UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1004

Stable and High-Order Finite Difference Methods for Multiphysics Flow Problems

JENS BERG

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2013

ISSN 1651-6214
ISBN 978-91-554-8557-3
urn:nbn:se:uu:diva-187204

Dissertation presented at Uppsala University to be publicly examined in Room 2446, Lägerhyddsvägen 2, Uppsala, Friday, February 1, 2013 at 10:15 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract
Berg, J. 2013. Stable and High-Order Finite Difference Methods for Multiphysics Flow Problems. Acta Universitatis Upsaliensis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 1004. 35 pp. Uppsala. ISBN 978-91-554-8557-3.

Partial differential equations (PDEs) are used to model various phenomena in nature and society, ranging from the motion of fluids and electromagnetic waves to the stock market and traffic jams. There are many methods for numerically approximating solutions to PDEs. Some of the most commonly used ones are the finite volume method, the finite element method, and the finite difference method. All methods have their strengths and weaknesses, and it is the problem at hand that determines which method that is suitable. In this thesis, we focus on the finite difference method which is conceptually easy to understand, has high-order accuracy, and can be efficiently implemented in computer software.

We use the finite difference method on summation-by-parts (SBP) form, together with a weak implementation of the boundary conditions called the simultaneous approximation term (SAT). Together, SBP and SAT provide a technique for overcoming most of the drawbacks of the finite difference method. The SBP-SAT technique can be used to derive energy stable schemes for any linearly well-posed initial boundary value problem. The stability is not restricted by the order of accuracy, as long as the numerical scheme can be written in SBP form. The weak boundary conditions can be extended to interfaces which are used either in domain decomposition for geometric flexibility, or for coupling of different physics models.

The contributions in this thesis are twofold. The first part, papers I-IV, develops stable boundary and interface procedures for computational fluid dynamics problems, in particular for problems related to the Navier-Stokes equations and conjugate heat transfer. The second part, papers V-VI, utilizes duality to construct numerical schemes which are not only energy stable, but also dual consistent. Dual consistency alone ensures superconvergence of linear integral functionals from the solutions of SBP-SAT discretizations. By simultaneously considering well-posedness of the primal and dual problems, new advanced boundary conditions can be derived. The new duality based boundary conditions are imposed by SATs, which by construction of the continuous boundary conditions ensure energy stability, dual consistency, and functional superconvergence of the SBP-SAT schemes.

Keywords: Summation-by-parts, Simultaneous Approximation Term, Stability, High-order accuracy, Finite difference methods, Dual consistency

Jens Berg, Uppsala University, Department of Information Technology, Division of Scientific Computing, Box 337, SE-751 05 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-554-8557-3
urn:nbn:se:uu:diva-187204 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-187204)

List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I J. Lindström and J. Nordström. A stable and high-order accurate conjugate heat transfer problem. *Journal of Computational Physics, 229(14):5440–5456, 2010.*

II J. Berg and J. Nordström. Spectral analysis of the continuous and discretized heat and advection equation on single and multiple domains. *Applied Numerical Mathematics, 62(11):1620–1638, 2012.*

III J. Berg and J. Nordström. Stable Robin solid wall boundary conditions for the Navier–Stokes equations. *Journal of Computational Physics, 230(19):7519–7532, 2011.*

IV J. Nordström and J. Berg. Conjugate heat transfer for the unsteady compressible Navier–Stokes equations using a multi-block coupling. *Accepted for publication in Computers & Fluids, 2012.*

V J. Berg and J. Nordström. Superconvergent functional output for time-dependent problems using finite differences on summation-by-parts form. *Journal of Computational Physics, 231(20):6846–6860, 2012.*

VI J. Berg and J. Nordström. On the impact of boundary conditions on dual consistent finite difference discretizations. *Accepted for publication in Journal of Computational Physics, 2012.*

Contents

1. Introduction

Many problems in the natural sciences can be described in the language of mathematics as systems of partial differential equations (PDEs). A system of PDEs typically describes the time-evolution of physical quantities such as velocity, momentum, and energy in a coupled manner. There are no general methods to compute analytical solutions to PDEs, and even when there are analytical solutions available, they are often not suitable for practical applications due to their complexity. Numerical methods for solving the PDEs are therefore the preferred and often only choice.

The increase in computing power over the past decades has helped to establish numerical simulations as the third cornerstone of science, alongside theoretical analysis and practical experiments. As the usage of computers grow, the algorithms which produce the numerical results become increasingly important. In particular for solving PDEs, there is a multitude of available methods. Each of them have their strengths and weaknesses, and the problem at hand determines which method that is suitable.

In this thesis, the problems under consideration usually appear in computational fluid dynamics (CFD) applications. The typical and most general example is the compressible Navier–Stokes equations which describe the motion of a compressible fluid. The Navier–Stokes equations provide a challenge for both mathematicians and numerical analysts. From a mathematical point of view, it has not yet been proven that a global smooth solution exists in three space dimensions. From a numerical point of view, the treatment of boundary conditions and high complexity make the construction of numerical schemes highly non-trivial.

It is common in CFD to derive numerical methods for model problems which are subsequently applied to more complicated equations. Model problems are constructed so that the main mathematical properties of the real problem are preserved, but the analysis is simplified. Also, when implementing the solution algorithms for a model problem, flaws in the algorithms are not hidden by the algebraic complexity of the equations to be solved.

Whatever numerical method used to solve PDEs, the following requirements have to be satisfied;

1. Consistency
2. Stability
3. Efficiency

By the famous theorem of Lax and Richtmeyer [26], the solution of a linear PDE given by a numerical method converges to the solution of the PDE if,

and only if, the method is consistent and stable. All schemes which are used in practice are consistent by construction. Far from all schemes are, however, stable. That is what brings us to the main topic of this thesis—the construction of stable and high-order accurate numerical schemes for solving time-dependent partial differential equations.

