High Order Finite Difference Approximations for Parabolic and Hyperbolic-Parabolic Problems with Variable Coefficients

Katharina Kormann
Martin Kronbichler

29th March 2006

Supervisor: Bernhard Müller
Reviewer: Jan Nordström

Abstract

We consider stable high order finite difference approximations for the parabolic term with a variable coefficient in a hyperbolic-parabolic equation. We present three different approaches to account for this problem. Hyperbolic-parabolic equations with variable coefficients arise in many application, for example when linearizing the Navier-Stokes equations. The parabolic term models diffusion and the hyperbolic term convection, which is why the resulting equation is often called convection diffusion equation.

The diffusion term is of self-adjoint form. It is desirable that a discretization of the parabolic term is also self-adjoint in order to maintain physical properties that arise from this form, i.e. a decrease in energy. Since the hyperbolic terms can be treated using summation by parts (SBP) operators that have been derived earlier, we devised SBP operators that can be combined with them. We suggest how a self-adjoint SBP operator that is strictly stable could look like. A simpler and more efficient way to approximate the diffusive term is to approximate all appearing derivatives separately using the known SBP operators. Here the self-adjoint form is not conserved and the approach yields a stable but not strictly stable operator.

As an alternative we propose a fourth order accurate operator based on finite elements. Using a technique called mass lumping, we obtain a diagonal mass matrix which is why we can interpret the operator as a finite difference method. This operator is strictly stable for the parabolic term, but requires hyperbolic terms to be approximated in the same way.

Contents

1 Introduction

Our objective is to find strictly stable high order finite difference approximations for the spatial part of the parabolic differential equation

$$\begin{aligned} &u_t(x,t) = (a(x,t)u_x(x,t))_x \\ &u(0,x) = f(x) \end{aligned} \tag{1.1}$$

in the domain $\Omega \times I$. The function a is assumed to be positive, i.e. $a(x,t) \geq a_{\min} > 0$.

In technical applications, the right hand side of (1.1) models diffusive processes, e.g. mass diffusion and heat conduction, and constitutes the diffusive part of the convection diffusion equation

$$\begin{aligned} &u_t(x,t) = (a(x,t)u_x(x,t))_x + b(x,t)u_x(x,t) + (c(x,t)u(x,t))_x \\ &u(0,x) = f(x) \end{aligned} \tag{1.2}$$

in the domain $\Omega \times I$. Hence a numerical method for solving (1.1) should be applicable to (1.2).

The method shall be of high order and the bandwidth of the operator shall be as small as possible in order to yield a consistent and efficient method. For a consistent method, convergence is equivalent to stability by the Lax-Richtmyer equivalence theorem.

Kreiss and Wu [9] showed that if a stable semi-discretized system is additionally discretized in time using a Runge-Kutta method, the fully discrete scheme will be stable as well. Carpenter et al., however, pointed out that the error might nevertheless grow exponentially in time, which can be avoided for strictly stable semi-discrete schemes [2].

Kreiss and Scherer [8] proposed to design finite difference methods that satisfy a summation by parts property using some specially defined discrete norms. Operators based on this idea are called summation by parts (SBP) operators. The summation by parts rule – a discrete analogon to integration by parts – yields an energy estimate, which guaranties stability of the difference scheme according to the theory developed in [7].

This idea has been refined by Strand [16] for hyperbolic systems, i.e. for approximations of the first derivative. Mattsson and Nordström [11] take up this idea and extend it to parabolic and mixed hyperbolic-parabolic systems by developing summation by parts operators based on the same norms as the SBP operators for the first derivative. A SBP operator does not handle the physical boundary data itself, which is why the boundary conditions have to be implemented additionally. There are different possibilities how to implement the boundary data [10], among which the Simultaneous Approximation Term (SAT) method is the most common one (cf. [3] for details about SAT).

However, in [11] only parabolic terms with constant coefficients are considered. The objective of this work is to devise SBP operators for a parabolic term with non-constant coefficients. There have been SBP operators based on different types of norms. We concentrate on diagonal norms here since they are most feasible in practical applications. We propose three different possibilities of how to design an operator discretizing $(au_x)_x$.

In order to distinguish the different ideas, we first give two different definitions of a summation by parts operator. In a strict sense a SBP operator shall fulfill a summation by parts rule where we can transfer one derivative from one vector to the other. As proposed in [11], it is however reasonable to weaken this strict definition to some extend.

Let $(\cdot,\cdot)_H$ be an inner product based on the diagonal matrix H. H should be the matrix that is needed to show stability for the SBP operators proposed in [16] and [11]. Denote the operator approximating $(au_x)_x$ with $Q(a)$.

Definition 1.1. *A difference operator $Q(a) = -H^{-1}P(a) + R(a)$, where $R(a)$ operates only on the boundary, approximating $\frac{\partial}{\partial x}a\frac{\partial}{\partial x}$ is a complete SBP operator for a self-adjoint parabolic term, if $P(a) = D_1^T HaD_1$, where D_1 is a consistent approximation of $\partial/\partial x$.*

This definition is reasonable since $Q(a)$ satisfies the following summation by parts rule

$$\begin{aligned}(v, Q(a)v)_H - \text{boundary terms} &= v^T HQ(a)v - \text{boundary terms} = -v^T P(a)v \\ &= -v^T D_1^T HaD_1 v = -(D_1 v)^T HaD_1 v \\ &= -(D_1 v, aD_1 v)_H \leq 0\end{aligned}$$

This property can be used to obtain an energy estimate

$$\frac{d}{dt}\|u\|_H = -v^T(P(a) + P(a)^T)v + \text{boundary terms} \leq \text{boundary terms}$$

The first equation shows that we essentially need $P(a) + P(a)^T \geq 0$ to obtain an energy estimate. This yields the weaker definition

Definition 1.2. *A difference operator $Q(a) = -H^{-1}P(a) + R(a)$, where $R(a)$ operates only on the boundary, approximating $\frac{\partial}{\partial x}a\frac{\partial}{\partial x}$ is a SBP operator for a self-adjoint parabolic term, if $P(a) + P(a)^T \geq 0$.*

These definitions are a generalization of the definitions given in [11] for $\partial^2/\partial x^2$. They are not very precise since they shall only give a rough idea of SBP operators and are motivated more carefully in the sequel.

In the following section we investigate the continuous problem in detail to give a basis for the numerical treatment of the problem.

In section 3 we discuss the possibility of applying the product rule on $(au_x)_x$, which yields $au_{xx} + a_x u_x$, where we can use the SBP operators for $\partial/\partial x$ and $\partial^2/\partial x^2$. Such a method is based on the weaker definition of an SBP operator. In contrast the operators presented in section 4 are complete summation by parts operators. In both cases we handle the physical boundary conditions with the SAT method. The first method leads to simpler stencils, on the other hand some stability problems arise.

In section 5 we use a different approach, namely we use the theory of finite element methods to devise the operator. Using a technique called mass lumping however, we obtain an operator which can be interpreted as a summation by parts difference operator. The advantage of this operator is that we can treat the physical boundary data in a natural way based on the finite element theory and thus avoid difficulties with the boundary treatment. In section 6 we sketch how this ansatz can be extended to the convection diffusion equation.

2 The Continuous Problem

2.1 Initial-Boundary Value Problem

When considering an initial-boundary value problem (IBVP) governed by the partial differential equation (1.2), we distinguish between two types of boundary conditions. In the case of *Dirichlet boundary conditions*, u is prescribed on $\partial\Omega \times I$. Then we get the following IBVP:

$$\begin{aligned} &u_t(x,t) = (a(x,t)u_x(x,t))_x + b(x,t)u_x(x,t) + (c(x,t)u(x,t))_x, \quad x \in \Omega,\ t \in I, \\ &u(x,0) = f(x), \\ &u(0,t) = g_0(t), \qquad u(1,t) = g_1(t), \end{aligned} \tag{2.1}$$

where $a(x,t) \geq a_{\min} > 0$ ($a_{\min}$ shall be some constant) and $I = [0,\infty)$. For our analysis we further put $\Omega = [0,1]$.

The other type of boundary conditions prescribes u_x on $\partial\Omega \times I$, a so-called *Neumann boundary condition*. Moreover, combinations of both types of boundary conditions are possible which yield an IBVP of the form

$$\begin{aligned} &u_t(x,t) = (a(x,t)u_x(x,t))_x + b(x,t)u_x(x,t) + (c(x,t)u(x,t))_x, \quad x \in \Omega,\ t \in I, \\ &u(x,0) = f(x), \\ &\beta_0 u(0,t) + u_x(0,t) = g_0(t), \qquad \beta_1 u(1,t) + u_x(1,t) = g_1(t), \end{aligned} \tag{2.2}$$

where $a(x,t) \geq a_{\min} > 0$ and $\Omega = [0,1],\ I = [0,\infty)$. Such boundary conditions are called *Robin boundary conditions*. Note that $\beta_0 = \beta_1 = 0$ yields Neumann boundary conditions.

2.2 Energy Estimate

The main requirement on an initial boundary value problem is well-posedness. This means that a unique solution shall exist and the solution shall depend continuously on initial and boundary data. A convenient way of showing well-posedness is the energy method. For this purpose we consider a certain norm of the solution, physically identified with an energy.

Let the inner product for real-valued functions $u, v \in \mathcal{L}^2(\Omega)$ be defined by $(u,v) = \int_\Omega u\,v\,\mathrm{d}x$ and the corresponding norm by $\|u\|^2 = (u,u)$.
Using the energy method, well-posedness is defined as follows (cf. [7]):

Definition 2.1. *We call problem* (2.1) *or* (2.2)*, respectively, with* $g_0 \equiv g_1 \equiv 0$ well-posed*, if for every* $f \in \mathcal{C}^\infty$ *that vanishes in a neighborhood of* $\partial\Omega$*, there are constants* α *and* K *that do not depend on* f *such that*

$$\|u(\cdot\,,t)\|^2 \leq K\mathrm{e}^{\alpha t}\|f\|^2 \tag{2.3}$$

Inhomogeneous problems can be transformed into homogeneous problems using some smooth function that satisfies the boundary conditions. In that way the inhomogeneous boundary data can be transformed to a forcing term. In order to get an energy estimate, some regularity requirements on the boundary data arise, i.e. g_0 and g_1 should be differentiable (cf. [7]). This reduction process can be avoided using the following definition.

Definition 2.2. *Problem* (2.1) *or* (2.2)*, respectively, is called* strongly well-posed*, if it is well-posed and*

$$\|u(\cdot\,,t)\|^2 \leq K\mathrm{e}^{\alpha t}\left(\|f\|^2 + \int_0^t \left(g_0(\tau)^2 + g_1(\tau)^2\right)\mathrm{d}\tau\right) \tag{2.4}$$

with some constants K *and* α *that do not depend on* f *and* g*.*

Not every problem allows getting such a stringent estimate. Using the energy method, for Dirichlet boundary conditions one can only show well-posedness and not strong well-posedness as we will see later. For well-posedness we will therefore require differentiable Dirichlet data.

To get an easier criterion for strong well-posedness, we require the differentiated form of (2.4) to hold:

$$\frac{\mathrm{d}}{\mathrm{d}t}\|u(\cdot\,,t)\|^2 \leq \alpha K \mathrm{e}^{\alpha t}\left(\|f\|^2 + \int_0^t \left(g_0(\tau)^2 + g_1(\tau)^2\right)\mathrm{d}\tau\right) + K\mathrm{e}^{\alpha t}\left(g_0(t)^2 + g_1(t)^2\right).$$

If $\alpha \geq 0$, we can replace the expression in the big brackets by $\|u(\cdot,t)\|^2$ using equation (2.4). For the second term we use the estimate $1 \leq \mathrm{e}^{\alpha t}$.
Thus, if we show

$$\frac{\mathrm{d}}{\mathrm{d}t}\|u(\cdot\,,t)\|^2 \leq \alpha\|u(\cdot\,,t)\|^2 + K\left(g_0(t)^2 + g_1(t)^2\right), \tag{2.5}$$

then (2.4) is satisfied and the problem is strongly well-posed.

We analyze the norm of the solution to (1.2). Taking the time derivative of $\|u\|$ leads to

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|u\|^2 &= (u_t, u) + (u, u_t)\\ &= ((au_x)_x + bu_x + (cu)_x, u) + (u, (au_x)_x + bu_x + (cu)_x)\\ &= 2\int_0^1 u((au_x)_x + bu_x + (cu)_x)\mathrm{d}x\end{aligned} \tag{2.6}$$

There are several possibilities to estimate the hyperbolic terms bu_x and $(cu)_x$. We will make use of two different estimates in the sequel. The first version will be used for the operators based on finite difference methods derived in the sections 3 and 4, while the second version for the operator based on finite elements in section 5.

2.2.1 Estimate for the Difference Methods

We assume that the derivatives of a, b and c exist in order to be able to use the identities $2u(cu)_x = c_x u^2 + (cu^2)_x$ and $2ubu_x = (bu^2)_x - b_x u^2$ and apply integration by parts on $(au_x)_x$. (2.6) becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}\|u\|^2 = \int_0^1 (-b+c)_x u^2 \mathrm{d}x - 2\int_0^1 a u_x^2 \mathrm{d}x + u(2au_x + (b+c)u)\Big|_0^1. \tag{2.7}$$

The the first integral can be estimated by

$$\left|\int_0^1 (-b+c)_x u^2 \mathrm{d}x\right| \leq \|(-b+c)_x\|_\infty \int_0^1 u^2 \mathrm{d}x = \|(-b+c)_x\|_\infty \|u\|^2,$$

while the second integral in (2.7) is bounded by

$$-\int_0^1 a u_x^2 \mathrm{d}x \leq -a_{\min}\|u_x\|^2. \tag{2.8}$$

Inserting these estimates into (2.6) we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\|u\|^2 \leq -2a_{\min}\|u_x\|^2 + \|(-b+c)_x\|_\infty \|u\|^2 + u(2au_x + (b+c)u)\Big|_0^1. \tag{2.9}$$

Concerning the boundary terms, we distinguish between Dirichlet and Robin boundary conditions.

Dirichlet Boundary Conditions.
We consider the term $u(2au_x+(b+c)u)\big|_0^1$ in (2.9) and insert $u(0,t)=g_0(t)$ and $u(1,t)=g_1(t)$. We set $d(x,t)=\frac{1}{2}(b(x,t)+c(x,t))$

$$u(au_x+du)\Big|_0^1 = g_1(au_x+dg_1)\big|_1 - g_0(au_x+dg_0)\big|_0$$

When trying to get an estimate of the form (2.5), we have to use some technique to split the terms $g_0u_x|_0$ and $g_1u_x|_1$ into u_x^2 and g_1^2. Using the inequality $2\big|g_iu_x|_i \le g_i^2+u_x^2|_i$, $i=0,1$, we obtain the term $\|u_x\|_\infty^2$. However, $\|u_x\|_\infty^2$ cannot be estimated by only using $\|u\|^2$ and $\|u_x\|^2$. Using the energy method, the strong well-posedness cannot be shown. Thus we set homogeneous boundary conditions and show well-posedness instead.

We neglect the negative term $-2a_{\min}\|u_x\|^2$ and obtain an estimate of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\|u(\cdot\,,t)\|^2 \le \|(-b+c)_x\|_\infty \|u(\cdot\,,t)\|^2$$

which means that (2.1) is well-posed with a constant

$$\alpha = \|(-b+c)_x\|_\infty. \tag{2.10}$$

Robin Boundary Conditions.
Again, we consider the term $u(2au_x+(b+c)u)\big|_0^1$ in (2.9) and insert $u_x(0,t)=g_0(t)-\beta_0u(0,t)$ and $u_x(1,t)=g_1(t)-\beta_1u(1,t)$:

$$\begin{aligned}
2\left|u(au_x+du)\big|_0^1\right| &= 2\left|u(ag_1-\beta_1au+du)\big|_1 - u(ag_0-\beta_0au+du)\big|_0\right| \\
&\le \left(a(1,t)+2|d(1,t)-\beta_1a(1,t)|\right)u(1,t)^2 + a(1,t)g_1(t)^2 \\
&\quad + \left(a(0,t)+2|d(0,t)-\beta_0a(0,t)|\right)u(0,t)^2 + a(0,t)g_0(t)^2,
\end{aligned}$$

using the algebraic inequality $2rs \le \frac{1}{\epsilon}r^2+\epsilon s^2$ with $\varepsilon=1$.
Set $\gamma = 2(1+|\beta_0|+|\beta_1|)\|a\|_\infty + 2\|b+c\|_\infty$. Then,

$$2\left|u(au_x+du)\big|_0 1\right| \le \gamma\|u\|_\infty^2 + \|a\|_\infty\left(g_0(t)^2+g_1(t)^2\right).$$

We need a so called *Sobolev inequality* to get the desired estimate.

Lemma 2.3. *Let the real-valued function f be continuous and piecewise in $\mathcal{C}^1$. Then it holds for every $\varepsilon>0$*

$$\|f\|_\infty^2 \le \varepsilon\|f_x\|^2 + \left(\varepsilon^{-1}+1\right)\|f\|^2,$$

where the derivative is to be interpreted in the weak sense.

Proof. This inequality is shown for a continuously differentiable function f in [7]. In section 6 we need it also for piecewise continuously differentiable functions, which is why we give this more general proof here.
Let x_1 and x_2 be points with

$$|f(x_1)| = \min_x |f(x)|, \qquad |f(x_2)| = \max_x |f(x)| = \|f\|_\infty.$$

Without restriction, we can assume that $x_1 < x_2$. Let k denote the number of discontinuities of f_x in the interval $[x_1, x_2]$ and y_j, $j = 1, \ldots, k$ the positions where the discontinuities occur. Then

$$\int_{x_1}^{x_2} 2ff_x \mathrm{d}x = \int_{x_1}^{y_1} 2ff_x \mathrm{d}x + \sum_{j=1}^{k-1} \int_{y_j}^{y_{j+1}} 2ff_x \mathrm{d}x + \int_{y_k}^{x_2} 2ff_x \mathrm{d}x$$

The function $2ff_x = (f^2)_x$ is continuous in each one of these interval, which implies

$$\int_{x_1}^{x_2} 2ff_x \mathrm{d}x = f^2\Big|_{x_1}^{y_1} + \sum_{j=1}^{k-1} f^2\Big|_{y_j}^{y_{j+1}} + f^2\Big|_{y_k}^{x_2} = f^2\Big|_{x_1}^{x_2},$$

that is

$$\begin{aligned}\|f\|_\infty^2 - f(x_1)^2 &\leq 2\int_{x_1}^{x_2} |f|\,|f_x| \mathrm{d}x \leq 2\int_0^1 |f|\,|f_x| \mathrm{d}x \\ &\leq 2\sqrt{\varepsilon}\|f_x\| \frac{1}{\varepsilon}\|f\| \leq \varepsilon\|f_x\|^2 + \varepsilon^{-1}\|f\|^2.\end{aligned}$$

Since $f(x_1)^2 \leq \|f\|^2$, it follows

$$\|f\|_\infty^2 \leq \varepsilon\|f_x\|^2 + \left(\varepsilon^{-1} + 1\right)\|f\|^2.$$

□

Using lemma 2.3 with $\varepsilon = \frac{2a_{\min}}{\gamma}$, we get

$$2\left|u(au_x + cu)\big|_0^1\right| \leq 2a_{\min}\|u_x\|^2 + \left(\frac{\gamma^2}{2a_{\min}} + \gamma\right)\|u\|^2 + \|a\|_\infty \left(g_0(t)^2 + g_1(t)^2\right).$$

Using this result in (2.9), we obtain the estimate

$$\frac{\mathrm{d}}{\mathrm{d}t}\|u(\cdot\,,t)\|^2 \leq \alpha\|u(\cdot\,,t)\|^2 + K\left(g_0(t)^2 + g_1(t)^2\right)$$

with $K = \|a\|_\infty$ and

$$\alpha = \|(-b + c)_x\|_\infty + \frac{\gamma^2}{2a_{\min}} + \gamma.$$

We see that in this case the problem is strongly well-posed independently on β_0 and β_1.

2.2.2 Estimate for the Operator based on Finite Elements

We apply integration by parts to the terms $u(au_x)_x$ and $u(cu)_x$. This leads to

$$\frac{\mathrm{d}}{\mathrm{d}t}\|u\|^2 = 2\int_0^1 (b - c)uu_x \mathrm{d}x - 2\int_0^1 au_x^2 \mathrm{d}x + 2u(au_x + cu)\Big|_0^1. \tag{2.11}$$

The the first integral can be estimated by

$$\left|\int_0^1 (b - c)uu_x \mathrm{d}x\right| \leq \|b - c\|_\infty \int_0^1 |uu_x| \mathrm{d}x \leq \|b - c\|_\infty \|u\|\,\|u_x\|,$$

where the last inequality is obtained using the Cauchy-Schwarz inequality. Using the algebraic inequality $2rs \leq \frac{1}{\epsilon}r^2 + \epsilon s^2$, we obtain

$$\left| \int_0^1 (b-c)uu_x \mathrm{d}x \right| \leq \frac{\|b-c\|_\infty^2}{2a_{\min}} \|u\|^2 + \frac{a_{\min}}{2} \|u_x\|^2 \tag{2.12}$$

for $\epsilon = \frac{a_{\min}}{\|b-c\|_\infty}$ and

$$\left| \int_0^1 (b-c)uu_x \mathrm{d}x \right| \leq \frac{\|b-c\|_\infty^2}{4a_{\min}} \|u\|^2 + a_{\min} \|u_x\|^2 \tag{2.13}$$

for $\epsilon = \frac{2a_{\min}}{\|b-c\|_\infty}$. We need these two different estimates because of the different treatment of Dirichlet and Robin boundary conditions.

Dirichlet Boundary Conditions.
We use the estimate (2.13) as an estimate of the hyperbolic part and (2.8). All boundary terms vanish in the case of homogeneous boundary conditions, and we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t} \|u(\cdot\,,t)\|^2 \leq \frac{\|(b-c)\|_\infty^2}{2a_{\min}} \|u(\cdot\,,t)\|^2,$$

which shows the well-posedness with

$$\alpha = \frac{\|(b-c)\|_\infty^2}{2a_{\min}}. \tag{2.14}$$

Robin Boundary Conditions.
We consider the boundary term $u(au_x + cu)\big|_0^1$ in (2.11) and insert $u_x(0,t) = g_0(t) - \beta_0 u(0,t)$ and $u_x(1,t) = g_1(t) - \beta_1 u(1,t)$:

$$\begin{aligned} 2\left|u(au_x + cu)\big|_0^1\right| &= 2\left|u(ag_1 - \beta_1 au + cu)\big|_1 - u(ag_0 - \beta_0 au + cu)\big|_0\right| \\ &\leq \left(a(1,t) + 2|c(1,t) - \beta_1 a(1,t)|\right) u(1,t)^2 + a(1,t)g_1(t)^2 \\ &\quad + \left(a(0,t) + 2|c(0,t) - \beta_0 a(0,t)|\right) u(0,t)^2 + a(0,t)g_0(t)^2, \end{aligned}$$

using the algebraic inequality $2rs \leq \frac{1}{\epsilon}r^2 + \epsilon s^2$ with $\epsilon = 1$.
Set $\eta = 2(1 + |\beta_0| + |\beta_1|)\|a\|_\infty + 4\|c\|_\infty$. Then,

$$2\left|u(au_x + cu)\big|_0^1\right| \leq \eta \|u\|_\infty^2 + \|a\|_\infty \left(g_0(t)^2 + g_1(t)^2\right).$$

Using lemma 2.3 with $\varepsilon = \frac{a_{\min}}{\eta}$, we get

$$2\left|u(au_x + cu)\big|_0^1\right| \leq a_{\min}\|u_x\|^2 + \left(\frac{\eta^2}{a_{\min}} + \eta\right) \|u\|^2 + \|a\|_\infty \left(g_0(t)^2 + g_1(t)^2\right).$$

We use this estimate together with (2.12) and (2.8) and get

$$\frac{\mathrm{d}}{\mathrm{d}t}\|u(\cdot\,,t)\|^2 \leq \alpha \|u(\cdot\,,t)\|^2 + K\left(g_0(t)^2 + g_1(t)^2\right)$$

with $K = \|a\|_\infty$ and

$$\alpha = \frac{\eta^2 + \|(b-c)\|_\infty^2}{a_{\min}} + \eta. \tag{2.15}$$

If we have homogeneous boundary conditions, i.e. $g_0 \equiv g_1 \equiv 0$, we obtain the slightly sharper estimate

$$\frac{\mathrm{d}}{\mathrm{d}t}\|u(\cdot\,,t)\|^2 \leq \alpha\|u(\cdot\,,t)\|^2$$

with

$$\alpha = \frac{\widetilde{\eta}^2 + \|(b-c)\|_\infty^2}{a_{\min}} + \widetilde{\eta}, \tag{2.16}$$

where $\widetilde{\eta} = 2(|\beta_0| + |\beta_1|)\|a\|_\infty + 4\|c\|_\infty$.

3 Approximation Using the Product Rule

Finite difference operators satisfying a summation by parts rule have been derived by Strand [16] for the first derivative and Mattsson and Nordström [11] for the second derivative. Since equation (1.1) can also be written in the form

$$u_t = a_x(x,t)u_x(x,t) + a(x,t)u_{xx}(x,t)$$

if a and u_x are sufficiently smooth, these operators can also be used to solve (2.2).

When we are solving the convection diffusion equation (1.2), we write it in the form

$$u_t = a(x,t)u_{xx}(x,t) + (a_x(x,t) + b(x,t))u_x(x,t) + (c(x,t)u(x,t))_x. \tag{3.1}$$

In the following sections we will use a similar notation as Strand and Mattsson and Nordström [16, 11]. The domain $\Omega = [0,1]$ is discretized using $N+1$ equidistant grid points,

$$x_j = jh, \quad j = 0, 1, \ldots, N, \quad h = \frac{1}{N}$$

The numerical approximation at the grid point x_j is denoted v_j, and the discrete solution vector $v^T = [v_0, v_1, \ldots, v_N]$. We will use the matrices and vectors

$$e_0 = [1, 0, \ldots, 0]^T, \quad E_0 = \text{diag}([1, 0, \ldots, 0]),$$
$$e_N = [0, \ldots, 0, 1]^T, \quad E_N = \text{diag}([0, \ldots, 0, 1]).$$

3.1 Construction

To apply the energy method on the discretization, we first need a suitable norm $\|\cdot\|_H$. The mentioned papers cover operators for both diagonal norms and block norms. In applications the diagonal norms are most common, which is why we concentrate on them.

Consider equation (3.1). The semi-discretized equation has the following form

$$v_t = \Lambda_0 D_2 v + \Lambda_1 D_1 v + D_1 \Lambda_2 v, \tag{3.2}$$

where $D_1 = H^{-1}Q$ and $D_2 = H^{-1}(-A + BS)$ stand for summation by parts operators for the first and second derivative, respectively, $\Lambda_0 = \text{diag}([a(x_0,t), \ldots, a(x_N,t)])$ for a diagonal matrix containing the values of the function $a(x,t)$ at the grid points x_j, $\Lambda_1 = \text{diag}([\lambda_1(x_0,t), \ldots, \lambda_1(x_N,t)])$ for a diagonal matrix with $\lambda_1 \approx a_x + b$ and finally $\Lambda_2 = \text{diag}([c(x_0,t), \ldots, c(x_N,t)])$. The derivative a_x in λ_1 can either be given or an approximation in the grid points of sufficiently high order can be used.
The matrix A is positive semidefinite, $B = \text{diag}([-1, 0, \ldots, 0, 1])$, Q satisfies $Q + Q^T = B$ and S approximates the first derivative at x_0 and x_N. The matrix H defines a discrete diagonal norm via $\|v\|_H^2 = \sum_{j=0}^{N} H_{ii} v_i^2$.

3.2 Energy Estimate for the Semi-Discrete Problem

The fundamental property of any discretization of (1.2) is its convergence to the exact solution. By the *Lax-Richtmyer equivalence theorem*, convergence of a consistent numerical method is equivalent to stability. For this reason our goal is to show that the semi-discretization (3.2) is stable.

Definition 3.1. *Assume homogeneous boundary conditions. A semi-discretization is called* stable *if for some discrete norm* $\|\cdot\|_H$ *it holds*

$$\|v\|_H^2 \leq K\mathrm{e}^{\alpha_s t}\|f\|_H^2 \tag{3.3}$$

for any v*, where* K *and* α_s *are constants that do not depend on* h *and* v*.*

This definition (cf. [7]) is the discrete counterpart to the definition of well-posedness of the continuous problem (2.3). With such an estimate, we can ensure that the growth of the solution is bounded by the data and avoid thereby that roundoff errors could grow arbitrarily fast. Hence $\alpha_s/2$ is the *growth rate* of the semi-discretization.
Again, we claim the differentiated form of (3.3) to hold

$$\frac{\mathrm{d}}{\mathrm{d}t}\|v\|_H^2 \leq \alpha_s\|v\|_H^2.$$

As in the continuous case, if this estimate is satisfied, then (3.3) holds.

For inhomogeneous boundary conditions, we can define strong stability.

Definition 3.2. *A semi-discretization is called* strongly stable *if it is stable and*

$$\|v\|_H^2 \leq K\mathrm{e}^{\alpha_s t}\left(\|f\|_H^2 + \int_0^t (g_0(\tau)^2 + g_1(\tau)^2)\mathrm{d}\tau\right) \tag{3.4}$$

for any v*, where* K *and* α_s *are constants that do not depend on* h *and* v*.*

It would be ideal if the growth of the approximation would be related to the growth of the exact solution. This justifies the next definition (cf. [7, 12]):

Definition 3.3. *A semi-discretization is called* strictly stable *if it is stable and*

$$\alpha_s = \alpha + \mathcal{O}(h),$$

where $\alpha/2$ *is the growth rate of the continuous problem.*

However, showing strict stability is not always possible, as we will see in the sequel.

The discrete approximation of (2.2) using the numerical method (3.2) needs an additional boundary treatment in order to implement the physical boundary conditions correctly. This is done using the Simultaneous Approximation Term (SAT) as proposed by Carpenter et. al., cf. [3]. For constant coefficients this procedure leads to a strictly stable approximation in the H-norm.

Here we do not use the norm induced by H, but define the discrete inner product $(\cdot,\cdot)_{\widetilde{H}}$ with a matrix $\widetilde{H}$ by $(u,v)_{\widetilde{H}} = u^T\widetilde{H}v$, where $\widetilde{H} = H\Lambda_0^{-1}$. As $\Lambda_0^{-1} > 0$, the inner product $\|\cdot\|_{\widetilde{H}}$ is well-defined.

In the following section, we derive an energy estimate neglecting the boundary. The boundary is treated in the sections 3.2.2 and 3.2.3.

3.2.1 Stability in the Interior

We start with the semi-discretization (3.2) and consider the $\widetilde{H}$-norm of the discrete solution vector v:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\|v\|_{\widetilde{H}}^2 &= \big(v, v_t\big)_{\widetilde{H}} + \big(v_t, v\big)_{\widetilde{H}} \\
&= -v^T\big(\widetilde{H}\Lambda_0 H^{-1}A + (\Lambda_0 H^{-1}A)^T\widetilde{H}\big)v + v^T\big(\widetilde{H}\Lambda_1 H^{-1}Q + (\Lambda_1 H^{-1}Q)^T\widetilde{H}\big)v \\
&\quad + v^T\left(\widetilde{H}H^{-1}Q\Lambda_2 + (\widetilde{H}H^{-1}Q\Lambda_2)^T\right)v \\
&\quad + v^T\big(\widetilde{H}\Lambda_0 H^{-1}BS + (\Lambda_0 H^{-1}BS)^T\widetilde{H}\big)v.
\end{aligned}$$

Since Λ_0^{-1}, Λ_1 and H are all diagonal matrices, it holds $\widetilde{H}\Lambda_1 H^{-1} = H\Lambda_0^{-1}\Lambda_1 H^{-1} = \Lambda_0^{-1}\Lambda_1$ which yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|v\|_{\widetilde{H}}^2 = &-v^T(A+A^T)v + v^T\left(\Lambda_0^{-1}\Lambda_1 Q + (\Lambda_0^{-1}\Lambda_1 Q)^T\right)v \\ &+ v^T\left(\Lambda_0^{-1}Q\Lambda_2 + (\Lambda_0^{-1}Q\Lambda_2)^T\right)v + v^T\left(BS + (BS)^T\right)v.\end{aligned}$$

We can rewrite $Q = R + \frac{1}{2}B$ with an skew-symmetric part R, i.e. $R = -R^T$, and the boundary matrix B. Then,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|v\|_{\widetilde{H}}^2 = &-v^T\left(A+A^T\right)v + v^T\left(\Lambda_0^{-1}\Lambda_1 R + (\Lambda_0^{-1}\Lambda_1 R)^T\right)v \\ &+ v^T\left(\Lambda_0^{-1}R\Lambda_2 + (\Lambda_0^{-1}R\Lambda_2)^T\right)v + v^T\left(BS + (BS)^T\right)v \\ &+ v^T\left(\Lambda_0^{-1}(\Lambda_1+\Lambda_2)B\right)v.\end{aligned}$$

In this section, we neglect the two terms $v^T(BS + (BS)^T)v$ and $v^T(\Lambda_0^{-1}(\Lambda_1+\Lambda_2)B)v$ since they contribute to the boundary part of the operator. These terms will be treated in the sections 3.2.2 and 3.2.3 in combination with the implementation of the physical boundary conditions.

For the continuous problem, we have an estimate of the form (when neglecting boundary terms)

$$\frac{\mathrm{d}}{\mathrm{d}t}\|u\|^2 \leq -2a_{\min}\|u_x\|^2 + \|(-b+c)_x\|_\infty \|u\|^2.$$

Now we want to derive a similar estimate for the discrete problem.
The matrix A is constructed such that it is positive semidefinite, i.e.

$$-v^T\left(A+A^T\right)v \leq 0 \quad \forall\, v. \tag{3.5}$$

Let $C_1 = \Lambda_0^{-1}\Lambda_1 R + (\Lambda_0^{-1}\Lambda_1 R)^T$ and $C_2 = \Lambda_0^{-1}R\Lambda_2 + (\Lambda_0^{-1}R\Lambda_2)^T$. We want to obtain an estimate for the eigenvalues of C_1 and C_2. First we analyze the order of magnitude of their entries. We assume that all functions are sufficiently smooth such that all appearing derivatives exist.

Lemma 3.4. *If $\lambda_1(x,t) = a_x(x,t) + b(x,t)$ and $c(x,t)$ are Lipschitz continuous with respect to x, then the entries $c_{jk}^{(1)}$ and $c_{jk}^{(2)}$ of both C_1 and C_2 satisfy the estimate*

$$|c_{jk}^{(i)}| \leq K_i h|j-k|\,|r_{jk}|, \qquad i = 1,2$$

with some constants K_i.

Proof. By construction, the elements r_{jk} of R satisfy $r_{jk} = -r_{kj}$. Then for C_1 holds

$$\begin{aligned}c_{jk}^{(1)} &= \frac{(\lambda_1)_j}{a_j}r_{jk} + \frac{(\lambda_1)_k}{a_k}r_{kj} = r_{jk}\left(\frac{(\lambda_1)_j}{a_j} - \frac{(\lambda_1)_k}{a_k}\right) \quad \Rightarrow \\ |c_{jk}^{(1)}| &= |r_{jk}|\left|\frac{(\lambda_1)_j}{a_j} - \frac{(\lambda_1)_k}{a_j} + \frac{(\lambda_1)_k}{a_j} - \frac{(\lambda_1)_k}{a_k}\right| \\ &\leq |r_{jk}|\left(\left|\frac{(\lambda_1)_j}{a_j} - \frac{(\lambda_1)_k}{a_j}\right| + \left|\frac{(\lambda_1)_k}{a_j} - \frac{(\lambda_1)_k}{a_k}\right|\right) \\ &= |r_{jk}|\left(\frac{1}{a_j}\left|(\lambda_1)_j - (\lambda_1)_k\right| + |(\lambda_1)_k|\left|\frac{1}{a_j} - \frac{1}{a_k}\right|\right).\end{aligned} \tag{3.6}$$

Similarily, for C_2 holds:

$$|c_{jk}^{(2)}| \leq |r_{jk}| \left(\frac{1}{a_j} \left| (\lambda_2)_k - (\lambda_2)_j \right| + |(\lambda_2)_j| \left| \frac{1}{a_j} - \frac{1}{a_k} \right| \right). \tag{3.7}$$

Because of the Lipschitz continuity of λ_i, it holds $|(\lambda_i)_j - (\lambda_i)_k| \leq L_i \frac{|j-k|}{N}$ with a Lipschitz constant $L_i = \|(\lambda_i)_x\|_\infty$ for $i = 1, 2$.
The function a is also Lipschitz continuous with respect to x with Lipschitz constant $L_a = \|a_x\|_\infty$. Moreover, the function $f(x) = 1/x$ is Lipschitz continuous in any compact interval which does not contain 0 (with Lipschitz constant $L_s = \frac{1}{a_{\min}^2}$ for $\frac{1}{a(x)}$). Then it follows

$$\left| \frac{1}{a_j} - \frac{1}{a_k} \right| \leq L_s |a_j - a_k| \leq L_s L_a h \, |j - k|.$$

Back in (3.6) we get

$$|c_{jk}^{(1)}| \leq |r_{jk}| K_1 h |j - k|$$

with a constant

$$K_1 = \frac{\|(\lambda_1)_x\|_\infty}{a_{\min}} + \|\lambda_1\|_\infty \frac{\|a_x\|_\infty}{a_{\min}^2}.$$

For C_2 we get from (3.7)

$$|c_{jk}^{(2)}| \leq |r_{jk}| K_2 h |j - k|$$

with a constant

$$K_2 = \frac{\|(\lambda_2)_x\|_\infty}{a_{\min}} + \|\lambda_2\|_\infty \frac{\|a_x\|_\infty}{a_{\min}^2}.$$

□

Using lemma 3.4, we get an estimate for the eigenvalues of C_1 and C_2 and thus for their spectral radius.

