frac{2}{rB}\, id \right) F_m\, P_m\varphi \\ &= P^{-1}\, F \left((-\Delta_{Nm})^{1/2} + \frac{2}{rB}\, id \right) F\, P\, \varphi \end{aligned} \tag{3.112}$$

instead of (3.111) at first. Due to (3.104)-(3.106), F_m has the same set of trigonometric eigenfunctions as $(-\Delta_{Nm})^{1/2} + c\, id$ and the eigenvalues

$$1 - \exp\left(-\frac{\pi kr}{\sqrt{2(1-\cos\gamma_m)}} \right).$$

Thus $F_m\,((-\Delta_{Nm})^{1/2}+\frac{2}{rB}\,id)\,F_m$ possesses the same trigonometric eigenfunctions on Γ_m. Therefore $C_{conv,equ}$ has the same eigenfunctions (3.60) as A, and the eigenvalues

$$g_{k,m} := \frac{1}{rB}\,(2+\pi kr)\,\left(1-\exp\left(-\frac{\pi kr}{\sqrt{2(1-\cos\gamma_m)}}\right)\right)^2. \tag{3.113}$$

The preconditioner $C_{conv,equ}$ is constructed in such a way that its eigenvalues $g_{k,m}$ are approximations of the reciprocals of the eigenvalues $\lambda_{k,m}$ (see (3.64)/(3.65)) of the operator A; in fact we will prove in Theorem 3.14 that

$$c_1 \le g_{k,m}\,\lambda_{k,m} \le c_2, \tag{3.114}$$

with c_1, c_2 not only independent of p, B, L, but also *independent of* r. So, different from Section 3.1, we can expect a condition number independent of r from a discretization of (3.112).

To get a preconditioner which is easy to evaluate on a Gauss-Lobatto grid, we substitute $(-\Delta_{Nm})^{1/2}$ in (3.112) by its approximation $C_{GL,Nm}$ and get (3.111). The independence of the condition number on r is transferred from (3.112) to (3.111): The operators $(-\Delta_{Nm})^{1/2}$ and $C_{GL,Nm}$ are defined locally on the Γ_i, i.e. the relation between their two norms does not depend on r. This is formalized in part (iii) of Theorem 3.14.

Let us remark that for the evaluation of the convolution integrals, a non-equidistant grid is no obstacle:

Evaluation of the convolution integrals (3.105) and the convolution kernel (3.109). To apply the operator F in (3.111) the convolution integrals (3.105) with kernel $\psi=\psi_m$ from (3.109) have to be calculated. In order to save computation time, the series ψ_m was chosen such that its limit can be computed explicitely: Using the abbreviations

$$\omega_1 := \frac{r\pi}{\sqrt{2(1-\cos\gamma_m)}}, \quad \omega_2 := \frac{\pi y}{B}$$

we get the expression

$$\begin{aligned}\psi_m(y) &= 2\sum_{k=1}^{\infty} e^{-\omega_1 k}\cos\omega_2 k = 2\,\mathrm{Re}\left(\sum_{k=1}^{\infty}\exp((-\omega_1+i\,\omega_2)k)\right)\\ &= 2\,\mathrm{Re}\left(\frac{\exp(-\omega_1+i\,\omega_2)}{1-\exp(-\omega_1+i\,\omega_2)}\right) = 2\,\frac{e^{-\omega_1}\cos\omega_2 - e^{-2\omega_1}}{1-2\,e^{-\omega_1}\cos\omega_2+e^{-2\omega_1}}\end{aligned} \tag{3.115}$$

for the convolution kernel. ψ_m is visualized in Fig. 3.23.

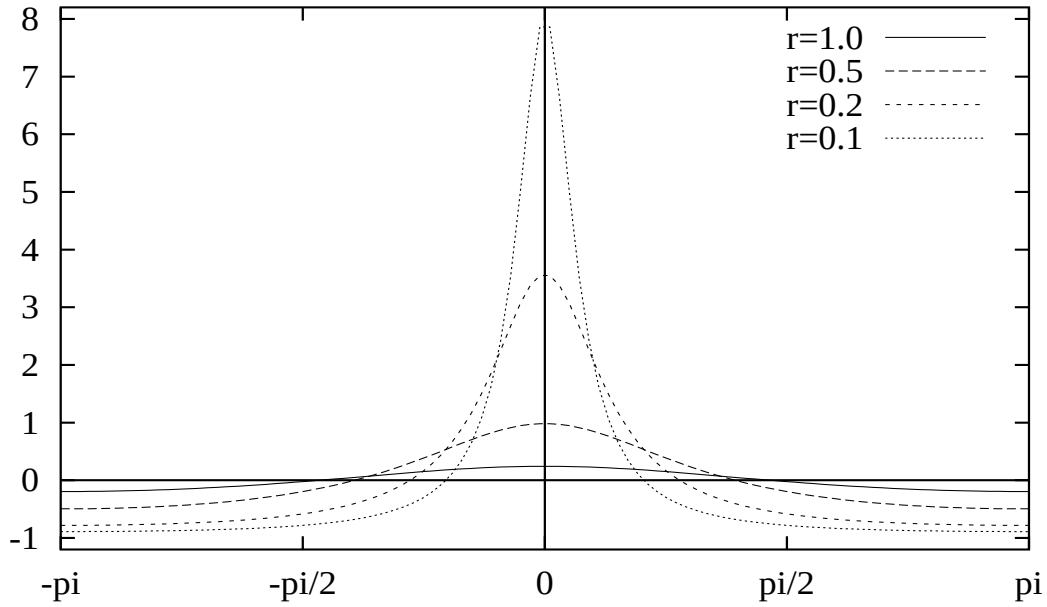


Figure 3.23: Visualization of the convolution kernel ψ_m for $m = p/2$ and $r = 1.0, 0.5, 0.2, 0.1$. The smaller r the steeper ψ_m and the higher discretization errors are to expect.

To compute the convolution integrals (3.105) with high accuracy, we use the transformation $z = \frac{B}{2}(1+\cos\bar{z})$ which maps the Gauss-Lobatto points $z_j \in [0, B]$ onto equidistant points $\bar{z}_j \in [0, \pi]$:

$$\varphi * \psi_m(y) = \frac{1}{2} \int_0^\pi \varphi(\frac{B}{2}(1+\cos z))\ \psi_m(\frac{B}{2}(1+\cos z) - y)\ \sin z\ dz. \tag{3.116}$$

This is discretized by the trapezoid rule with respect to the equidistant grid:

$$\varphi * \psi_m(y_k) = \frac{\pi}{2N} \sum_{j=1}^{N-1} \varphi(y_j)\ \psi_m(y_j - y_k)\ \sin\frac{j\pi}{N} \tag{3.117}$$

Bevore CGBI is started the values of $\psi_m(y_j - y_k)$ are stored in an array to optimize the evaluation.

Let us mention that $\varphi_m \in H^{1/2}_{mv}(\Gamma_m)$ implies

$$\varphi_m * \psi_m \in H^{1/2}_{mv}(\Gamma_m)\ : \tag{3.118}$$

$\varphi_m * \psi_m \in H^{1/2}(\Gamma_m)$ follows from the fact that ψ_m is in $C^\infty(\mathbb{R})$ (see (3.115)). $\int_0^{2B} \varphi_m * \psi_m(y)\, dy = 0$ is a consequence of (3.104)-(3.106).

Remark on the costs of this preconditioner. As this preconditioner should be used in the case $r \ll 1$ only, we should consider the case that the number of grid

points in x-direction N_x differs from the number of grid points in y-direction N_y. The application of FFT with respect to m in steps 1 and 3 requires $O(p\, N_y\, \log p)$ operations. FFT with respect to y requires $O(N_y\, \log N_y)$ operations per processor. The calculation of the convolution integrals for the N_y+1 grid points $x \in \Gamma_m$ requires $O(N_y^2)$ operations, as each integral is numerically calculated with respect to the N_y+1 grid points. So we arrive at

$$O(N_y\, (N_y + p\, \log p)) \tag{3.119}$$

operations which is - under the reasonable assumption $p \log p \lesssim N_X N_Y$ - neglectable compared to

$$O(N_x\, N_y\, (N_x + N_y)) \tag{3.120}$$

of the local Chebyshev solver.

In the case of equidistant grids the calculation of the convolution integrals is dropped as the multiplication of the eigenfunctions by $\lambda_{k,m}^{-1}$ can be applied directly. Therefore in this case the total preconditioning costs are

$$O(N_y\, (\log N_y + p\, \log p)). \tag{3.121}$$

This is tolerable compared to the costs of the local Multigrid solvers as long as $\log N_Y \lesssim N_X$ and $p \lesssim N_X$.

In (3.119) and (3.121), the term $O(p \log p)$ can be reduced to $O(\log p)$ by distributing the application of FFT among the p processors. However, this requires additional communication.

Theorem 3.14 *Let us consider Neumann boundary conditions on $\partial\Omega$.*

(i) Let $\sigma = 0$. The preconditioner $C_{conv,equ}$ (3.112) concatenated with the operator A^{-1} produces an operator with real eigenvalues μ_k which are bounded by

$$0 < c_1 \le \mu_k \le c_2$$

with c_1, c_2 independent of r, p, B, L.

(ii) Let $\sigma \ge 0$. Let F^σ be defined similar to F in (3.108)-(3.110), but with ψ_m replaced by $exp(-\frac{rB\sqrt{\sigma}}{\sqrt{2q}}) \cdot \psi_m$. The preconditioner $C_{conv,equ,\sigma} = P^{-1}\, F^\sigma\, ((-\Delta)^{1/2} + (\frac{2}{rB} + \sqrt{\sigma})\, id)\, F^\sigma\, P$ concatenated with the operator A^{-1} produces an operator with real eigenvalues μ_k which are bounded by

$$0 < c_1 \le \mu_k \le c_2$$

with c_1, c_2 independent of r, p, B, L and σ.

(iii) *Let* $\sigma = 0$. *If on a subspace* $\mathcal{S}(\Gamma_i) \subset H^{1/2}_{mv}(\Gamma_i)$ *the estimate*

$$\underline{c}_1(\mathcal{S})\ \langle(-\Delta_{Nm})^{1/2}\varphi,\varphi)\rangle_{\Gamma_i} \le \langle C_{GL,Nm}\varphi,\varphi\rangle_{\Gamma_i} \le \overline{c}_1(\mathcal{S})\ \langle(-\Delta_{Nm})^{1/2}\varphi,\varphi)\rangle_{\Gamma_i} \quad \forall\ \varphi\in\mathcal{S}(\Gamma_i) \qquad (3.122)$$

holds, then

$$\underline{c}_2\ \underline{c}_1(\mathcal{S})\ \langle A^{-1}\varphi,\varphi)\rangle_{\Gamma_i} \le \langle C_{conv,GL}\varphi,\varphi\rangle_{\Gamma_i}) \le \overline{c}_2\ \overline{c}_1(\mathcal{S})\ \langle A^{-1}\varphi,\varphi)\rangle_{\Gamma_i} \quad \forall\ \varphi\in\mathcal{S}(\Gamma_i) \qquad (3.123)$$

holds.

(iv) *Let* $\sigma \ge 0$. *If (3.122) holds, then estimate (3.123) holds for* $C_{conv,GL}$ *replaced by* $C_{conv,GL,\sigma} := P^{-1}\,F^{\sigma}\,(C_{GL,Nm} + (\frac{2}{rB} + \sqrt{\sigma})\,id)\,F^{\sigma}\,P$.

Proof. Ad (i). Obviously, $C_{conv,equ}$ multiplies each eigenfunction $\varphi_{k,m}$ by the factor $g_{k,m}$ from (3.113). We have to check that

$$0 < c_1 \le \lambda_{k,m}\,g_{k,m} \le c_2 \qquad (3.124)$$

with c_1, c_2 independent of k, m, r, p, B, L. If we substitute $K := \pi kr$ and $q := 1 - \cos\frac{\pi m}{p}$, it is sufficient to prove that the function

$$F(K,q) := (2+K)\ \left(1 - \exp\left(-\frac{K}{\sqrt{2q}}\right)\right)^2\ \frac{\cosh K - 1 + q}{K\ \sinh K} \qquad (3.125)$$

is bounded from above and below by positive constants on $(0,\infty)\times(0,2)$. Assigning $\exp(-\frac{c}{0}) := 0$ for all $c>0$, F is defined and continuous on $S := (0,\infty)\times[0,2]$. It remains to prove that for $(K,q) \to (0,q_0)$ and $(K,q) \to (\infty,q_0)$, $q_0 \in [0,2]$, $F(K,q)$ is bounded by constants $0 < c_1 \le F(K,q) \le c_2$.

Let us consider $(K,q) \to (\infty,q_0)$, $q_0 \in [0,2]$. Obviously, $F(K,q) \to 1$ in this case. Now the case $(K,q) \to (0,q_0)$, $q_0 \in (0,2]$. A Taylor expansion of exp, sinh, cosh shows that again $F(K,q) \to 1$. The case $(K,q) \to (0,0)$ remains: Suppose that there is a sequence $(K_j,q_j)_{j\in\mathbb{N}}$ in S with $(K_j,q_j) \to (0,0)$ and $F(K_j,q_j) \to 0$ or $F(K_j,q_j)$ unbounded. There is a subsequence, again denoted by (K_j,q_j), such that

$$\frac{K_j}{\sqrt{2q_j}} \to c \in [0,\infty].$$

If $c=\infty$, $F(K_j,q_j) \to 1$ again. If $c=0$, Taylor expansion of exp, sinh, cosh shows that $F(K_j,q_j) \to 1$. If $c\in(0,\infty)$, then

$$\lim_{j\to\infty} F(K_j,q_j) = 2\,(1-e^{-c})^2\ \left(\frac{1}{2} + \frac{1}{2c^2}\right)$$

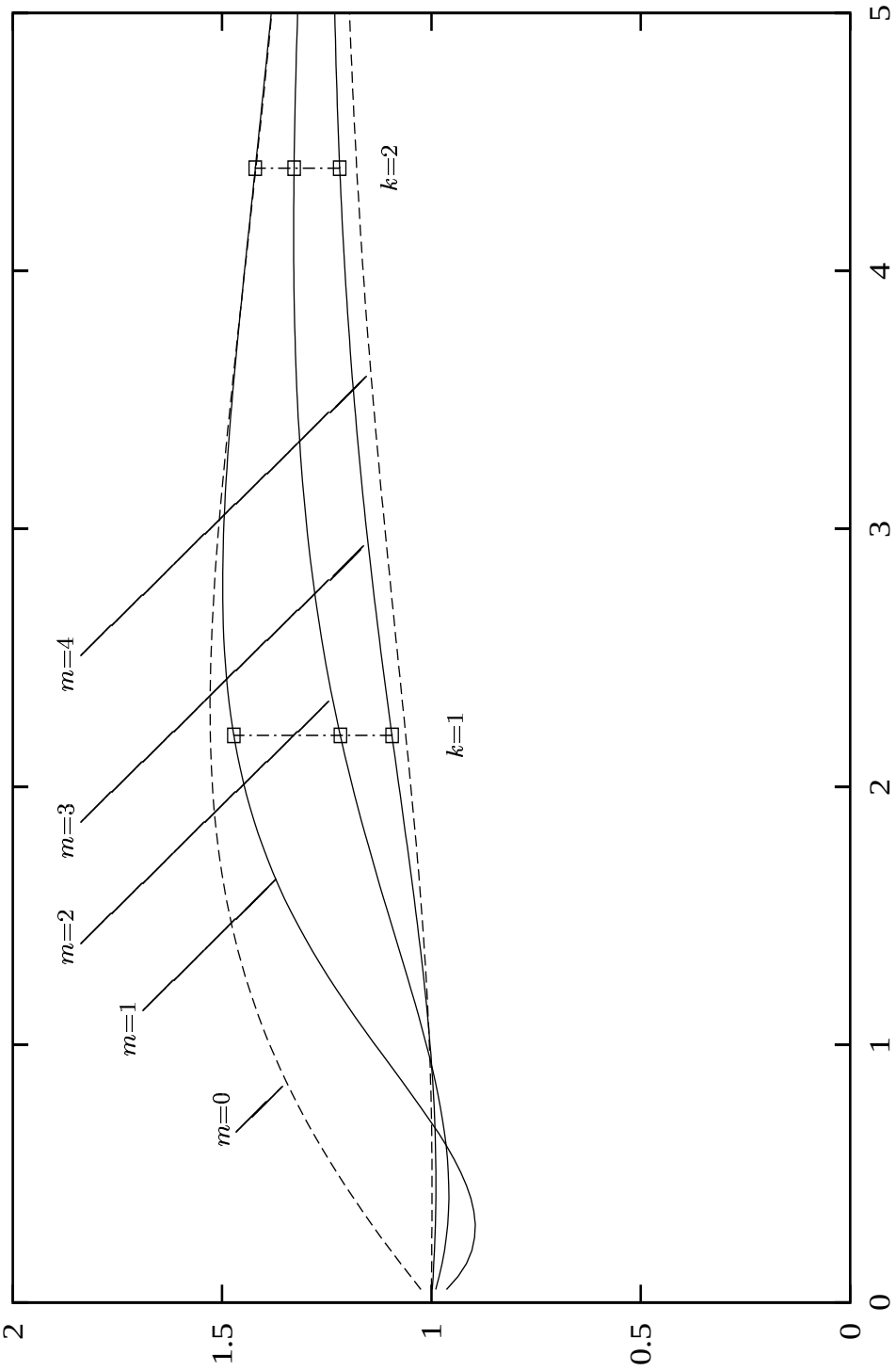


Figure 3.24: Visualizations of the eigenvalues ('dots') of the preconditioned operator $C_{equ,GL}A^{-1}$ for $\sigma = 0$, $r = 0.7$, $p = 4$. $K := \pi kr$ on the horizontal axis. Compare to Fig 3.9 where the simpler preconditioner $(-\Delta_{Nm})^{1/2}$ is used. As r is rather large, both preconditioners produce a similar distribution of the eigenvalues. This is not the case for $r = 0.1$ (Fig. 3.25).

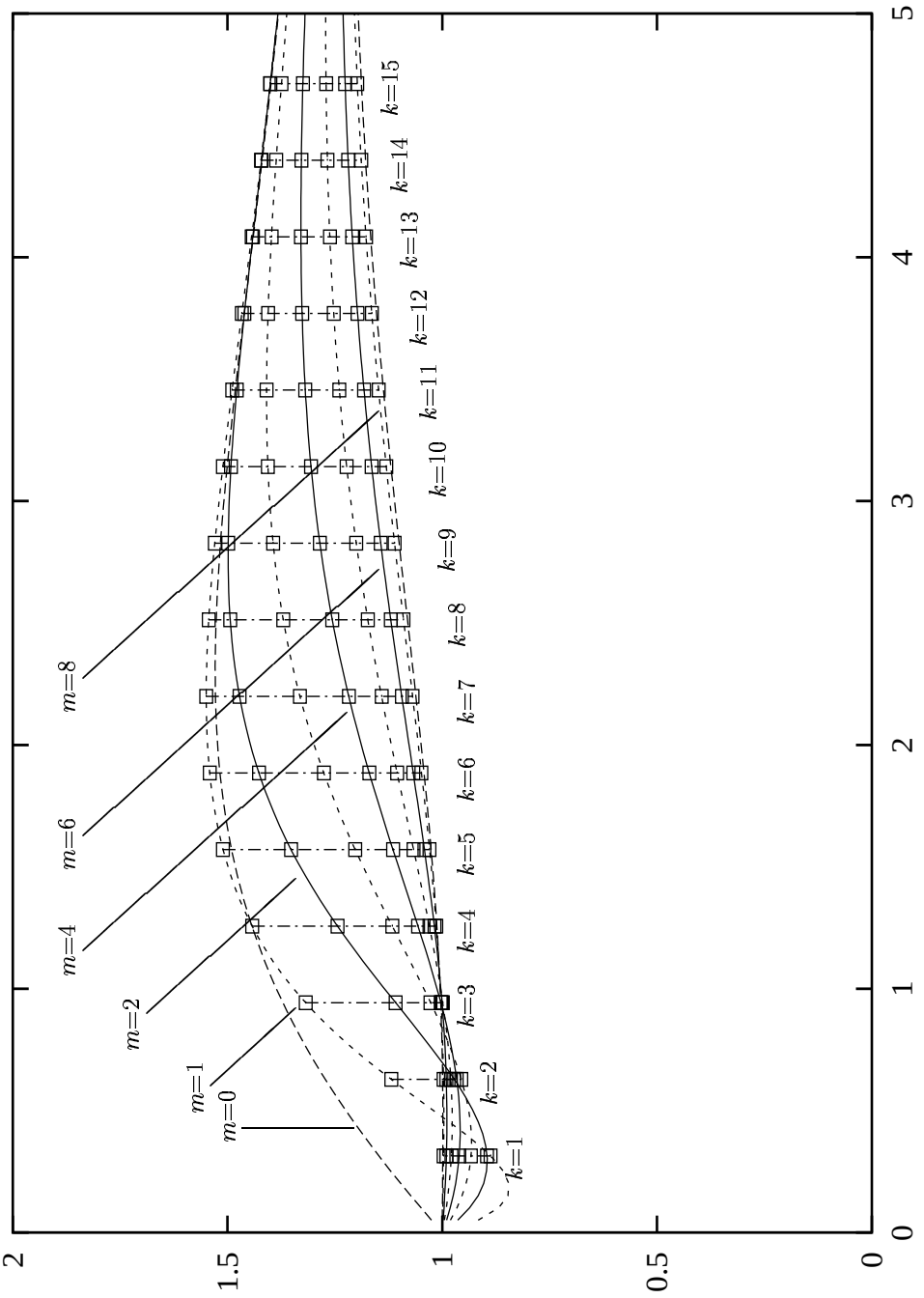


Figure 3.25: Visualizations of the eigenvalues ('dots') of the preconditioned operator $C_{conv,equ}A^{-1}$ for $\sigma = 0$, $r = 0.1$, $p = 8$. $K := \pi kr$ on the horizontal axis. Compare to Fig 3.10 where the simpler preconditioner $(-\Delta_{Nm})^{1/2}$ is used and to Fig. 3.24 where a larger r is used. As Theorem 3.14 predicted the ratio of the biggest and the smallest eigenvalue now is bounded independent of r. The figure shows that $c_2/c_1 \approx 2$. The comparison with Fig. 3.24 shows that we should expect that the condition number using the convolution operator is better than the condition number using simply $C_{GL,Nm}$ if r is smaller than ≈ 0.6.

which is bounded from above and below for all $c \in (0, \infty)$. So the assumption is false.

Ad (ii). Obviously $C_{conv,equ,\sigma}$ multiplies each eigenfunction (3.60) by

$$\frac{1}{rB}\,(2 + rB\sqrt{\sigma} + r\pi k)\left(1 - \exp\left(-\frac{rB\sqrt{\sigma} + r\pi k}{\sqrt{2q}}\right)\right)^2.$$

So complying with (3.64)-(3.65) we have to investigate the bounds of the function

$$F(K_1, K_2, q) = (2 + K_2)\,(1 - e^{-\frac{K_2}{\sqrt{2q}}})^2\,\frac{\cosh K_1 - 1 + q}{K_1 \sinh K_1} \tag{3.126}$$

where we have put

$$K_1 := r\sqrt{B^2\sigma + \pi^2 k^2}, \quad K_2 := rB\sqrt{\sigma} + r\pi k.$$

As already stated in the proof of Theorem 3.13 (ii), $1 \le \frac{K_2}{K_1} \le \sqrt{2}$ holds. So it is sufficient to find the infimum and the supremum of F on

$$S := \{(K_1, K_2, q) \in (0, \infty) \times (0, \infty) \times [0, 2] \mid 1 \le \frac{K_2}{K_1} \le \sqrt{2}\}.$$

F is continuous and strictly positive on S. Suppose F is not bounded on S by strictly positive constants. Then there is a sequence $(K_{1,i}, K_{2,i}, q_i)$ in S with $\lim K_{1,i} = \lim K_{2,i} \in \{0, \infty\}$, $\lim q_i = q_0 \in [0, 2]$ and $\lim F(K_{1,i}, K_{2,i}, q) \in \{0, \infty\}$.

Let $K_{1,i} \to \infty$, $K_{2,i} \to \infty$ at first. Then, due to (3.126),

$$\lim F(K_{1,i}, K_{2,i}, q) \le \limsup \frac{K_{2,i}}{K_{1,i}} \le \sqrt{2}$$

and

$$\lim F(K_{1,i}, K_{2,i}, q) \ge \liminf \frac{K_{2,i}}{K_{1,i}} \ge 1.$$

Now let $K_{1,i} \to 0$, $K_{2,i} \to 0$. There are three cases:

If $\lim \frac{K_{2,i}}{\sqrt{2q_i}} = 0$ then

$$\lim F(K_{1,i}, K_{2,i}, q_i) \le 2 \limsup \frac{K_{2,i}^2}{2q_i} \cdot \frac{\frac{K_{1,i}^2}{2} + q_i}{K_{1,i}^2} = \limsup \frac{K_{2,i}^2}{K_{1,i}^2} \le 2$$

and analogously

$$\lim F(K_{1,i}, K_{2,i}, q_i) \ge 1.$$

If $\lim \frac{K_{2,i}}{\sqrt{2q_i}} = \infty$ then

$$\lim \frac{K_{1,i}}{\sqrt{2q_i}} = \infty$$

and

$$\lim F(K_{1,i}, K_{2,i}, q_i) = 2 \lim \frac{\frac{K_{1,i}^2}{2} + q_i}{K_{1,i}^2} = 1.$$

If $\lim \frac{K_{2,i}}{\sqrt{2q_i}} = c \in (0, \infty)$ then

$$\lim F(K_{1,i}, K_{2,i}, q_i) = 2\,(1 - e^{-c})^2 \left(\frac{1}{2} + \lim \frac{q_i}{K_{1,i}^2}\right)$$

where $\frac{1}{2c^2} \leq \lim \frac{q_i}{K_{1,i}^2} \leq \frac{1}{c^2}$. So $F(K_{1,i}, K_{2,i}, q_i)$ is bounded in this case, too, and the assumption must be false.

ad (iii). We already know that $\psi := FP\varphi \in H_{mv}^{1/2}(\Gamma)$ for $\varphi \in H_{mv}^{1/2}(\Gamma)$ (see (3.118)). Due to the assumption,

$$\langle (C_{GL,Nm} + 2\,id)\psi, \psi \rangle_\Gamma \leq \overline{c}_1 \, \langle ((-\Delta_{Nm})^{1/2} + 2\,id)\psi, \psi \rangle_\Gamma$$

follows. Furthermore,

$$\langle C_{conv,eq}\varphi, \varphi \rangle_\Gamma \leq \overline{c}_2 \, \langle A^{-1}\varphi, \varphi \rangle_\Gamma$$

is known from (i). As $P^{-1} = P^T$ and $F^T = F$ we get

$$\begin{aligned} \langle C_{conv,GL}\varphi, \varphi \rangle_\Gamma &= \langle (C_{GL,Nm} + 2\,id)\, FP\varphi, FP\varphi \rangle_\Gamma \\ &\leq \overline{c}_1 \, \langle ((-\Delta_{Nm})^{1/2} + 2\,id)\, FP\varphi, FP\varphi \rangle_\Gamma \\ &= \overline{c}_1 \, \langle C_{conv,equ}\varphi, \varphi \rangle_\Gamma \leq \overline{c}_1\, \overline{c}_2 \, \langle A^{-1}\varphi, \varphi \rangle_\Gamma. \end{aligned}$$

The other direction

$$\langle C_{conv,GL}\varphi, \varphi \rangle_\Gamma \geq \underline{c}_1\, \underline{c}_2 \, \langle A^{-1}\varphi, \varphi \rangle_\Gamma$$

is proved analogously.

Ad (iv). Analogous to (iii). ∎

Remark. Theorem 3.14 lets us expect that the discrete preconditioners associated with (i)-(iv) produce a condition number independent of r; and at least for preconditioners in (i) and (iii), independent of p, B, L, $N(, \sigma)$. (iii) and (iv) show that the replacement of $(-\Delta)^{1/2}$ in $C_{conv,equ}$, $C_{conv,equ,\sigma}$ by a spectrally equivalent operator does not influence the quality of the preconditioner much. As it is not sure if (3.122) holds on $H_{mv}^{1/2}(\Gamma)$ (see Sec. 3.1.3.3), (iii) and (iv) are formulated on a (e.g. finite-dimensional 'discrete') subspace of $H_{mv}^{1/2}(\Gamma)$.

The assertions of Theorem 3.14 still hold if the term $2\,id$ is replaced by $c\,id$, $c > 0$, if in the definition (3.109) of ψ_m the factor 2 is replaced by the same factor c. Numerical evaluation of the eigenvalues show that for $c = 2$ the condition number is quite small compared to very large or very small values of c.

Test runs. A comparison of the convergence rate of the two preconditioners

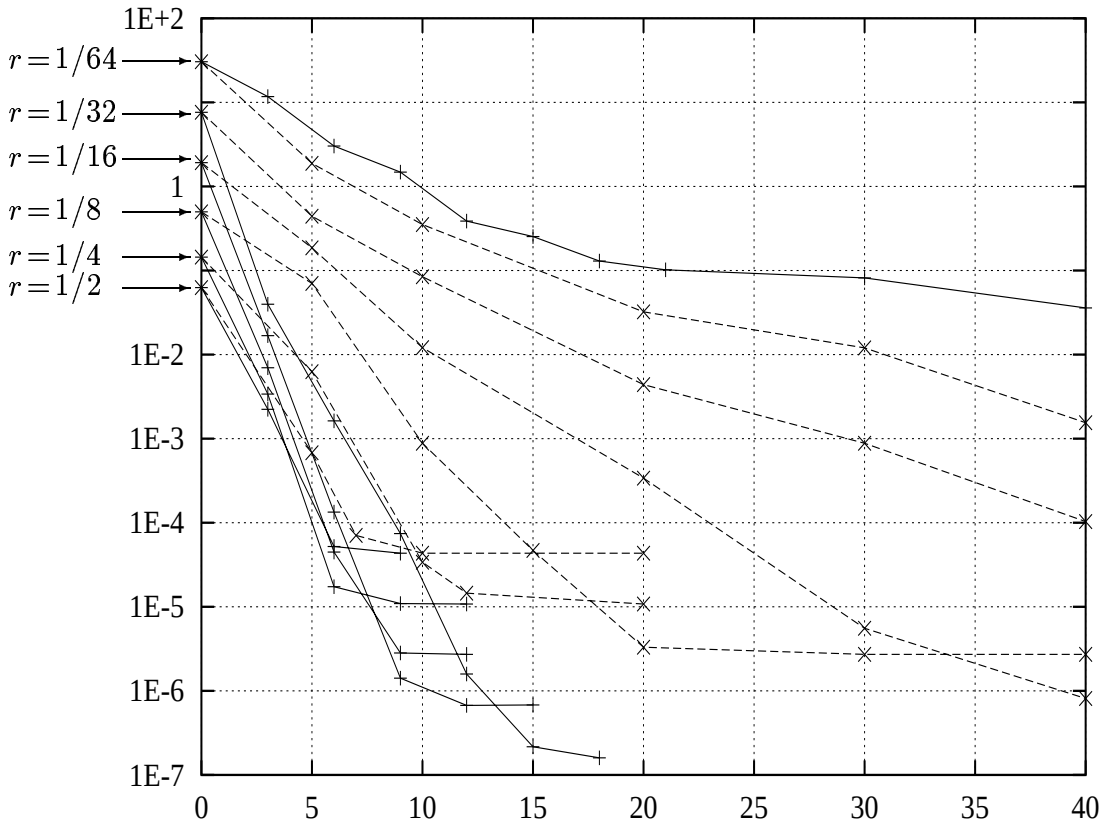


Figure 3.26: Error decay with preconditioner $C_{conv,GL}$ (full lines) compared to $C_{GL,Nm}$ (broken lines) for the solution (2.81), $p=8$ processors, $N=64$, oscillating CG starting vector and $r = 1/2, 1/4, 1/8, 1/16, 1/32, 1/64$.

$C_{GL,Nm}$ and $C_{conv,GL}$ was made in Fig. 3.26. For all the tests, $N=N_x=N_y=64$, $p = 8$ and an oscillating CG starting vector was used ('worst case' test run). The diagramme shows for example that for $r = 1/16$ there are 15 CG steps with $C_{GL,Nm}$ necessary, but only 4 with $C_{conv,GL}$ to get an error reduction of three powers of 10. To investigate the behaviour for $r \to 0$, different values $r = \frac{1}{2}, \frac{1}{4}, ..., \frac{1}{64} = N^{-1}$ were used. As expected (Theorem 3.13 (ii)) the convergence rate drops for $C_{GL,Nm}$, $r \to 0$. For $C_{conv,GL}$, however, the convergence rate is constant up to $r \approx 32^{-1} = 2N^{-1}$. Only for smaller r (which is not reasonable in practical applications anyway) the numerical evaluation errors of the convolution integrals for the highly oscillating modes cause a bad result. Calculations with various N show that the minimum r_{min} which is necessary to obtain good results depends linearly on N^{-1}, $r_{min} \approx 2N^{-1}$.

Fig. 3.27 handles the equidistant mesh case. The new preconditioner $C=A^{-1}$ (3.103) is compared with the old preconditioner $C=(-\Delta_{Nm})^{1/2}$ for a subdomain aspect ratio $r = 1$. For such a moderate r, the old preconditioner is already

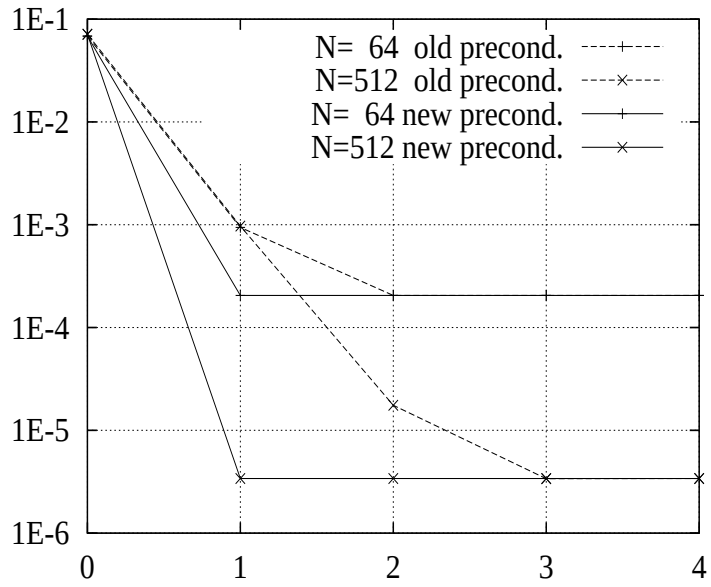


Figure 3.27: The equidistant mesh case. Error decay with preconditioner (3.103) (full lines) compared to $(-\Delta_{Nm})^{1/2}$ (broken lines) for the solution (2.81), $p=4$ subdomains with FDM solver, $r=1$, $N=64$ and $N=512$, zero CG starting vector.

very efficient; the solution is found after 2 rsp. 3 CGBI steps. However, the new preconditioner always finds the solution in the *first* CGBI step. In fact, this preconditioner turns CGBI into an *explicit* method.

Remark on the symmetry of $\mathbf{C_{conv,GL}}$. The symmetry of F_m in $H^{1/2}_{mv}(0,B)$ follows directly from the symmetry of the convolution. Thus, F is symmetric in $H^{1/2}_{mv}(\Gamma)$. $C_{conv,GL}$ is symmetric because $C_{GL,Nm}+c\,id$ and F are symmetric.

The replacement of the *two* time-consuming applications of the convolutions F_m by one convolution $\tilde{F}_m$ seems to be difficult: The sole concatenation of $C_{GL,Nm}+\frac{2}{rB}\,id$ and F_m^2 is *not* symmetric, as the systems of eigenfunctions differ: $C_{GL,Nm}+\frac{2}{rB}\,id$ has polynomial eigenfunctions, F_m has trigonometric eigenfunctions; and the same holds if we replace F_m^2 by the approximation[13] $\tilde{F}_m$,

$$\begin{aligned}
\tilde{F}_m\varphi &:= \varphi - \varphi * \tilde{\psi}_m, \\
\tilde{\psi}_m(x) &:= 2\sum_{k=1}^{\infty} \exp\left(-\frac{(\pi kr)^2}{2(1-\cos\gamma_m)}\right)\,\cos\frac{\pi kx}{B}, \\
\tilde{F}\varphi &:= (\tilde{F}_1\varphi_1, ..., \tilde{F}_{p-1}\varphi_{p-1}).
\end{aligned} \tag{3.127}$$

[13] $\tilde{F}_m$ is an approximation of F_m^2, in the sense that both operators have the same set of cosine eigenfunctions, and the eigenvalues $(1-\exp(-\omega_1 k))^2$ of F_m^2 are approximated by the eigenvalues $1-\exp(-\omega_1^2 k^2)$ of $\tilde{F}_m$. In fact, part (i) of Theorem 3.14 stays valid if we replace the two applications of F by one application of $\tilde{F}$ as in (3.130), (3.131). The proof is similar.

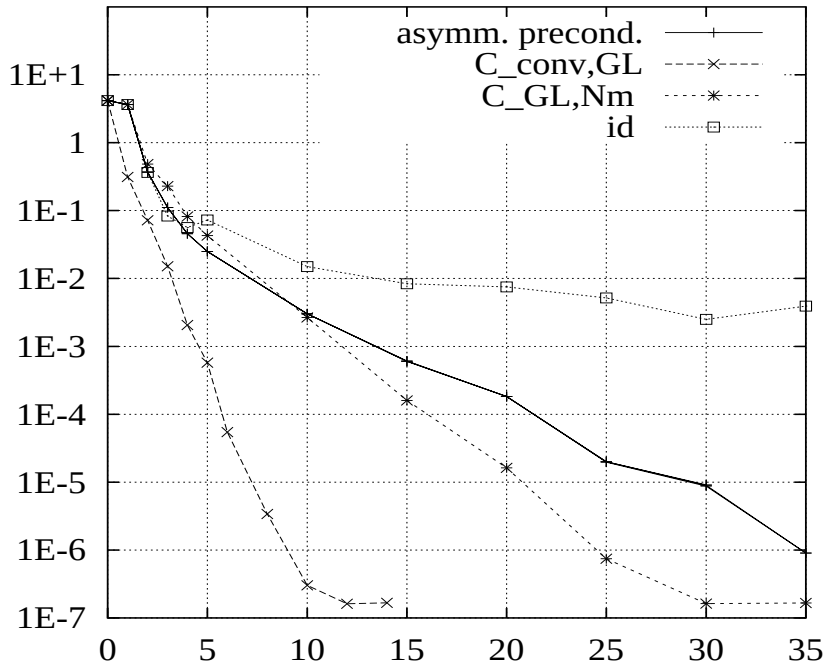


Figure 3.28: Comparison of the different preconditioners id, $C_{GL,Nm}$, $C_{conv,GL}$ and the asymmetric preconditioners (3.128)-(3.131). All the asymmetric preconditioners produce very similar results, so there is only one curve. $p=4$ subdomains, $N_X = N_Y = 64$, solution (2.81), oscillating CG starting vector and $r = 1/16$.

Nevertheless, numerical test were made with

$$C_{conv,GL,1} := P^{-1} \left(C_{GL,Nm} + \frac{2}{rB}\, id \right) F^2\, P\, \varphi \tag{3.128}$$

$$C_{conv,GL,2} := P^{-1}\, F^2 \left(C_{GL,Nm} + \frac{2}{rB}\, id \right) P\, \varphi \tag{3.129}$$

$$C_{conv,GL,3} := P^{-1} \left(C_{GL,Nm} + \frac{2}{rB}\, id \right) \tilde{F}\, P\, \varphi \tag{3.130}$$

$$C_{conv,GL,4} := P^{-1}\, \tilde{F} \left(C_{GL,Nm} + \frac{2}{rB}\, id \right) P\, \varphi \tag{3.131}$$

(see Fig. 3.28) though these operators are not really symmetric in $H^{1/2}_{mv}(\Gamma)$.

The substitution of the two operators F in (3.111), (3.128), (3.129) by the *one* operator $\tilde{F}$ in (3.130), (3.131) reduces the number of operations for the preconditioning to roughly one half. For $\tilde{F}$ in (3.130), (3.131), The values $\tilde{\psi}_m(z_j, z_k)$ have to be computed numerically. This is justifiable because the series (3.127) decreases rapidly.

However, Fig. 3.28 shows that the convergence rate drops drastically for these non-symmetric preconditioning matrices: The non-symmetric convolution pre-

conditioners (3.128)-(3.131) all produce very similar results, so there appears only one line. Especially after some CG steps, the convergence rate drops even below the convergence rate of the simple $C_{GL,Nm}$ preconditioner. Although the equidistant case analogons of (3.128)-(3.131),

$$\begin{aligned} &P^{-1}\,F^2\,((-\Delta_{Nm})^{1/2}+2\,id)\,P, \qquad P^{-1}\,F^2\,((-\Delta_{Nm})^{1/2}+2\,id)\,P, \\ &P^{-1}\,\tilde{F}\,((-\Delta_{Nm})^{1/2}+2\,id)\,P, \qquad P^{-1}\,((-\Delta_{Nm})^{1/2}+2\,id)\,\tilde{F}\,P, \end{aligned} \tag{3.132}$$

all are *identical* to $P^{-1}F\,((-\Delta_{Nm})^{1/2}+2id)\,FP$. We see that the symmetry of the preconditioning matrix is distorted too much to get a good convergence rate.