2. The summation-by-parts technique

A finite difference method for solving differential equations is constructed by approximating the derivatives in discrete points as weighted sums of solution values in neighboring points. Recall the mathematical definition of the first derivative;

$$u'(x) = \lim_{h \to 0} \frac{u(x+h) - u(x)}{h}. \tag{2.1}$$

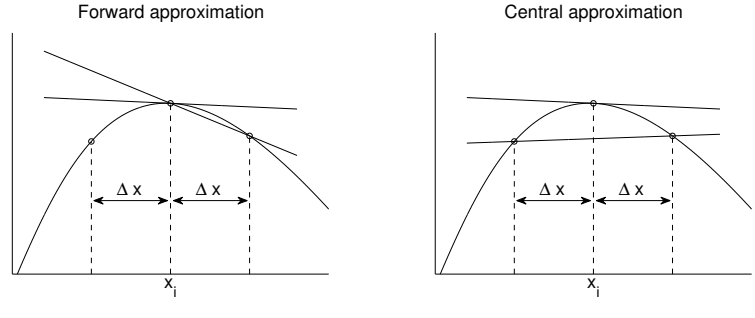
A computer has finite precision and hence h in (2.1) can not be made arbitrarily small. Instead, a computational grid is introduced where $h = \Delta x > \delta > 0$ and the first derivative at the point $x = x_i$ in (2.1) becomes approximated as

$$u'(x_i) \approx \frac{u(x_i + \Delta x) - u(x_i)}{\Delta x}. \tag{2.2}$$

In (2.2), only one neighbor-point is used. More points can be included to obtain more accurate approximations of the first derivative. For example the central approximation, where two neighbor-points are used, given by

$$u'(x_i) \approx \frac{u(x_i + \Delta x) - u(x_i - \Delta x)}{2\Delta x}. \tag{2.3}$$

The geometric interpretations of (2.2) and (2.3) can be seen in Figure 2.1.



(a) Forward difference using one neighbor-point

(b) Central difference using two neighbor-points

Figure 2.1. Geometric interpretation of first derivative approximation using forward and central differences

We say that (2.3) is second-order accurate since substituting the Taylor series expansion of $u(x)$ around $x = x_i$ gives

$$\frac{u(x_i+\Delta x)-u(x_i-\Delta x)}{2\Delta x} = u'(x_i) + \frac{\Delta x^2}{6}u^{(3)}(\xi) + \ldots,$$

where $x_i - \Delta x \leq \xi \leq x_i + \Delta x$. Thus if the third derivative of u is sufficiently smooth, the error term will behave as Δx^2 and tend to zero as $\Delta x \to 0$.

For Cauchy problems, the stability criteria for a given numerical scheme can be analyzed with von Neumann analysis. For an initial boundary value problem (IBVP), however, the formula (2.3) reveals difficulties. For example, if the point $x_i = x_0$ is a boundary point, then $x_0 - \Delta x$ is not included in the discretization and special care has to be taken.

The difficulties at the boundaries for IBVPs using finite difference methods is what gave birth to the summation-by-parts (SBP) form [23, 24]. We say that;

Definition 2.1. A finite difference matrix D_1 is an SBP operator for the first derivative if

$$\begin{aligned} D_1 &= P^{-1}Q, \\ Q + Q^T &= E_N - E_0 = \text{diag}[0,\ldots,0,1] - \text{diag}[1,0,\ldots,0], \end{aligned}$$

and the matrix P defines an inner product and norm by

$$(u_h, v_h)_h = u_h^T P v_h, \quad ||u_h||^2 = u_h^T P u_h,$$

for any discrete grid functions u_h, v_h.

Given these definitions, we have

$$(u_h, D_1 v_h)_h = u_h^T (E_N - E_0) v_h - (D_1 u_h, v_h)_h,$$

which mimics integration by parts in the continuous sense and motivates the SBP terminology. Essentially, an SBP operator is a central finite difference operator in the interior while the boundaries have been modified so that the operator is one-sided. For example, the second-order accurate operator is given by

$$D_1 = P^{-1}Q = \frac{1}{2\Delta x}\begin{bmatrix} -2 & 2 & 0 & 0 & \ldots & 0 \\ -1 & 0 & 1 & 0 & \ldots & 0 \\ 0 & -1 & 0 & 1 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & -1 & 0 & 1 \\ 0 & \ldots & 0 & 0 & -2 & 2 \end{bmatrix},$$

where

$$P = \Delta x \begin{bmatrix} \frac{1}{2} & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \dots & 0 & 1 & 0 \\ 0 & \dots & 0 & 0 & \frac{1}{2} \end{bmatrix}, \quad Q = \frac{1}{2} \begin{bmatrix} -1 & 1 & 0 & \dots & 0 \\ -1 & 0 & 1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & -1 & 0 & 1 \\ 0 & \dots & 0 & -1 & 1 \end{bmatrix}.$$

There are SBP operators for the first derivative with interior order of accuracy $2p$ for $p = 1,2,3,4$. The global accuracy depends on the choice of the norm matrix P. With the requirement of P being diagonal, the order of accuracy at the boundaries needs to be reduced to p. The global order of accuracy then becomes $p+1$. There are also block-diagonal matrices which give $2p$-order global accuracy [46]. While a diagonal P gives less accuracy, it has more flexibility. For example, a diagonal norm is required to derive energy estimates under curvilinear coordinate transforms since P has to commute with the (diagonal) Jacobian matrix of the coordinate transform [37, 48]. In this thesis, a diagonal matrix P has been consistently used.

Once an energy estimate has been derived, a higher order accurate solution can be obtained by simply replacing the difference operator with one of higher order.

SBP operators can also be used to approximate the second derivative. The most direct way is to apply the first derivative twice, $D_2 = D_1 D_1$, which results in a wide difference stencil. The order of accuracy is the same as for the first derivative. A compact stencil can be obtained by considering a second derivative operator of the form

$$D_2 = P^{-1}(-A + (E_N - E_0)S),$$

where $A + A^T \geq 0$ and S approximates the first derivative at the boundary. In this case, S can be chosen to be accurate of order $p+1$ instead of p and the global accuracy increases from $p+1$ to $p+2$ for pointwise-stable discretizations [4, 31, 50].

Several attempts to include the boundary conditions were made after the construction of the SBP finite difference operator. Injection of the boundary values destroy the SBP properties and stability is restricted to low-order accurate schemes. An orthonormal projection method which preserves the SBP properties was proposed in [41, 42], but is not in practical use because of other complications [28, 30]. The current state-of-the art method for imposing the boundary conditions was proposed by Carpenter et al. in [3] and has become known as the Simultaneous Approximation Term (SAT). Together, the SBP-SAT technique provides a method for constructing stable and high-order accurate approximations of IBVPs.