Lemma 3.5. *The spectral radius of both C_1 and C_2 is in $\mathcal{O}(h)$.*

Proof. Let

$$s = \max_{j=0,\ldots,N} \left\{ \sum_{k=0}^{N} |j - k| \, |r_{jk}| \right\}.$$

s does not depend on N because r_{jk} is zero outside $2p + 1$ diagonals for interior points and $3p - 1$ points around the main diagonal for points in the boundary $2p \times 2p$ block.
Using *Gershgorin's theorem*, we find that all eigenvalues of C_i are lying in a disk around 0 (all diagonal elements of C are 0) with radius

$$K_i s h$$

Thus the spectral radius of C_i can be estimated by

$$\rho(C_i) \leq K_i s h,$$

where

$$K_i = \frac{\|(\lambda_i)_x\|_\infty}{a_{\min}} + \|\lambda_i\|_\infty \frac{\|a_x\|_\infty}{a_{\min}^2}.$$

□

Using lemma 3.5, we get

$$v^T C_i v \leq \rho(C_i) v^T v \leq K_i s\kappa \|a\|_\infty \|v\|^2_{\widetilde{H}} \qquad i = 1, 2, \tag{3.8}$$

where $\kappa = h\rho(H^{-1}) = \mathcal{O}(1)$ is the largest entry in the diagonal matrix hH^{-1}.
With the estimates in (3.5) and (3.8), we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\|v\|^2_{\widetilde{H}} \leq \alpha_s \|v\|^2_{\widetilde{H}} + v^T \left(BS + (BS)^T\right) v + v^T \left(\Lambda_0^{-1}(\Lambda_1 + \Lambda_2)B\right) v. \tag{3.9}$$

where

$$\alpha_s = \frac{s\kappa\|a\|_\infty}{a_{\min}} \left(\|(a_x + b)_x\|_\infty + \|c_x\|_\infty + (\|a_x + b\|_\infty + \|c\|_\infty) \frac{\|a_x\|_\infty}{a_{\min}} \right). \tag{3.10}$$

We have now shown stability for the operator in the interior.

3.2.2 Dirichlet Boundary Conditions

For Dirichlet boundary conditions the physical data is implemented with the SAT terms

$$\widetilde{H}^{-1}(\tau_0 S^T + \sigma_0 I)(E_0 v - e_0 g_0(t))$$

for the left boundary and

$$\widetilde{H}^{-1}(\tau_1 S^T + \sigma_1 I)(E_N v - e_N g_1(t))$$

for the right boundary. Here $\tau_0, \sigma_0, \tau_1, \sigma_1$ are constants that are to be determined such that the semi-discretization including the boundary is stable. If we insert the exact solution, the boundary terms vanish and hence the accuracy of (3.2) is not affected.
This yields the following semi-discretized system:

$$\begin{aligned} v_t = {} & \Lambda_0 H^{-1}(-A + BS)v + \Lambda_1 H^{-1} Q v + H^{-1} Q \Lambda_2 v \\ & - \widetilde{H}^{-1}(\tau_0 S^T + \sigma_0 I)\big(E_0 v - e_0 g_0(t)\big) - \widetilde{H}^{-1}(\tau_1 S^T + \sigma_1 I)\big(E_N v - e_N g_1(t)\big), \\ v(0) = {} & f. \end{aligned} \tag{3.11}$$

Taking the time derivative of the discrete $\widetilde{H}$-norm leads to

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\|v\|^2_{\widetilde{H}} &= \big(v, v_t\big)_{\widetilde{H}} + \big(v_t, v\big)_{\widetilde{H}} \\ &= -v^T \left(A + A^T\right) v + v^T C_1 v + v^T C_2 v + v^T \left(BS + (BS)^T\right) v \\ &\quad + v^T \left(\Lambda_0^{-1}(\Lambda_1 + \Lambda_2)B\right) v - 2(\tau_0 (Sv)_0 + \sigma_0 v_0)\big(v_0 - g_0(t)\big) \\ &\quad - 2(\tau_1 (Sv)_N + \sigma_1 v_N)\big(v_N - g_1(t)\big). \end{aligned} \tag{3.12}$$

Setting homogeneous boundary conditions $g_0(t) = g_1(t) = 0$, we can derive a condition on the constants τ_0, σ_0, τ_1 and σ_1:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\|v\|^2_{\widetilde{H}} = {} & -v^T \left(A + A^T\right) v + v^T C_1 v + v^T C_2 v - 2v_0^2 \left(\frac{(a_x)_0 + b_0 + c_0}{2a_0} + \sigma_0 \right) \\ & - 2v_N^2 \left(-\frac{(a_x)_N + b_N + c_N}{2a_N} + \sigma_1 \right) - 2v_0 (Sv)_0 (1 + \tau_0) \\ & + 2v_N (Sv)_N (1 - \tau_1). \end{aligned}$$

The first three terms have been discussed in section 3.2.1. The remaining terms represent the boundary treatment of the operator. For stability it is required that

$$\begin{aligned}&\tau_0 = -1 \quad \text{and} \quad \tau_1 = 1 \quad \text{as well as}\\ &\sigma_0 = -\frac{(a_x)_0 + b_0 + c_0}{2a_0} \quad \text{and} \quad \sigma_1 = \frac{(a_x)_N + b_N + c_N}{2a_N}.\end{aligned} \tag{3.13}$$

In this case we get the estimate

$$\frac{\mathrm{d}}{\mathrm{d}t}\|v\|_{\widetilde{H}}^2 \leq \alpha_s \|v\|_{\widetilde{H}} \tag{3.14}$$

with α_s given by (3.10).

3.2.3 Robin Boundary Conditions

For Robin boundary conditions the semi-discretized system with SAT boundary treatment has the following form

$$\begin{aligned}v_t = {} & \Lambda_0 H^{-1}(-A + BS)v + \Lambda_1 H^{-1} Q v + H^{-1} Q \Lambda_2 v - \widetilde{H}^{-1}\tau_0\big(E_0(\beta_0 I + S)v - e_0 g_0(t)\big)\\ & - \widetilde{H}^{-1}\tau_1\big(E_N(\beta_1 I + S)v - e_N g_1(t)\big), \quad v(0) = f,\end{aligned} \tag{3.15}$$

where τ_0 and τ_1 are some constants that are to be determined.
To investigate the accuracy of the added boundary terms in (3.15), we insert the exact solution u. The term $\beta_i u$ is exact while $\widetilde{H}^{-1}Su = \widetilde{H}^{-1}u_x + \mathcal{O}(h^{\tau+1})$, where τ is the accuracy of S. The order of accuracy of S can be one order less than the global order of accuracy of the scheme, see section 3.3 and [11].
If we take the $\widetilde{H}$-norm of (3.15), we obtain

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|v\|_{\widetilde{H}}^2 &= \big(v, v_t\big)_{\widetilde{H}} + \big(v_t, v\big)_{\widetilde{H}}\\ &= -v^T\left(A + A^T\right)v + v^T C_1 v + v^T C_2 v + v^T\left(BS + (BS)^T\right)v\\ &\quad + v^T\left(\Lambda_0^{-1}(\Lambda_1 + \Lambda_2)B\right)v - 2\tau_0 v_0\big(\beta_0 v_0 + (Sv)_0 - g_0(t)\big)\\ &\quad - 2\tau_1 v_N\big(\beta_1 v_N + (Sv)_N - g_1(t)\big),\end{aligned}$$

using the notation from the previous sections. We regroup the terms and set $\mu_i = \frac{(a_x)_i + b_i + c_i}{2a_i}$, $i = 0, N$, to get

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|v\|_{\widetilde{H}}^2 = {} & -v^T\left(A + A^T\right)v + v^T C_1 v + v^T C_2 v - 2v_0(Sv)_0(1 + \tau_0) + 2v_N(Sv)_N(1 - \tau_1)\\ & - 2(\mu_0 + \beta_0\tau_0)v_0^2 - 2(-\mu_N + \beta_1\tau_1)v_N^2 + 2\tau_0 v_0 g_0 + 2\tau_1 v_N g_1.\end{aligned} \tag{3.16}$$

We expand the terms in the second line to obtain an expression with the boundary terms separated:

$$\begin{aligned}&-2(\mu_0 + \tau_0\beta_0)v_0^2 - 2v_N^2(-\mu_N + \beta_1\tau_1) + 2\tau_0 v_0 g_0 + 2\tau_1 v_N g_1\\ &\quad = -2(\mu_0 + \beta_0\tau_0)\left(v_0 - \frac{\tau_0}{2(\mu_0 + \beta_0\tau_0)}g_0\right)^2 + \frac{\tau_0^2}{2(\mu_0 + \beta_0\tau_0)}g_0^2\\ &\qquad - 2(-\mu_N + \beta_1\tau_1)\left(v_N - \frac{\tau_1}{2(-\mu_N + \beta_1\tau_1)}g_1\right)^2 + \frac{\tau_1^2}{2(-\mu_N + \beta_1\tau_1)}g_1^2.\end{aligned} \tag{3.17}$$

Back in (3.16), we get

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|v\|_{\widetilde{H}}^2 = &-v^T\left(A+A^T\right)v + v^TC_1v + v^TC_2v - 2v_0(Sv)_0(1+\tau_0) + 2v_N(Sv)_N(1-\tau_1)\\ &- 2(\mu_0+\beta_0\tau_0)\left(v_0 - \frac{\tau_0}{2(\mu_0+\beta_0\tau_0)}g_0\right)^2 + \frac{\tau_0^2}{2(\mu_0+\beta_0\tau_0)}g_0^2\\ &- 2(-\mu_N+\beta_1\tau_1)\left(v_N - \frac{\tau_1}{2(-\mu_N+\beta_1\tau_1)}g_1\right)^2 + \frac{\tau_1^2}{2(-\mu_N+\beta_1\tau_1)}g_1^2.\end{aligned}$$

For stability is required that all boundary terms involving v are non-positive, i.e.

$$\begin{aligned}&\tau_0 = -1 \quad \text{and} \quad \tau_1 = 1 \quad \text{as well as}\\ &\beta_0 \le \frac{(a_x)_0 + b_0 + c_0}{2a_0} \quad \text{and} \quad \beta_1 \ge \frac{(a_x)_N + b_N + c_N}{2a_N}.\end{aligned} \tag{3.18}$$

We obtain the energy estimate

$$\frac{\mathrm{d}}{\mathrm{d}t}\|v\|_{\widetilde{H}}^2 \le \alpha_s\|v\|_{\widetilde{H}} + K\left(g_0(\tau)^2 + g_1(\tau)^2\right), \tag{3.19}$$

where

$$K = \max\left\{\frac{1}{2(\mu_0-\beta_0)}, \frac{1}{2(-\mu_N+\beta_1)}\right\}.$$

Remark 3.6. *By this procedure we get some requirements on the boundary conditions that do not occur in this form for the continuous problem. The reason is that we did not prove a discrete counterpart of the Sobolev inequality in our case which enables us to set positive boundary terms against the terms $-v^T(A+A^T)v$ and $\|v\|_{\widetilde{H}}$. In the continuous case the maximum norm can be estimated by the $\mathcal{L}^2$-norm of a function and its derivative. In [7] such an inequality is shown for D_1^2 approximating the second derivative.*
On the other hand, we can obtain a similar criterion as (3.18) *for the continuous case: We start off from equation* (2.9)*, use $d(x,t) = b(x,t) + c(x,t)$ and insert $u_x(0,t) = g_0(t) - \beta_0 u(0,t)$ and $u_x(1,t) = g_1(t) - \beta_1 u(1,t)$. Similarly to* (3.17)*, we obtain:*

$$\begin{aligned}&2u(ag_1 - \beta_1 au + du)\big|_1 - 2u(ag_0 - \beta_0 au + du)\big|_0\\ &\quad= -2(d-\beta_0 a)\left(u + \frac{a}{2(d-\beta_0 a)}g_0\right)^2\Big|_0 + \frac{a^2}{2(d-\beta_0 a)}g_0^2\Big|_0\\ &\qquad- 2(-d+\beta_1 a)\left(u - \frac{a}{2(-d+\beta_1 a)}g_1\right)^2\Big|_1 + \frac{a^2}{2(-d+\beta_1 a)}g_1^2\Big|_1\end{aligned}$$

We get the estimate

$$\frac{\mathrm{d}}{\mathrm{d}t}\|u(\cdot,t)\|^2 \le \|(-b+c)_x\|_\infty\|u(\cdot,t)\|^2 + K\left(g_0(t)^2 + g_1(t)^2\right)$$

in the case

$$\beta_0 \le \max_t \frac{b(0,t)+c(0,t)}{2a(0,t)} \quad \textit{and} \quad \beta_1 \ge \min_t \frac{b(1,t)+c(1,t)}{2a(1,t)} \tag{3.20}$$

where

$$K = \max\left\{\max_t \frac{a(0,t)^2}{2(d(0,t)-\beta_0 a(0,t))}, \max_t \frac{a(1,t)^2}{2(-d(1,t)+\beta_1 a(1,t))}\right\}.$$

The difference between the continuous condition on β_0, β_1 and the discrete one is due to the application of the product rule for the parabolic term and due to the different estimate of the hyperbolic terms.

3.2.4 Remark on Strict Stability

We consider *Dirichlet boundary conditions*. For Neumann boundary conditions the results are similar.

If we compare the constants $\alpha = \|(-b+c)_x\|_\infty$ for the continuous problem in (2.10) and α_s in (3.10), we note some differences.
We see that the application of the triangle inequality in (3.9) splits the term $(-b+c)_x$ into two terms. Additionally, the term a_x is added due to the application of the product rule.
The application of the norm induced by $\widetilde{H}$ causes some additional terms in the estimate. When using this norm, they cannot be avoided. Moreover, the norm makes it difficult to derive an estimate for the term $(-b+c)_x$ without using the triangle inequality.
Alternatively, the norm H introduced by Strand [16] could be used. However, we cannot show stability for the term $-v(\Lambda_0 A + A^T\Lambda_0)v$ (since we loose the positive definiteness of A), but for the first derivative parts we could get the term $\widetilde{\alpha}_s = \|(a_x+b-c)_x\|_\infty$ which is – apart from a_x – the same as in the continuous case.

If $b \equiv c \equiv 0$, the energy for the continuous problem is non-increasing. For a strictly stable approximation, the energy should also be either non-increasing or increase only by $\mathcal{O}(h)$. However, our theoretical analysis shows that this is not the case. An analysis of the matrix $X = -(A+A^T) + \left(\Lambda_0^{-1}\Lambda_1 R + (\Lambda_0^{-1}\Lambda_1 R)^T\right)$ with $a(x) = 0.2(1+x(x-1))$ and $\Lambda_1^{(i)} = a_x(x_i)$ shows that it has only one positive eigenvalue which is in $\mathcal{O}(h)$, while all others are negative. This eigenvalue results in a possibly positive term $v^T X v = \mathcal{O}(1)$, i.e. here we *do not have strict stability*.
The eigenvector corresponding to the positive eigenvalue is of the form

$$[1,1,\ldots,1] + \mathcal{O}(h)[f(0), f(h), \ldots, f(1)],$$

where $f(jh) = \cos(j\pi h)$ + some low frequency perturbation.
This means that it represents a low frequency wave. Such waves are usually uncritical for the error-growth because roundoff-errors are statistical errors and therefore more likely to be of high frequency. On the other hand, the truncation error can be reduced effectively by taking a finer mesh. This means that the not optimal energy estimate is still acceptable in terms of the overall-performance of the operators (the operators have minimal bandwidth and are thus very efficient).

The numerical results in section 3.5 approve of this appraisement.

3.3 Accuracy

The accuracy of the method to solve (3.2) is determined by the accuracy of D_1 and D_2. Let the operators D_1 and D_2 be of the order $2p$ in the interior and p at the boundary. The approximation of a_x shall be $2p$th order accurate . Then the scheme (3.2) approximates the right hand side of equation (3.1) with $2p$th order accuracy in the interior and pth order accuracy at the boundary.

We assume that the coefficient functions and the solution are sufficiently smooth.
Let $e = u - v$ be the difference between the exact solution at the grid points and the solution of the semi-discretization. It satisfies the differential equation

$$e_t = Me + T, \quad e(0) = 0, \tag{3.21}$$

where

$$M = H^{-1}(-\Lambda_0 A + \Lambda_1 R + R\Lambda_2)$$

for Dirichlet boundary conditions and

$$M = H^{-1}(-\Lambda_0 A + \Lambda_1 Q + Q\Lambda_2 + \beta_0 E_0 - \beta_1 E_N)$$

for Robin boundary conditions. The vector

$$T = \left[\mathcal{O}(h^p), \ldots, \mathcal{O}(h^p), \mathcal{O}(h^{2p}), \ldots, \mathcal{O}(h^{2p}), \mathcal{O}(h^p), \ldots, \mathcal{O}(h^p)\right]^T$$

denotes the truncation error with contributions from the approximation of the derivatives and the approximation of boundary derivatives (Su) in the SAT penalty term in the case of Robin boundary conditions.

The energy estimate from section 3.2 allows as an immediate consequence the estimate $\|e\|_{\widetilde{H}} \leq \mathcal{O}(h^p)$, which is however not sharp.
Since the approximation (3.2) is stable, a general result by Gustafsson [6] can be applied which ensures that the global order of accuracy is at least $p+1$, i.e. we gain at least one power at the boundary.
Nordström and Svärd [13] showed that for a parabolic problem with constant coefficients, two powers at the boundary can be gained.
Mattsson and Nordström [11] considered the convection-diffusion equation $u_t + au_x = \epsilon u_{xx}$ with constant a and $\epsilon > 0$ and proved that the semi-discretized scheme $v_t + aD_1 v = \epsilon D_2 v + C$ with a SAT boundary term C has the global order of accuracy $p+2$.

We suppose that this result can also be applied to our problem with variable coefficients. Indeed, we can apply a similar proof as Mattsson and Nordström to show that we gain two powers at the boundaries. We assume here that $a = a(x)$ does not depend on time and the operators are pointwise bounded in order to make it possible to apply the Laplace transform technique, cf. [7].

Theorem 3.7. *Consider the convection-diffusion equation with Dirichlet or Robin boundary conditions* (2.1), (2.2) *and the corresponding semi-discrete problems* (3.11), (3.15). *The error given by* (3.21) *satisfies the estimate* $\|e\|_{\widetilde{H}} = \mathcal{O}\left(h^{p+2}\right)$.

Proof. We split the error into three parts $e = e_{\mathrm{i}} + e_{\mathrm{b}}^{(\mathrm{l})} + e_{\mathrm{b}}^{(\mathrm{r})}$, where the subscripts (i, b) denote the inner and left and right boundary points, respectively. Similarily, the truncation error is divided into $T = T_{\mathrm{i}} + T_{\mathrm{b}}^{(\mathrm{l})} + T_{\mathrm{b}}^{(\mathrm{r})}$, where

$$\begin{aligned}
T_{\mathrm{i}} &= \left[0, \ldots, 0, \mathcal{O}(h^{2p}), \ldots, \mathcal{O}(h^{2p}), 0, \ldots, 0\right]^T, \\
T_{\mathrm{b}}^{(\mathrm{l})} &= \left[\mathcal{O}(h^p), \ldots, \mathcal{O}(h^p), 0, \ldots, 0, 0, \ldots, 0\right]^T, \\
T_{\mathrm{b}}^{(\mathrm{r})} &= \left[0, \ldots, 0, 0, \ldots, 0, \mathcal{O}(h^p), \ldots, \mathcal{O}(h^p)\right]^T.
\end{aligned}$$

Concerning e_{i}, we use the energy estimates (3.14) and (3.19) (note that for Robin boundary data condition (3.18) must hold) and complete them by the term $\|T_{\mathrm{i}}\|_{\widetilde{H}}$. Then we get the estimate

$$\|e_i\|_{\widetilde{H}} \leq \frac{\mathrm{e}^{\alpha_s t_0}}{\alpha_s} t_0 \left(\|T_{\mathrm{i}}\|_{\widetilde{H}}\right)_{\max([0,t])} = \mathcal{O}\left(h^{2p}\right)$$

at the time t_0.
To estimate $e_{\mathrm{b}}^{(\mathrm{l})}$ and $e_{\mathrm{b}}^{(\mathrm{r})}$, we use the Laplace transformation [7] of (3.21). We look at the error equations for $e_{\mathrm{b}}^{(\mathrm{l})}$ and $e_{\mathrm{b}}^{(\mathrm{r})}$ separately.
First we consider only the left boundary. We obtain

$$s\hat{e}_{\mathrm{b}}^{(\mathrm{l})} - M\hat{e}_{\mathrm{b}}^{(\mathrm{l})} = T_{\mathrm{b}}^{(\mathrm{l})}.$$

We multiply this equation by h^2 and introduce $\widetilde{s} = h^2 s$, $\widetilde{T}_{\mathrm{b}}^{(\mathrm{l})} = h^2 T_{\mathrm{b}}^{(\mathrm{l})}$, $\widetilde{M} = h^2 M$ as well as $P = \widetilde{s} I - \widetilde{M}$ in order to rewrite it as

$$P\hat{e}_{\mathrm{b}}^{(\mathrm{l})} = \widetilde{T}_{\mathrm{b}}^{(\mathrm{l})}. \tag{3.22}$$

For constant coefficients, we can find the solution to (3.22) by solving the characteristic equation determined by the internal difference scheme

$$\left(\hat{e}_{\mathrm{b}}^{(\mathrm{l})}\right)_j = \sum_{i=1}^{2p} \sigma_i \kappa_i^j,$$

where the κ_i are the roots of the characteristic equation and the unknowns σ_i are determined by the remaining equations from the boundary. For variable coefficients, the characteristic equations still depend on the position via a_j, i.e. we have to solve the equations

$$\widetilde{s}(\hat{e}_{\mathrm{b}}^{(\mathrm{l})})_j - (\widetilde{M}\hat{e}_{\mathrm{b}}^{(\mathrm{l})})_j = (\widetilde{T}_{\mathrm{b}}^{(\mathrm{l})})_j$$

with a matrix $\widetilde{M}$ whose entries can be different in each row. When considering interior points, $(\widetilde{T}_{\mathrm{b}}^{(\mathrm{l})})_j = 0$ and we have to solve a homogeneous equation of the form

$$\widetilde{s}(\hat{e}_{\mathrm{b}}^{(\mathrm{l})})_j = (\widetilde{M}\hat{e}_{\mathrm{b}}^{(\mathrm{l})})_j$$

We freeze the coefficients at the left boundary and derive an expression for the roots of the characteristic equation. Let $\kappa = \kappa_i$ be a root of the characteristic equation

$$\widetilde{s}\kappa^j = a_0 \bar{D}_2 \kappa^j + ((a_x)_0 + b_0 + c_0)\bar{D}_1 \kappa^j, \tag{3.23}$$

where the operator $\bar{D}_l$, $l = 1, 2$, is defined by

$$\bar{D}_l \kappa^j = \sum_{k=0}^{N} h^2 (D_l)_{j,k} \kappa^k,$$

i.e. it operates on κ^j just as the jth line of $h^2 D_l$ on the vector $v = [\kappa^0, \kappa^1, \ldots, \kappa^N]$.
Denote the roots of the variable coefficient problem as $\kappa + \Delta\kappa$. We show that $\Delta\kappa = \mathcal{O}(h)$: Inserting $\kappa + \Delta\kappa$ into the characteristic equation leads to:

$$\hat{s}(\kappa + \Delta\kappa)^j = \Lambda_0 \bar{D}_2 (\kappa + \Delta\kappa)^j + \Lambda_1 \bar{D}_1 (\kappa + \Delta\kappa)^j + \bar{D}_1 \Lambda_2 (\kappa + \Delta\kappa)^j.$$

If we expand this equation around x_0 into its Taylor series and use $(\kappa + \Delta\kappa)^j = \kappa^j + j\Delta\kappa \cdot \kappa^{j-1} +$ higher order terms, we obtain:

$$\begin{aligned}\hat{s}(\kappa^j + j\Delta\kappa \cdot \kappa^{j-1}) = {} & a_0 \bar{D}_2 \kappa^j + ((a_x)_0 + b_0 + c_0)\bar{D}_1 \kappa^j + h\partial\Lambda_0 \bar{D}_2 \kappa^j \\ & + h(\partial\Lambda_1 \bar{D}_1 + \bar{D}_1 \partial\Lambda_2)\kappa^j \\ & + a_0 \Delta\kappa \bar{D}_2 j \kappa^{j-1} + ((a_x)_0 + b_0 + c_0)\Delta\kappa \bar{D}_1 j \kappa^{j-1} + \text{h.o.t.},\end{aligned}$$

where $\partial\Lambda_i$, $i = 0, 1, 2$ denote rest terms.
Using (3.23), we can eliminate the leading terms to get

$$\begin{aligned}\hat{s} j \Delta\kappa \cdot \kappa^{j-1} = {} & h\partial\Lambda_0 \bar{D}_2 \kappa^j + a_0 j \Delta\kappa \bar{D}_2 \kappa^{j-1} + ((a_x)_0 + b_0 + c_0) j \Delta\kappa \bar{D}_1 \kappa^{j-1} \\ & + \frac{2a_0 \Delta\kappa}{h} \bar{D}_1 \kappa^{j-1} + h((a_x)_0 + b_0 + c_0)\Delta\kappa \bar{Y} \kappa^{j-1} + \text{h.o.t.},\end{aligned}$$

where we used the identities $\bar{D}_2 j\kappa^{j-1} = j\bar{D}_2\kappa^{j-1} + \frac{2}{h}\bar{D}_1\kappa^{j-1}$ and $\bar{D}_1 j\kappa^{j-1} = j\bar{D}_1\kappa^{j-1} + h\bar{Y}\kappa^{j-1}$ with some operator $\bar{Y} = \mathcal{O}(1)$. Since $\bar{D}_1 = \mathcal{O}(h)$, the term $h(\partial\Lambda_1\bar{D}_1 + \bar{D}_1\partial\Lambda_2)\kappa^j$ was neglected because it is in h^2.
Again we use (3.23) to eliminate the terms involving $j\Delta\kappa$ and get

$$0 = h\partial\Lambda_0\bar{D}_2\kappa^j + \frac{2a_0\Delta\kappa}{h}\bar{D}_1\kappa^{j-1} + h((a_x)_0 + b_0 + c_0)\Delta\kappa\bar{Y}\kappa^{j-1} + \text{h.o.t.}$$

Since the terms $\partial\Lambda_0\bar{D}_2\kappa^j$ and $\frac{1}{h}\bar{D}_1\kappa^{j-1}$ do not depend explicitly on h, $\Delta\kappa = \mathcal{O}(h)$ must hold and the term with $\bar{Y}$ is a higher order term.
Thus we can get a similar estimate as in the case of constant coefficients

$$\left(\hat{e}_{\mathrm{b}}^{(\mathrm{l})}\right)_j = \sum_{i=1}^{2p} \sigma_i(\kappa_i + \mathcal{O}(h))^j. \tag{3.24}$$

We seek conditions for which σ_i, $i = 1, \ldots, 2p$, are bounded and proportional to h^{p+2}, since this leads to $\|e_{\mathrm{b}}^{(\mathrm{l})}\|_{\widetilde{H}} = \mathcal{O}\left(h^{p+2}\right)$.
We need exactly $2p$ conditions to solve for the $2p$ unknowns σ_i. However, each boundary block $P^{(\mathrm{l,r})}$ has $2p$ rows (resulting in $4p$ rows totally) and the set of equations must be reduced. Let

$$\begin{aligned}
\hat{e}_1^{(\mathrm{ll})} &= \left[(\hat{e}_{\mathrm{b}}^{(\mathrm{l})})_1, \ldots, \hat{e}_{\mathrm{b}}^{(\mathrm{l})})_p\right]^T, \widetilde{T}_1^{(\mathrm{ll})} = \left[(\widetilde{T}_{\mathrm{b}}^{(\mathrm{l})})_1, \ldots, (\widetilde{T}_{\mathrm{b}}^{(\mathrm{l})})_p\right]^T, \\
\hat{e}_2^{(\mathrm{ll})} &= \left[(\hat{e}_{\mathrm{b}}^{(\mathrm{l})})_{p+1}, \ldots, \hat{e}_{\mathrm{b}}^{(\mathrm{l})})_{3p}\right]^T, \widetilde{T}_2^{(\mathrm{ll})} = \left[(\widetilde{T}_{\mathrm{b}}^{(\mathrm{l})})_{p+1}, \ldots, (\widetilde{T}_{\mathrm{b}}^{(\mathrm{l})})_{3p}\right]^T, \\
\hat{e}_1^{(\mathrm{lr})} &= \left[(\hat{e}_{\mathrm{b}}^{(\mathrm{l})})_{N-p+1}, \ldots, \hat{e}_{\mathrm{b}}^{(\mathrm{l})})_{N}\right]^T, \widetilde{T}_1^{(\mathrm{lr})} = \left[(\widetilde{T}_{\mathrm{b}}^{(\mathrm{l})})_{N-p+1}, \ldots, (\widetilde{T}_{\mathrm{b}}^{(\mathrm{l})})_{N}\right]^T, \\
\hat{e}_2^{(\mathrm{lr})} &= \left[(\hat{e}_{\mathrm{b}}^{(\mathrm{l})})_{N-3p+1}, \ldots, \hat{e}_{\mathrm{b}}^{(\mathrm{l})})_{N-p}\right]^T, \widetilde{T}_2^{(\mathrm{lr})} = \left[(\widetilde{T}_{\mathrm{b}}^{(\mathrm{l})})_{N-3p+1}, \ldots, (\widetilde{T}_{\mathrm{b}}^{(\mathrm{l})})_{N-p}\right]^T
\end{aligned}$$

and

$$P^{(\mathrm{l})} = \begin{bmatrix} P_{11}^{(\mathrm{l})} & P_{12}^{(\mathrm{l})} \\ P_{21}^{(\mathrm{l})} & P_{22}^{(\mathrm{l})} \end{bmatrix}, \qquad P^{(\mathrm{r})} = \begin{bmatrix} P_{22}^{(\mathrm{r})} & P_{21}^{(\mathrm{r})} \\ P_{12}^{(\mathrm{r})} & P_{11}^{(\mathrm{r})} \end{bmatrix},$$

where $P_{11}^{(\mathrm{l,r})}, P_{21}^{(\mathrm{l,r})}$ are $p \times p$ coefficient matrices and $P_{12}^{(\mathrm{l,r})}, P_{22}^{(\mathrm{l,r})}$ are $p \times 2p$ matrices. The $4p$ equations can be written as

$$\begin{aligned}
P_{11}^{(\mathrm{l,r})}\hat{e}_1^{(\mathrm{ll,lr})} + P_{12}^{(\mathrm{l,r})}\hat{e}_2^{(\mathrm{ll,lr})} &= \widetilde{T}_1^{(\mathrm{ll,lr})} \\
P_{21}^{(\mathrm{l,r})}\hat{e}_1^{(\mathrm{ll,lr})} + P_{22}^{(\mathrm{l,r})}\hat{e}_2^{(\mathrm{ll,lr})} &= \widetilde{T}_2^{(\mathrm{ll,lr})}
\end{aligned}$$

If $P_{21}^{(\mathrm{l,r})}$ is non-singular (notice that $P_{21}^{(\mathrm{l,r})}$ is independent of $\widetilde{s}$), this equation system can be reduced to

$$B^{(\mathrm{l,r})}\hat{e}_2^{(\mathrm{ll,lr})} = \widetilde{T}_f^{(\mathrm{ll,lr})},$$

where

$$B^{(\mathrm{l,r})} = P_{12}^{(\mathrm{l,r})} - P_{11}^{(\mathrm{l,r})}\left(P_{21}^{(\mathrm{l,r})}\right)^{-1}P_{22}^{(\mathrm{l,r})}, \quad \widetilde{T}_f^{(\mathrm{ll,lr})} = \widetilde{T}_1^{(\mathrm{ll,lr})} - P_{11}^{(\mathrm{l,r})}\left(P_{21}^{(\mathrm{l,r})}\right)^{-1}\widetilde{T}_2^{(\mathrm{ll,lr})}.$$

This leads to a linear equation system

$$C(\widetilde{s})\sigma = \widetilde{T}_f, \tag{3.25}$$

where

$$\widetilde{T}_f^T = \left[\left[\widetilde{T}_f^{(\mathrm{ll})} \right]^T, \left[\widetilde{T}_f^{(\mathrm{lr})} \right]^T \right],$$

$$C(\widetilde{s}) = \begin{bmatrix} \sum_{i=1}^{2p} b_{1,i}^{(\mathrm{l})} (\kappa_1 + \mathcal{O}(h))^{p+i} & \dots & \sum_{i=1}^{2p} b_{1,i}^{(\mathrm{l})} (\kappa_{2p} + \mathcal{O}(h))^{p+i} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{2p} b_{p,i}^{(\mathrm{l})} (\kappa_1 + \mathcal{O}(h))^{p+i} & \dots & \sum_{i=1}^{2p} b_{p,i}^{(\mathrm{l})} (\kappa_{2p} + \mathcal{O}(h))^{p+i} \\ \sum_{i=1}^{2p} b_{1,i}^{(\mathrm{r})} (\kappa_1 + \mathcal{O}(h))^{N-3p+i} & \dots & \sum_{i=1}^{2p} b_{1,i}^{(\mathrm{r})} (\kappa_{2p} + \mathcal{O}(h))^{N-3p+i} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{2p} b_{p,i}^{(\mathrm{r})} (\kappa_1 + \mathcal{O}(h))^{N-3p+i} & \dots & \sum_{i=1}^{2p} b_{p,i}^{(\mathrm{r})} (\kappa_{2p} + \mathcal{O}(h))^{N-3p+i} \end{bmatrix}$$

and

$$\sigma = \begin{bmatrix} \sigma_1 \\ \vdots \\ \sigma_{2p} \end{bmatrix}.$$

Now we can look at the solution to (3.25). The coefficients in $\widetilde{M}$ are proportional to $(\gamma_1 + \gamma_2 h)$, where γ_1 and γ_2 are of order one. This means that the solution to the system has terms proportional to $(\gamma_1 + \gamma_2 h)^{-1} h^{p+2}$ if $C(\widetilde{s}) \neq 0$. It holds $(\gamma_1 + \gamma_2 h)^{-1} \approx 1/\gamma_1 - (\gamma_2/\gamma_1^2)h$.
This means that the solution $\hat{e}_{\mathrm{b}}^{(\mathrm{l})}$ to (3.22) is proportional to h^{p+2}. Using Parseval's relation $\|\hat{e}_{\mathrm{b}}^{(\mathrm{l})}\|_{\widetilde{H}} = \|e_{\mathrm{b}}^{(\mathrm{l})}\|_{\widetilde{H}}$, $e_{\mathrm{b}}^{(\mathrm{l})}$ is of order h^{p+2}.
The right boundary can be estimated in the same way using the grid point x_N instead of x_0 in (3.23).
Then,

$$\|e\|_{\widetilde{H}} = \|e_{\mathrm{i}} + e_{\mathrm{b}}^{(\mathrm{l})} + e_{\mathrm{b}}^{(\mathrm{r})}\|_{\widetilde{H}} \leq \|e_{\mathrm{i}}\|_{\widetilde{H}} + \|e_{\mathrm{b}}^{(\mathrm{l})}\|_{\widetilde{H}} + \|e_{\mathrm{b}}^{(\mathrm{r})}\|_{\widetilde{H}} = \mathcal{O}\left(h^{p+2}\right),$$

where the triangle inequality has been used.
This proves the theorem. □

3.4 Damping of π-Modes

A desirable property of an operator is the damping of the highest-frequency waves, the so called π-modes.

We assume periodic boundary conditions (i.e. $u(0,t) = u(1,t)$) and constant coefficients such that the PDE (1.2) has the form

$$u_t = a u_{xx} + b u_x \tag{3.26}$$

with constants a and b.