Preconditioning for *Dirichlet* boundary conditions. For Dirichlet conditions posed on Γ^W the results of this section, especially Theorem 3.14, are still valid (with C_{GL} instead of $C_{GL,Nm}$) if we replace $(-\Delta_{Nm})^{1/2}$ by $(-\Delta_0)^{1/2}$ (having sine instead of cosine eigenfunctions). This exchange does *not* influence the definition (3.109) of ψ_m as a *cosine* series; property (3.104)-(3.106) is replaced by

$$\begin{aligned} &\varphi(y) = \sum_{k\in\mathbb{N}} \alpha_k \,\sin\frac{\pi k y}{B}, \qquad \psi(y) = \sum_{k\in\mathbb{N}} \beta_k \,\cos\frac{\pi k y}{B} \\ &\Longrightarrow \quad (\varphi * \psi)(y) = \frac{1}{2}\sum_{k\in\mathbb{N}} \alpha_k\beta_k \,\sin\frac{\pi k y}{B} \end{aligned}$$

If Dirichlet boundary conditions are posed on $\Gamma^I \cup \Gamma^O$, the algorithm developed in this section fails because the decomposition (3.107) cannot be performed easily. Instead, the *complete* decomposition

$$\varphi(x,y) = \sum_{m=1}^{p-1}\sum_{k=1}^{\infty} \alpha_{k,m}\,\varphi_{k,m}(x,y) \tag{3.133}$$

into the eigenfunctions (3.62) rsp. (3.63) can be calculated by performing the numerical integration

$$\alpha_{k,m} = \frac{(\varphi_{k,m},\varphi)_{L^2(\Gamma)}}{\|\varphi_{k,m}\|^2_{L^2(\Gamma)}} = \frac{\sum_{i=1}^{p-1}\int_{\Gamma_i}\varphi_{k,m}(x,y)\,\varphi(x,y)\,dy}{\sum_{i=1}^{p-1}\int_{\Gamma_i}\varphi^2_{k,m}(x,y)\,dy}, \quad x = irB$$

after applying the frequently (e.g. in (3.116)) used integral transformation. Then, each summand is multiplied by $\lambda^{-1}_{k,m}$ from (3.64), (3.66)-(3.67) and the summation on the right hand side of (3.103) is performed.

This procedure takes $O(p^2N_y^2)$ operations. Distributed among the p processors $O(pN_y^2)$. This may be more than the costs of the local Chebyshev solver (3.120) if $p \gg N_x$. But the amount of CG iteration steps reduces from $O(p)$ (see Theorem 3.13 (ii) for $p\,r = const$) to $O(1)$ (Theorem 3.14). So the total work is

$$O(p\,N_x\,N_y^2)) \quad (\text{for } N_X \le N_Y)$$

with the conventional preconditioner C_{GL} compared to

$$O(N_x N_y^2 + p N_y^2)$$

with this new preconditioner. So the new preconditioner is efficient if $p \gg 1$ or $N_x \gg 1$. However, the 'elegance' of the Neumann case preconditioner is obviously lost.

3.4 Preconditioning by sparse matrices

3.4.1 A first approach by tridiagonal matrices

As pointed out in the previous chapters the most important thing to construct a preconditioner is to find a numerical approximation of $(-\Delta)^{1/2}$ on a given grid. Beside the spectral approach of Section 3.1 there is the possibility to use a matrix approach: We may search for a symmetric positive definite matrix C_{FD} being an approximation of the square root of the negative Laplacian operator. Furthermore, it would be nice if C_{FD} is a matrix with limited bandwidth. This would reduce the number of operations necessary to apply C_{FD}.

For the sake of simplicity we assume that for the width of the channel $B=\pi$ holds.

To compare the operator[14] $(-\Delta)^{1/2}$ to a discrete approximation C_{FD}, we introduce the function space

$$\mathcal{T}_{N,0} := \{\sum_{k=1}^{N-1} c_k \sin kx \mid c_k \in \mathbb{R}\}$$

for the case of Dirichlet b.c. on Γ^W and

$$\mathcal{T}_{N,Nm} := \{\sum_{k=1}^{N} c_k \cos kx \mid c_k \in \mathbb{R}\}$$

for the case of Neumann b.c. on Γ^W. We define the restriction operators

$$\begin{aligned} R_{N,0}: &\quad \mathcal{T}_{N,0} \to \mathbb{R}^{N-1}, &\quad \varphi \mapsto (\varphi(\pi/N), \varphi(2\pi/N), ..., \varphi(\pi(N-1)/N))^t. \\ R_{N,Nm}: &\quad \mathcal{T}_{N,Nm} \to K_N, &\quad \varphi \mapsto (\varphi(0), \varphi(\pi/N), ..., \varphi(\pi))^t, \\ &\text{where } K_N := \{(x_0, ..., x_N) \in \mathbb{R}^{N+1} \mid \frac{x_0}{2} + \sum_{k=1}^{N-1} x_k + \frac{x_N}{2} = 0\} \end{aligned}$$

Obviously $R_{N,0}$, $R_{N,Nm}$ are well defined and bijective an have inverse operators (interpolation operators)

$$\begin{aligned} I_{N,0} &: \mathbb{R}^{N-1} \to \mathcal{T}_{N,0}, \\ I_{N,Nm} &: K_N \to \mathcal{T}_{N,Nm}. \end{aligned}$$

Now we define the discrete analogon $(-\Delta_0)_N^{1/2}$ to $(-\Delta_0)^{1/2}$ and $(-\Delta_{Nm})_N^{1/2}$ to $(-\Delta_{Nm})^{1/2}$ by

$$\begin{aligned} (-\Delta_0)_N^{1/2} &:= R_{N,0}\,(-\Delta_0)^{1/2}\,I_{N,0}, \\ (-\Delta_{Nm})_N^{1/2} &:= R_{N,Nm}\,(-\Delta_{Nm})^{1/2}\,I_{N,Nm}. \end{aligned} \tag{3.134}$$

[14] $(-\Delta)^{1/2}$ as in Lemma 3.4

At first we are going to search a tridiagonal matrix which approximates (3.134). Let us focus on the Dirichlet case at first. We will use the approach

$$C_{FD} := \begin{pmatrix} \alpha & \beta & & \\ \beta & \ddots & \ddots & \\ & \ddots & \ddots & \beta \\ & & \beta & \alpha \end{pmatrix} \tag{3.135}$$

for the approximation of $(-\Delta_0)_N^{1/2}$.

It is well known that the $(N-1) \times (N-1)$-matrix (3.135) has the eigenvectors

$$v_k = (v_k^i)_{i=1..N-1} = \left(\sin \frac{ik\pi}{N} \right)_{i=1..N-1}, \quad k = 1, ..., N-1 \tag{3.136}$$

and the eigenvalues

$$\mu_k = \alpha + 2\beta \cos \frac{k\pi}{N}, \quad k = 1, ..., N-1. \tag{3.137}$$

The eigenvectors v_k and eigenvalues μ_k correspond to the eigenfunctions

$$\psi_k = \sin kx \tag{3.138}$$

and the eigenvalues

$$\nu_k = k \tag{3.139}$$

of the *exact* operator $(-\Delta_0)^{1/2}$ defined on the intervall $(0, B) = (0, \pi)$. We conclude that $(-\Delta_0)_N^{1/2}$ has *the same eigenvectors* (3.136) as C_{FD}. That means that we only have to find $\alpha, \beta \in \mathbb{R}$ such that the μ_k are a good approximation of the ν_k in the sense that the condition number $\kappa = \kappa(N)$ of the symmetric positive definite matrix $((-\Delta_0)_N^{1/2})^{-1} \, C_{FD} = C_{FD} \, ((-\Delta_0)_N^{1/2})^{-1}$ is small. Obviously,

$$\kappa = \frac{\max\limits_{k=1,...,N-1} \frac{\mu_k}{\nu_k}}{\min\limits_{k=1,...,N-1} \frac{\mu_k}{\nu_k}}. \tag{3.140}$$

If we would take the most simple choice $C_{FD} := id$ ($\alpha := 1$, $\beta := 0$) we would have $\mu_k = 1$ for all k and therefore $\kappa(N) = N-1$. In order to get a formulation less dependent on discrete values, we introduce $x := \frac{k\pi}{N}$ and substitute problem (3.140) by the (slightly stronger) problem to find $\alpha, \beta \in \mathbb{R}$ for which

$$\bar{\kappa} := \frac{\max\limits_{x \in [\frac{\pi}{N}, \pi]} \frac{g(x)}{x}}{\min\limits_{x \in [\frac{\pi}{N}, \pi]} \frac{g(x)}{x}} \tag{3.141}$$

is small (possibly independent of N), where

$$g(x) := \alpha + 2\,\beta\,\cos x.$$

This problem is visualized in the left part of Fig. 3.29. That diagram suggests to focus on the points $x = \frac{\pi}{N}$ and $x = \pi$ and to postulate

$$\frac{g(\pi/N)}{\pi/N} \approx \frac{g(\pi)}{\pi}$$

to determine $\alpha = \alpha(N)$, $\beta = \beta(N)$. So we choose $\alpha = \frac{1}{2} + \frac{1}{N}$, $\beta = -\frac{1}{4}$ because for this choice of α, β we have

$$\begin{aligned} \frac{g(\pi)}{\pi} &= \frac{1 + 1/N}{\pi}, \\ \frac{g(\pi/N)}{\pi/N} &= \pi^{-1}\,(1 + \frac{N}{2}\,(1 - \cos\frac{\pi}{N})) = \pi^{-1}\,(1 + O(N^{-1})). \end{aligned}$$

The resulting g is displayed in the left part of Fig. 3.29 (upper dotted line). The approximation quality of C_{FD} for this choice of α, β is investigated in the following lemma.

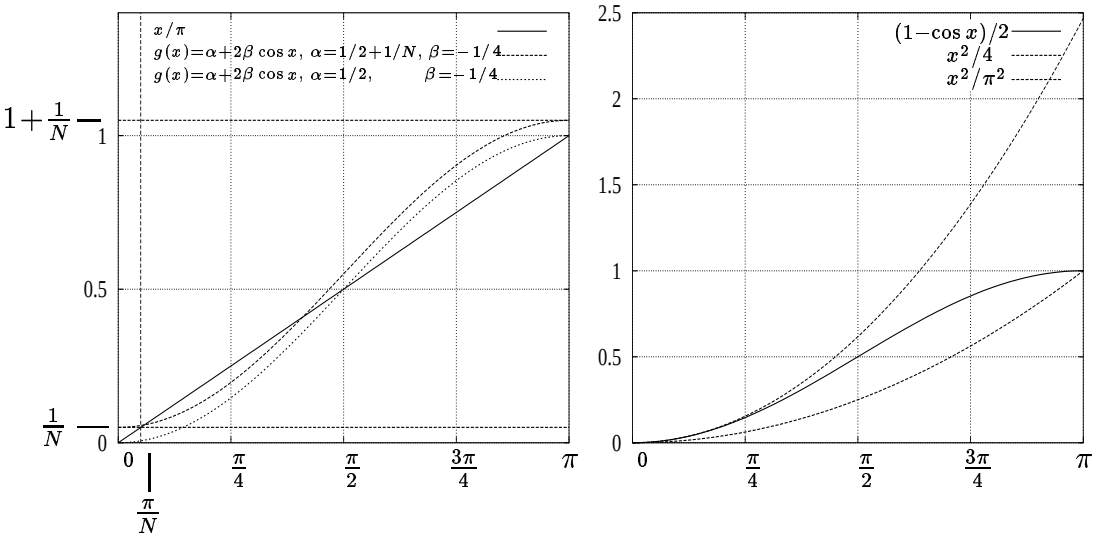


Figure 3.29: Left part: Visualization of function g. The term $1/N$ in the definition of α is necessary because otherwise the ratio $g(x)/x$ would behave like $O(N^{-1})$ for $x = \pi/N$ (see lower dotted curve). – Right part: Visualization of estimate (3.143).

Lemma 3.15 *Let $\alpha = \frac{1}{2} + \frac{1}{N}$, $\beta = -\frac{1}{4}$. Then the symmetric positive definite $(N-1) \times (N-1)$-matrix $C_{FD}\,((-\Delta_0)_N^{1/2})^{-1}$ has the condition number*

$$\kappa = O(N^{1/2}). \tag{3.142}$$

Proof. Due to the explanations above it is sufficient to show that $\bar{\kappa} = O(N^{1/2})$. For this, we have to find the extreme values of $\frac{g(x)}{x}$ on $[\frac{\pi}{N}, \pi]$. Using the estimate

$$\frac{x^2}{\pi^2} \leq \frac{1-\cos x}{2} \leq \frac{x^2}{4} \quad \text{for all } 0 \leq x \leq \pi \tag{3.143}$$

(see right part of Fig. 3.29) on $\frac{g(x)}{x} = \frac{1}{Nx} + \frac{1}{2x}(1-\cos x)$ we get

$$\frac{x}{\pi^2} + \frac{1}{Nx} \leq \frac{g(x)}{x} \leq \frac{x}{4} + \frac{1}{Nx}, \quad \frac{\pi}{N} \leq x \leq \pi. \tag{3.144}$$

The lower bound in (3.144) takes its minimum at $x = \pi\, N^{-1/2}$, the upper bound in (3.144) takes its maximum at $x = \pi$. So (3.144) leads to

$$\frac{2}{\pi N^{1/2}} \leq \frac{g(x)}{x} \leq \frac{\pi}{4} + \frac{1}{\pi N}$$

and therefore

$$\kappa \leq \bar{\kappa} \leq \frac{\pi^2}{8}\, N^{1/2} + \frac{1}{2}\, N^{-1/2}. \tag{3.145}$$

■

N	κ without precond.	κ est. by (3.145)	κ acc. to (3.140)
$16 = 2^4$	15	5.060	2.332
$256 = 2^8$	255	19.770	11.402
$4048 = 2^{12}$	4047	78.965	45.648

Table 3.1: The condition number $\kappa = O(N^{1/2})$. The comparison between the unconditioned value, the estimated value (3.145) and the true value according to (3.140).

The Neumann case. Obviously, Lemma 3.15 and its derivation stays valid in the Neumann case. We just have to replace β in the first and in the last line of matrix C_{FD} by 2β. This is necessary because C_{FD} has the same eigenvectors

$$v_k = (v_k^i)_{i=0..N} = \left(\cos\frac{ik\pi}{N}\right)_{i=0..N}, \quad k = 1, ..., N \tag{3.146}$$

as $(-\Delta_{Nm})_N^{1/2}$, then.

3.4.2 Multidiagonal matrices

Sec. 3.4.1 showed that a tridiagonal matrix preconditioner can reduce the condition number from $O(N)$ to $O(N^{1/2})$.[15] Now we will investigate how the condition number can be reduced if matrices with larger bandwidth are used. We construct the matrix C_{FD}^M as a symmetric $(2M+1)$-bandwidth matrix with the structure

$$\ldots \quad 0 \quad 0 \quad 0 \quad \alpha_M \quad \ldots \quad \alpha_2 \quad \alpha_1 \quad \alpha_0 \quad \alpha_1 \quad \alpha_2 \quad \ldots \quad \alpha_M \quad 0 \quad 0 \quad 0 \quad \ldots \tag{3.147}$$

Let us focus on the Dirichlet case. As we want C_{FD}^M to have the eigenvectors (3.136), we have to modify the first and the last $M-1$ lines of the matrix (3.147) by taking into account the symmetry of these vectors (after odd extension). As an example, (3.148) displays C_{FD}^M in the case $M=4$:

$$\begin{pmatrix}
\alpha_0-\alpha_2 & \alpha_1-\alpha_3 & \alpha_2-\alpha_4 & \alpha_3 & \alpha_4 & & \\
\alpha_1-\alpha_3 & \alpha_0-\alpha_4 & \alpha_1 & \alpha_2 & \ddots & \ddots & \\
\alpha_2-\alpha_4 & \alpha_1 & \alpha_0 & \ddots & \ddots & \ddots & \alpha_4 \\
\alpha_3 & \alpha_2 & \ddots & \ddots & \ddots & \alpha_2 & \alpha_3 \\
\alpha_4 & \ddots & \ddots & \ddots & \alpha_0 & \alpha_1 & \alpha_2-\alpha_4 \\
 & \ddots & \ddots & \alpha_2 & \alpha_1 & \alpha_0-\alpha_4 & \alpha_1-\alpha_3 \\
 & & \alpha_4 & \alpha_3 & \alpha_2-\alpha_4 & \alpha_1-\alpha_3 & \alpha_0-\alpha_2
\end{pmatrix} \tag{3.148}$$

The eigenvalues of this matrix are obviously

$$\mu_k = \alpha_0 + 2\sum_{j=1}^{M} \alpha_j \cos\frac{kj\pi}{N}, \quad k = 1, \ldots, N-1. \tag{3.149}$$

So we have to find α_j such that

$$g(x) = \alpha_0 + 2\sum_{j=1}^{M} \alpha_j \cos jx \tag{3.150}$$

is a good approximation of the function $f(x) = x$ on $[\frac{\pi}{N}, \pi]$ in the sense that the expression (3.141) is small. An idea which suggests itself is to compare (3.150) with the truncated Fourier series f_M of the 2π-periodic extended function

$$f(x) = \begin{cases} x, & 0 \le x < \pi \\ 2\pi - x, & \pi \le x < 2\pi \end{cases}$$

To improve the approximation property in the sense of (3.141) we add (similar to the case of the tridiagonal matrix) a constant such that $g(0) = \frac{\pi}{N}$, so

$$g(x) = f_M(x) - f_M(0) + \frac{\pi}{N}. \tag{3.151}$$

[15] equidistant boundary mesh supposed

See Fig 3.30 for f and g. f has the Fourier coefficients

$$\alpha'_k := \frac{1}{\pi}\int_0^{2\pi} f(x)\cos kx\,dx = \begin{cases} \pi, & k=0 \\ 0, & k=2,4,6,\ldots \\ -\frac{4}{\pi k^2}, & k=1,3,5,\ldots \end{cases}$$

so the truncated Fourier series f_M of f is

$$f_M(x) := \frac{\alpha'_0}{2} + \sum_{k=1}^{M} \alpha'_k \cos kx = \frac{\pi}{2} - \sum_{\substack{k=1 \\ k \text{ odd}}}^{M} \frac{4}{\pi k^2} + \sum_{\substack{k=1 \\ k \text{ odd}}}^{M} \frac{4}{\pi k^2}(1-\cos kx). \quad (3.152)$$

From the representation (3.152) we conclude

$$f_M(0) = \frac{4}{\pi}\left(\frac{\pi^2}{8} - \sum_{\substack{k=1 \\ k \text{ odd}}}^{M} \frac{1}{k^2}\right) \quad (3.153)$$

and therefore

$$g(x) = \frac{\pi}{N} + \sum_{\substack{k=1 \\ k \text{ odd}}}^{M} \frac{4}{\pi k^2}(1-\cos kx). \quad (3.154)$$

A comparison of (3.150) and (3.154) gives the matrix entries

$$\alpha_k = \begin{cases} \frac{\pi}{N} + \sum_{\substack{j=1 \\ j \text{ odd}}}^{M} \frac{4}{\pi j^2}, & k=0 \\ 0, & k=2,4,6,\ldots \\ -\frac{2}{\pi k^2}, & k=1,3,5,\ldots \end{cases}$$

for C_{FD}^M.

We now take

$$M := \lceil N^\alpha \rceil \quad \text{with } 0<\alpha<1. \quad (3.155)$$

The costs of the application of the preconditioner C_{FD}^M behaves like $NM \approx N^{1+\alpha}$. The following lemma shows the dependence of the condition number on α and N:

Lemma 3.16 *Let the matrix C_{FD}^M with M from (3.155) be constructed as above discribed. Then the symmetric positive definite $(N-1)\times(N-1)$-matrix $C_{FD}^M\,((-\Delta_0)_N^{1/2})^{-1}$ has the condition number*

$$\kappa = O(N^{(1-\alpha)/2}). \quad (3.156)$$

Proof. We have to proof that expression (3.141) is bounded by $c\,N^{(1-\alpha)/2}$.
(i). Consider $\frac{\pi}{M} \le x \le \pi$ at first. From the approach (3.151) we get

$$\begin{aligned} |g(x) - f(x)| &= |f_M(x) - f(x) - f_M(0) + \frac{\pi}{N}| \\ &\le \frac{\pi}{N} + |f_M(0)| + \sum_{\substack{k=M+1 \\ k \text{ odd}}}^{\infty} |\alpha'_k|. \end{aligned} \tag{3.157}$$

As $\sum\limits_{\substack{k=1 \\ k \text{ odd}}}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{8}$, (3.153) yields

$$f_M(0) = \frac{4}{\pi} \sum_{\substack{k=M+1 \\ k \text{ odd}}}^{\infty} \frac{1}{k^2}.$$

So we get

$$f_M(0) + \sum_{\substack{k=M+1 \\ k \text{ odd}}}^{\infty} |\alpha'_k| = \sum_{k=\lceil \frac{M}{2} \rceil}^{\infty} \frac{8}{\pi\,(2k+1)^2} \le \frac{8}{\pi} \int_{\frac{M-1}{2}}^{\infty} \frac{dx}{(2x+1)^2} = \frac{4}{M\pi}$$

This in (3.157) yields

$$\frac{g(x)}{x} = 1 + \frac{g(x) - x}{x} \le 1 + \frac{\frac{\pi}{N} + \frac{4}{M\pi}}{\frac{\pi}{M}} = 1 + \frac{4}{\pi^2} + \frac{M}{N} \longrightarrow 1 + \frac{4}{\pi^2}$$

and analogously

$$\frac{g(x)}{x} \ge 1 - \frac{|g(x) - x|}{x} \ge 1 - \frac{\frac{\pi}{N} + \frac{4}{M\pi}}{\frac{\pi}{M}} = 1 - \frac{4}{\pi^2} - \frac{M}{N} \longrightarrow 1 - \frac{4}{\pi^2}.$$

(ii). Now let $\frac{\pi}{N} \le x \le \frac{\pi}{M}$. As $kx \le \pi$ for all $k \le M$, we can use (3.143) on (3.154) and get

$$\frac{\pi}{Nx} + \frac{4Mx}{\pi^3} \le \frac{g(x)}{x} \le \frac{\pi}{Nx} + \frac{(M+1)\,x}{\pi}. \tag{3.158}$$

For $x \in \mathbb{R}^+$, the left part of (3.158) takes its minimum at $x = \frac{\pi^2}{2\sqrt{MN}}$. The right part of (3.158) takes its maximum for $x \in [\pi/N, \pi/M]$ at $x = \frac{\pi}{M}$. So we get

$$\frac{4}{\pi}\sqrt{\frac{M}{N}} \le \frac{g(x)}{x} \le 1 + \frac{M}{N} + \frac{1}{M}$$

on $[\frac{\pi}{N}, \frac{\pi}{M}]$.
(iii). (i) and (ii) together yield

$$\frac{4}{\pi}\sqrt{\frac{M}{N}} \le \frac{g(x)}{x} \le 1 + \frac{4}{\pi^2} + \frac{M}{N}$$

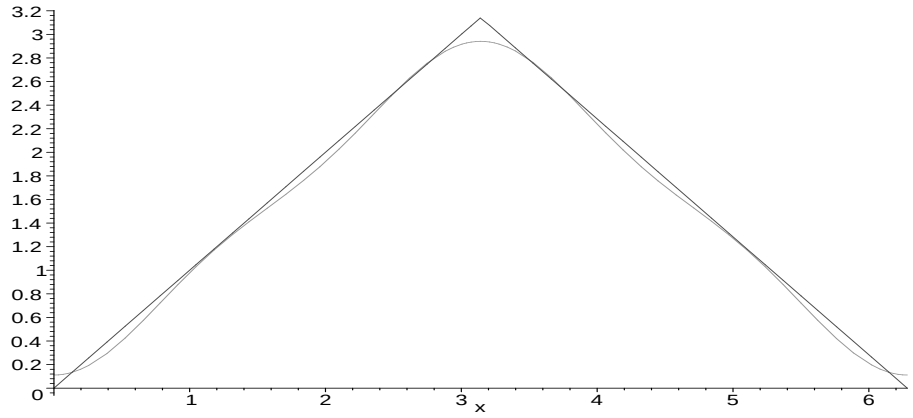


Figure 3.30: Visualisation of f and g with $M = 3$, $N = 9$ on the intervall $[0, 2\pi]$. As expected, the approximation quality measured by $g(x)/f(x)$ becomes worst for small values of x because g is a linear combination of cosine functions and behaves like a second order polynomial for small x. See Fig. 3.31 for the ratio $g(x)/f(x) = g(x)/x$ on $[0, \pi]$.

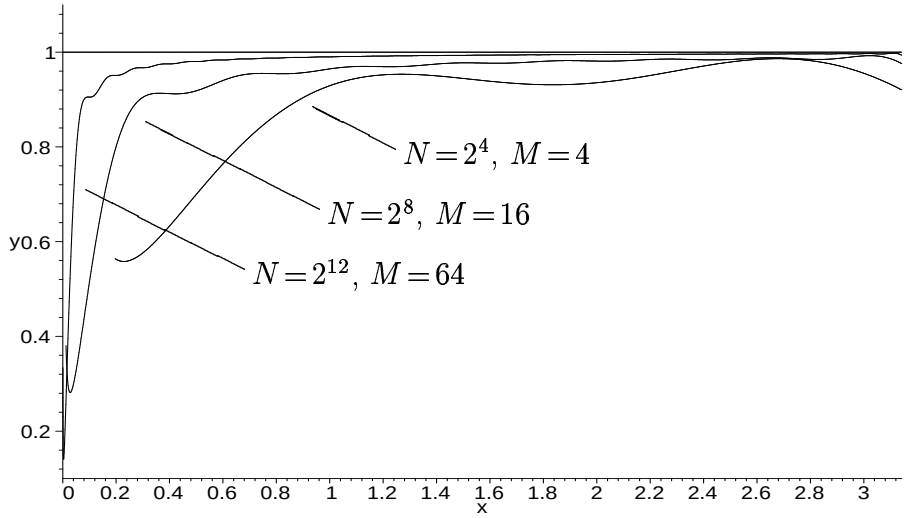


Figure 3.31: Visualisation of $g(x)/x$ on the intervall $[\pi/N, \pi]$ with $M = N^{1/2}$ (i.e. $\alpha = 0.5$). The three lines represent the cases $N = 2^4, 2^8, 2^{12}$. As expected, the approximation is worst (i.e. $g(x)/x \ll 1$) for small values of x, see Fig. 3.30.

for all $x \in [\frac{\pi}{N}, \pi]$ and N, M sufficiently large, and therefore (3.141) can be estimated by

$$\begin{aligned} \frac{\pi^2+4}{4\pi}\left(\frac{N}{M}\right)^{1/2} + \frac{\pi}{4}\left(\frac{N}{M}\right)^{-1/2} &= \frac{\pi^2+4}{4\pi} N^{(1-\alpha)/2} + \frac{\pi}{4} N^{-(1-\alpha)/2} \\ &\approx 1.104\, N^{(1-\alpha)/2}. \end{aligned} \tag{3.159}$$

■

N	$M = N^{1/2}$	κ est. by (3.159)	κ acc. to (3.141)	κ acc. to (3.140)
$16 = 2^4$	4	2.207	1.75	1.745
$256 = 2^8$	16	4.415	3.52	3.500
$4048 = 2^{12}$	64	8.830	7.08	7.045

Table 3.2: The condition number κ in the case $\alpha = 0.5$. We get $\kappa = O(N^{1/4})$. The comparison between the estimation (3.159) and the true value according to (3.140) shows that the estimation is very sharp. Compare to Tab. 3.1.

The Neumann problem. In the case of Neumann boundary conditions the matrix C_{FM}^M should have the eigenvectors (3.146); taking into account their symmetry (after even extension) we get

$$C_{FD}^M = \begin{pmatrix} \alpha_0 & 2\,\alpha_1 & 2\,\alpha_2 & 2\,\alpha_3 & 2\,\alpha_4 & & & \\ \alpha_1 & \alpha_0+\alpha_2 & \alpha_1+\alpha_3 & \alpha_2+\alpha_4 & \alpha_3 & \alpha_4 & & \\ \alpha_2 & \alpha_1+\alpha_3 & \alpha_0+\alpha_4 & \alpha_1 & \ddots & \ddots & \ddots & \\ \alpha_3 & \alpha_2+\alpha_4 & \alpha_1 & \ddots & \ddots & \ddots & \alpha_3 & \alpha_4 \\ \alpha_4 & \alpha_3 & \ddots & \ddots & \ddots & \alpha_1 & \alpha_2+\alpha_4 & \alpha_3 \\ & \ddots & \ddots & \ddots & \alpha_1 & \alpha_0+\alpha_4 & \alpha_1+\alpha_3 & \alpha_2 \\ & & \alpha_4 & \alpha_3 & \alpha_2+\alpha_4 & \alpha_1+\alpha_3 & \alpha_0+\alpha_2 & \alpha_1 \\ & & & 2\,\alpha_4 & 2\,\alpha_3 & 2\,\alpha_2 & 2\,\alpha_1 & \alpha_0 \end{pmatrix} \tag{3.160}$$

The $(N+1) \times (N+1)$-matrix (3.160) is symmetric in the scalar product

$$(x, y) := \frac{x_0\, y_0}{2} + \sum_{i=1}^{N-1} x_i\, y_i + \frac{x_N\, y_N}{2} \tag{3.161}$$

and has the same eigenvalues (3.149) as in the Dirichlet case, but now for $k = 0, ..., N$ instead of $k = 1, ..., N-1$. For the definition of the condition number (3.140) the index $k = 0$ does not have to be taken into account because the related eigenvector is constant and not in K_N; the related continuous function φ is not in $H_{mv}^{1/2}(0, \pi)$. Thus, in (3.140) k runs from 1 to N.

We can use the same α, β rsp. α_j and, obviously, get the same estimate (3.156) for the condition number as in the Dirichlet case.

3.4.3 Condition number independent of N

In the last section we found matrix type preconditioners consuming $O(N^{\alpha+1})$ operations producing a condition number $\kappa = O(N^{(1-\alpha)/2})$, $0<\alpha<1$. We will show that it is possible to construct a matrix type preconditioner with less operations and a condition number independent of N! We just have to drop the requirement that the non-zero bands of the preconditioner matrix are side by side.

Lemma 3.17 *Let C^*_{FD} be the matrix of type (3.148) ($M := N$) in the case of Dirichlet boundary conditions rsp. (3.160) in the case of Neumann boundary conditions with the entries*

$$\alpha_j = \begin{cases} \pi + c_N, & j = 0 \\ -\frac{\pi}{2}, & j = 1 \\ -\frac{1}{2j}, & j = 2, 4, 8, 16, ..., \lfloor \log_2 N \rfloor \\ 0, & else, \end{cases}$$

$$c_N := 1 - \left(\frac{1}{2}\right)^{\lfloor \log_2 N \rfloor}.$$

Then, the symmetric positive definite matrix $C^*_{FD}\,((-\Delta_0)^{1/2}_N)^{-1}$ has a condition number

$$\kappa \leq \frac{\pi^3}{4} + 2 \approx 9.75 \tag{3.162}$$

independent of N.

Proof. Using (3.149), the matrix C^*_{FD} from Lemma 3.17 has the eigenvalues

$$\mu_k = \pi \left(1 - \cos\frac{k\pi}{N}\right) + \sum_{j=1}^{\lfloor \log_2 N \rfloor} \frac{1 - \cos\frac{2^j k\pi}{N}}{2^j} \tag{3.163}$$

So, by replacing $x := \frac{\pi k}{N}$, it is sufficient to show that $\frac{g(x)}{x}$ with

$$g(x) = \pi\,(1 - \cos x) + \sum_{j=1}^{\lfloor \log_2 N \rfloor} \frac{1 - \cos 2^j x}{2^j}, \quad x \in [\frac{\pi}{N}, \pi],$$

is bounded by constants c_1, c_2 from above and below with $\frac{c_2}{c_1} = \frac{\pi^3}{4} + 2$. As $\frac{\pi}{x} \leq N$, we can write

$$g(x) = \pi\,(1 - \cos x) + \sum_{j=1}^{\lfloor \log_2 \frac{\pi}{x} \rfloor} \frac{1 - \cos 2^j x}{2^j} + \sum_{j=\lfloor \log_2 \frac{\pi}{x} \rfloor + 1}^{\lfloor \log_2 N \rfloor} \frac{1 - \cos 2^j x}{2^j}$$

and use (3.143) on the first term and on the first sum. We get

$$\frac{2x^2}{\pi} + \sum_{j=1}^{\lfloor \log_2 \frac{\pi}{x} \rfloor} \frac{2^{j+1}x^2}{\pi^2} \leq g(x) \leq \frac{\pi x^2}{2} + \sum_{j=1}^{\lfloor \log_2 \frac{\pi}{x} \rfloor} 2^{j-1}x^2 + \sum_{j=\lfloor \log_2 \frac{\pi}{x} \rfloor + 1}^{\infty} 2^{1-j},$$

thus

$$\frac{2x}{\pi} + \frac{4x}{\pi^2}\left(2^{\lfloor \log_2 \frac{\pi}{x} \rfloor} - 1\right) \leq \frac{g(x)}{x} \leq \frac{x\pi}{2} + x\left(2^{\lfloor \log_2 \frac{\pi}{x} \rfloor} - 1\right) + \frac{1}{x}\, 2^{1-\lfloor \log_2 \frac{\pi}{x} \rfloor}.$$

Using $\xi - 1 \leq \lfloor \xi \rfloor \leq \xi$ we get

$$\frac{2x}{\pi} + \frac{4x}{\pi^2}\left(\frac{\pi}{2x} - 1\right) \leq \frac{g(x)}{x} \leq \frac{x\pi}{2} + x\left(\frac{\pi}{x} - 1\right) + \frac{4}{\pi}.$$

So,

$$\frac{2}{\pi} \leq \frac{g(x)}{x} \leq \frac{\pi^2}{2} + \frac{4}{\pi}$$

for $x \in [\frac{\pi}{N}, \pi]$.

This yields (3.162). ∎

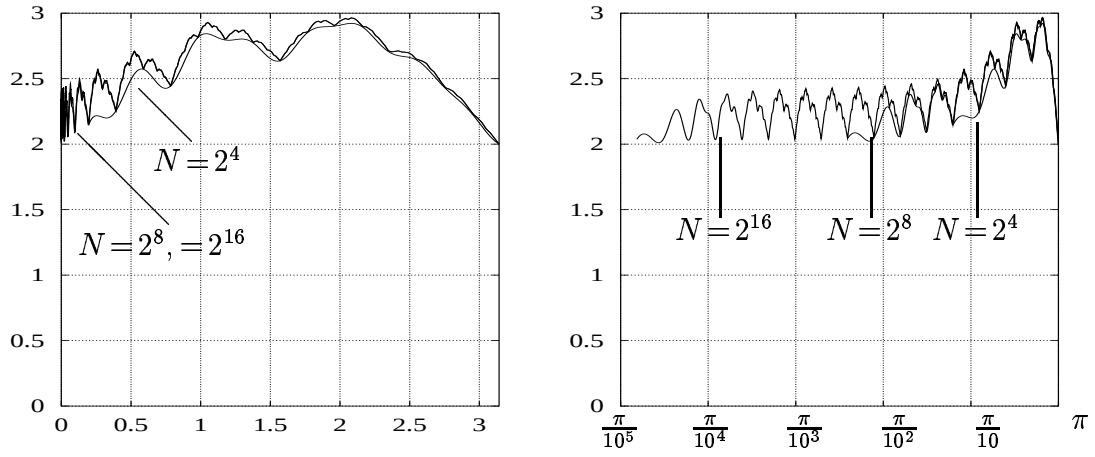


Figure 3.32: Visualization of $\frac{g(x)}{x}$ on the intervall $[\frac{\pi}{N}, \pi]$. The three lines represent the cases $N = 2^4, 2^8, 2^{16}$. For the left figure a linear scale and for the right figure a logarithmic scale is used on the x-axis. Obviously, $2 \leq \frac{g(x)}{x} \leq 3$. That means $\kappa \leq 1.5$, which is even better than (3.162). The value of κ is evaluated more exactly in Table 3.3. The advantage of C^*_{FD} over C_{FD} is obvious by comparing Fig. 3.32 to Fig. 3.31. Furthermore, C^*_{FD} is even less costly than C_{FD}.

N	κ
2^5	1.47559
2^8	1.48126
2^{11}	1.48298
2^{20}	1.48307

Table 3.3: The 'true' condition number κ of the matrix $C_{FD}^*((-\Delta)_N^{1/2})^{-1}$ calculated by evaluating (3.140) with μ_k from (3.163).

Application of the results. In Section 3.4 approximations C_{FD}, C_{FD}^* of $(-\Delta)^{1/2}$ were derived. In the equidistant grid case, these matrices can be used to construct preconditions by replacing $(-\Delta)^{1/2}$ by C_{FD}, C_{FD}^* in Section 3.1. We will get the condition numbers of Tab. 3.1 (right column) and Tab. 3.3.

But also in the Gauss-Lobatto mesh case these matrices apply: Using Theorem 3.9, we can use C_{FD}, C_{FD}^* to replace C_{GL}. In this case, we should expect slightly larger condition numbers $\kappa\frac{c_2}{c_1}$, where κ is from Tab. 3.1,3.3 and the c_i are the (not explicitely known) equivalence constants of Theorem 3.9.

Test runs. In the following we are comparing the effectivity of the multidiagonal preconditioner C_{FD}^*, the tridiagonal preconditioner C_{FD} and the spectral preconditioner C from Sec. 3.1.

Fig. 3.33 uses 4 FD subdomains and an equidistant boundary mesh. As expected, the multidiagonal and the spectral preconditioner show a better performance than the tridiagonal. Comparing the upper and the lower diagram, the effectivity of the multidiagonal and the spectral preconditioner does not depend on the discretization parameter N. The tridiagonal preconditioner looses efficiency for large N. This complies with the theoretical results of Lemmas 3.15, 3.17.

In Fig. 3.34 four *spectral* subdomains with Chebyshev-Gauss-Lobatto boundary mesh are used. As explained, C_{FD}, C_{FD}^* are used to replace the spectral realization of $(-\Delta_0)^{1/2}$ in (3.35).

According to theory and to the numerical tests, the multidiagonal matrix preconditioner and the spectral preconditioner generate (in the Dirichlet case) a condition number independent from N. Comparing both, the spectral preconditioner needs 0-2 CGBI iteration steps less and the multidiagonal. This can be explained by the approximation error of the approximation $C_{FD}^* \approx (-\Delta_0)_N^{1/2}$ displayed in Table 3.3.

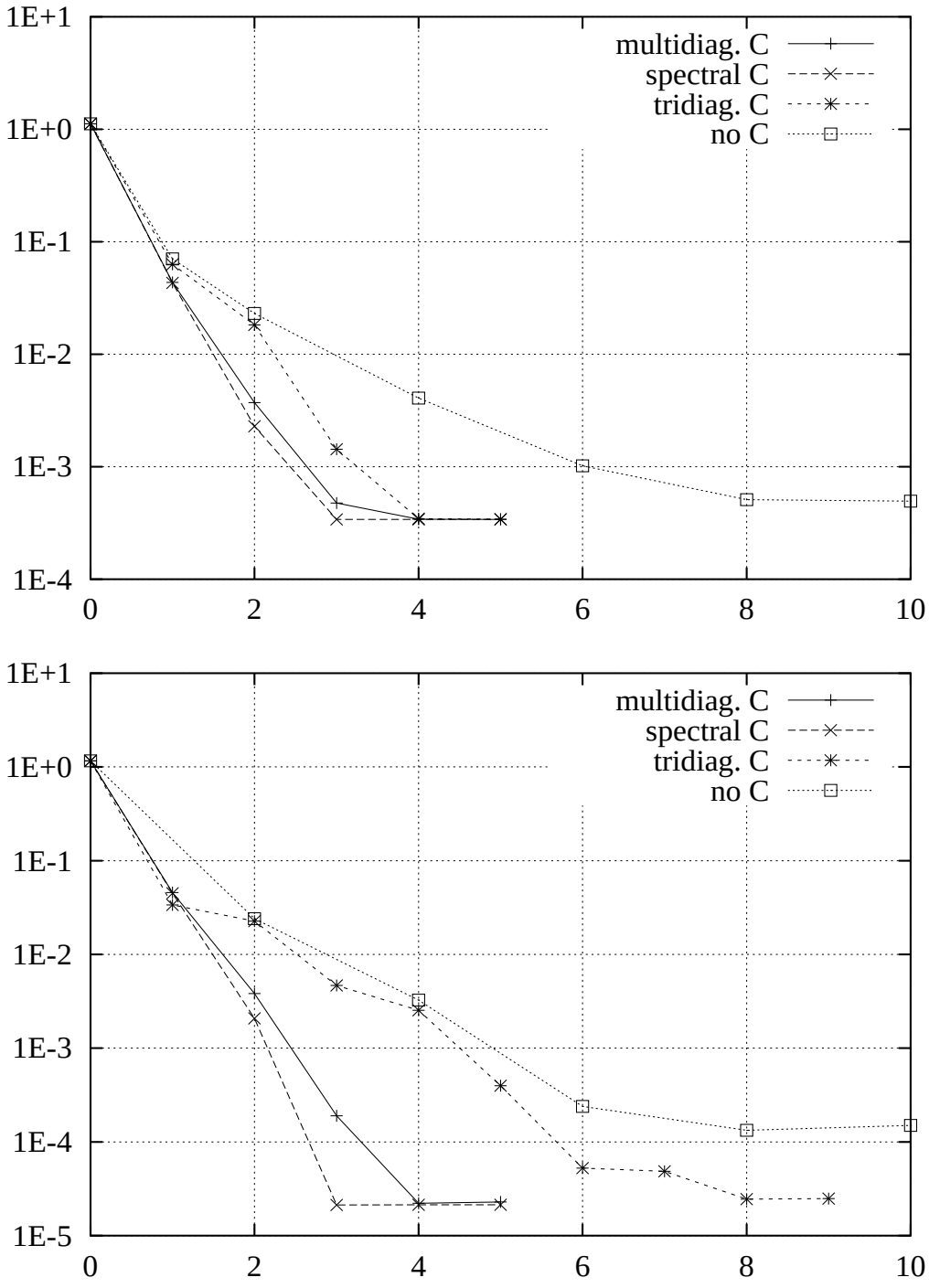


Figure 3.33: The Dirichlet problem on 4 FD subdomains for the exact solution (2.83). Comparison of the multidiagonal preconditioner C^*_{FD}, the tridiagonal preconditioner C_{FD} and the spectral preconditioner from Sec. 3.1. $\sigma = 0$, $r = 1$. – Upper fig.: $N = 64$. – Lower fig.: $N = 256$.