2.1 Initial boundary value problems

Analyzing the stability requirements for a full time and space discretization is difficult. The analysis can be simplified by only discretizing in space while keeping time continuous. The stability analysis can then be done by using the energy method which is applicable to complicated problems. Such a semi-discretization is called a method of lines [47].

2.1.1 Well-posedness of the continuous problem

The semi-discrete energy estimates are closely related to well-posedness of the continuous problem. One of the earliest definitions of well-posedness were given by Hadamard [16] in the 1920's and can be stated as;

Definition 2.2 (Hadamard)**.** A problem is called well-posed if

1. A solution exists
2. The solution is unique
3. The solution depends smoothly on the data of the problem

The two first statements are obvious for a problem to be computable. The third statement is somewhat vaguely formulated. In Hadamard's original texts, data refers to everything from initial and boundary data to the boundary conditions. Even so, it is clear that such a definition is necessary from a numerical point of view. Every numerical computation produces discretization and round-off errors. These errors can be thought of as data of the problem, and perturbations due to finite precision arithmetic can not be allowed to affect the solution too much.

When studying finite difference discretizations of IBVPs, Kreiss [25] made another definition of well-posedness which became very influential for numerical solutions of PDEs. The definition can be stated as;

Definition 2.3 (Kreiss)**.** A homogeneous IBVP is well-posed if a unique solution u exists and satisfies the energy estimate

$$||u|| \leq K_c e^{\alpha_c t} ||f||, \quad \forall t > 0,$$

where f is the initial data. The parameters K_c and α_c are not allowed to depend on neither t nor f.

By the principle of Duhamel, it is sufficient to study the homogeneous problem since well-posedness of the inhomogeneous problem follows. Moreover, the boundary conditions can also be assumed to be homogeneous [15]. Definition 2.3 quantified the vague definition of Hadamard, and also allowed the solution to be stable against lower-order perturbations. The later property has

extensive use in numerics since non-linear and variable coefficient problems can be treated using linearizations and localizations [22].

2.1.2 Stability of the semi-discrete problem

The definition of Kreiss did not only improve the theory of PDEs in general, it also suggested a method for proving stability of semi-discretizations. The same reasoning can namely be applied in the discrete sense as described in;

Definition 2.4 (Kreiss)**.** A semi-discretization of a homogeneous IBVP is called stable if the discrete solution u_h satisfies the energy estimate

$$||u_h|| \leq K_d e^{\alpha_d t}||f||, \quad \forall t > 0,$$

where f is the initial data. The parameters K_d and α_d are not allowed to depend on neither t nor f.

Kreiss and Wu [25] showed that when time is kept continuous and a space discretization is stable according to definition 2.4, a Runge-Kutta time integration scheme can be used to integrate the solution in time while maintaining stability.

The outlined procedure of assuming that there is no forcing function and that the boundary conditions are homogeneous gives the most basic sufficient requirements of well-posedness and stability. There are other definitions where the data of the problem is included in the estimates, in which case the continuous problem is called strongly well-posed and the semi-discretization is called strongly stable. Moreover, the semi-discrete problem is called strictly stable if $\alpha_d = \alpha_c + O(\Delta x)$. More details on the definitions and their usage can be found in [15, 36, 39].

We can now exemplify the whole idea of the SBP-SAT method by considering the advection equation with wavespeed $\bar{u} > 0$,

$$\begin{aligned} u_t + \bar{u}u_x &= 0, \qquad 0 \leq x \leq 1, \\ u(x,0) &= f(x), \\ u(0,t) &= g_L(t). \end{aligned} \tag{2.4}$$

Assuming that a unique solution exists, we let $g_L = 0$ and integrate (2.4) over the spatial domain. We obtain

$$\frac{d}{dt}||u||^2 = -\bar{u}u(1,t)^2 \leq 0,$$

which leads to an energy estimate and hence (2.4) is well-posed. An SBP-SAT discretization of (2.4) can be written as

$$\frac{d}{dt}u_h + \bar{u}D_1 u_h = \sigma P^{-1} e_0 (e_0^T u_h - g_L(t)) \tag{2.5}$$

where $e_0 = [1,0,\ldots,0]^T$. The parameter σ has to be determined such that (2.5) is stable in the norm defined by P. By multiplying (2.5) with $u_h^T P$, assuming $g_L = 0$, we get

$$\frac{d}{dt}||u_h||^2 = (\bar{u}+2\sigma)u_h^T E_0 u_h - \bar{u}u_h^T E_N u_h \tag{2.6}$$

and a discrete energy estimate is obtained for $\sigma \leq -\bar{u}/2$. For those values of σ, the scheme is stable. Note that there is no restriction on the order of accuracy for stability in the energy estimate (2.6). Once a discrete energy estimate has been obtained, the same requirements are valid for all orders of accuracy.

The construction of stable boundary procedures for the compressible Navier–Stokes equations with Robin solid wall boundary conditions is the topic of paper III.

2.2 Coupled problems

To study complex flow phenomena, such as conjugate heat transfer, the flow equations need to be coupled with the equations for heat transfer [17, 8, 45]. For model problems, well-posed coupling conditions can be derived using the standard energy method. When the coupling conditions are derived from first principles of physics, the energy method in its standard setting might be insufficient. The reason is that the energy estimates are derived in the L^2-norm which might not capture the physics of the problem. A simple example is the coupled heat equations in one dimension, given by

$$\begin{aligned} u_t &= \alpha_L u_{xx}, & -1 \leq x \leq 0, \\ v_t &= \alpha_R v_{xx}, & 0 \leq x \leq 1, \\ u(-1,t) &= g_L(t), & v(1,t) = g_R(t), \\ u(0,t) &= v(0,t), & \kappa_L u_x(0,t) = \kappa_R v_x(0,t), \end{aligned}$$

where $\alpha_{L,R} = \frac{\kappa_{L,R}}{c_{L,R}\rho_{L,R}}$ are the thermal diffusivities and $\kappa_{L,R}$, $c_{L,R}$, and $\rho_{L,R}$ are the thermal conductivities, specific heat capacities, and densities, respectively. The coupling conditions require continuity of temperature and heat fluxes. In order to obtain an energy estimate, it is necessary to modify the norms as

$$||u||_L^2 = \int_{-1}^{0} u^2 \delta_L dx, \quad ||v||_R^2 = \int_{0}^{1} v^2 \delta_R dx,$$

where $\delta_{L,R} > 0$ are to be determined. The energy method (assuming $g_L = g_R = 0$) results in

$$\frac{d}{dt}(||u||_L^2 + ||v||_R^2) + 2\alpha_L||u_x||_L^2 + 2\alpha_R||v_x||_R^2 = [\delta_L \alpha_L u u_x - \delta_R \alpha_R v v_x]_{x=0}. \tag{2.7}$$