First we expand the exact solution into its Fourier series,

$$u(x,t) = \sum_{\omega=-\infty}^{\infty} \hat{u}_\omega(t) \mathrm{e}^{2\pi \mathrm{i} \omega x}.$$

Inserting this into (3.26), we get

$$\sum_{\omega=-\infty}^{\infty} \frac{\mathrm{d}\hat{u}_\omega(t)}{\mathrm{d}t} \mathrm{e}^{2\pi \mathrm{i} \omega x} = \sum_{\omega=-\infty}^{\infty} \hat{u}_\omega(t) \mathrm{e}^{2\pi \mathrm{i} \omega x} \left(-a(2\pi)^2 \omega^2 + b 2\pi \mathrm{i} \omega \right)$$

Since $\left\{e^{2\pi i\omega x},\ \omega \in \mathbb{Z}\right\}$ is a linearly independent set, this equality must be satisfied for each summand, i.e.

$$\frac{d\hat{u}_\omega(t)}{dt} = \hat{u}_\omega(t)\left(-a\,(2\pi)^2\omega^2 + b\,2\pi i\omega\right) \quad \text{for all } \omega \tag{3.27}$$

with the exact solution

$$\hat{u}_\omega(t) = e^{\left(-a\,(2\pi)^2\omega^2 + b\,2\pi i\omega\right)t} \quad \text{for all } \omega.$$

When considering a semi-discretization of (3.26), we modify the operators D_2 and D_1 to $\widetilde{D}_2$ and $\widetilde{D}_1$ such that they approximate problems with periodic boundary conditions.
We assume $N = 2r$ and expand the numerical approximation into its Fourier series

$$v_j(t) = \sum_{\omega=-r}^{r} \hat{v}_\omega(t) e^{2\pi i\omega x_j}. \tag{3.28}$$

We look at $\widetilde{D}_i e^{2\pi i\omega x_j}$, $i = 1, 2$ and $j \in \{0, \ldots, N\}$:

- 2nd order accurate scheme:

$$\begin{aligned}
\widetilde{D}_1 e^{2\pi i\omega x_j} &= \frac{1}{2h}\left(e^{2\pi i\omega h} - e^{2\pi i\omega(-h)}\right) e^{2\pi i\omega x_j} = \frac{i}{h}\sin(\xi) e^{2\pi i\omega x_j} = \hat{D}_{1,2}(\xi) e^{2\pi i\omega x_j} \\
\widetilde{D}_2 e^{2\pi i\omega x_j} &= \frac{2}{h^2}(-1 + \cos(\xi)) e^{2\pi i\omega x_j} = \hat{D}_{2,2}(\xi) e^{2\pi i\omega x_j},
\end{aligned}$$

  where $\xi = 2\pi\omega h$.

- 4th order accurate scheme:

$$\begin{aligned}
\widetilde{D}_1 e^{2\pi i\omega x_j} &= \frac{i}{h}\left(\frac{4}{3}\sin(\xi) - \frac{1}{6}\sin(2\xi)\right) e^{2\pi i\omega x_j} = \hat{D}_{1,4}(\xi) e^{2\pi i\omega x_j} \\
\widetilde{D}_2 e^{2\pi i\omega x_j} &= \frac{1}{h^2}\left(-\frac{5}{2} + \frac{8}{3}\cos(\xi) - \frac{1}{6}\cos(2\xi)\right) e^{2\pi i\omega x_j} = \hat{D}_{2,4}(\xi) e^{2\pi i\omega x_j}.
\end{aligned}$$

- 6th order accurate scheme:

$$\begin{aligned}
\widetilde{D}_1 e^{2\pi i\omega x_j} &= \frac{i}{h}\left(\frac{3}{2}\sin(\xi) - \frac{3}{10}\sin(2\xi) + \frac{1}{30}\sin(3\xi)\right) e^{2\pi i\omega x_j} = \hat{D}_{1,6}(\xi) e^{2\pi i\omega x_j} \\
\widetilde{D}_2 e^{2\pi i\omega x_j} &= \frac{1}{h^2}\left(-\frac{49}{18} + 3\cos(\xi) - \frac{3}{10}\cos(2\xi) + \frac{1}{45}\cos(3\xi)\right) e^{2\pi i\omega x_j} = \hat{D}_{2,6}(\xi) e^{2\pi i\omega x_j}.
\end{aligned}$$

- 8th order accurate scheme:

$$\begin{aligned}
\widetilde{D}_1 e^{2\pi i\omega x_j} &= \frac{i}{h}\left(\frac{8}{5}\sin(\xi) - \frac{2}{5}\sin(2\xi) + \frac{8}{105}\sin(3\xi) - \frac{1}{140}\sin(4\xi)\right) e^{2\pi i\omega x_j} \\
&= \hat{D}_{1,8}(\xi) e^{2\pi i\omega x_j} \\
\widetilde{D}_2 e^{2\pi i\omega x_j} &= \frac{1}{h^2}\left(-\frac{205}{72} + \frac{16}{5}\cos(\xi) - \frac{2}{5}\cos(2\xi) + \frac{16}{315}\cos(3\xi) - \frac{1}{280}\cos(4\xi)\right) e^{2\pi i\omega x_j} \\
&= \hat{D}_{2,8}(\xi) e^{2\pi i\omega x_j}.
\end{aligned}$$

When inserting (3.28) into the semi-discretization $v_t = a\widetilde{D}_2 v + b\widetilde{D}_1 v$, we get as before

$$\frac{d\hat{v}_\omega(t)}{dt} = \hat{v}_\omega(t)\left(a\,\hat{D}_{2,2p}(\xi) + b\,\hat{D}_{1,2p}(\xi)\right) \quad \text{for all } \omega,\ p = 1, 2, 3, 4. \tag{3.29}$$

This differential equation has the analytic solution

$$\hat{v}_\omega(t) = e^{\left(a\,\hat{D}_{2,2p}(\xi) + b\,\hat{D}_{1,2p}(\xi)\right)t} \quad \text{for all } \omega,\ p = 1, 2, 3, 4.$$

Concerning the damping of certain modes, we first note that the hyperbolic term is purely imaginary and does therefore not change the discrete norm of v. In the following we assume therefore $b = 0$.
To compare the properties of the exact solution with those of the semi-discretization, we define $\hat{D}_0(\xi) = -(2\pi)^2\omega^2 = -\frac{\xi^2}{h^2}$. We consider the wave numbers $\omega = 0, 1, \ldots, r$ which yield $\xi = 0, \pi/r, \ldots, \pi$.

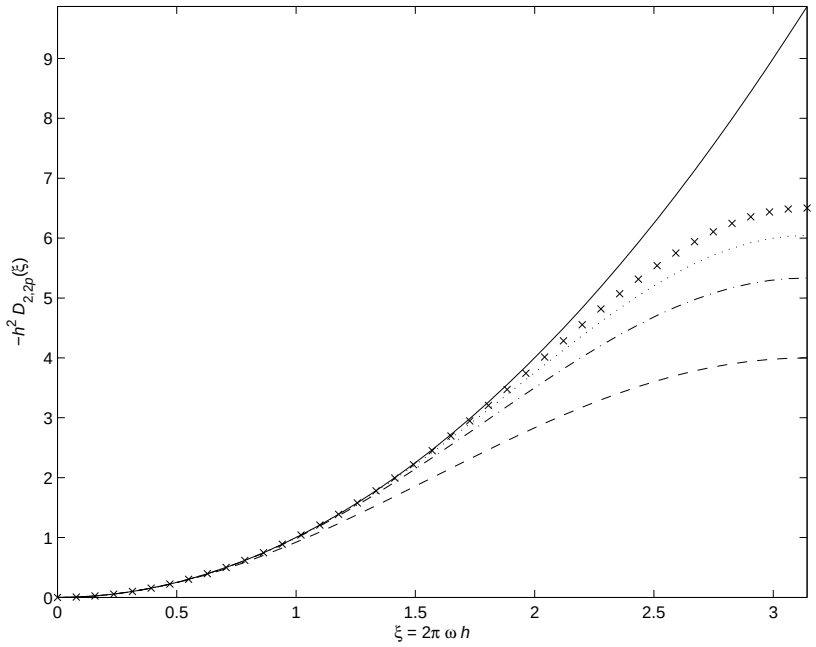

Figure 1: Damping of modes with wave number ξ of the exact solution (-), 2nd (- -), 4th (-·), 6th (·) and 8th (x) order accurate operator D_2.

We see in figure 1 that for low wave numbers, the damping of the approximation corresponds almost perfectly to the damping of the exact operator. For higher wave numbers, the damping of the approximations is somewhat lower than it should be, but still high frequency error modes get damped very quickly. This is one of the main advantages of a minimal width operator (an operator that uses $2p + 1$ points for an order of accuracy of $2p$, i.e. as few as possible) over an operator arising from using the operator D_1 twice as an approximation of u_{xx}, see also [11].

3.5 Computational Results

We apply the above operators to some test cases.

For the time integration we use the classical fourth order Runge-Kutta method. Since it is an explicit method, we get a restriction on the time step. To calculate the limit on the time step, we perform a Fourier stability analysis:
We start with the equation $v_t = aD_2v$. When considering the Fourier-transform of this equation, we get $\frac{\mathrm{d}}{\mathrm{d}t}\hat{v}_\omega(t) = a\hat{D}_{2,2p}(\omega)\hat{v}_\omega(t)$, i.e. for each ω we have the scalar test equation $y_t = \alpha y$. For negative real α, the stability limit for the 4th order Runge-Kutta method is $-2\sqrt{2} \approx -2.8$. This means that $a\left|\hat{D}_{2,2p}(\omega)\Delta t\right| \leq 2.8$ for all ω. We see in figure 1 that $|\hat{D}_{2,2p}|$ takes its maximum for $\omega = r$. The maxima are given in table 1. Using them, we can derive a limit on the time step $\frac{\Delta t}{h^2} \leq \frac{2.8}{a\left|h^2\hat{D}_{2,2p}(\omega)\right|}$. This result for constant coefficients can be extended

p	$h^2\lvert\hat{D}_{2,2p}(\omega = r)\rvert$	$\frac{2.8}{a\lvert h^2\hat{D}_{2,2p}(\omega=r)\rvert}$
1	4	$0.70/a$
2	$\frac{16}{3} \approx 5.33$	$0.53/a$
3	$\frac{272}{45} \approx 6.04$	$0.46/a$
4	$\frac{2048}{315} \approx 6.50$	$0.43/a$

Table 1: Maximum of $h^2\hat{D}_{2,2p}$ for $p = 1, 2, 3, 4$.

to the case of variable coefficients. In that case, we substitute a in table 1 with $\|a\|_\infty$. If we are dealing with the general convection-diffusion equation (1.2), we choose the time step slightly smaller to account for the additional term $\frac{\Delta t}{h}|\hat{D}_{1,2p}|$ from the approximation of the first derivative.

3.5.1 Results for the Parabolic Term

Example 1.
We consider the simple parabolic equation

$$u_t = 0.2u_{xx} \quad \text{with initial condition}$$

$$u(x, 0) = \sin(\pi x) + \cos(\pi x).$$

We choose Dirichlet boundary conditions

$$u(0, t) = \mathrm{e}^{-0.2\pi^2 t}, \qquad u(1, t) = -\mathrm{e}^{-0.2\pi^2 t}$$

and Neumann boundary conditions

$$u_x(0, t) = \pi\mathrm{e}^{-0.2\pi^2 t}, \qquad u_x(1, t) = -\pi\mathrm{e}^{-0.2\pi^2 t}.$$

This problem has the exact solution

$$u(x, t) = (\sin(\pi x) + \cos(\pi x))\mathrm{e}^{-0.2\pi^2 t}$$

for both Dirichlet and Neumann boundary conditions.
We solve the problem at the time $t = 1$ and use a time step of $\Delta t = 3h^2$ for the second order method and $\Delta t = 2h^2$ for the fourth order method (the stability limit is $3.5h^2$ and $2.65h^2$, respectively). We compare the numerical solution with the exact solution at the grid points by $\|u - v\|_h$, where $\|v\|_h^2 = h \cdot v^T v$ denotes the discrete l_2-norm.
The numerical order of accuracy q_D and q_N for both types of boundary conditions is calculated between two subsequent grid spacings h_1 and h_2 using the formula

$$q_{D,N} = \log\left(\frac{\|u - v\|_{h_2}}{\|u - v\|_{h_1}}\right) \Big/ \log\left(\frac{h_2}{h_1}\right)$$

The numerical results are given in table 2. As expected, the numerical convergence rates correspond to the accuracy in the interior. For the fourth order method, the boundary closure is only second order accurate, which means that we gain two orders at the boundary. This coincides with the results by [11].

Example 2.
We consider the problem

$$u_t(x, t) = (a(x)u_x(x, t))_x, \qquad u(x, 0) = \sin(\pi x) + \cos(\pi x)$$

	2nd order accurate scheme				4th order accurate scheme			
	Dirichlet B.C.		Neumann B.C.		Dirichlet B.C.		Neumann B.C.	
N	$\|u-v\|_h$	q_D	$\|u-v\|_h$	q_N	$\|u-v\|_h$	q_D	$\|u-v\|_h$	q_N
10	$6.52 \cdot 10^{-3}$		$8.58 \cdot 10^{-3}$		$1.97 \cdot 10^{-4}$		$6.70 \cdot 10^{-5}$	
20	$1.35 \cdot 10^{-3}$	2.27	$2.12 \cdot 10^{-3}$	2.01	$1.01 \cdot 10^{-5}$	4.29	$3.43 \cdot 10^{-6}$	4.29
40	$3.03 \cdot 10^{-4}$	2.15	$5.27 \cdot 10^{-4}$	2.01	$5.99 \cdot 10^{-7}$	4.08	$1.79 \cdot 10^{-7}$	4.26
80	$7.16 \cdot 10^{-5}$	2.08	$1.31 \cdot 10^{-4}$	2.01	$3.75 \cdot 10^{-8}$	4.00	$1.00 \cdot 10^{-8}$	4.16
160	$1.74 \cdot 10^{-5}$	2.04	$3.27 \cdot 10^{-5}$	2.00	$2.36 \cdot 10^{-9}$	3.99	$5.92 \cdot 10^{-10}$	4.08
320	$4.29 \cdot 10^{-6}$	2.02	$8.18 \cdot 10^{-6}$	2.00	$1.48 \cdot 10^{-10}$	3.99	$3.50 \cdot 10^{-11}$	4.08

Table 2: Numerical results for the equation $u_t = 0.2\, u_{xx}$ for Dirichlet and Neumann boundary conditions with the second and fourth order accurate schemes

with

$$a(x) = 0.2 + 0.4x(x-1).$$

We consider Dirichlet boundary conditions of the form

$$u(0,t) = \mathrm{e}^{-0.2\pi(\pi+2)t} \text{ and } u(1,t) = -\mathrm{e}^{-0.2\pi(\pi-2)t}.$$

The boundary conditions are chosen such that $u(0,0)$ and $u(1,0)$ are well defined (i.e. $f(0) = g_0(0)$ and $f(1) = g_1(0)$) and the differential equation can be satisfied in $(0,0)$ and $(1,0)$. Additionally, we include Robin boundary conditions. To mind the stability condition (3.18), we choose $\beta_0 = -1$ and $\beta_1 = 1$. Thus we have the boundary conditions:

$$-u(0,t) + u_x(0,t) = (-1+\pi)\mathrm{e}^{-0.2\pi(\pi+2)t} \text{ and } u(1,t) + u_x(1,t) = (-1-\pi)\mathrm{e}^{-0.2\pi(\pi-2)t}$$

Since we cannot get an analytic solution here, we take the numerical solution with the more accurate fourth order method on a very fine grid ($N = 960$) as a reference and calculate the error compared to that.
The derivative of the coefficient function a is approximated with fourth order accuracy using the standard approximation in the interior and one-sided difference operators at the boundary.
The results are given in table 3. We observe that for both the 2nd and 4th order accurate semi-discretization the numerical convergence rate is 2nd and 4th order, respectively. Hence we gain two orders at the boundary here as well.

	2nd order accurate scheme				4th order accurate scheme			
	Dirichlet B.C.		Robin B.C.		Dirichlet B.C.		Robin B.C.	
N	$\|u-v\|_h$	q_D	$\|u-v\|_h$	q_N	$\|u-v\|_h$	q_D	$\|u-v\|_h$	q_N
10	$6.27 \cdot 10^{-3}$		$1.27 \cdot 10^{-2}$		$1.10 \cdot 10^{-3}$		$2.81 \cdot 10^{-4}$	
20	$1.10 \cdot 10^{-3}$	2.42	$3.10 \cdot 10^{-3}$	2.03	$7.09 \cdot 10^{-5}$	3.95	$1.60 \cdot 10^{-5}$	4.14
40	$2.25 \cdot 10^{-4}$	2.29	$7.65 \cdot 10^{-4}$	2.02	$4.29 \cdot 10^{-6}$	4.04	$9.32 \cdot 10^{-7}$	4.10
80	$5.07 \cdot 10^{-5}$	2.15	$1.90 \cdot 10^{-4}$	2.01	$2.62 \cdot 10^{-7}$	4.03	$5.63 \cdot 10^{-8}$	4.05
160	$1.20 \cdot 10^{-5}$	2.07	$4.73 \cdot 10^{-5}$	2.00	$1.62 \cdot 10^{-8}$	4.02	$3.46 \cdot 10^{-9}$	4.02
320	$2.93 \cdot 10^{-6}$	2.04	$1.18 \cdot 10^{-5}$	2.00	$9.92 \cdot 10^{-10}$	4.03	$2.17 \cdot 10^{-10}$	4.00

Table 3: Numerical results for the equation $u_t = (a(x)u_x)_x$, $a(x) = 0.2 + 0.4x(x-1)$ for Dirichlet and Robin boundary conditions with the second and fourth order accurate schemes

We try to solve the same problem with Neumann boundary conditions. We have not shown

stability with the used operators, but we guess that some discrete form of the Sobolev inequality exists which gives stability for all β_0, β_1 in an approximation of (2.2) (cf. remark 3.6). We use

$$u_x(0,t) = \pi \mathrm{e}^{-0.2\pi(\pi+2)t} \text{ and } u_x(1,t) = -\pi \mathrm{e}^{-0.2\pi(\pi-2)t}.$$

The results are given in table 4. Indeed we see that we get a convergent method which is second or fourth order accurate, respectively.

	2nd order accurate scheme		4th order accurate scheme	
N	$\|u-v\|_h$	q_N	$\|u-v\|_h$	q_N
10	$1.12 \cdot 10^{-2}$		$2.94 \cdot 10^{-4}$	
20	$2.73 \cdot 10^{-3}$	2.03	$1.60 \cdot 10^{-5}$	4.20
40	$6.73 \cdot 10^{-4}$	2.02	$9.23 \cdot 10^{-7}$	4.12
80	$1.67 \cdot 10^{-4}$	2.01	$5.55 \cdot 10^{-8}$	4.05
160	$4.17 \cdot 10^{-5}$	2.00	$3.41 \cdot 10^{-9}$	4.02
320	$1.04 \cdot 10^{-5}$	2.00	$2.13 \cdot 10^{-10}$	4.00

Table 4: Numerical results for the equation $u_t = (a(x)u_x)_x$, $a(x) = 0.2 + 0.4x(x-1)$ for Neumann boundary conditions with the second and fourth order accurate schemes

Example 3.
Next we consider a non-polynomial a using

$$a(x) = 0.2(1 + \sin(\pi x))$$

in the equation

$$u_t(x,t) = (a(x)u_x(x,t))_x, \qquad u(x,0) = \sin(\pi x) + \cos(\pi x).$$

As boundary conditions we examine the Dirichlet data

$$u(0,t) = 1 \text{ and } u(1,t) = -\mathrm{e}^{-0.4\pi^2 t}$$

as well as the Neumann data

$$u_x(0,t) = \pi \text{ and } u_x(1,t) = -\pi \mathrm{e}^{-0.4\pi^2 t}.$$

Note that for this a condition (3.18) is satisfied.
Concerning the integration of the semi-discretized system, we note that $\|a\|_\infty = 0.4$ and hence we choose a time step of $1.5h^2$ and h^2 for the 2nd and 4th order method, respectively.

Example 4.
We conclude the experiments with the parabolic term by an example of a time-dependent a, namely

$$a(x,t) = 0.5\,\mathrm{e}^{x-2} \cdot (1 + \sin(\pi t))$$

in the partial differential equation

$$u_t(x,t) = (a(x)u_x(x,t))_x, \qquad u(x,0) = (\sin(\pi x))^2 + 2x.$$

Our Dirichlet boundary conditions look like

$$u(0,t) = 0 \text{ and } u(1,t) = 2\,\mathrm{e}^{-1} \cdot \mathrm{e}^{(\pi^2+1)t},$$

	2nd order accurate scheme				4th order accurate scheme			
	Dirichlet B.C.		Neumann B.C.		Dirichlet B.C.		Neumann B.C.	
N	$\|u-v\|_h$	q_D	$\|u-v\|_h$	q_N	$\|u-v\|_h$	q_D	$\|u-v\|_h$	q_N
10	$9.62 \cdot 10^{-3}$		$4.07 \cdot 10^{-3}$		$2.81 \cdot 10^{-4}$		$8.66 \cdot 10^{-4}$	
20	$2.63 \cdot 10^{-3}$	2.03	$5.12 \cdot 10^{-4}$	2.99	$1.71 \cdot 10^{-5}$	4.04	$8.15 \cdot 10^{-5}$	3.41
40	$5.86 \cdot 10^{-4}$	2.01	$6.40 \cdot 10^{-5}$	3.00	$9.44 \cdot 10^{-7}$	4.17	$6.45 \cdot 10^{-6}$	3.66
80	$1.46 \cdot 10^{-4}$	2.01	$1.16 \cdot 10^{-5}$	2.46	$5.21 \cdot 10^{-7}$	4.18	$4.58 \cdot 10^{-7}$	3.82
160	$3.64 \cdot 10^{-5}$	2.00	$3.14 \cdot 10^{-6}$	1.89	$2.96 \cdot 10^{-9}$	4.14	$3.05 \cdot 10^{-8}$	3.91
320	$9.09 \cdot 10^{-6}$	2.00	$8.68 \cdot 10^{-7}$	1.86	$1.74 \cdot 10^{-10}$	4.09	$1.95 \cdot 10^{-9}$	3.97

Table 5: Numerical results for the equation $u_t = (a(x)u_x)_x$, $a(x) = 0.2(1+\sin(\pi x))$ for Dirichlet and Neumann boundary conditions with the second and fourth order accurate schemes

and our Neumann boundary conditions are

$$u_x(0,t) = 0 \text{ and } u_x(1,t) = 2\,\mathrm{e}^{(\pi^2+1)\mathrm{e}^{-1}t(1+\sin(\pi t))}.$$

It holds $\max_t\{\|a(\cdot,t)\|\} = \mathrm{e}^{-1} \approx 0.37$ which is why we choose the time step $\Delta t = 1.5h^2$ for the second order scheme and $\Delta t = h^2$ for the fourth order scheme. The results are given in table 6. In this example the error is relatively large compared with the previous examples. The main reason is not the more difficult problem, but the fact that the l_2-norm of the solution is about 100 time as large as above due to the growing boundary value.

	2nd order accurate scheme				4th order accurate scheme			
	Dirichlet B.C.		Neumann B.C.		Dirichlet B.C.		Neumann B.C.	
N	$\|u-v\|_h$	q_D	$\|u-v\|_h$	q_N	$\|u-v\|_h$	q_D	$\|u-v\|_h$	q_N
10	$3.48 \cdot 10^{-1}$		$4.16 \cdot 10^{-2}$		$1.31 \cdot 10^{-1}$		$1.09 \cdot 10^{-2}$	
20	$1.05 \cdot 10^{-1}$	1.73	$9.51 \cdot 10^{-3}$	2.13	$9.30 \cdot 10^{-3}$	3.81	$5.73 \cdot 10^{-4}$	4.25
40	$3.60 \cdot 10^{-2}$	1.54	$2.22 \cdot 10^{-3}$	2.10	$6.51 \cdot 10^{-4}$	3.84	$2.67 \cdot 10^{-5}$	4.42
80	$1.05 \cdot 10^{-2}$	1.77	$5.35 \cdot 10^{-4}$	2.05	$4.48 \cdot 10^{-5}$	3.86	$1.22 \cdot 10^{-6}$	4.45
160	$2.84 \cdot 10^{-3}$	1.89	$1.31 \cdot 10^{-4}$	2.03	$2.99 \cdot 10^{-6}$	3.91	$5.62 \cdot 10^{-8}$	4.45
320	$7.38 \cdot 10^{-3}$	1.95	$3.25 \cdot 10^{-5}$	2.01	$1.90 \cdot 10^{-7}$	3.97	$2.51 \cdot 10^{-9}$	4.48

Table 6: Numerical results for the equation $u_t = (a(x,t)u_x)_x$, $a(x,t) = 0.5\,\mathrm{e}^{x-2}(1+\sin(\pi t))$ for Dirichlet and Neumann boundary conditions with the second and fourth order accurate schemes

3.5.2 Results for the Convection Diffusion Equation

Example 1.
We consider the example $u_t = (a(x,t)u_x(x,t))_x + b(x,t)u_x(x,t) + (c(x,t)u(x,t))_x$ with coefficients

$$a(x,t) = \frac{1}{10}\left(1 + \frac{1}{2}\sin\left(\frac{\pi}{2}(x+t)\right)\right),$$
$$b(x,t) = 2\sinh\left(-3x + \frac{3}{2}\right)(1+t) \quad \text{and}$$
$$c(x,t) = 4x(x-1)\left(xt^2+1\right).$$

Let the initial condition $u(x,0) = f(x)$ be given by $f(x) = 4(\sin(\pi x) + \cos(10x))$.
As Dirichlet boundary conditions we choose

$$u(0,t) = g_0(t) = 4\mathrm{e}^{4(-14+1/40\cdot\pi^2+2\pi\sinh(3/2))t} \approx 4\,\mathrm{e}^{-1.48t} \quad \text{and}$$
$$u(1,t) = g_1(t) = 4\cos(10)\mathrm{e}^{4(-11\cos(10)+2\sinh(3/2)(\pi+10\sin(10)))t} \approx 4\,\cos(10)\mathrm{e}^{-2.24t},$$

while we use

$$u_x(0,t) = g_0(t) \approx 4\pi\mathrm{e}^{-1.48t} \quad \text{and}$$
$$u_x(1,t) = g_1(t) \approx 4(-\pi - 10\sin(10))\mathrm{e}^{-2.24t}$$

as Neumann boundary conditions.
Note that again we choose the boundary conditions such that the PDE is satisfied in $(0,0)$ and $(1,0)$ and $g_0(0) = f(0)$ as well as $g_1(0) = f(1)$.
The numerical results are given in table 7. We use a time step of $3h^2$ for the second order method and $2h^2$ for the fourth order method. The results show that we gain two powers at the boundaries in the case of the fourth order accurate scheme even for the convection diffusion equation.

	2nd order accurate scheme				4th order accurate scheme			
	Dirichlet B.C.		Neumann B.C.		Dirichlet B.C.		Neumann B.C.	
N	$\|u-v\|_h$	q_D	$\|u-v\|_h$	q_N	$\|u-v\|_h$	q_D	$\|u-v\|_h$	q_N
10	2.62		$7.76 \cdot 10^{-2}$		1.95		$7.00 \cdot 10^{-3}$	
20	$9.94 \cdot 10^{-1}$	1.40	$1.83 \cdot 10^{-2}$	2.08	$4.94 \cdot 10^{-1}$	1.98	$1.82 \cdot 10^{-3}$	1.95
40	$2.05 \cdot 10^{-1}$	2.28	$4.49 \cdot 10^{-3}$	2.03	$7.22 \cdot 10^{-2}$	2.78	$2.76 \cdot 10^{-4}$	2.72
80	$3.19 \cdot 10^{-2}$	2.68	$1.10 \cdot 10^{-3}$	2.03	$8.01 \cdot 10^{-3}$	3.17	$3.05 \cdot 10^{-5}$	3.18
160	$5.06 \cdot 10^{-3}$	2.66	$2.64 \cdot 10^{-4}$	2.06	$7.04 \cdot 10^{-4}$	3.51	$2.43 \cdot 10^{-6}$	3.65
320	$8.72 \cdot 10^{-4}$	2.54	$5.51 \cdot 10^{-5}$	2.26	$5.15 \cdot 10^{-5}$	3.78	$1.55 \cdot 10^{-7}$	3.97

Table 7: Numerical results for the convection diffusion equation using Dirichlet and Neumann boundary conditions with the second and fourth order accurate schemes

The relatively high errors and bad convergence rates in the case of Dirichlet boundary conditions are due to the properties of the solution. At the right boundary, the solution is very steep. This steepness cannot be resolved with a low number of grid points in space and influences the convergence rate especially when the number of grid points is small.

Example 2.
We consider the problem

$$u_t(x,t) = (a(x)u_x(x,t))_x + b(x)u_x(x,t) \quad \text{with}$$
$$a(x) = \frac{1}{10}\left(1 + \frac{1}{2}\sin\left(\frac{\pi}{2}x\right)\right) \quad \text{and}$$
$$b(x) = 2\sinh\left(-3x + \frac{3}{2}\right).$$

Let the initial condition $u(x,0) = f(x)$ be given by

$$f(x) = \sin(\pi x) + \frac{7}{5}\cos(10x).$$

As Dirichlet boundary conditions we choose

$$u(0,t) = g_0(t) = \frac{7}{5}\,\mathrm{e}^{(-14+1/40\cdot\pi^2+2\pi\sinh(3/2))t} \approx 1.4\,\mathrm{e}^{-0.37t} \quad \text{and}$$
$$u(1,t) = g_1(t) = \frac{7}{5}\,\cos(10)\mathrm{e}^{(-21\cos(10)+2\sinh(3/2)(\pi+14\sin(10)))t} \approx 1.4\,\cos(10)\mathrm{e}^{-1.43t},$$

while we use

$$u_x(0,t) = g_0(t) \approx \pi\mathrm{e}^{-0.37t} \quad \text{and}$$
$$u_x(1,t) = g_1(t) \approx (-\pi - 14\sin(10))\mathrm{e}^{-1.43t}$$

as Neumann boundary conditions.
Note that again we choose the boundary conditions such that the PDE is satisfied in $(0,0)$ and $(1,0)$ and $g_0(0) = f(0)$ as well as $g_1(0) = f(1)$.
The numerical results are given in table 8. We use a time step of $3h^2$ for the second order method and $2h^2$ for the fourth order method. It can be seen that the results are very similar to the previous test case. Again, the numerical order of accuracy is two for scheme that is second order accurate in the interior and four for the fourth order accurate one.

	2nd order accurate scheme				4th order accurate scheme			
	Dirichlet B.C.		Neumann B.C.		Dirichlet B.C.		Neumann B.C.	
N	$\|u-v\|_h$	q_D	$\|u-v\|_h$	q_N	$\|u-v\|_h$	q_D	$\|u-v\|_h$	q_N
10	$7.02\cdot10^{-1}$		$1.34\cdot10^{-2}$		$2.36\cdot10^{-1}$		$4.64\cdot10^{-3}$	
20	$1.01\cdot10^{-1}$	2.79	$3.12\cdot10^{-3}$	2.10	$3.00\cdot10^{-2}$	2.97	$8.87\cdot10^{-4}$	2.39
40	$1.56\cdot10^{-2}$	2.69	$7.54\cdot10^{-4}$	2.05	$3.07\cdot10^{-3}$	3.29	$1.26\cdot10^{-4}$	2.82
80	$2.56\cdot10^{-3}$	2.61	$1.85\cdot10^{-4}$	2.03	$2.58\cdot10^{-4}$	3.57	$1.20\cdot10^{-5}$	3.39
160	$4.53\cdot10^{-4}$	2.50	$4.60\cdot10^{-5}$	2.01	$1.88\cdot10^{-5}$	3.78	$8.52\cdot10^{-7}$	3.81
320	$8.88\cdot10^{-4}$	2.35	$1.15\cdot10^{-5}$	2.00	$1.26\cdot10^{-6}$	3.91	$5.09\cdot10^{-8}$	4.07

Table 8: Numerical results for the convection diffusion equation ($c \equiv 0$) using Dirichlet and Neumann boundary conditions with the second and fourth order accurate schemes

4 Self-Adjoint Form

Equation (1.1) is in self-adjoint form, which makes it possible to obtain an energy estimate by simply applying integration by parts. In this section we derive operators which imitate this property in the sense that they allow us to apply a discrete analogon.

4.1 Properties of the Operator

We want to devise a general form for an operator that mimics the summation by parts property for the parabolic term in (2.2). Let $2p$ be the order of accuracy in the interior and p the accuracy at the boundary.

In the continuous case we have the $\mathcal{L}_2$-inner product and the operator $\frac{\partial}{\partial x}$. Integration by parts gives

$$(u, \frac{\partial}{\partial x} a \frac{\partial}{\partial x} u) = -(\frac{\partial}{\partial x} u, a \frac{\partial}{\partial x} u) + \text{boundary term} \tag{4.1}$$

If we want to mimic this behavior, we have to define two discrete operators Q_1, Q_2 approximating $\frac{\partial}{\partial x}$, i.e. in this case we are looking for two separate operators for the two first derivatives, not for one operator for the whole problem. We also want to include the physical boundary data using SAT. Thus we look at a semi-discrete scheme of the form

$$v_t = Q_1 a Q_2 v + \text{boundary term} \tag{4.2}$$

In order to achieve the accuracy in the interior even for variable a, the inner part of Q_1 has to be $2p$th order accurate. Since in the interior we want to have that $Q_1^T = -Q_2$, also Q_2 will have to be $2p$th order accurate.
We want to obtain a discrete analogon of (4.1). In the discrete case the $\mathcal{L}_2$ inner product is replaced by the inner product induced by H, where H is the diagonal norm introduced in [16]. Thus we want to have

$$(v, Q_1 a Q_2 v)_H = -(Q_2 v, a Q_2 v)_H + \text{boundary term} \tag{4.3}$$

We can transform the left hand side:

$$\begin{aligned}(v, Q_1 a Q_2 v)_H &= v^T H Q_1 a Q_2 v = v^T H Q_1 a H^{-1} H Q_2 v \\ &= \left((H Q_1 H^{-1})^T v\right)^T a H Q_2 v = ((H Q_1 H^{-1})^T v, a Q_2 v)_H\end{aligned}$$

Thus Q_1 and Q_2 shall be connected in the following way:

$$H Q_1 = -(H Q_2)^T + \text{boundary term}$$

This relation can be generalized with a positive definite diagonal matrix K

$$H Q_1 = -(H K Q_2)^T + \text{boundary term}$$

without loosing the essential property of (4.3) which becomes

$$(v, Q_1 a Q_2 v)_H = -(K Q_2 v, a Q_2 v)_H + \text{boundary term} \tag{4.4}$$

Since $a(x,t) \geq a_{\min} > 0$ and K is positive definite and diagonal, it still holds

$$-(K Q_2 v, a Q_2 v)_H = -(K Q_2 v, a K^{-1} K Q_2 v) \leq 0$$

which was the desired property of the summation by parts rule. Thus we can obtain an energy estimate in the H-norm when neglecting the boundary part

$$\begin{aligned}\frac{d}{dt}\|v\|_H^2 - \text{boundary term} &= v^T H Q_1 a Q_2 v + (Q_1 a Q_2 v)^T H v \\ &= -(KQ_2v)^T \left(aK^{-1} + (K^{-1})^T a\right)(KQ_2v) \leq 0.\end{aligned}$$

We now look at the additional boundary part. For summation by parts operators we implement the *Robin physical boundary conditions* by a SAT term of the form

$$-\tau_0 H^{-1}(E_0(\beta_0 I + S)v - e_0 a_0 g_0(t)) - \tau_N H^{-1}(E_N(\beta_1 I + S)v - e_N a_N g_1(t)), \tag{4.5}$$

where S is an approximation of $a\frac{\partial}{\partial x}$ in the first and last line. Its accuracy can be one order less than the global order of the summation by parts operator without effecting its accuracy. *Dirichlet boundary conditions* are implemented using a SAT term of the form

$$-\tau_0 H^{-1} S^T(E_0 v - e_0 g_0(t)) - \tau_N H^{-1} S^T(E_N v - e_N g_1(t)), \tag{4.6}$$

where S is an approximation of $a\frac{\partial}{\partial x}$ at x_0 and x_N as above. When calculating the time derivative of $\|v\|_H^2$, we get for both Dirichlet and Robin boundary conditions the same boundary term $v^T E_i S v + v^T S^T E_i v$, $i = 0, N$, which is why in the following we only consider Robin boundary conditions where the additional terms with β_0, β_1 occur.