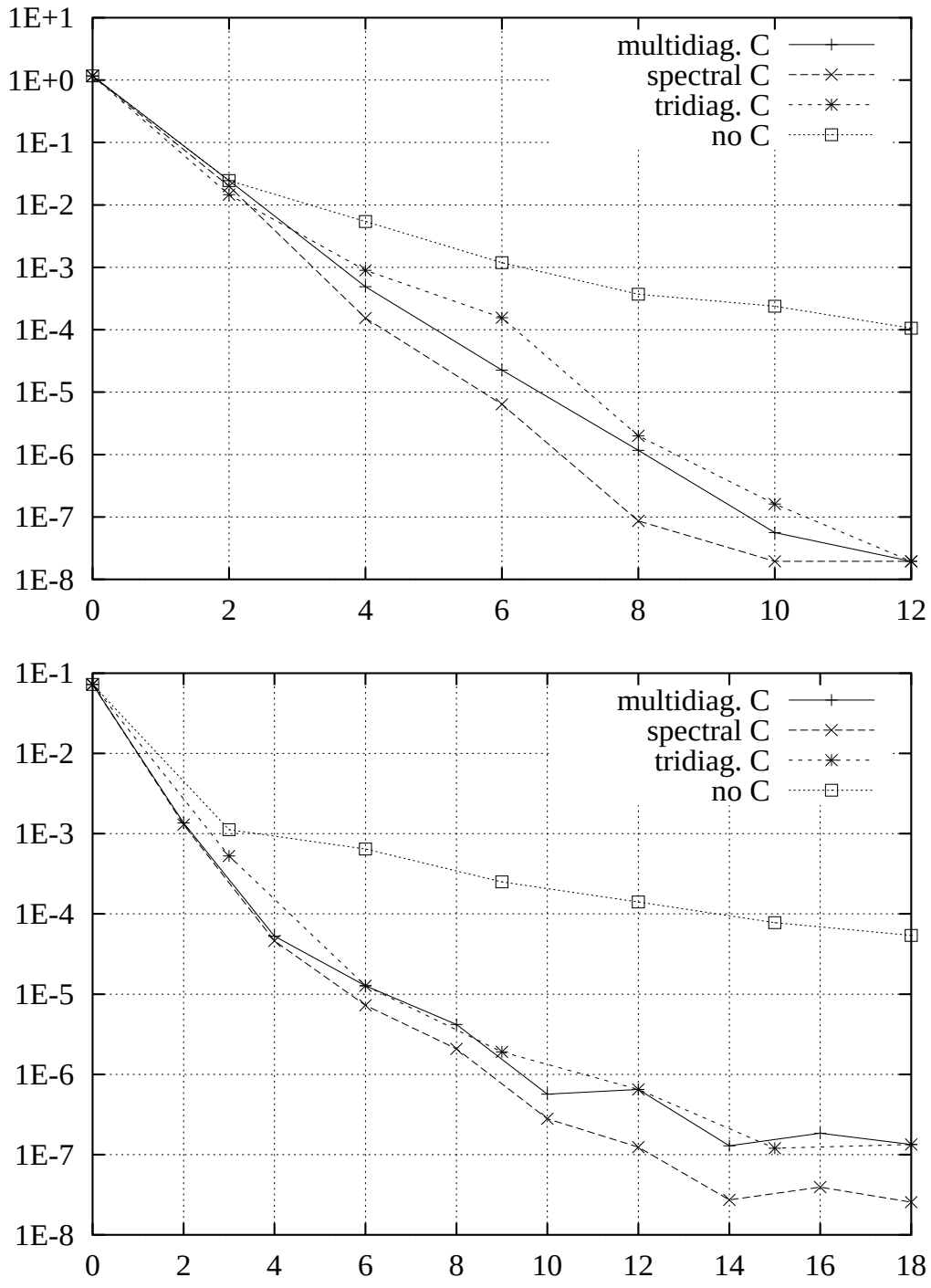


Figure 3.34: A comparison of the multidiagonal preconditioner C^*_{FD}, the tridiagonal preconditioner C_{FD} and the spectral preconditioner on 4 *Chebyshev* subdomains. $\sigma = 0$, $r = 1$. – Upper fig.: Dirichlet problem for solution (2.83), $N = 64$. – Lower fig.: Neumann problem for solution (2.81), $N = 256$.

3.5 Irregular meshes

Up to now, preconditioners for equidistant boundary meshes and for Chebyshev-Gauss-Lobatto boundary meshes were developed in this Chapter. In the context of FE solvers, different boundary meshes may occur. In this section we handle the case of *quasi-uniform* boundary meshes, which are defined as follows: Let us consider a sequence of meshes $(\mathcal{M})$ on Γ. For each mesh, let $0 = x_0 < x_1 < \ldots < x_N = B$, $N = N(\mathcal{M}, i)$, $x_j = x_j(\mathcal{M}, i)$ be the boundary mesh on Γ_i and $h_j = h_j(\mathcal{M}) := x_{j+1} - x_j$ the local mesh size. We postulate that there is a constant $c > 0$ such that

$$h_{min}(\mathcal{M}) \geq c\, h_{max}(\mathcal{M}) \quad \forall\, \mathcal{M}, \tag{3.164}$$

where

$$h_{min}(\mathcal{M}) := \inf_{i,j} h_j(\mathcal{M}, i), \quad h_{max}(\mathcal{M}) := \sup_{i,j} h_j(\mathcal{M}, i),$$

i.e. the ratio of the largest mesh size and the smallest mesh size is bounded independent of $\mathcal{M}$.

We will show that under the assumption (3.164), any discretization (the spectral approach of Sec. 3.1, the matrix approach of Sec. 3.4.3) of the preconditioner C_{glob} (3.59) can be used on the non-equidistant data set $(\varphi_i(x_j))_{i=1,\ldots,p-1, j=0,\ldots,N}$.

Let us prove the following lemma at first:

Lemma 3.18 *Let us identify $\Gamma_i = (0, B)$ for all $i = 1, \ldots, p-1$. For any function $w : [0, B] \to [0, B]$ with*

$$l \leq \frac{w(x) - w(y)}{x - y} \leq L \quad \forall\, x \neq y \in [0, B],$$

and $w(0) = 0$, $w(B) = B$, the norms generated by $C_{glob}(\cdot)$ and by $C_{glob}(\cdot \circ w)$ (see (3.59)) are equivalent, i.e. there are c_1, $c_2 > 0$ depending on l, L, but independent of φ such that

$$c_1\, \langle C_{glob}(\varphi \circ w), \varphi \circ w \rangle_\Gamma \leq \langle C_{glob}\varphi, \varphi \rangle_\Gamma \leq c_2\, \langle C_{glob}(\varphi \circ w), \varphi \circ w \rangle_\Gamma \quad \forall\, \varphi \in R(A).$$

Proof. **(i)** Let us consider the case of Neumann conditions on Γ^W. $C_{glob} = (-\Delta_{Nm})^{1/2} + id$ is a norm equivalent to the $H^{1/2}_{mv}(\Gamma)$-norm. It is well known that the $H_{mv}(\Gamma)$-norm is equivalent to the norm

$$\left(|||\varphi|||^2 + c \int_0^B \varphi^2\, dx \right)^{1/2}, \quad c > 0 \tag{3.165}$$

$$|||\varphi||| := \left(\int_0^B \int_0^B \left(\frac{\varphi(x) - \varphi(y)}{x - y} \right)^2 dx\, dy \right)^{1/2}$$

(see Sec. 3.1.3.1). (3.165) can be rewritten as

$$\left(\int_0^B \int_0^B \left(\frac{\varphi(w(x)) - \varphi(w(y))}{w(x) - w(y)} \right)^2 w'(x)w'(y)\,dx\,dy \;+\; c \int_0^B (\varphi \circ w)^2\, w'\,dx \right)^{1/2} \tag{3.166}$$

$$\leq \left(\frac{L^2}{l^2} \int_0^B \int_0^B \left(\frac{\varphi \circ w(x) - \varphi \circ w(y)}{x - y} \right)^2 dx\,dy \;+\; cL \int_0^B (\varphi \circ w)^2\,dx \right)^{1/2}$$

$$\leq c \max\{L/l, \sqrt{L}\}\, \|\varphi \circ w\|_{H^{1/2}_{mv}(0,B)}.$$

Similarly we get the lower bound $c \min\{l/L, \sqrt{l}\}\, \|\varphi \circ w\|_{H^{1/2}_{mv}(0,B)}$ for (3.166).

(ii). In the Dirichlet case, C_{glob} is equivalent to the $H^{1/2}_{00}(\Gamma)$-norm, which is, by using (2.12)-(2.13), equivalent to $(|||\varphi|||^2 + \|\varphi/\sqrt{x(B-x)}\|^2_{L^2(0,B)})^{1/2}$. This can be estimated similarly to the Neumann case (i); the constants

$$\max\{L/l, \sqrt{L}\}, \quad \min\{l/L, \sqrt{l}\}$$

in (i) are replaced by

$$\max\{L/l, \sqrt{L}/l\}, \quad \min\{l/L, \sqrt{l}/L\},$$

respectively. ∎

Let us return to the sequence of quasi-uniform meshes $(\mathcal{M})$. To apply the lemma, let us introduce the mesh distribution function $w = w(\mathcal{M}, i) : [0, B] \to [0, B]$ with $w(jB/N) = x_j$, which is piecewise linear on the intervals $[jB/N, (j+1)B/N]$.

Obviously, w meets the assumptions of Lemma 3.18 with

$$l = h_{min} N/B, \quad L = h_{max} N/B.$$

By appling the lemma we conclude that $C_{global}(\cdot)$ and $C_{global}(\cdot \circ w)$ are spectrally equivalent operators with equivalence constants bounded independently of $\mathcal{M}$. So we have to proceed similarly as in the Gauss-Lobatto case of Sec. 3.1.3: Instead of applying the preconditioner C_{global} on φ we apply C_{global} on $\varphi \circ w$.

Let us mention that the equivalence constants in the proof of Lemma 3.18 can be estimated as follows: As $l \leq 1$, $L \geq 1$, $L/l \geq 1$,

$$\begin{aligned}
\max\{L/l, \sqrt{L}\} &\leq \max\{L/l, \sqrt{L/l}\} = L/l = h_{max}/h_{min}, \leq c^{-1}, \\
\min\{l/L, \sqrt{l}\} &\geq \min\{l/L, \sqrt{l/L}\} = l/L = h_{min}/h_{max}, \geq c, \\
\max\{L/l, \sqrt{L}/l\} &= L/l \leq c^{-1}, \\
\min\{l/L, \sqrt{l}/L\} &= l/L \geq c,
\end{aligned}$$

with c from (3.164) which is independent of $\mathcal{M}$.

Chapter 4

The Characteristics Method

4.1 Introduction

In the past decades a variety of schemes for solving transport dominated problems were invented. Especially the Lagrange-Galerkin methods associated with the names Morton and Süli (e.g. [13] [38] [39] [48] [49], but also [14] [44]) are of great importance. They are combining the characteristics (Lagrange-) method with the finite element method. Another approach combining a *spectral* method with the method of characteristics is given in [50].

Error estimates for the Lagrange-Galerkin method are e.g. available [13] [39] for the transport equation

$$u_t + a\, u_x = 0 \tag{4.1}$$

with $a = a(t, x)$ or $a = a(u)$ and in [44] [48] for the Navier-Stokes equations.

As pointed out in the previous chapters, we want to involve a spectral collocation method for the solution of the diffusion equation (5.3). To be consistent with this

- we have to use a *high* order method of characteristics
- we prefer a *collocation* approach instead of a Galerkin method.

For our Navier-Stokes solver, the elliptic solvers determine the mesh which should also be used for the hyperbolic step. Therefore one main aspect of our study is to extend the theoretical and numerical behaviour of the characteristics solver to non-equidistant and non-quasi-uniform (Chebyshev-Gauss-Lobatto) meshes.

For the Lagrange-Galerkin method numerical integration may be a source of instability, even in the linear case $a = a(t, x)$ and using linear elements, if standard quadrature formulars are applied [39] [49]. Thus, a detailed stability analysis of *our* approach seems to be indispensable and is carried out in the course of this

chapter. Let us mention that the lack of diffusivity in (4.1) may aggravate the question of stability.

As we are interested in a highly accurate characteristics solver, we use high order interpolation in space and time for the trace-back of the characteristic lines. The spatial interpolation may be regarded as the representation of a function with respect to certain (finite element) piecewise polynomial ansatz functions (see right parts of Figs. 4.1, 4.2). For the sake of simplicity, the theoretical investigations of this chapter are restricted to the case of one space dimension. In Chapter 5, the method is also applied to 2d flow problems.

An error estimate of type

$$\|u - u^h\|_\infty \leq c\,\Delta t^2 + c \min\left\{\frac{h^2}{\Delta t}, h\right\} \tag{4.2}$$

can be proven for the nonlinear (quasi-linear) equation (4.1) with

$$a = a(t, x, u(t, x)) \tag{4.3}$$

and periodic or Dirichlet inflow boundary conditions when linear ansatz functions are used. No restriction on the time step size Δt is necessary, i.e. the method is inconditional stable (see Section 4.3).

In Section 4.4 higher order ansatz functions are used. Let p be the polynomial order in space and q the polynomial order in time. Then we get the error estimate

$$\|u - u^h\|_2 \leq c\,\Delta t^{q+1} + c \min\left\{\frac{h^{p+1}}{\Delta t}, h^p\right\} \tag{4.4}$$

for the nonlinear equation and periodic boundary conditions if a certain stability condition is met. An interesting property of the scheme is that, though the differential equation may be nonlinear, the stability of the spatial interpolation operator (which is linear) implies the stability of the whole scheme. Sections 4.4.2-4.4.6 deal with the question under which restrictions on the Courant number and on the mesh type this stability condition is fulfilled. It turns out that on a quasi-uniform mesh a bounded Courant number is sufficient whereas on a Gauss-Lobatto mesh a very severe stability condition occurs.

Numerical tests (Sec. 4.4.6) confirm that the error estimates (4.2), (4.4) are optimal. The fact that the estimates reveal a lower order than the best-approximation is coherent with the previously mentioned publications on the Lagrange-Galerkin method. Concerning stability, our numerical tests on the Gauss-Lobatto mesh revealed better stability properties than could be expected due to theory.

A different approach to combine a spectral method with the characteristics method is presented in [50]. That approach uses a projection onto the space of shape functions which consists of *trigonometric* functions; the underlying mesh is equidistant. Both methods ([50] and our) have in common that polynomial

interpolation is used to avoid the expensive 'full order' evaluation of a function at the characteristics foot points (Sec. 5 in [50]). A main difference is that the scheme [50] performs a spectral decomposition into the ansatz functions.

As in our method no global systems of equations occur, our method is easy to parallelize even when used on 2d or 3d domains.

4.2 The scheme

In this section we will define some operators describing the extrapolation in time and the interpolation in space of the numerical solution which is a cruxial point of the algorithm. After that, the numerical scheme is given. In the Sections 4.3 and 4.4 the stability and the convergence of the schemes is investigated.

Let $\Omega = (0, L) \subset \mathbb{R}$ be an intervall and Ω_h be the set of grid points x_k, $0 = x_0 < ... < x_{l-1} = L$ and $h := \max\{x_{k+1} - x_k\}$ be the mesh size. Let $u^n(x_k)$ be the numerical solution at the grid point $x_k \in \Omega_h$ at the time step $t_n = n\,\Delta t$ and $u^{h,n} = (u^0, ..., u^n)$ the totality of the previously calculated time steps.

Spatial interpolation. For linear interpolation, we define $\mathcal{S}_1 u^n : \mathbb{R} \to \mathbb{R}$ as the piecewise linear function with $(\mathcal{S}_1 u^n)(x_k) = u^n(x_k)$ for all grid points $x_k \in \Omega_h$ (Fig. 4.1).
For higher order interpolation, we define the interpolant as follows:
For $p \geq 2$ let $\mathcal{S}_{p,\Delta k,k} u^n : \mathbb{R} \to \mathbb{R}$ be the polynomial of degree p which is defined[1] by

$$(\mathcal{S}_{p,\Delta k,k} u^n)(x_i) = u^n(x_i) \text{ for } i = k + \Delta k - p, ..., k + \Delta k. \tag{4.5}$$

Let

$$\mathcal{S}_{p,\Delta k} u^n(x) := \mathcal{S}_{p,\Delta k,k} u^n(x) \text{ for } x \in [x_k, x_{k+1}) \tag{4.6}$$

be our spatial interpolation function (see left part of Fig. 4.2). [2] We notice that the definition of $\mathcal{S}_1$ is consistent with definition (4.5)-(4.6) for $p=1$, $\Delta k=1$.

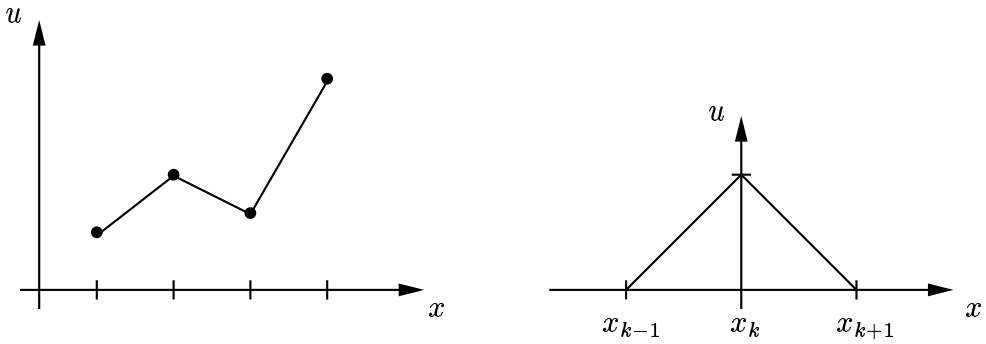


Figure 4.1: Linear interpolation ($p=1$) and related ansatz function.

Spatial ansatz functions. We want to identify the ansatz functions $\alpha_k = \alpha_{k,p,\Delta k}$ related to the interpolation operator $\mathcal{S}_{p,\Delta k}$ and the mesh (x_k), i.e. the functions α_k such that the expansion

$$\mathcal{S}_{p,\Delta k} u(x) = \sum_k u(x_k)\, \alpha_k(x) \tag{4.7}$$

[1] If x_k lies close to the boundary, a boundary condition may be necessary to define $\mathcal{S}_{p,k,\Delta k}$ properly. We will assume periodic boundary conditions.

[2] It is possible to replace the interval $[x_k, x_{k+1})$ by $(x_{k-1}, x_k]$. This would flip the ansatz functions (Fig. 4.2, right part).

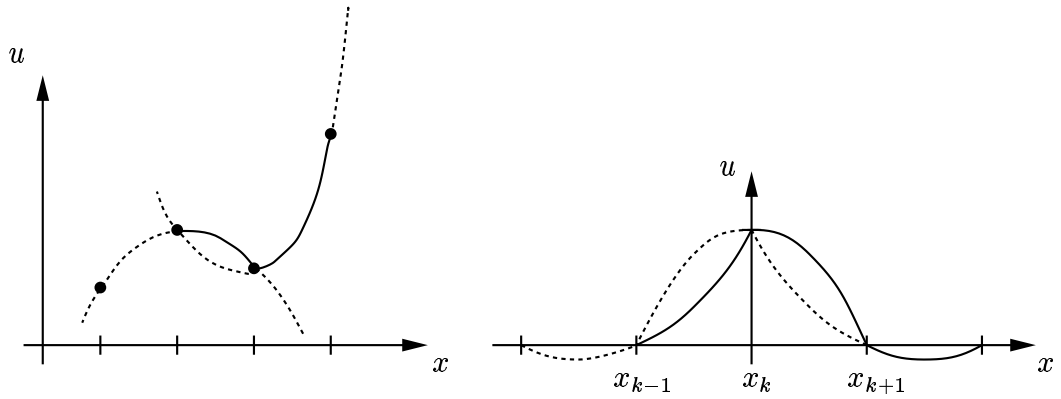


Figure 4.2: Left part: Second order interpolation ($p=2$, $\Delta k=1$). Right part: related ansatz function $\alpha_{k,2,1}$. For $\Delta k=2$ the ansatz function would be reflected at the y-axis. Only for odd p (and equidistant mesh) the ansatz functions are symmetric.

holds.

Let $L^j_{a,b}$ be the Lagrangian polynomial satisfying

$$\begin{aligned} L^j_{a,b}(x_j) &= 1, \\ L^j_{a,b}(x_i) &= 0 \quad \forall\, i=a,...,b;\ i \neq j \end{aligned} \tag{4.8}$$

if $a \leq j \leq b$ and $L^j_{a,b} \equiv 0$ for $j<a$, $j>b$.

Then it is easy to verify that

$$\alpha_k(x) = L^k_{j+\Delta k-p,j+\Delta k}(x) \quad \text{for all } x_j \leq x < x_{j+1}. \tag{4.9}$$

In Figs. 4.1, 4.2, α_k for $p=1$ rsp. $p=2$, $\Delta k=1$ are displayed. For equidistant mesh, all the ansatz functions have the same shape:

$$\alpha_k(x) = \alpha_{k+i}(x+ih)$$

Polynomial extrapolation in time. Using the numerical solution at the time steps t_n, t_{n-1},...,t_{n-q}, we define for $q \in \mathbb{N}_0$ and each $x \in \Omega$ the interpolant $(I^n_{q,p,\Delta k} u^h)(\cdot, x)$ as the polynomial of degree q with

$$(I^n_{q,p,\Delta k} u^{h,n})(t_\nu, x) = \mathcal{S}_{p,\Delta k} u^{h,n}(t_\nu, x) \text{ for } n \geq \nu \geq n-q \tag{4.10}$$

for $p \geq 2$. For $p=1$ we define correspondingly $(I^n_{q,1} u^{h,n})(\cdot, x)$ as the polynomial of degree q with

$$(I^n_{q,1} u^{h,n})(t_\nu, x) = \mathcal{S}_1 u^{h,n}(t_\nu, x) \text{ for } n \geq \nu \geq n-q. \tag{4.11}$$

We will use the abbreviation $I^n = I^n_{q,p,\Delta k}$ rsp. $I^n = I^n_{q,1}$.

Now we can state the algorithm: For given $u^{h,n}$, u^{n+1} is determined as follows:

(i) At each grid point x_k calculate $X(t_n; t_{n+1}, x_k)$, where the characteristic $X(\cdot) = X(\cdot; t_{n+1}, x_k)$ is the solution of the initial value problem

$$\begin{aligned} \frac{d}{dt}X(t) &= a(t, X(t), I^n u^{h,n}(t, X(t))) \\ X(t_{n+1}) &= x_k. \end{aligned} \tag{4.12}$$

(ii) Set

$$u^{n+1}(x_k) := I^n u^{h,n}(t_n, X(t_n)) = \mathcal{S}u^n(X(t_n)). \tag{4.13}$$

So we are using the extrapolated field $I^n u^{h,n}$ to calculate backward in time the characteristic starting at the point x_k at time t_{n+1}.

For a visualization of this process see Figs. 4.3, 4.4.

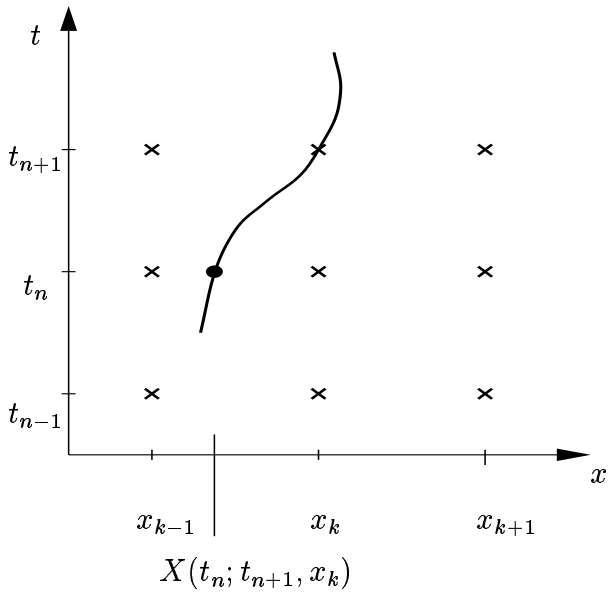


Figure 4.3: Visualization of a timestep.

Remarks.

- Our method has features both of an implicit and of an explicit method: It is implicit in the sense that the characteristics are calculated backward in time ($t \le t_{n+1}$) from a point (t_{n+1}, x_k). It is explicit in the sense that an extrapolation of the flow field in time is used.
- The described spatial interpolation by $\mathcal{S}$ may be substituted by spline interpolation, but that would mean to solve a global system of equation in each timestep which is also a disadvantage in case of parallel computing.

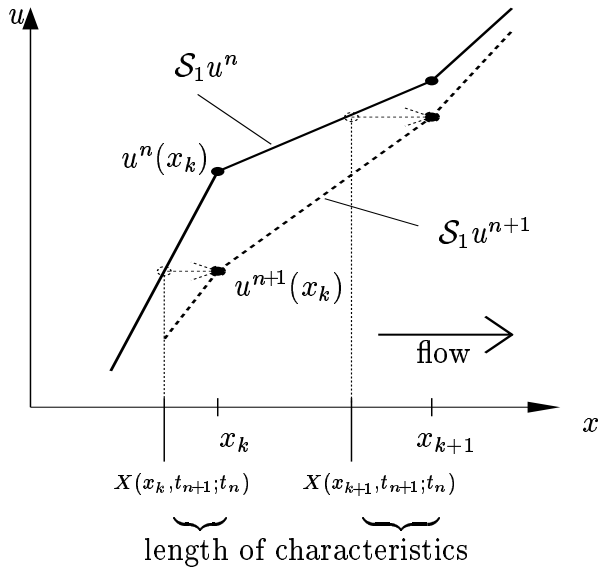


Figure 4.4: A timestep with $q=0$, $p=1$. The two graphs show the velocity distribution at time t_n and t_{n+1}. The hollow dots are the endpoints $X(t_n; t_{n+1}, x_k)$, $X(t_n; t_{n+1}, x_{k+1})$ of the characteristics. The two dotted horizontal arrows show the action of the update (4.13).

However, tests do not show any improvement of the spline interpolation compared to the application of our piecewise polynomial interpolation operator $\mathcal{S}$.

4.3 A convergence theorem for linear ansatz functions

In this section we will prove the unconditional stability and the convergence of the scheme for linear spatial interpolation ($\mathcal{S} = \mathcal{S}_1$, $I^n = I^n_{1,1}$, see Sec. 4.2) for the nonlinear transport equation (4.1) in the L^∞-norm.

We assume that for the time step size Δt taken from an interval $(0, \Delta t_0]$ the numerical integration method for the ordinary differential equation (4.12) has the following two properties:

If $y(\Delta t)$ and $\tilde{y}(\Delta t)$ are the numerical solutions of $y' = f(t,y)$ rsp. $\tilde{y}' = \tilde{f}(t,\tilde{y})$, $y(0) = \tilde{y}(0) = y_0$, and f, $\tilde{f}$ are smooth, then after one timestep

(i) $$|y(\Delta t) - \tilde{y}(\Delta t)| \le P(\Delta t)\,\Delta t \sup_{(t,x)\in[0,\Delta t]\times\mathbb{R}} |(f-\tilde{f})(t,x)| \;\forall\, \Delta t \in [0,T] \quad (4.14)$$

where $P(t) = P(t,f)$ is a bounded function for $t \in [0, \Delta t_0]$ independent of $\tilde{f}$, and

(ii) $$|y(\Delta t) - y(0)| \le C_0\,\Delta t \sup_{(t,x)\in[0,\Delta t]\times\mathbb{R}} |f(t,x)| \quad (4.15)$$

for a positive constant C_0.[3]

For the Runge-Kutta methods, (4.14) holds with $P(t) = 1 + \frac{t}{2}\|f_x\|_\infty$ for the 2^{nd} order Runge-Kutta method

$$y(t_0 + \Delta t) = y_0 + \Delta t\, f(t_0 + \frac{\Delta t}{2}, y_0 + \frac{\Delta t}{2} f(t_0, y_0)) \quad (4.16)$$

and $P(t) = \sum_{k=0}^{3} \frac{t^k \|f_x\|_\infty^k}{(k+1)!}$ for the classical 4^{th} order Runge-Kutta method.

For the *first* timestep we may use $I^n_{0,1}$ instead of $I^n_{1,1}$ or we may assume that we know an approximation of $u(-\Delta t)$.

Theorem 4.1 *Let $u \in C^2(\bar{Q})$ be the exact solution of (4.1), $a \in C^2(\bar{Q}\times\mathbb{R})$, $Q = [0,T]\times\Omega$, $\partial a/\partial u$ bounded, u^h the numerical solution according to Section 4.2 with $I^n = I^n_{1,1}$ (i.e. linear interpolation in space and time). Let us consider a numerical integration method fulfilling (4.14), (4.15) of order ≥ 2. For Δt, $h^2/\Delta t$ small enough, the following error estimate holds:*

$$\|u - u^h\|_{L^\infty_h} \le C\,\Delta t^2 + C\,\min\left\{h, \frac{h^2}{\Delta t}\right\} \quad (4.17)$$

where $\|\cdot\|_{L^\infty_h}$ denotes the maximum over all grid points $x_k \in \Omega_h$ and all time steps $n = 0, ..., T/\Delta t$ and $h := \max_k x_{k+1} - x_k$.

[3] For all methods of type $y(\Delta t) = y(0) + \Delta t \sum_i \alpha_i f(\xi_i)$ (e.g. Runge-Kutta methods, Adams methods), (4.15) holds with $C_0 = \sum_i |\alpha_i|$. For all consistent methods of this type with *positive* coefficients α_i, (4.15) holds with $C_0 = 1$.

Remark. This theorem can easily be generalized to higher order estimates *in time*: If we replace $I^n_{1,1}$ by $I^n_{q,1}$ ($q \in \mathbb{N}_0$) and tighten the regularity assumptions on a and u, then (4.17) holds with Δt^2 replaced by Δt^{q+1}. The increase of the *spatial* interpolation order requires a special investigation of the stability. This is pointed out at the end of this section.

Proof of Theorem 4.1. Let us assume a being bounded on $Q \times \mathbb{R}$ at first. We will use the abbreviations $U = \max_Q |u(t,x)|$, $A = \sup_{Q\times\mathbb{R}} |a(t,x,u)|$, $A_u = \max_Q |\frac{\partial a(t,x,u)}{\partial u}|$, $U_{tt} = \max_Q |\frac{\partial^2 u(t,x)}{\partial t^2}|$ and so on.

We suppose that we have already computed u^0,...,u^n with an error

$$|\mathcal{S}u^\nu(x) - u(t_\nu, x)| \le E_n \tag{4.18}$$

for $\nu = n-1, n$ and all points $x \in \tilde{\Omega}_h$ with[4]

$$\tilde{\Omega}_h := \Omega \cap \bigcup_{x_k \in \Omega_h} [x_k - C_0 A \Delta t, x_k + C_0 A \Delta t].$$

We have to derive how E_{n+1} depends on E_n.

(i) *Error of the characteristics.* Let $X = X(t; t_{n+1}, x_k)$ be the exact characteristic according to the exact flow $a(t,x,u(t,x))$, X_{num} the numerical characteristic for the same a and $X_{h,num}$ the numerical characteristic according to the numerical flow $a(t,x,I^n u^{h,n}(t,x))$.

Using (4.15) with $f(t,x) = a(t,x,u(t,x))$ rsp. $f(t,x) = a(t,x,I^n u^{h,n}(t,x))$, we get $X_{num}(t_n; t_{n+1}, x_k), X_{h,num}(t_n; t_{n+1}, x_k) \in \tilde{\Omega}_h$. Using (4.14) with $f(t,x) = a(t,x,u(t,x))$, $\tilde{f}(t,x) = a(t,x,I^n u^{h,n}(t,x))$, we get

$$|(f - \tilde{f})(t,x)| \le A_u \, |I^n u^{h,n}(t,x) - u(t,x)|$$

and

$$\begin{aligned} &|X_{h,num}(t_n; t_{n+1}, x_k) - X_{num}(t_n; t_{n+1}, x_k)| \\ &\quad \le A_u \, \Delta t \, P(\Delta t) \sup_{(t,x)\in[t_n,t_{n+1}]\times\tilde{\Omega}_h} |I^n u^{h,n}(t,x) - u(t,x)| \end{aligned} \tag{4.19}$$

with $P(\Delta t)$ depending on a, u, Δt, but independent of $u^{h,n}$.

By standard estimates for polynomial interpolation (see e.g. [47] Satz 3.4), we get for $x \in \tilde{\Omega}_h$

$$\begin{aligned} &\sup_{t\in[t_n,t_{n+1}]} |I^n u^{h,n}(t,x) - u(t,x)| \\ &\quad \le \sup_{t\in[t_n,t_{n+1}]} |I^n(u^{h,n} - u(t))(x)| + \sup_{t\in[t_n,t_{n+1}]} |(I^n u - u)(t,x)| \\ &\quad \le 3E_n + U_{tt}\Delta t^2 \end{aligned} \tag{4.20}$$

[4] This definition of $\tilde{\Omega}_h$ turns out to be useful in (4.23) in part 3 of the proof; obviously all endpoints $X(t_n; t_{n+1}, x_k)$ (and, due to (4.15), also its numerical versions X_{num}, $X_{h,num}$ defined below) are contained in $\tilde{\Omega}_h$.

for all $x \in \tilde{\Omega}_h$. The factor 3 results from the extrapolation in time. Hence, the error in computing the characteristics is

$$\begin{aligned}
&|X_{h,num}(t_n; t_{n+1}, x_k) - X(t_n; t_{n+1}, x_k)| \\
&\quad \leq |X_{h,num}(t_n; t_{n+1}, x_k) - X_{num}(t_n; t_{n+1}, x_k)| \\
&\qquad + |X_{num}(t_n; t_{n+1}, x_k) - X(t_n; t_{n+1}, x_k)| \\
&\quad \leq (3E_n + U_{tt}\Delta t^2)\, A_u \Delta t\, P(\Delta t) + C_2 \Delta t^3 \qquad (4.21)
\end{aligned}$$

where C_2 depends on the numerical integration method for the calculation of the characteristics.

(ii) *Error of u_h at the grid points x_k.* With (4.13) and (4.21) we obtain

$$\begin{aligned}
&|u^{n+1}(x_k) - u(t_{n+1}, x_k)| \\
&\quad = |\mathcal{S}u^n(X_{h,num}(t_n; t_{n+1}, x_k)) - u(t_n, X(t_n; t_{n+1}, x_k))| \\
&\quad \leq |\mathcal{S}u^n(X_{h,num}(t_n; t_{n+1}, x_k)) - u(t_n, X_{h,num}(t_n; t_{n+1}, x_k))| \\
&\qquad + |u(t_n, X_{h,num}(t_n; t_{n+1}, x_k)) - u(t_n, X(t_n; t_{n+1}, x_k))| \\
&\quad \leq E_n + U_x\, |X_{h,num}(t_n; t_{n+1}, x_k) - X(t_n; t_{n+1}, x_k)| \\
&\quad \leq (1 + 3\, U_x A_u \Delta t\, P(\Delta t))\, E_n + C_3 \Delta t^3 \qquad (4.22)
\end{aligned}$$

with $C_3 = U_x(C_2 + U_{tt} A_u\, P(\Delta t))$.

(iii) *Error of Iu_h for $x \in \tilde{\Omega}_h$.* Using standard approximation results (again [47] Satz 3.4) and the definition of $\tilde{\Omega}_h$, we get for arbitrary $x \in \tilde{\Omega}_h$:

$$\begin{aligned}
&|\mathcal{S}u^{n+1}(t_{n+1}, x) - u(t_{n+1}, x)| \\
&\quad \leq |\mathcal{S}(u^{n+1} - u(t_{n+1}))(x)| + |\mathcal{S}u(t_{n+1}, x) - u(t_{n+1}, x)| \\
&\quad \leq \max_{x_k \in \Omega_h} |(u^{n+1} - u(t_{n+1}))(x_k)| + \frac{1}{2} U_{xx} h \min\{A\Delta t, \frac{h}{4}\}. \qquad (4.23)
\end{aligned}$$

Taking the maximum of (4.23) over all $x \in \tilde{\Omega}_h$ and (4.22) we get the recurrency inequality

$$E_{n+1} \leq \alpha E_n + \beta \qquad (4.24)$$

with

$$\alpha = 1 + 3\, U_x A_u \Delta t\, P(\Delta t), \qquad \beta = C_3 \Delta t^3 + \frac{1}{2} U_{xx} h \min\{A\Delta t, \frac{h}{4}\}. \qquad (4.25)$$

Assuming $E_0 \leq \beta$ we have

$$\begin{aligned}
E_n &\leq \beta \sum_{k=0}^{n} \alpha^k = \beta\, \frac{\alpha^{n+1} - 1}{\alpha - 1} \qquad (4.26) \\
&\leq \frac{\exp(3\, U_x A_u (T + \Delta t)\, P(\Delta t))}{3\, U_x A_u\, P(\Delta t)} \left(C_3\, \Delta t^2 + \frac{U_{xx}}{2} \frac{h}{\Delta t} \min\{A\, \Delta t, \frac{h}{4}\} \right).
\end{aligned}$$

We now may drop the assumption that a is bounded. If a is unbounded on $Q\times\mathbb{R}$ we just have to replace A in the proof by $A := \sup_{\substack{(t,x)\in Q\\ |u|\le U+\bar U}} |a(t,x,u)|$ which is $<\infty$.

Here we have set $\bar U := \frac{\exp(3\,U_x\,A_u(T+\Delta t)\,P(\Delta t))}{3U_x\,A_u\,P(\Delta t)}\left(C_3\,\Delta t^2 + \frac{U_{xx}h^2}{8\,\Delta t}\right)$ ∎

Unfortunately, this proof fails for higher order spatial interpolation $\mathcal{S}_{p,\Delta k}$ with $p>1$. Then, the estimate

$$|\mathcal{S}_{p,\Delta k}e(x)| \le \max_k |e(x_k)| \tag{4.27}$$

which was used in (4.23) no longer holds (see Fig. 4.5),

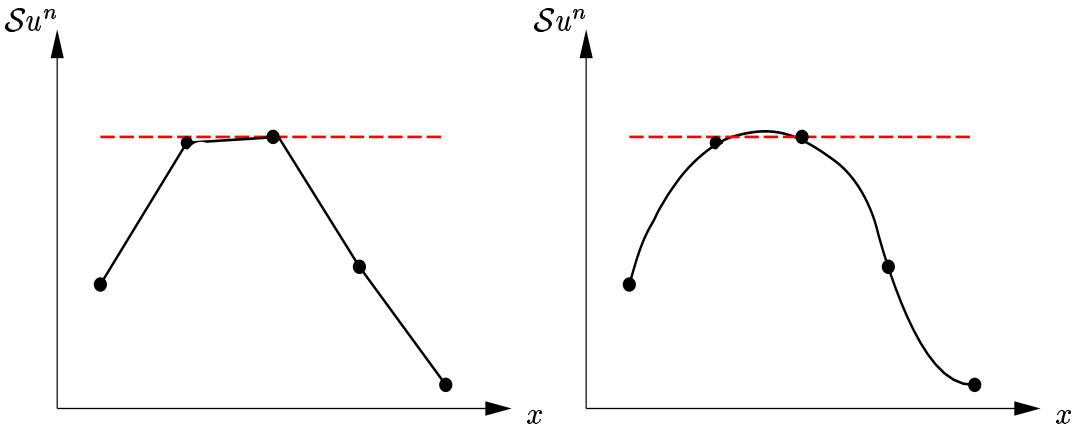


Figure 4.5: For linear interpolation, the maximum is always taken at a grid point. For higher order interpolation, this is false. This can be a cause of instability.

but only

$$|\mathcal{S}_{p,\Delta k}e(x)| \le C_{max}(p)\max_k |e(x_k)|. \tag{4.28}$$

See Lemma 4.4 (i) where (4.28) is proved.

However, numerical tests indicated the stability also for higher order ansatz functions. This is investigated in Section 4.4.

4.4 Stability and convergence for higher order ansatz functions

In this section we will prove the stability and the convergence of the scheme for higher order ansatz functions ($p \geq 2$). To avoid the problem of (4.27)-(4.28) we will use a discrete L^2-norm

$$\|f\|_{L_h^2} := \left(\sum_{k=0}^{l-1} w_{k,h} \, |f(x_k)|^2 \right)^{1/2} \tag{4.29}$$

instead of the L_h^∞-norm because in L_h^∞ we are not able to prove stability. In (4.29), the $w_{k,h}$ are positive weigths with

$$C_{w1} \geq \sum_{k=0}^{l-1} w_{k,h}, \qquad \frac{w_{k+1,h}}{w_{k,h}} \leq C_{w2}, \quad w_{k,h} \geq C_{w3} \min_k x_{k+1} - x_k \quad \forall \, k, h \tag{4.30}$$

where the constants C_{w1}, C_{w2}, C_{w3} are independent of k, h. An appropriate choice of $w_{k,h}$ could be[5] $w_{k,h} = \frac{1}{2}(x_{k+1} - x_{k-1})$; see Sections 4.4.3, 4.4.5 for this point.