To obtain an energy estimate, using the interface conditions, it is required that $\delta_{L,R} = c_{L,R}\rho_{L,R}$ since then (2.7) reduces to

$$\frac{d}{dt}(||u||_L^2 + ||v||_R^2) + 2\alpha_L||u_x||_L^2 + 2\alpha_R||v_x||_R^2 = [(\kappa_L u_x - \kappa_R v_x)u]_{x=0} = 0,$$

and an energy estimate is obtained.

The modifications of the norms are also seen in the discretization of the coupled problem. A discretization using the SBP-SAT method can be written as

$$\begin{aligned}
\frac{d}{dt}u_h &= \alpha_L D_1^2 u_h + \sigma_1 P^{-1} D_1^T e_0(e_0^T u_h - g_L) \\
&\quad + \sigma_2 P^{-1} D_1^T e_N(e_N^T u_h - e_0^T v_h) + \sigma_3 P^{-1} e_N(\kappa_L e_N^T(D_1 u_h) - \kappa_R e_0^T(D_1 v_h)) \\
\frac{d}{dt}v_h &= \alpha_R D_1^2 v_h + \tau_1 P^{-1} D_1^T e_N(e_N^T v_h - g_R) \\
&\quad + \tau_2 P^{-1} D_1^T e_0(e_0^T v_h - e_N^T u_h) + \tau_3 P^{-1} e_0(\kappa_R e_0^T(D_1 v_h) - \kappa_L e_N^T(D_1 u_h))
\end{aligned} \tag{2.8}$$

and we have to choose $\sigma_{1,2,3}$ and $\tau_{1,2,3}$ such that the scheme is stable. For simplicity, we have assumed that both domains have the same number of grid points since then the same operators can be used in both domains. This is to simplify the notation and in general the domains can have different discretizations. Since a modified norm was required to obtain an energy estimate in the continuous case, the same modification is required to obtain a discrete energy estimate. The modified discrete norms are defined analogously as

$$||u_h||_L^2 = \delta_L u_h^T P u_h, \quad ||v_h||_R^2 = \delta_R v_h^T P v_h,$$

with $\delta_{L,R}$ determined from the continuous energy estimate. To highlight the relation to the continuous energy estimate, we consider only the interface terms and apply the modified energy method with general $\delta_{L,R}$. We get

$$\frac{d}{dt}(||u_h||_L^2 + ||v_h||_R^2) + 2\alpha_L||D_1 u_h||_L^2 + 2\alpha_R||D_1 v_h||_R^2 = q_h^T M q_h,$$

where $q_h = [e_N^T u_h, e_0^T v_h, e_N^T(D_1 u_h), e_0^T(D_1 v_h)]^T$ and

$$M = \begin{bmatrix} 0 & 0 & m_1 & m_2 \\ 0 & 0 & m_3 & m_4 \\ m_1 & m_3 & 0 & 0 \\ m_2 & m_4 & 0 & 0 \end{bmatrix},$$

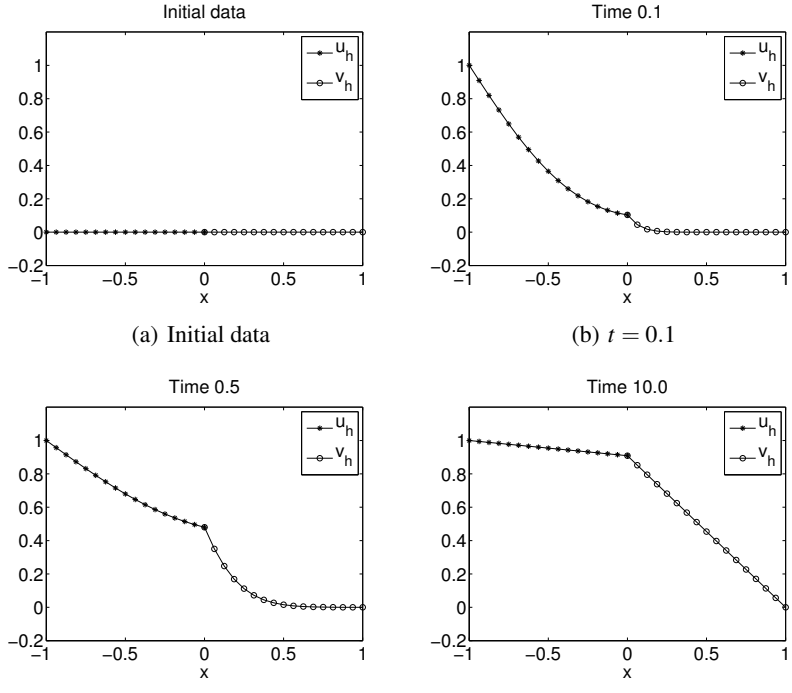
with

$$\begin{aligned}
m_1 &= (\alpha_L + \sigma_2 + \sigma_3\kappa_L)\delta_L, & m_2 &= -\sigma_3\delta_L\kappa_R - \tau_2\delta_R, \\
m_3 &= -\sigma_2\delta_L - \tau_3\delta_R\kappa_L, & m_4 &= (-\alpha_R + \tau_2 + \tau_3\kappa_R)\delta_R.
\end{aligned}$$

In order to obtain an energy estimate, it is required that all parameters are chosen such that $M \leq 0$. Since the main diagonal of M consists of zeros, the only option is to choose the parameters such that $m_{1,2,3,4} = 0$. A little bit of algebra shows that this requirement is possible if, and only if,

$$\frac{\delta_L}{\delta_R} = \frac{c_L \rho_L}{c_R \rho_R},$$

which is satisfied by the choices of $\delta_{L,R}$ from the continuous energy estimate. Thus, if a modified norm is required to obtain an energy estimate in the continuous case, the same modification has to be done to the discrete norm. An example of an implementation of the scheme (2.8) can be seen in Figure 2.2, where we have chosen the problem parameters such that $\alpha_L/\alpha_R = \kappa_L/\kappa_R = 10$.