Now we want to calculate the energy estimate for the whole scheme including the boundary part. We make up the semi-discrete scheme by the SAT boundary part, i.e.

$$\begin{aligned}v_t = {} & Q_1 a Q_2 v - \tau_0 H^{-1}(E_0(\beta_0 I + S)v - e_0 a_0 g_0(t)) \\ & - \tau_N H^{-1}(E_N(\beta_1 I + S)v - e_N a_N g_1(t)).\end{aligned}$$

At the inner points Q_1 and Q_2 shall approximate $\frac{\partial}{\partial x}$ $2p$th order accurate. At the boundary we cannot achieve the same accuracy. Furthermore we allow some asymmetry represented by a matrix R and require only the total operator $Q_1(\cdot)Q_2$ to be a pth order accurate approximation of $\frac{\partial}{\partial x}\left(\cdot \frac{\partial}{\partial x}\right)$. Let r be the size of the boundary part. Then R has only non-zero entries in an $r \times r$ part in the upper left and lower right corner. It is defined through

$$HQ_1 = -(HKQ_2)^T + R, \tag{4.7}$$

This yields the following energy estimate

$$\begin{aligned}\frac{d}{dt}\|v\|_H^2 &= -(KQ_2v)^T(aK^{-1} + (K^{-1})^T a)(KQ_2v) + v^T R a Q_2 v + v^T Q_2^T a R^T v \\ &\quad - 2\tau_0 v_0(\beta_0 v_0 + (Sv)_0 - a_0 g_0(t)) - 2\tau_1 v_N(\beta_1 v_N + (Sv)_N - a_N g_1(t)) \\ &= -(KQ_2v)^T(aK^{-1} + (K^{-1})^T a)(KQ_2v) + 2v^T R a Q_2 v \\ &\quad - 2\tau_0\beta_0\left(v_0 - \frac{1}{2\tau_0\beta_0} a_N g_0(t)\right)^2 - 2\tau_1\beta_1\left(v_N - \frac{1}{2\tau_1\beta_1} a_N g_1(t)\right)^2 \\ &\quad + \frac{\tau_0}{2\beta_0} a_0^2 g_0(t)^2 + \frac{\tau_1}{2\beta_1} a_N^2 g_1(t)^2 - 2v_0(Sv)_0\tau_0 - 2v_N(Sv)_N\tau_1.\end{aligned}$$

We can obtain an estimate if R satisfies

$$\begin{aligned}& v_0\left(\tau_0(Sv)_0 - (RaQ_2v)_0\right) = C_0 v_0^2,\ C_0 \geq 0, \\ & v_i(RaQ_2v)_i = -C_i v_i^2,\ C_i \geq 0, \qquad 1 \leq i \leq r \ \vee\ N - r \leq i \leq N - 1, \\ & v_N\left(\tau_N(Sv)_N - (RaQ_2v)_N\right) = C_N v_N^2,\ C_N \geq 0.\end{aligned} \tag{4.8}$$

The easiest way to obtain this is to choose the summation by parts operator for the first derivative derived by Strand [16] for both Q_1 and Q_2. However, this procedure has two drawbacks: First the operator for the whole problem has quite a wide stencil and relatively large error constants and second the π-modes are not damped.
It is not possible to approximate the first derivative to some order $2p$ using less points than Strand on a normal grid. However one can reduce the bandwidth by using a staggered grid. The idea is now to design two operators with the following properties:

- Q_1 approximates the first derivative at the point $x_{j+1/2}$ in the interior ($2p$th order accurate)
- Q_2 approximates the first derivative at the point $x_{j-1/2}$ in the interior ($2p$th order accurate)
- $Q_a = Q_1(\cdot)Q_2$ approximates $\frac{\partial}{\partial x}\left(\cdot \frac{\partial}{\partial x}\right)$ at the boundary (pth order accurate)
- relation (4.7) holds and
- R is such that (4.8) holds.

This requires that we apply Q_1 on a at intermediate points $x_{j-1/2}$ while u is approximated at the actual grid points.
Combining Q_1 and Q_2 with a SAT treatment of the boundary data leads to a strictly stable high order approximation (even for non-smooth a) for the semi-discretization (1.1).

4.2 Summation by Parts Operator of Order 2

4.2.1 Construction

In this section we will construct a summation by parts operator that is second order accurate in the interior and first order accurate on the boundary. We will not follow the general procedure derived in section 4.1 here but make the connection to that afterwards.

The following operator is a discretization of the spatial part of (1.1) imitating the summation by parts property of the continuous problem.
Let $a_{j-1/2} = a(x_j - \frac{1}{2}h, t), \quad j = 1, ..., N$ and $a_{-1/2} = 0$, $\Lambda = \text{diag}([a_{-1/2}, a_{1/2}, \ldots, a_{N-1/2}])$,

$$D_- = \frac{1}{h}\begin{pmatrix} 1 & 0 & \ldots & & & 0 \\ -1 & 1 & 0 & \ldots & & \vdots \\ 0 & -1 & 1 & 0 & \ldots & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \\ & \ldots & 0 & -1 & 1 & 0 \\ 0 & & \ldots & 0 & -1 & 1 \end{pmatrix},$$

$$D_+ = \frac{1}{h}\begin{pmatrix} -1 & 1 & 0 & \ldots & & 0 \\ 0 & -1 & 1 & 0 & \ldots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \\ & \ldots & 0 & -1 & 1 & 0 \\ & & \ldots & 0 & -1 & 1 \\ 0 & & & \ldots & 0 & -1 \end{pmatrix}$$

and

$$BS = \frac{1}{h}\begin{pmatrix} \frac{3}{2}a_{1/2} & -(\frac{3}{2}a_{1/2}+\frac{1}{2}a_{3/2}) & \frac{1}{2}a_{3/2} & & & \\ & 0 & & & & \\ & & \ddots & & & \\ & & & 0 & & \\ & & \frac{1}{2}a_{N-3/2} & -(\frac{3}{2}a_{N-1/2}+\frac{1}{2}a_{N-3/2}) & \frac{3}{2}a_{N-1/2} & \end{pmatrix}.$$

Define $Q_a = H^{-1}(-A+BS)$, where $-A = D_+\Lambda D_-$. Hence

$$Q_a = \frac{1}{h^2}\begin{pmatrix} a_{1/2} & -\widetilde{a}_1 & a_{3/2} & 0 & & & \\ a_{1/2} & -\widetilde{a}_1 & a_{3/2} & 0 & & & \\ 0 & a_{3/2} & -\widetilde{a}_2 & a_{5/2} & 0 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 0 & a_{N-5/2} & -\widetilde{a}_{N-2} & a_{N-3/2} & 0 \\ & & & 0 & a_{N-3/2} & -\widetilde{a}_{N-1} & a_{N-1/2} \\ & & & 0 & a_{N-3/2} & -\widetilde{a}_{N-1} & a_{N-1/2} \end{pmatrix}, \tag{4.9}$$

where $\widetilde{a}_j = a_{j-1/2} + a_{j+1/2}, \quad j = 1, \ldots, N-1.$
This is a discretization of $(au_x)_x$ that is stable and second order accurate in the interior and first order accurate at the boundary as we will see in the next sections.

Before showing strict stability and the accuracy, we want to complete this section by looking at the operator in terms of the notation of section 4.1.

- $Q_1 = D_+ + \frac{1}{h}\begin{pmatrix} 0 & -2 & 1 & 0 & \\ & 0 & & & \\ & & \ddots & & \\ & & & 0 & \\ & & 0 & -1 & 2 \end{pmatrix}$
- $Q_2 = D_-$
- $K = I$
- $R = \begin{pmatrix} -1 & \frac{1}{2} & 0 & \\ 0 & & & \\ & \ddots & & \\ & & 0 & \\ & 0 & -\frac{1}{2} & 1 \end{pmatrix}.$

4.2.2 Consistency

Consider the jth line of the operator for $j \neq 0, N$. A Taylor expansion of $u(x_{j-1})$ and $u(x_{j+1})$ around x_j yields

$$\begin{aligned}
&\frac{1}{h^2}(a_{j-1/2}u_{j-1} - (a_{j-1/2} + a_{j+1/2})u_j + a_{j+1/2}u_{j+1}) \\
&\quad = \frac{1}{h^2}(a_{j-1/2} - a_{j-1/2} - a_{j+1/2} + a_{j+1/2})u_j + \frac{1}{h}(-a_{j-1/2} + a_{j+1/2})u_x(x_j) \\
&\qquad + \frac{a_{j-1/2} + a_{j+1/2}}{2}u_{xx}(x_j) + \frac{h^2}{6}\frac{a_{j+1/2} - a_{j-1/2}}{h}u_{xxx}(x_j) + \mathcal{O}(h^2) \\
&\quad = (a_x(x_j) + \mathcal{O}(h^2)))u_x(x_j) + (a(x_j) + \mathcal{O}(h^2))u_{xx}(x_j) \\
&\qquad + \frac{h^2}{6}(a_x(x_j) + \mathcal{O}(h^2))u_{xxx}(x_j) + \mathcal{O}(h^2) \\
&\quad = a_x(x_j)u_x(x_j) + a(x_j)u_{xx}(x_j) + \mathcal{O}(h^2)
\end{aligned} \tag{4.10}$$

For the second equality the following identities are obtained by Taylor expansion of a around x_j:

$$\begin{aligned}
\frac{a_{j+1/2} - a_{j-1/2}}{h} &= \frac{1}{h}(1-1)a(x_j) + a_x(x_j) + h(1-1)a_{xx}(x_j) + \mathcal{O}(h^2) \\
&= a_x(x_j) + \mathcal{O}(h^2) \\
\frac{a_{j+1/2} + a_{j-1/2}}{2} &= \frac{1}{2}(1+1)a(x_j) + \frac{1}{2}\left(\frac{h}{2} - \frac{h}{2}\right)a_x(x_j) + \mathcal{O}(h^2) \\
&= a(x_j) + \mathcal{O}(h^2)
\end{aligned}$$

Now we look at the accuracy at the boundary:

$$\begin{aligned}
&\frac{a_{1/2}u_0 - (a_{1/2} + a_{3/2})u_1 + a_{3/2}u_2}{h^2} = \frac{a_{3/2} - a_{1/2}}{h}u_x(x_0) + \frac{3a_{3/2} - a_{1/2}}{2}u_{xx}(x_0) + \mathcal{O}(h) \\
&\quad = (a_x(x_0) + \mathcal{O}(h))u_x(x_0) + (a_0 + \mathcal{O}(h))u_{xx}(x_0) + \mathcal{O}(h) = (au_x)_x|_{x_0} + \mathcal{O}(h)
\end{aligned}$$

for the left boundary and

$$\begin{aligned}
&\frac{a_{N-1/2}u_N - (a_{N-1/2} + a_{N-3/2})u_{N-1} + a_{N-3/2}u_{N-2}}{h^2} \\
&\quad = \frac{a_{N-3/2} - a_{N-1/2}}{h}u_x(x_N) + \frac{3a_{N-3/2} - a_{N-1/2}}{2}u_{xx}(x_N) + \mathcal{O}(h) \\
&\quad = (a_x(x_N) + \mathcal{O}(h))u_x(x_N) + (a_N + \mathcal{O}(h))u_{xx}(x_N) + \mathcal{O}(h) = (au_x)_x|_{x_N} + \mathcal{O}(h)
\end{aligned}$$

for the right boundary. Thus the operator is second order accurate in the interior and first order accurate at the boundary.
The boundary values can be implemented using the SAT technique. Therefore we show that the first line of S is a second order accurate approximation of $(au_x)|_{x_0}$ and for the last row analogous:

$$\begin{aligned}
&\frac{-3a_{1/2}u_0 + (3a_{1/2} + a_{3/2})u_1 - a_{3/2}u_2}{2h} \\
&\quad = \frac{3a_{1/2} - a_{3/2}}{2}u_x(x_0) + \frac{3a_{1/2} - 3a_{3/2}}{2}\frac{h}{2}u_{xx}(x_0) + \mathcal{O}\left(h^2\right) \\
&\quad = \left(a(x_0) + \mathcal{O}\left(h^2\right)\right)u_x(x_0) + h^2\left(\frac{3}{4}a_x(x_1) + \mathcal{O}\left(h^2\right)\right)u_{xx}(x_0) + \mathcal{O}\left(h^2\right) \\
&\quad = a(x_0)u_x(x_0) + \mathcal{O}\left(h^2\right)
\end{aligned}$$

4.2.3 Stability

Consider the initial boundary value problem (2.2) with $b \equiv c \equiv 0$. This can be approximated using the above operator and the SAT method for the treatment of the boundary conditions:

$$\begin{aligned} v_t = {} & H^{-1}(-A + BS)v - H^{-1}\tau_0 \left(E_0(\beta_0 I + S)v - e_0 g_0(t)\right) \\ & - H^{-1}\tau_1 \left(E_N(\beta_1 I + S)v - e_N g_1(t)\right), \quad v(0) = f. \end{aligned} \tag{4.11}$$

The energy method applied on this approximation leads to

$$\begin{aligned} \frac{d}{dt}\|v\|_H^2 = {} & -v^T(A + A^T)v + v^T(BS + (BS)^T)v \\ & - 2\tau_0 v_0 \left(\beta_0 v_0 + (Sv)_0 - a_0 g_0(t)\right) - 2\tau_1 v_N \left(\beta_1 v_N + (Sv)_N - a_N g_1(t)\right). \end{aligned}$$

Using the above notation gives

$$\begin{aligned} \frac{d}{dt}\|v\|_H^2 = {} & v^T \left(D_+ \Lambda D_- + (D_+ \Lambda D_-)^T\right) v \\ & - 2v_0(Sv)_0 + 2v_N(Sv)_N - 2\tau_0 v_0(\beta_0 v_0 + (Sv)_0 - a_0 g_0(t) \\ & - 2\tau_1 v_N(\beta_1 v_N - (Sv)_N + a_N g_1(t) \\ = {} & -2(D_- v)^T \Lambda (D_- v) - 2\tau_0\beta_0 \left(v_0 - \frac{a_0}{2\beta_0} g_0(t)\right)^2 \\ & + \frac{\tau_0}{2\beta_0} a_0^2 g_0(t)^2 - 2\tau_1\beta_1 \left(v_N - \frac{a_N}{2\beta_1} g_1(t)\right)^2 + \frac{\tau_1}{2\beta_1} a_N^2 g_1(t)^2 \\ & - 2v_0(Sv)_0(1 + \tau_0) + 2v_N(Sv)_N(1 - \tau_1) \end{aligned} \tag{4.12}$$

Hence an energy estimate exists if the condition

$$\beta_0 \leq 0 \quad \text{and} \quad \beta_1 \geq 0$$

(compare also (3.20) for $c \equiv 0$) holds and additionally:

$$\tau_0 = -1 \quad \text{and} \quad \tau_1 = 1. \tag{4.13}$$

4.3 Summation by Parts Operators of Higher Order

The system that has to be solved for the boundary part of an SBP-operator of order $2p$, where $p \geq 2$, is rather complicated since it is non-linear. Therefore we did not solve the system and cannot tell how accurate the boundary part can be designed.
A possible inner stencil is a minimal width stencil of order $2p$ on a staggered grid, which can be calculated using a formula given by Fornberg [4].

As an example we give the inner stencil for Q_1 and Q_2, respectively, in the case of $2p = 4$:

$$(Q_1 v)_i = \frac{1}{24} v_{j-1} - \frac{8}{9} v_j + \frac{8}{9} v_{j+1} - \frac{1}{24} v_{j+2} \tag{4.14}$$

$$(Q_2 v)_i = \frac{1}{24} v_{j-2} - \frac{8}{9} v_{j-1} + \frac{8}{9} v_j - \frac{1}{24} v_{j+1} \tag{4.15}$$

This leads to the following stencil for the whole operator $Q_a = Q_1 a Q_2 v$ in the interior:

$$\begin{aligned} & \frac{1}{576} a_{j-3/2} v_{j-3} - \frac{3}{64}\Big(a_{j-3/2} + a_{j-1/2}\Big) v_{j-2} + \frac{3}{64}\Big(a_{j-3/2} + 27a_{j-1/2} + a_{j+1/2}\Big) v_{j-1} \\ & - \frac{1}{64}\Big(\frac{1}{9} a_{j-3/2} + 81 a_{j-1/2} + 81 a_{j+1/2} + \frac{1}{9} a_{j+3/2}\Big) v_j + \\ & \frac{3}{64}\Big(a_{j-1/2} + 27a_{j+1/2} + a_{j+3/2}\Big) v_{j+1} - \frac{3}{64}\Big(a_{j+1/2} + a_{j+3/2}\Big) v_{j+2} + \frac{1}{576} a_{j+3/2} v_{j+3} \end{aligned}$$

4.4 Damping of π-Modes

As done in section 3.4 for the operator based on the product rule, we want to consider the damping of different wave numbers also for the operators presented above.

For a constant coefficient a, the second order operator Q_a in (4.9) reduces to the second order operator proposed by Mattsson and Nordström [11]. Hence the Fourier transform $\widehat{Q}_2$ looks the same as for the D_2 operator, which is already presented in 3.4

$$\widehat{Q}_2 = \frac{2}{h^2}\left(-1+\cos(\xi)\right), \qquad \xi = 2\pi\omega h.$$

Since we only look at the inner stencil, we can also consider the 4th order scheme, which we denote by Q_4. The Fourier transform for $a \equiv 1$ is given by:

$$\widehat{Q}_4 = \frac{1}{h^2}\left(-\frac{365}{144}+\frac{87}{32}\cos(\xi)-\frac{3}{16}\cos(2x)+\frac{1}{288}\cos(3x)\right), \qquad \xi = 2\pi\omega h.$$

Figure 2 shows the damping of different wave numbers by the operators compared with the

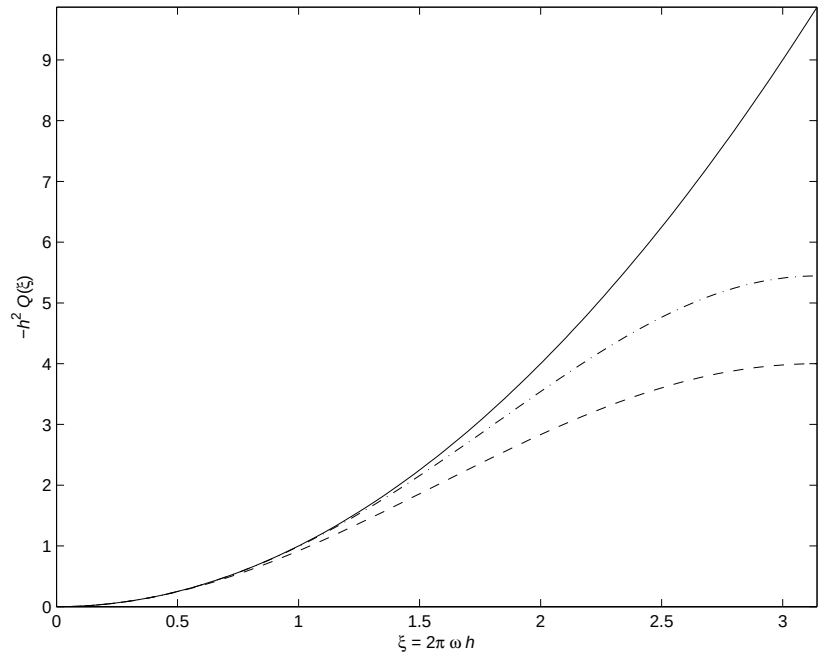

Figure 2: Damping of modes with wave number ξ of the exact solution (-), 2nd (- -) and 4th (-·) order accurate operator Q.

damping of the exact solution. Like for the operators based on the product rule, the damping gets worse for growing wave numbers. But at least all wave numbers are damped to some amount.

4.5 Computational Results

We now present the results of some numerical experiments we have performed on equation (1.1) using the second operator derived above for the space discretization and the classical fourth order Runge-Kutta for time integration.
The stability limit is the same as the one for the second order operator based on the product rule, since both operators are reduced to the same operator for constant a. If we assume a to be Lipschitz continuous the a depending coefficients of the term in the Fourier transform

of the a depending operator can by estimated by $a(x_j) + \mathcal{O}(h^2)$. Therefore we use the same time steps as in section (3.5.1).
We choose the same examples as in section 4. In the cases where we do not know the exact solution, we use the solution calculated with the product rule operator of 4th order on a grid with 320 points.

Example 1 We consider the problem

$$
\begin{aligned}
&u_t = 0.2u_{xx} \\
&u(x,0) = \sin(\pi x) + \cos(\pi x) \\
&\text{Dirichlet: } u_x(0,t) = \mathrm{e}^{-0.2\pi^2 t}, \quad u_x(1,t) = -\mathrm{e}^{-0.2\pi^2 t} \text{ or} \\
&\text{Neumann: } u_x(0,t) = \pi\mathrm{e}^{-0.2\pi^2 t}, \quad u_x(1,t) = -\pi\mathrm{e}^{-0.2\pi^2 t}
\end{aligned}
$$

with the exact solution

$$u(x,t) = (\sin(\pi x) + \cos(\pi x))e^{-0.2\pi^2 t}.$$

In this case a is constant, which is why the operator is of the same form as the second order method based on the product rule. The result are hence given in table 2.

Example 2 We consider the problem

$$
\begin{aligned}
&u_t = \frac{\partial}{\partial x}\left((0.2 + 0.4x(x-1))u_x\right) \\
&u(x,0) = \sin(\pi x) + \cos(\pi x) \\
&\text{Dirichlet: } u(0,t) = \mathrm{e}^{-0.2\pi(\pi+2)t}, \quad u(1,t) = -\mathrm{e}^{-0.2\pi(\pi-2)t} \text{ or} \\
&\text{Neumann: } u_x(0,t) = \pi\mathrm{e}^{-0.2\pi(\pi+2)t}, \quad u_x(1,t) = -\pi\mathrm{e}^{-0.2\pi(\pi-2)t}.
\end{aligned}
$$

Table 9 shows the results at time $t = 1$.

	Dirichlet		Neumann	
N	$\|u-v\|_h$	q_D	$\|u-v\|_h$	q_N
10	$2.38 \cdot 10^{-3}$		$5.03 \cdot 10^{-4}$	
20	$5.24 \cdot 10^{-4}$	2.18	$1.19 \cdot 10^{-3}$	2.08
40	$1.63 \cdot 10^{-4}$	1.68	$2.88 \cdot 10^{-4}$	2.04
80	$4.59 \cdot 10^{-5}$	1.83	$7.10 \cdot 10^{-5}$	2.02
160	$1.21 \cdot 10^{-5}$	1.92	$1.76 \cdot 10^{-5}$	2.01
320	$3.21 \cdot 10^{-6}$	1.96	$4.38 \cdot 10^{-6}$	2.01

Table 9: Numerical results for the equation $u_t = ((0.2 + 0.4x(x-1))u_x)_x$

Example 3 We consider the problem

$$
\begin{aligned}
&u_t = \frac{\partial}{\partial x}\left(0.2(1 + \sin(\pi x))u_x\right) \\
&u(x,0) = \sin(\pi x) + \cos(\pi x) \\
&\text{Dirichlet: } u(0,t) = 1, \quad u(1,t) = -\mathrm{e}^{-0.4\pi^2 t} \text{ or} \\
&\text{Neumann: } u_x(0,t) = \pi, \quad u_x(1,t) = -\pi\mathrm{e}^{-0.4\pi^2 t}.
\end{aligned}
$$

	Dirichlet		Neumann	
N	$\|u-v\|_h$	q_D	$\|u-v\|_h$	q_N
10	$8.14 \cdot 10^{-4}$		$3.69 \cdot 10^{-3}$	
20	$7.22 \cdot 10^{-5}$	3.50	$8.77 \cdot 10^{-4}$	2.07
40	$2.78 \cdot 10^{-5}$	1.38	$2.14 \cdot 10^{-4}$	2.03
80	$9.81 \cdot 10^{-6}$	1.50	$5.30 \cdot 10^{-5}$	2.02
160	$2.85 \cdot 10^{-6}$	1.78	$1.32 \cdot 10^{-5}$	2.01
320	$7.66 \cdot 10^{-7}$	1.90	$3.29 \cdot 10^{-6}$	2.00

Table 10: Numerical results for the equation $u_t = (0.2(1+\sin(\pi x))u_x)_x$

Table 10 shows the results at time $t = 1$.

Example 4 We consider the problem

$$u_t = \frac{\partial}{\partial x}\left(0.5\mathrm{e}^{x-2}(1+\sin(\pi t))u_x\right) \quad \text{with} \quad u(x,0) = (\sin(\pi x))^2 + 2x$$

Dirichlet: $u(0,t) = 0, \quad u(1,t) = 2\mathrm{e}^{e^{-1}(\pi^2+1)t}$ or

Neumann: $u_x(0,t) = u_x(0,t) = 2\,\mathrm{e}^{\mathrm{e}^{-2}(\pi^2+1)t}, \quad u_x(1,t) = 2\,\mathrm{e}^{e^{-1}(\pi^2+1)t}.$

Table 11 shows the results at time $t = 1$.

	Dirichlet		Neumann	
N	$\|u-v\|_h$	q_D	$\|u-v\|_h$	q_N
10	1.21		$2.14 \cdot 10^{-1}$	
20	$1.55 \cdot 10^{-1}$	2.96	$4.86 \cdot 10^{-2}$	2.14
40	$1.82 \cdot 10^{-2}$	3.10	$1.14 \cdot 10^{-2}$	2.09
80	$2.36 \cdot 10^{-3}$	2.94	$2.75 \cdot 10^{-3}$	2.05
160	$5.12 \cdot 10^{-4}$	2.21	$6.75 \cdot 10^{-4}$	2.03
320	$1.41 \cdot 10^{-4}$	1.86	$1.67 \cdot 10^{-4}$	2.01

Table 11: Numerical results for the equation $u_t = \left(0.5\mathrm{e}^{x-2}(1+\sin(\pi t))u_x\right)_x$

The experiments show that the solution converges with second order, which is in accordance with the theoretical analysis.

4.6 Extension to the Convection Diffusion Equation

The approximation of the parabolic term $(au_x)_x$ can be used in an approximation of the convection diffusion equation (1.2) combined with the operator D_1 derived by Strand [16] for the hyperbolic terms. However, the straight-forward implementation of the equation in the form $v_t = Q_a v + \Lambda_b D_1 v + D_1(\Lambda_c v)$ is not optimal since it does not lead to a strictly stable approximation if we use a higher order ($2p > 2$) operator D_1 as an approximation of the first derivative (cf. [14]). Hence we assume that b and c are both differentiable with respect to x and use the identities $bu_x = \frac{1}{2}\big((bu)_x + bu_x - b_x u\big)$ as well as $(cu)_x = \frac{1}{2}\big((cu)_x + cu_x + c_x u\big)$. Denote $d(x,t) = \frac{1}{2}(b(x,t) + c(x,t))$, $\Lambda = \mathrm{diag}([d(x_0,t),\ldots,d(x_N,t)])$ the diagonal matrix containing the values of d and $\Gamma = \frac{1}{2}\,\mathrm{diag}([-b_x(x_0,t) + c_x(x_0,t),\ldots,-b_x(x_N,t) + c_x(x_N,t)])$. Then (1.2) can be approximated by (cf. [14])

$$v_t = Q_a v + \Lambda D_1 v + D_1 \Lambda v + \Gamma v, \tag{4.16}$$

where Q_a is the operator derived in the previous sections that approximates $\frac{\partial}{\partial x}a\frac{\partial}{\partial x}$. The derivative $-b_x + c_x$ can either be given or approximated to sufficiently high order.

4.6.1 Stability

To prove stability, we use the energy method. Before giving details on the implementation of boundary conditions, we analyze the terms arising from the semi-discretization in the interior.

Taking the discrete H-norm of some vector $v \in \mathbb{R}^{N+1}$ leads to

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|v\|_H^2 &= (v, v_t)_H + (v_t, v)_H \\ &= (v, Q_1 a Q_2 v)_H + (Q_1 a Q_2 v, v)_H + (v, \Lambda D_1 v)_H + (\Lambda D_1 v, v)_H \\ &\quad + (v, D_1 \Lambda v)_H + (D_1 \Lambda v, v)_H + (v, \Gamma v)_H + (\Gamma v, v)_H.\end{aligned}$$

We use property (4.4) for the parabolic terms with the boundary term BS. Then,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|v\|_H^2 &= -(KQ_2 v, aQ_2 v)_H - (aQ_2 v, KQ_2 v)_H + v^T \left(BS + (BS)^T\right) v \\ &\quad + v^T H \Lambda D_1 v + v^T D_1^T \Lambda H v + v^T H D_1 \Lambda v + v^T \Lambda D_1^T H v + 2v^T \Gamma v.\end{aligned}$$

For the terms $v^T H D_1 \Lambda v$ and $v^T \Lambda D_1^T H v$, we use $HD_1 = Q = -Q^T + B = -D_1^T H + B$ and $D_1^T H = Q^T = -Q + B = -HD_1 + B$, respectively. This leads to

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|v\|_H^2 &= -(KQ_2 v, aQ_2 v)_H - (aQ_2 v, KQ_2 v)_H + v^T \left(BS + (BS)^T\right) v \\ &\quad + v^T H \Lambda D_1 v - v^T D_1^T H \Lambda v + v^T D_1^T \Lambda H v - v^T \Lambda H D_1 v \\ &\quad + v^T B \Lambda v + v^T \Lambda B v + 2v^T \Gamma v.\end{aligned}$$

Since both H and Λ are diagonal matrices, they commute, and we get

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|v\|_H^2 &= -(KQ_2 v, aQ_2 v)_H - (aQ_2 v, KQ_2 v)_H + v^T \left(BS + (BS)^T\right) v \\ &\quad + v^T H \Lambda D_1 v - \left(v^T H \Lambda D_1 v\right)^T + v^T D_1^T \Lambda H v - \left(v^T D_1^T \Lambda H v\right)^T \\ &\quad + v^T B \Lambda v + v^T \Lambda B v + 2v^T \Gamma v.\end{aligned}$$

The terms $v^T H \Lambda D_1 v$ and $v^T D_1^T \Lambda H v$ are real numbers, which is why they equal their transpose. This means that they cancel each other. Hence

$$\frac{\mathrm{d}}{\mathrm{d}t}\|v\|_H^2 = -2(KQ_2 v, aQ_2 v)_H + 2v^T BSv + 2v^T B\Lambda v + 2v^T \Gamma v. \tag{4.17}$$

In the following we consider both types of boundary conditions separately.

Dirichlet boundary conditions are implemented by the SAT terms

$$H^{-1}(\tau_0 S^T + \sigma_0 I)(E_0 v - e_0 g_0) \quad \text{and} \quad H^{-1}(\tau_1 S^T + \sigma_1 I)(E_N v - e_N g_1),$$

where τ_0, τ_1, σ_0 and σ_1 are constants that are to be determined such that the approximation is strictly stable. These boundary terms are just those in (4.6) extended by the additional terms $\sigma_i I$, $i = 0, 1$ arising from the approximation of hyperbolic terms.
The approximation of (2.1) has the form

$$\begin{aligned}v_t = Q_a v + \Lambda D_1 v + D_1 \Lambda v + \Gamma v &- H^{-1}\left(\tau_0 S^T + \sigma_0 I\right)(E_0 v - e_0 g_0) \\ &- H^{-1}\left(\tau_1 S^T + \sigma_1 I\right)(E_N v - e_N g_1).\end{aligned} \tag{4.18}$$

For showing stability, we set homogeneous boundary conditions as for the error estimate in the continuous case in section 2.2.1 and also used in section 3.2.2 for the ansatz with the product rule.
Taking the time-derivative of the discrete H-norm of v leads to

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|v\|_H^2 &= -2(KQ_2v, aQ_2v)_H + 2v^TBSv + 2v^TB\Lambda v + 2v^T\Gamma v - 2\tau_0 v^TS^TE_0v \\ &\quad - 2\tau_1 v^TS^TE_Nv - 2\sigma_0 v^TE_0v - 2\sigma_1 v^TE_Nv \\ &= -2(KQ_2v, aQ_2v)_H + 2v^T\Gamma v + 2v_0(Sv)_0(-1-\tau_0) + 2v_N(Sv)_N(1-\tau_1) \\ &\quad + 2v_0^2(-d_0-\sigma_0) + 2v_N^2(d_N-\sigma_1).\end{aligned}$$

If we require

$$\tau_0 = -1, \qquad \tau_1 = 1, \qquad \sigma_0 = -d_0 = -\frac{1}{2}(b_0+c_0), \qquad \sigma_1 = d_N = \frac{1}{2}(b_N+c_N),$$

we get the estimate

$$\frac{\mathrm{d}}{\mathrm{d}t}\|v\|_H^2 \leq \alpha_s \|v\|_H^2,$$

where $\alpha_s = \|(-b+c)_x\|_\infty$, which is exactly the same constant as α in the continuous case. Hence (4.18) is a strictly stable approximation of (2.1).

Robin boundary conditions are implemented by the SAT terms

$$H^{-1}\tau_0(E_0(a_0\beta_0 I + S)v - a_0e_0g_0) \quad \text{and} \quad H^{-1}\tau_1(E_N(a_N\beta_1 I + S)v - a_Ne_Ng_1),$$

where τ_0 and τ_1 are constants that are to be determined such that the approximation is stable. These boundary terms are exactly those in (4.5).
The approximation of (2.2) has the form

$$\begin{aligned}v_t &= Q_av + \Lambda D_1v + D_1\Lambda v + \Gamma v - H^{-1}\tau_0(E_0(a_0\beta_0 I + S)v - a_0e_0g_0) \\ &\quad - H^{-1}\tau_1(E_N(a_N\beta_1 I + S)v - a_Ne_Ng_1).\end{aligned}$$

Applying the energy method to v gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|v\|_H^2 &= -2(KQ_2v, aQ_2v)_H + 2v^TBSv + 2v^TB\Lambda v + 2v^T\Gamma v \\ &\quad - 2\left(\tau_0 v^TE_0Sv + \tau_1 v^TE_NSv\right) \\ &\quad - 2\left(\tau_0\beta_0a_0v^TE_0v + \tau_1\beta_1a_Nv^TE_Nv + \tau_0a_0v_0g_0 + \tau_1a_Nv_Ng_1\right) \\ &= -2(KQ_2v, aQ_2v)_H + 2v^T\Gamma v + 2v_0(Sv)_0(-1-\tau_0) + 2v_N(Sv)_N(1-\tau_1) \\ &\quad + 2(-d_0-\tau_0\beta_0a_0)\left(v_0 - \frac{\tau_0a_0}{-2(d_0+\tau_0\beta_0a_0)}g_0\right)^2 - \frac{\tau_0^2a_0^2}{-2(d_0+\tau_0\beta_0a_0)}g_0^2 \\ &\quad + 2(d_N-\tau_1\beta_1a_N)\left(v_N - \frac{\tau_1a_N}{2(d_N-\tau_1\beta_1a_N)}g_1\right)^2 - \frac{\tau_1^2a_N^2}{2(d_N-\tau_1\beta_1a_N)}g_2^2.\end{aligned}$$

We require

$$\tau_0 = -1, \qquad \tau_1 = 1$$

and condition (3.20) on β_0, β_1 to hold:

$$\beta_0 \leq \min_t \frac{b(0,t)+c(0,t)}{2a(0,t)} \quad \text{and} \quad \beta_1 \geq \max_t \frac{b(1,t)+c(1,t)}{2a(1,t)}.$$

Then we obtain a growth constant of $\alpha_s = \|(-b+c)_x\|_\infty = \alpha$, i.e. the constant is the same as in the continuous case which was treated in the remark 3.6. This means that the approximation is strictly stable.

4.6.2 Accuracy

The accuracy of the method is determined by the accuracy of Q_a and D_1. If Q_a and D_1 are both $2p$th order accurate in the interior and pth order accurate at the boundary, the overall scheme has the local order of accuracy p at the boundary and $2p$ in the interior. Since there is a parabolic term, we suspect that a similar proof as in theorem 3.7 shows global order of accuracy $p+2$. However, we do not give it here since we have not derived an operator with $p > 1$.

4.6.3 Computations

We use the same example as in section 3.5.2 to test the implementation of the convection-diffusion equation:

$$
\begin{aligned}
&u_t = (a(x,t)u_x(x,t))_x + b(x,t)u_x(x,t) + (c(x,t)u(x,t))_x, \text{ where}\\
&a(x,t) = \frac{1}{10}\left(1 + \frac{1}{2}\sin\left(\frac{\pi}{2}(x+t)\right)\right), \quad b(x,t) = 2\sinh\left(-3x + \frac{3}{2}\right)(1+t),\\
&c(x,t) = 4x(x-1)(xt^2+1) \text{ with initial condition}\\
&u(x,0) = 4(\sin(\pi x) + \cos(10x))
\end{aligned}
$$

As boundary conditions we use Dirichlet and Neumann boundary conditions of the form

$$
\begin{aligned}
&u(0,t) = g_0(t) = 4\mathrm{e}^{4(-14+1/40\cdot\pi 2+2\pi\sinh(3/2))t} \approx \mathrm{e}^{-1.48t} \quad \text{and}\\
&u(1,t) = g_1(t) = 4\cos(10)\mathrm{e}^{4(-11\cos(10)+2\sinh(3/2)(\pi+10\sin(10)))t} \approx \cos(10)\mathrm{e}^{-2.24t}
\end{aligned}
$$

and

$$
\begin{aligned}
&u_x(0,t) = g_0(t) \approx 4\pi\mathrm{e}^{-1.48t} \quad \text{and}\\
&u_x(1,t) = g_1(t) \approx 4(-\pi - 10\sin(10))\mathrm{e}^{-2.24t},
\end{aligned}
$$

respectively.
The time step for the Runge-Kutta method is chosen as in the case of the product rule to be $3h^2$.
Table 12 shows the results at time $t = 1$. For Dirichlet boundary conditions we see as in the section 3.5.2 that the steep boundary part causes difficulties when the number of grid points is small. When N increases, the numerical rate of convergence tends to the expected value of 2.

	Dirichlet		Neumann	
N	$\|u-v\|_h$	q_D	$\|u-v\|_h$	q_N
10	$2.29 \cdot 10^{3}$		$2.76 \cdot 10^{-2}$	
20	3.11	9.53	$5.09 \cdot 10^{-3}$	2.44
40	$2.49 \cdot 10^{-1}$	3.64	$1.19 \cdot 10^{-3}$	2.10
80	$3.47 \cdot 10^{-2}$	2.84	$2.89 \cdot 10^{-4}$	2.04
160	$5.37 \cdot 10^{-3}$	2.69	$7.17 \cdot 10^{-5}$	2.01
320	$8.89 \cdot 10^{-4}$	2.60	$1.79 \cdot 10^{-5}$	2.00

Table 12: Numerical results for the convection diffusion equation

5 Approximation Using Finite Elements

5.1 Introduction

In this section we present an operator approximating the spatial part of equation (1.1) using a finite element ansatz.
With the so called mass lumping, the mass matrix can be reduced to a diagonal matrix, which makes it possible to interpret the finite element operator as a finite difference operator. This ansatz is based on an idea of Zemui, presented in his PhD-thesis [17].

5.2 Variational Formulation of the Semi-Discretization in Space

The semi-discretization is based on a variational formulation of (1.1) when using a finite element ansatz.