Throughout this section we will assume periodic boundary conditions.

In contrast to the linear interpolation $\mathcal{S}_1$ in Section 4.3 the use of higher order interpolation $\mathcal{S}_{p,\Delta k,k}$ will lead to a restriction on the time step size Δt. The validity of the stability condition (4.34) (taking the role of (4.27) in the linear case) is investigated in the Sections 4.4.2-4.4.6.

Before this, in section 4.4.1 we will show that this stability condition implies the convergence of the scheme in the L_h^2-norm.

4.4.1 Convergence

We are going to prove an error estimate for our characteristics method for higher order interpolation.

Let $h_k := x_{k+1} - x_k$, $h := \max h_k$ and $h_{min} := \min h_k$. Let us assume that the 'local variation' of the mesh size is bounded: There is a constant $C_{mesh} = C_{mesh}(p)$ independent of h, h_{min} for all the meshes under consideration such that[6]

$$\frac{\max\limits_{j=k+1,\ldots,k+p} h_j}{\min\limits_{j=k+1,\ldots,k+p} h_j} \leq C_{mesh} \quad \forall \, k = 0, \ldots, l. \tag{4.31}$$

[5] For this choice of $w_{k,h}$, $\|u\|_{L_h^2}^2$ is the approximation of $\|u\|_{L^2}^2$ with the trapezoid rule. As we are dealing with periodic functions, the trapezoid rule provides extra high accuracy.

[6] For the Gauss-Lobatto mesh, (4.31) holds with $C_{mesh} = p^2$.

Theorem 4.2 *Let $\Omega = (0, L)$. Let $u \in C^{q,p}(\bar{Q})$ be the exact solution of (4.1), $a \in C^{\max\{q+1,p+1\}}(\bar{Q} \times \mathbb{R})$, $Q = [0,T] \times \Omega$ and $u^{h,n}$ the numerical solution according to Section 4.2 with $I = I_{q,p,\Delta k}$, $p \geq 2$, $q \geq 0$ and*

$$\Delta t \leq c\, h_{min}, \tag{4.32}$$

$$\Delta t^{q+1} + h^p \leq c\, h_{min}^{3/2+\epsilon}, \quad \epsilon > 0. \tag{4.33}$$

We propose that we are using a numerical integration method fulfilling (4.14), (4.15) of order $\geq q+1$ and that (4.30) holds. Suppose that the stability condition

$$\|\mathcal{S} f(X(t_n; t_{n+1}, \cdot))\|_{L_h^2} \leq (1 + C_{stab}\, \Delta t)\, \|f\|_{L_h^2} \tag{4.34}$$

holds for any discrete function f given on the grid points with C_{stab} independent of h, Δt, f. Then for all $h \leq h_0$, $h_0 > 0$, the following error estimate holds:

$$\max_n \|u(t_n) - u^{h,n}\|_{L_h^2} \leq c\, \Delta t^{q+1} + c \min\left\{ \frac{h^{p+1}}{\Delta t}, h^p \right\} \tag{4.35}$$

In (4.35) the maximum is taken over all time steps $t_0 = 0$, $t_1 = \Delta t, \ldots, t_N = N\,\Delta t = T$ and the c are independent of h, Δt, u^h.

Before the proof we present two preparatory lemmas which are proved at the end of the section:

Lemma 4.3 *Let*

$$\mathbb{P}_{p,[0,s]} := \{P : [0,s] \to \mathbb{R} \mid P \text{ is polynomial with degree} \leq p\}$$

and $0 = x_0 < \ldots < x_p = s$ be grid points on $[0,s]$. Then there is a constant $C_4 = C_4(p)$ independent of s, P and the distribution of the x_k such that

$$\max_{x \in [0,s]} |P'(x)| \leq C_4 \max_{k=1,\ldots,p} \frac{|P(x_k) - P(x_{k-1})|}{x_k - x_{k-1}} \tag{4.36}$$

for all $P \in \mathbb{P}_{p,[0,s]}$.

Lemma 4.4 *Under the assumption (4.31), for all functions e given discretely on the mesh,*

(i) (4.28) holds with $C_{max} = (p+1)\,(pC_{mesh})^p$,

(ii) $$\sup_{x \in [0,L]} |(\mathcal{S}_{p,\Delta k} e)'(x)| \leq C_5\, h_{min}^{-1} \max_{k=0,\ldots,l} |e(x_k)|$$

holds with $C_5 = 2\, C_4$.

Proof of Theorem 4.2. Let us define the discrete error

$$e^\nu(x_k) := u^\nu(x_k) - u(t_\nu, x_k), \quad \nu = 0, ..., T/\Delta t, \; x_k \in \Omega_h$$

and its norm

$$\tilde{E}_n := \max_{0 \le \nu \le n} \|e^\nu\|_{L_h^2}, \quad n = 0, ..., T/\Delta t.$$

The estimate

$$E_n \le C_{w3}^{-1/2} h_{min}^{-1/2} \; \tilde{E}_n \tag{4.37}$$

holds due to (4.30).

Let us assume that a is bounded, at first.

(i) *The decomposition of the error.* We have to find an estimate of $\tilde{E}_{n+1}$ in dependence on $\tilde{E}_n$. So we have to estimate e^{n+1}. We perform the decomposition

$$\begin{aligned}
\|e^{n+1}\|_{L_h^2} &= \|u^{n+1}(\cdot) - u(t_{n+1}, \cdot)\|_{L_h^2} \\
&= \|\mathcal{S}u^n(X_{h,num}(t_n; t_{n+1}, \cdot)) - u(t_n, X(t_n; t_{n+1}, \cdot))\|_{L_h^2} \\
&\le \|\mathcal{S}u^n(X_{h,num}(t_n; t_{n+1}, \cdot)) - \mathcal{S}u^n(X(t_n; t_{n+1}, \cdot))\|_{L_h^2} \\
&\quad + \|\mathcal{S}u^n(X(t_n; t_{n+1}, \cdot)) - \mathcal{S}u(X(t_n; t_{n+1}, \cdot))\|_{L_h^2} \\
&\quad + \|\mathcal{S}u(X(t_n; t_{n+1}, \cdot)) - u(X(t_n; t_{n+1}, \cdot))\|_{L_h^2}.
\end{aligned} \tag{4.38}$$

Similar to (4.23), the last term in (4.38) is estimated by $\frac{U_{x^{p+1}}}{(p+1)} h^p \min\{A\,\Delta t, h\}$. Using the stability, the second but last term in (4.38) is estimated by $(1 + C_{stab}\,\Delta t)\tilde{E}_n$. Let

$$L(f) := \sup_{x \ne y \in \Omega} \frac{|f(x) - f(y)|}{|x - y|}$$

be the Lipschitz constant of a function f. The remaining term in (4.38) can be estimated by the Lipschitz constant $L(\mathcal{S}u^n)$ and the error of the characteristics:

$$\begin{aligned}
&\|\mathcal{S}u^n(X_{h,num}(t_n; t_{n+1}, \cdot)) - \mathcal{S}u^n(X(t_n; t_{n+1}, \cdot))\|_{L_h^2} \\
&\quad \le L(\mathcal{S}u^n)\, \|X_{h,num}(t_n; t_{n+1}, \cdot) - X(t_n; t_{n+1}, \cdot)\|_{L_h^2}.
\end{aligned} \tag{4.39}$$

(ii) *An estimate for the Lipschitz constant in (4.39).* It is

$$L(\mathcal{S}u^n) \le L(\mathcal{S}u(t_n)) + L(\mathcal{S}e^n).$$

$L(\mathcal{S}u(t_n))$ is bounded due to the regularity of u: With help of Lemma 4.3 and the mean value theorem we get $L(\mathcal{S}u(t_n)) \le C_4 U_x$. The Lipschitz constant of $\mathcal{S}e^n$ is estimated by Lemma 4.4 (ii) and (4.37):

$$L(\mathcal{S}e^n) \le C_5\, h_{min}^{-1}\, E_n \le C_5 C_{w3}^{-1/2}\, h_{min}^{-3/2}\; \tilde{E}_n \tag{4.40}$$

(iii) *The error of the characteristics.* Similar to (4.19)-(4.21) we get

$$|X_{num}(t_n;t_{n+1},x_k) - X(t_n;t_{n+1},x_k)| \le C_2\,\Delta t^{q+2} \tag{4.41}$$

and

$$\begin{aligned}
&|X_{h,num}(t_n;t_{n+1},x_k) - X_{num}(t_n;t_{n+1},x_k)| \\
&\quad\le A_u\,\Delta t\,P(\Delta t) \sup_{\substack{|x-x_k|\le C_0 A\,\Delta t \\ t\in[t_n,t_{n+1}]}} |I^n u^{h,n}(t,x) - u(t,x)| \\
&\quad\le A_u\,\Delta t\,P(\Delta t)\ \ (C'(q) \sup_{\substack{|x-x_k|\le C_0 A\,\Delta t \\ \nu\le n}} |\mathcal{S}e^\nu(x)|\ +\ U_{t^{q+1}}\Delta t^{q+1})
\end{aligned} \tag{4.42}$$

where $C'(q)$ is a constant such that $|P(q+1)| \le C'(q) \max\limits_{\nu=0,\dots,q} |P(\nu)|$ for all polynomials P of degree $\le q$. The interpolation operator $\mathcal{S}$ has local character: The values of the above supremum only depend on the nodal values $e^\nu(x_{k+j})$, $j\in G$,

$$G := \{-\lfloor C_0 A\,\Delta t\rfloor/h_{min} + \Delta k - p, \dots, \lceil C_0 A\,\Delta t\rceil/h_{min} + \Delta k\}.$$

Due to (4.32), $|G|$ is bounded independent of h, Δt (the Courant number is bounded). So with (4.28), (4.30) we get

$$\begin{aligned}
&\sum_k w_{k,h} \sup_{\substack{|x-x_k|\le C_0 A\,\Delta t \\ \nu\le n}} |\mathcal{S}e^\nu(x)|^2 \le C_{max}(p)^2 \sum_k w_{k,h} \sup_{\substack{j\in G \\ \nu\le n}} |e^\nu(x_{k+j})|^2 \\
&\le\ C_{max}(p)^2\, C_{w2}^{|G|} \sum_k \sup_{\substack{j\in G \\ \nu\le n}} w_{k+j,h}|e^\nu(x_{k+j})|^2 \\
&\le\ C_{max}(p)^2\, C_{w2}^{|G|}\, |G|\, \tilde{E}_n^2.
\end{aligned} \tag{4.43}$$

(iv) *Collecting the results.* (4.39)-(4.43) yield

$$\begin{aligned}
&\|\mathcal{S}u^n(X_{h,num}(t_n;t_{n+1},\cdot)) - \mathcal{S}u^n(X(t_n;t_{n+1},\cdot))\|_{L_h^2} \\
&\le\ c\,\Delta t^{q+2} + c\,\Delta t^{q+2}\,h_{min}^{-3/2}\tilde{E}_n + c\,\Delta t\,\tilde{E}_n + c\,\Delta t\,h_{min}^{-3/2}\,\tilde{E}_n^2 \\
&\le\ c\,\Delta t^{q+2} + c\,\Delta t\,\tilde{E}_n + c\,\Delta t\,h_{min}^{-3/2}\,\tilde{E}_n^2
\end{aligned} \tag{4.44}$$

where the constants c are generic and where we have used (4.33) in the last step. All in all we get from (i) and (4.44)

$$\tilde{E}_{n+1} \le C_6\,\Delta t\,h_{min}^{-3/2}\tilde{E}_n^2 + (1 + C_7\,\Delta t)\tilde{E}_n + C_8\,\Delta t^{q+2} + C_9\,h^p\,\min\{A\,\Delta t, h\}. \tag{4.45}$$

Unlike the proof of Theorem 4.1 we have to treat the term $\Delta t\,h_{min}^{-3/2}\tilde{E}_n^2$.
(v) *Solving the recurrency inequality.* Let $C_{10} > 0$ be an arbitrary constant. As an abbreviation we use

$$\alpha := C_7 + C_6 C_{10}, \quad \beta := C_8\,\Delta t^{q+2} + C_9\,h^p\,\min\{A\,\Delta t, h\}.$$

Suppose that h, Δt are small enough that

$$\beta \frac{e^{\alpha T}}{\alpha \, \Delta t} \leq C_{10} h_{min}^{3/2} \tag{4.46}$$

(here we have used (4.33)). Using the inductive hypothesis

$$\tilde{E}_n \leq \beta \frac{(1+\alpha\Delta t)^n - 1}{\alpha \Delta t} \tag{4.47}$$

and the estimate $(1+\alpha\Delta t)^n \leq e^{\alpha T}$ we get

$$\tilde{E}_n \leq \beta \frac{e^{\alpha T}}{\alpha \Delta t} \leq C_{10} h_{min}^{3/2} \tag{4.48}$$

and therefore the first summand in (4.45) is estimated by $C_6 C_{10} \, \Delta t \, \tilde{E}_n$ and

$$\tilde{E}_{n+1} \leq (1+\alpha\Delta t) \, \tilde{E}_n + \beta$$

follows from (4.45). With (4.47) we get

$$\tilde{E}_{n+1} \leq \beta \frac{(1+\alpha\Delta t)^{n+1} - 1}{\alpha \Delta t}.$$

Hence (4.47) holds for *all* time steps n, and the left part of (4.48) shows that (4.35) holds.

The generalization to unbounded a follows analogously to the proof of Theorem 4.1. ∎

Proof of Lemma 4.3.
(i). On the finite dimensional quotient space

$$Q_{p,[0,s]} := \mathbb{P}_{p,[0,s]} / \mathbb{P}_{0,[0,s]}$$

both $\|[P]\|_c := s \max\limits_{x\in[0,s]} |P'(x)|$ and $\|[P]\|_d := s \max\limits_{k=1,\ldots,p} \frac{|P(x_k)-P(x_{k-1})|}{x_k - x_{k-1}}$ are norms. Therefore there is a constant $C_4 = C_4(p, s, x_1, \ldots, x_{p-1})$ with

$$\max_{x\in[0,s]} |P'(x)| \leq C_4 \max_{k=1,\ldots,p} \frac{|P(x_k) - P(x_{k-1})|}{x_k - x_{k-1}} \tag{4.49}$$

for all $P \in \mathbb{P}_{p,[0,s]}$.
(ii). Let us fix s at first. We have to show that C_4 in (4.49) can be chosen independently of the distribution of the mesh points.

Suppose that there is no constant C_4 independent of $x_1, \ldots, x_{p-1}$ such that (4.36) holds. Then there is a sequence of meshes $0 = x_0^{(n)} < \ldots < x_p^{(n)} = s$, $n \in \mathbb{N}$ and a sequence of polynomials $P_n \in \mathbb{P}_{p,[0,s]}$ such that $\max\limits_{x\in[0,s]} |P'(x)| = 1$ for all n

and $\max_k \frac{|P_n(x_k^{(n)})-P_n(x_{k-1}^{(n)})|}{x_k^{(n)}-x_{k-1}^{(n)}} \overset{n\to\infty}{\longrightarrow} 0$. As the sequences $x_k^{(n)}$ are bounded, there are convergent subsequences, again denoted by $x_k^{(n)}$, $x_k^{(n)} \overset{n\to\infty}{\longrightarrow} x_k$ for all $k=0,...,p$. As $\mathbb{P}_{p,[0,s]}$ has a finite dimension, there is a subsequence such that additionally $P_n \overset{n\to\infty}{\longrightarrow} P \in \mathbb{P}_{p,[0,s]}$ in $\|\cdot\|_{L^\infty([0,b])}$. For arbitrary $\epsilon>0$ and n sufficiently large we get

$$
\begin{aligned}
&|P(x_k)-P(x_{k-1})| \\
&\le |P(x_k)-P_n(x_k)| + s\,\frac{|P_n(x_k)-P_n(x_{k-1})|}{x_k-x_{k-1}} + |P_n(x_{k-1})-P(x_{k-1})| \le 2\epsilon + s\epsilon.
\end{aligned}
$$

Hence, $P(x_0)=...=P(x_p)=const$. So P is a constant polynomial. As $\mathbb{P}_{p,[0,s]}$ is of finite dimension, the convergence $P_n \to P=const$ holds in arbitrary norms, especially $\max_{x\in[0,s]} |P_n'(x)| \overset{n\to\infty}{\longrightarrow} 0$ which is a contradiction.

(iii). We have to show that $C_4=C_4(p,s)$ in (4.49) can be chosen independently of s. For every linear transformation $\mathcal{L}:[0,s']\to[0,s]$, the mapping $P\mapsto P\circ\mathcal{L}$, $x_k \mapsto x_k\, s'/s$ is a norm preserving isomorphism $Q_{p,[0,s]} \to Q_{p,[0,s']}$ both in the sense of $\|\cdot\|_c$ and $\|\cdot\|_d$. So for all $s'>0$, (4.49) holds with the same constant $C_4(p,s)$. ∎

Proof of Lemma 4.4.

ad (i). Without loss of generality we can assume $|e(x_j)| \le 1$. Let $x^* := \arg\max_{x\in[0,L]} |\mathcal{S}_{p,\Delta k} e(x)|$. Then $\mathcal{S}_{p,\Delta k} e(x^*) = \mathcal{S}_{p,\Delta k,k_0}(x^*)$ for a certain k_0 (see (4.6)). Using the representation of $\mathcal{S}_{p,\Delta k,k_0}$ through Lagrangian polynomials

$$
\mathcal{S}_{p,\Delta k,k_0} e(x^*) = \sum_{j=k_0+\Delta k-p}^{k_0+\Delta k} e(x_j) \prod_{\substack{i=k_0+\Delta k-p \\ i\ne j}}^{k_0+\Delta k} \frac{x^*-x_i}{x_j-x_i}
$$

we arrive at

$$
|\mathcal{S}_{p,\Delta k} e(x^*)| \le (p+1)\,(pC_{mesh})^p.
$$

ad (ii). Let $x^* := \arg\max_{x\in[0,L]} |(\mathcal{S}_{p,\Delta k} e)'(x)|$ and let k_0 such that $(\mathcal{S}_{p,\Delta k} e)'(x^*) = (\mathcal{S}_{p,\Delta k,k_0} e)'(x^*)$. Let $\tilde{x}_0 := x_{k_0+\Delta k-p}, ..., \tilde{x}_p := x_{k_0+\Delta k}$ be the grid points on the interval $[x_{k_0+\Delta k-p}, x_{k_0+\Delta k}]$. With Lemma 4.3 we get

$$
\begin{aligned}
&|(\mathcal{S}_{p,\Delta k} e)'(x^*)| = |(\mathcal{S}_{p,\Delta k,k_0} e)'(x^*)| \\
&\le C_4 \max_{k=1,...,p} \frac{|\mathcal{S}_{p,\Delta k,k_0} e(\tilde{x}_k) - \mathcal{S}_{p,\Delta k,k_0} e(\tilde{x}_{k-1})|}{h_{min}} \le \frac{2\,C_4}{h_{min}} \max_{k=0,...,p} |\mathcal{S}_{p,\Delta k,k_0} e(\tilde{x}_k)| \\
&= \frac{2\,C_4}{h_{min}} \max_{k=0,...,p} |e(\tilde{x}_k)|.
\end{aligned}
$$

∎

4.4.2 Stability on equidistant grid

In this and the following sections we investigate the question of stability of the update step (4.13) in the sense of (4.34). We will make use of the Lax stability theory [32].

Throughout Section 4.4.2 we assume the grid points to be equidistantly distributed:

$$x_k = \frac{Lk}{l}, \quad k=0,...,l, \quad h = \frac{L}{l}, \quad \Omega_h := \{x_k \,|\, k=0,...,l\}$$

Let $v : \Omega \times [0,T] \to \mathbb{R}$ be the flow field which is used to calculate the end point $X(t_n; t_{n+1}, x_k)$ of the characteristic starting at (t_{n+1}, x_k). We will assume that $v(t)$ is C^1 in x. This regularity assumption is justified by our purpose: For the application of the stability (4.34) (on the second but last term in (4.38)), the characteristics X are computed with respect to the *exact* flow

$$v(t,x) = a(t,x,u(t,x));$$

the less regular approximation $a(t,x,I^n u^{h,n}(t,x))$ is not involved.

We assume that $v(t,\cdot)$ is L-periodic. So for the sake of clarity we will use the writing $x_{l-1} = x_{-1}$, $x_l = x_0$, $x_{l+1} = x_1$,...

Furthermore, we write

$$\delta(x_k) := X(t_n; t_{n+1}, x_k) - x_k,$$

and we define the Courant-Friedrichs-Lewy number

$$C_{CFL} := \frac{\max_k\{|\delta(x_k)|\}}{h}. \tag{4.50}$$

which is bounded by $\frac{A\,\Delta t}{h}$, $A = \sup_{(t,x)\in Q} |a(t,x,u(t,x))|$.

The stability problem (4.34) can be written

$$\|\mathcal{S}u(\cdot + \delta(\cdot))\|_{L_h^2} \le (1 + C_{stab}\,\Delta t)\, \|u\|_{L_h^2} \tag{4.51}$$

or

$$\|D_v u\|_{L_h^2} \le (1 + C_{stab}\,\Delta t)\, \|u\|_{L_h^2} \tag{4.52}$$

for all $u : \bar{\Omega}_h \to \mathbb{R}$ being l-periodic. Here,

$$D_v u(x_k) := \mathcal{S}u(x_k + \delta(x_k)). \tag{4.53}$$

We cannot assume any regularity[7] of the u because (4.51) is applied to the *error* e^n in part (i) of the proof of Theorem 4.2. Therefore we will have to focus on the regularity of the $\delta(x_k)$ using the regularity of the flow field v (see proof of Lemma 4.5).

[7] like e.g. a bound for $|u(x_{k+1}) - u(x_k)|/|x_{k+1} - x_k|$

Lemma 4.5 *On an equidistant grid the coefficients c_j in the representation*

$$D_v u(x_k) = \sum_{j=-N}^{N} c_j(x_k)\, u(x_{k-j}) \tag{4.54}$$

have bounded differential quotients, and the bound $L(c_i)$ depends linearly on $\Delta t/h$ for all $\Delta t \le \Delta t_0$, $\Delta t_0 > 0$:

$$|c_j(x) - c_j(x')| \le L(c_j)\,|x - x'| \quad \text{for all } x \in \Omega_h,\ j = 0, ..., l-1, \tag{4.55}$$

$L(c_j) \le c\,\frac{\Delta t}{h}$.

Proof. Let $\bar{k}$ be the index such that $x_k + \delta(x_k) \in [x_{\bar{k}}, x_{\bar{k}+1})$. Let $L^j_{a,b}$ be the Lagrangian polynomial defined in (4.8) and α_k the ansatz function (4.7)/(4.9).

So we get the representation in the following (finite) sum:

$$\begin{aligned} D_v u(x_k) &= \mathcal{S}_{p,\Delta k} u(x_k + \delta(x_k)) \\ &= \sum_{j \in \mathbf{Z}} u(x_j)\, \alpha_j(x_k + \delta(x_k)) \\ &= \sum_{j \in \mathbf{Z}} u(x_j)\, L^j_{\bar{k}+\Delta k-p,\bar{k}+\Delta k}(x_k + \delta(x_k)) \\ &= \sum_{j \in \mathbf{Z}} u(x_{k-j})\, L^{k-j}_{\bar{k}+\Delta k-p,\bar{k}+\Delta k}(x_k + \delta(x_k)) \end{aligned} \tag{4.56}$$

As the mesh is equidistant,

$$L^j_{a,b}(x) = L^{j+i}_{a+i,b+i}(x + ih) \tag{4.57}$$

holds. So

$$D_v u(x_k) = \sum_{j \in \mathbf{Z}} L^{-j}_{\bar{k}-k+\Delta k-p,\bar{k}-k+\Delta k}(\delta(x_k))\ u(x_{k-j}) \tag{4.58}$$

and the coefficients in (4.54) are

$$c_j(x_k) = L^{-j}_{\bar{k}-k+\Delta k-p,\bar{k}-k+\Delta k}(\delta(x_k)). \tag{4.59}$$

We can set $N := \lceil C_{CFL} \rceil + p$ in (4.54).

We have to estimate $L(c_i)$. Obviously

$$L(c_i) \le L(\delta) \max_{j=0,...,p} L(L^j_{0,p}) \tag{4.60}$$

and $L(L^j_{0,p}) \le c(p)\, h^{-1}$ where $c(p)$ is a constant only depending on p. Hence, we just need an estimate for $L(\delta)$:

Define $X_k(t) := X(t; t_{n+1}, x_k)$,

$$\Phi(t) := (X_{k+1}(t) - x_{k+1}) - (X_k(t) - x_k)$$

and $\alpha := A_x + A_u U_x$.[8] Then,

$$\begin{aligned}\frac{d}{dt}\Phi(t) &= a(t, X_{k+1}(t), u(t, X_{k+1}(t))) - a(t, X_k(t), u(t, X_k(t))) \\ &\leq \alpha\,(X_{k+1} - X_k)(t) \\ &= \alpha\,\Phi(t) + \alpha\,(x_{k+1} - x_k)\end{aligned}$$

and $\Phi(t_{n+1}) = 0$. Using Gronwall's lemma (e.g. [56] Chapter I.1) we get

$$\begin{aligned}\delta(x_{k+1}) - \delta(x_k) &= \Phi(t_n) \leq \alpha\,(x_{k+1} - x_k) \int_0^{\Delta t} e^{\alpha(\Delta t - s)}\,ds \\ &= (x_{k+1} - x_k)(e^{\alpha \Delta t} - 1) \leq (x_{k+1} - x_k)\,\alpha\,\Delta t\,e^{\alpha \Delta t}.\end{aligned}$$

Thus

$$\begin{aligned}L(\delta) &= \alpha\,\Delta t\,e^{\alpha \Delta t}, \qquad (4.61)\\ L(c_j) &= \alpha\,c(p)\,\frac{\Delta t}{h}\,e^{\alpha \Delta t}.\end{aligned}$$

■

We are going to apply the result of Lemma 4.5 in the context of a stability theorem by Lax[9]:

Theorem 4.6 *For any operator $\mathcal{D}$ of type*

$$\mathcal{D}u(x_k) = \sum_{j=0}^{l-1} c_j(x_k)\,u(x_{k-j}) \qquad (4.62)$$

we define the so-called amplification factor

$$C(x, \xi) := \sum_{j=0,\ldots,l-1} c_j(x)\,e^{ij\xi}, \qquad (4.63)$$

and let us restrict ourselves to the non-weighted case $w_{k,h} := h$ in the definition (4.29) of the L_h^2-norm. $\mathcal{D}$ is stable in the sense that

$$\|\mathcal{D}u\|_{L_h^2} \leq (1 + c\,h)\,\|u\|_{L_h^2} \qquad (4.64)$$

holds for any L-periodic $u : \bar{\Omega}_h \to \mathbb{R}$ if the following conditions are met:

(i) C has bounded difference quotients with respect to x, i.e.

$$\|C(x) - C(x')\|_* \leq c\,|x - x'| \quad \forall\, x, x' \in \Omega_h \qquad (4.65)$$

where the norm $\|C(x)\|_*$ is defined as the maximum of $|C(x, \xi)|$ with respect to ξ in some strip around the real ξ-axis.

[8] See definitions in the beginning of the proof of Theorem 4.2.

[9] A short overview over the development of stability results on difference schemes is given in [36] Chapter 8.1.

(ii) For all $x \in \Omega_h$ and all real ξ

$$|C(x,\xi)| \leq 1. \tag{4.66}$$

(iii) There is a representation

$$1 - |C(x,\xi)|^2 = Q(x)\,\xi^{2q} + O(\xi^{2q+1}) \tag{4.67}$$

with $q \in \mathbb{N}$ independent of x and $Q(x) > 0$ for all x.

The same result holds for arbitrary (non-periodic) l^2-summable u if the summation in (4.29), (4.62), (4.63) is adapted to an infinite sum.

Proof. Theorem 4.6 is a modification of Theorem 3.1 in [32] by Lax. The Lax theorem handles the higher dimensional case, i.e. the c_j and the C are matrices. When reduced to the scalar case, condition (ii) of the Lax Theorem would become

$$|C(x,\xi)| < 1 \quad \forall\, \xi \neq 0 \bmod 2\pi$$

instead of (4.66) which would be insufficient for our purpose. But condition (ii) of the Lax Theorem is only used to guarantee (3.3) in [32]. However, in the one-dimensional case our weaker condition (4.66) is enough to guarantee (3.3) (with $M \equiv 1$ in [32]). ∎

The formulation of (i) guarantees the exponential decay of the $c_j(x)$ for $|j| \to \infty$ und thus the analyticity of $C(x,\xi)$ with respect to ξ. Therefore the boundedness of the diffence quotients with respect to x not only holds for C, but also for the ξ-derivatives of C. In the case that only a *bounded* number of c_j are non-vanishing (i.e. the difference scheme is explicit), (4.65) can be replaced by

$$\max_{\xi \in [0,2\pi]} |C(x,\xi) - C(x',\xi)| \leq c\,|x - x'| \tag{4.68}$$

or by[10]

$$|c_j(x) - c_j(x')| \leq c\,|x - x'|. \tag{4.69}$$

We will prove that the amplification factors $C(x,\xi)$ related to our interpolation operator $\mathcal{S}_{p,\Delta k}$ fulfil condition (ii) of the previous theorem:

[10] See Theorem 4.8, a new version of the stability theorem 4.6.

Lemma 4.7 *For the following combinations of p and Δk, $1-|C(x_k,\xi)|^2$ has the representation*

p	Δk	$1-\lvert C\rvert^2$
1	1	$(1-cos(\xi))\ \ 2\,\bar{\delta}\,(1-\bar{\delta})$
2	1	$(1-cos(\xi))^2\ \ (1-\bar{\delta})\,\bar{\delta}^2\,(1+\bar{\delta})$
2	2	$(1-cos(\xi))^2\ \ (2-\bar{\delta})\,(1-\bar{\delta})^2\,\bar{\delta}$
3	2	$(1-cos(\xi))^2\ \ \frac{1}{3}\,(2-\bar{\delta})\,(1-\bar{\delta})\,\bar{\delta}\,(1+\bar{\delta})\ +$ $(1-cos(\xi))^3\ \ \frac{2}{9}\,(2-\bar{\delta})\,(1-\bar{\delta})^2\,\bar{\delta}^2\,(1+\bar{\delta})$
4	2	$(1-cos(\xi))^3\ \ \frac{1}{9}\,(2-\bar{\delta})\,(1-\bar{\delta})\,\bar{\delta}^2\,(1+\bar{\delta})\,(2+\bar{\delta})\ +$ $(1-cos(\xi))^4\ \ \frac{1}{36}\,(2-\bar{\delta})\,(1-\bar{\delta})^2\,\bar{\delta}^2\,(1+\bar{\delta})^2\,(2+\bar{\delta})$
4	3	$(1-cos(\xi))^3\ \ \frac{1}{9}\,(3-\bar{\delta})\,(2-\bar{\delta})\,(1-\bar{\delta})^2\,\bar{\delta}\,(1+\bar{\delta})\ +$ $(1-cos(\xi))^4\ \ \frac{1}{36}\,(3-\bar{\delta})\,(2-\bar{\delta})^2\,(1-\bar{\delta})^2\,\bar{\delta}^2\,(1+\bar{\delta})$
5	3	$(1-cos(\xi))^3\ \ (-\frac{1}{450})\,(\bar{\delta}-3)\,(\bar{\delta}-2)\,(\bar{\delta}-1)\,\delta\,(\bar{\delta}+1)\,(\bar{\delta}+2)\ +$ $(1-cos(\xi))^4\ \ \frac{1}{60}\,(\bar{\delta}-3)\,(\bar{\delta}-2)\,(\bar{\delta}-1)^2\,\bar{\delta}^2\,(\bar{\delta}+1)\,(\bar{\delta}+2)\ +$ $(1-cos(\xi))^5\ \ (-\frac{1}{60})\,(\bar{\delta}-3)\,(\bar{\delta}-2)^2\,(\bar{\delta}-1)^2\,\bar{\delta}^2\,(\bar{\delta}+1)^2\,(\bar{\delta}+2)$

Table 4.1: The amplification factor of our scheme $\mathcal{S}_{p,\Delta k}$.

where $\bar{\delta} := \bar{\delta}(x_k) := \delta(x_k)/h - \lfloor \delta(x_k)/h \rfloor \in [0,1)$. So for these combinations of Δk and p, condition (ii) in Theorem 4.6 is met.[11] *Condition (ii) is not met for any other couple $(p,\Delta k)$ with $p \leq 5$.*

Proof. By simple calculus using Lagrangian polynomials (see (4.8), (4.9), (4.59)). ∎

Remark. Condition (iii) is not met for any $p = 2,3,4,5$ and arbitrary Δk, as $\delta(x_k) = 0$ is possible which means that $1 - C(x_k,\cdot) \equiv 0$. To avoid condition (iii), we will derive another, less restrictive version of the stability Theorem 4.6. We will also give a concrete estimate for the constant c in (4.64):

[11] See Figs. 4.6, 4.7

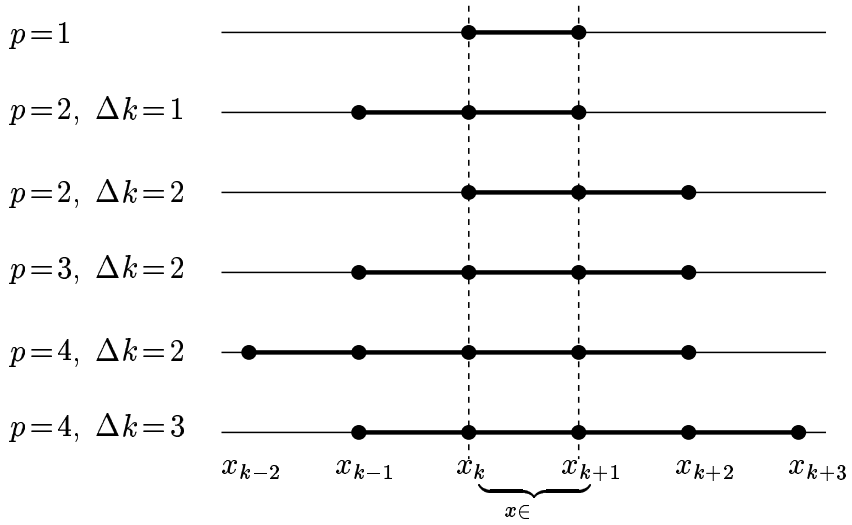


Figure 4.6: Let $x \in [x_k, x_{k+1})$. The figure shows which mesh points $x_{k+\Delta k-p},\ldots,x_{k+\Delta k}$ are to be taken for the interpolation $\mathcal{S}_{p,\Delta k}$, i.e. which combinations of p and Δk ($p \leq 4$) are possible to comply with stability condition (ii) in Theorem 4.6.

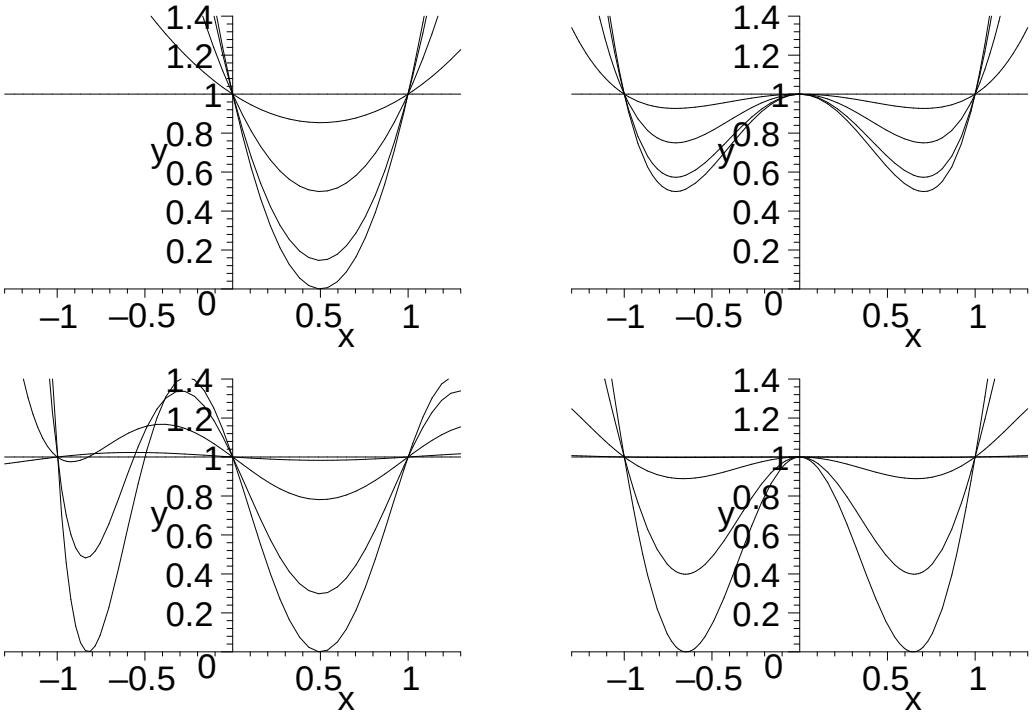


Figure 4.7: Visualization of $|C(x,\xi)|$ which is a function of $\bar{\delta}(x)$ and ξ (see Table 4.1). $\bar{\delta}$ on the horizontal axis. For $p=1$ (upper left), $p=2$ (upper right), $p=3$ (lower left), $p=4$ (lower right), $|C|$ is displayed as a function of $\bar{\delta}$. In each diagram, the curves for $\xi = 0, \pi/4, \pi/2, 3\pi/4, \pi$ are displayed. As $\bar{\delta}(x_k) \in [0,1)\ \forall\ x_k$, the diagrams show that property (ii) is fulfilled.

Theorem 4.8 *Suppose that the following two conditions hold:*

(i)
$$\begin{aligned} |c_j(x) - c_j(x')| &\le L\,|x-x'| && (4.70)\\ c_j(x) &= 0 \qquad \forall\, x\in\Omega_h,\ |j|>M && (4.71)\\ |c_j(x)| &\le C_{11} \qquad \forall\, x\in\Omega_h,\ |j|\le M && (4.72)\end{aligned}$$

(ii)
$$|C(x,\xi)| \le 1 \quad \forall\, x,x'\in\Omega_h,\ \xi\in\mathbb{R}$$

Then (4.64) holds.

Proof. We are following (only) partially the idea of Lax in [32]. From definition (4.63) and condition (ii) we derive that, for arbitrary fixed $x\in\Omega_h$, $1-|C(x,\cdot)|^2$ is an analytic, non-negative function. As a consequence, its Taylor series with respect to ξ starts with an *even* power of ξ, i.e. there is a representation

$$1-|C(x,\xi)|^2 = Q(x)\,(\xi^{2q(x)} + r_x(\xi)) \quad \forall\, x\in\Omega_h,\ \xi\in\mathbb{R} \tag{4.73}$$

with $q(x)\in\mathbb{N}_0$, $Q(x)>0$, $r_x(\xi) = O(\xi^{2q(x)+1})$. Now let us define

$$D(x,\xi) := (e^{i\xi}-1)^{q(x)} \sqrt{\frac{\xi^{2q(x)} + r_x(\xi)}{|e^{i\xi}-1|^{2q(x)}}}\,\sqrt{Q(x)} \quad \forall\, x,\xi\in\mathbb{R}. \tag{4.74}$$

The argument of the first root in (4.74) is an analytic function, and therefore, as $\lim\limits_{\xi\to 0}\sqrt{\frac{\xi^{2q(x)}+r_x(\xi)}{|e^{i\xi}-1|^{2q(x)}}} = 1$, the first root in (4.74) itself is analytic in ξ, too.[12] So D is analytic in ξ. Let[13]

$$\begin{aligned} K(x;\xi,\eta) &:= \bar{C}(x,\xi)\,C(x,\eta),\\ \tilde{K}(x;\xi,\eta) &:= \bar{D}(x,\xi)\,D(x,\eta). \end{aligned} \tag{4.75}$$

It is

$$|D(x,\xi)|^2 = 1-|C(x,\xi)|^2,$$

hence

$$\begin{aligned} K(x;\xi,\xi) + \tilde{K}(x;\xi,\xi) &= 1, && (4.76)\\ \tilde{K}(x;\xi,\xi) &\ge 0. \end{aligned}$$

[12] Here it is essential that the argument of the first root in (4.74) is bounded away from zero to derive the analyticity of the root. To illustrate this point, let us mention that the square root $|\xi|$ of the non-negative $C^\infty(\mathbb{R})$-function $f(\xi)=\xi^2$ is not $C^\infty(\mathbb{R})$.

[13] In [32], $K(x;\xi,\eta) := \bar{C}(x,\xi)C(x,\eta) + \bar{D}(x,\xi)D(x,\eta)$. The regularity requirements on K in [32] then lead to regularity requirements on D; in particular the Lipschitz continuity with respect to x of ξ-derivatives of D. Our definition (4.75) of K avoids any regularity assumptions on D with respect to x and allows $q = q(x)$, especially $1-|C(x,\cdot)|^2 \equiv 0$. See the remark on p. 157.