(a) Initial data (b) $t = 0.1$

(c) $t = 0.5$ (d) Steady-state

Figure 2.2. A sequence of solutions for two coupled heat equations with different problem parameters

In Figure 2.2, the initial data did not match the boundary data at the left boundary. This causes instabilities for schemes with strong implementation of the boundary conditions. With weak boundary conditions and energy stability, the solution attains the boundary value and the scheme remains stable throughout the computation. In this example, we have used 33 grid points in each subdomain and second-order accurate SBP operators.

In paper I and paper IV we investigate coupling procedures for computing conjugate heat transfer problems. In the first case for a one-dimensional model problem, and in the second case for the two-dimensional compressible Navier–Stokes equations. In paper II, the coupling procedure itself is studied using model problems.

3. Functionals and dual problems

The solution of the governing equations might not be the output of primary interest in many CFD applications. Of equal, or even greater, importance is the computation of functionals from the solution. In general, a functional is defined as any map from a vector space V into the underlying scalar field $\mathbb{K}$. Every vector space has an associated vector space called its dual (or adjoint) space. The dual space is denoted by V^* and is defined as the space of all linear functionals $V \to \mathbb{K}$.

The adjoint, or dual, operator L^* of a linear operator L is the (unique) operator satisfying

$$(v, Lu)_V = (L^* v, u)_V, \tag{3.1}$$

where $(.,.)_V$ denotes the inner product on the space V. The study of linear functionals and dual spaces is the topic of functional analysis and additional preliminaries can be found in any functional analysis textbook, for example the classical works [43, 44].

In this section, we consider initial boundary value problems of the form

$$\begin{aligned} u_t + \mathscr{L}(u) &= F, \quad && x \in \Omega, \\ \mathscr{B}(u) &= g_\Gamma, && x \in \Gamma \subseteq \partial\Omega, \\ u &= f, && t = 0. \end{aligned} \tag{3.2}$$

For applications in CFD, a linear functional of interest usually represents the lift or drag on a solid body in a fluid, which is computed in terms of an integral of the solution of (3.2). The functional can be represented in terms of an integral inner product as

$$J(u) = (g, u) = \int_\Omega g^T u d\Omega,$$

where g is a weight function. A main complication in CFD is that no physically relevant solutions have compact support in the computational domain. The dual operator is obtained through integration by parts which will introduce boundary terms that must be removed. The dual PDE has thus to be supplied with dual boundary conditions to close the system.

The associated dual problem has been extensively studied [11, 12] and used in the context of error control and adaptive mesh refinement [1, 2, 6, 14, 10, 7] as well as within optimization and control problems [21, 13]. In error control

and mesh adaptation, the dual problem is derived and treated as a variational problem. In optimization and control problems, the dual problem is derived and treated as a sensitivity problem with respect to design parameters. In the end, the two different formulations yield the same dual problem. A similarity for the different areas of applications is that most of them are based on unstructured methods, such as finite elements or discontinuous Galerkin.

3.1 Quadrature accuracy

Only recently was the study of duality introduced to structured methods, such as the SBP-SAT technique. Recall that the SBP operator was constructed to satisfy

$$(v_h, D_1 u_h)_h = u_h^T (E_N - E_0) v_h - (D_1 v_h, u_h)_h,$$

which mimics an integration property, rather than a differentiation property. While the differentiation properties of the SBP operator has been extensively studied and used [46, 49, 32, 50, 20, 38, 5, 33, 29], the integration properties of the matrix P have been much less explored. The integration properties of P was thoroughly investigated by Hicken and Zingg [19]. It was shown that the requirements on P to obtain an accurate SBP operator include, and extend, the Gregory formulas for quadrature rules using equidistant points. Two main results were proven in [19], which are restated here for convenience. The first theorem establishes the accuracy of P as an integration operator;

Theorem 3.1. *Let P be a full, restricted-full, or diagonal mass matrix from an SBP first-derivative operator $D_1 = P^{-1}Q$, which is a $2p$-order accurate approximation to the first derivative in the interior. Then the mass matrix P constitutes a $2p$-order accurate quadrature for integrands $u \in C^{2p}(\Omega)$.*

The second theorem extends the results to include discrete integrands computed from an SBP differentiation;

Theorem 3.2. *Let $D_1 = P^{-1}Q$ be a an SBP first derivative operator with a diagonal mass matrix P and $2p$-order interior accuracy. Then $(v_h, D_1 u_h)_h$ is a $2p$-order accurate approximation of (v, u_x).*

These theorems proved in summary that it is possible to retain the full order of accuracy when computing integrals from an SBP discretization, even with a diagonal P.

3.2 Dual consistency

For IBVPs, it is not sufficient to integrate the solution obtained by an SBP-SAT discretization using P to obtain a functional of $2p$-order accuracy. It was shown in [18] that an additional property of the discretization was required—the so called dual consistency property. The main result in [18] extends the results in [19] to include SBP-SAT solutions to IBVPs. Even though the solution u_h to an IBVP using SBP-SAT is accurate of order $p+1$ when using a diagonal P, any linear functional of u_h is accurate of order $2p$ when integrated using P, if the discretization is dual consistent.