First we consider Neumann boundary conditions, i.e.

$$\frac{\partial}{\partial x}u(x_0,t) = g_0(t), \quad \frac{\partial}{\partial x}u(x_N,t) = g_1(t), \quad t \in I. \tag{5.1}$$

Let $V = H_0^1(\Omega)$ be the Sobolev space of functions with first derivatives in the weak sense and compact support.
In order to obtain a weak formulation of the problem, we multiply (1.1) for a fixed $t \in I$ by a function $v \in V$ and integrate over Ω:

$$\begin{aligned}\int_\Omega u_t v dx &= \int_\Omega (au_x)_x v \,\mathrm{d}x \\ &= -\int_\Omega au_x v_x \,\mathrm{d}x + a(x_N)g_1(t)v(x_N) - a(x_0)g_0(t)v(x_0)\end{aligned} \tag{5.2}$$

For the last equality we use integration by parts and the boundary conditions (5.1).
Let $(\cdot,\cdot)$ denote the $\mathcal{L}^2(\Omega)$ inner product. If $a(x,t) > 0$, we can define the following positive semidefinite bilinear form:

$$s(u,v) = \int_\Omega a(x,t)u_x v_x dx, \quad u,v \in V \tag{5.3}$$

i.e. the $\mathcal{L}_2$ inner product for the derivatives with weight function a. Using this notation we get the variational formulation of (1.1):
Find $u(t) \in V, t \in I$, such that

$$\begin{aligned}&(u_t, v) = -s(u,v) + a(x_N)g_1(t)v(x_N) - a(x_0)g_0(t)v(x_0) \quad \forall\, v \in V,\, t \in I \\ &u(0) = f(0)\end{aligned} \tag{5.4}$$

Now let V_h be a finite-dimensional subspace of V with basis $\{\varphi_0, \ldots, \varphi_N\}$. Replacing V be its subspace V_h we can obtain an analogue variational formulation for the semi-discrete problem:
Find $u_h(t) \in V_h, t \in I$, such that

$$\begin{aligned}&(\frac{\mathrm{d}u_h}{\mathrm{d}t}, v_h) = -s(u_h, v_h) + a(x_N)g_1(t)v_h(x_N) - a(x_0)g_0(t)v_h(x_0) \quad \forall\, v_h \in V_h,\, t \in I \\ &u_h(0) = f(0)\end{aligned} \tag{5.5}$$

We can now rewrite u_h as a linear combination of the basis functions with time-dependent coefficients:

$$u_h(x,t) = \sum_{i=0}^{N} b_i(t)\varphi_i(x), \quad t \in I \tag{5.6}$$

Using this representation of u_h and the basis functions as test functions, we get

$$\sum_{i=0}^{N} b_i'(t)(\varphi_i, \varphi_j) = -\sum_{i=0}^{N} b_i(t) s(\varphi_i, \varphi_j) + a(x_N) g_1(t) \varphi_j(x_N) - a(x_0) g_0(t) \varphi_j(x_0),$$
$$j = 0, \ldots, N, \quad t \in I \tag{5.7}$$
$$\sum_{i=0}^{N} b_i(0)(\varphi_i, \varphi_j) = (f, \varphi_j)$$

We can write this system using matrices:

$$M^h \frac{\mathrm{d}}{\mathrm{d}t} b = S^h b + a_N g_1 e_N - a_0 g_0 e_0, \tag{5.8}$$

where

$$M^h = (\varphi_i, \varphi_j)$$

is the mass matrix,

$$S^h = -s(\varphi_i, \varphi_j)$$

the stiffness matrix and b the coefficient vector.
If we choose a nodal basis, b is a discrete representation of u and will in accordance with the notation in the prior sections be denoted v in the following.

Remark:
In the case of Dirichlet boundary conditions

$$u(0,t) = g_0(t), \qquad u(1,t) = g_1(t),$$

we get the same variational formulation except for the two boundary points. Thus we solve the system (5.8) at the grid points x_j, $j = 1, \ldots, N-1$, and impose the boundary conditions at x_0 and x_N.

For our further analysis we assume $\Omega = [0, 1]$ for simplicity.

5.3 Construction

5.3.1 Basis Functions

In order to obtain a fourth order accurate scheme, cubic polynomials are chosen as basis functions. We choose the basis functions such that (5.6) is a piece-wise cubic Lagrange interpolant of $u(x,t)$, if $v_i = u(x_i)$.
From the Lagrange interpolation formula we get the following expression for the basis polynomials φ_j, $j = 4, 5, \ldots, N-4$

$$\varphi_j(x) = \varphi\left(\frac{x}{h} - j\right)$$

where

$$\varphi(x) = \begin{cases} \frac{1}{6}(x+1)(x+2)(x+3), & -2 \le x \le -1 \\ -\frac{1}{2}(x-1)(x+1)(x+2), & -1 \le x \le 0 \\ \frac{1}{2}(x-1)(x-2)(x+1), & 0 \le x \le 1 \\ -\frac{1}{6}(x-1)(x-2)(x-3), & 1 \le x \le 2 \\ 0, & \text{elsewhere} \end{cases}$$

The basis functions φ_j, $j = 0, 1, 2, 3$, have to be modified in the interval $[x_0, x_1]$ – and similarly φ_{N-j}, $j = 0, 1, 2, 3$, in the interval $[x_{N-1}, x_N]$ – such that (5.6) corresponds to the one-sided cubic Lagrange polynomial. This yields the following basis functions φ_j, $j = 0, 1, 2, 3$,

$$\varphi_j(x) = \begin{cases} \varphi_j^b(\frac{x}{h}), & x_0 \le x \le x_1 \\ \varphi(\frac{x}{h} - j), & x_1 \le x \le x_N \end{cases}$$

where

$$\begin{aligned} \varphi_0^b(x) &= -\frac{1}{6}(x-1)(x-2)(x-3), && 0 \le x \le 1 \\ \varphi_1^b(x) &= \frac{1}{2}x(x-2)(x-3), && 0 \le x \le 1 \\ \varphi_2^b(x) &= -\frac{1}{2}x(x-1)(x-3), && 0 \le x \le 1 \\ \varphi_3^b(x) &= -\frac{1}{6}x(x-1)(x-2), && 0 \le x \le 1 \end{aligned}$$

and analogously for the right boundary.

We want to collect some properties of the basis function that are useful for the further analysis. They can also be found in [17].

1. $\{\varphi_j, j = 0, \ldots, N\}$ is a nodal basis.

2. The support of φ_j is $[\max(x_{j-2}, x_0), \min(x_{j+2}, x_N)]$, $j \neq 3, N-3$ and $x_0 \le x \le x_5$ or $x_{N-5} \le x \le x_N$ for $j = 3$ and $j = N-3$, respectively.

3. $\varphi_j(x)$ is continuous and its derivative is piece-wise continuous in $0 \le x \le 1$, which means that it is a conforming finite element basis for second order systems.

   We now look at the approximation property of the basis set.

4. The error of (5.6) is given by

   $$u_h(x) - u(x) = \frac{u^{(4)}(\zeta)}{24} p(x), \quad x_i \le x, \zeta \le x_{i+1},$$

   where

   $$p(x) = \begin{cases} (x - x_{i-2})(x - x_{i-1})(x - x_i)(x - x_{i+1}), & i \neq 0, i \neq N \\ (x - x_0)(x - x_1)(x - x_2)(x - x_3), & i = 0 \\ (x - x_{N-3})(x - x_{N-2})(x - x_{N-1})(x - x_N), & i = N \end{cases}$$

5. Property 4 implies that cubic polynomials are exactly reproduced by the interpolant (5.6). Thus the approximation power of the basis is $\mathcal{O}(h^4)$ which is equivalent to

   $$\sum_{i=0}^{N} \varphi_i(x) = 1 \tag{5.9}$$

   $$\sum_{i=0}^{N} (x_i - \alpha)^n \varphi_i(x) = (x - \alpha)^n, \quad n = 0, 1, 2, 3, \quad \text{any } \alpha. \tag{5.10}$$

6. By differentiation of (5.9) we get

$$\sum_{i=0}^{N} \varphi_i'(x) = 0. \tag{5.11}$$

7. For $4 \leq i \leq N-4$ the following moment conditions are valid

$$\int_{x_{i-2}}^{x_{i+2}} \varphi_i(x)\mathrm{d}x = h \tag{5.12}$$

$$\int_{x_{i-2}}^{x_{i+2}} (x-x_i)^n \varphi_i(x)\mathrm{d}x = 0, \quad n = 1,2,3. \tag{5.13}$$

8. From property 7 it follows

$$\int_{x_{i-2}}^{x_{i+2}} u(x)(x-x_i)^n \varphi_i(x)\mathrm{d}x = \mathcal{O}(h^{n+1}), \quad n \geq 0 \tag{5.14}$$

since the support of φ_i is $\mathcal{O}(h)$ around x_i.
For the derivative φ_i we get the estimate with one order less.

5.3.2 Mass Matrix

The elements of the mass matrix are given by

$$M^h = (\varphi_i, \varphi_j)$$

as we have seen in section 5.2. This means that the matrix has a bandwidth of 7 and system (5.8) is an ordinary differential equation system for v that is implicit. This requires solving linear systems, which is computationally quite costly.

Therefore we now apply a technique called mass lumping to replace the matrix by a diagonal one.
Let $w(x,t) = \frac{\partial}{\partial t}u(x,t)$. Then we can write the ith row of the left hand side of system (5.8) in the following way:

$$\sum_{j=0}^{N} M_{ij}^h w_j = \sum_{j=0}^{N} \int \varphi_i(x)\varphi_j(x)\,\mathrm{d}x\ w_j \tag{5.15}$$

Expanding $w(x_j)$ into its Taylor series around x_i yields:

$$\begin{aligned}
\sum_{j=0}^{N} M_{ij}^h w_j &= \sum_{j=0}^{N} \left(\int \varphi_i(x)\varphi_j(x)\mathrm{d}x \left(\sum_{k=0}^{3} \frac{1}{k!}(x_j - x_i)^k w_i^{(k)} + \mathcal{O}(h^4) \right) \right) \\
&= \int \varphi_i(x) \sum_{k=0}^{3} \left(\frac{1}{k!} w_i^{(k)} \sum_{j=0}^{N} \varphi_j(x)(x_j - x_i)^k \right) \mathrm{d}x + \mathcal{O}(h^5) \\
&= \int \varphi_i(x) \left(w_i + (x-x_i)w_i' + \frac{1}{2}(x-x_i)^2 w_i'' + \frac{1}{6}(x-x_i)^3 w_i''' \right) \mathrm{d}x \\
&\quad + \mathcal{O}(h^5),
\end{aligned} \tag{5.16}$$

where the identity (5.10) was used for the last equality. Here $w^{(k)}$ denotes the kth derivative with respect to x.
For the inner points, i.e. $4 \leq i \leq N-4$, we get

$$\sum_{j=0}^{N} M_{ij}^h w_j = w_i \int \varphi_i(x) \mathrm{d}x + \mathcal{O}(h^5) \tag{5.17}$$

using the moment condition (5.13). Thus we can achieve diagonal lumping for the interior points by defining

$$M_{ii}^L w_i = w_i \int_{x_{i-2}}^{x_{i+2}} \varphi_i(x)\, \mathrm{d}x.$$

This integral can be directly calculated using (5.12):

$$M_{ii}^L = h. \tag{5.18}$$

We now look at the boundary part. Here it is more difficult to achieve diagonal lumping as we are lacking an identity similar to (5.13). Therefore we give up two orders of accuracy locally in order to achieve a diagonal matrix. We will see later on that this does not reduce the global accuracy (see section 5.5). We consider only the left boundary, but the right one can be treated in the same way. For the boundary treatment we have to consider Neumann and Dirichlet data separately.

Neumann Boundary Conditions
We start off from equation (5.16) and apply $w_i' = w_0' + ihw_0'' + \mathcal{O}(h^2)$ and $w_i'' = w_0'' + \mathcal{O}(h)$ in the cases $i = 1, 2, 3$. We find that

$$\begin{aligned}\sum_{j=0}^{N} M_{ij}^h w_j = {} & w_i \int_{x_0}^{x_{i+2}} \varphi_i(x)\, \mathrm{d}x + w_0' \int_{x_0}^{x_{i+2}} \varphi_i(x)(x - x_i)\, \mathrm{d}x \\ & + w_0'' \int_{x_0}^{x_{i+2}} \varphi_i(x_i) \left((x - x_i) ih + \frac{1}{2}(x - x_i)^2 \right) \mathrm{d}x + \mathcal{O}(h^4) \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i = 0, 1, 2, 3.\end{aligned} \tag{5.19}$$

Hence we could obtain a diagonal mass matrix with mass lumping of third order defining

$$M_{ii}^L w_i = w_i \int_{x_0}^{x_{i+2}} \varphi_i(x)\, \mathrm{d}x + w_0' \int_{x_0}^{x_{i+2}} \varphi_i(x)(x - x_i)\, \mathrm{d}x, \tag{5.20}$$

where we need to know $w_0' = \frac{\partial^2}{\partial x \partial t} u(0, t)$. In the case of Neumann boundary conditions, u_x is given at $x = 0$ as a function of t. Hence we can calculate w_0' from the physical boundary data by taking its derivative with respect to t. In practical applications where the exact derivative is not available, it can be approximated second order accurate.

We can rewrite the lumped mass matrix as

$$M^L w = \widetilde{M} w - \widetilde{m}_0 \frac{\mathrm{d}}{\mathrm{d}t} g_0(t) - \widetilde{m}_1 \frac{\mathrm{d}}{\mathrm{d}t} g_1(t) \tag{5.21}$$

where

$$\widetilde{M}_{ij} = \delta_{ij} \int \varphi_i(x)\ \mathrm{d}x \tag{5.22}$$

and

$$\widetilde{m}_0^{(i)} = \begin{cases} -\int_{x_0}^{x_{i+2}} \varphi_i(x)(x-x_i)\,\mathrm{d}x, & i \in \{0,1,2,3\} \\ 0, & \text{else} \end{cases}$$
$$\widetilde{m}_1^{(i)} = \begin{cases} -\int_{x_{i-2}}^{x_N} \varphi_i(x)(x-x_i)\,\mathrm{d}x, & i \in \{N-3,N-2,N-1,N\} \\ 0, & \text{else.} \end{cases}$$

Dirichlet Boundary Conditions

In the case of Dirichlet boundary conditions the first and last row of the operator are neglected since the values are given in advance and can be imposed directly. Thus in this section we only consider $i = 1, 2, 3$.

Again we start off from equation (5.16). This time we use

$$w_i' = \frac{w_i - w_0}{x_i - x_0} + \frac{1}{2}(x_i - x_0)w_i'' + \mathcal{O}(h^2) = \frac{w_i - w_0}{ih} + \frac{1}{2}ihw_0'' + \mathcal{O}(h^2) \tag{5.23}$$

and $w_i'' = w_0'' + \mathcal{O}(h)$ to obtain

$$\sum_{j=0}^{N} M_{ij}^h w_j = w_i \frac{1}{ih}\int_{x_0}^{x_{i+2}} \varphi_i(x)(x-x_0)\,\mathrm{d}x - w_0\frac{1}{ih}\int_{x_0}^{x_{i+2}} \varphi_i(x)(x-x_i)\,\mathrm{d}x$$
$$+ w_0''\frac{1}{2}\int_{x_0}^{x_{i+2}} \varphi_i(x)((x-x_i)^2 + (x-x_i)ih)\,\mathrm{d}x + \mathcal{O}(h^4). \tag{5.24}$$

Like in the case of Neumann boundary conditions, we can use the physical boundary data to gain accuracy by calculating $w_0 = \frac{\partial}{\partial t}u(0,t) = \frac{\partial}{\partial t}g_0(t)$. This leads us to the following mass lumping of third order

$$M_{ii}^L w_i = w_i \frac{1}{ih}\int_{x_0}^{x_{i+2}} \varphi_i(x)(x-x_0)\,\mathrm{d}x - w_0\frac{1}{ih}\int_{x_0}^{x_{i+2}} \varphi_i(x)(x-x_i)\,\mathrm{d}x \tag{5.25}$$

Treating the right boundary in the same way we can write the lumped mass matrix as

$$M^L w = \widetilde{M}w - \widetilde{m}_0\frac{\mathrm{d}}{\mathrm{d}t}g_0(t) - \widetilde{m}_1\frac{\mathrm{d}}{\mathrm{d}t}g_1(t)$$

where

$$\widetilde{M}_{ij} = \begin{cases} \frac{1}{ih}\int_{x_0}^{x_{i+2}} \varphi_i(x)(x-x_0)\,\mathrm{d}x & i = 1,2,3 \\ \delta_{ij}\int_{x_{i-2}}^{x_{i+2}} \varphi_i(x)\,\mathrm{d}x & 4 \le i \le N-4 \\ \frac{1}{(i-N)h}\int_{x_{i-2}}^{x_N} \varphi_i(x)(x-x_N)\,\mathrm{d}x & i = N-3,N-2,N-1 \end{cases} \tag{5.26}$$

and

$$\widetilde{m}_0^{(i)} = \begin{cases} -\frac{1}{ih}\int_{x_0}^{x_{i+2}} \varphi_i(x)(x-x_i)\,\mathrm{d}x, & i \in \{1,2,3\} \\ 0, & \text{else} \end{cases}$$
$$\widetilde{m}_1^{(i)} = \begin{cases} -\frac{1}{(i-N)h}\int_{x_{i-2}}^{x_N} \varphi_i(x)(x-x_i)\,\mathrm{d}x, & i \in \{N-3,N-2,N-1\} \\ 0, & \text{else.} \end{cases}$$

5.3.3 Stiffness Matrix

The elements of the stiffness matrix shall approximate

$$S^h = -s(\varphi_i, \varphi_j) \tag{5.27}$$

Since $a(x,t)$ is usually only given at some discrete points, this integral cannot be solved analytically. Thus it has to be approximated in some way. We choose to evaluate the integrals by numerical integration. Using the cubic basis functions, we can eventually obtain a fourth order accurate finite difference scheme. When approximating the integral in (5.27), we have to choose a quadrature formula which is at least fifth order accurate because $M^{-1}Sv$ should be fourth order accurate and M is $\mathcal{O}(h)$. Since a is given at the grid points, it is appropriate to use a Newton-Cotes quadrature formula because it is based on a regular grid. Another difficulty is that the derivative of the basis functions is not continuous at the grid points. Hence we have to approximate the integral for each interval between two grid points separately.

For these reasons we choose the Simpson rule in each interval of length h, i.e. the integral over $[x_i, x_{i+1}]$ for some j,k is calculated by the formula:

$$\frac{h}{6}\big(a(x_i)\varphi_j'(x_i^+)\varphi_k'(x_i^+) + 4a(x_{i+1/2})\varphi_j'(x_{i+1/2})\varphi_k'(x_{i+1/2}) + a(x_{i+1})\varphi_j'(x_{i+1}^-)\varphi_k'(x_{i+1}^-)\big), \tag{5.28}$$

where $\varphi_l'(x_m^+) = \lim_{x\to x_m, x>x_m} \varphi_l'(x)$ and $\varphi_l'(x_m^-) = \lim_{x\to x_m, x<x_m} \varphi_l'(x)$.
This means that the value of a is needed on a finer grid with step size $\frac{h}{2}$.
Let S denote the stiffness matrix obtained by numerical integration in the following.

Hence we replace the inner product s by a discretized one, which means that we are using nonconforming finite elements. Therefore we have to investigate the influences of numerical integration on accuracy and stability. We will do this with the help of finite difference theory. In the interior we do not loose accuracy due to the numerical integration (see section 5.4). At the boundary however we have to do some corrections in the mass matrix in order not to loose accuracy (see section 5.3.4).

Remark 5.1. *One could also choose a different ansatz in order to approximate the elements of the stiffness matrix by approximating the function $a(x,t)$ in a Lagrange polynomial using the basis functions. The resulting integrals could then be calculated analytically. This has the advantage that we would not have to care about influences of numerical integration. However we did not choose this ansatz because stability problems may arise in this case. Since the basis functions are less than 0 in some intervals, it might happen that the approximation for a is smaller than 0, which means that $s(\cdot,\cdot)$ is no longer positive semidefinite.*

5.3.4 Influences of Numerical Integration

When using numerical integration, some problems arise at the boundary. The objective of this section is to correct the lumped mass matrix such that the local error of the difference scheme at the boundary, that we obtain by evaluating the integral exactly, is kept.

We consider the left boundary. The right boundary can be treated in the same way.
Let thus $i \in \{0,1,2,3\}$. We first show that the local error at the boundary is of the order h^2 in the case of exact evaluation of the integral in the stiffness matrix. Let $u(x,t)$ and $a(x,t)$ be sufficiently smooth functions in x for this analysis.

We start with some transformation of the stiffness matrix:

$$\begin{aligned}
\sum_{j=0}^{N} S_{ij}^h v_j &= -\sum_{j=0}^{i+3} \int_{x_0}^{x_{i+2}} a(x)\varphi_i'(x)\varphi_j'(x)\, v_j \,\mathrm{d}x \\
&= -\sum_{\substack{j=0 \\ j\neq i}}^{i+3} \int_{x_0}^{x_{i+2}} a(x)\varphi_i'(x)\varphi_j'(x) v_j \,\mathrm{d}x + v_i \int_{x_0}^{x_{i+2}} a(x)\varphi_i'(x) \sum_{\substack{k=0 \\ k\neq i}}^{N} \varphi_k'(x) \,\mathrm{d}x \\
&= -\sum_{k=1}^{3} (v_{i+k} - v_i) \int a(x)\varphi_i'(x)\varphi_{i+k}'(x) \,\mathrm{d}x \\
&\quad -\sum_{k=1}^{i} (v_{i-k} - v_i) \int a(x)\varphi_i'(x)\varphi_{i-k}'(x) \,\mathrm{d}x
\end{aligned}$$

The second equality follows from (5.11). Expanding the exact solution $u(x)$ in its Taylor series, we obtain

$$\begin{aligned}
\sum_{j=0}^{N} S_{ij}^h u_j = &-\int_{x_0}^{x_{i+2}} a(x)\varphi_i'(x) h \left(\sum_{k=0}^{i+3} (k-i)\varphi_k'(x) \right) u_i' \mathrm{d}x \\
&-\int_{x_0}^{x_{i+2}} a(x)\varphi_i'(x) \frac{1}{2} h^2 \left(\sum_{k=0}^{i+3} (k-i)^2 \varphi_k'(x) \right) u_i'' \mathrm{d}x \\
&-\int_{x_0}^{x_{i+2}} a(x)\varphi_i'(x) \frac{1}{6} h^3 \left(\sum_{k=0}^{i+3} (k-i)^3 \varphi_k'(x) \right) u_i''' \mathrm{d}x \\
&-\int_{x_0}^{x_{i+2}} a(x)\varphi_i'(x) \frac{1}{24} h^4 \left(\sum_{k=0}^{i+3} (k-i)^4 \varphi_k'(x) \right) u_i^{(4)} \mathrm{d}x + \mathcal{O}(h^5)
\end{aligned}$$

For $x_0 \le x \le x_2$ we can derive the following identities by differentiation of (5.10):

$$\begin{aligned}
&\sum_{k=0}^{i+3} h(k-i)\varphi_k'(x) = \sum_{k=0}^{N} h(k-i)\varphi_k'(x) = \frac{\mathrm{d}}{\mathrm{d}x}(x - x_i) = 1 \\
&\frac{1}{2}\sum_{k=0}^{i+3} h^2(k-i)^2\varphi_k'(x) = \frac{1}{2}\sum_{k=0}^{N} h^2(k-i)^2\varphi_k'(x) = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}x}(x - x_i)^2 = (x - x_i) \\
&\frac{1}{6}\sum_{k=0}^{i+3} h^3(k-i)^3\varphi_k'(x) = \frac{1}{6}\sum_{k=0}^{N} h^3(k-i)^3\varphi_k'(x) = \frac{1}{6}\frac{\mathrm{d}}{\mathrm{d}x}(x - x_i)^3 = \frac{1}{2}(x - x_i)^2
\end{aligned}$$

Using these identities yields

$$\begin{aligned}
\sum_{j=0}^{N} S_{ij}^h u_j = -\int_{x_0}^{x_{i+2}} a(x)\varphi_i'(x) \Big(&u_i' + (x - x_i)u_i'' + \frac{1}{2}(x - x_i)^2 u_i''' \\
&+ \frac{1}{6}(x - x_i)^3 u_i^{(4)} \Big) u_i' \mathrm{d}x + R_i^S u_i^{(4)}(x) + \mathcal{O}(h^5),
\end{aligned}$$

where

$$R_i^S = -\int_{x_0}^{x_{i+2}} a(x)\varphi_i'(x) \left[\frac{1}{24} h^4 \left(\sum_{k=0}^{i+3} (k-i)^4 \varphi_k'(x) \right) - \frac{1}{6}(x - x_i)^3 \right] \mathrm{d}x. \tag{5.29}$$

We now use

$$u'(x) = u_i' + (x - x_i)u_i'' + \frac{1}{2}(x - x_i)^2 u_i''' + \frac{1}{6}(x - x_i)^3 u_i^{(4)} + \mathcal{O}(h^4) \tag{5.30}$$

which leads us to

$$\sum_{j=0}^{N} S_{ij}^h u_j = -\int_{x_0}^{x_{i+2}} a(x)\varphi_i'(x)u'(x)\,\mathrm{d}x + R_i^S u_i^{(4)}(x) + \mathcal{O}(h^5). \tag{5.31}$$

Using integration by parts, the first term in (5.31) becomes

$$\begin{aligned} -\int_{x_0}^{x_2} \varphi_0'(x)a(x)u'(x)\,\mathrm{d}x &= a_0 u_0' + \int_{x_0}^{x_2} \varphi_0(x)(a(x)u'(x))'\,\mathrm{d}x \\ -\int_{x_0}^{x_{i+2}} \varphi_i'(x)a(x)u'(x)\,\mathrm{d}x &= \int_{x_0}^{x_{i+2}} \varphi_i(x)(a(x)u'(x))'\,\mathrm{d}x \end{aligned}$$

Now we turn to the mass matrix. Here we have to consider Neumann and Dirichlet boundary conditions separately.
First assume **Neumann** data.
We start off from equation (5.19) combined with the definition of M^L in (5.20) and M^h in (5.15), which yields

$$\begin{aligned} M_{ii}^L w_i &= \sum_{j=0}^{N} \int_{x_0}^{x_{i+2}} \varphi_i(x)\varphi_j(x)\mathrm{d}x\ w_j + R_i^M w_0'' + \mathcal{O}(h^4) \\ &= \int_{x_0}^{x_{i+2}} \varphi_i(x) \sum_{j=0}^{N} \varphi_j(x) w_j\,\mathrm{d}x + R_i^M w_0'' + \mathcal{O}(h^4), \end{aligned} \tag{5.32}$$

where

$$R_i^M = -\int_{x_0}^{x_{i+2}} \varphi_i(x_i)\Big((x - x_i) + \frac{1}{2}(x - x_i)^2\Big)\,\mathrm{d}x \tag{5.33}$$

Using the interpolation property $w(x) = \sum_{j=0}^{N} \varphi_j(x)w_j + \mathcal{O}(h^4)$, we get

$$M_{ii}^L w_i = \int_{x_0}^{x_{i+2}} \varphi_i(x)w(x)\,\mathrm{d}x + R_i^M w_0'' + \mathcal{O}(h^4).$$

We use the differential equation $w = u_t = (au_x)_x$ which yields

$$M_{ii}^L w_i = \int_{x_0}^{x_{i+2}} \varphi_i(x)(a(x)u_x(x))_x\,\mathrm{d}x + R_i^M w_0'' + \mathcal{O}(h^4) \tag{5.34}$$

Finally we can conclude that

$$\frac{1}{h}(Mu - S^h u)_i = \mathcal{O}(h^2), \tag{5.35}$$

i.e. that the local truncation error at the boundary is found to be of order h^2 in each row in the case of exactly evaluated stiffness matrix elements. The scaling with $\frac{1}{h}$ is necessary because we consider second derivatives.

Now we consider the additional error because of the numerical integration. In order to identify it, we expand $\sum_{j=0}^{N} S_{ij}^h u_j$ in its Taylor series around x_i:

$$\begin{aligned}
\sum_{j=0}^{N} S_{0j}^h u(x_j) &= a(x_0)u_x(x_0) + \frac{1}{3}h(au_x)_x|_{x_0} + \frac{2}{45}h^2(au_x)_{xx}|_{x_0} - \frac{1}{45}h^3(au_x)_{xxx}|_{x_0} \\
&\quad + h^3\frac{1}{30}a(x_0)u^{(4)}(x_0) + \mathcal{O}(h^4) \\
\sum_{j=0}^{N} S_{1j}^h u(x_j) &= \frac{31}{24}h(au_x)_x|_{x_1} - \frac{7}{36}h^2(au_x)_{xx}|_{x_1} + \frac{7}{80}h^3(au_x)_{xxx}|_{x_1} \\
&\quad - h^3\frac{139}{1440}a(x_1)u^{(4)}(x_1) + \mathcal{O}(h^4) \\
\sum_{j=0}^{N} S_{2j}^h u(x_j) &= \frac{5}{6}h(au_x)_x|_{x_2} + \frac{23}{90}h^2(au_x)_{xx}|_{x_2} - \frac{1}{5}h^3(au_x)_{xxx}|_{x_1} \\
&\quad + h^3\frac{7}{90}a(x_2)u^{(4)}(x_2) + \mathcal{O}(h^4) \\
\sum_{j=0}^{N} S_{3j}^h u(x_j) &= \frac{25}{24}h(au_x)_x|_{x_3} - \frac{19}{180}h^2(au_x)_{xx}|_{x_3} + \frac{79}{720}h^3(au_x)_{xxx}|_{x_3} \\
&\quad - h^3\frac{7}{480}a(x_3)u^{(4)}(x_3) + \mathcal{O}(h^4)
\end{aligned} \tag{5.36}$$

and compare the results with the Taylor series for the stiffness matrix Sv_i based on numerical integration

$$\begin{aligned}
\sum_{j=0}^{N} S_{0j} u(x_j) &= a(x_0)u_x(x_0) + \frac{1}{3}h(au_x)_x|_{x_0} + \frac{7}{144}h^2(au_x)_{xx}|_{x_0} - \frac{1}{48}h^3(au_x)_{xxx}|_{x_0} \\
&\quad + h^3\frac{11}{288}a(x_0)u^{(4)}(x_0) + \mathcal{O}(h^4) \\
\sum_{j=0}^{N} S_{1j} u(x_j) &= \frac{31}{24}h(au_x)_x|_{x_1} - \frac{59}{288}h^2(au_x)_{xx}|_{x_1} + \frac{55}{576}h^3(au_x)_{xxx}|_{x_1} \\
&\quad - h^3\frac{31}{288}a(x_1)u^{(4)}(x_1) + \mathcal{O}(h^4) \\
\sum_{j=0}^{N} S_{2j} u(x_j) &= \frac{5}{6}h(au_x)_x|_{x_2} + \frac{19}{72}h^2(au_x)_{xx}|_{x_2} - \frac{31}{144}h^3(au_x)_{xxx}|_{x_1} \\
&\quad + h^3\frac{25}{288}a(x_2)u^{(4)}(x_2) + \mathcal{O}(h^4) \\
\sum_{j=0}^{N} S_{3j} u(x_j) &= \frac{25}{24}h(au_x)_x|_{x_3} - \frac{31}{288}h^2(au_x)_{xx}|_{x_3} + \frac{9}{64}h^3(au_x)_{xxx}|_{x_3} \\
&\quad - h^3\frac{5}{288}a(x_3)u^{(4)}(x_3) + \mathcal{O}(h^4)
\end{aligned} \tag{5.37}$$

The comparison shows that already the coefficients in front of h^2 differ, which means that the local error (after scaling) will be reduced by one order. Let C_i be the error coefficients of h^2, i.e.

$$C_0 = \frac{1}{240}, \qquad C_1 = -\frac{1}{96}, \qquad C_2 = \frac{1}{120}, \qquad C_3 = -\frac{1}{480}$$

It holds

$$C_i(au_x)_{xx}(x_i) = C_i(u_t)_x(x_i) = C_i u_{tx}(x_0) + \mathcal{O}(h). \tag{5.38}$$

Thus we can avoid loosing local accuracy at the boundary if we add C_i to the correction vector $\widetilde{m}_0^{(i)}$.
In the following, let us denote the corrected mass matrix with M and its correction vectors with m_0 and m_1, respectively.

Now we turn to **Dirichlet** data. The mass matrix is given by (5.25) and its leading error term can be obtained by comparing equation (5.25) with (5.24). If we calculate the occurring integrals, we get the following expression for the lumped mass matrix in each irregular row:

$$\begin{aligned}
\sum_{j=0}^{N} M_{1j}^L u_t(x_j) &= h\frac{79}{72}u_t(x_1) + h\frac{7}{36}u_t(x_0) \\
&= h\frac{31}{24}u_t(x_1) - h^2\frac{7}{36}u_{tx}(x_1) + h^3\frac{7}{72}u_{txx}(x_0) + \mathcal{O}(h^4) \\
\sum_{j=0}^{N} M_{2j}^L u_t(x_j) &= h\frac{173}{180}u_t(x_2) - h\frac{23}{180}u_t(x_0) \\
&= h\frac{5}{6}u_t(x_2) + h^2\frac{23}{90}u_{tx}(x_2) - h^3\frac{23}{90}u_{txx}(x_0) + \mathcal{O}(h^4) \\
\sum_{j=0}^{N} M_{3j}^L u_t(x_j) &= h\frac{1087}{1080}u_t(x_3) + h\frac{19}{540}u_t(x_0) \\
&= h\frac{25}{24}u_t(x_3) - h^2\frac{19}{180}u_{tx}(x_3) + h^3\frac{19}{120}u_{txx}(x_0) + \mathcal{O}(h^4)
\end{aligned} \tag{5.39}$$

The stiffness matrix is the same as in the Neumann case except that the 0th row is missing. Comparing the expression for the mass matrix (5.39) with the one for the stiffness matrix with exact integration (5.36) derived above, we see that the local truncation error is of order h^2 like in the case with Neumann data, whereas with the numerically integrated stiffness matrix (5.37) we again have a local error of order h.
We can correct the mass matrix such that the local truncation error is not reduced by one order.
Like in the Neumann case we want to correct with a multiple of $(au_x)_x = (u_t)_x$. Let C_i denote the coefficient in front of $(au_x)_x$ in the expansion of S, i.e.

$$C_1 = -\frac{59}{288}, \qquad C_2 = \frac{19}{72}, \qquad C_3 = -\frac{31}{288}.$$

However, in the present case this is not that easy since we have given u instead of u_x as a function of t. Therefore we approximate the x-derivative by the one-sided difference operator, i.e. $(u_t)_x(x_i) = (u_t(x_i) - u_t(x_0))/(ih) + \mathcal{O}(h)$, like we did it in the derivation of the mass matrix (cf. (5.23)). Hence we replace the exact value of the integral $\int_{x_0}^{x_{i+2}} \varphi_i(x)(x - x_i)\,\mathrm{d}x$ by $C_i/(ih)$ in both the mass matrix and its correction term. This leads to a slightly different mass matrix, which we denote by M with the correction terms m_0 and m_1 in the following.

5.3.5 Finite Difference Method

We want to round off the section about the construction by stating the operator that has been derived above.

The operator has been developed using the theory of finite elements. However, using mass lumping, we obtained a diagonal mass matrix M, which can be inverted by taking the reciprocal of the diagonal elements. Thus we obtain the operator

$$L(\cdot) = M^{-1}S(\cdot) + M^{-1}\left(m_0\frac{\mathrm{d}}{\mathrm{d}t}g_0(t) + m_1\frac{\mathrm{d}}{\mathrm{d}t}g_1(t)\right) \tag{5.40}$$

which can be interpreted as a finite difference method.

Note: For simplicity we have chosen the same notation for mass and stiffness matrix for both different cases of boundary data even though the matrices are slightly different at the boundary.

For Neumann data we have the following system

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}v(t) = Lv(t) + M_{NN}^{-1}a(1)g_1(t) - M_{00}^{-1}a(0)g_0(t) \\ &v(t) = f^h(t), \quad f^h(j)(t) = f(x_j, t), \quad j = 0, 1, \ldots, N. \end{aligned} \tag{5.41}$$

and for Dirichlet data

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}v(t) = Lv(t) \\ &v_0(t) = g_0(t), \quad v_N(t) = g_1(t), \\ &v(t) = f^h(t), \quad f^h(j)(t) = f(x_j, t), \quad j = 0, 1, \ldots, N. \end{aligned} \tag{5.42}$$

We present the operator in more detail in appendix A.
The following convergence analysis is based on typical techniques for finite difference methods.

5.4 Local Order of Accuracy

We now look at the order of accuracy that can be obtained using the finite element system with the lumped mass matrix (5.41) and (5.42), respectively.
Let $u(x,t)$ and $a(x,t)$ be sufficiently smooth functions in x.
We use here the conventional error analysis for finite difference methods and therefore first look at the local order of accuracy at each grid point in order to obtain the global error.