We get

$$\begin{aligned}
\int_0^{2\pi}\int_0^{2\pi} \bar{U}(\xi)\, K(x;\xi,\eta)\, U(\eta)\, d\xi\, d\eta &= \left| \int_0^{2\pi} C(x,\xi)\, U(\xi)\, d\xi \right|^2, \\
\int_0^{2\pi}\int_0^{2\pi} \bar{U}(\xi)\, \tilde{K}(x;\xi,\eta)\, U(\eta)\, d\xi\, d\eta &= \left| \int_0^{2\pi} D(x,\xi)\, U(\xi)\, d\xi \right|^2
\end{aligned} \tag{4.77}$$

for all $U \in L^2(0,2\pi)$. Let us consider the Fourier expansion of K, $\tilde{K}$:

$$\begin{aligned}
K(x;\xi,\eta) &= \sum_{l,m\in\mathbf{Z}} K_{l,m}(x)\, e^{i(m\eta - l\xi)} \\
\tilde{K}(x;\xi,\eta) &= \sum_{l,m\in\mathbf{Z}} \tilde{K}_{l,m}(x)\, e^{i(m\eta - l\xi)}
\end{aligned} \tag{4.78}$$

From (4.75) and (4.63) we know that

$$K_{l,m}(x) = \bar{c}_l(x)\, c_m(x) \quad \forall\, l,m \in \mathbb{Z}. \tag{4.79}$$

Thus, using (4.70), (4.72), we have

$$|K_{l,m}(x) - K_{l,m}(x')| \le 2\, C_{11} L\, |x - x'|. \tag{4.80}$$

K and $\tilde{K}$ are analytic in ξ, η, therefore their Fourier series may be rearranged. The combination of (4.78) and (4.76) yields

$$\sum_{l,m\in\mathbf{Z}} (K_{l,m}(x) + \tilde{K}_{l,m}(x))\, e^{i(m-l)\xi} = 1. \tag{4.81}$$

A comparison of the Fourier coefficients of both sides of the equation results in

$$\sum_{k\in\mathbf{Z}} K_{k-r,k-s}(x) + \tilde{K}_{k-r,k-s}(x) = \begin{cases} 1, & r = s \\ 0, & r \neq s. \end{cases} \tag{4.82}$$

For every real-valued square-summable sequence (w_j), $U(\xi) := \sum_{j\in\mathbf{Z}} w_j\, e^{-ij\xi}$ is in $L^2(0,2\pi)$. Using this fact and (4.78) on (4.77) we get

$$\begin{aligned}
&\int_0^{2\pi}\int_0^{2\pi} \sum_{j_1\in\mathbf{Z}} w_{j_1}\, e^{ij_1\xi} \sum_{l,m\in\mathbf{Z}} K_{l,m}(x)\, e^{i(m\eta - l\xi)} \sum_{j_2\in\mathbf{Z}} w_{j_2} e^{-ij_2\eta}\, d\xi\, d\eta \\
&\quad = \left| \int_0^{2\pi} \sum_{l\in\mathbf{Z}} c_l(x)\, e^{il\xi} \sum_{j\in\mathbf{Z}} w_j\, e^{-ij\xi} \right|^2
\end{aligned}$$

Using orthogonalities, this equation simplifies to

$$\sum_{l,m\in\mathbf{Z}} w_l\, K_{l,m}(x)\, w_m = |\sum_{j\in\mathbf{Z}} c_j(x)\, w_j|^2. \tag{4.83}$$

Analogously we get

$$\sum_{l,m\in\mathbf{Z}} w_l\, \tilde{K}_{l,m}(x)\, w_m = \left| \int_0^{2\pi} D(x,\xi)\, U(\xi)\, d\xi \right|^2 \geq 0. \tag{4.84}$$

(4.83) reads for $w_j := u(x_{k-j})$

$$|(\mathcal{D}u)(x_k)|^2 = |\sum_{j\in\mathbf{Z}} c_j(x_k)\, u(x_{k-j})|^2 = \sum_{l,m\in\mathbf{Z}} u(x_{k-l})\, K_{l,m}(x_k)\, u(x_{k-m}). \tag{4.85}$$

Summing over k we get

$$\begin{aligned}
\frac{1}{h}\,\|\mathcal{D}u\|_{L_h^2}^2 &= \sum_{k\in\mathbf{Z}} \sum_{l,m\in\mathbf{Z}} u(x_{k-l})\, K_{l,m}(x_k)\, u(x_{k-m}) \\
&= \sum_{k\in\mathbf{Z}} \sum_{l,m\in\mathbf{Z}} u(x_{k-l})\, K_{l,m}(x_{k-l})\, u(x_{k-m}) \\
&\quad + \sum_{k\in\mathbf{Z}} \sum_{l,m\in\mathbf{Z}} u(x_{k-l})\, (K_{l,m}(x_k) - K_{l,m}(x_{k-l}))\, u(x_{k-m}) \\
&= \sum_{k\in\mathbf{Z}} \sum_{r,s\in\mathbf{Z}} u(x_r)\, K_{k-r,k-s}(x_r)\, u(x_s) \\
&\quad + \sum_{k\in\mathbf{Z}} \sum_{r,s\in\mathbf{Z}} u(x_r)\, (K_{k-r,k-s}(x_k) - K_{k-r,k-s}(x_r))\, u(x_s),
\end{aligned} \tag{4.86}$$

where in the last step $r := k-l$, $s := k-m$ was put.

The first summand of (4.86) is due to (4.84) and (4.82) estimated by

$$\sum_{r,s\in\mathbf{Z}} \sum_{k\in\mathbf{Z}} u(x_r)\, (K_{k-r,k-s}(x_r) + \tilde{K}_{k-r,k-s}(x_r))\, u(x_s) = \frac{1}{h}\,\|u\|_{L_h^2}^2.$$

The second summand of (4.86) is estimated by

$$\sum_{k\in\mathbf{Z}} \sum_{r,s\in\mathbf{Z}} \frac{1}{2}\,(u(x_r)^2 + u(x_s)^2)\, |K_{k-r,k-s}(x_k) - K_{k-r,k-s}(x_r)|. \tag{4.87}$$

The terms containing $u(x_r)^2$ are estimated (see (4.80)) by

$$\begin{aligned}
&\frac{1}{2} \sum_{r\in\mathbf{Z}} u(x_r)^2 \sum_{k=r-M}^{r+M} \sum_{s=k-M}^{k+M} 2\, C_{11} L\, |x_k - x_r| \\
&\quad \leq \sum_{r\in\mathbf{Z}} u(x_r)^2 \sum_{k=r-M}^{r+M} (2M+1)\, C_{11} L\, |x_k - x_r|.
\end{aligned}$$

For an equidistant mesh this is equal to

$$\sum_{r\in\mathbf{Z}} u(x_r)^2\, (2M+1)\, (M+1)\, M\, C_{11} L\, h. \tag{4.88}$$

The estimate for the terms in (4.87) containing $u(x_s)^2$ is similar; all in all we get

$$\|Du\|_{L_h^2}^2 \leq (1 + 2\,M\,(M+1)\,(2M+1)\,C_{11}L\,h)\,\|u\|_{L_h^2}^2. \tag{4.89}$$

For periodic u the sums $\sum_{k\in\mathbf{Z}}$ must be replaced by $\sum_{k=0,\ldots,l-1}$. ∎

Applying this theorem to our interpolation scheme (4.53) we get:

Corollary 4.9 *For the scheme (4.53)*

$$\|D_v u\|_{L_h^2} \leq (1 + 2\,M\,(M+1)\,(2M+1)\,C_{11}\tilde{L}\,\Delta t)\,\|u\|_{L_h^2}$$

holds where $\tilde{L} = \alpha\,c(p)\,e^{\alpha\Delta t}$ *from the proof of Lemma 4.5 and* $M = \lceil C_{CFL}\rceil + p$ *and* $C_{11} = C_{max}(p)$ *from (4.28)/Lemma 4.4 (i).*

Proof. Lemma 4.5 and Theorem 4.8. ∎

Cor. 4.9 leads directly to the following improvement of Lemma 4.7, which is in fact the main result of this section on stability for equidistant meshes:

Corollary 4.10 *For all the combinations of p, Δk mentioned in Table 4.1/Lemma 4.7 and bounded Courant number, the stability assumptions of Theorem 4.8 are met, i.e. the convergence result of Theorem 4.2 holds.*

4.4.3 Stability on quasi-uniform grid

In this and in the following section we investigate if the stability results from the equidistant mesh case (Section 4.4.2) can be applied to the non-equidistant case. In the non-equidistant case, the following problems arise:

- The property (4.57) for the Lagrange polynomials and the shape functions gets lost. As a consequence, (4.60) is no longer valid.
- The explicit knowledge of the $C(x,\xi)$ (see Lemma 4.7) gets lost.

To avoid these problems, we introduce new coefficient functions c_j^* which are close approximations of the c_j. It is possible to check all the required conditions for the c_j^*. This will imply the stability of the scheme also for the coefficients c_j.

Let us define the quasi-uniformity of the given mesh at first.

We assume that for a sequence of meshes $(\mathcal{M})$ the grid points $x_k = x_k(\mathcal{M})$ are given by

$$x_k := x_{k,\mathcal{M}} := \omega(kL/N), \quad k = 0, \ldots, N.$$

ω is a function $\omega : [0, L] \to [0, L]$ independent of $\mathcal{M}$ with Lipschitz continuous derivative ω',

$$\omega' \geq C_{\omega 1} > 0, \quad \left|\frac{\omega'(x) - \omega'(x')}{x - x'}\right| \leq C_{\omega 2}. \tag{4.90}$$

We will call such a sequence of meshes fulfilling (4.90) *quasi-uniform*. Let $N+1 = N(\mathcal{M})+1$ be the number of mesh points of the mesh $\mathcal{M}$ and let us denote the average mesh size $\bar{h} := L/N$. The *local* mesh size is given by ω':

$$h_k = x_{k+1} - x_k = \omega((k+1)\bar{h}) - \omega(k\bar{h}) = \omega'(\xi)\,\bar{h} \approx \omega'(k\bar{h})\,\bar{h}$$

Obviously, the following very important estimates follow:

$$h_{min} \geq C_{\omega 1}\,\bar{h}, \quad h_{max} \leq C_{\omega 3}\,\bar{h}, \tag{4.91}$$

$C_{\omega 3} := \max \omega'$. We may define the Courant number by

$$C_{CFL} := \max_k \frac{1}{\bar{h}} |\omega^{-1}(x_k + \delta(x_k)) - \omega^{-1}(x_k)|. \tag{4.92}$$

This definition is consistent with definition (4.50) for the equidistant case.

Another, very graphic property of these quasi-uniform meshes is that the ratio of adjoint mesh cells tends to 1 for $\bar{h} \to 0$ as[14]

$$\begin{aligned} \frac{x_{k+i+1} - x_{k+i}}{x_{k+1} - x_k} &= \frac{\omega'(\xi')\bar{h}}{\omega'(\xi)\bar{h}} \leq \frac{\omega'(\xi) + C_{\omega 2}\,(i+1)\,\bar{h}}{\omega'(\xi)} \\ &\leq 1 + \frac{C_{\omega 2}}{C_{\omega 1}}\,(i+1)\,\bar{h} \end{aligned} \tag{4.93}$$

As a consequence of (4.93), the shape of ansatz functions tends to the shape of equidistant mesh ansatz functions for $\bar{h} \to 0$. This important property is investigated in the following lemma which proves that the 'error' between two Lagrangian polynomials can be estimated by a constant depending *linearly* on the difference of the affiliated mesh points:

Lemma 4.11 *Let $h > 0$.*

(i) Let us consider the Lagrangian polynomials with respect to the grid points $0, h, 2h, ..., ph$ rsp. $\epsilon_0, h+\epsilon_1, ..., ph+\epsilon_p$, $\epsilon_i \in \mathbb{R}$,

$$P_j^0(x) = \prod_{\substack{k=0 \\ k \neq j}}^{p} \frac{x - kh}{jh - kh}, \quad P_j(x) = \prod_{\substack{k=0 \\ k \neq j}}^{p} \frac{x - kh - \epsilon_k}{jh + \epsilon_j - kh - \epsilon_k}$$

with $\epsilon_{max} := \max_k |\epsilon_k| \leq C_{12}\,\frac{h}{2}$, $C_{12} < 1$. Then[15]

$$|P_j^0(x) - P_j(x)| \leq C_{13}\,\frac{\epsilon_{max}}{h} \quad \textit{for all } x \in [\epsilon_0, hp + \epsilon_p] \tag{4.94}$$

with $C_{13} = C_{13}(p, C_{12})$ independent of h, ϵ_{max}.

[14] This is untrue for Gauss-Lobatto meshes

[15] From the fact that the minimum and the maximum of $P_j(x)$ depend continuously on the $p+1$ mesh points we can conclude that $|P_j^0(x) - P_j(x)| \to 0$ uniformly for $\epsilon_{max} \to 0$, h fixed. The estimates (4.94) and (4.95), however, require a closer investigation.

(ii) Let α_k^0 rsp. α_k, $k=0,...,p$, be the ansatz functions (see (4.7)/(4.9)) with respect to the grid points $(ih)_{i\in\mathbf{Z}}$ rsp. $(ih+\epsilon_i)_{i\in\mathbf{Z}}$, $\epsilon_i\in\mathbb{R}$. Again, an estimate

$$|\alpha_k^0(x)-\alpha_k(x)| \le C_{14}\,\frac{\epsilon_{max}}{h} \quad \textit{for all } x\in[\epsilon_0, hp+\epsilon_p] \tag{4.95}$$

holds with $C_{14}=C_{14}(p,C_{12})$ independent of h, ϵ_{max}.

Proof.
ad (i). Let us consider the expansion

$$P_j(x)-P_j^0(x) = \sum_{\substack{k=0\\k\neq j}}^{p}\left(\prod_{\substack{i=0\\i\neq j}}^{k-1}\frac{x-ih}{jh-ih}\;\Delta_k^j(x)\prod_{\substack{i=k+1\\i\neq j}}^{p}\frac{x-ih-\epsilon_i}{jh+\epsilon_j-ih-\epsilon_i}\right) \tag{4.96}$$

where

$$\Delta_k^j(x) = \frac{x-kh-\epsilon_k}{jh+\epsilon_j-kh-\epsilon_k}-\frac{x-kh}{jh-kh}.$$

The last line evaluates to

$$\Delta_k^j(x) = \frac{x\,(\epsilon_k-\epsilon_j)+kh\,\epsilon_j-jh\,\epsilon_k}{(jh-kh)\,(jh+\epsilon_j-kh-\epsilon_k)}$$

Thus,

$$|\Delta_k^j(x)| \le \frac{(4ph+2\epsilon_{max})\,\epsilon_{max}}{h\,(h-2\epsilon_{max})}.$$

The other factors from (4.96) can be estimated by p rsp. $\frac{ph+2\epsilon_{max}}{h-2\epsilon_{max}}$, each. So (4.96) is estimated by

$$p\left(\frac{ph+2\epsilon_{max}}{h-2\epsilon_{max}}\right)^{p-1}\frac{4ph+2\epsilon_{max}}{h-2\epsilon_{max}}\,\frac{\epsilon_{max}}{h} \le p\left(\frac{p+C_{12}}{1-C_{12}}\right)^{p-1}\frac{4p+C_{12}}{1-C_{12}}\,\frac{\epsilon_{max}}{h}.$$

ad (ii). The interval $[\epsilon_0, ph+\epsilon_p]$ decomposes into the sets

$$\begin{aligned} M_1 &:= \{x\,|\,\exists\, j : jh+\epsilon_j \le x \le (j+1)h+\epsilon_{j+1} \wedge jh \le x \le (j+1)h\} \quad \text{and}\\ M_2 &:= \{x\,|\,\exists\, j : jh < x < jh+\epsilon_j \vee jh+\epsilon_j < x < jh\} \end{aligned}$$

(see left part of Fig. 4.8 for M_1, M_2). Let $x\in M_1$. Then the estimate follows directly from (i):

$$|\alpha_k^0(x)-\alpha_k(x)| = |L^{k,0}_{j+\Delta k-p,j+\Delta k}(x)-L^{k}_{j+\Delta k-p,j+\Delta k}(x)| \le C_{13}\,\frac{\epsilon_{max}}{h}$$

($L^{k,0}$ is defined as L^k but with respect to equidistant grid points).

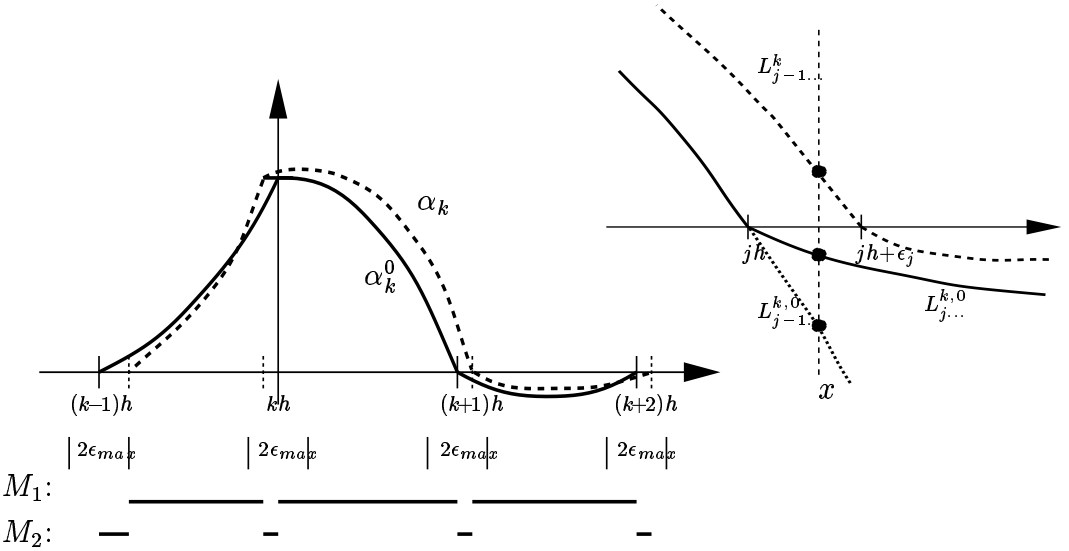


Figure 4.8: Left hand part: Visualization of two ansatz functions α_k, α_k^0; the first related to a non-equidistant mesh (broken line), the latter related to an equidistant mesh (full line). Right hand part: Visualization of estimate (4.97).

Now let $x \in M_2$. Let $jh < x < jh + \epsilon_j$. (The case $jh + \epsilon_j < x < jh$ is analogous.) Then[16]

$$\begin{aligned} |\alpha_k^0(x) - \alpha_k(x)| &= |L^{k,0}_{j+\Delta k-p,j+\Delta k}(x) - L^k_{j-1+\Delta k-p,j-1+\Delta k}(x)| \\ &\leq |L^{k,0}_{j+\Delta k-p,j+\Delta k}(x) - L^{k,0}_{j-1+\Delta k-p,j-1+\Delta k}(x)| \\ &\quad + |L^{k,0}_{j-1+\Delta k-p,j-1+\Delta k}(x) - L^k_{j-1+\Delta k-p,j-1+\Delta k}(x)|. \end{aligned} \tag{4.97}$$

The second difference is estimated as in the case $x \in M_1$. The first difference is estimated by

$$\begin{aligned} &|L^{k,0}_{j+\Delta k-p,j+\Delta k}(x) - L^{k,0}_{j-1+\Delta k-p,j-1+\Delta k}(x)| \\ &\quad \leq |L^{k,0}_{j+\Delta k-p,j+\Delta k}(jh) - L^{k,0}_{j-1+\Delta k-p,j-1+\Delta k}(jh)| \\ &\qquad + L(L^{k,0}_{j+\Delta k-p,j+\Delta k})\,\epsilon_{max} + L(L^{k,0}_{j-1+\Delta k-p,j-1+\Delta k})\,\epsilon_{max} \end{aligned}$$

where L denotes the Lipschitz constant of the polynomials. Due to the equidistance of the related mesh, both Lipschitz constants are $\frac{c(p)}{h}$, further we know that $L^{k,0}_{j+\Delta k-p,j+\Delta k}(jh) = \delta_{kj} = L^{k,0}_{j-1+\Delta k-p,j-1+\Delta k}(jh)$. (4.95) follows. ∎

Let $\alpha_j^k(x)$ be the piecewise polynomial shape functions with respect to the equidistant mesh $x_k \pm ih_k$, $i \in \mathbb{Z}$. Analogously to the $c_j(x_k) = \alpha_{k-j}(x_k + \delta(x_k))$ let

$$c_j^*(x_k) := \alpha^k_{k-j}(x_k + \delta(x_k)) \tag{4.98}$$

[16] See right hand part of Fig. 4.8.

be the corresponding coefficient functions,

$$D_v^* u(x_k) =:= \sum_j c_j^*(x_k)\, u(x_{k-j})$$

the corresponding scheme and

$$C^*(x,\xi) := \sum_j c_j^*(x)\, e^{ij\xi}$$

the corresponding amplification factor.

We show

- that the c_j^* are approximations of the c_j (Lemma 4.12) and
- that the difference scheme with the coefficient functions c_j^* instead of the c_j is stable (Lemma 4.13).

This will imply the stability of the original scheme with the coefficient functions c_j (Theorem 4.14).

Lemma 4.12 (approximation property) *There is a constant such that the estimate*

$$|c_j^*(x_k) - c_j(x_k)| \le c\,\bar{h}$$

holds for all j, k.

Proof. From definition (4.98) and Lemma 4.11 (ii) we get

$$|c_j^*(x_k) - c_j(x_k)| \le C_{14}\,\frac{\epsilon_{max}}{h_k},$$

where $\epsilon_{max} = \max_{|j|\le M} |x_{k-j} - x_k + jh_k|$, $M = \lceil C_{CFL}\rceil + p$. A Taylor expansion of $x_{k-j} = \omega((k-j)\bar{h})$ and the mean value theorem for $h_k = \omega((k+1)\bar{h}) - \omega(k\bar{h})$ gives

$$|x_{k-j} - x_k + jh_k| = |-\omega'(\xi)j\bar{h} + \omega'(\xi')j\bar{h}| \le C_{\omega 2}(|j|+1)\,|j|\bar{h}^2.$$

So, using (4.91),

$$|c_j^*(x_k) - c_j(x_k)| \le C_{14}C_{\omega 2}\,M(M+1)\,\frac{\bar{h}^2}{h_k} = \frac{C_{14}C_{\omega 2}M(M+1)}{C_{\omega 1}}\,\bar{h}. \qquad \blacksquare$$

Lemma 4.13 (stability of c_j^*-scheme) *The difference scheme with the coefficient functions c_j replaced by c_j^* fulfils the requirements of Theorem 4.8 with*

$$L \le c\,\frac{\Delta t}{\bar{h}}. \tag{4.99}$$

Proof. Due to the equidistance of the mesh which is used for the definition of the c_j^*, properties (4.71), (4.72) and (ii) of Theorem 4.8 are obviously met (see Sec. 4.4.2 for the equidistant case). The Lipschitz continuity (4.70) follows like this: For $k \neq k'$,

$$\begin{aligned}
|c_j^*(x_k) - c_j^*(x_{k'})| &= |\alpha_{k-j}^k(x_k+\delta(x_k)) - \alpha_{k'-j}^{k'}(x_{k'}+\delta(x_{k'}))| \\
&= |\alpha_{k-j}^k(x_k+\delta(x_k)) - \alpha_{k-j}^k(x_k+\delta(x_{k'})\,\frac{h_k}{h_{k'}})| \\
&\leq L(\alpha_{k-j}^k)\;|\delta(x_k) - \delta(x_{k'})\,\frac{h_k}{h_{k'}}| \\
&\leq L(\alpha_{k-j}^k)\;\left(L(\delta)\,|x_k - x_{k'}| + \max_{x\in\Omega_h}|\delta(x)|\;\left(1-\frac{h_k}{h_{k'}}\right)\right) \\
&\leq c h_k^{-1}\;\left(c\,\Delta t\,|x_k - x_{k'}| + c\,\Delta t\;(|k-k'|+1)\,\bar{h}\right)
\end{aligned}$$

where the equidistance of the meshes was used in the second and (4.61), (4.93) were used in the last step. (4.70) with (4.99) follows with help of the estimate

$$|k\bar{h} - k'\bar{h}| \leq C_{\omega 1}^{-1}\,|\omega(k\bar{h}) - \omega(k'\bar{h})| = C_{\omega 1}^{-1}\,|x_k - x_{k'}|.$$ ∎

Theorem 4.14 *If*

$$\bar{h} \leq c\,\Delta t \tag{4.100}$$

is respected, the difference scheme with the coefficient functions c_j is stable in the sense (4.64) for all combinations of Δk, p given in Table 4.1/Lemma 4.7. Under these conditions, the Convergence Theorem 4.2 holds.

Proof. Due to Lemma 4.13, the scheme D_v^* with the coefficient functions c_j^* meets (4.51). By Lemma 4.12 and (4.100) we see that the scheme D_v with the coefficients c_j fulfils (4.51), too. ∎

Remark. Of course, the convergence result also holds in any equivalent norm. Due to the quasi-uniformity of the mesh, any discretization of the $L^2(\Omega)$-norm, e.g. (4.29) with $w_{k,h} = \omega(k\bar{h})\bar{h}$, is an equivalent norm.

4.4.4 No stability on Gauss-Lobatto grid

Many discretizations do not fulfil the condition $\omega' \geq \text{const} > 0$ rsp. $h_{min} \geq \text{const}\,\bar{h}$ which was made in the previous section, e.g. when adaptive remeshing is used or when a Chebyshev-Gauss-Lobatto mesh is used; in the latter case we have

$$\omega_{GL}(x) = \frac{L}{2}\,\left(1 - \cos\frac{\pi x}{L}\right) \tag{4.101}$$

and $\omega'_{GL}(0) = \omega'_{GL}(L) = 0$. The question arises if the stability conditions of the previous section, especially the regularity of the coefficient functions c_j (4.70), are *necessary* to derive (4.52). In the following we will show that for a Gauss-Lobatto mesh and $p=2$ the stability statement (4.52) is *not* met even if arbitrary weighted L^q-norms are used. We will use a grid point distribution function ω which is $\omega(x)=x^2$ for small x. The result also applies for the Gauss-Lobatto case (4.101) as $\omega_{GL}(x)=cx^2+O(x^4)$ for small x.

Lemma 4.15 (instability) *Let $\Omega=(0,1)$, $\omega(x)=x^2$ on $(0,c)$, $c>0$, i.e. $x_k = k^2/N^2$ for k small enough. Let $N=N(\Delta t)$ be given for each $\Delta t>0$ with $N(\Delta t) \to \infty$ for $\Delta t \to 0$. We assume that the velocity field for the calculation of the characteristics is $v \equiv -1$, that means $\delta(x_k) = \Delta t$ for all k. Let the Courant number (4.92) be bounded by 1. Let $1 \le s \le \infty$ and $w : [0,1] \to \mathbb{R}$ be an arbitrary weight function with $w(x) > 0 \, \forall \, x \in (0,1)$. Let the order of the interpolation polynomials be $p=2$ and $\Delta k=1$.*[17]

Then there is no constant $c>0$ such that

$$\|D_v u\|_{s,w} \le (1+c\,\Delta t)\, \|u\|_{s,w} \tag{4.102}$$

holds for all discretely given $u : \Omega_h \to \mathbb{R}$. Here,

$$\begin{aligned} \|r\|_{s,w} &:= \left(N^{-1} \sum_{k=0}^{N} w(x_k)\, |r(x_k)|^s \right)^{1/s}, \quad 1 \le s < \infty, \\ \|r\|_{\infty,w} &:= \max_{k=0,\ldots,N} w(x_k)\, |r(x_k)|. \end{aligned}$$

Proof. As $v=\text{const}$ the maximum in (4.92) is taken for $k=0$. So $1 \ge C_{CFL} = N\omega^{-1}(\delta(0)) = N\omega^{-1}(\Delta t) = N\sqrt{\Delta t}$. Thus,

$$N^2 \Delta t \le 1. \tag{4.103}$$

We will show that (4.102) is false for u defined by $u(x_2)=1$, $u(x_k)=0$ otherwise. As $C_{CFL} \le 1$, $(D_v u)(x_2) = (\mathcal{S}u)(x_2+\delta(x_2)) = P(x_2+\delta(x_2))$ where P is the Lagrangian polynomial $L^2_{1,3}$ (see Fig. 4.9).

It is

$$P(x) = \frac{(x-x_1)(x-x_3)}{(x_2-x_1)(x_2-x_3)},$$

so

$$P(x_2) = 1, \quad P'(x_2) = \frac{2}{15}\, N^2, \quad P''(x_2) = -\frac{2}{15}\, N^4.$$

[17] For $\Delta k=2$ a similar situation for instability can be found.

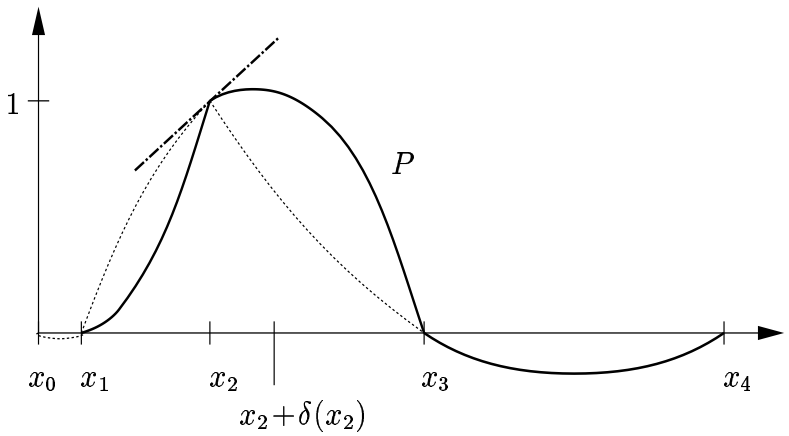


Figure 4.9: The interpolation curve $\mathcal{S}u$ for the Gauss-Lobatto mesh. The ratio of adjoint mesh cells does not tend to 1 for $N \to \infty$ (see the remarks on (4.93) in Section 4.4.3), i.e. the ansatz function P has a non-horizontal tangent at $(x_2, 1)$ even in the limit $N \to \infty$. This property is used in the proof.

Therefore

$$\begin{aligned} P(x_2+\delta(x_2)) &= P(x_2+\Delta t) = 1 + \frac{1}{15} N^2 \, \Delta t \, (2 - N^2 \Delta t) \\ &\geq 1 + \frac{1}{15} N^2 \Delta t \end{aligned}$$

because of (4.103). Together with

$$\begin{aligned} \|u\|_{s,w}^s &= N^{-1} w(x_2), \\ \|D_v u\|_{s,w}^s &\geq N^{-1} w(x_2) \, |(D_v u)(x_2)|^s \end{aligned}$$

$(1 \leq s < \infty)$ we have

$$\frac{\|D_v u\|_{s,w}}{\|u\|_{s,w}} \geq |(D_v u)(x_2)| \geq 1 + \frac{1}{15} N^2 \Delta t$$

even for all $1 \leq s \leq \infty$. The contradiction to (4.102) follows as $N(\Delta t) \to \infty$ for $\Delta t \to 0$. ∎

4.4.5 A 'stable' scheme on Gauss-Lobatto grids

In the last section we have seen that there is no stability for our scheme if a Chebyshev-Gauss-Lobatto mesh is used. The source of this instability on meshes which do not fulfil (4.90) seems to be (see proof of Lemma 4.15 and Fig. 4.9) that the shape functions (for $p = 2$) have a nonzero derivative at the related mesh points. In this section we will introduce and investigate a modified scheme

not having this defect. This scheme is based on local transformation onto an equidistant mesh.

Another point is the norm in which the stability is investigated. The norm (4.29) with $w_{k,h} = \bar{h}$ on the one hand and e.g. $w_{k,h} = \bar{h}\,\omega(k\bar{h})$ (which is a discretization of the $L^2(\Omega)$-norm) on the other hand cannot be estimated mutually with constants independent of $\bar{h}$ any more. It is obvious that L^2_h with $w_{k,h}=\bar{h}$ is the wrong norm to expect stability on Gauss-Lobatto grids: Even for $v \equiv 1$ it is easy to find examples where the exact flow u (restricted to the mesh) is instable in this norm whereas it is stable in the L^2-norm.

Let us assume the following requirements on the grid point distribution function ω: Let $0=r_0<r_1<\ldots<r_m=L$, $m\geq 1$ be a number of points, fixed for all considered discretizations.[18] Let us assume that ω is piecewise C^2 with respect to the r_i, i.e. $\omega \in C^1([0,L])$, $\omega|_{[r_{i-1},r_i]}$ is a C^2 function (i.e. the first and second derivatives from the left and from the right exist at the points r_i). Let us assume

$$\omega' \leq c, \quad \omega'' \leq c, \tag{4.104}$$

$$x_{k+1} - x_k = \omega((k+1)\bar{h}) - \omega(k\bar{h}) \geq c\,\bar{h}^2 \quad \forall\, k \tag{4.105}$$

holds.

We may remark that (4.104)-(4.105) are weaker than the assumptions on ω in Section 4.4.3 in the sense that we do *not* assume that $\omega' \geq c > 0$; in the case of a Gauss-Lobatto mesh (or a stringing together of Gauss-Lobatto domains) the conditions (4.104)-(4.105) are met.

As we are expecting a strong stability restriction on Δt, we may restrict ourself to the case that the local Courant number is bounded by 1:

$$C_{CFL} \leq 1 \tag{4.106}$$

Furthermore, we restrict ourself to second order ansatz functions ($p=2$).

Let us introduce the new interpolation scheme at first. To avoid the problem mentioned above we will use a mapping of the non-equidistantly meshed domain onto an equidistantly meshed domain. Then, we will perform the interpolation on the equidistant mesh. Let $\omega_{k,h}^{-1}$ be that mapping where $\omega_{k,h}$ denotes a smooth function defined on the interval $[(k-1)\bar{h},(k+1)\bar{h}] =: [\bar{x}_{k-1},\bar{x}_{k+1}]$ with values in $[x_{k-1},x_{k+1}]=[\omega((k-1)\bar{h}),\omega((k+1)\bar{h})]$, with

$$\omega_{k,h}(j\bar{h}) \;=\; \omega(j\bar{h}) \quad \text{for } j=k-1,k,k+1. \tag{4.107}$$

So our mapping $\omega_{k,h}$ is locally defined on the two adjacent mesh cells of the grid point x_k and it may depend on the discretization parameter $\bar{h}$. Furthermore, we assume that

$$\omega'_{k,h} > c\,\bar{h} \quad \text{at least on } [k\bar{h}-c\bar{h}, k\bar{h}+c\bar{h}] \tag{4.108}$$

[18] In case of a multi-domain approach, the r_i are the interfaces between the subdomains.

and that the properties

$$\omega'_{k,h} \le c, \quad \omega''_{k,h} \le c, \quad \omega'''_{k,h} \le c \tag{4.109}$$

hold for $w_{k,h}$ on $[(k-c)\bar{h}, (k+c)\bar{h}]$. All the c are generic and independent of $\bar{h}$, k. The existence of such a $\omega_{k,h}$ is checked in the next lemma.

The new interpolation operator $\mathcal{S}^* = \mathcal{S}^*_h$ is then defined by

$$\begin{aligned} \mathcal{S}^* u(x) &:= (\mathcal{S}(u \circ \omega_{k,h}))\,(\omega_{k,h}^{-1}(x)) \\ &= (\mathcal{S}(u \circ \omega))\,(\omega_{k,h}^{-1}(x)) \quad \text{on } [(k-c)\bar{h}, (k+c)\bar{h}] \end{aligned} \tag{4.110}$$

where $\mathcal{S} = \mathcal{S}_{2,1}$ denotes the interpolation operator defined in Section 4.2 with respect to the equidistant grid $\bar{x}_k = k\bar{h}$.[19] So instead of interpolation e.g. at a point $x = x_k + \delta(x_k)$ we interpolate at $\bar{x} = \omega_{k,h}^{-1}(x_k + \delta(x_k))$ with respect to the equidistant grid, see Fig. 4.10. In the transformed space the length of the characteristics is

$$\begin{aligned} \bar{\delta}(\bar{x}_k) &:= \omega_{k,h}^{-1}(x_k + \delta(x_k)) - \omega_{k,h}^{-1}(x_k) = \omega_{k,h}^{-1}(\omega_{k,h}(\bar{x}_k) + \delta(\omega_{k,h}(\bar{x}_k))) - \bar{x}_k, \\ \bar{x}_k &:= \omega_{k,h}^{-1}(x_k). \end{aligned} \tag{4.111}$$

This length can be used to define a local and a global Courant number:

$$C_{CFL}(x_k) := |\bar{\delta}(x_k)|, \quad C_{CFL} := \max_k \; C_{CFL}(x_k) \tag{4.112}$$

The reasons for using $\omega_{k,h}^{-1}$ instead of ω^{-1} for the local mapping and the definition (4.110) of the scheme are enumerated after Theorem 4.17.

The stability condition (4.52) has to be replaced by

$$\|D_v^* u\|_{L_h^2} \le (1 + C_{stab}\,\Delta t)\,\|u\|_{L_h^2}, \tag{4.113}$$

$$(D_v^* u)(x_k) := (\mathcal{S}^* u)(x_k + \delta(x_k)) \tag{4.114}$$

$$=: \sum_{j=-1}^{1} c_j^*(x_k)\, u(x_{k-j}). \tag{4.115}$$

Concerning the norm (4.29) we will use the weight function $w_{k,h} : \bar{h}\,\omega'_{k,h}(k\bar{h})$.

Lemma 4.16 (existence of $\omega_{k,h}$) *Under the assumptions (4.104), (4.105) on ω there is a function $\omega_{k,h}$ fulfilling the requirements (4.107), (4.108), (4.109). Herein, the constants c do not depend on k, h. As a consequence, $\omega_{k,h}$ is monotoneous on $[k\bar{h} - c\bar{h}, k\bar{h} + c\bar{h}]$, $c > 0$, and $w_{k,h}^{-1}$ exists locally.*

[19] The equivalence in (4.110) is a consequence of (4.107).

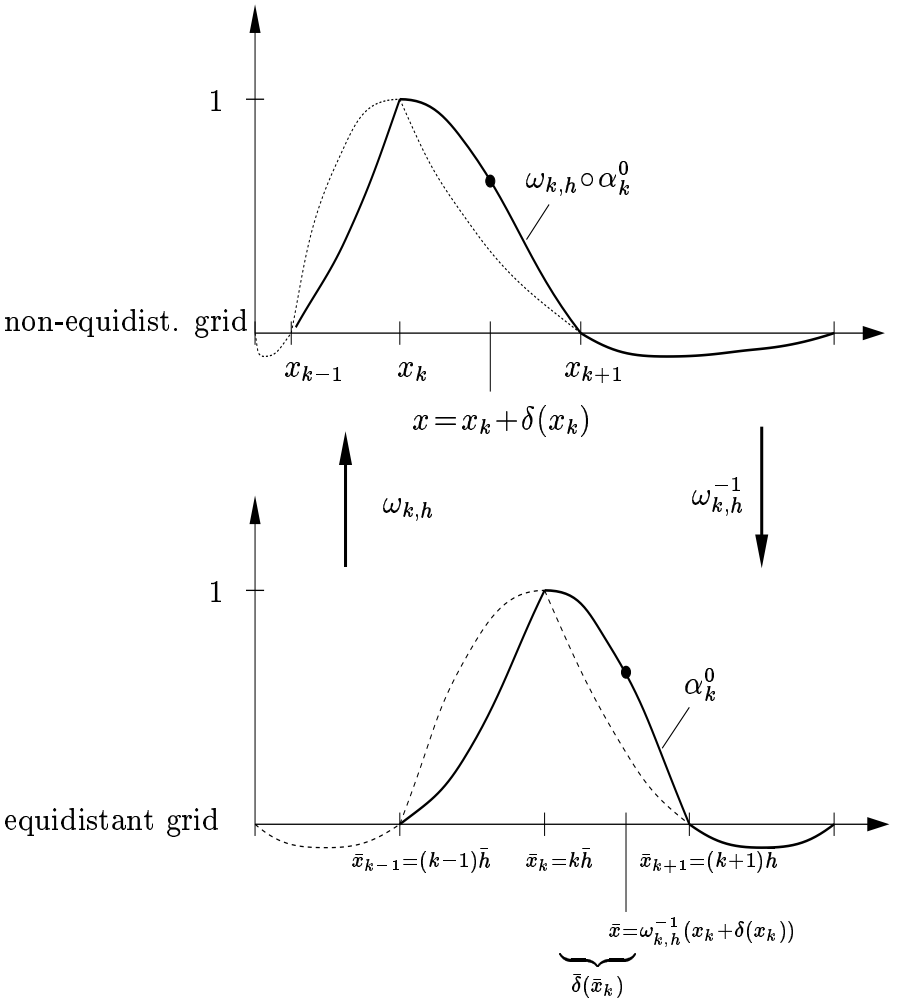


Figure 4.10: In the lower part the piecewise Lagrangian ansatz function α_k^0 in the transformed space, in the upper part the related ansatz function $\alpha_k^* = \omega_{k,h}\circ\alpha_k^0$. The problem on the non-vanishing gradient of the ansatz functions at the related mesh points (see Fig. 4.9) disappears for $\omega_{k,h}\circ\alpha_k^0$.