As suggested by the name, dual consistency requires that the discretization of the primal problem is also a consistent approximation of the dual problem. In order to construct a dual consistent discretization, one first have to derive the dual problem and work with both the primal and dual problems simultaneously. To obtain the dual differential operator we consider the linear, or linearized, Cauchy problem,

$$\begin{aligned} u_t + Lu &= f, \qquad x \in \Omega, \\ u &= 0, \qquad t = 0, \\ J(u) &= (g,u), \end{aligned}$$

where $J(u)$ is a linear functional of interest. We seek a function θ, in some appropriate function space, such that

$$\int_0^T J(u)dt = \int_0^T (\theta, f)dt.$$

Using integration by parts, we can write

$$\begin{aligned} \int_0^T J(u)dt &= \int_0^T J(u)dt - \int_0^T (\theta, u_t + Lu - f)dt \\ &= \int_0^T (\theta_t - L^*\theta + g, u)dt - [(\theta,u)]_{t=T} + \int_0^T (\theta, f)dt \end{aligned}$$

and it is clear that $\theta = 0$ at $t = T$ is needed, and that θ has to satisfy the dual equation $-\theta_t + L^*\theta = g$. The time transform $\tau = T - t$ is usually introduced, and the dual Cauchy problem becomes

$$\begin{aligned} \theta_\tau + L^*\theta &= g, \quad x \in \Omega, \\ \theta &= 0, \quad \tau = 0. \end{aligned}$$

The situation is more complicated for IBVPs. Since the primal equation does not have compact support in general, the boundary terms resulting from the

integration by parts procedure has to be properly taken care of by the homogeneous primal boundary conditions. The dual boundary conditions are defined as the minimal set of homogeneous conditions such that the boundary terms vanish after the homogeneous primal boundary conditions have been applied. Still, one needs to investigate the well-posedness of the dual equation with the resulting dual boundary conditions. A well-posed set of boundary conditions for the primal problem does not necessary lead to a well-posed dual problem.

A discretization of a problem with a functional of interest can be written as

$$\begin{aligned} \frac{d}{dt}u_h + L_h u_h &= f, \\ J_h(u_h) &= (g, u_h)_h, \end{aligned} \tag{3.3}$$

where the entire spatial discretization, including the boundary conditions, has been collected into the discrete operator L_h. Recall that the inner product is defined as

$$(v_h, u_h)_h = v_h^T P u_h \tag{3.4}$$

in an SBP-SAT framework. The discrete adjoint operator L_h^* is defined, analogously to (3.1), as the unique operator satisfying

$$(v_h, L_h u_h)_h = (L_h^* v_h, u_h)_h. \tag{3.5}$$

The discrete adjoint operator can hence be explicitly computed, using (3.4) and (3.5), as

$$L_h^* = P^{-1} L_h^T P. \tag{3.6}$$

The discrete dual problem is obtained analogously to the continuous case by finding θ_h such that $\int_0^T J_h(u_h)dt = \int_0^T (\theta_h, f)dt$. Integration by parts and (3.6) gives

$$\begin{aligned} \int_0^T J_h(u_h)dt &= \int_0^T (g, u_h)_h dt - \int_0^T (\theta_h, \frac{d}{dt}u_h + L_h u_h - f)_h dt \\ &= \int_0^T (\frac{d}{dt}\theta_h - L_h^*\theta_h + g, u_h)_h dt - [(\theta_h, u_h)_h]_{t=T} + \int_0^T (\theta_h, f)_h dt \end{aligned}$$

and hence the θ_h has to satisfy the discrete dual problem

$$\begin{aligned} \frac{d}{d\tau}\theta_h + L_h^*\theta_h &= g, \\ \theta_h &= 0, \quad \tau = 0, \end{aligned}$$

where $\tau = T - t$. Dual consistency can now be defined in terms of L_h^* and L^*;

Definition 3.3. A discretization is called dual consistent if L_h^* is a consistent approximation of L^* and the continuous dual boundary conditions.

The above definition is not specific for SBP-SAT discretizations. Any discretization which can be written in the form (3.3) is applicable. The SBP-SAT technique is particularly well-suited for this framework because of the well-defined inner product and operator form.

It is common, in optimization for example, that continuous and discrete adjoint methods are distinguished [34, 35, 9]. This is because the discrete adjoint operator does not approximate the continuous adjoint operator and boundary conditions in general. In the SBP-SAT framework, the dual consistency property can allow for very efficient use of adjoint based techniques due to the unification of the continuous and discrete adjoints. SBP-SAT is not the only method which offers consistency with the dual equations. It was shown that, for example, the discontinuous Galerkin method can also exhibit this property [27, 40].

The dual consistency property can be easily exemplified using the model problem (2.4). Dual consistency does not depend on any data of the problem but only the differential operator and the form of the boundary conditions. We hence consider the inhomogeneous problem with homogeneous boundary and initial conditions,

$$\begin{aligned} u_t + \bar{u}u_x &= f, \qquad 0 \leq x \leq 1 \\ u(0,t) &= 0, \\ u(x,0) &= 0, \\ J(u) &= (g,u), \end{aligned} \tag{3.7}$$

where $J(u)$ is a linear functional of interest. We seek a function θ so that $\int_0^T J(u)dt = \int_0^T (\theta, f)dt$ and integration by parts gives

$$\begin{aligned} \int_0^T J(u)dt &= \int_0^T J(u)dt - \int_0^T (\theta, u_t + \bar{u}u_x - f)dt \\ &= \int_0^T (\theta_t + \bar{u}\theta_x + g, u)dt - \int_0^1 [\theta u]_{t=T}dx - \int_0^T [\bar{u}\theta u]_{x=1}dt + \int_0^T (\theta, f)dt. \end{aligned}$$

It is clear that θ has to satisfy the dual problem

$$\begin{aligned} \theta_\tau - \bar{u}\theta_x &= g, \quad 0 \leq x \leq 1, \\ \theta(1,\tau) &= 0, \\ \theta(x,0) &= 0, \end{aligned} \tag{3.8}$$

where we have introduced the time transform $\tau = T - t$.

The model problem (3.7) can be discretized as

$$\begin{aligned} \frac{d}{dt}u_h + \bar{u}D_1 u_h &= \sigma P^{-1}(e_0^T u_h - 0)e_0 + f, \\ J_h(u_h) &= (g, u_h)_h, \end{aligned} \tag{3.9}$$

and the parameter σ has to be determined so that the scheme is not only stable, but also a consistent approximation of the dual problem (3.8). It is convenient to rewrite (3.9) in operator form as

$$\frac{d}{dt}u_h + L_h u_h = f,$$

where the spatial discretization, including the boundary condition, is included in the operator

$$L_h = \bar{u}D_1 - \sigma P^{-1}E_0.$$

The discrete dual operator can be directly computed as

$$L_h^* = P^{-1}L_h^T P = -\bar{u}D_1 + \bar{u}P^{-1}E_N - (\sigma + \bar{u})P^{-1}E_0, \tag{3.10}$$

and it is seen that L_h^* imposes a boundary condition at $x = 0$, due to the last term in (3.10), unless $\sigma = -\bar{u}$. With $\sigma = -\bar{u}$, the discrete dual problem becomes

$$\frac{d}{d\tau}\theta_h - \bar{u}D_1\theta_h = -\bar{u}P^{-1}E_N\theta_h + g,$$

which is a consistent approximation of the dual problem (3.8). Since $\sigma = -\bar{u}$ does not contradict the stability condition ($\sigma \leq -\bar{u}/2$), the scheme is both stable and dual consistent. In Table 3.1 we show the convergence rates q for the solution and the functionals, together with the functional error, using the dual inconsistent and consistent schemes.