5.4.1 Local Error at the Inner Points

We can show fourth order accuracy in the interior using Taylor expansion.
Let $4 \leq j \leq N-4$. We look at the difference operator L:

$$\begin{aligned}
\sum_{k=j-3}^{j+3} L_{jk} v_k = \frac{1}{6h^2}\Big[&\Big(\frac{1}{18}a_{j-2} - \frac{1}{144}a_{j-3/2} + \frac{1}{18}a_{j-1}\Big)v_{j-3} + \\
&\Big(\frac{1}{12}a_{j-2} + \frac{3}{16}a_{j-3/2} - \frac{2}{3}a_{j-1} + \frac{3}{16}a_{j-1/2} + \frac{1}{12}a_j\Big)v_{j-2} + \\
&\Big(-\frac{1}{6}a_{j-2} - \frac{3}{16}a_{j-3/2} - \frac{1}{3}a_{j-1} - \frac{81}{16}a_{j-1/2} - \frac{1}{3}a_j - \frac{3}{16}a_{j+1/2} - \frac{1}{6}a_{j+1}\Big)v_{j-1} + \\
&\Big(\frac{1}{36}a_{j-2} + \frac{1}{144}a_{j-3/2} + \frac{10}{9}a_{j-1} + \frac{81}{16}a_{j-1/2} + \frac{1}{2}a_j + \frac{81}{16}a_{j+1/2} + \frac{10}{9}a_{j+1} \\
&\qquad + \frac{1}{144}a_{j+3/2} + \frac{1}{36}a_{j+2}\Big)v_j + \\
&\Big(-\frac{1}{6}a_{j-1} - \frac{3}{16}a_{j-1/2} - \frac{1}{3}a_j - \frac{81}{16}a_{j+1/2} - \frac{1}{3}a_{j+1} - \frac{3}{16}a_{j+3/2} - \frac{1}{6}a_{j+2}\Big)v_{j+1} + \\
&\Big(\frac{1}{12}a_j + \frac{3}{16}a_{j+1/2} - \frac{2}{3}a_{j+1} + \frac{3}{16}a_{j+3/2} + \frac{1}{12}a_{j+2}\Big)v_{j+2} + \\
&\Big(\frac{1}{18}a_{j+1} - \frac{1}{144}a_{j+3/2} + \frac{1}{18}a_{j+2}\Big)v_{j+3}\Big]
\end{aligned}$$

If we now insert the sufficiently smooth functions u and a and expand u in its Taylor series around x_j we get:

$$\begin{aligned}
&\sum_{k=j-3}^{j+3} L_{jk} u(x_k) = -\frac{1}{6h^2}\Big[\\
&h\Big(-\frac{1}{6}a_{j-2} - \frac{1}{6}a_{j-3/2} + \frac{4}{3}a_{j-1} + \frac{9}{2}a_{j-1/2} - \frac{9}{2}a_{j+1/2} - \frac{4}{3}a_{j+1} + \frac{1}{6}a_{j+3/2} + \frac{1}{6}a_{j+2}\Big)u_x(x_j) + \\
&\frac{h^2}{2}\Big(\frac{2}{3}a_{j-2} + \frac{1}{2}a_{j-3/2} - \frac{8}{3}a_{j-1} - \frac{9}{2}a_{j-1/2} - \frac{9}{2}a_{j+1/2} - \frac{8}{3}a_{j+1} + \frac{1}{2}a_{j+3/2} + \frac{2}{3}a_{j+2}\Big)u_{xx}(x_j) + \\
&\frac{h^3}{6}\Big(-2a_{j-2} - \frac{9}{8}a_{j-3/2} + 4a_{j-1} + \frac{27}{8}a_{j-1/2} - \frac{27}{8}a_{j+1/2} - 4a_{j+1} + \frac{9}{8}a_{j+3/2} + 2a_{j+2}\Big)u_{xxx}(x_j) + \\
&\frac{h^4}{24}\Big(\frac{17}{3}a_{j-2} + \frac{9}{4}a_{j-3/2} - \frac{20}{3}a_{j-1} - \frac{9}{4}a_{j+1/2} + 2a_j - \frac{9}{4}a_{j+1/2} - \frac{20}{3}a_{j+1} + \frac{9}{4}a_{j+3/2} + \frac{17}{3}a_{j+2}\Big) \\
&\qquad u^{(4)}(x_j) + \\
&\frac{h^5}{120}\Big(-16a_{j-2} - \frac{33}{8}a_{j-3/2} + 8a_{j-1} - \frac{9}{8}a_{j-1/2} + \frac{9}{8}a_{j+1/2} - 8a_{j+1} + \frac{33}{8}a_{j+3/2} + 16a_{j+2}\Big)u^{(5)}(x_j)\Big] \\
&+ \mathcal{O}(h^4)
\end{aligned}$$

We insert the Taylor series of $a(x)$ around x_j and obtain

$$\begin{aligned}
\sum_{k=j-3}^{j+3} L_{jk} u(x_k) &= \big(a_x(x_j) + \mathcal{O}(h^4)\big)\, u_x(x_j) + \big(a(x_j) + \mathcal{O}(h^4)\big)\, u_{xx}(x_j) \\
&\quad + \left(\frac{123}{864}\, h^4 a_{xxx}(x_j) + \mathcal{O}(h^5)\right) u_{xxx}(x_j) + \left(\frac{41}{288}\, h^4 a_{xx}(x_j) + \mathcal{O}(h^5)\right) u^{(4)}(x_j) \\
&\quad + \left(\frac{41}{480}\, h^4 a_x(x_j) + \mathcal{O}(h^5)\right) u^{(5)}(x_j) + \mathcal{O}(h^4) \\
&= (a(x)u_x(x))_x\big|_{x_j} + \mathcal{O}(h^4)
\end{aligned}$$

Thus the operator (5.40) is 4th order accurate in the interior. Let $T_j = \mathcal{O}(h^4)$ denote the local truncation error at a regular point x_j.

5.4.2 Local Error at the Boundary

In section 5.3.4 we have already considered the local order of accuracy at the boundary. In this section, however, we want to define the local truncation error a bit more detailed since

we need it for the analysis of the global error in the next section.
For this purpose we expand Lu. The expansion of Su is given in section 5.3.4 by equation (5.37). We multiply the jth row by M_{jj}^{-1} and add $m_0^{(j)}\mathrm{d}/\mathrm{d}tg_0 + m_1^{(j)}\mathrm{d}/\mathrm{d}tg_1$ to get an expansion for Lu.

In the **Neumann** case this yields

$$\begin{aligned}
\sum_{j=0}^{N} L_{0j}u(x_j) &= 3\frac{1}{h}a(x_0)u_x(x_0) + (au_x)_x|_{x_0} + \frac{7}{48}h(au_x)_{xx}|_{x_0} - \frac{1}{16}h(au_x)_{xxx}|_{x_0} \\
&\quad + h^2\frac{11}{96}a(x_0)u^{(4)}(x_0) + \mathcal{O}(h^3) - \frac{7}{48}h(au_x)_{xx}|_{x_0} \\
&= 3\frac{1}{h}a(x_0)u_x(x_0) + (au_x)_x|_{x_0} - \frac{1}{16}h(au_x)_{xxx}|_{x_0} + h^2\frac{11}{96}a(x_0)u^{(4)}(x_0) \\
&\quad + \mathcal{O}(h^3) \\
\sum_{j=0}^{N} L_{1j}u(x_j) &= (au_x)_x|_{x_1} - \frac{59}{372}h(au_x)_{xx}|_{x_1} + \frac{55}{744}h^2(au_x)_{xxx}|_{x_1} \\
&\quad - h^2\frac{1}{12}a(x_1)u^{(4)}(x_1) + \mathcal{O}(h^3) + \frac{59}{372}h(au_x)_{xx}|_{x_0} \\
&= (au_x)_x|_{x_1} - \frac{21}{248}h^2(au_x)_{xxx}|_{x_0} - h^2\frac{1}{12}a(x_0)u^{(4)}(x_0) + \mathcal{O}(h^3) \\
\sum_{j=0}^{N} L_{2j}u(x_j) &= (au_x)_x|_{x_2} + \frac{19}{60}h(au_x)_{xx}|_{x_2} - \frac{31}{120}h^2(au_x)_{xxx}|_{x_2} \\
&\quad + h^2\frac{5}{48}a(x_2)u^{(4)}(x_2) + \mathcal{O}(h^3) - \frac{19}{60}h(au_x)_{xx}|_{x_0} \\
&= (au_x)_x|_{x_2} + \frac{3}{8}h^2(au_x)_{xxx}|_{x_0} + h^2\frac{5}{48}a(x_0)u^{(4)}(x_0) + \mathcal{O}(h^3) \\
\sum_{j=0}^{N} L_{3j}u(x_j) &= (au_x)_x|_{x_3} - \frac{31}{300}h(au_x)_{xx}|_{x_3} + \frac{27}{200}h^2(au_x)_{xxx}|_{x_3} \\
&\quad - h^2\frac{1}{60}a(x_3)u^{(4)}(x_3) + \mathcal{O}(h^3) + \frac{31}{300}h(au_x)_{xx}|_{x_0} \\
&= (au_x)_x|_{x_3} - \frac{7}{40}h^2(au_x)_{xxx}|_{x_0} - h^2\frac{1}{60}a(x_0)u^{(4)}(x_0) + \mathcal{O}(h^3)
\end{aligned} \tag{5.43}$$

Finally we get the local truncation error T_j at each irregular grid point (left boundary) by

$$u_t - (Lu)_j - M_{NN}^{-1}a(1)g_1(t) + M_{00}^{-1}a(0)g_0(t) = T_j, \tag{5.44}$$

where

$$\begin{aligned}
T_0 &= \frac{1}{16}h^2(au_x)_{xxx}|_{x_0} - h^2\frac{11}{96}a(x_0)u^{(4)}(x_0) + \mathcal{O}(h^3) \\
T_1 &= \frac{21}{248}h^2(au_x)_{xxx}|_{x_0} + h^2\frac{1}{12}a(x_0)u^{(4)}(x_0) + \mathcal{O}(h^3) \\
T_2 &= -\frac{3}{8}h^2(au_x)_{xxx}|_{x_0} - h^2\frac{5}{48}a(x_0)u^{(4)}(x_0) + \mathcal{O}(h^3) \\
T_3 &= \frac{7}{40}h^2(au_x)_{xxx}|_{x_0} + h^2\frac{1}{60}a(x_0)u^{(4)}(x_0) - \mathcal{O}(h^3)
\end{aligned}$$

The truncation error for the right boundary can be calculated analogously.

Remark 5.2. *In the first row u_x is approximated additionally. Therefore it might be reasonable not to scale the error in this row. Nevertheless we prefer to do so, since we treat the operator as a finite difference scheme of the form* (5.41) *and use the scaling given in this form. However we state that the error in the first derivative at $x = 0$ and $x = 1$ would be one order less than the error of the scheme as we treated it above.*

Now we consider **Dirichlet** data, where we get the following expansion

$$\begin{aligned}
\sum_{j=0}^{N} L_{1j}u(x_j) &= \frac{372}{313}(au_x)_x|_{x_1} - \frac{59}{313}h(au_x)_{xx}|_{x_1} + \frac{55}{626}h^2(au_x)_{xxx}|_{x_1} \\
&\quad - h^2\frac{31}{313}a(x_1)u^{(4)}(x_1) + \mathcal{O}(h^3) - \frac{59}{313}h(au_x)_x|_{x_0} \\
&= (au_x)_x|_{x_1} - \frac{2}{313}h^2(au_x)_{xxx}|_{x_0} - h^2\frac{31}{313}a(x_0)u^{(4)}(x_0) + \mathcal{O}(h^3) \\
\sum_{j=0}^{N} L_{2j}u(x_j) &= \frac{120}{139}(au_x)_x|_{x_2} + \frac{38}{139}h(au_x)_{xx}|_{x_2} - \frac{31}{139}h^2(au_x)_{xxx}|_{x_2} \\
&\quad + h^2\frac{25}{278}a(x_2)u^{(4)}(x_2) + \mathcal{O}(h^3) + \frac{19}{139}h(au_x)_x|_{x_0} \\
&= (au_x)_x|_{x_2} + \frac{7}{139}h^2(au_x)_{xxx}|_{x_0} + h^2\frac{25}{278}a(x_0)u^{(4)}(x_0) + \mathcal{O}(h^3) \\
\sum_{j=0}^{N} L_{3j}u(x_j) &= \frac{900}{869}(au_x)_x|_{x_3} - \frac{93}{869}h(au_x)_{xx}|_{x_3} + \frac{243}{1738}h^2(au_x)_{xxx}|_{x_3} \\
&\quad - h^2\frac{15}{869}a(x_3)u^{(4)}(x_3) + \mathcal{O}(h^3) - \frac{31}{869}h(au_x)_x|_{x_0} \\
&= (au_x)_x|_{x_3} - \frac{18}{869}h^2(au_x)_{xxx}|_{x_0} - h^2\frac{15}{869}a(x_0)u^{(4)}(x_0) + \mathcal{O}(h^3)
\end{aligned} \tag{5.45}$$

Finally we get the local truncation error T_j at each irregular grid point (left boundary) by

$$u_t - (Lu)_j = T_j, \tag{5.46}$$

where

$$\begin{aligned}
T_0 &= 0 \\
T_1 &= \frac{2}{313}h^2(au_x)_{xxx}|_{x_0} + h^2\frac{31}{313}a(x_0)u^{(4)}(x_0) + \mathcal{O}(h^3) \\
T_2 &= -\frac{7}{139}h^2(au_x)_{xxx}|_{x_0} - h^2\frac{25}{278}a(x_0)u^{(4)}(x_0) + \mathcal{O}(h^3) \\
T_3 &= \frac{18}{869}h^2(au_x)_{xxx}|_{x_0} + h^2\frac{15}{869}a(x_0)u^{(4)}(x_0) + \mathcal{O}(h^3)
\end{aligned}$$

The truncation error for the right boundary can be calculated analogously.

5.5 Global Error

In section 5.4 we have derived the local error T at each grid point. Now we are interested in an estimate of the global error $w_i = u(x_i) - v_i$. In order to obtain an estimate for the global error, we insert the exact solution u into the discrete scheme and subtract Lv, i.e. $Lu - Lv$.

This yields the error equation

$$\begin{aligned} &w_t = Lw + T \\ &w(0) = 0, \\ &\text{Neumann: } w_x(0) = 0, \quad w_x(1) = 0. \\ &\text{Dirichlet: } w(0) = 0, \quad w(1) = 0. \end{aligned} \tag{5.47}$$

For Dirichlet data we impose the boundary conditions, which is why the boundary conditions for the solution and its approximation are the same and the error vanishes at the boundary. In case of Neumann data, however, the derivative of the error at the boundary is not exactly zero, but of order h^3 (cf. remark 5.2). This is due to the fact that we approximate the first derivative in the boundary lines of the operator and subtract the given data for the derivative. Therefore we prescribe homogeneous boundary conditions for the global error, while the error in its derivative is part of the truncation error that we have found at the boundary. In the following, we will denote the boundary conditions by

$$B_0 w(0) = 0, \quad B_1 w(1) = 0,$$

where $B_{0,1}$ may stand for Dirichlet or Neumann data.
For the semi-discrete approximation (5.47) we have an energy estimate as it will be shown in section 5.6, which leads us to the following error estimate

$$\|w\|_M \leq \mathcal{O}(h^2) \tag{5.48}$$

However this error estimate is not optimal. For finite difference schemes it is often the case that the global error is of higher order than the local truncation error in the neighborhood of boundaries. As shown by Gustafsson [6], the local error at the boundary can be one order less than the global error, if there exists an energy estimate (cf. section 5.6). However, even this estimate is not sharp in our case.
In order to obtain an optimal error estimate, we split the error into interior and boundary part, i.e. $w = w^i + w^b$, where

$$\begin{aligned} &w_t^i = Lw^i + T^i \\ &w^i(0) = 0, \\ &B_0 w^i(0) = 0, \quad B_1 w^i(1) = 0, \end{aligned} \tag{5.49}$$

with $T^i = [0, 0, 0, 0, T_4, ..., T_{N-4}, 0, 0, 0, 0]$ for Neumann and $T^i = [0, 0, 0, T_4, ..., T_{N-4}, 0, 0, 0]$ for Dirichlet data, respectively, and

$$\begin{aligned} &w_t^b = Lw^b + T^b, \\ &w^b(0) = 0, \\ &B_0 w^b(0) = 0, \quad B_1 w^b(1) = 0, \end{aligned}$$

with $T^b = [T_0, T_1, T_2, T_3, 0, ..., 0, T_{N-3}, T_{N-2}, T_{N-1}, T_N]$ for Neumann boundary data and $T^b = [T_1, T_2, T_3, 0, ..., 0, T_{N-3}, T_{N-2}, T_{N-1}]$ for Dirichlet data, respectively.
For (5.49) we get an error estimate analogously to (5.48):

$$\|w\|_M \leq \mathcal{O}(h^4) \tag{5.50}$$

To investigate w^b, we first simplify the semi-discrete problem by Laplace transforming in time, i.e. we consider the problem

$$\begin{aligned} &\widehat{L}\widehat{w}^b = T, \\ &B_0 \widehat{w}^b(0) = 0, \quad B_1 \widehat{w}^b(1) = 0, \end{aligned}$$

where

$$\widehat{L} = -L + s \tag{5.51}$$

where $s \in \mathbb{C}$, $\operatorname{Re} s \geq \text{const} > 0$. We assume a to be independent of time.
In order to get an estimate for $\widehat{w}^b$, we investigate the operator $\widehat{L}$ and prove some properties. For this purpose we introduce the continuous analogon of (5.51) by

$$\widehat{L}_c = \left(-\frac{\mathrm{d}}{\mathrm{d}x} a \frac{\mathrm{d}}{\mathrm{d}x} + s \right)$$

We want to characterize the solution to the equation

$$\begin{aligned} &\widehat{L}_c u(x) = f(x) \\ &B_0 u(0) = 0, \quad B_1 u(1) = 0, \end{aligned} \tag{5.52}$$

with some integrable function f. This can be done with the help of the Green's function defined by

$$\begin{aligned} &\widehat{L}_c G(x,\xi) = \delta(x-\xi) \\ &B_0 G(0,\xi) = 0, \quad B_1 G(1,\xi) = 0 \end{aligned} \tag{5.53}$$

where δ stands for the Dirac distribution. The solution to (5.52) would then be given by

$$u(x) = \int_0^1 G(x,\xi) f(\xi) \mathrm{d}\xi$$

The Green's function can be expanded in terms of the eigenfunctions of the operator $L_c = \frac{\mathrm{d}}{\mathrm{d}x}\left(a\frac{\mathrm{d}}{\mathrm{d}x}\right)$ (cf. [15], for Green's function of self-adjoint operators see [5]). Let μ_n be an eigenfunction of the real valued, self-adjoint operator L_c to the eigenvalue λ_n, i.e.

$$\begin{aligned} &L_c \mu_n(x) = \lambda_n \mu_n(x) \\ &B_0 \mu_n(0) = 0, \quad B_1 \mu_n(1) = 0. \end{aligned}$$

In [15] it is shown that all μ_n as well as all λ_n are real and that the Green's function defined by (5.53) is given by the bilinear series

$$G(x,\xi) = \sum_n \frac{\mu_n(x)\mu_n(\xi)}{s - \lambda_n}.$$

Furthermore all eigenvalues of L_c are nonpositive, if $a(x) \geq a_{\min} > 0$, since

$$(L_c u, u) = \int_0^1 a(x) u \Delta u \, \mathrm{d}x \leq -a_{\min} \int_0^1 \nabla u \nabla u \, \mathrm{d}x = -a_{\min} \|\nabla u\|^2 \leq 0.$$

Since λ_n is a nonpositive real number for all n and $\operatorname{Re} s > 0$, the bilinear series has no singularity.
From this representation it can be seen that the Green's function is symmetric in its arguments x and ξ. Furthermore G is continuous even for $x = \xi$ and its derivative is continuous except for $x = \xi$ where it has a jump of magnitude $-\frac{1}{a(x)}$. If we assume $a(x) \geq a_{\min} > 0$ the jump of the derivative is bounded by $\frac{1}{a_{\min}}$ (cf. [15]).
Let $y \in [0,1]$ be fixed and $\eta = 0 + \mathcal{O}(h)$, where h is the step size of the discrete problem. Then it holds

$$\frac{\mathrm{d}}{\mathrm{d}x} G(\eta, y) = \frac{\mathrm{d}}{\mathrm{d}x} G(0, y) + \mathcal{O}(1)$$

because the jump of the derivative is bounded by $\frac{1}{a_{\min}}$, which is independent of h. If $y > \eta$, $\frac{\partial}{\partial x}G(\cdot, y)$ is continuous in the interval $[0, \eta]$ and thus we even get an estimate of the oder h. Using this argument and that G is symmetric, we get

$$G(y, \eta) = G(\eta, y) = G(0, y) + h\frac{\mathrm{d}}{\mathrm{d}x}G(0, y) + \mathcal{O}(h) \tag{5.54}$$

For homogeneous Neumann boundary conditions we have

$$G(x_j, \xi) = G(x_j, x_0) + \mathcal{O}(h), \quad j = 0, \ldots, N,$$

and for homogeneous Dirichlet data

$$G(x_j, \xi) = G(x_j, x_0) + \mathcal{O}(h) = G(x_0, x_j) + \mathcal{O}(h) = \mathcal{O}(h), \quad j = 0, \ldots, N.$$

Now we consider the discrete problem. Like in the continuous case, we define the discrete Green's function G^h by

$$\widehat{L}G^h_{jl} = \delta^h_{jl},$$

where

$$\delta^h_{jl} = \begin{cases} \frac{1}{h}, & \text{if } j = l \\ 0, & \text{else.} \end{cases}$$

We want to investigate the relation between the discrete and the continuous Green's function. Let $l \in \{0, 1, 2, 3\}$. Consider the following continuous problem

$$\widehat{L}\Gamma_l(x) = \frac{1}{h}\chi_{[x_l, x_l+h[}(x), \tag{5.55}$$

where

$$\chi_{[a,b]}(x) = \begin{cases} 1, & x \in [a, b] \\ 0, & \text{elsewhere.} \end{cases}$$

Since for all $j \in \{0, 1, 2, ..., N-1, N\}$ it holds that

$$\delta^h_{jl} = \frac{1}{h}\chi_{[x_l, x_l+1[}(x_j),$$

we can use (5.48) to obtain

$$\Gamma_l(x_j) - G^h_{jl} = \mathcal{O}(h^2) \text{ for all } j \in \{0, 1, 2, ..., N-1, N\}. \tag{5.56}$$

The solution to (5.55) can be given using the Green's function

$$\Gamma_l(x) = \frac{1}{h}\int_{x_l}^{x_l+1} G(x, \xi)\mathrm{d}\xi \tag{5.57}$$

Since $x_l = x_0 + \mathcal{O}(h)$, we can apply the estimate (5.54), which yields

$$\Gamma_l(x) = G(0, x) + h\frac{\mathrm{d}}{\mathrm{d}x}G(0, x) + \mathcal{O}(h) \tag{5.58}$$

Combining (5.56) and (5.58) yields the following estimate for the discrete Green's function

$$G_{jl} = G(0, x_j) + h\frac{\mathrm{d}}{\mathrm{d}x}G(0, x_j) + \mathcal{O}(h)$$

or for the special case of homogeneous Neumann conditions

$$G_{jl} = G(0, x_j) + \mathcal{O}(h)$$

and for homogeneous Dirichlet conditions

$$G_{jl} = \mathcal{O}(h)$$

For $l \in \{N-3, N-2, N-1, N\}$ we can obtain an analoguous estimate by defining Γ_l by

$$\widehat{L}\Gamma_l(x) = \frac{1}{h}\chi_{]x_l - h, x_l]}(x).$$

Now we can use the discrete Green's function to describe the solution to (5.47). We first look at the Neumann case

$$\begin{aligned}\widehat{w}^b(x_j) &= \sum_{l=0}^{N} hG_{jl}T_l = \sum_{l=0}^{3} hG_{jl}T_l + \sum_{l=N-3}^{N} hG_{jl}T_l \\ &= \sum_{l=0}^{3} h\left(G(0, x_j) + \mathcal{O}(h)\right)T_l + \sum_{l=N-3}^{N} h\left(G(1, x_j) + \mathcal{O}(h)\right)T_l = \mathcal{O}(h^4)\end{aligned}$$

In the Dirichlet case it is

$$\begin{aligned}\widehat{w}^b(x_j) &= \sum_{l=0}^{N} hG_{jl}T_l = \sum_{l=0}^{3} hG_{jl}T_l + \sum_{l=N-3}^{N} hG_{jl}T_l = \mathcal{O}(h)\left(\sum_{l=0}^{3} hT_l + \sum_{l=N-3}^{N} hT_l\right) \\ &= \mathcal{O}(h^4).\end{aligned}$$

Hence we can estimate the boundary error w^b in the discrete M-norm using Parseval's relation by

$$\|w^b\|_M \leq \mathcal{O}(h^4). \tag{5.59}$$

Combining this with the estimate (5.50), it follows applying the triangle inequality that

$$\|w\|_M \leq \mathcal{O}(h^4). \tag{5.60}$$

The same estimate also holds in the discrete l_2-norm.

5.6 Stability

In order to prove stability of the discretization, we use the energy method derived in [7]. We consider homogeneous boundary conditions here, which is sufficient in the case of boundary condition which are differentiable with a bounded first derivative (cf. [7]). We prefer this method since our method requires differentiable boundary data anyway and this poof has the advantage that we do not require any regularity on the function a.

We define an inner product using the lumped mass matrix M in the following way:

$$(u, v)_M = u^T M v$$

Let $||u||_M = \sqrt{(u, u)_M}$ be the corresponding norm. Using this energy norm, we have:

$$\frac{\mathrm{d}}{\mathrm{d}t}\|v\|_M^2 = v^T(S + S^T)v$$

We treat both boundary conditions together because the 0th and Nth row is put to zero in the case of homogeneous Dirichlet data anyway.
Since S is symmetric, stability is equivalent to $-S = R$ positive semidefinite. The element (i, j) of R approximates the following integral:

$$\int_0^1 a(x,t)\varphi_i'\varphi_j' \,\mathrm{d}x$$

For sufficiently smooth a, stability follows immediately. However, we will give a proof in which we do not have to claim any smoothness of a.

The above integral is approximated using the Simpson rule between two grid points, i.e.

$$\begin{aligned}\int_0^1 a(x,t)\varphi_i'\varphi_j' \,\mathrm{d}x = \frac{h}{6}\sum_{k=-1}^{2} \big(&a(x_{i-k})\varphi_i'(x_{i-k}^+)\varphi_j'(x_{i-k}^+) + 4a(x_{i-k+1/2})\varphi_i'(x_{i-k+1/2})\varphi_j'(x_{i-k+1/2}) \\ &+ a(x_{j-k+1})\varphi_i'(x_{i-k+1}^-)\varphi_j'(x_{i-k+1}^-)\big) + \mathcal{O}(h^5) = (\phi^{(i)})^T\Lambda\phi^{(j)} + \mathcal{O}(h^5),\end{aligned}$$

where

$$\phi_{3k+l}^{(i)} = \begin{cases} (\varphi^{(i)})'(x_k^+), & l = 0, k = 0, 1, \ldots, N-1 \\ (\varphi^{(i)})'(x_{k+1/2}), & l = 1, k = 0, 1, \ldots, N-1 \\ (\varphi^{(i)})'(x_{k+1}^-), & l = 2, k = 0, 1, \ldots, N-1. \end{cases}$$

and

$$\Lambda = \frac{h}{6}\,\mathrm{diag}(a(x_0), 4a(x_{1/2}), a(x_1^-), a(x_1^+), 4a(x_{1+1/2}), a(x_2^-), \ldots, a(x_N)).$$

Using this vector notation, we have

$$R_{ij} = (\phi^{(i)})^T\Lambda\phi^{(j)}.$$

For $v \in \mathbb{R}^{N+1}$ this leads to

$$v^TRv = \sum_i\sum_j v_iv_j((\phi^{(i)})^T\Lambda\phi^{(j)}) = \sum_i\sum_j((v_i\phi^{(i)})^T\Lambda(v_j\phi^{(j)})) = \Phi^TP\Phi, \tag{5.61}$$

where

$$\Phi_{i+k} = v_i\phi_k^{(i)}, \quad k = 0, 1, 2, \ldots, 3N-1, \quad i = 0, 1, 2, \ldots, N$$

and $P \in \mathbb{R}^{(N+1)(3N)\times(N+1)(3N)}$ with

$$P_{i+k,j+l} = \Lambda_{k,l}, \quad k, l = 0, 1, 2, \ldots, 3N-1, \quad i, j = 0, 1, 2, \ldots, N.$$

Since the matrix P is symmetric, it is positive semi-definite if and only if all eigenvalues are non-negative.
First we state that $\mathrm{rank}(P) = \mathrm{rank}(\Lambda)$. Thus P has at most $3N$ eigenvalues that are different from 0.
Next we prove the following

Lemma 5.3. *Let λ_i be the ith diagonal element of Λ ($i \in \{0, \ldots, 3N-1\}$). Then $(3N)\lambda_i$ is an eigenvalue of P with corresponding eigenvector $v^{(i)}$,*

$$v_k^{(i)} = \begin{cases} 1, & k = i + (3N)l, \quad l = 0, 1, \ldots, N \\ 0, & \text{else} \end{cases}$$

Proof.

$$
\begin{aligned}
(Pv^{(i)})_k &= \sum_{l=0}^{N} P_{k,l} v_l^{(i)} = \sum_{m=0}^{N} P_{k,i+m(3N)} \\
&= \begin{cases} (3N)\lambda_i, & k = i + (3N)l, \quad l = 0, 1, \ldots, N \\ 0, & \text{else} \end{cases} \\
&= ((3N)\lambda_i v^{(i)})_k.
\end{aligned}
$$

□

For $a(x,t) > 0$ all $\lambda_i = a(x_{i/2}) > 0$. Thus all eigenvalues of P are either positive or 0.

From that we can finally conclude that R is positive semi-definite and thus the operator is strictly stable, since we have shown that the energy is diminishing.

5.7 Damping of π-Modes

In this section we want to investigate how the highest-frequency waves, the π-modes, are damped by the operator derived with finite elements. The theory was derived in section 3.4. The Fourier transform $\widehat{L}$ of the inner stencil of the operator L for constant a is given by

$$
\widehat{L}(\xi) = a \frac{1}{h^2} \left(-\frac{5}{144} \cos(3\xi) + \frac{1}{24} \cos(2\xi) + \frac{103}{48} \cos(\xi) - \frac{155}{72} \right), \tag{5.62}
$$

where $\xi = 2\pi\omega h$. Figure 3 shows how the different wave numbers are damped by the operator L compared to the the continuous operator ($a \equiv 1$). The agreement gets worse with increasing wave number. But there is a damping even for the high wave numbers.

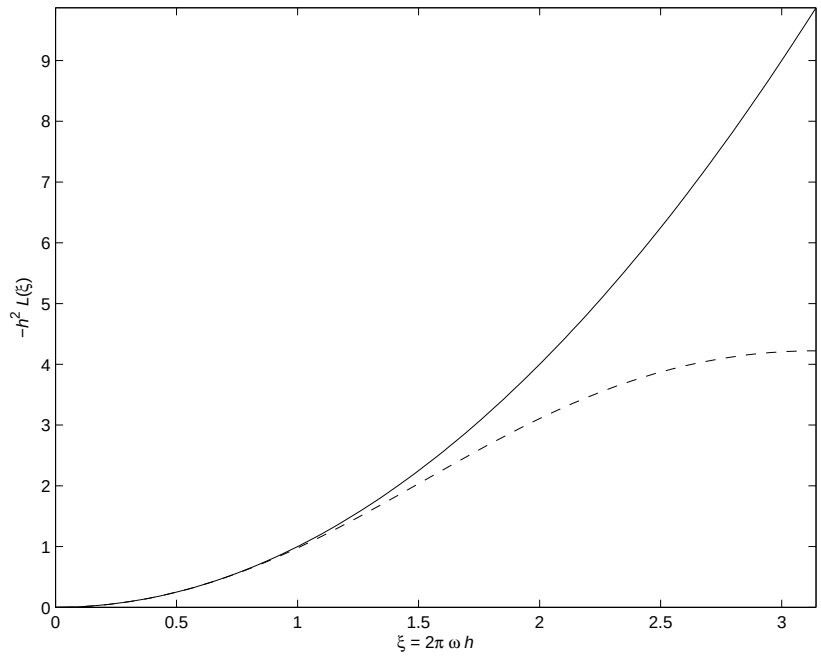

Figure 3: Damping of modes with wave number ξ of the exact solution (-) and the FEM operator with mass lumping (- -).

5.8 Computational Results

In this section we present some numerical experiments which have been performed using the operator L and the classical fourth order Runge-Kutta method for time integration.

Since the time integrator is an explicit method, we are faced with a stability limit for the time integration. The limit on the time step can be detected by Fourier transforming the space variable of our difference equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{v}_\omega(t) = \hat{L}(\omega)\hat{v}_\omega(t)$$

Equation (5.62) gives the Fourier transformed operator for constant a at the interior points. If a is variable, the coefficients will depend on a. These coefficients can be expanded into their Taylor series. If we neglect all terms in $\mathcal{O}(h)$, we only have to take into account the term with a and none of its derivatives.
However for Lipschitz continuous a it is sufficient to consider the problem with $a_{\max} = \max_{x\in[0,1]} a(x)$ instead of a variable a because for each row we can estimate the a-depending coefficients by $a(x_j) + \mathcal{O}(h)$ in the jth row. Thus we consider the operator L for $a \equiv a_{\max}$.
From (5.62) we get an estimate for $\hat{L}$ based on the stencil for the interior points

$$0 \geq \hat{L} \geq -\frac{38}{9}$$

As described in section 3.5, this leads to the following limit on the time step

$$\Delta t \leq \frac{0.66}{a_{\max}}$$

However computations close to this limit show that this estimate is not rigid enough. Plotting the error over the grid points in space shows that the instability comes from the boundaries. Therefore it is necessary to consider also the Fourier transform of the irregular stencil in the neighborhood of the boundaries.
Since the irregular stencils are not symmetric around the diagonal, $\hat{L}(\xi)$ is a function with complex codomain and negative real part. Table 13 gives the value of $|\hat{L}|$ in each irregular row for Neumann and Dirichlet data: The stability domain of the Runge-Kutta method has

row	Neumann	Dirichlet
0	1.02	-
1	8.43	7.10
2	4.95	7.97
3	4.77	4.61

Table 13: Maximum for $\hat{L}$ (irregular lines)

a little complicated shape. However, except for the first row where the radius is very small anyway, the imaginary part of $\widehat{L}$ is smaller than 1. Close to the negative real axis the stability domain comprises the semi-circle around 0 with radius 2.8. Therefore we require $|\hat{L}| \leq 2.8$. From these data we can finally get a feasible limit on the time step

$$\text{Neumann: } \Delta t \leq \frac{0.33}{a_{\max}} \quad \text{and} \quad \text{Dirichlet: } \Delta t \leq \frac{0.36}{a_{\max}}.$$

In our experiments we multiplied the limit by 0.9 in order to keep it a bit away from the stability limit in order to avoid difficulties.
The global error $\|u - v\|_h$ is calculated in the discrete h-norm, where we use as reference u

either the exact solution, if it can easily be calculated analytically, or the numerical solution calculated with $N = 960$. Out of these results we calculate the convergence rate using the formula

$$q_{D,N} = \log\left(\frac{\|u-v\|_{h_1}}{\|u-v\|_{h_2}}\right) \Big/ \log(2), \quad h_1 = 2h_2$$

Remark:
Since the time step is of the order h^2, the main error will be the one made by the discretization in space. Hence the convergence analysis for the operator L will not be disturbed be the time discretization error.
We will now consider the same test problems as for the product rule operator.

Example 1 We consider the problem

$u_t = 0.2u_{xx}$
$u(x,0) = \sin(\pi x) + \cos(\pi x)$
Neumann: $u_x(0,t) = \pi e^{-0.2\pi^2 t}, \quad u_x(1,t) = -\pi e^{-0.2\pi^2 t}$ or
Dirichlet: $u_x(0,t) = e^{-0.2\pi^2 t}, \quad u_x(1,t) = -e^{-0.2\pi^2 t}$

with the exact solution

$u(x,t) = (\sin(\pi x) + \cos(\pi x))e^{-0.2\pi^2 t}$.

Table 14 shows the results at time $t = 1$.

	Neumann		Dirichlet	
N	$\|u-v\|_h$	q_N	$\|u-v\|_h$	q_D
10	$2.28 \cdot 10^{-4}$		$7.39 \cdot 10^{-5}$	
20	$1.13 \cdot 10^{-5}$	4.33	$4.51 \cdot 10^{-6}$	4.03
40	$5.60 \cdot 10^{-7}$	4.34	$2.80 \cdot 10^{-7}$	4.01
80	$2.98 \cdot 10^{-8}$	4.23	$1.74 \cdot 10^{-8}$	4.00
160	$1.70 \cdot 10^{-9}$	4.13	$1.09 \cdot 10^{-9}$	4.00
320	$1.02 \cdot 10^{-10}$	4.06	$6.76 \cdot 10^{-11}$	4.01

Table 14: Numerical results for the equation $u_t = 0.2u_{xx}$

Example 2 We consider the problem

$u_t = \frac{\partial}{\partial x}((0.2 + 0.4x(x-1))u_x)$
$u(x,0) = \sin(\pi x) + \cos(\pi x)$
Neumann: $u_x(0,t) = \pi e^{-0.2\pi(\pi+2)t}, \quad u_x(1,t) = -\pi e^{-0.2\pi(\pi-2)t}$ or
Dirichlet: $u(0,t) = e^{-0.2\pi(\pi+2)t}, \quad u(1,t) = -e^{-0.2\pi(\pi-2)t}$.