Proof. Let $w_{k,h}$ be the second order polynomial matching (4.107). It is

$$\omega'_{k,h}(\bar{x}_k) = \frac{\omega_{k,h}(\bar{x}_k+\bar{h}) - \omega_{k,h}(\bar{x}_k-\bar{h})}{2\,\bar{h}} = \frac{\omega(\bar{x}_k+\bar{h}) - \omega(\bar{x}_k-\bar{h})}{2\,\bar{h}}, \tag{4.116}$$

$$\begin{aligned} \omega''_{k,h}(\bar{x}) &= \omega''_{k,h}(\bar{x}_k) = \frac{\omega_{k,h}(\bar{x}_k+\bar{h}) - 2\omega_{k,h}(\bar{x}_k) + \omega_{k,h}(\bar{x}_k-\bar{h})}{\bar{h}^2} \\ &= \frac{\omega(\bar{x}_k+\bar{h}) - 2\omega(\bar{x}_k) + \omega(\bar{x}_k-\bar{h})}{\bar{h}^2} \quad \forall\,\bar{x}, \end{aligned} \tag{4.117}$$

$$\omega'''_{k,h} \equiv 0.$$

From (4.105) and (4.116) we derive that $\omega'_{k,h}(\bar{x}_k) \geq c\bar{h}$. Taylor expansions of ω around $\bar{x}_k$ in the right hand sides of (4.116), (4.117) yield $\omega'_{k,h}(\bar{x}_k) = \frac{1}{2}(\omega'(\bar{\xi}_1)+$

$\omega'(\bar{\xi}_2))$, $\omega''_{k,h}(\bar{x}) = \omega''(\bar{\eta}_1)+\omega''(\bar{\eta}_2)$. Thus, $\omega'_{k,h}(\bar{x}_k) \le c$, $|\omega''_{k,h}(\bar{x})| \le c$. Hence

$$\begin{aligned} |\omega'_{k,h}(\bar{x})| &= |\omega'(\bar{x}_k) + (\bar{x}-\bar{x}_k)\,\omega''(\bar{\xi})| \\ &\ge c\bar{h} - |\bar{x}-\bar{x}_k|\,c \ge c\bar{h} \end{aligned}$$

and

$$\le c\bar{h} + c\,|\bar{x}-\bar{x}_k| \le c\bar{h}$$

for $|\bar{x}-\bar{x}_k| \le c\bar{h}$, for a $c>0$. All c are generic and independent of k, h. ∎

Theorem 4.17 (convergence) *The interpolation operator $\mathcal{S}^* = \mathcal{S}^*_{k,h}$ from (4.110) is of order 3 in $\bar{h}$ in a neighbourhood of x_k if $u \in C^3$.*

Proof. Setting $\bar{x} = w^{-1}_{k,h}(x)$ we have

$$(\mathcal{S}^* u)(x) - u(x) = (\mathcal{S}(u \circ w_{k,h}))(\bar{x}) - u \circ w_{k,h}(\bar{x}).$$

Thus,

$$|(\mathcal{S}^* u)(x) - u(x)| \le c\,\bar{h}^3 \sup_{\xi} |(u \circ w_{k,h})'''(\bar{\xi})|.$$

$\omega_{k,h}$, $\omega'_{k,h}$, $\omega''_{k,h}$, $\omega'''_{k,h}$ are bounded. ∎

The reasons to introduce $\omega_{k,h}$ instead of using ω in (4.110) are:

- Property (4.108) is required in the stability investigation. For a Gauss-Lobatto mesh, ω does not meet this condition.
- ω may not be given explicitely, or it may be more difficult to evaluate ω^{-1} than $\omega^{-1}_{k,h}$, as $\omega_{k,h}$ can be chosen as a second order polynomial (Lemma 4.16).
- The use of ω instead of $\omega_{k,h}$ in the definition (4.110) would lead to a loss of convergence order (see proof of Theorem 4.17) because ω is less regular at $r_0,\ldots,r_m$.

Lemma 4.18, 4.19 and Theorem 4.20 deal with the stability of the scheme D^*_v and therefore check some conditions of the coefficient functions c^*_j of this scheme. The new estimate (4.120) takes the role of the 'Lipschitz' condition (4.70). Unfortunately, the stability proof reqires a severe stability restraint (4.119) which apparenty cannot be weakened (see Lemma 4.21).

Lemma 4.18 *Suppose the assumptions (4.104)-(4.109) hold. Let D^*_v, $c^*_j(\bar{x})$ be defined in (4.114), (4.115). Then, (4.71), (4.72), (ii), (iii') in Theorem 4.8 hold with c_j replaced by c^*_j.*

Proof. (4.71) with $M=1$ is a consequence of the definition of the interpolation operator $\mathcal{S}^*$, $\mathcal{S}=\mathcal{S}_{2,1}$ and (4.106). Now (4.72), (ii), (iii'): Similar to (4.56)-(4.58) we get the representation

$$\begin{aligned} D_v^* u(x_k) &= \mathcal{S}^* u(x_k+\delta(x_k)) = \mathcal{S}_{2,1}(u\circ\omega_{k,h})(\omega_{k,h}^{-1}(x_k+\delta(x_k))) \\ &= \sum_{j=k-1}^{k+1} u(x_j)\ \alpha_j^0(\omega_{k,h}^{-1}(x_k+\delta(x_k))) \\ &= \sum_{j=-1}^{1} u(x_{k-j})\ L_{k-1,k+1}^{k-j,0}(\omega_{k,h}^{-1}(x_k+\delta(x_k))) \end{aligned}$$

where the α_j^0 are the piecewise Lagrangian ansatz functions $L_{a,b}^{j,0}$ which are defined with respect to the equidistant grid $(j\bar{h})_{j\in\mathbf{Z}}$. So the coefficients in the scheme D_v^* are

$$\begin{aligned} c_j^*(x_k) &= L_{k-1,k+1}^{k-j,0}(\omega_{k,h}^{-1}(x_k+\delta(x_k))) = L_{-1,+1}^{-j,0}(\omega_{k,h}^{-1}(x_k+\delta(x_k)) - k\,\bar{h}) \\ &= L_{-1,+1}^{-j,0}\circ\bar{\delta}(\bar{x}_k) \end{aligned} \tag{4.118}$$

where (4.111) was used. Using Lemma 4.4 (i) and Table 4.1, we get (4.72), (ii), (iii'). ∎

Lemma 4.19 *Suppose the assumptions of the previous lemma and the stability restraint*

$$\Delta t \le \bar{h}^4 \tag{4.119}$$

hold. Let the Courant number (4.112) be so small that $w_{k,h}^{-1}\circ\omega$ exists on $[(k-C_{CFL})\bar{h},(k+C_{CFL})\bar{h}]$ (see 4.108), especially $C_{CFL}\le 1$. Then the estimate

$$\mathcal{W}_k\ |\omega'_{k,h}(\bar{x}_k)\,c_i^*(x_k)\,c_j^*(x_k) - \omega'_{k+1,h}(\bar{x}_{k+1})\,c_i^*(x_{k+1})\,c_j^*(x_{k+1})| \le c\ \Delta t \tag{4.120}$$

with

$$\mathcal{W}_k := \mathcal{W}_{k,n_1,n_2} := \frac{1}{\sqrt{\omega'_{k+n_1,h}(\bar{x}_{k+n_1})\ \omega'_{k+n_2,h}(\bar{x}_{k+n_2})}}$$

holds for all $k,i,j\in\mathbf{Z}$, $i^2+j^2\neq 0$, $|n_1|,|n_2|\le 1$, the constant c being independent of Δt, $\bar{h}$.

The rather technical proof of this lemma is given after the following main theorem of this section.

Theorem 4.20 (stability) *Under the assumptions of the previous lemma the interpolation operator D_v^* is stable in the sense (4.113) where the L_h^2-norm is defined using the weight $w_{k,h}:=\omega'_k(\bar{x}_k)\bar{h}$.*

Proof. We are following the proof of Theorem 4.8. The estimate (4.86) is replaced by

$$\begin{aligned}\frac{1}{\bar{h}}\,\|D_v^* u\|_{L_h^2}^2 &= \sum_{k\in\mathbb{Z}}\sum_{l,m\in\mathbb{Z}} u(x_{k-l})\; w_{k,h}\; K_{l,m}(x_k)\; u(x_{k-m}) \\ &= \sum_{k\in\mathbb{Z}}\sum_{l,m\in\mathbb{Z}} u(x_{k-l})\; w_{k-l,h}\; K_{l,m}(x_{k-l})\; u(x_{k-m}) \\ &\quad + \sum_{k\in\mathbb{Z}}\sum_{\substack{l,m\in\mathbb{Z}\\ l\neq 0}} u(x_{k-l})\; (w_{k,h}\, K_{l,m}(x_k) - w_{k-l,h}\, K_{l,m}(x_{k-l}))\; u(x_{k-m}) \\ &= \sum_{k\in\mathbb{Z}}\sum_{r,s\in\mathbb{Z}} u(x_r)\; w_{r,h}\; K_{k-r,k-s}(x_r)\; u(x_s) \qquad (4.121) \\ &\quad + \sum_{k\in\mathbb{Z}}\sum_{\substack{r,s\in\mathbb{Z}\\ r\neq k}} u(x_r)\; (w_{k,h}\, K_{k-r,k-s}(x_k) - w_{r,h}\, K_{k-r,k-s}(x_r))\; u(x_s),\end{aligned}$$

where again $r:=k-l$, $s:=k-m$ was set. Again, the first summand of (4.121) can be estimated by $\|u\|_{L_h^2}^2$. The second summand of (4.121) is estimated by

$$\sum_{k\in\mathbb{Z}}\sum_{\substack{r,s\in\mathbb{Z}\\ r\neq k}} \frac{w_{r,h}u(x_r)^2 + w_{s,h}u(x_s)^2}{2\,\sqrt{w_{r,h}w_{s,h}}}\; |w_{k,h}K_{k-r,k-s}(x_k) - w_{r,h}K_{k-r,k-s}(x_r)|. \qquad (4.122)$$

As the Courant number is bounded by 1, only those indices r, s with $|r-k|=1$, $|s-k|\leq 1$ have to be considered in (4.122). Using (4.120) instead of (4.70) we follow the rest of the proof of Theorem 4.8. ∎

Proof of Lemma 4.19. From (4.118) and Lemma 4.7 we derive that the $c_i^*(x_k)\, c_j^*(x_k)$ are polynomials of order 4 in $\bar{\delta}(\bar{x}_k)/\bar{h}$. Due to the condition $i^2+j^2\neq 0$ and Lemma 4.7 the trailing coefficient of these polynomials vanishes. So it is sufficient to prove

$$\mathcal{W}_k\; \left[\omega'_{k,h}(\bar{x}_k)\,\frac{\bar{\delta}^n(\bar{x}_k)}{\bar{h}^n} - \omega'_{k+1,h}(\bar{x}_{k+1})\,\frac{\bar{\delta}^n(\bar{x}_{k+1})}{\bar{h}^n}\right] \leq c\;\Delta t \qquad (4.123)$$

for $n=1,2,3,4$.

Before, let us state that

$$\frac{\omega'_{k,h}(\bar{x}_k+\epsilon\bar{h})}{\omega'_{k,h}(\bar{x}_k)} = 1 + \frac{\omega''_{k,h}(\bar{x}_k+\xi)\;\epsilon\,\bar{h}}{\omega'_{k,h}(\bar{x}_k)} \leq 1 + c\,\epsilon \qquad (4.124)$$

which is bounded for bounded ϵ.[20] Furthermore, a Taylor expansion of (4.111) with respect to δ leads to

$$\begin{aligned}\bar{\delta}(\bar{x}_k) &= (\omega_{k,h}^{-1})'(\omega_{k,h}(\bar{x}_k))\,\delta(\omega_{k,h}(\bar{x}_k)) + \frac{1}{2}(\omega_{k,h}^{-1})''(\xi)\,\delta^2(\omega_{k,h}(\bar{x}_k)) \\ &= A(\bar{x}_k) + B(\bar{x}_k),\end{aligned}$$

[20] Here we have used (4.108).

where

$$A(\bar{x}_k) := \frac{1}{\omega'_{k,h}(\bar{x}_k)}\,\delta(\omega_{k,h}(\bar{x}_k)), \quad B(\bar{x}_k) = -\frac{\omega''_{k,h}(\omega^{-1}_{k,h}(\xi))}{2\,\omega'_{k,h}(\omega^{-1}_{k,h}(\xi))^3}\,\delta^2(\omega_{k,h}(\bar{x}_k)),$$

$$|A(\bar{x}_k)| \le c\,\frac{\Delta t}{\bar{h}}, \quad |B(\bar{x}_k)| \le c\,\frac{\Delta t^2}{\bar{h}^3} \tag{4.125}$$

where (4.61) and, again, (4.108), (4.109) were used. With this notation, the left hand side of (4.123) is equal to

$$\frac{\mathcal{W}_k}{\bar{h}^n}\sum_{i=0}^{n}\binom{n}{i}[\omega'_{k,h}(\bar{x}_k)\,A(\bar{x}_k)^i\,B(\bar{x}_k)^{n-i} - \omega'_{k+1,h}(\bar{x}_{k+1})\,A(\bar{x}_{k+1})^i\,B(\bar{x}_{k+1})^{n-i}], \tag{4.126}$$

$n = 1, 2, 3, 4$. We are using (4.124) and the estimates (4.125) on the summands

$$\frac{\mathcal{W}_k\;\omega'_{j,h}(\bar{x}_j)\;A^i(\bar{x}_j)\;B^{n-i}(\bar{x}_j)}{\bar{h}^n}, \quad j = k, k{+}1, \quad n{=}1, 2, 3, 4, \quad 0{\le}i{\le}n,$$

from (4.126) to get the bounds

$$c\,\frac{\Delta t^{2n-i}}{\bar{h}^{4n-2i}}.$$

Using the stability constraint (4.119), this is estimated by

$$\Delta t\;\bar{h}^{4n-2i-4}.$$

Only in the case $n{=}1$, $i{=}1$, the exponent of $\bar{h}$ becomes negative. So to prove the boundedness of (4.126) by $c\,\Delta t$, we only have to estimate the summand for $n{=}i{=}1$ in (4.126), i.e.

$$\frac{\mathcal{W}_k}{\bar{h}}\,|\omega'_{k,h}(\bar{x}_k)A(\bar{x}_k) - \omega'_{k+1,h}(\bar{x}_{k+1})A(\bar{x}_{k+1})| = \frac{\mathcal{W}_k}{\bar{h}}\,|\delta(\omega_{k,h}(\bar{x}_k)) - \delta(\omega_{k+1,h}(\bar{x}_{k+1}))|,$$

by

$$c\,\frac{\Delta t\;\mathcal{W}_k}{\bar{h}}\,|\omega_{k,h}(\bar{x}_k) - \omega_{k+1,h}(\bar{x}_{k+1})| = c\;\Delta t\;\mathcal{W}_k\;\omega'_{k,h}(\bar{\xi}) \le c\;\Delta t$$

where we have used (4.61) and (4.124). ∎

Of course, the question arises if the stability result of this section can be improved, i.e. if the severe stability restraint (4.119) can be weakened. We cannot answer the question here. However, we can prove that the 'Lipschitz continuity' of the coefficient functions in the sense of (4.120) is violated whenever (4.119) is weakened:

Lemma 4.21 *Under the assumptions of Lemma 4.19, but with (4.119) replaced by*

$$\Delta t := c\,\bar{h}^{4-\epsilon}, \quad 0<\epsilon<1,$$

and using the grid point distribution function $\omega(x)$ which is equal to x^2 at least on an interval $[0,\zeta]$, $\zeta>0$, there is no constant $c>0$ independent of $\bar{h}$, Δt such that the estimate (4.120) holds.

Proof. Let us consider the velocity field $v(t,x) := -(1+x)e^t/(2-e^t)$ in a neighbourhood of $(t,x)=(0,0)$. The length of the characteristics according to (4.12) is $\delta(x_k) = (1+x_k)\,(1-e^{-\Delta t})$, $x_k = \bar{x}_k^2 = k^2\bar{h}^2$. For grid points $x_k>0$ ($k>0$ small enough) we can obviously choose $\omega_{k,h}\equiv\omega$ (Lemma 4.16), and therefore the length of the characteristics in transformed coordinates is

$$\bar{\delta}(\bar{x}_k) = \sqrt{\bar{x}_k^2 + (1+\bar{x}_k^2)\,(1-e^{-\Delta t})} - \bar{x}_k$$

(see (4.111)) for $\Delta t>0$ small enough. A Taylor expansion of $\bar{\delta}$ with respect to Δt yields

$$\bar{\delta}(\bar{x}_k) = \frac{1+\bar{x}_k^2}{2\bar{x}_k}\,\Delta t - R_1, \quad R_1 \geq 0,$$

as well as

$$\bar{\delta}(\bar{x}_k) = \frac{1+\bar{x}_k^2}{2\bar{x}_k}\,\Delta t - \left[\frac{(1+\bar{x}_k^2)^2}{8\bar{x}_k^3} + \frac{1+\bar{x}_k^2}{4\,\bar{x}_k}\right]\,\Delta t^2 + R_2, \quad R_2 \geq 0$$

for small Δt. Therefore the estimates

$$\bar{\delta}(\bar{x}_k) \leq A(\bar{x}_k), \quad \bar{\delta}(\bar{x}_k) \geq A(\bar{x}_k) - B(\bar{x}_k) \tag{4.127}$$

hold where we have defined

$$A(\bar{x}_k) := \frac{1+\bar{x}_k^2}{2\bar{x}_k}\,\Delta t, \quad B(\bar{x}_k) := \frac{(1+\bar{x}_k^2)^2}{8\bar{x}_k^3}\,\Delta t^2 + \frac{1+\bar{x}_k^2}{4\,\bar{x}_k}\,\Delta t^2. \tag{4.128}$$

Evaluating the Lagrangian polynomials in the left hand side of estimate (4.120) for $k=2$, $n_1=n_2=0$, $\Delta k=1$ by (4.118) and then applying (4.127) we get

$$\begin{aligned}
&-\frac{1}{\sqrt{\omega'_{2,h}(2\bar{h})\omega'_{3,h}(3\bar{h})}}\,(\omega'_{3,h}(3\bar{h})\,c_0^*(x_3)\,c_1^*(x_3) - \omega'_{2,h}(2\bar{h})\,c_0^*(x_2)\,c_1^*(x_2)) \\
&= -\frac{1}{\sqrt{6}}\,(3L^{0,0}_{-1,1}(\bar{\delta}(3\bar{h}))\,L^{-1,0}_{-1,1}(\bar{\delta}(3\bar{h})) - 2L^{0,0}_{-1,1}(\bar{\delta}(2\bar{h}))\,L^{-1,0}_{-1,1}(\bar{\delta}(2\bar{h}))) \\
&= \frac{1}{\sqrt{6}}\left[3\,\frac{\bar{\delta}(3\bar{h})}{2\bar{h}}\left(1-\frac{\bar{\delta}(3\bar{h})}{\bar{h}}\right)\left(1-\frac{\bar{\delta}(3\bar{h})^2}{\bar{h}^2}\right) - 2\,\frac{\bar{\delta}(2\bar{h})}{2\bar{h}}\left(1-\frac{\bar{\delta}(2\bar{h})}{\bar{h}}\right)\left(1-\frac{\bar{\delta}(2\bar{h})^2}{\bar{h}^2}\right)\right] \\
&\geq \frac{1}{2\sqrt{6}}\,\left[3\,\frac{A(3\bar{h})-B(3\bar{h})}{\bar{h}}\left(1-\frac{A(3\bar{h})}{\bar{h}}\right)\left(1-\frac{A(3\bar{h})^2}{\bar{h}^2}\right)\right. \\
&\qquad\left. -2\,\frac{A(2\bar{h})}{\bar{h}}\left(1-\frac{A(2\bar{h})-B(2\bar{h})}{\bar{h}}\right)\left(1-\frac{(A(2\bar{h})-B(2\bar{h}))^2}{\bar{h}^2}\right)\right]
\end{aligned} \tag{4.129}$$

Introducing $\Delta t := \bar{h}^{4-\epsilon}$ (4.129) is equal to

$$\frac{1}{\sqrt{6}} \left[\frac{5}{4} + \frac{1}{72} \bar{h}^{-\epsilon} + R \right] \Delta t \tag{4.130}$$

where the rest R has the shape $R = \sum_{i=2}^{24} \sum_{j=1}^{6} c_{ij} \bar{h}^{i-j\epsilon}$, $c_{ij} \in \mathbb{R}$ independent of $\bar{h}$, Δt. (4.130) cannot be bounded by any expression $c\,\Delta t$ for $\bar{h} \to 0$. ∎

4.4.6 Summary and numerical results

Stability. In Chapter 4 it was pointed out that the stability criteria of our scheme depend on

- whether the interpolation in space is linear or not ($p=1$, $p>1$) and on
- the mesh

Let us summarize the main results of this chapter:

If linear interpolation in space is used, the scheme is stable in L^∞ for arbitrary meshes and arbitrary Δt.

If the interpolation is space is nonlinear and the mesh is equidistant or quasi-uniform, L^2-stability can be proved for a bounded Courant number. This weak restriction is rather unimportant for a practical use of the method.

In the context of our Navier-Stokes solver, the stability on a Chebyshev-Gauss-Lobatto mesh (which is not quasi-uniform) is of significance. For the original interpolation scheme, *in*stability (in the sense of Lemma 4.15) can be proven even for arbitrary small timesteps. When the interpolation is modified using a local mapping onto an equidistant grid, stability is gained if the severe condition

$$\Delta t = O(h_{mean}^4) = O(h_{max}^2) \tag{4.131}$$

is fulfilled.

spatial interpol.	grid	stability cond.
linear ($p=1$)	arbitrary	none
higher order ($p \geq 2$)	quasi-uniform	CFL no. bounded
higher order ($p=2$)	G.L.	always unstable
higher order ($p=2$) modified scheme	G.L.	see (4.131)

Table 4.2: Stability restrictions depending on the spatial order and the mesh type.

We are faced with the question if the condition (4.131) is necessary. An indication that the condition cannot be improved was given in the last section:

When (4.131) in weakened, the coefficients of the interpolation scheme lose their Lipschitz property which is usually required for stability proofs.

However, numerical stability tests seem to be useful. For our tests, the space interval consists of a concatenation of three Gauss-Lobatto domains. We used several smooth initial conditions and the nonlinear equation

$$u_t + u\,\nabla u = 0 \tag{4.132}$$

in one space dimension. We could *not* find any stability problems, even if the exact solution becomes discontinuous within the time interval $[0,T]$, and even if the non-modified scheme is used! Even if the initial values at the collocation points were independently chosen by random (i.e. the transport field a is, at least at time $t=t_0$, at random)[21], the scheme was observed to be stable.

Let us discuss this discrepancy of theory and numerics. Obviously, a local (in space and time) violation of the Lipschitz condition rsp. of the stability condition (4.51) does not necessarily mean that the calculation becomes instable. Let us concentrate on the example which was used in Lemma 4.15 to show the violation of the Lipschitz condition. If we start a test run with the *linear* equation (i.e. a independent of u in (4.1)) with the *non-modified* scheme, the initial flow field $u(t_0,x_k)=\delta_{k,k_0}$ (k_0 fixed) of 'Dirac type', the fixed transport field $a(t,x)=u(t_0,x)$, then the situation of proof 4.15 and Fig. 4.9 is recovered at *each timestep* at $x=x_{k_0}$: In each timestep, the characteristic starting at x_{k_0} ends at a fixed position $\tilde{x}$ where $\mathcal{S}u^n(\tilde{x}) > u^n(x_{k_0})$. As expected, in this test run, $u^n(x_{k_0}) = (\mathcal{S}u^n(\tilde{x})/u^n(x_{k_0}))^n$ goes to infinity rapidly for $n\to\infty$ for arbitrary Δt. If we replace the transport field a by $a(t,x)=u(t,x)$ then the described situation only takes place in the first timestep or in a limited number of timesteps, and the solution stays bounded. (In the case $a(t,x)=u(t,x)$, the value of $u(t,x_{k_0})$) and therefore the length of the characteristic at x_{k_0} increases at first from timestep to timestep. But when the end point $\tilde{x}_k$ of the characteristic approaches x_{k_0+1}, then $u(\tilde{x}_k)$ is not any longer larger than $u(x_{k_0})$ (see Fig. 4.9) and the solution does not increase any more.) In that situation the violation of the Lipschitz condition does not lead to instability.

Accuracy. Let us consider the initial condition

$$u_0(x) := 1 - x^2$$

on $[-1,1]$. The exact solution of (4.1) with $a\equiv u$ for the initial condition $u(0,x)=u_0(x)$ is

$$u(t,x) := 1 - \left(\frac{2\,(x-t)}{1+\sqrt{1-4tx+4t^2}}\right)^2 \tag{4.133}$$

(see Fig. 4.11) for $x\in(-1,1)$, $t\geq 0$. We made test runs for this problem on

[21] For the stability proofs in Sec. 4.4, we assumed the transport field to be smooth.

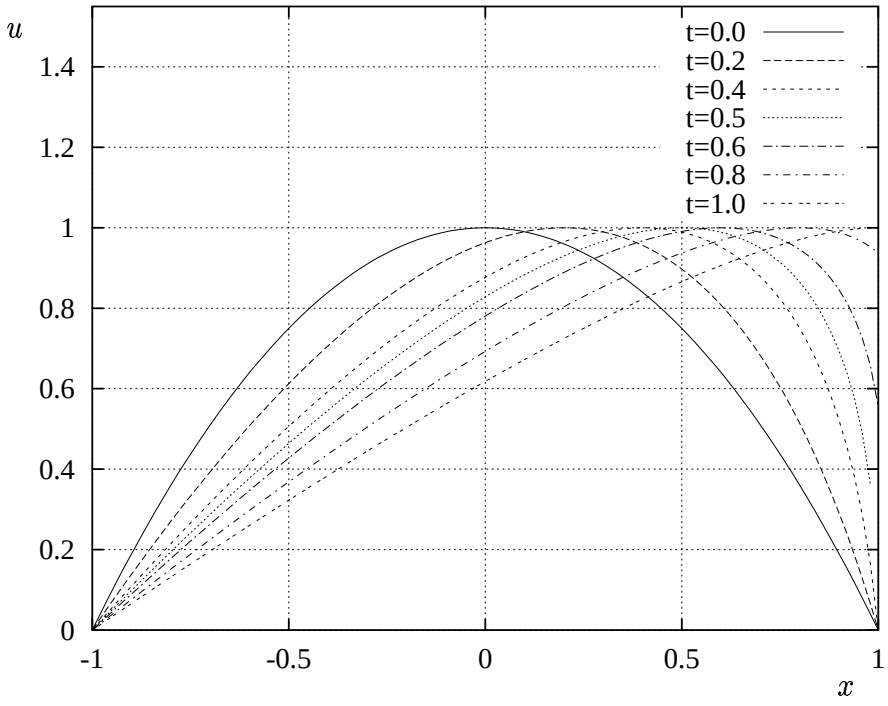


Figure 4.11: Exact solution (4.133).

the spatial interval $(-1,0)$ on an equidistant grid. Figs. 4.12-4.14 show the $L^\infty(-1,0)$-error at time $T=1$. In Fig. 4.12 we set $\Delta t = h$, i.e. the global CFL number is 1. As predicted by the estimates (4.17) and (4.35), we get an error of order $\min\{p, q+1\} = 1, 2, 3$ for the parameter sets (p,q) equal to $(1,1)$,$(2,1)$,$(3,2)$.

To investigate more closely the accuracy of the error estimates (4.17), (4.35), Figs. 4.13-4.14 show test runs where we have dropped the coupling $\Delta t = h$. In Fig. 4.13, $\Delta t = 2^{-9}$ is fixed and the error decay for $h \to 0$ is studied. For $h \gg \Delta t$, we recover the same error decay as in Fig. 4.12, as the error in (4.17), (4.35) is governed by the term h^p. For $h \ll \Delta t$, the error estimates are dominated by the term Δt^{q+1}. Therefore the error in Fig. 4.13 becomes constant for $h \to 0$. For higher order methods ($p=2$, $p=3$) the transition from the h-dependent to the h-independent behaviour takes place at a Courant number of 1. This could be expected, as for smaller h, the characteristics begin to cross adjacent grid points where the spatial interpolation function is only C^0-regular. For our test runs with *linear* interpolation, the error becomes stationary for $h \approx 2^{-12}$, i.e. $CFL \approx 8$. So for the lower order method, the crossing of adjacent grid points by the characteristics is unproblematic.

However, let us point out that our higher order methods ($p \geq 2$) are *not restricted* to Courant numbers less than 1. It is just *more efficient* to chose h, Δt such that the Courant number is not bigger than 1.

A closer look at the curves of Fig. 4.13 shows that the error decay is stronger

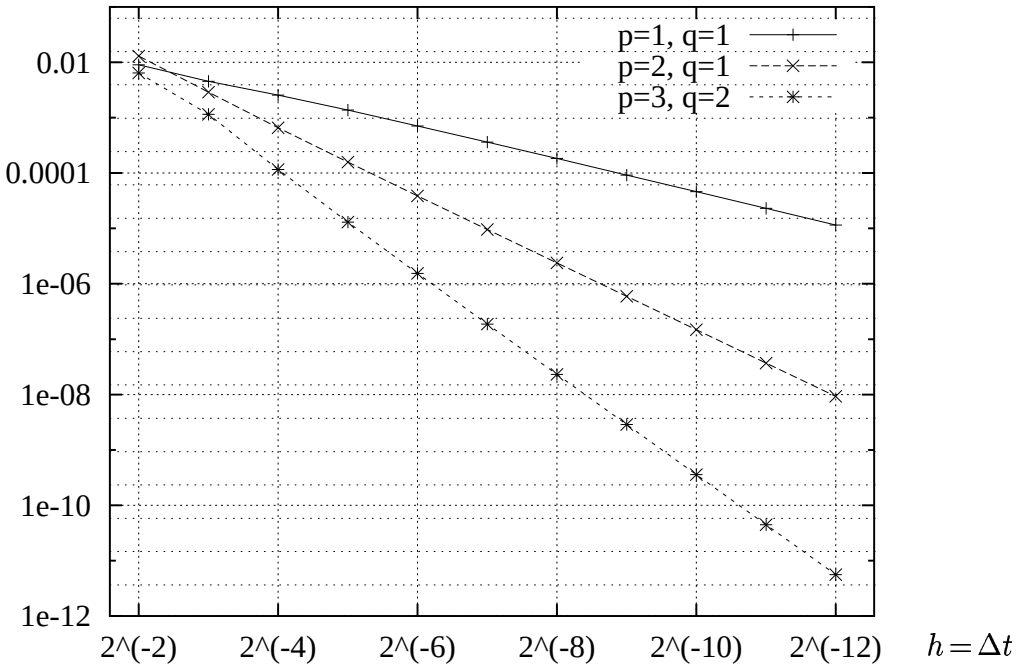


Figure 4.12: L^∞-error for $h = \Delta t \to 0$. In this and the following figures, dotted mesh lines to indicate the (even) powers of 2 are displayed. They serve to facilitate the check of convergence order.

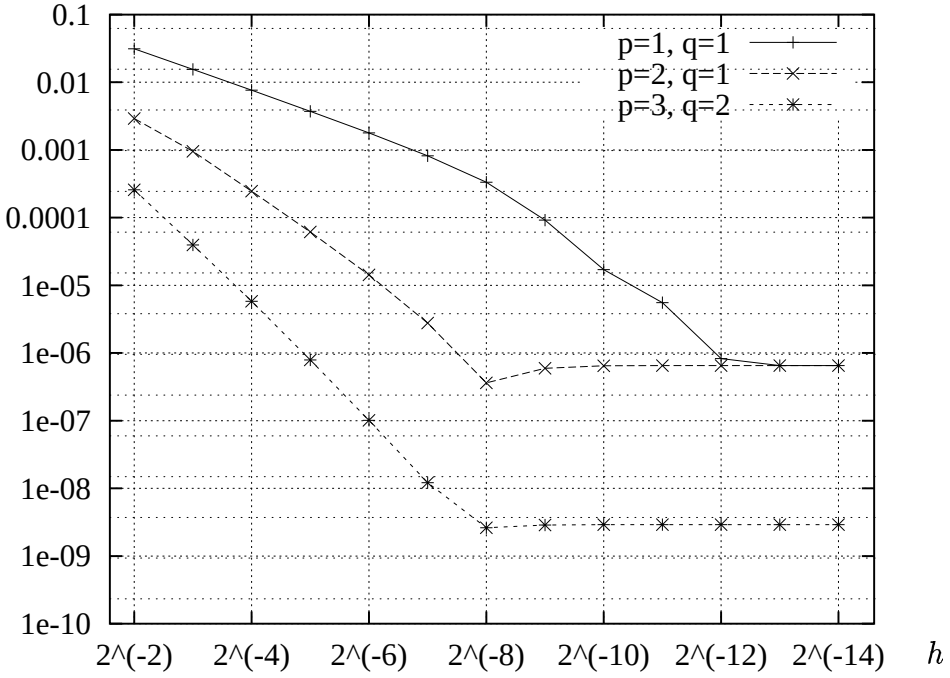


Figure 4.13: L^∞-error for $h \to 0$, $\Delta t = const.$

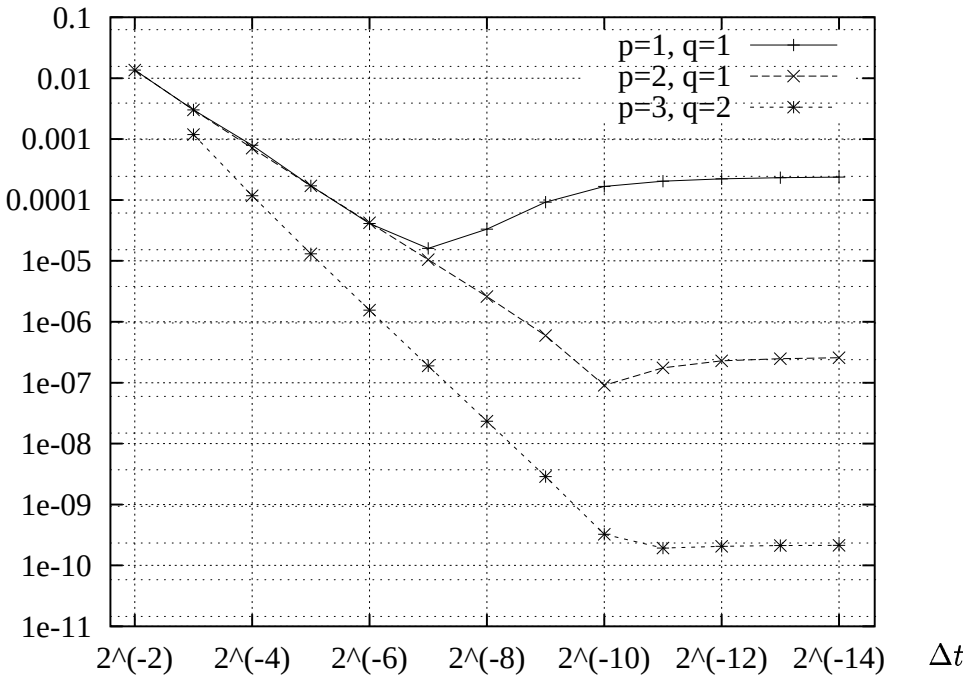


Figure 4.14: L^∞-error for $\Delta t \to 0$, $h = const$.

in the intermediate part than in the left part of the curves. Even this behaviour is reflected by the estimates (4.17), (4.35): For certain 'intermediate' h, the term $h^{p+1}/\Delta t$ dominates h^p causing a stronger error decay than the term h^p which dominates the left part of the curves.

Fig. 4.14 shows test runs for fixed $h = 2^{-9} = const$ and $\Delta t \to 0$. For $\Delta t \gg h$, the term Δt^{q+1} is dominant in the error estimates (4.17), (4.35). This causes the error decay in the left part of the curves. For certain Δt, the term $h^{p+1}/\Delta t$ governs the estimates, i.e. the error increases. For $\Delta t \ll h$, the term h^p is dominant, i.e. the error becomes independent of Δt.

All these observations show that the error estimates describe very accurately the behaviour of the error; they seem to be 'optimal'.

Fig. 4.15 shows a comparison between the equidistant and the Chebyshev-Gauss-Lobatto discretization. The test example (4.133) with $\Delta t \sim h$ and $T = 1$ is used on the spatial interval $[-1, 1]$. [22] The full line in Fig 4.15 represents the use of an equidistant grid with $h = \Delta t$. The two broken lines represent the use of a Gauss-Lobatto mesh, the upper line with $h_{mean} = \Delta t$, the lower line with $h_{max} = \Delta t$. For the latter, about three times more grid points than for the first (and for the equidistant) case have to be used.

Of course, if the *maximum* mesh size of the Gauss-Lobatto mesh is equal to

[22] On this interval, the gradient of the exact solution becomes unbounded at $t = 1/2$. Obviously, this singularity does not perturb our numerical scheme.

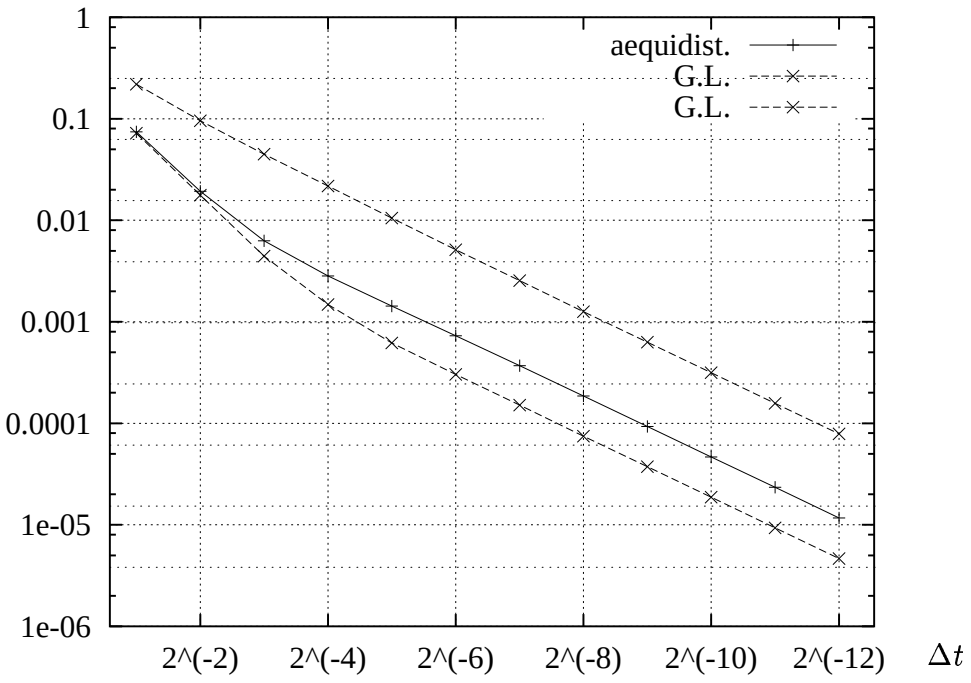


Figure 4.15: L^∞-error in dependence on Δt for $h \sim \Delta t \to 0$, $p = 1$, $q = 1$; comparison of equidistant grid and Gauss-Lobatto grid. In the equidistant case, $h := \Delta t$ was set. For the two Gauss-Lobatto curves, $h_{mean} := \Delta t$ (upper curve) rsp. $h_{max} := \Delta t$ (lower curve) was set.

the equidistant mesh size, the Gauss-Lobatto mesh causes a smaller error. But if the *mean* Gauss-Lobatto mesh size is equal to the equidistant mesh size (i.e. the same number of grid points is used), the equidistant mesh causes a smaller error.

Chapter 5

The Navier-Stokes Solver

5.1 Remarks on the code

In a first stage, the CGBI solver was implemented with a Chebyshev spectral solver and a FD solver (see test runs in Sec. 2.8, 3.1.5, 3.3, 3.4.3). In *this* chapter, test runs with the newly implemented FE solver replacing the FD solver are presented. However, the FD solver still may be used for comparisons of the test results.

Let us begin with the description of the Navier-Stokes time scheme (fractional step scheme). Afterwards, we will remark on some parts of the parallel Navier-Stokes solver.

1. The Navier-Stokes splitting scheme. Our time splitting scheme (proposed in [5]) is of the pressure correction type.[1]

Let $\vec{u}^n$ be a numerical approximation of the velocity field $\vec{u}(t_n)$ at time t_n and p^n an approximation of the pressure. The Lagrangian ('material') derivative

$$D_t\vec{u} := \vec{u}_t + \vec{u}\,\nabla\vec{u}$$

of the velocity field at time $t = t^{n+1}$ in (1.1) is approximated by the first order backward Euler method:

$$D_t\vec{u}(t_{n+1}) \approx \frac{\vec{u}_*^{n+1} - \vec{u}^n}{\Delta t} \tag{5.1}$$

Here, we have put $\vec{u}_*^{n+1}(x) := \vec{u}^n(\vec{X}(t_n; t_{n+1}, x))$ where $\vec{X}(t_n; t_{n+1}; x)$ is the so-called *foot point* of the characteristics $\vec{X}(t; t_{n+1}, x)$ starting at (t_{n+1}, x) and ending at time $t = t_n$. The computation of the characteristics, however, consists of the

[1] Pressure correction schemes ('fractional step methods') where introduced by Temam and Corin in the 1960s, see [51].

numerical solution of an ordinary differential equation

$$\begin{aligned} \frac{d}{dt}\vec{X}(t) &= \vec{u}(t,\vec{X}(t)), \\ \vec{X}(t_{n+1}) &= x \end{aligned} \tag{5.2}$$

for each mesh point x. The numerical approximation of the flow field $\vec{u}$ in (5.2) may be *time-dependent*; instead of $\vec{u}^n$, a combination of $\vec{u}^n$, $\vec{u}^{n-1}$,...,$\vec{u}^{n-q}$ may be used to extrapolate $\vec{u}$ for $t \in [t_n, t_{n+1}]$.[2] In this chapter we use $q = 1$. A higher value for q is reasonable only if a higher approximation order in (5.1) is used.

When $\vec{u}_*^{n+1}$ is known, an implicit 'diffusion step' is performed. The resulting velocity field $\vec{u}_{**}^{n+1}$ is not yet solenoidal. Hence, a pressure correction step yielding $\vec{u}^{n+1}$ and p^{n+1} is performed.

So each Navier-Stokes time step is splitted into the following three problems:

(i) Solve, for each mesh point x, the initial value problem (5.2) on the time interval $[t_n, t_{n+1}]$ and set $\vec{u}_*^{n+1}(x) := \vec{u}^n \circ \vec{X}^{n+1}(t_n; t_{n+1}, x)$.