Table 3.1. *Convergence rates q, and functional errors for the dual inconsistent and consistent schemes*

5th-order ($2p = 8$)						
	$\sigma = -1/2$			$\sigma = -1$		
N	$q(u_h)$	$q(J_h(u_h))$	Error	$q(u_h)$	$q(J_h(u_h))$	Error
96	4.58	4.51	1.87e-05	5.14	8.20	7.54e-09
128	4.87	4.80	3.02e-06	5.34	7.96	2.71e-10
160	4.97	4.91	7.58e-07	5.41	8.02	2.74e-11
192	5.02	4.97	2.53e-07	5.44	8.06	4.58e-12
224	5.05	5.01	1.02e-07	5.46	8.21	1.05e-12
256	5.06	5.04	4.72e-08	5.46	8.62	2.97e-13

As we can see from Table 3.1, the convergence rate for the linear functional increases from $p+1$ to $2p$ when using the dual consistent discretization. Also

notice that dual consistency is merely a choice of parameters. The solution of the dual problem is never required and hence the increased rate of convergence for linear functionals comes at no extra computational cost.

In paper V we establish the dual consistency theory for time-dependent problems, and relate the dual consistency property to stability using several model problems of different types. A general proof is presented which shows that stable and dual consistent SBP-SAT schemes produces superconvergent linear integral functionals. In paper VI we extend the theory to include advanced boundary conditions which further enhances the performance of dual consistent schemes.

4. Summary of papers

4.1 Contributions

The ideas of the papers in this thesis have been developed in close collaboration between the authors. The papers have been written by the author of this thesis. The computations in the papers have been performed by the author of this thesis. The analysis in the papers have been done to large extent by the author of this thesis in close collaboration with the co-author.

4.2 Paper I

J. Lindström and J. Nordström. A stable and high-order accurate conjugate heat transfer problem. *Journal of Computational Physics, 229(14):5440–5456, 2010.*

This paper was a first attempt to compute conjugate heat transfer problems using the SBP-SAT framework. In previous work, the coupling procedures have been focused on multi-block couplings to split the computational domain. For conjugate heat transfer problems, not only is the domain split, there are also different governing equations in the blocks describing the fluid and solid, respectively.

A one-dimensional model problem was analyzed. An incompletely parabolic system of equations was coupled to the scalar heat equation and well-posed interface conditions were derived for the continuous problem. The coupled problem was discretized and it was shown how to construct an SAT so that the coupling is stable, and that the target high-order accuracy was obtained.

The stable discrete coupling was derived as a function of one parameter describing the weight between Dirichlet and Neumann conditions, showing that there are no restrictions on how the coupling is done. The results extends earlier results where restrictions on the coupling were required for stability.

The spectral properties of the discretization were investigated as a function of the interface parameter and it was shown that both the rate of convergence to steady-state as well as the stiffness of the coupled system could be enhanced compared to having pure Dirichlet or Neumann conditions.

4.3 Paper II

J. Berg and J. Nordström. Spectral analysis of the continuous and discretized heat and advection equation on single and multiple domains. *Applied Numerical Mathematics, 62(11):1620–1638, 2012.*

To obtain further insights in how numerical coupling procedures affect the overall discretization, two model problems were investigated. Most physical problems consist of advective and/or diffusive terms which have very different mathematical and numerical properties. We hence considered the advection and heat equation on single and multiple domains. This was done to see which effects the multi-block coupling have, compared to a single domain discretization. Second-order accurate SBP operators were used since they allow analytic computations of spectral properties.

For the heat equation, we derived a closed form expression of the eigenvalues for the discrete single domain operator and showed asymptotical second-order convergence of all discrete eigenvalues. For the multi-block domain we showed that the eigenvalues from the single domain operator were included in the set of eigenvalues of the multi-block operator. The stable coupling conditions were derived as a function of one coupling parameter, similarly as in paper I, for which the discretization properties were studied.

For the advection equation, we showed how the eigenvalues of the multi-block operator are again included in the set of eigenvalues of the single domain operator. The multi-block coupling was derived as a function of one semi-bounded parameter for which the discretization is both stable and conservative. Two different values of the parameter could be distinguished. One value which gives minimal interface dissipation, and another which gives a fully upwinded scheme. It was shown by several examples that the upwinded interface treatment is the preferred choice since it improves the errors, stiffness, and rate of convergence to steady-state. In the latter case, adding several interfaces can enhance the convergence rate to steady-state by several orders of magnitude.

4.4 Paper III

J. Berg and J. Nordström. Stable Robin solid wall boundary conditions for the Navier–Stokes equations. *Journal of Computational Physics, 230(19):7519–7532, 2011.*

There are multiple choices of well-posed solid wall boundary conditions for the compressible Navier–Stokes equations. The most commonly used ones are

the no-slip conditions for the velocity with an isothermal or adiabatic temperature condition. These boundary conditions make sure that there are no velocities in neither the normal nor tangential directions, and that the temperature or temperature gradient is specified. It is well-known that the no-slip conditions are accurate as long as the characteristic length scale is large enough. For flows on the micro or nano scale, molecular interactions have to be taken into account, and the Navier–Stokes equations do no longer give an accurate description of the physics. The effects of molecular dynamics can be modeled by slip-flow boundary conditions where the tangential velocities are allowed to be non-zero.

All of the above mentioned boundary conditions can be represented by Robin solid wall boundary conditions on the tangential velocity and temperature. This allows for a transition from no-slip to slip, and from isothermal to adiabatic, by varying parameters. We have proved that the SBP-SAT method can be made stable for all choices of parameters, using sharp energy estimates. All physically relevant solid wall boundary conditions for the compressible Navier–Stokes equations are thus contained within one uniform, energy stable, formulation.