Table 15 shows the results at time $t = 1$.

Example 3 We consider the problem

$u_t = \frac{\partial}{\partial x}(0.2(1 + sin(\pi x))u_x)$
$u(x,0) = \sin(\pi x) + \cos(\pi x)$
Neumann: $u_x(0,t) = \pi, \quad u_x(1,t) = -\pi e^{-0.4\pi^2 t}$ or
Dirichlet: $u(0,t) = 1, \quad u(1,t) = -e^{-0.4\pi^2 t}$.

	Neumann		Dirichlet	
N	$\|u-v\|_h$	q_N	$\|u-v\|_h$	q_D
10	$4.10 \cdot 10^{-4}$		$2.36 \cdot 10^{-4}$	
20	$2.04 \cdot 10^{-5}$	4.33	$1.48 \cdot 10^{-5}$	4.00
40	$1.09 \cdot 10^{-6}$	4.22	$9.36 \cdot 10^{-7}$	3.99
80	$6.33 \cdot 10^{-8}$	4.11	$5.86 \cdot 10^{-8}$	4.00
160	$3.81 \cdot 10^{-9}$	4.06	$3.66 \cdot 10^{-9}$	4.00
320	$2.30 \cdot 10^{-10}$	4.05	$2.27 \cdot 10^{-10}$	4.01

Table 15: Numerical results for the equation $u_t = ((0.2 + 0.4x(x-1))u_x)_x$

	Neumann		Dirichlet	
N	$\|u-v\|_h$	q_N	$\|u-v\|_h$	q_D
10	$2.12 \cdot 10^{-4}$		$9.02 \cdot 10^{-5}$	
20	$1.84 \cdot 10^{-5}$	3.53	$7.60 \cdot 10^{-6}$	3.57
40	$1.27 \cdot 10^{-6}$	3.85	$5.20 \cdot 10^{-7}$	3.87
80	$8.02 \cdot 10^{-8}$	3.99	$3.24 \cdot 10^{-8}$	4.01
160	$4.91 \cdot 10^{-9}$	4.03	$2.03 \cdot 10^{-9}$	4.00
320	$2.83 \cdot 10^{-10}$	4.12	$1.48 \cdot 10^{-10}$	3.78

Table 16: Numerical results for the equation $u_t = (0.2(1 + \sin(\pi x))u_x)_x$

Table 16 shows the results at time $t = 1$.

Example 4 We consider the problem

$$u_t = \frac{\partial}{\partial x}\left(0.5\mathrm{e}^{x-2}(1 + sin(\pi t))u_x\right)$$
$$u(x,0) = (\sin(\pi x))^2 + 2x$$
Neumann: $u_x(0,t) = 2\,\mathrm{e}^{\mathrm{e}^{-2}(\pi^2+1)t}, \quad u_x(1,t) = 2\,\mathrm{e}^{e^{-1}(\pi^2+1)t}$ or
Dirichlet: $u(0,t) = 0, \quad u(1,t) = 2\mathrm{e}^{e^{-1}(\pi^2+1)t}$.

Table 17 shows the results at time $t = 1$

	Neumann		Dirichlet	
N	$\|u-v\|_h$	q_N	$\|u-v\|_h$	q_D
10	$3.55 \cdot 10^{-2}$		$4.65 \cdot 10^{-2}$	
20	$2.06 \cdot 10^{-3}$	4.11	$3.67 \cdot 10^{-3}$	3.66
40	$1.11 \cdot 10^{-4}$	4.22	$2.53 \cdot 10^{-4}$	3.87
80	$5.67 \cdot 10^{-6}$	4.29	$1.62 \cdot 10^{-5}$	3.97
160	$2.78 \cdot 10^{-7}$	4.35	$1.01 \cdot 10^{-6}$	4.00
320	$1.34 \cdot 10^{-8}$	4.38	$6.19 \cdot 10^{-8}$	4.03

Table 17: Numerical results for the equation $u_t = \left(0.5\mathrm{e}^{x-2}(1 + \sin(\pi t))u_x\right)_x$

6 Finite Element Ansatz for the Convection Diffusion Equation

In this section we give a short overview of how the finite element ansatz can be extended to the convection diffusion problem. The operator L derived in section 5 is strictly stable in the M-norm. If we want to apply it to the convection diffusion equation, we have to design an operator for the first derivative that is strictly stable in the same norm. Therefore we cannot use the SBP operator for the first derivative derived by Strand [16].

We can however use the finite element ansatz that has led us to L also for the first derivative. We discuss two different possibilities of doing so.

The first ansatz is to handle the hyperbolic parts like the parabolic part by evaluating the stiffness matrix using numerical integration. This leads to an operator that is strictly stable for the convection diffusion equation if we require continuous and piecewise continuously differentiable coefficient functions a, b and c. However the local order of accuracy at the boundary will be reduced to order h only because of the additional errors due to numerical integration.

Therefore we propose an alternative treatment, which leads to a strictly stable operator if all coefficients are continuous and piecewise continuously differentiable. In this case the coefficient functions are approximated by their Lagrange interpolant and the resulting integrals are evaluated exactly. This idea was already described in remark 5.1.

Before discussing the numerical treatment, we will extend the variational formulation of the semi-discrete problem to the convection diffusion equation.

6.1 Variational Formulation

We want to derive the variational formulation of the partial differential equation (1.2) with Neumann boundary conditions. We define the Sobolev space V like in section 5.2. We first multiply (1.2) by a function $v \in V$ and integrate over $[0, 1]$, which yields

$$\begin{aligned}\int_0^1 u_t v \,\mathrm{d}x &= \int_0^1 \left((au_x)_x + bu_x + (cu)_x\right) v \,\mathrm{d}x \\ &= \int_0^1 \left(-au_x v_x + bu_x v - cuv_x\right) \,\mathrm{d}x + v(au_x + cu)\big|_0^1\end{aligned}$$

If we introduce V_h as a subspace of V with the basis functions φ_j, $j = 0, 1, \ldots, N$, we obtain an ODE system with mass and stiffness matrix analogously as in section 5.2 of the form

$$M^h \frac{\mathrm{d}}{\mathrm{d}t} v = T^h v + a_N g_1 e_N - a_0 g_0 e_0 + c_N v_N - c_0 v_0,$$

where the mass matrix M^h is the same as in the pure parabolic case and the new stiffness matrix T^h is given by

$$T^h_{kl} = -\int_0^1 a(x,t)\varphi_k'(x)\varphi_l'(x)\,\mathrm{d}x + \int_0^1 b(x,t)\varphi_k(x)\varphi_l'(x)\,\mathrm{d}x - \int_0^1 c(x,t)\varphi_k'(x)\varphi_l(x)\,\mathrm{d}x.$$

Remark:
Like in the case of the pure parabolic problem, we handle Dirichlet data by neglecting the first and the last line of the system and imposing the physical boundary conditions at the boundary points.

6.2 Operator Based on Numerical Integration

6.2.1 Construction

We want to solve the convection diffusion equation using the operator derived in section 5 for the parabolic part. Therefore we use the mass matrix M with its correction terms $m_{0,1}$ and the stiffness matrix S (notation from section 5). The stiffness matrix S has to be extended by some part T^b and T^c taking care of the two hyperbolic terms in equation (1.2). The integrals appearing in T^b and T^c shall also be approximated using the Simpson rule.

6.2.2 Stability

The stability proof will be similar to that in section 5.6. Also here we will consider homogeneous boundary conditions. We show that the method is strictly stable by finding an estimate for the discrete energy in the M-norm identical to the estimate for the continuous problem neglecting higher order terms. The estimate for the analytic problem is given by (2.14) and (2.16) for Dirichlet and Neumann data, respectively.

Differentiating the discrete energy of some vector $v \in \mathbb{R}^{N+1}$ yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\|v\|_M^2 = v^T(T+T^T)v + 2v_N(a_N(v_x)_N + c_N v_N) - 2v_0(a_0(v_x)_0 + c_0 v_0) \tag{6.1}$$

We first look at the inner part. The matrix $T = S + T^b + T^c$ is an approximation of the stiffness matrix using Simpson's rule. A Taylor expansion shows that the error of this approximation is in $\mathcal{O}(h^2)$ at the boundary and in $\mathcal{O}(h^4)$ in the interior if all coefficients are sufficiently smooth. We require the coefficients to be continuous and piecewise continuously differentiable such that the integration error of each single integral is in $\mathcal{O}\left(h^2\right)$. Hence it holds

$$\begin{aligned}
v^T(T+T^T)v = \sum_{k,l=0}^{N} v_k v_l \Bigg(& -2\int_0^1 a(x,t)\varphi_k'(x)\varphi_l'(x)\,\mathrm{d}x \\
& + \int_0^1 b(x,t)\left(\varphi_k(x)\varphi_l'(x) + \varphi_k'(x)\varphi_l(x)\right)\,\mathrm{d}x \\
& - \int_0^1 c(x,t)\left(\varphi_k'(x)\varphi_l(x) + \varphi_k(x)\varphi_l'(x)\right)\,\mathrm{d}x\Bigg) + \mathcal{O}(h) \\
= 2\Bigg(& -\int_0^1 a(x,t) z_x(x,t) z_x(x,t)\,\mathrm{d}x \\
& + \int_0^1 (b(x,t) - c(x,t))\left(z(x,t) z_x(x,t)\right)\,\mathrm{d}x\Bigg) + \mathcal{O}(h),
\end{aligned}$$

where we define the function $z(x,t) = \sum_{l=0}^{N} v_l \varphi_l(x,t)$, which equals v_l at the grid points x_l. We use the Cauchy-Schwarz inequality and estimate the functions a, b and c, which yields

$$v^T(T+T^T)v \le 2\left(-a_{\min}\|z_x(\cdot,t)\|^2 + \|b(\cdot,t) - c(\cdot,t)\|_\infty \|z(\cdot,t)\| \, \|z_x(\cdot,t)\|\right) + \mathcal{O}(h),$$

where $\|\cdot\|$ denotes the $\mathcal{L}^2$-norm.
We use the algebraic equality $2\|z\| \, \|z_x\| \le \epsilon\|z_x\|^2 + \frac{1}{\epsilon}\|z\|^2$ for some value ϵ, which yields

$$v^T(T+T^T)v \le \left(-2a_{\min} + \|b-c\|_\infty \epsilon\right)\|z_x\|^2 + \|b-c\|_\infty \frac{1}{\epsilon}\|z\|^2 + \mathcal{O}(h). \tag{6.2}$$

We consider the boundary part in equation (6.1). In the case of homogeneous Dirichlet boundary conditions it is zero. Therefore we choose $\epsilon = \frac{2a_{\min}}{\|b-c\|_\infty}$ such that the norm of the derivative vanishes.

In the case of homogeneous Neumann data we have to consider the term $-c_0 v_0^2 + c_N v_N^2$ for the boundary. In the continuous case we have used the Sobolev inequality (see lemma 2.3), which is valid for piecewise continuously differentiable functions. In order to be able to use this inequality here as well, we use that $z(x_j) = v_j$ and z is piecewise continuously differentiable. Thus it holds

$$-c_0 v_0^2 + c_N v_N^2 \leq 2\|c\|_\infty \|z\|_\infty^2 \leq 2\|c\|_\infty \left(\delta \|z_x\|^2 + \left(1 + \frac{1}{\delta}\right)\|z\|^2\right) \tag{6.3}$$

Combining (6.2) and (6.3), we get

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|v\|_M^2 \leq{}& (-2a_{\min} + \epsilon\|b-c\|_\infty + 4\delta\|c\|_\infty)\|z_x\|^2 \\ &+ \left(\|b-c\|_\infty \frac{1}{\epsilon} + 4\|c\|_\infty\left(1+\frac{1}{\delta}\right)\right)\|z\|^2 + \mathcal{O}(h).\end{aligned}$$

If we choose $\epsilon = \frac{a_{\min}}{\|b-c\|_\infty}$ and $\delta = \frac{a_{\min}}{4\|c\|_\infty}$, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\|v\|_M^2 \leq \alpha_s \|z\|^2 + \mathcal{O}(h),$$

where $\alpha_s = \frac{16\|c\|_\infty^2 + \|b-c\|_\infty^2}{a_{\min}} + 4\|c\|_\infty$.
Finally we substitute the $\mathcal{L}^2$-norm of z by the M-norm of v. The norm can be written as

$$\|z\|^2 = \sum_{k,l=0}^{N} v_k v_l \int_0^1 \varphi_k(x)\varphi_l(x)\,\mathrm{d}x$$

Therefore we use the arguments form mass lumping the other way round to get

$$\begin{aligned}&\sum_{k=0}^{N} v_k v_l \int_0^1 \varphi_k(x)\varphi_l(x)\,\mathrm{d}x \\ &\quad= \begin{cases} v_l \int_0^1 \varphi_l(x)\,\mathrm{d}x + \mathcal{O}(h^3), & l \in \{0,1,2,3,N-3,N-2,N-1,N\} \\ v_l \int_0^1 \varphi_l(x)\,\mathrm{d}x + \mathcal{O}(h^5), & \text{else.}\end{cases}\end{aligned}$$

Thus we can state the following connection to the M-norm

$$\|z\|^2 = \|v\|_M^2 + \mathcal{O}(h^3).$$

Using this relation, we obtain the following estimate for Dirichlet data

$$\frac{\mathrm{d}}{\mathrm{d}t}\|v\|_M^2 \leq \frac{\|b-c\|_\infty^2}{2a_{\min}}\|v\|_M^2 + \mathcal{O}(h),$$

and for Neumann data

$$\frac{\mathrm{d}}{\mathrm{d}t}\|v\|_M^2 \leq \alpha_s \|v\|_M^2 + \mathcal{O}(h),$$

In both cases we have identical estimates as in (2.14) and (2.16) for the continuous case if we neglect higher order terms. Hence the method is strictly stable.

6.3 Operator Based on Interpolated Coefficient Functions

In the following we derive the operator for the convection diffusion equation and show that this leads to a strictly stable semi-discretization.

6.3.1 Construction

For this operator we do not use numerical integration, which is why we apply the lumped mass matrix M without the corrections that we did to account for the numerical integration. Now we discuss the calculation of the stiffness matrix.

Stiffness Matrix for $(au_x)_x$
We have to approximate the integrals $-\int_0^1 a(x,t)\varphi_j'(x)\varphi_k'(x)$ for $j,k = 0,...,N$. As proposed in remark 5.1, we define an interpolation a^h to a by $a^h(x,t) = \sum_{l=0}^N a(x_l,t)\varphi_l(x)$. If a is sufficiently smooth, it holds $a^h = a + \mathcal{O}(h^4)$, which is why the accuracy of the operator will not be affected by this approximation. Since we will only require $a^h = a + \mathcal{O}(h)$ for stability, we assume here a to be continuous and to have a piecewise continuous first derivative.
Let A be our stiffness matrix for the parabolic part $(au_x)_x$. Then the elements will be given by

$$A_{ij} = -\int_0^1 \varphi_i'(x)\varphi_j'(x)a^h(x,t)\mathrm{d}x = -\sum_{l=0}^N a(x_l,t)\int_0^1 \varphi_l(x)\varphi_i'(x)\varphi_j'(x)\mathrm{d}x.$$

Stiffness matrix for bu_x
Let B be the stiffness matrix for the part bu_x. We calculate the integrals using the Lagrange interpolation for b, i.e. $b^h(x,t) = \sum_{l=0}^N b(x_l,t)\varphi_l(x)$. Thus an element of B looks the following

$$B_{ij} = \int_\Omega b^h(x,t)\varphi_i(x)\varphi_j'(x)\,\mathrm{d}x = \sum_{l=0}^N b(x_l,t)\int_0^1 \varphi_l(x)\varphi_i(x)\varphi_j'(x)\mathrm{d}x.$$

Stiffness matrix for $(cu)_x$
Let C be the stiffness matrix for the part $(cu)_x$. Again we approximate the matrix by introducing an interpolant: $c^h = \sum_{l=0}^N c(x_l,t)\varphi_l(x)$.
Using c^h, we get

$$C_{ij} = -\int_\Omega c^h(x,t)\varphi_j(x)\varphi_i'(x)\,\mathrm{d}x = -\sum_{l=0}^N c(x_l,t)\int_\Omega \varphi_l(x)\varphi_j(x)\varphi_i'(x)\,\mathrm{d}x.$$

to be the stiffness matrix.

All occurring integrals can be evaluated exactly.

6.3.2 Order of Accuracy

By Taylor expansion it can be shown that this approximation has a local truncation error of the order h^4 in the interior and of the order h^2 at the boundary. In section 5.3.4 we have already considered the case of a stiffness matrix without error terms due to numerical integration. This has to be extended to the convection diffusion equation.

Since the convection diffusion equation is not an equation in self-adjoint form, the global accuracy proof based on the Green's function for the pure parabolic problem cannot be carried over straightforward to the convection diffusion problem. Using Sturm-Liouville theory it is, however, be possible to transform the equation into self-adjoint form (cf. [1]). It is to be examined whether the proof can be applied to the problem after such a transformation. Alternatively, it might be possible to use a proof based on the Laplace transform like in section 3.3 for the product rule. Anyway it is clear that the global error is at least of order h^3 since an energy estimate holds (cf. section 6.3.3) as shown by Gustafsson [6].

6.3.3 Stability

Like for the stability in the pure parabolic case, we use the energy method. We assume that a, b and c are continuous and piecewise continuously differentiable, so that $a^h(x,t) = a(x,t) + \mathcal{O}(h)$, $b^h(x,t) = b(x,t) + \mathcal{O}(h)$ and $c^h(x,t) = c(x,t) + \mathcal{O}(h)$.
We define an energy using the norm induced by the mass matrix $\widetilde{M}$.

We first consider **Neumann boundary conditions**.
In the $\widetilde{M}$-norm it holds

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|v\|^2_{\widetilde{M}} &= v^T(A+A^T)v + v^T(B+B^T)v + v^T(C+C^T)v + 2v^T\widetilde{m_0}\frac{\mathrm{d}}{\mathrm{d}x}g_0 + 2v^T\widetilde{m_1}\frac{\mathrm{d}}{\mathrm{d}x}g_1 \\ &\quad + v_N a_N g_1 - v_0 a_0 g_0 + 2c_N v_N^2 - 2c_0 v_0^2\end{aligned}$$

If we define $z(x,t) = \sum_{j=0}^N v_j \varphi_j(x)$, we get

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|v\|^2_{\widetilde{M}} &= -\int_0^1 a^h(x,t) z_x^2(x,t)\mathrm{d}x + \int_0^1 b^h(x,t) z_x(x,t) z(x,t)\mathrm{d}x \\ &\quad - \int_0^1 c^h(x,t) z_x(x,t) z(x,t)\mathrm{d}x + v^T\widetilde{m_0}\frac{\mathrm{d}}{\mathrm{d}x}g_0 + v^T\widetilde{m_1}\frac{\mathrm{d}}{\mathrm{d}x}g_1 \\ &\quad + v_N a_N g_1 - v_0 a_0 g_0 + c_N v_N^2 - c_0 v_0^2 \\ &= -\int_0^1 a^h(x,t) z_x^2(x,t)\mathrm{d}x + \int_0^1 (b^h(x,t) - c^h(x,t)) z_x(x,t) z(x,t)\mathrm{d}x \\ &\quad + v^T\widetilde{m_0}\frac{\mathrm{d}}{\mathrm{d}x}g_0 + v^T\widetilde{m_1}\frac{\mathrm{d}}{\mathrm{d}x}g_1 + v_N a_N g_1 - v_0 a_0 g_0 + c_N v_N^2 - c_0 v_0^2.\end{aligned}$$

Using the approximation property of a^h, b^h and c^h, we obtain

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|v\|^2_{\widetilde{M}} &\le -(a_{\min} + \mathcal{O}(h))\int_0^1 z_x^2(x,t)\mathrm{d}x + \|b-c\|_\infty \int_0^1 |z_x(x,t) z(x,t)|\mathrm{d}x \\ &\quad + v^T\widetilde{m_0}\frac{\mathrm{d}}{\mathrm{d}x}g_0 + v^T\widetilde{m_1}\frac{\mathrm{d}}{\mathrm{d}x}g_1 + v_N a_N g_1 - v_0 a_0 g_0 + c_N v_N^2 - c_0 v_0^2 \\ &\le -(a_{\min} + \mathcal{O}(h))\|z_x\|^2 + \|b-c\|_\infty \|z\|\|z_x\| + v^T\widetilde{m_0}\frac{\mathrm{d}}{\mathrm{d}x}g_0 + v^T\widetilde{m_1}\frac{\mathrm{d}}{\mathrm{d}x}g_1 \\ &\quad + v_N a_N g_1 - v_0 a_0 g_0 + c_N v_N^2 - c_0 v_0^2, \qquad (6.4)\end{aligned}$$

where $\|\cdot\|$ denotes the $\mathcal{L}^2$-norm.

Now we consider the boundary part:
For the part $a_0 v_0 g_0 + c_0 v_0$ we use $v_0 = z(x_0)$, which yields

$$-c_0 v_0^2 - a_0 v_0 g_0 \le \|a\|_\infty \|z\|_\infty |g_0| + \|c\|_\infty \|z\|_\infty^2 \le \frac{1}{2}\|a\|_\infty (\|z\|_\infty^2 + |g_0|^2) + \|c\|_\infty \|z\|_\infty^2$$

The mass matrix correction term can be estimated in the following way

$$v^T\widetilde{m}_0\frac{\mathrm{d}}{\mathrm{d}x}g_0 = \sum_{j=0}^3 \widetilde{m}_0^{(j)} v_j \frac{\mathrm{d}}{\mathrm{d}x}g_0 \le \frac{3}{5}h^2\|z\|_\infty |\frac{\mathrm{d}}{\mathrm{d}x}g_0| \le \frac{3}{10}h^2\left(\|z\|_\infty^2 + |\frac{\mathrm{d}}{\mathrm{d}x}g_0|^2\right)$$

The right boundary can be treated in the same way. If we assume the derivative of the boundary conditions to be bounded, we get the following estimate for the whole boundary part:

$$\begin{aligned}&v^T\widetilde{m_0}\frac{\mathrm{d}}{\mathrm{d}x}g_0 + v^T\widetilde{m_1}\frac{\mathrm{d}}{\mathrm{d}x}g_1 + v_N a_N g_1 - v_0 a_0 g_0 + c_N v_N^2 - c_0 v_0^2 \\ &\quad \le \frac{1}{2}\|a\|_\infty \left(g_0^2 + g_1^2\right) + (\|a\|_\infty + 2\|c\|_\infty)\|z\|_\infty^2 + \mathcal{O}(h^2)\end{aligned}$$

If we insert this estimate into (6.4), we obtain

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|v\|^2_{\widetilde{M}} \leq -(a_{\min}+\mathcal{O}(h))\|z_x\|^2 + (\|b-c\|_\infty+\mathcal{O}(h))\|z\|\|z_x\| + \frac{1}{2}\|a\|_\infty\left(g_0^2+g_1^2\right) + \gamma\|z\|^2_\infty + \mathcal{O}(h^2),$$

where $\gamma = \|a\|_\infty + 2\|c\|_\infty$.

As a next step, we estimate the $\mathcal{L}^2$-norm of z_x and the maximum-norm of z by the norm of z. For this purpose we use an algebraic inequality for the inner part and the Sobolev inequality for the boundary part, i.e.

$$2\|z\|\|z_x\| \leq \epsilon\|z_x\|^2 + \frac{1}{\epsilon}\|z\|^2 \quad \text{and} \quad \|z\|^2_\infty \leq \delta\|z_x\|^2 + \left(1+\frac{1}{\delta}\right)\|z\|^2$$

with some constants ϵ and δ. This yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\|v\|^2_{\widetilde{M}} \leq \|z_x\|^2\left(-(2a_{\min}+\mathcal{O}(h)) + (\|b-c\|_\infty+\mathcal{O}(h))\epsilon + 2\gamma\delta\right) + \|z\|^2\left(\frac{\|b-c\|_\infty+\mathcal{O}(h)}{\epsilon} + 2\left(1+\frac{1}{\delta}\right)\gamma\right) + \|a\|_\infty\left(g_0^2+g_1^2\right) + \mathcal{O}(h^2)$$

We choose $\epsilon = \frac{a_{\min}+\mathcal{O}(h)}{\|b-c\|_\infty+\mathcal{O}(h)}$ and $\delta = \frac{a_{\min}+\mathcal{O}(h)}{2\gamma}$, such that the norm of the derivative vanishes in the estimate. This leads to

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|v\|^2_{\widetilde{M}} &\leq \|z\|^2\left(\frac{\|b-c\|^2_\infty}{a_{\min}} + \gamma + \frac{4\gamma^2}{a_{\min}} + \mathcal{O}(h)\right) + \|a\|_\infty\left(g_0^2+g_1^2\right) + \mathcal{O}\left(h^2\right)\\ &= \alpha_s\|z\|^2 + \|a\|_\infty\left(g_0^2+g_1^2\right) + \mathcal{O}\left(h^2\right),\end{aligned}$$

where $\alpha_s = \frac{\|b-c\|^2_\infty}{a_{\min}} + 2\gamma + \frac{4\gamma^2}{a_{\min}} + \mathcal{O}(h)$. This is the same constant as in the continuous case, see (2.15).

Finally we convert the $\mathcal{L}^2$-norm of z to the $\widetilde{M}$-norm of v. Like in the stability proof for the other method, we get

$$\|z\|^2 = \|v\|^2_{\widetilde{M}} + \mathcal{O}(h).$$

The error is of order h in the present case since we are considering inhomogeneous boundary conditions.

This yields the following energy estimate

$$\frac{\mathrm{d}}{\mathrm{d}t}\|v\|^2_{\widetilde{M}} \leq \alpha_s\|v\|^2_{\widetilde{M}} + \|a\|_\infty\left(g_0^2+g_1^2\right) + \mathcal{O}(h),$$

which is identical to the analytic estimate if we neglect higher order terms. Hence the operator is strictly stable.

In the case of **Dirichlet** boundary conditions we consider homogeneous boundary conditions to prove stability like for well-posedness in the continuous case.

We get a similar estimate as with Neumann data except that we have no boundary terms and we put the first and last line of each matrix to zero. The latter modification has however no effects since we have homogeneous boundary data which means that the boundary part will be set to zero anyway, if we consider the product of the vector v with the stiffness matrix. The same transformations for the inner part as above yield

$$\frac{\mathrm{d}}{\mathrm{d}t}\|v\|^2_{\widetilde{M}} \leq \left(-(a_{\min}+\mathcal{O}(h)) + (\|b-c\|_\infty+\mathcal{O}(h))\frac{\epsilon}{2}\right)\|z_x\|^2 + \left(\frac{\|b-c\|_\infty+\mathcal{O}(h)}{2\epsilon}\right)\|z\|^2,$$

where we choose the constant $\epsilon = \frac{2a_{\min}+\mathcal{O}(h)}{\|b-c\|_\infty+\mathcal{O}(h)}$ in this case. Thus we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\|v\|^2_{\widetilde{M}} \leq \left(\frac{\|b-c\|^2_\infty}{4a_{\min}} + \mathcal{O}(h)\right)\|z\|^2,$$

Again we can transfer the $\mathcal{L}^2$-norm to the $\widetilde{M}$-norm, which yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\|v\|^2_{\widetilde{M}} \leq \alpha_s\|v\|^2_{\widetilde{M}} + \mathcal{O}(h^2),$$

with $\alpha_s = \frac{\|b-c\|^2_\infty}{4a_{\min}} + \mathcal{O}(h)$. Since this is identical to the analytic estimate in (2.14) if we neglect higher order terms, the operator for homogeneous boundary conditions is strictly stable as well.

6.4 Computational Results

We give here one example for the performance of the operator, where the stiffness matrix is calculated using an interpolation of the coefficient functions. Let us consider the following problem

$$\begin{aligned}
&u_t(x,t) = (a(x)u_x(x,t))_x + b(x)u_x(x,t) \quad \text{with}\\
&a(x) = \frac{1}{10}\left(1+\frac{1}{2}\sin\left(\frac{\pi}{2}x\right)\right) \quad \text{and} \quad b(x) = 2\sinh\left(-3x+\frac{3}{2}\right).\\
&f(x) = \sin(\pi x) + \frac{7}{5}\cos(10x)\\
&\text{Neumann: } u_x(0,t) = g_0(t) \approx 3.14\,\mathrm{e}^{-0.37t} \quad \text{and} \quad u_x(1,t) = g_1(t) \approx 4.74\,\mathrm{e}^{-1.43t}\\
&\text{Dirichlet: } u(0,t) = g_0(t) \approx 1.40\,\mathrm{e}^{-0.37t} \quad \text{and} \quad u(1,t) = g_1(t) \approx -11.8\,\mathrm{e}^{-1.43t}.
\end{aligned}$$

This problem has already been treated with the product rule based operator.
The numerical results are given in table 18. We use a time step of $2h^2$, since the Fourier transform of this operator is rather similar to the one where we used numerical integration and $a_{\min} = 0.15$ in our example.

	Neumann		Dirichlet	
N	$\|u-v\|_h$	q_N	$\|u-v\|_h$	q_D
10	$7.42\cdot10^{-3}$		$2.72\cdot10^{-1}$	
20	$8.90\cdot10^{-4}$	3.06	$7.74\cdot10^{-2}$	1.81
40	$1.06\cdot10^{-4}$	3.07	$1.53\cdot10^{-2}$	2.33
80	$8.83\cdot10^{-6}$	3.58	$1.83\cdot10^{-3}$	3.07
160	$5.97\cdot10^{-7}$	3.89	$1.48\cdot10^{-4}$	3.62
320	$4.20\cdot10^{-8}$	3.83	$9.76\cdot10^{-6}$	3.92

Table 18: Numerical results for the convection diffusion equation ($c \equiv 0$) using Dirichlet and Neumann conditions

In the experiment we observe that the solution converges with a rate of 4.

7 Concluding Remarks

We have discussed three different approaches to discretize the parabolic term $(au_x)_x$ appearing in the convection diffusion equation. To round off our work, we compare some properties of the three operators. For this purpose we consider efficiency, stability and damping properties. We finish our investigation with an outlook on open questions that could be considered in future work.

7.1 Comparison of the Operators

An efficient method should be designed such that the number of floating point operations to obtain a certain accuracy is minimal. Therefore we want to minimize the number of function evaluations of the diffusion coefficient a and the band width of the stencil to approximate the derivatives. The latter property also simplifies a parallel implementation of the operator.

Concerning efficiency, the operator based on the product rule is the best choice. Both the operator for the first and second derivative used in this approximation have minimal band width. The operator is especially efficient if the diffusion term is not time-dependent. In that case an approximation of the derivative of the diffusion coefficient can be calculated in advance and therefore the work for approximating the convection diffusion equation is only slightly increased when going from constant to variable coefficients. For the two other methods, the band width is not that small. The fourth order method derived by the product rule has a band width of 5 (the main diagonal and two subdiagonals on each side), while the operator for the self-adjoint form and the one based on finite elements both have a band width of 7. Moreover, the approximation of a is about twice as costly for the self-adjoint operators.

However, the local efficiency of the operator based on the product rule has a drawback. In contrast to the other two operators, it is not strictly stable. This means that the growth of the numerical solution is not directly linked to the growth of the analytical solution of the partial differential equation. Even though the Lax-Richtmyer equivalence theorem guarantees the convergence of a consistent approximation, roundoff errors could grow exponentially in time. This makes the error estimate in terms of $\mathrm{e}^{\alpha_s t}\mathcal{O}(h^{p+2})$ (where α_s is the growth rate of the semi-discretization and p the order of accuracy of the operator at the boundary) useless for practical purposes, where h cannot be chosen as small as necessary. However, in our experiments such problems did not occur.

Both the other ideas rely on the self-adjoint form, which is why the approximations are strictly stable. This is the advantage of these operator. Strict stability holds even if the coefficients do not fit the regularity requirements needed for stability of the operator derived from the product rule. A self-adjoint operator designed according to the suggestions in section 4 would have all wished properties, but it is unsure whether the system for the weights of the boundary section can be solved.

The last ansatz is based on a different theory, namely finite elements, which is why it cannot be combined with the known SBP operators as the other two methods. Nevertheless it is possible to apply the operator to the convection diffusion equation by designing an operator for the first derivative that is also based on the finite element ansatz.

As a third feature we require good damping properties. The natural way for approximating $(au_x)_x$ would be to apply a known SBP operator for the first derivative twice. Besides efficiency lacks, such an operator shows clear disadvantages in damping high frequency error modes. In figure 4 it can be seen that this operator has a very poor damping for high wave numbers.

This effect can be seen in numerical results where high frequency errors represent the

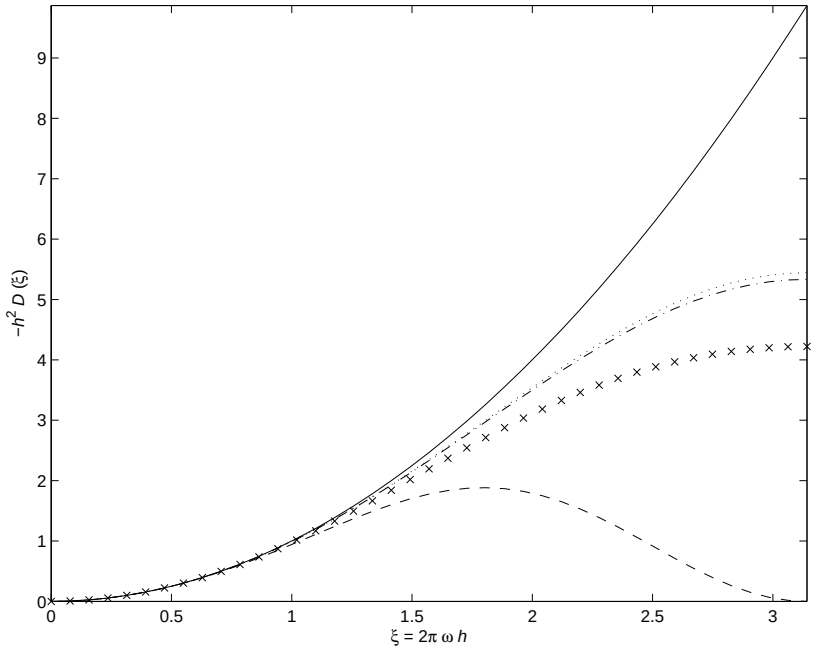

Figure 4: Damping of modes with wave number ξ of the exact solution (-), D_1 applied twice (- -), the operator D_2 as base for the product rule (-·), the operator Q derived from the self-adjoint form (·) and the operator L based on finite elements with mass lumping (x). All operators are fourth order accurate.

main error source. This problem was one of the motivations to devise new operators for this problem. As seen in figure 4, all operators proposed in this work overcome this problem. The damping is however not as strong as provided by the exact differential operator, but it is the better the higher the order of the method.

7.2 Outlook

When we consider the stability of the two methods based on finite differences, we get a stronger requirement on the boundary data than for the continuous problem. This is because we cannot apply the Sobolev inequality onto the semi-discretization. However, in the experiments we did not observe stability problems when violating the additional stability conditions. Therefore it would be interesting to prove a discrete analogon of the Sobolev inequality.

For the operator derived from the self-adjoint form we have only done a theoretical analysis of the boundary treatment for higher order methods. However, we did not succeed in calculating the weights for the boundary part of the operator except for the order two.

The finite element ansatz has only been investigated for fourth order accuracy in this work. It would be interesting to use higher order polynomials as basis functions in order to get sixth or eighth order accurate operators. The question is how feasible the idea of mass lumping is in these cases. Furthermore, the extension of this operator to the convection diffusion equation requires a more thorough investigation, especially concerning accuracy. Furthermore, the accuracy proof using the Green's function might also be applicable for showing the order of accuracy of finite difference methods.

8 Summary in Swedish

Differensapproximationer av högre ordningen för paraboliska och hyperbolisk-paraboliska problem med variabla koefficienter

Denna rapport handlar om den numeriska lösningen av tidsberoende partiella differentialekvationer med variabla koefficienter. Ett exempel på en hyperbolisk-parabolisk ekvation är den linearisерade Navier-Stokes-ekvationen som styr flöden av gaser och vätskor. Ekvationen dyker till exempel upp inom aerodynamiken och i simuleringen av förbränningsprocesser. Huvuddelen av arbetet är undersökningen av en parabolisk diffusionsterm konstruerad av den första derivatan av produkten av koefficienten och första derivatan av lösningen. Denna form kallas själv-adjungerad. För att karakterisera den analytiska lösningen definerar man en energi som en integral över kvadraten på lösningen. Den själv-adjungerade formen ger upphov till att diffusionstermens bidrag till lösningens energi minskar i tiden, som beror på möjligheten att använda partiell integration på den diffusiva termen.

I tidigare arbete av Strand och Mattsson/Nordström härleddes högre ordningens finita differens-metoder för hyperboliska ekvationer och för paraboliska termer med konstanta koefficienter. Dessa metoder är strikt stabila, dvs den diskreta operatorn förhåller sig som den analytiska derivatan vad gäller tillväxt och dämpning. Stabilitet byggs på en partiell summationsregel (som ersätter partiell integration för diskretiseringen), som operatorerna uppfyller i en viss norm. Stabilitet i det diskreta fallet kan betraktas som den diskreta versionen till en energiuppskattning. Sådana metoder kallas för SBP (“Summation By Parts”)-metoder. Målet med detta arbete är att hitta differensmetoder som är effektiva och imiterar de fysikaliska egenskaperna så bra som möjligt även för variabla koefficienter.