(ii) Solve the elliptic partial differential equation[3]

$$\frac{\vec{u}_{**}^{n+1} - \vec{u}_*^{n+1}}{\Delta t} - \nu \Delta \vec{u}_{**}^{n+1} + \nabla p^n = \vec{f}. \tag{5.3}$$

Each component of the vector equation (5.3) is of the Helmholtz resolvent type

$$(\sigma I - \Delta) w = \tilde{f}, \quad \sigma > 0. \tag{5.4}$$

In fact, $\sigma = \frac{1}{\nu \Delta t}$.

(iii) The pressure correction step: Solve

$$-\Delta \Pi^n = -\frac{1}{\Delta t} \text{ div } \vec{u}_{**}^{n+1} \tag{5.5}$$

and do the update

$$\vec{u}^{n+1} := \vec{u}_{**}^{n+1} - \Delta t \, \nabla \Pi^n, \tag{5.6}$$

$$p^{n+1} := p^n + \Pi^n. \tag{5.7}$$

The new velocity field $\vec{u}^{n+1}$ is solenoidal, then.

[2] The parameter q is the same as in Chapter 4 (p. 140).

[3] In the following simulations we use no external force ($\vec{f} = 0$ in (5.3)).

Let us mention that in (5.3) the 'old' pressure p^n is used. This deviation from the original Temam/Chorin pressure correction method is known as a measure to increase the accuracy.

The characteristics solver. On each subdomain the characteristics solver uses the same grid as the local elliptic solver, i.e. the characteristics scheme presented in Chapter 2 is implemented for two-dimensional equidistant, Chebyshev-Gauss-Lobatto and finite element meshes. Each processor computes the characteristic lines for those grid points x_0 which are situated in its domain $\bar{\Omega}_i$. The problem of characteristics crossing the interface to Ω_{i-1} or Ω_{i+1} (Fig. 5.1) is handled by introduction of overlapping stripes (Fig. 5.2): The flow field data of these stripes are exchanged before each timestep (rsp. subtimestep, see below) so that during the computation of the characteristics, no interprocessor communication is needed. This technique reduces the number of communication startups drastically.

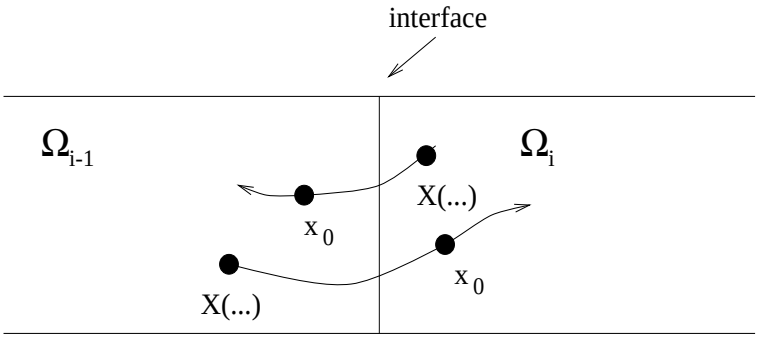


Figure 5.1: Characteristic lines crossing an interface.

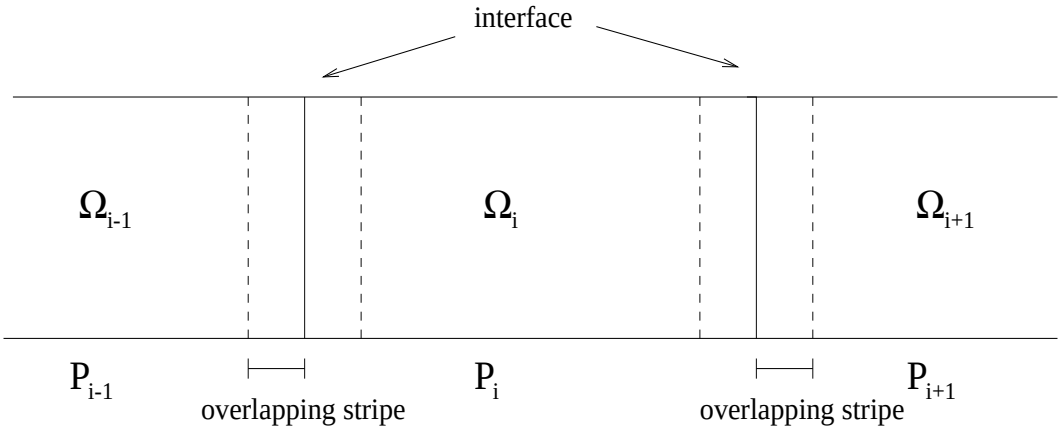


Figure 5.2: Overlapping stripes to reduce communication.

Each Navier-Stokes timestep may be subdivided into substeps ('subcycling'). This enables us to limit the length of the characteristics (even for large Δt) such that

- this length is smaller than the chosen width of the overlapping stripes and
- that, on FE subdomains, the characteristics do not leave the elements adjacent to its starting point which simplifies the tracing of the characteristics on FE domains.

So substeps are introduced to simplify the computation/data management, but they are *not* a consequence of any stability restriction. As a matter of fact, we could *not* observe any restriction of Δt due to stability problems although also higher order interpolation ($p=1,2,3$) on Gauss-Lobatto meshes was used.

Calculation of derivatives. The pressure correction algorithm involves the calculation of derivatives:

- The calculation of the divergence for the intermediate velocity in (5.5),
- the calculation of the gradient of the pressure p^n and of the pressure difference Π^n in (5.3), (5.6).

On the Gauss-Lobatto mesh, these derivatives are calculated by using the differentiation matrix (see [12] p. 69). This ensures that the velocity $\vec{u}^n$ is solenoidal for all *inner* grid points of the subdomains Ω_i. On the interfaces, the divergence is not forced to be absolutely equal to zero, as our spectral collocation approach imposes the differential equation only on *inner* grid points. However, due to the regularity of the solution and the fineness of the mesh at the interfaces this effect was not observed to influence the flow rate across the interfaces.

On the FD subdomains, the gradient and the divergence are approximated by using forward and backward difference quotients, respectively. So the concatenation of both discrete operators is equal to the discrete Laplacian operator (2.74). This ensures that the velocity field $\vec{u}^n$ is solenoidal in the sense of the discrete divergence operator. This holds even on the interfaces between FD subdomains, as the FD approach imposes the differential equation not only on interior nodes, but also on Neumann boundary mesh points.

On FE subdomains, derivatives of the piecewise linear functions are piecewise constant functions which are only well defined on the elements and discontinuous at the edges and the nodes. For the divergence in (5.5), this is not a problem, as the equation is solved in the *weak* formulation, i.e. the divergence of the intermediate velocity field has to be known in the sense that integrals over $div\, \vec{u}_{**}^{n+1}\, \varphi_i$ (φ_i being a FE ansatz function) have to be calculated. The same holds for the gradient of the pressure in (5.3). Equation (5.6) requires the nodal values of $\nabla\Pi^n$ which are not well-defined.[4] To handle this problem we use (5.6) in a *weak* formulation in combination with the so-called 'mass-lumping'[5] technique for the

[4] $\nabla\Pi^n$ is discontinuous

[5] The mass element matrix is replaced by a certain diagonal matrix.

terms $\vec{u}^{n+1}$ and $\vec{u}_{**}^{n+1}$ which avoids the solution of an additional system to get the nodal values of $\vec{u}^{n+1}$. This approach leads to a nodal value being an average over the (constant) values on the surrounding elements, weighted by the size of the elements. Another possibility is to weight the values on the elements by angles instead of elements size. Both methods for the FE case lead to a discrete velocity field which is at least approximately solenoidal.

Boundary conditions. For the characteristics part of the Navier-Stokes solver we have to pose boundary conditions only on the inflow part Γ^I of the boundary. If a characteristic crosses the inflow boundary we use a constant prolongation of the velocity field outside the domain.

For the elliptic problems (5.3), (5.5) we are using the following boundary conditions:

	Γ^W	Γ^I	Γ^O
$\vec{u}\cdot\vec{n}$	$\vec{u}\cdot\vec{n}=0$	$\vec{u}\cdot\vec{n}=u_{\Gamma^I}$	$\partial\vec{u}/\partial\vec{n}=0$
$\vec{u}\cdot\vec{\tau}$	$\vec{u}\cdot\vec{\tau}=0$	$\vec{u}\cdot\vec{\tau}=0$	$\partial\vec{u}/\partial\vec{\tau}=0$
p	$\partial p/\partial\vec{n}=0$	$\partial p/\partial\vec{n}=\varphi_{\Gamma^I}$	$p=0$

Here, $\vec{n}$ denotes an outward normal vector field on $\partial\Omega$ and $\vec{\tau}$ a tangential vector field. So we have 'no-slip' boundary conditions on the physical wall Γ^W. u_{Γ^I} and φ_{Γ^I} are chosen according to a Poiseuille flow, i.e. if Γ^I is identified with the interval $(0,B)$ we set

$$u_{\Gamma^I}(y)=U_{max}\,\frac{4\,y\,(B-y)}{B^2},\quad \varphi_{\Gamma^I}(y)=\nu\,U_{max}\,\frac{8}{B^2}. \tag{5.8}$$

Initial conditions. We have used two different methods to start the flow. The first one is to start with a zero flow field and to increase the inflow velocity smoothly over a time interval $(0,\tau)$: For $t\geq\tau$ we use (5.8). For $0\leq t\leq\tau$ we use

$$C(t)\,u_{\Gamma^I}(y)$$

instead of $u_{\Gamma^I}(y)$, where the time-dependent amplifier $C(t)$ is the third order polynomial with $C(0)=0$, $C(\tau)=1$, $C'(0)=C'(\tau)=0$. We took $\tau=0.5$.

The second method to start the flow is to assume that $\vec{u}=0$ in Ω, but to impose the inflow condition (5.8) immediatly at $t=0$. By discretization, this discontinuous initial field is represented by a smooth field, so the solver can be applied. As this initial velocity field is far away from being solenoidal, a 'shock' (large values for the pressure during the first time steps) occurs. The CGBI-characteristics-pressure-correction solver turned out to be robust enough to handle this shock without any problems.

5.2 Test runs

5.2.1 Flow past a backward facing step

As a first example, we present the computation of a flow past a backward facing step [31]. The Reynolds number is moderate so that the flow becomes stationary after some time. The reason to choose this example is the following: As we focus on the stationary state of the flow, the time discretization error of our scheme is of minor importance; we choose the rather large time step size $\Delta t = 0.08$. This test case enables us to investigate the spatial accuracy of the Navier-Stokes solver.

In a first stage, our domain is rectangular of size

$$\Omega = (0, 6) \times (0, 1), \tag{5.9}$$

and the step of height 0.5 is modelled by posing a Poiseuille inflow profile on *the half* of the left edge of the computational domain. This simplified geometry of the computational domain enables us to use the spectral solver on *all* subdomains. The maximum inflow velocity is 1.0, and the Reynolds number with respect to the step size and the maximum velocity at the inlet is $Re = 150$. We compare the test runs of

1. Chebyshev solvers on all 6 subdomains ('CC'-coupling), interpolation in the transport step by polynomials of order $p = 1$
2. Chebyshev solvers on all 6 subdomains ('CC'-coupling), interpolation in the transport step by polynomials of order $p = 3$
3. FE solver on first subdomain, Chebyshev solvers on the other subdomains ('FC'-coupling), $p = 3$ on the Chebyshev domains, $p = 1$ on FE domain,
4. FD solvers on all subdomains ('DD'-coupling), $p = 1$.

In a second stage, we include the upstream part of the channel into the computational domain in order to make a computation comparable to the benchmark [37]; our new computational domain Ω is L-shaped:

$$\Omega = (0, 0.75) \times (0.5, 1) \cup (0.75, 6) \times (0, 1). \tag{5.10}$$

It is devided into 6 subdomains of the same width; on the first, the FE solver is used. This geometry and the Reynolds number correspond to the benchmark tests in [37]; except that the length of our channel is smaller. However, our computational results showed that the influence of a longer channel in downstream direction on the numerical result is neglectable. The spectral solvers use $(N+1) \times (N+1)$ grid points on each subdomain, $N = 8, 16, 32, 64, 128$, and on the FE domain, a regular mesh of triangles with a similar number of nodes is

used. See Fig. 5.4 for the visualization of the x-component of the flow for both geometries ($N=64$, $t=51.2$) and Fig. 5.5 for the pressure.

For the analysis of the results, we focus on the length L of the recirculation zone at time $t=50$. At this time, the recirculation length is already stationary up to ± 0.005. Fig. 5.3 shows L in dependence on the method and on the spatial discretization parameter N. Fig. 5.3 reads that all test runs 1.-4. using the rectangular domain (5.9) lead, for $N \to \infty$, to a limit of about 2.67.[6] Furthermore, we see that for the case of Chebyshev solvers on all subdomains, the order of the spatial interpolation in the transport step is essential for the accuracy of the whole Navier-Stokes solver: For the high order interpolation $p=3$, the result for $N=8$ is more accurate than for the lower order interpolation $p=1$ and $N=64$! Fig. 5.3 also shows that the influence of the lower order FE domain onto the error of L is limited; the 'FC' result is clearly more accurate than the 'DD' result.

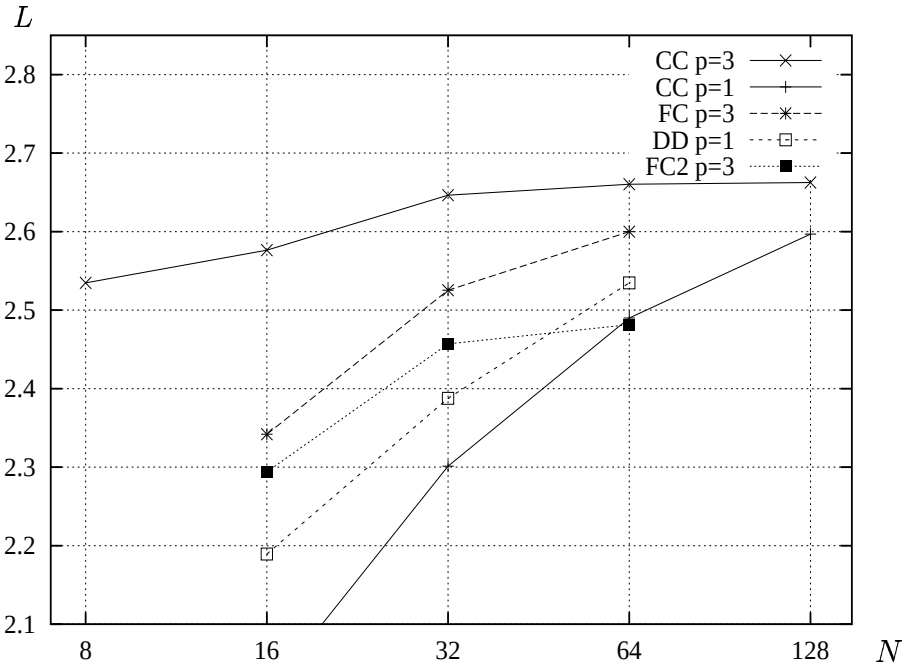


Figure 5.3: The length of the recirculation zone for different combinations of local solvers. CC=Chebyshev method on all subdomains, FC=finite element method on first subdomain and Chebyshev on the other subdomains, DD=finite difference method on all subdomains, FC2: as FE, but with the different geometry (5.10). $p=$ order of polynomial for spatial interpolation on Chebyshev domains for the method of characteristics.

Let u_N denote the numerical solution for $(N+1)\times(N+1)$ grid points per

[6] Recently, the result $L=2.67$ was confirmed by a computation with another parallel Chebyshev collocation code which uses an explicit matrix representation of the operator (2.77) [43].

subdomain. Assuming a simplified[7] law

$$\|u_N - u_{exact}\|_{L^2(\Omega)} \approx c\, N^{-\alpha} + f(\Delta t) \tag{5.11}$$

for the error, we can compute the order α approximately without knowing u_{exact} by

$$\alpha \approx \log_2 \frac{\|u_N - u_{N/2}\|_{L^2(\Omega)}}{\|u_{2N} - u_N\|_{L^2(\Omega)}}.$$

We get the following results:

method	order α
CC (p=3)	2.6
CC (p=1)	0.8
DD	0.7

Table 5.1: Approximate order of the different methods for the backward facing step

Results very similar to Table 5.1 are gained by regarding $L(N)$ instead of the $L^2(\Omega)$-error of u_N.

The fact that $\alpha < p$ seems to be a drawback of the fact that the Δt-dependent terms of (5.11) for different N do not extinct each other completely, pretending a smaller value for α. The lack of regularity of the solution near the step seems to be of minor importance, as tests with a smoother solution suggest.

Finally, let us compare our test runs 'FC2' on the L-shaped computational domain (5.10) with the benchmark [37]. We find a good correspondence of our results; in [37], most results are situated between 2.2 and 2.6; using a coarse discretization comparable to [37] ($N=32$), 'FC2' in Fig. 5.3 reads 2.46.

[7] The simplification consists in the substitution of the '$\leq$'-sign by the '$\approx$'/'='-sign.

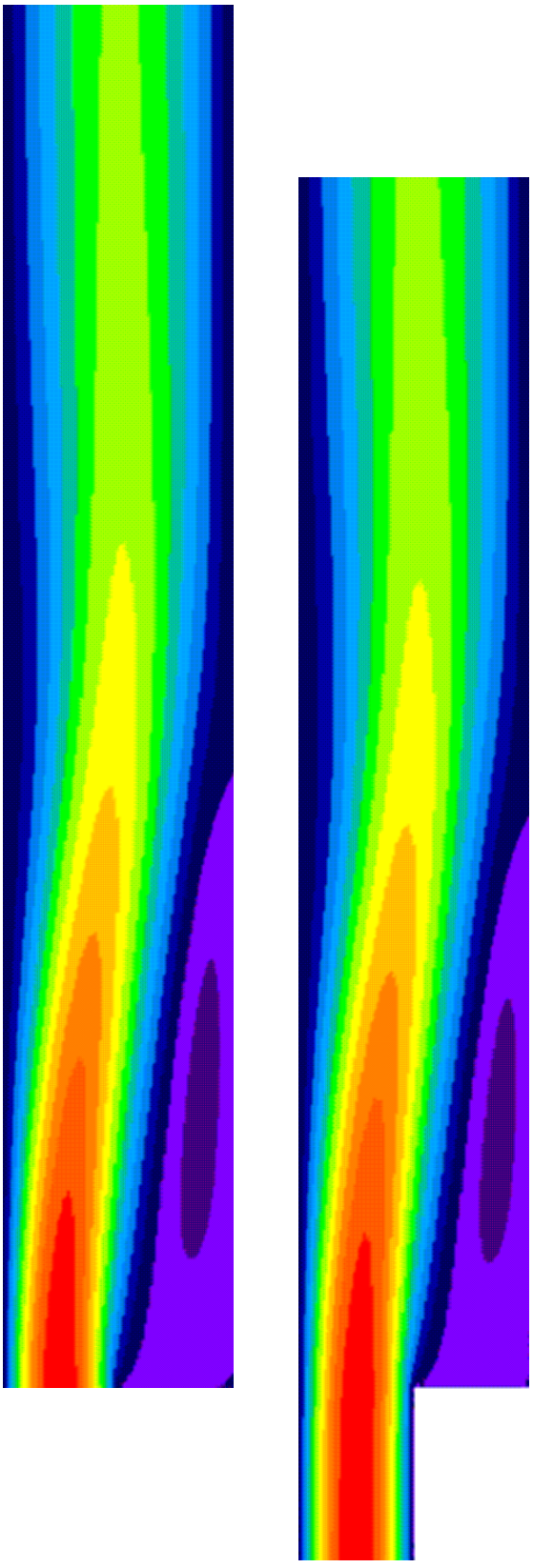


Figure 5.4: x-component of the velocity for the flow over a backward facing step for the two different computational domains (5.9) and (5.10) when the stationary state is reached. Each colour maps a range of $0.08\bar{3}$. Red indicates high velocity, the violet hues indicate negative velocity. Upper part: the computation on 6 spectral subdomains ('CC'), lower part: the computation on 1 FE and 5 spectral subdomains ('FC2'). The latter geometry leads to a shorter recirculation zone.

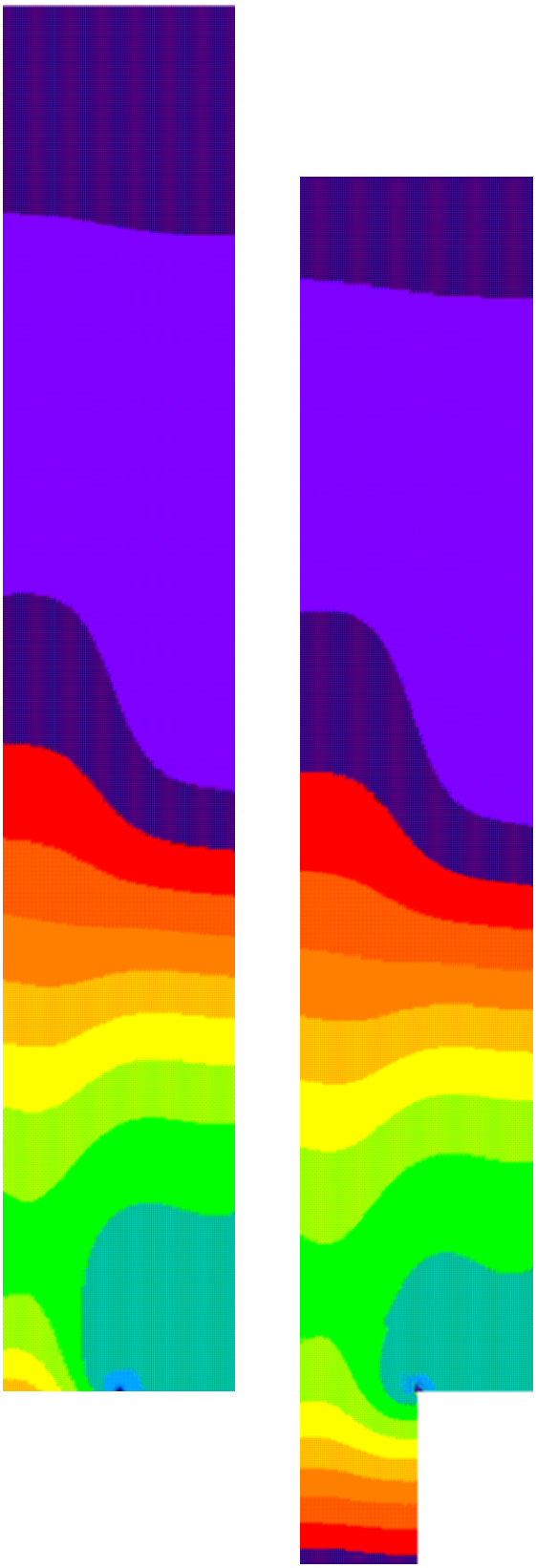


Figure 5.5: The pressure for the flow over a backward facing step for the two different computational domains (5.9) and (5.10) when the stationary state is reached. Each colour maps a range of 0.011. Violet indicates high pressure, blue and green indicate low pressure. Upper part: the computation on 6 spectral subdomains ('CC'), lower part: the computation on 1 FE and 5 spectral subdomains ('FC2'). At the inward facing edge, a pressure singularity occurs. We get realistic results although the presently used mesh has no refinement there.

5.2.2 Channel flow past a cylinder I

In the following we present the computation of a non-stationary 2d flow past a cylinder. The channel hight is 1, and the length of the computational domain 6. A circular obstacle of diameter 0.2 is situated on the symmetry axis. The center of the obstacle has a distance of 0.7 from the inlet. At the inflow, a Poiseuille profile scaled to $U_{max}=1$ is prescribed for the x-component of the velocity:

$$u(y) = 6\,U_{mean}\,y\,(1-y), \quad U_{mean} = 2/3$$

The Reynolds number $Re = U_{mean}D/\nu$ with respect to the mean velocity at the inlet and the diameter D of the obstacle is $Re=133$.

Our computation uses 6 square subdomains. The first, containing the obstacle, uses the FE solver and the other the spectral solver. Within $0 \leq t \leq 0.5$, $U_{mean}=U_{mean}(t)$ is increased smoothly from 0 to $2/3$ and is constant afterwards. To break the symmetry and accelerate the beginning of the periodic vortex shedding, the inflow profile is disturbed asymmetrically, but only within the time interval $[0, 0.5]$.

To make a comparison between different Δt, N, p, q and to get an impression of the accuracy of the numerical result, we are focussing on the Strouhal number

$$St = \frac{f\,D}{U_{mean}} = \frac{D}{T\,U_{mean}}$$

where f is the frequency of the vortex shedding and T its reciprocal. We compute T and St by monitoring the y-component u_y of the velocity at a fixed point $x=1.5$, $y=0.5$ behind the obstacle. If t_1, t_2, t_3,... are the moments when u_y is zero, then the difference $t_{k+2}-t_k$ is a value for the period T.

In Fig. 5.6 and also Fig. 5.7, u_y as a function of t is displayed for several settings of the numerical parameters. Fig. 5.6 shows that a periodic flow establishes rapidly. Several tests were made for the discretization parameter $N=64$. Those which are using linear spatial elements in the transport step on the spectral subdomains (i.e. $p=1$) show an amplitude of u_y which is obviously much too small if we compare it with the higher order/finer discretization test runs: For both $p=2$ and $p=3$, the amplitude is much higher, and both curves are very similar. Both curves show 'double peaks'; a feature that is missing in the result for $p=1$. We can conclude that for the chosen test problem higher order interpolation ($p \geq 2$) is essential to get realistic results. If we increase the number of mesh points to $N=128$, the shape of the u_y curve only changes moderately, but the amplitude of one of the two peaks increases.

To give a clearer view, Fig. 5.7 shows a magnification of Fig. 5.6 for certain settings of the numerical parameters. Now, also the dependence of u_y on the timestep size Δt is investigated. We see that the shape of the curves is not influenced by Δt, but only the period length. Focussing on the three computations

for $p=1$ we see that the period depends linearly on Δt. This impression seems reasonable as the whole Navier-Stokes time scheme is of first order.

As already explained, the u_y-curves are analyzed to find the period T and the Strouhal number St (Fig. 5.8). Again, the $p=1$-computations give rather inaccurate results of $T \approx 1.4$ with a linear convergence for $\Delta t \to 0$. The use of $q=1$ instead of $q=0$ improves the result slightly, which can be explained easily: If the larger value $q=1$ is used, one of the two leading error terms of the Navier-Stokes time scheme vanishes.

Higher precision is gained by $p \geq 2$. Comparing test runs for different Δt in Fig. 5.8, we can conclude that the error in time of the Strouhal number is about $2-3\%$ for $\Delta t=0.01$. The refinement of the mesh from $N=64$ to 128 diminishes the Strouhal number again about 4%.

The results are very satisfactory if we take into account that the discretization of the FE mesh is rather coarse; no refinement is used close to the obstacle. As long as the discretization on the FE domain is not finer than the discretization on the spectral domains, the global spatial discretization error is of course governed by the FE discretization error.

Finally, Fig. 5.9 visualizes the flow field. The x- and the y-component of the velocity, the Euclidian norm of the velocity and the pressure for $N=128$, $\Delta t=0.01$, $p=3$ [8] at time $t=13.1$ are displayed. In the left part of Fig. 5.13 the accompanying vorticity field is displayed. Very clearly, the Kàrmàn vortex street appears. Especially the vorticity field shows many details which are resolved by the highly accurate spectral method.

[8] $p=3$ on the Gauss-Lobatto domains only

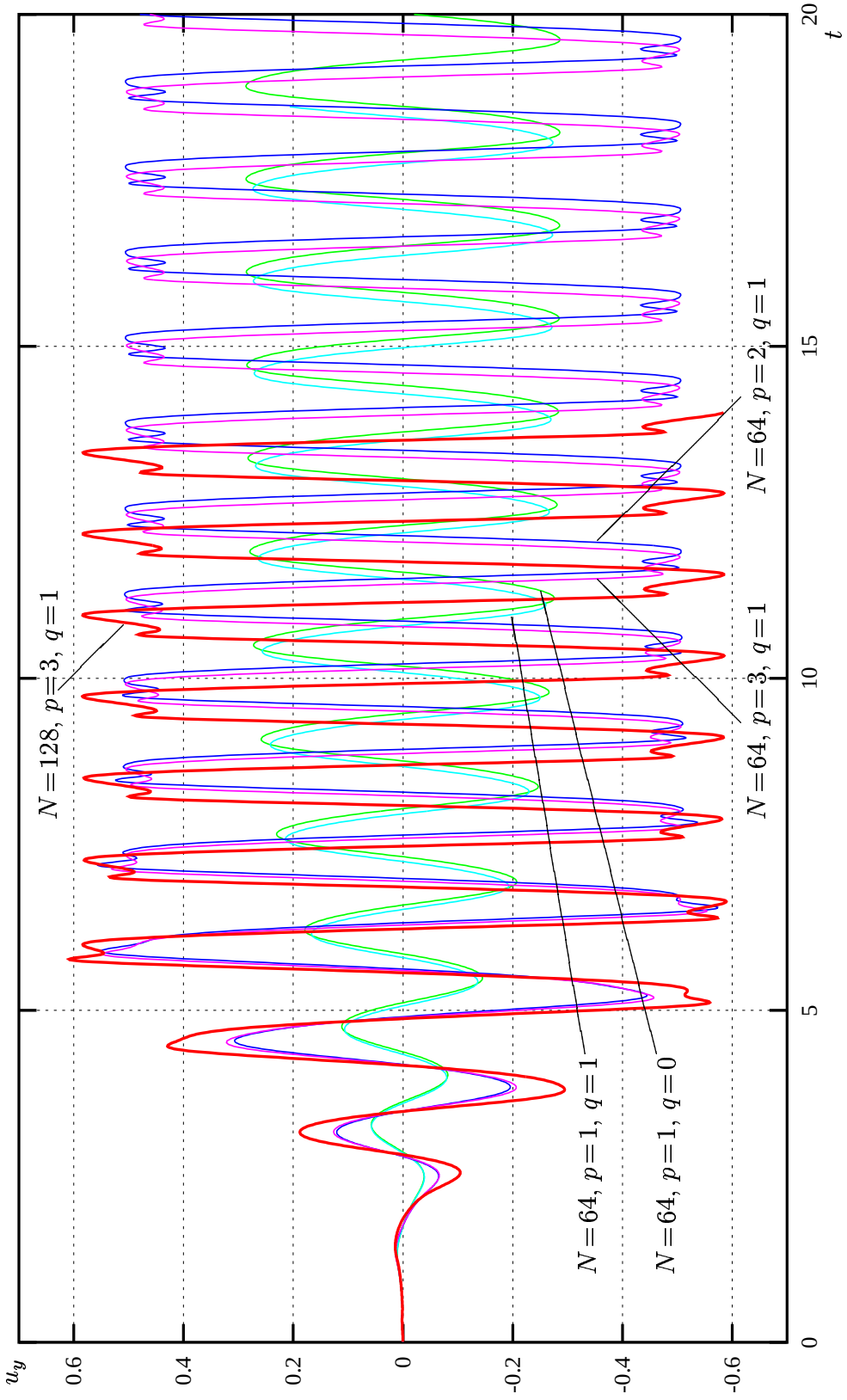


Figure 5.6: The transverse component $u_y(t)$ of the velocity at a fixed point on the symmetry axis behind the obstacle for different interpolation orders (p, q) in the transport step. $\Delta t = 0.01$.

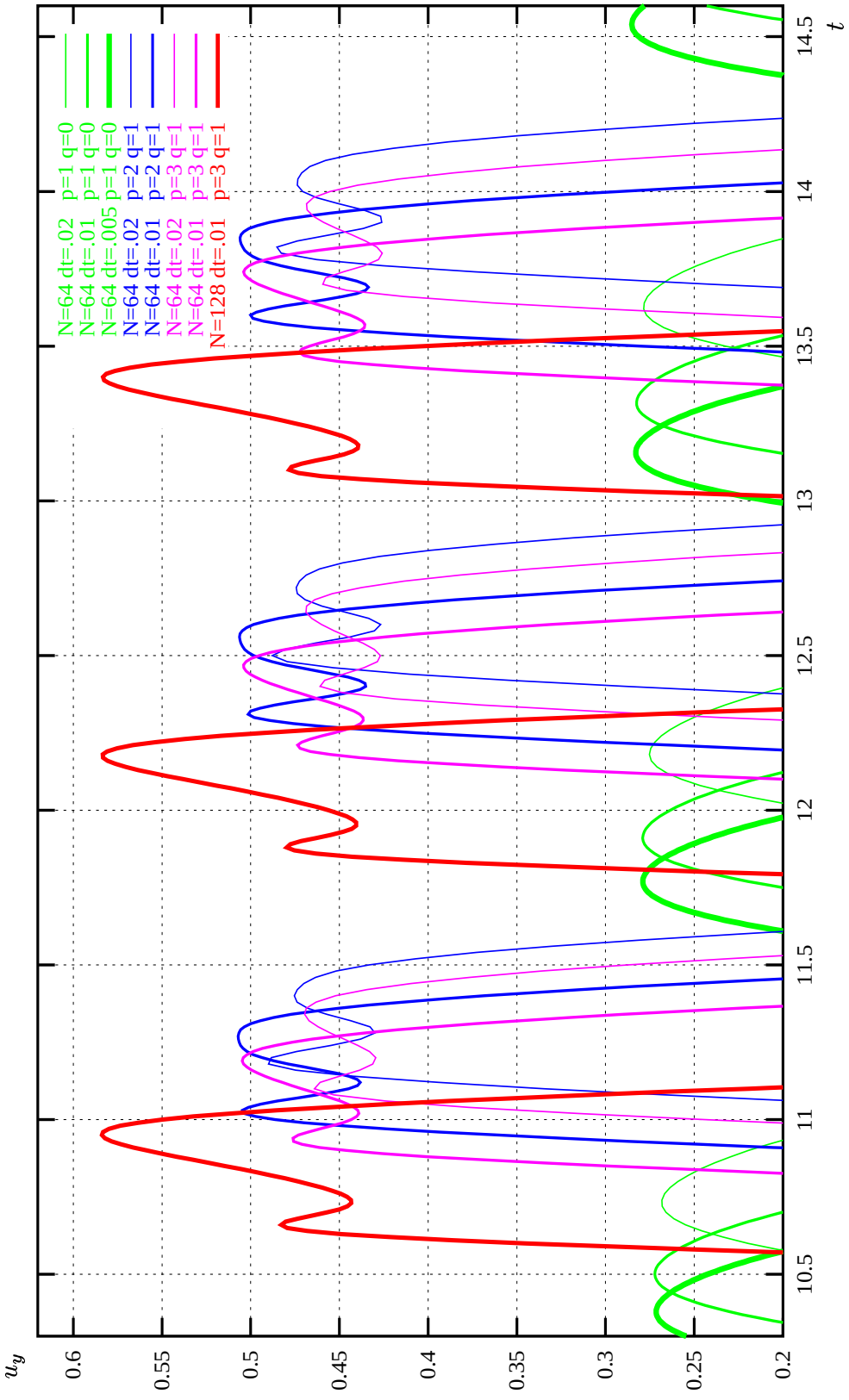


Figure 5.7: Magnification of a part of Fig. 5.6. In Fig. 5.7, not only different interpolation orders, but also different time step sizes are compared. The colours (indicating the numerical scheme) coincide with those of Fig. 5.6.

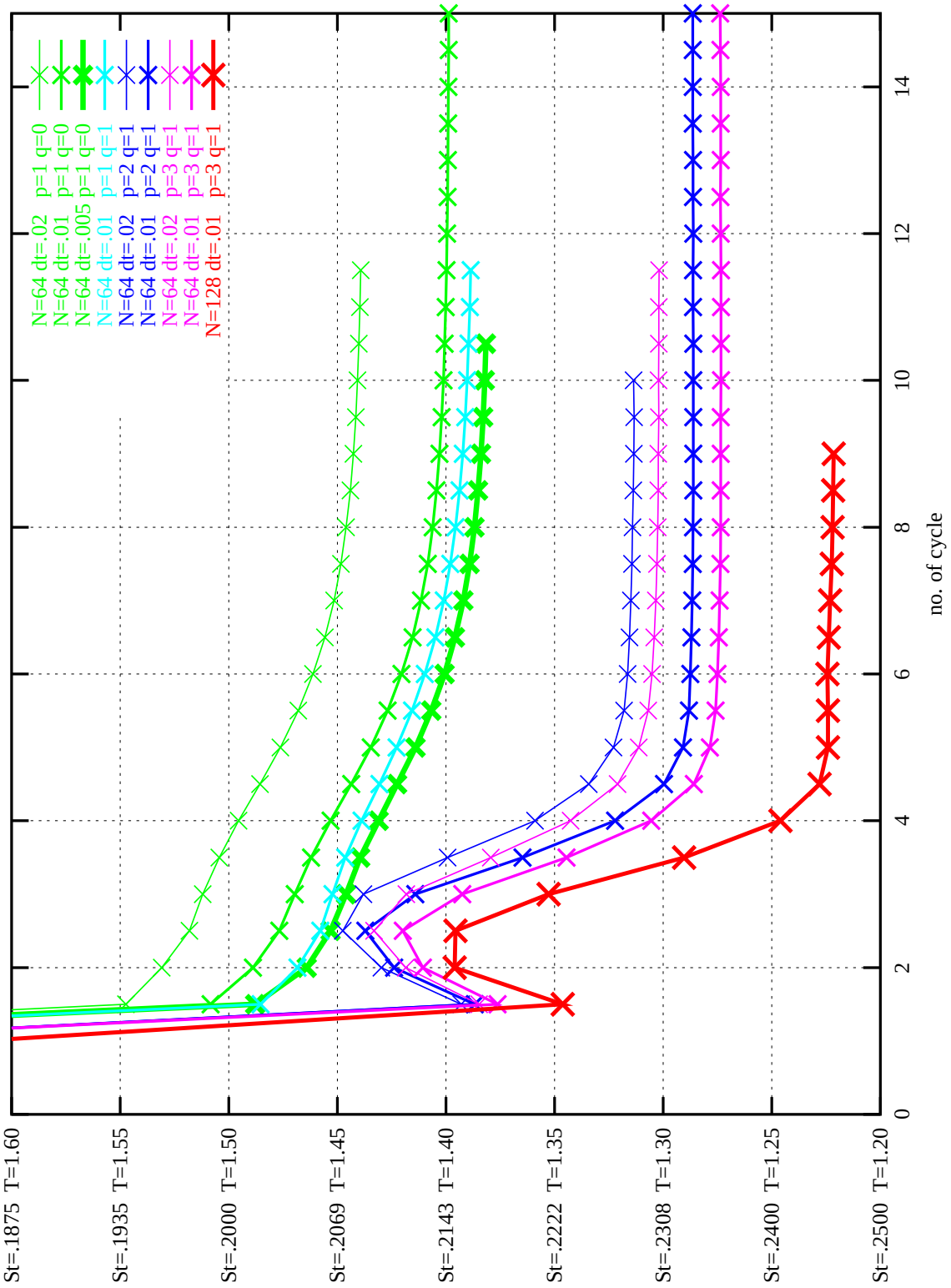


Figure 5.8: Cycle length T (and the related Strouhal number) as a function of the cycle number.

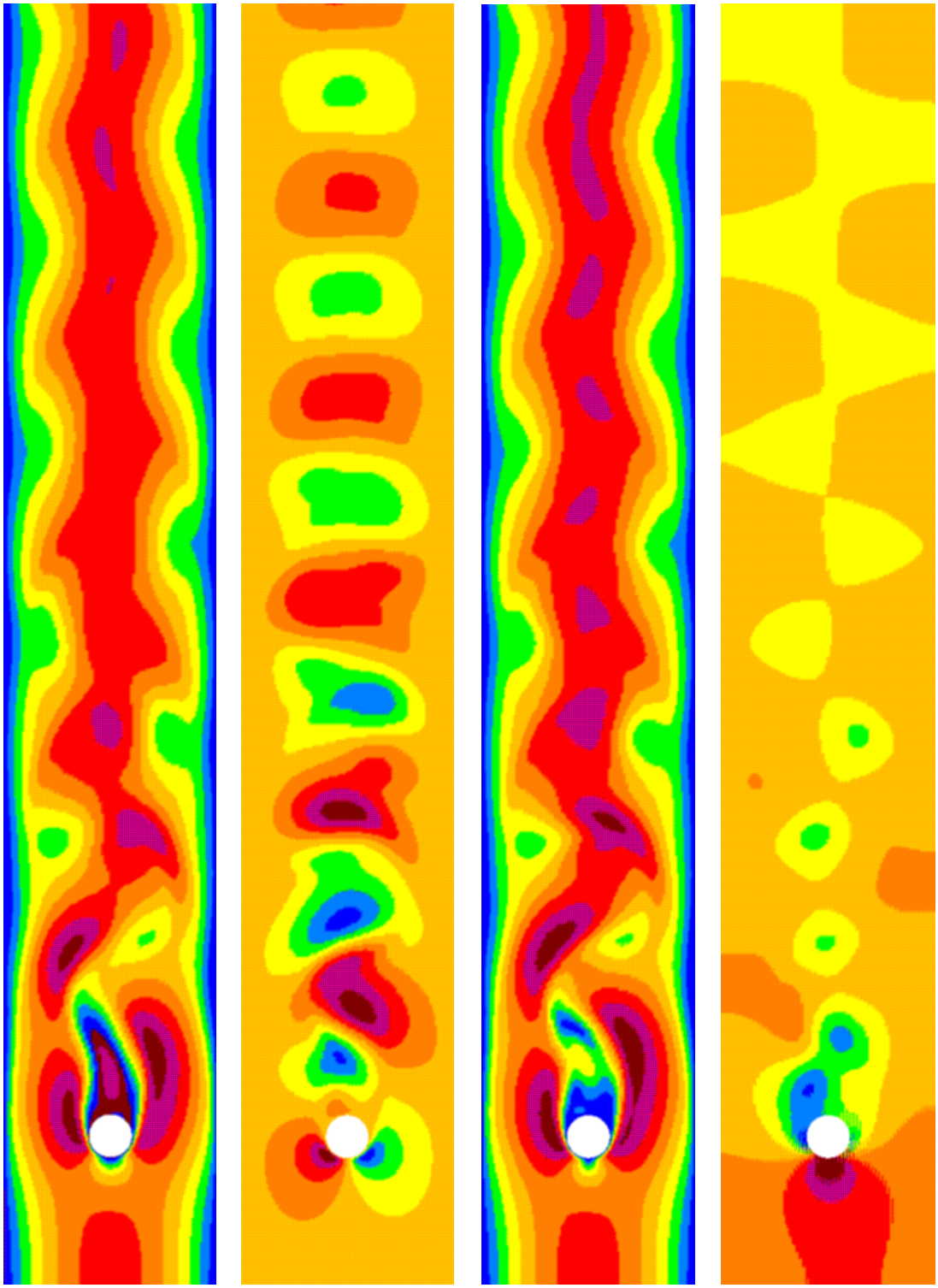


Figure 5.9: From left to right: the x-, the y-component, the norm of the velocity and the vorticity for $N = 128$, $\Delta t = 0.01$, $t = 13.1$. For the first three columns, each colour maps a range of 0.152; for the last column, a range of 0.12.