Vi undersökte tre olika ansatser för att uppnå målet. I första ansatsen diskretiserar vi koefficienterna och lösningen var för sig genom att använda produktregeln på den diffusiva termen. På så sätt kan man använda samma operatorer som för problemet med konstanta koefficienter. Denna metod är mycket effektiv, men operatorerna uppfyller inte längre en partiell summationsregel eftersom vi inte utnyttjar den själv-adjungerade formen och inte får en energiuppskattning som liknar den kontinuerliga. Detta kan leda till stabiltetsproblem fast våra experiment inte visade sådana.

Vi undersökte också möjligheten att konstruera en metod som utnyttjar den själv-adjungerade formen. Detta ledde till ett icke-linjärt ekvationssystem för randoperatorn, vilket vi inte lyckades lösa för metoder av högre ordning än två.

Den tredje ansatsen är att använda teorin från finita element-metoden, dvs vi interpolerar lösningen med polynom och integrerar differentialekvationen för den approximativa lösningen. På så sätt kan man utnyttja den själv-adjungerade formen genom att använda partiell integration på integralerna. Om man beräknar approximationernas integraler, får man ett ordinärt differentialekvationssystem i tiden. Men systemet blir implicit för denna ansats, vilket vi undviker genom att reducera det till en explicit form. Denna process kallas för “mass-lumping”. Det explicita systemet kan tolkas som en differensmetod. Metoden är stabil, men stabiliteten bygger inte på samma norm som de tidigare operatorerna. Därför kan man inte kombinera metoderna för att lösa en hyperbolisk-parabolisk ekvation. För att kunna använda denna metod krävs att man härleder en diskretisering av den hyperboliska termen med samma metod som vi också gjorde. För både den paraboliska och den hyperbolisk-paraboliska ekvationen härledde vi med hälp av finita elementmetoden 4:e ordningens finita differensmetoder.

Acknowledgements

We would like to thank our supervisor Bernhard Müller for guidance and feedback during our work. We also want to thank Jan Nordström for commenting our work and Bertil Gustafsson for his course on higher order finite difference methods.
Finally our thank goes to all employees at Uppsala Universitet that made our instructive stay in Sweden possible.

A Finite elements with Numerical Integration

We now present the the difference scheme $L(\cdot) = M^{-1}S(\cdot) + m_0\frac{\mathrm{d}}{\mathrm{d}t}g_0 + m_1\frac{\mathrm{d}}{\mathrm{d}t}g_1$ developed in section 5.

Stiffness matrix:
Left boundary part:

$$
\begin{aligned}
h\cdot S_{0,0} &= -\frac{121}{216}a_0 - \frac{529}{24}a_{1/2} - \frac{1}{864}a_1 - \frac{1}{864}a_{3/2} - \frac{1}{216}a_2\\
h\cdot S_{0,1} &= \frac{11}{12}a_0 + \frac{161}{288}a_{1/2} - \frac{1}{18}a_1 + \frac{1}{32}a_{3/2} + \frac{1}{36}a_2\\
h\cdot S_{0,2} &= -\frac{11}{24}a_0 + \frac{23}{288}a_{1/2} + \frac{1}{9}a_1 - \frac{1}{32}a_{3/2} - \frac{1}{72}a_2\\
h\cdot S_{0,3} &= \frac{11}{108}a_0 - \frac{23}{864}a_{1/2} - \frac{1}{54}a_1 + \frac{1}{864}a_{3/2} - \frac{1}{108}a_2
\end{aligned}
$$

$$
\begin{aligned}
h\cdot S_{1,0} &= \frac{11}{12}a_0 + \frac{161}{288}a_{1/2} - \frac{1}{18}a_1 + \frac{1}{32}a_{3/2} + \frac{1}{36}a_2\\
h\cdot S_{1,1} &= -\frac{3}{2}a_0 - \frac{49}{96}a_{1/2} - \frac{1}{12}a_1 - \frac{27}{32}a_{3/2} - \frac{5}{27}a_2 - \frac{1}{864}a_{5/2} - \frac{1}{216}a_3\\
h\cdot S_{1,2} &= \frac{3}{4}a_0 - \frac{7}{96}a_{1/2} + \frac{1}{6}a_1 + \frac{27}{32}a_{3/2} + \frac{1}{18}a_2 + \frac{3}{32}a_{5/2} + \frac{1}{36}a_3\\
h\cdot S_{1,3} &= -\frac{1}{6}a_0 + \frac{7}{288}a_{1/2} - \frac{1}{36}a_1 - \frac{1}{32}a_{3/2} + \frac{1}{9}a_2 - \frac{3}{32}a_{5/2} - \frac{1}{72}a_3\\
h\cdot S_{1,4} &= -\frac{1}{108}a_2 + \frac{1}{864}a_{5/2} - \frac{1}{108}a_3
\end{aligned}
$$

$$
\begin{aligned}
h\cdot S_{2,0} &= -\frac{11}{24}a_0 + \frac{23}{288}a_{1/2} + \frac{1}{9}a_1 - \frac{1}{32}a_{3/2} - \frac{1}{72}a_2\\
h\cdot S_{2,1} &= \frac{3}{4}a_0 - \frac{7}{96}a_{1/2} + \frac{1}{6}a_1 + \frac{27}{32}a_{3/2} + \frac{1}{18}a_2 + \frac{3}{32}a_{5/2} + \frac{1}{36}a_3\\
h\cdot S_{2,2} &= -\frac{3}{8}a_0 - \frac{1}{96}a_{1/2} - \frac{1}{3}a_1 - \frac{27}{32}a_{3/2} - \frac{1}{12}a_2 - \frac{27}{32}a_{5/2} - \frac{1}{54}a_3 - \frac{1}{864}a_{7/2} - \frac{1}{216}a_4\\
h\cdot S_{2,3} &= \frac{1}{12}a_0 + \frac{1}{288}a_{1/2} + \frac{1}{18}a_1 + \frac{1}{32}a_{3/2} + \frac{1}{18}a_2 + \frac{27}{32}a_{5/2} + \frac{1}{18}a_3 + \frac{1}{32}a_{7/2} + \frac{1}{36}a_4\\
h\cdot S_{2,4} &= -\frac{1}{72}a_2 - \frac{1}{32}a_{5/2} + \frac{1}{9}a_3 - \frac{1}{32}a_{7/2} - \frac{1}{72}a_4\\
h\cdot S_{2,5} &= -\frac{1}{108}a_3 + \frac{1}{864}a_{7/2} - \frac{1}{108}a_4
\end{aligned}
$$

$$
\begin{aligned}
h\cdot S_{3,0} &= \frac{11}{108}a_0 - \frac{23}{864}a_{1/2} - \frac{1}{54}a_1 + \frac{1}{864}a_{3/2} - \frac{1}{108}a_2\\
h\cdot S_{3,1} &= -\frac{1}{6}a_0 + \frac{7}{288}a_{1/2} - \frac{1}{36}a_1 - \frac{1}{32}a_{3/2} + \frac{1}{9}a_2 - \frac{3}{32}a_{5/2} - \frac{1}{72}a_3\\
h\cdot S_{3,2} &= \frac{1}{12}a_0 + \frac{1}{288}a_{1/2} + \frac{1}{18}a_1 + \frac{1}{32}a_{3/2} + \frac{1}{18}a_2 + \frac{27}{32}a_{5/2} + \frac{1}{18}a_3 + \frac{1}{32}a_{7/2} + \frac{1}{36}a_4\\
h\cdot S_{3,3} &= -\frac{1}{54}a_0 - \frac{1}{864}a_{1/2} - \frac{1}{108}a_1 - \frac{1}{864}a_{3/2} - \frac{5}{27}a_2 - \frac{27}{32}a_{5/2} - \frac{1}{12}a_3\\
&\quad - \frac{27}{32}a_{7/2} - \frac{5}{27}a_4 - \frac{1}{864}a_{9/2} - \frac{1}{216}a_5\\
h\cdot S_{3,4} &= \frac{1}{36}a_2 + \frac{1}{32}a_{5/2} + \frac{1}{18}a_3 + \frac{27}{32}a_{7/2} + \frac{1}{18}a_4 + \frac{1}{32}a_{9/2} + \frac{1}{36}a_5\\
h\cdot S_{3,5} &= -\frac{1}{72}a_3 - \frac{1}{32}a_{7/2} + \frac{1}{9}a_4 - \frac{1}{32}a_{9/2} - \frac{1}{72}a_5\\
h\cdot S_{3,6} &= -\frac{1}{108}a_4 + \frac{1}{864}a_{9/2} - \frac{1}{108}a_5
\end{aligned}
$$

The inner scheme ($l = 4, 5, \ldots, N-4$) is given by

$$
\begin{aligned}
h \cdot S_{l,l-3} &= -\frac{1}{108}a_{l-2} + \frac{1}{864}a_{l-3/2} - \frac{1}{108}a_{l-1} \\
h \cdot S_{l,l-2} &= -\frac{1}{72}a_{l-2} - \frac{1}{32}a_{l-3/2} + \frac{1}{9}a_{l-1} - \frac{1}{32}a_{l-1/2} - \frac{1}{72}a_{l} \\
h \cdot S_{l,l-1} &= \frac{1}{36}a_{l-2} + \frac{1}{32}a_{l-3/2} + \frac{1}{18}a_{l-1} + \frac{27}{32}a_{l-1/2} + \frac{1}{18}a_{l} + \frac{1}{32}a_{l+1/2} + \frac{1}{36}a_{l+1} \\
h \cdot S_{l,l} &= -\frac{1}{216}a_{l-2} - \frac{1}{864}a_{l-3/2} - \frac{5}{27}a_{l-1} - \frac{27}{32}a_{l-1/2} - \frac{1}{12}a_{l} \\
&\quad - \frac{27}{32}a_{l+1/2} - \frac{5}{27}a_{l+1} - \frac{1}{864}a_{l+3/2} - \frac{1}{216}a_{l+2} \\
h \cdot S_{l,l+1} &= \frac{1}{36}a_{l-1} + \frac{1}{32}a_{l-1/2} + \frac{1}{18}a_{l} + \frac{27}{32}a_{l+1/2} + \frac{1}{18}a_{l+1} + \frac{1}{32}a_{l+3/2} + \frac{1}{36}a_{l+2} \\
h \cdot S_{l,l+2} &= -\frac{1}{72}a_{l} - \frac{1}{32}a_{l+1/2} + \frac{1}{9}a_{l+1} - \frac{1}{32}a_{l+3/2} - \frac{1}{72}a_{l+2} \\
h \cdot S_{l,l+3} &= -\frac{1}{108}a_{l+1} + \frac{1}{864}a_{l+3/2} - \frac{1}{108}a_{l+2}
\end{aligned}
$$

Right boundary part:

$$
\begin{aligned}
h \cdot S_{N,N} &= -\frac{121}{216}a_{N} - \frac{529}{24}a_{N-1/2} - \frac{1}{864}a_{N-1} - \frac{1}{864}a_{N-3/2} - \frac{1}{216}a_{N-2} \\
h \cdot S_{N,N-1} &= \frac{11}{12}a_{N} + \frac{161}{288}a_{N-1/2} - \frac{1}{18}a_{N-1} + \frac{1}{32}a_{N-3/2} + \frac{1}{36}a_{N-2} \\
h \cdot S_{N,N-2} &= -\frac{11}{24}a_{N} + \frac{23}{288}a_{N-1/2} + \frac{1}{9}a_{N-1} - \frac{1}{32}a_{N-3/2} - \frac{1}{72}a_{N-2} \\
h \cdot S_{N,N-3} &= \frac{11}{108}a_{N} - \frac{23}{864}a_{N-1/2} - \frac{1}{54}a_{N-1} + \frac{1}{864}a_{N-3/2} - \frac{1}{108}a_{N-2}
\end{aligned}
$$

$$
\begin{aligned}
h \cdot S_{N-1,N} &= \frac{11}{12}a_{N} + \frac{161}{288}a_{N-1/2} - \frac{1}{18}a_{N-1} + \frac{1}{32}a_{N-3/2} + \frac{1}{36}a_{N-2} \\
h \cdot S_{N-1,N-1} &= -\frac{3}{2}a_{N} - \frac{49}{96}a_{N-1/2} - \frac{1}{12}a_{N-1} - \frac{27}{32}a_{N-3/2} - \frac{5}{27}a_{N-2} \\
&\quad - \frac{1}{864}a_{N-5/2} - \frac{1}{216}a_{N-3} \\
h \cdot S_{N-1,N-2} &= \frac{3}{4}a_{N} - \frac{7}{96}a_{N-1/2} + \frac{1}{6}a_{N-1} + \frac{27}{32}a_{N-3/2} + \frac{1}{18}a_{N-2} + \frac{3}{32}a_{N-5/2} + \frac{1}{36}a_{N-3} \\
h \cdot S_{N-1,N-3} &= -\frac{1}{6}a_{N} + \frac{7}{288}a_{N-1/2} - \frac{1}{36}a_{N-1} - \frac{1}{32}a_{N-3/2} + \frac{1}{9}a_{N-2} - \frac{3}{32}a_{N-5/2} - \frac{1}{72}a_{N-3} \\
h \cdot S_{N-1,N-4} &= -\frac{1}{108}a_{N-2} + \frac{1}{864}a_{N-5/2} - \frac{1}{108}a_{N-3}
\end{aligned}
$$

$$
\begin{aligned}
h \cdot S_{N-2,N} &= -\frac{11}{24}a_{N} + \frac{23}{288}a_{N-1/2} + \frac{1}{9}a_{N-1} - \frac{1}{32}a_{N-3/2} - \frac{1}{72}a_{N-2} \\
h \cdot S_{N-2,N-1} &= \frac{3}{4}a_{N} - \frac{7}{96}a_{N-1/2} + \frac{1}{6}a_{N-1} + \frac{27}{32}a_{N-3/2} + \frac{1}{18}a_{N-2} + \frac{3}{32}a_{N-5/2} + \frac{1}{36}a_{N-3} \\
h \cdot S_{N-2,N-2} &= -\frac{3}{8}a_{N} - \frac{1}{96}a_{N-1/2} - \frac{1}{3}a_{N-1} - \frac{27}{32}a_{N-3/2} - \frac{1}{12}a_{N-2} - \frac{27}{32}a_{N-5/2} \\
&\quad - \frac{1}{54}a_{N-3} - \frac{1}{864}a_{N-7/2} - \frac{1}{216}a_{N-4} \\
h \cdot S_{N-2,N-3} &= \frac{1}{12}a_{N} + \frac{1}{288}a_{N-1/2} + \frac{1}{18}a_{N-1} + \frac{1}{32}a_{N-3/2} + \frac{1}{18}a_{N-2} + \frac{27}{32}a_{N-5/2} \\
&\quad + \frac{1}{18}a_{N-3} + \frac{1}{32}a_{N-7/2} + \frac{1}{36}a_{N-4} \\
h \cdot S_{N-2,N-4} &= -\frac{1}{72}a_{N-2} - \frac{1}{32}a_{N-5/2} + \frac{1}{9}a_{N-3} - \frac{1}{32}a_{N-7/2} - \frac{1}{72}a_{N-4} \\
h \cdot S_{N-2,N-5} &= -\frac{1}{108}a_{N-3} + \frac{1}{864}a_{N-7/2} - \frac{1}{108}a_{N-4}
\end{aligned}
$$

$$
\begin{aligned}
h \cdot S_{N-3,N} &= \frac{11}{108}a_N - \frac{23}{864}a_{N-1/2} - \frac{1}{54}a_{N-1} + \frac{1}{864}a_{N-3/2} - \frac{1}{108}a_{N-2} \\
h \cdot S_{N-3,N-1} &= -\frac{1}{6}a_N + \frac{7}{288}a_{N-1/2} - \frac{1}{36}a_{N-1} - \frac{1}{32}a_{N-3/2} + \frac{1}{9}a_{N-2} - \frac{3}{32}a_{N-5/2} - \frac{1}{72}a_{N-3} \\
h \cdot S_{N-3,N-2} &= \frac{1}{12}a_N + \frac{1}{288}a_{N-1/2} + \frac{1}{18}a_{N-1} + \frac{1}{32}a_{N-3/2} + \frac{1}{18}a_{N-2} + \frac{27}{32}a_{N-5/2} \\
&\quad + \frac{1}{18}a_{N-3} + \frac{1}{32}a_{N-7/2} + \frac{1}{36}a_{N-4} \\
h \cdot S_{N-3,N-3} &= -\frac{1}{54}a_N - \frac{1}{864}a_{N-1/2} - \frac{1}{108}a_{N-1} - \frac{1}{864}a_{N-3/2} - \frac{5}{27}a_{N-2} - \frac{27}{32}a_{N-5/2} \\
&\quad - \frac{1}{12}a_{N-3} - \frac{27}{32}a_{N-7/2} - \frac{5}{27}a_{N-4} - \frac{1}{864}a_{N-9/2} - \frac{1}{216}a_{N-5} \\
h \cdot S_{N-3,N-4} &= \frac{1}{36}a_{N-2} + \frac{1}{32}a_{N-5/2} + \frac{1}{18}a_{N-3} + \frac{27}{32}a_{N-7/2} + \frac{1}{18}a_{N-4} + \frac{1}{32}a_{N-9/2} + \frac{1}{36}a_{N-5} \\
h \cdot S_{N-3,N-5} &= -\frac{1}{72}a_{N-3} - \frac{1}{32}a_{N-7/2} + \frac{1}{9}a_{N-4} - \frac{1}{32}a_{N-9/2} - \frac{1}{72}a_{N-5} \\
h \cdot S_{N-3,N-6} &= -\frac{1}{108}a_{N-4} + \frac{1}{864}a_{N-9/2} - \frac{1}{108}a_{N-5}
\end{aligned}
$$

The mass matrix is depending on the kind of physical boundary data:
For **Neumann data** we have the following mass matrix:

$$
M = h \begin{pmatrix}
\frac{1}{3} & & & & & & & & & & \\
& \frac{31}{24} & & & & & & & & & \\
& & \frac{5}{6} & & & & & & & & \\
& & & \frac{25}{24} & & & & & & & \\
& & & & 1 & & & & & & \\
& & & & & \ddots & & & & & \\
& & & & & & 1 & & & & \\
& & & & & & & \frac{25}{24} & & & \\
& & & & & & & & \frac{5}{6} & & \\
& & & & & & & & & \frac{31}{24} & \\
& & & & & & & & & & \frac{1}{3}
\end{pmatrix}
$$

Boundary correction of the mass matrix:

$$
\begin{aligned}
m_0 &= h^2 \left(-\frac{7}{144}, \frac{59}{288}, -\frac{19}{72}, \frac{31}{288}, 0, ... \right) \\
m_1 &= h^2 \left(0, ..., 0, -\frac{31}{288}, \frac{19}{72}, -\frac{59}{288}, \frac{7}{144} \right)
\end{aligned}
$$

The mass matrix for **Dirichlet data** is given by

$$
M = h \begin{pmatrix}
\frac{313}{288} & & & & & & & & & & \\
& \frac{139}{144} & & & & & & & & & \\
& & \frac{869}{864} & & & & & & & & \\
& & & 1 & & & & & & & \\
& & & & \ddots & & & & & & \\
& & & & & 1 & & & & & \\
& & & & & & \frac{869}{864} & & & & \\
& & & & & & & \frac{139}{144} & & & \\
& & & & & & & & \frac{313}{288} & &
\end{pmatrix}
$$

Boundary correction of the mass matrix:

$$m_0 = h\left(\frac{59}{288}, -\frac{19}{144}, \frac{31}{864}, 0, ...\right)$$
$$m_1 = h\left(0, ..., 0, \frac{31}{864}, -\frac{19}{144}, \frac{59}{288}\right)$$

B Finite Elements with Approximated Basis Functions

The mass matrix is depending on the kind of physical boundary data for this ansatz as well:
Neumann data
Mass matrix:

$$\widetilde{M} = h \begin{pmatrix} \frac{1}{3} & & & & & & & & & & \\ & \frac{31}{24} & & & & & & & & & \\ & & \frac{5}{6} & & & & & & & & \\ & & & \frac{25}{24} & & & & & & & \\ & & & & 1 & & & & & & \\ & & & & & \ddots & & & & & \\ & & & & & & 1 & & & & \\ & & & & & & & \frac{25}{24} & & & \\ & & & & & & & & \frac{5}{6} & & \\ & & & & & & & & & \frac{31}{24} & \\ & & & & & & & & & & \frac{1}{3} \end{pmatrix}$$

Boundary correction of the mass matrix:

$$\widetilde{m}_0 = h^2\left(-\frac{2}{45}, \frac{7}{36}, -\frac{23}{90}, \frac{19}{180}, 0, ...\right)$$
$$\widetilde{m}_1 = h^2\left(0, ..., 0, -\frac{19}{180}, \frac{23}{90}, -\frac{7}{36}, \frac{2}{45}\right)$$

Dirichlet data

$$\widetilde{M} = h \begin{pmatrix} \frac{79}{72} & & & & & & & & & & \\ & \frac{173}{180} & & & & & & & & & \\ & & \frac{1087}{1080} & & & & & & & & \\ & & & 1 & & & & & & & \\ & & & & \ddots & & & & & & \\ & & & & & 1 & & & & & \\ & & & & & & \frac{1087}{1080} & & & & \\ & & & & & & & \frac{173}{180} & & & \\ & & & & & & & & \frac{79}{72} & & \end{pmatrix}$$

Boundary correction of the mass matrix:

$$\widetilde{m}_0 = h\left(\frac{7}{36}, -\frac{23}{180}, \frac{19}{540}, 0, ...\right)$$
$$\widetilde{m}_1 = h\left(0, ..., 0, \frac{19}{540}, -\frac{23}{180}, \frac{7}{36}\right)$$

Stiffness matrix for the parabolic term $(au_x)_x$:

Left boundary part:

$$
\begin{aligned}
h \cdot A_{0,0} &= -\frac{77}{108}a_0 - \frac{641}{945}a_1 + \frac{173}{756}a_2 - \frac{46}{945}a_3 \\
h \cdot A_{0,1} &= \frac{947}{945}a_0 + \frac{22}{35}a_1 - \frac{23}{105}a_2 + \frac{52}{945}a_3 \\
h \cdot A_{0,2} &= -\frac{1381}{3780}a_0 + \frac{1}{15}a_1 + \frac{1}{140}a_2 - \frac{8}{945}a_3 \\
h \cdot A_{0,3} &= \frac{8}{105}a_0 - \frac{16}{945}a_1 - \frac{16}{945}a_2 + \frac{2}{945}a_3
\end{aligned}
$$

$$
\begin{aligned}
h \cdot A_{1,0} &= \frac{947}{945}a_0 + \frac{22}{35}a_1 - \frac{23}{105}a_2 + \frac{52}{945}a_3 \\
h \cdot A_{1,1} &= -\frac{89}{63}a_0 - \frac{2303}{1728}a_1 - \frac{18733}{60480}a_2 - \frac{2071}{60480}a_3 + \frac{11}{20160}a_4 \\
h \cdot A_{1,2} &= \frac{167}{315}a_0 + \frac{47191}{60480}a_1 + \frac{569}{1344}a_2 + \frac{107}{4032}a_3 - \frac{121}{60480}a_4 \\
h \cdot A_{1,3} &= -\frac{113}{945}a_0 - \frac{4631}{60480}a_1 + \frac{37}{320}a_2 - \frac{449}{12096}a_3 + \frac{59}{60480}a_4 \\
h \cdot A_{1,4} &= \frac{29}{60480}(a_1 + a_4) - \frac{617}{60480}(a_2 + a_3)
\end{aligned}
$$

$$
\begin{aligned}
h \cdot A_{2,0} &= -\frac{1381}{3780}a_0 + \frac{1}{15}a_1 + \frac{1}{140}a_2 - \frac{8}{945}a_3 \\
h \cdot A_{2,1} &= \frac{167}{315}a_0 + \frac{47191}{60480}a_1 + \frac{569}{1344}a_2 + \frac{107}{4032}a_3 - \frac{121}{60480}a_4 \\
h \cdot A_{2,2} &= -\frac{277}{1260}a_0 - \frac{2053}{2240}a_1 - \frac{389}{432}a_2 - \frac{3781}{6048}a_3 + \frac{593}{15120}a_4 + \frac{11}{20160}a_5 \\
h \cdot A_{2,3} &= \frac{52}{945}a_0 + \frac{461}{6720}a_1 + \frac{15203}{30240}a_2 + \frac{7853}{15120}a_3 + \frac{1}{160}a_4 - \frac{121}{60480}a_5 \\
h \cdot A_{2,4} &= \frac{59}{60480}(a_1 + a_5) - \frac{503}{15120}(a_2 + a_4) + \frac{47}{480}a_3 \\
h \cdot A_{2,5} &= \frac{29}{60480}(a_2 + a_5) - \frac{617}{60480}(a_3 + a_4)
\end{aligned}
$$

$$
\begin{aligned}
h \cdot A_{3,0} &= \frac{8}{105}a_0 - \frac{16}{945}a_1 - \frac{16}{945}a_2 + \frac{2}{945}a_3 \\
h \cdot A_{3,1} &= -\frac{113}{945}a_0 - \frac{4631}{60480}a_1 + \frac{37}{320}a_2 - \frac{449}{12096}a_3 + \frac{59}{60480}a_4 \\
h \cdot A_{3,2} &= \frac{52}{945}a_0 + \frac{461}{6720}a_1 + \frac{15203}{30240}a_2 + \frac{7853}{15120}a_3 + \frac{1}{160}a_4 - \frac{121}{60480}a_5 \\
h \cdot A_{3,3} &= -\frac{11}{945}a_0 + \frac{1627}{60480}a_1 - \frac{9203}{15120}a_2 - \frac{1673}{1728}a_3 - \frac{7397}{12096}a_4 + \frac{593}{15120}a_5 + \frac{11}{20160}a_6 \\
h \cdot A_{3,4} &= -\frac{121}{60480}(a_1 + a_6) + \frac{1}{160}(a_2 + a_5) + \frac{31243}{60480}(a_3 + a_4) \\
h \cdot A_{3,5} &= \frac{59}{60480}(a_2 + a_6) - \frac{503}{15120}(a_3 + a_5) + \frac{47}{480}a_4 \\
h \cdot A_{3,6} &= \frac{29}{60480}(a_3 + a_6) - \frac{617}{60480}(a_4 + a_5)
\end{aligned}
$$

The inner scheme ($l = 4, 5, \ldots, N-4$) is given by

$$
\begin{aligned}
h \cdot A_{l,l-3} &= \frac{29}{60480}\,(a_{l-3} + a_l) - \frac{617}{60480}\,(a_{l-2} + a_{l-1}) \\
h \cdot A_{l,l-2} &= \frac{59}{60480}\,(a_{l-3} + a_{l+1}) - \frac{503}{15120}\,(a_{l-2} + a_l) + \frac{47}{480}a_{l-1} \\
h \cdot A_{l,l-1} &= -\frac{121}{60480}\,(a_{l-3} + a_{l+2}) + \frac{1}{160}\,(a_{l-2} + a_{l+1}) + \frac{31243}{60480}\,(a_{l-1} + a_l) \\
h \cdot A_{l,l} &= \frac{11}{20160}\,(a_{l-3} + a_{l+3}) + \frac{593}{15120}\,(a_{l-2} + a_{l+2}) - \frac{7397}{12096}\,(a_{l-1} + a_{l+1}) - \frac{209}{216}a_l \\
h \cdot A_{l,l+1} &= -\frac{121}{60480}\,(a_{l-2} + a_{l+3}) + \frac{1}{160}\,(a_{l-1} + a_{l+2}) + \frac{31243}{60480}\,(a_{l+1} + a_l) \\
h \cdot A_{l,l+2} &= \frac{59}{60480}\,(a_{l-1} + a_{l+3}) - \frac{503}{15120}\,(a_l + a_{l+2}) + \frac{47}{480}a_{l+1} \\
h \cdot A_{l,l+3} &= \frac{29}{60480}\,(a_l + a_{l+3}) - \frac{617}{60480}\,(a_{l+1} + a_{l+2})
\end{aligned}
$$

Stiffness matrix for the hyperbolic term bu_x:

Left boundary part:

$$
\begin{aligned}
B_{0,0} &= -\frac{1}{3}b_0 - \frac{151}{630}b_1 + \frac{5}{54}b_2 - \frac{37}{1890}b_4 \\
B_{0,1} &= \frac{151}{315}b_0 + \frac{19}{63}b_1 - \frac{29}{315}b_2 + \frac{1}{45}b_3 \\
B_{0,2} &= -\frac{5}{27}b_0 - \frac{1}{14}b_1 + \frac{1}{210}b_2 - \frac{1}{270}b_3 \\
B_{0,3} &= \frac{37}{945}b_0 + \frac{1}{105}b_1 - \frac{1}{189}b_2 + \frac{1}{945}b_3
\end{aligned}
$$

$$
\begin{aligned}
B_{1,0} &= -\frac{151}{630}b_0 - \frac{38}{63}b_1 + \frac{103}{630}b_2 - \frac{2}{63}b_3 \\
B_{1,1} &= \frac{19}{63}b_0 - \frac{607}{1728}b_2 + \frac{1501}{30240}b_3 + \frac{1}{20160}b_4 \\
B_{1,2} &= -\frac{1}{14}b_0 + \frac{607}{1728}b_1 + \frac{4229}{20160}b_2 + \frac{11}{2520}b_3 - \frac{19}{8640}b_4 \\
B_{1,3} &= \frac{1}{105}b_0 - \frac{1501}{15120}b_1 - \frac{481}{20160}b_2 - \frac{221}{10080}b_3 + \frac{19}{8640}b_4 \\
B_{1,4} &= -\frac{1}{10080}b_1 + \frac{113}{60480}b_2 - \frac{1}{3024}b_3 - \frac{1}{20160}b_4
\end{aligned}
$$

$$
\begin{aligned}
B_{2,0} &= \frac{5}{54}b_0 + \frac{103}{630}b_1 - \frac{1}{105}b_2 + \frac{17}{1890}b_3 \\
B_{2,1} &= -\frac{29}{315}b_0 - \frac{607}{1728}b_1 - \frac{4229}{10080}b_2 + \frac{131}{6720}b_3 + \frac{1}{3024}b_4 \\
B_{2,2} &= \frac{1}{210}b_0 + \frac{4229}{20160}b_1 - \frac{7117}{30240}b_3 + \frac{157}{7560}b_4 - \frac{1}{20160}b_5 \\
B_{2,3} &= -\frac{1}{189}b_0 - \frac{481}{20160}b_1 + \frac{7117}{15120}b_2 + \frac{509}{2160}b_3 - \frac{19}{8640}b_5 \\
B_{2,4} &= \frac{113}{60480}b_1 - \frac{157}{3780}b_2 - \frac{103}{3360}b_3 - \frac{157}{7560}b_4 + \frac{19}{8640}b_5 \\
B_{2,5} &= -\frac{1}{10080}b_2 + \frac{113}{60480}b_3 - \frac{1}{3024}b_4 - \frac{1}{20160}b_5
\end{aligned}
$$

$$
\begin{aligned}
B_{3,0} &= -\frac{37}{1890}b_0 - \frac{2}{63}b_1 + \frac{17}{1890}b_2 - \frac{2}{945}b_3 \\
B_{3,1} &= \frac{1}{45}b_0 + \frac{1501}{30240}b_1 + \frac{131}{6720}b_2 + \frac{221}{5040}b_3 - \frac{113}{60480}b_4 \\
B_{3,2} &= -\frac{1}{270}b_0 + \frac{11}{2520}b_1 - \frac{7117}{30240}b_2 - \frac{509}{1080}b_3 + \frac{103}{3360}b_4 + \frac{1}{3024}b_5 \\
B_{3,3} &= \frac{1}{945}b_0 - \frac{221}{10080}b_1 + \frac{509}{2160}b_2 - \frac{14249}{60480}b_4 + \frac{157}{7560}b_5 + \frac{1}{20160}b_6 \\
B_{3,4} &= -\frac{1}{3024}b_1 - \frac{103}{3360}b_2 + \frac{14249}{30240}b_3 + \frac{14249}{60480}b_4 - \frac{19}{8640}b_6 \\
B_{3,5} &= \frac{113}{60480}b_2 - \frac{157}{3780}b_3 - \frac{103}{3360}b_4 - \frac{157}{7560}b_5 + \frac{19}{8640}b_6 \\
B_{3,6} &= -\frac{1}{10080}b_3 + \frac{113}{60480}b_4 - \frac{1}{3024}b_5 - \frac{1}{20160}b_6
\end{aligned}
$$

The inner scheme ($l = 4, 5, \ldots, N-4$) is given by

$$
\begin{aligned}
B_{l,l-3} &= \frac{1}{20160}b_{l-3} + \frac{1}{3024}b_{l-2} - \frac{113}{60480}b_{l-1} + \frac{1}{10080}b_l \\
B_{l,l-2} &= -\frac{19}{8640}b_{l-3} + \frac{157}{7560}b_{l-2} + \frac{103}{3360}b_{l-1} + \frac{157}{3780}b_l - \frac{113}{60480}b_{l+1} \\
B_{l,l-1} &= \frac{19}{8640}b_{l-3} - \frac{14249}{60480}b_{l-1} - \frac{14249}{30240}b_l + \frac{113}{3360}b_{l+1} + \frac{1}{3024}b_{l+2} \\
B_{l,l} &= -\frac{1}{20160}b_{l-3} - \frac{157}{7560}b_{l-2} + \frac{14249}{60480}b_{l-1} - \frac{14249}{60480}b_{l+1} + \frac{157}{7560}b_{l+2} + \frac{1}{20160}b_{l+3} \\
B_{l,l+1} &= -\frac{1}{3024}b_{l-2} - \frac{103}{3360}b_{l-1} + \frac{14249}{30240}b_l + \frac{14249}{60480}b_{l+1} - \frac{19}{8640}b_{l+3} \\
B_{l,l+2} &= \frac{113}{60480}b_l - \frac{157}{3780}b_l - \frac{103}{3360}b_{l+1} - \frac{157}{7560}b_{l+2} + \frac{19}{8640}b_{l+3} \\
B_{l,l+3} &= -\frac{1}{10080}b_l + \frac{113}{60480}b_{l+1} - \frac{1}{3024}b_{l+2} - \frac{1}{20160}b_{l+3}
\end{aligned}
$$

The right boundary part is constructed analogously to left one.
The stiffness matrix C for $(cu)_x$ can be calculated from B by interchanging the coefficient function and taking the negative transpose.

References

[1] George B. Arfken and Hans J. Weber. *Mathematical Methods for Physicists*. Academic Press, London, 2001.

[2] Mark H. Carpenter, David Gottlieb, and Saul Abarbanel. The Stability of Numerical Boundary Treatments for Compact High-Order Finite-Difference Schemes. *Journal of Computational Physics*, 108:272 – 295, 1993.

[3] Mark H. Carpenter, David Gottlieb, and Saul Abarbanel. Time-stable boundary conditions for finite-difference schemes solving hyperbolic systems: Methodology and application to high-order compact schemes. *Journal of Computational Physics*, 111(2):220 – 236, 1994.

[4] Bengt Fornberg. Calculation of Weights in Finite Difference Formulas. *SIAM Review*, 40:685 – 691, 1998.

[5] Bernard Friedman. *Principles and Techniques of Applied Mathematics*. Wiley, New York, 1966.

[6] Bertil Gustafsson. The convergence rate for difference approximations to general mixed initial boundary value problems. *SIAM Journal of Numerical Analysis*, 18(2):179 – 190, 1981.

[7] Bertil Gustafsson, Heinz-Otto Kreiss, and Joseph Oliger. *Time Dependent Problems and Difference Methods*. Wiley, New York, 1995.

[8] Heinz-Otto Kreiss and G. Scherer. Finite Element and Finite Difference Methods for Hyperbolic Partial Differential Equations. *Mathematical Aspects of Finite Elements in Partial Differential Equations*, 110, 1974.

[9] Heinz-Otto Kreiss and L. Wu. On the Stability Definition of Difference Approximations for the initial boundary value problem. *Applied Numerical Mathematics*, 12, 1993.

[10] Ken Mattsson. Boundary Procedures for Summation-by-Parts Operators. *Journal of Scientific Computing*, 18:133 – 153, 2003.

[11] Ken Mattsson and Jan Nordström. Summation by parts operators for finite difference approximations of second derivatives. *Journal of Computational Physics*, 199:503 – 540, 2004.

[12] Jan Nordström and Mark H. Carpenter. Boundary and Interface Conditions for High Order Finite Difference Methods applied to the Euler and Navier-Stokes Equations. *Journal of Computational Physics*, 148:621–645, 1999.

[13] Jan Nordström and Magnus Svärd. *On the Order of Accuracy for Difference Approximations of Initial-Boundary Value Problems*. Dept. of Scientific Computing, Uppsala University, 2004-040 edition, 2004.

[14] Pelle Olsson. Summation By Parts, Projections, and Stability. *Mathematics of Computation*, 64(211):1035 – 1065, 1995.

[15] Ivar Stakgold. *Green's Functions and Boundary Value Problems*. Wiley, New York, 1998.

[16] Bo Strand. Summation By Parts for Finite Difference Approximations for d/dx. *Journal of Computational Physics*, 110:47 – 67, 1994.

[17] Abraham Zemui. *High Order Symmetric Finite Difference Schemes for the Acoustic Wave Equation*. PhD thesis, Faculty of Science and Technology, Universitet Uppsala, 2003.