5.2.3 Channel flow past a cylinder II

In this section, another computation of a flow past a cylinder is presented where the parameters are adapted to the benchmark computation introduced in [46][9]. The computational domain and its substructuring into subdomains is the same as in Sec 5.2.2, but now, the circular obstacle is of diameter 0.244 and its center is positioned 0.49 from one channel wall and from the inlet. At the inflow, a Poiseuille profile with $U_{mean} = 1.0$ is prescribed for the x-component of the velocity. The Reynolds number is $Re = 100$ now.

Due to the asymmetric position of the obstacle the periodic vortex shedding establishes rapidly; no disturbance of the inflow profile is needed. In Fig. 5.12, the x-, the y-component, the norm of the velocity and the pressure for $N = 128$, $\Delta t = 0.005$ at time $t = 11.8$ are displayed. The right hand part of Fig. 5.13 shows the accompanying vorticity field.

Although a coarse spatial discretization on the FE domain (no local refinement close to the obstacle) and the first order time scheme was used, we found a Strouhal number of $St = 0.272$ for $N = 128$ ($St = 0.259$ for $N = 64$). These results were confirmed by a computation using a parallel code with Chebyshev spectral solvers on *all* subdomains, modelling the obstacle by a penalty method [25] and making use of a second order time scheme, which gave a Strouhal number of 0.288 [43]. the benchmark computation [46] gives $St = 0.300 \pm 0.005$ to be the 'exact' value. A further improvement of the numerical results can be expected by implementation of a second order time scheme, by improvement of the outflow boundary conditions and by grid refinement close to the obstacle.

[9] The parameters coincide up to a scaling in space and time.

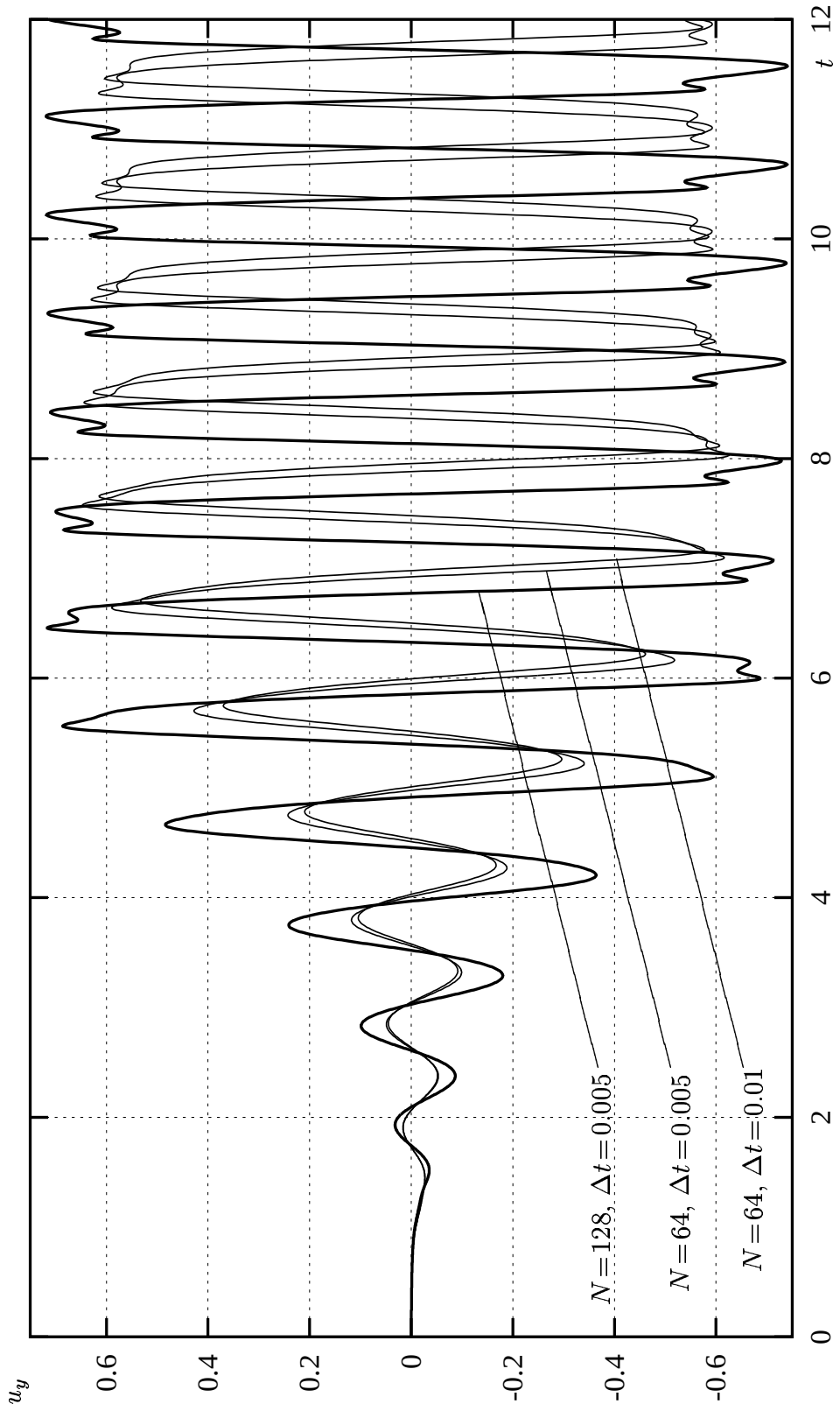


Figure 5.10: The transverse component $u_y(t)$ of the velocity at a fixed point on the symmetry axis behind the obstacle for different discretization parameters.

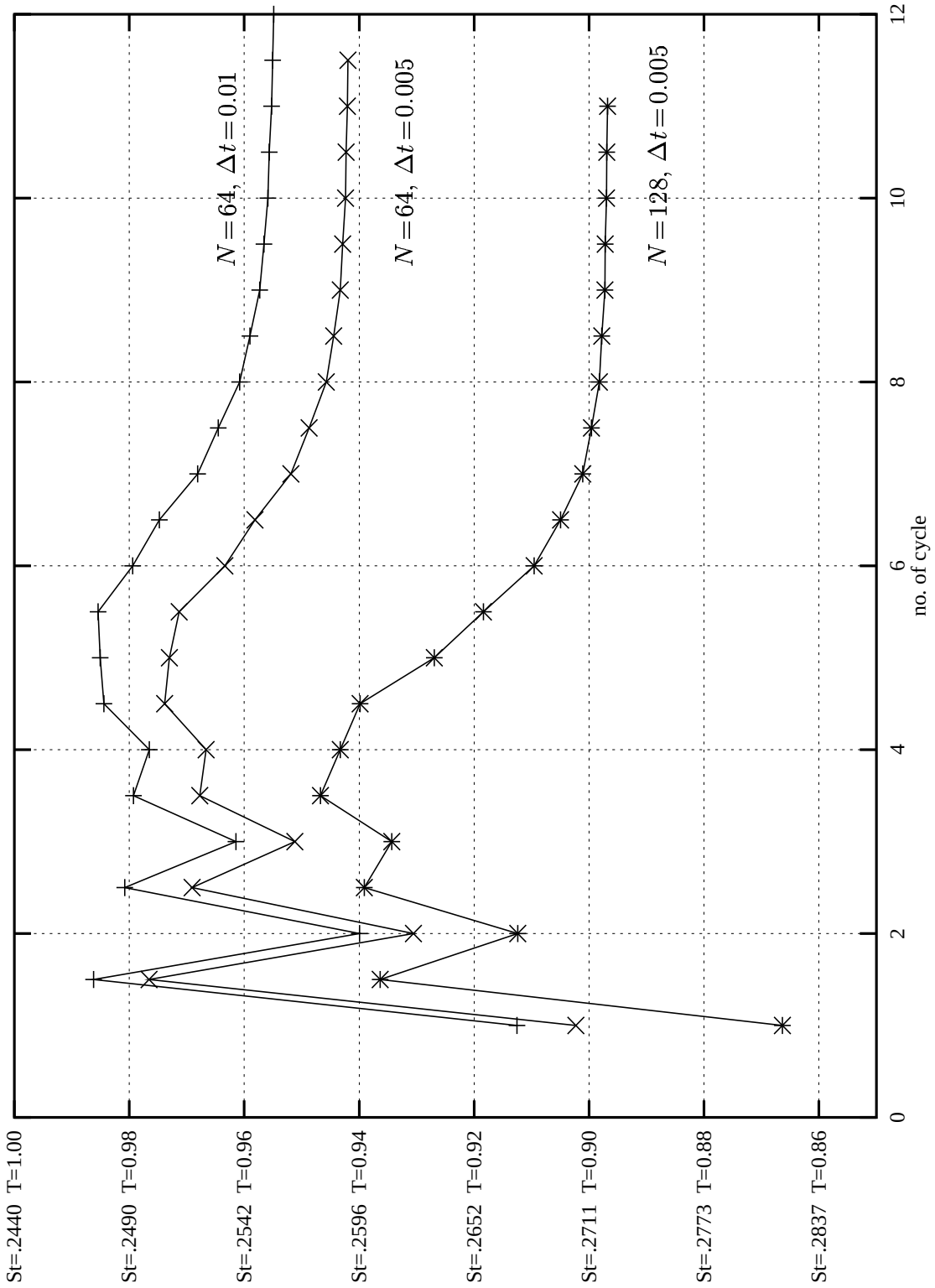


Figure 5.11: Cycle length T (and the related Strouhal number) as a function of the cycle number.

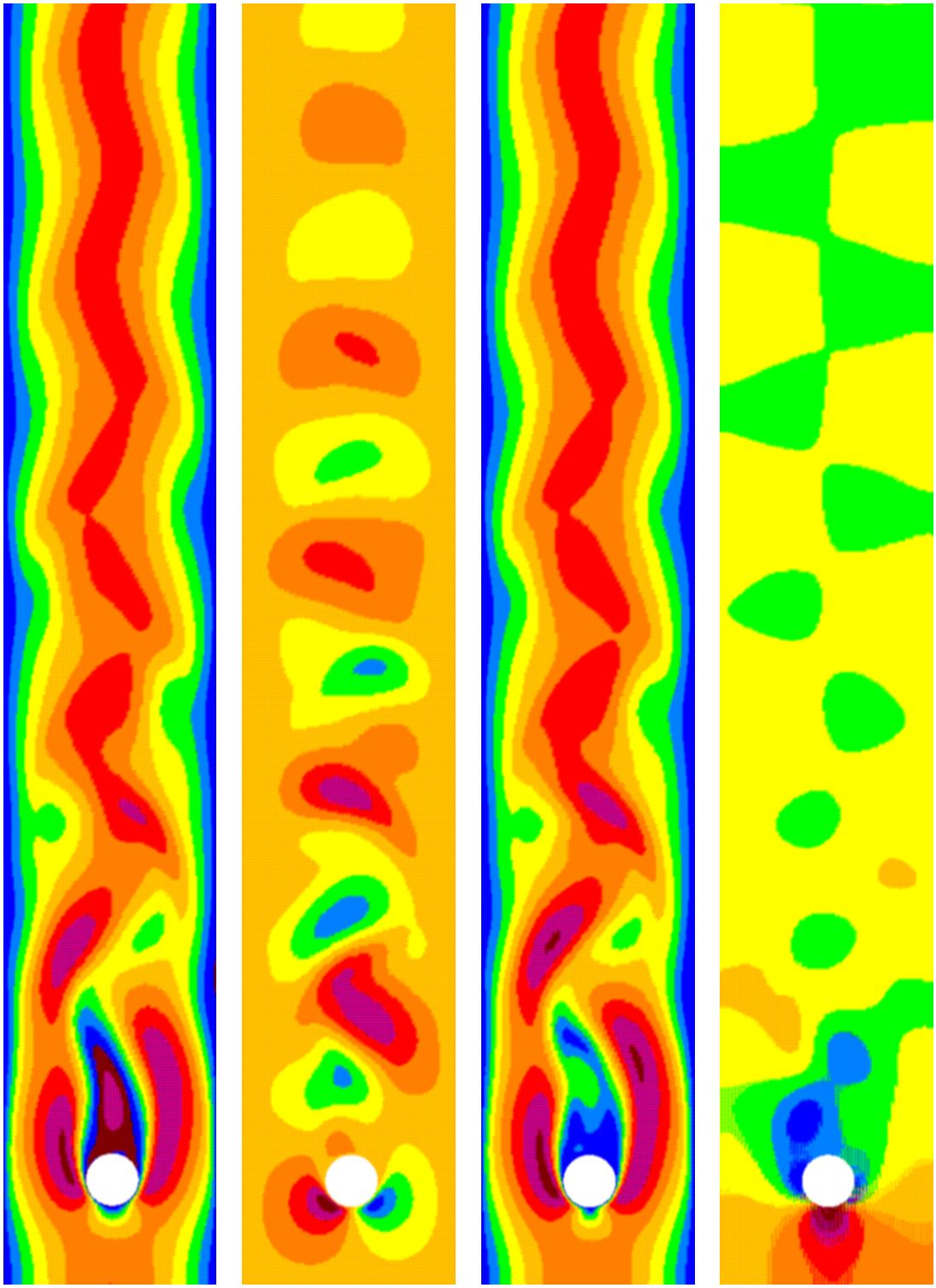


Figure 5.12: From left to right: the x-, the y-component, the norm of the velocity and the pressure for $N = 128$, $\Delta t = 0.005$, $t = 11.8$. For the first three columns, each colour maps a range of 0.25; for the last stripe, a range of 0.3. The time t within the vortex shedding period was chosen such that the pictures are comparable to Fig. 5.9.

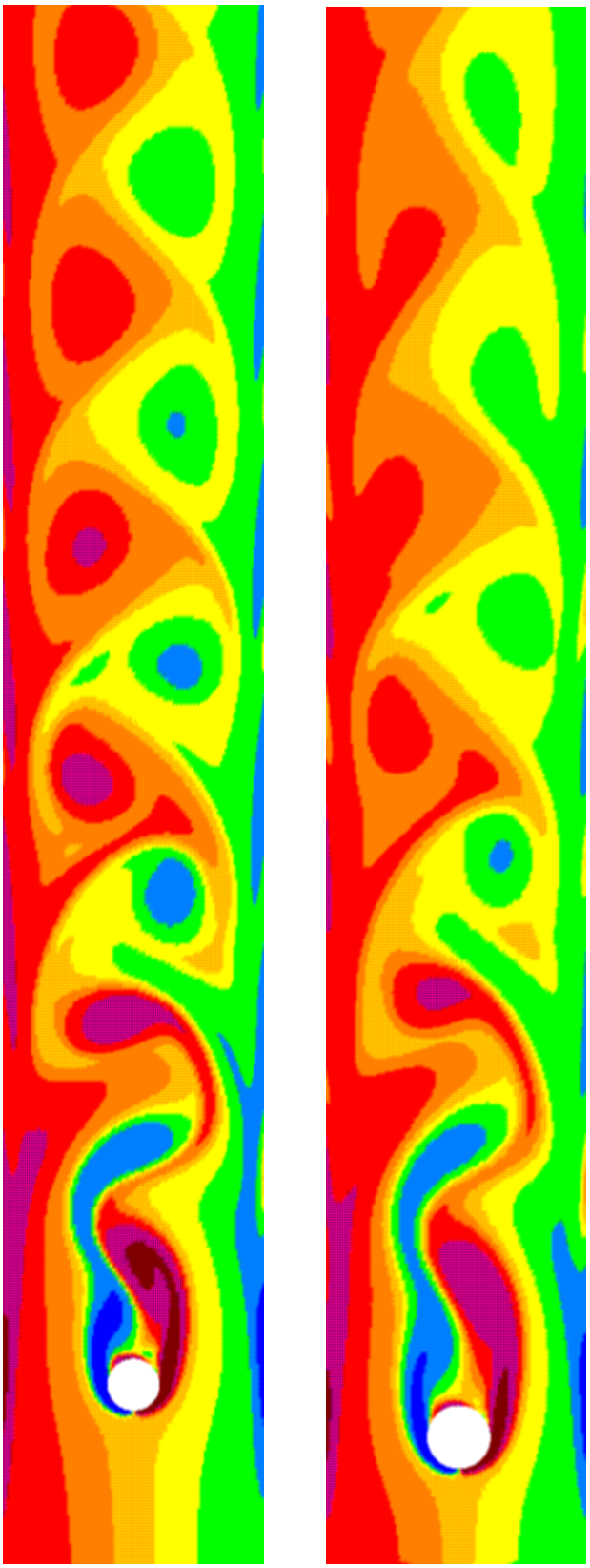


Figure 5.13: From left to right: The vorticity for the flow past cylinder I and flow past cylinder II. See captions of Fig. 5.9 and of Fig 5.12 for the program parameters. Blue and green indicate a circulation in the negative sense, red and violet circulation in the positive sense. A nonlinear colour scale is used to muffle the positive and negative extreme values.

5.3 Outlook

The results presented in this chapter show that CGBI is a well suited parallelization tool for the coupling of FE solvers and spectral collocation solvers, and the highly accurate computation of a Navier-Stokes flow. A high spatial accuracy is reached on the spectral domains. Presently, the FE solver uses approximately the same amount of grid points as the spectral solver. To balance the error and also the processor load, we intend to use a refined FE mesh (which requires a preconditioning or a multigrid method for the local FE problem) coupled with rather coarse Gauss-Lobatto grids for future computations.

The implementation of a higher order time scheme is presently under investigation. Without much modification of the code, a higher order method of characteristics (i.e. a higher order of the approximation of the Lagrangian derivative (5.1), (5.3)) is possible. Another possibility which is under current investigation is the introduction of a semi-Lagrangian method [42] which avoids costly spatial interpolations. However, the semi-Lagrangian method leads to a restriction of the timestep size used in the transport equation, but we intend to overcome this problem by 'subcycling': In [42] at least for a *sequential* Navier-Stokes solver based on the Chebyshev spectral method, it was shown that the stability restraint concerns only the subtimestep size, but not the global time step Δt.

Another possibility is to replace the collocation-type computation of the interface jumps by *weak* formulations as they are used by the *mortar methods* (e.g. [2]).

In my opinion the most important task for the next future is the investigation of CGBI for domain decompositions with *interior crosspoints*. The most important question is whether a preconditioner acting only on the interfaces then still leads to a condition number independent of the discretization parameter, and number/size of the subdomains. As mentioned in Sec. 2.9, the FETI method does *not* meet this desirable property.

Bibliography

[1] Alt, H.W.: *Lineare Funktionalanalysis*, 3. Aufl., Berlin, Heidelberg, New York, Springer 1999

[2] Ben Belgacem, F; Maday, Y: *Coupling spectral and finite elements for second order elliptic three-dimensional equations*, SIAM J. Numer. Anal., Vol. 36, No. 4, pp. 1234-1263, 1999

[3] Bhardwaj, M.; Day, D.; Farhat, C.; Lesoinne, M.; Pierson, K.; Rixen, D.: *Application of the FETI method to ASCI problems - scalability results on 1000 processors and discussion of highly heterogeneous problems*, Int. J. Numer. Methods Engin., Vol. 47, pp. 513-535 (2000)

[4] Bjørstad, P.E.; Widlund, O.B.: *Iterative methods for the solution of elliptic problems on regions partitioned into substructures*, SIAM J. Numer. Anal., Vol. 23, No. 6, pp. 1097-1120 (1986)

[5] Blazy, S.; Borchers, W.; Dralle, U.: *Parallelization Methods for a Characteristic's Pressure Correction Scheme*, Flow simulation with high-performance computers, II, 305–321, Notes Numer. Fluid Mech., 52, Vieweg, Braunschweig, 1996

[6] Blazy, S.; Borchers, W.; Kräutle, S.; Rautmann, R.; Roß, N.; Wielage, K.: *Strömungsberechnung: Eine Herausforderung für Mathematik und Informatik*, ForschungsForum Paderborn, pp. 100-104, 3/2000

[7] Borchers, W.; Forestier, M.Y.; Kräutle, S.; Pasquetti, R.; Peyret, R.; Rautmann, R.; Roß, N.; Sabbah, C.: *A Parallel Hybrid Highly Accurate Elliptic Solver for Viscous Flow Problems*, Numerical Flow Simulation I, Notes on Num. Fluid Mech. Vol. 66, Hirschel (ed.), pp. 3-24, Springer Verlag 1998

[8] Borchers, W.; Kräutle, S.; Pasquetti, R.; Rautmann, R.; Roß, N.; Wielage, K., Xu, C.J.: *Towards a Parallel Hybrid Highly Accurate Navier-Stokes Solver*, Numerical Flow Simulation II, Notes on Num. Fluid Mech. Vol. 75, CNRS-DFG Collab. Research Progr., Results 1998-2000, Hirschel (ed.), pp. 3-18 Springer Verlag 2001

[9] Bramble, J.H.; Pasciac, J.E.; Schatz, A.H.: *The construction of preconditioners for elliptic problems by substructuring I*, Math. Comp. 47 (1986), pp. 103-134

[10] Bramble, J.H.; Pasciac, J.E.; Schatz, A.H.: *The construction of preconditioners for elliptic problems by substructuring II*, Math. Comp. 49 (1987), pp. 1-16

[11] Brenner, S.C.: *The condition number of the Schur complement in domain decomposition*, Numer. Math. (1999) 83, pp. 187-203

[12] Canuto, C.; Hussaini, M.Y.; Quarteroni, A.; Zang, T.A.: *Spectral Methods in Fluid Dynamics*, Springer Verlag, New York 1988

[13] Childs, P.N.; Morton, K.W.: *Characteristic Galerkin methods for scalar conservation laws in one dimension*, SIAM J. Numer. Anal., Vol. 27, No. 3, pp. 553-594 (1990)

[14] Douglas, J.; Russel, T.: *Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures*, SIAM J. Numer. Anal., Vol. 19, No. 5, pp. 871-885

[15] Dryja, M.: *A capacitance matrix method for Dirichlet problem on polygon region*, Numer. Math., 39, pp. 51-64 (1982)

[16] Ehrenstein, U.: *Méthodes spectrales de résolution des équations de Stokes et de Navier-Stokes. Application à des écoulements de convection double-diffusive*, Doctoral thesis, Nice, 1986

[17] Ehrenstein, U.; Peyret, R.: *A Chebyshev collocation method for the Navier-Stokes equations with application to double diffusive convection*, Int. Journal for Numerical Methods in Fluids 9 (1989), pp. 427-452.

[18] Farhat, C: *A method of finite element tearing and interconnecting and its parallel solution algorithm*, Int. J. Num. Methods Engin., Vol. 32, pp. 1205-1227 (1991)

[19] Farhat, C; Roux, F.-X.: *An unconventional domain decomposition method for an efficient parallel solution of large-scale finite element systems*, SIAM J. Sci. Stat. Comput., Vol. 13, No. 1, pp. 379-396 (1992)

[20] Farhat, C.; Chen, P.-S.; Roux, F.-X.: *The dual Schur complement method with well-posed local Neumann problems: Regularization with perturbed Lagrangian formulation*, SIAM J. Sci. Comput., Vol. 14, No. 3, pp. 752-759 (1993)

[21] Farhat, C; Roux, F.-X.: *The dual Schur complement method with well-posed local Neumann problems*, Contemporary Mathematics, Vol. 157, pp. 193-201 (1994)

[22] Farhat, C; Crivelli, L; Roux, F.-X.: *A transient methodology for large-scale parallel implicit computations in structural mechanics*, Int. J. Numer. Methods Engin., Vol. 37, pp. 1945-1975 (1994)

[23] Favini, A.: *Sulla interpolazione di certi spazi di Sobolev con peso*, Rend. Semin. Mat. Univ. Padova 50 (1973), pp. 223-249

[24] Forestier, M.Y.; Pasquetti, R.; Peyret, R.; Sabbah, C.: *Spatial Development of Wakes using a Spectral Multi-Domain Method*, to be published

[25] Forestier, M.Y.; Pasquetti, R.; Peyret, R.: *Calculations of 3d wakes in stratified fluids*, ECCOMAS 2000, to be published

[26] Gilbarg, D.; Trudinger, N.S.: *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag Berlin Heidelberg New York 1977

[27] Goudjo, C.: *Problèmes aux limites dans les espaces de Sobolev avec poids*, Boll. Unione Mat. Ital. 8 (1973), pp. 468-493

[28] Klawonn, A; Widlund, O.B.: *FETI and Neumann-Neumann iterative substructuring methods: Connections and new results*, Comm. Pure Appl. Math., Vol. 54, pp. 57-90 (2001)

[29] Kräutle, S.: *A Higher Order Characteristic's Method for the Transport Equation*, Third Seminar on Euler and Navier-Stokes Equations (Proceedings), Institute of Thermomechanics AS CR, Prague, 1998

[30] Kräutle, S.; Wielage, K.: *The CGBI method for viscous channel flows and its preconditioning*, Nonlinear Analysis: Theory, Methods & Applications, 47 (6) (2001) pp. 4193-4203

[31] Kräutle, S.; Wielage, K.: *Numerical results for the CGBI method to viscous channel flow*, Navier Stokes Equations: Theory and Numerical Methods (proceedings), R. Salvi (ed.), pp. 247-255 (2001)

[32] Lax, P.D.: *On the Stability of Difference Approximations to Solutions of Hyperbolic Equations With Variable Coefficients*, Comm. Pure Appl. Math., Vol.XIV, pp. 497-520 (1961)

[33] Lax, P.D.; Nirenberg, L.: *On Stability for Difference Schemes; a Sharp Form of Gårding's Inequality*, Comm. Pure Appl. Math., Vol. XIX, No. 4, pp. 473-492 (1966)

[34] Lions, J.L.; Magenes, E.: *Non-homogeneous boundary value problems and applications I*, Springer Verlag, Berlin-Heidelberg-New York, 1982

[35] Mansfield, L.: *On the conjugate gradient solution of the Schur complement system obtained from domain decomposition*, SIAM J. Numer. Anal. Vol. 27, No. 6, pp. 1612-1620 (1990)

[36] Marchuk, G.I.: *Methods of Numerical Mathematics*, Springer Verlag, New York Heidelberg Berlin, 1975

[37] Morgan, K.; Periaux, J.; Thomasset, F. (Eds.): *Analysis of Laminar Flow Over a Backward Facing Step*, Notes on numerical fluid mechanics, Vol. 9, Vieweg, Braunschweig, Wiesbaden 1984

[38] Morton, K.W.; Süli, E.: *Evolution-Galerkin methods and their supraconvergence*, Numer. Math. 71, pp. 331-355 (1995)

[39] Morton, K.W.; Priestley, A.; Süli, E.: *Stability of the Lagrange-Galerkin Method with non-exact Integration*, M2AN Vol. 22, no. 4, pp. 625-653 (1988)

[40] Park, K.C.; Justino, M.R.; Felippa, JR. & C.A.: *An algebraically partitioned FETI method for parallel structural analysis: Algorithm description*, Int. J. Numer. Methods Engin., Vol. 40, pp. 2717-2737 (1997)

[41] Pasquetti, R.; Sabbah, C.: *A Divergence-free Multi-Domain Spectral Solver of the Navier-Stokes Equations in Geometries of High Aspect Ratio*, J. Comput. Phys. 139, No.2, pp. 359-379 (1998)

[42] Pasquetti, R.; Xu, C.: *On the efficiency of semi-implicit and semi-Lagrangian spectral methods for the calculation of incompressible flows*, Int. J. for Num. Methods in Fluids, to be published

[43] Pasquetti, R: private communication

[44] Pironneau, O.: *On the Transport-Diffusion Algorithm and Its Applications to the Navier-Stokes Equations*, Numer. Math. 38, pp. 309-332 (1982)

[45] Quarteroni, A.; Valli, A.: *Domain Decomposition Methods for Partial Differential Equations*, Clarendon Press, Oxford 1999

[46] Schaefer, M.; Turek, S.: *Benchmark computations of laminar flow around a cylinder. (With support of F. Durst, E. Krause and R. Rannacher.),* Hirschel, E.H. (ed.), Flow simulation with high-performance computers II. DFG priority research programme results 1993-1995. Vieweg, Wiesbaden. Notes Numer. Fluid Mech. 52, 547-566 (1996)

[47] Schwarz, H.R.: *Numerische Mathematik*, B. G. Teubner, Stuttgart 1993, 3. Aufl.

[48] Süli, E.: *Convergence and Nonlinear Stability of the Lagrange-Galerkin Method for the Navier-Stokes Equations*, Numer. Math. 53, pp. 459-483 (1988)

[49] Süli, E.: *Stability and Convergence of the Lagrange-Galerkin Method with non-exact Integration*, The mathematics of finite elements and applications, VI (Uxbridge, 1987), pp. 435-442, Academic Press, London, 1988

[50] Süli, A.; Ware, A.: *A spectral method of characteristics for hyperbolic problems*, SIAM J. Numer. Anal., Vol. 28, No. 2, pp. 423-445 (1991)

[51] Temam, R.: *Sur l'approximation de la solution des équations de Navier-Stokes par la méthode des pas fractionaires (II)*, Arch. Rational Mech. Anal. 32, pp. 377-383 (1969)

[52] Temam, R.: *Navier-Stokes Equations*, Rev. ed., North-Holland Publishing Company, Amsterdam - New York - Oxford 1979

[53] Tezaur, R.: *Analysis of Lagrange multiplier based domain decomposition.* Doctoral dissertation, University of Colorado at Denver, Denver, 1998, `http://www-math.cudenver.edu/graduate/thesis/rtezaur.ps.gz`

[54] Triebel, H.: *Interpolation Theory, Function Spaces, Differential Operators*, North-Holland Publishing Company Amsterdam New York Oxford 1978

[55] Velte, W.: *Direkte Methoden der Variationsrechnung*, Teubner, Stuttgart 1976

[56] Walter, W.: *Differential- und Integral-Ungleichungen*, Springer-Verlag, Berlin, Göttingen, Heidelberg, New York, 1964

[57] Wloka, J.: *Partielle Differentialgleichungen*, Teubner, Stuttgart 1982

Zusammenfassung in deutscher Sprache

Das Thema dieser Dissertation ist Beschreibung und Erforschung der *Conjugate Gradient Boundary Iteration* (CGBI) Methode. CGBI ist eine Gebietszerlegungsmethode zur Parallelisierung elliptischer symmetrischer partieller Differentialgleichungen und wurde vorgeschlagen von Borchers [5].

Die Lösung der globalen partiellen Differentialgleichung kann sehr leicht parallel gefunden werden, sobald ihre zugehörigen Randbedingungen auf den künstlichen Rändern zwischen den Subgebieten bekannt sind. CGBI ermittelt diese Randbedingungen vom Neumannschen Typ ('natürliche Randbedingungen') mittels einer CG (conjugate gradient)-Iteration. Die Iteration selbst sowie die in Kapitel 3 konstruierten Vorkonditionierer arbeiten nur auf den Rändern der Subgebiete; daher der Name 'boundary iteration'. In jedem CG-Schritt muss in jedem Subgebiet jeweils ein lokales Problem, das von den übrigen lokalen Problemen unabhängig ist, gelöst werden.

CGBI ermöglicht die *Kopplung* verschiedener lokaler Löser, die auf den verschiedenen Teilen des Rechengebietes arbeiten, auch unter Verwendung nichtkonformer Gitter. So können auf kompliziert geformten Teilen des Rechengebietes Finite Elemente (FE) Löser zum Einsatz kommen, wohingegen auf rechteckigen Subgebieten hocheffiziente Tschebyschev-Spektrallöser verwendet werden.

Als Anwendung von CGBI wird in Kapitel 5 ein paralleler Navier-Stokes-Löser konstruiert. Dieser zerlegt jeden Zeitschritt mit Hilfe der Druck-Korrekturmethode von Temam & Chorin in elliptische und ein hyperbolisches Problem. Die elliptischen Probleme (Poisson-Gleichung für den Druck, Helmholtz-Resolventengleichung für die Geschwindigkeitskomponenten) werden mittels CGBI, das hyperbolische durch ein Charakteristiken-Verfahren gelöst.

In Kapitel 2 wird der theoretische Hintergrund der CGBI-Methode dargestellt. Dabei wurde besonderen Wert auf die konsequente Verwendung der *schwachen* Formulierungen von Randwertproblemen gelegt, denn nur in dieser Formulierung sind die auftretenden Probleme tatsächlich lösbar. Ferner macht die Verwendung der schwachen Formulierungen deutlich, in welchen Funktionenräumen nach Lösungen gesucht werden muss, und sie zeigt auf, auf welche Weise effiziente Vorkonditionierer zu konstruieren sind (Kap. 2.2). Es stellt sich heraus, dass Diskretisierungen der Wurzel des negativen Laplace-Operators als Vorkonditionierer geeignet sind.

Kapitel 3 beschäftigt sich mit der Konstruktion solcher Vorkonditionierer. Ausgangspunkt ist immer der *exakte* vorzukonditionierende Operator, zu dessen Inversen mit Hilfe von Eigenvektorbasen ein spektral äquivalenter Vorkonditio-

nierungsoperator gesucht wird. Anschließend wird dieser *diskretisiert*. Falls die künstlichen Ränder, die bei der Gebietszerlegung auftreten, ein äquidistantes Gitternetz aufweisen, so ist diese Diskretisierung des Vorkonditionierers sehr leicht. Es kann gezeigt werden, dass die resultierende Konditionszahl bei Gebietszerlegungen ohne 'innere Kreuzungspunkte' unabhängig von der Anzahl der Subgebiete und vom Diskretisierungsparameter ist. Bei Verwendung einer vereinfachten Geometrie können explizite Schranken für die Konditionszahl angegeben werden.

Der Tschebyschev-Spektrallöser verwendet allerdings kein äquidistantes, sondern ein Gauß-Lobatto-Gitter. Mittels der Interpolationstheorie gewichteter Sobolev-Räume wird der Gauß-Lobatto-Fall auf den äquidistanten Fall zurückgeführt (Kap. 3.1.3). Der Nachweis, dass der resultierende Gauß-Lobatto-Vorkonditionierer Konditionszahlen unabhängig vom Diskretisierungsparameter erzeugt, gelingt allerdings nur für den Fall von Dirichlet-Randdaten. Nichtsdestotrotz wurde durchweg, sofern das Seitenverhältnis der Subgebiete nicht zu schlecht ist, eine Fehlerreduktion von etwa einer vollen Zehnerpotenz pro CGBI-Schritt beobachtet. Der Aufwand der Vorkonditionierung ist, verglichen mit dem der lokalen Löser, vernachlässigbar gering.

Alternative Vorkonditionierer, basierend auf Bandmatrizen, oder, um das Problem der Abhängigkeit der Konditionszahl vom Seitenverhältnis der Subgebiete zu lösen, basierend auf Faltungskernen, werden in Kap. 3.3 und 3.4 vorgestellt und untersucht.

CGBI wird verschiedensten numerischen Tests unterzogen, die auch die Kopplung verschiedener lokaler Löser (FDM–Spektrallöser, FEM–Spektrallöser) auf verschiedenen Gittern einschließen. Diese Tests belegen, dass CGBI ein robustes Verfahren zur effizienten Lösung elliptischer Probleme darstellt.

CGBI hat große Ähnlichkeit zu dem seit den frühen 90ern entwickelten FETI-Verfahren von Farhat & Roux (Kap. 2.9); beide nutzen *natürliche* Randbedingungen an den künstlichen Rändern. Ein wesentlicher Unterschied ist, dass unser Verfahren Vorkonditionierer benutzt, die ausschließlich auf den Rändern agieren und vernachlässigbar wenig Rechenzeit benötigen, wohingegen die FETI-Vorkonditionierer zusätzliche Probleme auf den Subgebieten lösen, was den Rechenaufwand pro Iteration etwa verdoppelt. Ein detaillierter Vergleich der Effizienz von CGBI und FETI kann jedoch erst dann erfolgen, wenn CGBI auf den Fall innerer Kreuzungspunkte ausgedehnt ist. Die wesentliche Frage ist, ob in diesem Fall die Unabhängigkeit der Kondition vom Diskretisierungsparameter und von Anzahl und Größe der Subgebiete erhalten bleibt. Bei FETI jedenfalls ergibt sich eine logarithmische Abhängigkeit der Konditionszahl von der Größe der lokalen Probleme.

Um das beim Navier-Stokes-Löser auftretende hyperbolische Problem zu lösen, wurde ein Charakteristiken-Verfahren höherer Ordnung programmiert und theoretisch untersucht. Dieses Verfahren basiert auf der Berechnung von

Charakteristiken startend an den Gitterpunkten. Die dabei nötige Auswertung des Stömungsfeldes geschieht mit polynomieller Interpolation in der Zeit und stückweise polynomieller Interpolation im Raum. Diese Methode benötigt *nicht* das Auflösen globaler Gleichungssysteme und ist u.a. deshalb leicht zu parallelisieren. Es werden Fehlerabschätzungen und Stabilitätskriterien hergeleitet. Während die Stabilität bei Verwendung *linearer* räumlicher Interpolation trivialerweise immer gegeben ist, ist ihr Nachweis bei Verwendung höherer räumlicher Interpolation an Bedingungen gebunden. Auf äquidistanten Gittern reicht die Beschränktheit der Courant-Zahl. Das gleiche gilt für quasi-uniforme Gitter. Auf Gauß-Lobatto-Gittern hingegen kann die Stabilität nur unter starken Einschränkungen an die Zeitschrittweite gezeigt werden. Numerische Tests hingegen zeigen auch auf diesen Gittern keinerlei Stabilitätsprobleme.

Das letzte Kapitel dokumentiert einige Rechnungen des vollen 2d-Navier-Stokes-Lösers. Als Testprobleme dienen die Kanalstömung über eine Stufe ('backward facing step') sowie die Kanalströmung hinter einem zylinderförmigen Hindernis (Ausbildung einer Kàrmànschen Wirbelstraße). Neben einer grafischen Darstellung der resultierenden Strömungen werden die Länge des Rückströmbereichs (im Fall der ersten Rechengeometrie) und die Strouhal-Zahl (im Fall der zweiten Rechengeometrie) ermittelt und mit Werten aus der Literatur verglichen.

Lebenslauf

Am 21.06.1970	wurde ich geboren in Bad Driburg, NRW. Meine Eltern sind Jürgen Kräutle und Marlies Kräutle, geborene Gehle.
Aug. 76 - März 89	Besuch der Kath. Grundschule Brakel und des Gymnasiums Brede in Brakel mit Abiturabschluß (Note 1.3)
Juni 89 - Aug. 90	habe ich den Grundwehrdienst bei der Bundeswehr abgeleistet.
Okt. 90 - Mai 96	Diplomstudium Mathematik an der Uni-GH Paderborn mit Nebenfach Informatik
Am 14.05.96	bestand ich die Diplomprüfung mit der Note "mit Auszeichnung". Die Diplomarbeit trägt den Titel *"Approximationen der Navier - Stokes - Gleichungen mit finiten Differenzen"*. Ab dem 8. Semester Abhalten von Übungsgruppen.
Juni 96 - Sept. 97	arbeitete ich an der Uni-GH Paderborn als Wiss. Hilfskraft in dem DFG - Projekt "Hybrid Finite Element - Spectral Solver for Viscous Flow Problems" mit dem Ziel der Promotion. Währenddessen weiterhin Abhalten von Übungsgruppen (Analysis, Numerik gewöhnlicher Differentialgleichungen)
Okt. 97 - Apr. 00	Fortsetzung des DFG-Projektes und der Promotion am Institut für Angewandte Mathematik I der Friedrich-Alexander Universität Erlangen. Abhalten von Übungsgruppen (Mathematik für Elektrotechniker 1-4, Mathematik für Informatiker 1)
Mai 00 - Jan. 01	Forschungsaufenthalt auf Einladung von Prof. Peyret und Dr. Pasquetti am Laboratoire Dieudonné der Université de Nice - Sophia-Antipolis in Nizza, Frankreich; Finanzierung der 'Post-Doc'-Stelle durch die CNRS (Centre National de la Recherche Scientifique).
Feb. 01 - Sommer 01	Abschluss der Dissertation, Titel der Dissertationsschrift *"A parallel Navier-Stokes solver based on CGBI and the method of characteristics"*, Abgabe im Juni.
19.02.2002	mündliche Promotionsprüfung