4.5 Paper IV

J. Nordström and J. Berg. Conjugate heat transfer for the unsteady compressible Navier–Stokes equations using a multi-block coupling. *Accepted for publication in Computers & Fluids, 2012.*

There are two possible choices for how to compute conjugate heat transfer problems: 1) the Navier–Stokes equations are coupled to the heat equation, and 2) the Navier–Stokes equations themselves govern heat transfer in the solid. The first is the most obvious choice due to the simplicity of the scalar heat equation. The latter is common for incompressible fluids because the energy component is decoupled from the momentum, and reduces exactly to the heat equation for zero velocities. For compressible fluids, the latter choice is less explored since stability and accuracy become problematic.

We used a modified multi-block coupling, where only the temperature is coupled over the interface, and let the compressible Navier–Stokes equations govern heat transfer also in the solid region. In the continuous case, we showed how to scale and choose the coefficients in the energy component of the Navier–Stokes equations, so that it becomes as similar to the heat equation as possible. Well-posedness of the modified multi-block coupling was shown using energy estimates in a modified norm.

In the discrete case, we showed that the coupling can be made stable using the same modified norm. Computations using both approaches were performed and it was shown that the differences can be made very small.

4.6 Paper V

J. Berg and J. Nordström. Superconvergent functional output for time-dependent problems using finite differences on summation-by-parts form. *Journal of Computational Physics, 231(20):6846–6860, 2012.*

The theory of dual consistency and functional superconvergence for SBP-SAT discretizations was first derived for steady problems. In this paper, we extended the theory to time-dependent problems and related dual consistency to stability. We gave a general proof that dual consistency and stability implies superconvergence for linear (integral) functionals. Several model problems of different kinds were analyzed. It was shown how to construct schemes which are stable and dual consistent, and that superconvergence was obtained for all cases.

4.7 Paper VI

J. Berg and J. Nordström. On the impact of boundary conditions on dual consistent finite difference discretizations. *Accepted for publication in Journal of Computational Physics, 2012.*

In paper V, the model PDEs were supplied with Dirichlet boundary conditions to simplify the analysis in the continuous case. Dirichlet boundary conditions automatically ensures that both the primal and dual problems are well-posed. The discretization, however, became more complicated as it was required to reduce the equations to first-order form to derive stability conditions. In realistic applications, Dirichlet boundary conditions are rarely suitable at far-field boundaries. It is well-known that they give large reflections which eventually will pollute the whole solution unless exact boundary data is known. Other kind of boundary conditions can significantly enhance both the stability and accuracy of a numerical scheme.

We considered a linear incompletely parabolic system of PDEs in one dimension. The boundary conditions of far-field type were derived using energy

estimates, under the restriction that both the primal and dual problems were well-posed. By simultaneously considering the primal and dual problems, the amount of free parameters could be reduced, which allowed the construction of new advanced boundary conditions.

The equations were discretized using the SBP-SAT technique, and it was shown that the construction of the continuous boundary conditions are sufficient for both stability and dual consistency. In fact, with the new boundary conditions, stability and dual consistency are equivalent. Several computations were performed with the new boundary conditions, and it was shown that they provide both error boundedness in time, fast convergence to steady-state, and superconvergence of linear integral functionals.

5. Acknowledgements

Before anything, I would like to express my sincerest gratitudes to my supervisor Prof. Jan Nordström. Working with you has been a pleasure. The work was hard and extremely rewarding. Your dedication and interest in my research has been a source of never-ending inspiration. Getting manuscripts accepted in the peer review process was the easy part, getting them accepted by you—that was the challenge. Your devotion has extended far beyond the expectations of a supervisor. I have enjoyed our running, travels, and last, but certainly not least, a few beers in nice pubs.

The division of scientific computing has been a great place to work. The atmosphere was always friendly with an open discussion environment and a great attitude towards coffee breaks. Tom Smedsaas, Gunilla Kreiss, and Carina Lindgren, you have certainly done a good job in keeping the division running smoothly.

I have enjoyed the many activities together with my fellow Ph.D students. I wish you all the best in your future careers. In particular, I am grateful to Sofia Eriksson, not only for proofreading and many scientific discussions, but also for being a great friend. The conferences and travels with you have been truly enjoyable. Many thanks to Martin Tillenius for proofreading, but mainly for great friendship and lots of fun both at work and outside work. Furthermore, I would like to thank Sven-Erik Ekström, Pavol Bauer, Magnus Grandin, Stefan Hellander and many other for making work more than only work.

Thanks to my friends who have kept me sane during these years. In particular Johan Kullberg and the wednesday crew; Dan, David, Edvin, and Martin. Without the encouragement from Andreas Eklind, Magnus Isomäki-Krondahl, Anders With, and Mikael Larsson, I would not even have written this thing.

To my family. Thank you for all your support, not only during these years but during my whole life. You say that you are proud of having me as a son and brother, and I am equally proud of having you as my family.

Finally, to my dear wife Gita. Your existence and uniqueness makes my life well-posed. Thank you for making every day a great day and always being there for me.

Financial support for this work has been granted by NanoSpace AB, the Swedish Royal Academy of Sciences, and Anna Maria Lundins resestipendier.

6. Summary in Swedish

Stabila finita differensmetoder med hög noggrannhetsordning för multifysik- och flödesproblem

Många problem inom teknik och naturvetenskap, och även inom andra områden, kan modelleras med hjälp av partiella differentialekvationer (PDE:er). Några exempel är dynamiken hos fluider och elektromagnetisk vågutbredning, men även problem från aktiemarknaden och trafikstockningar kan beskrivas med PDE:er. I allmänhet finns inga generella metoder för att hitta exakta lösningar till dessa problem. Även i de fall där det finns exakta lösningar så är de i allmänhet för komplexa för att vara praktiskt användbara. Numeriska metoder är därför ofta nödvändiga.

Under de senaste årtiondena har datorerna utvecklats till den grad att numeriska simuleringar har etablerat sig som ett av vetenskapens fundament, likvärdigt med teori och experiment. Den ökade användningen av datorer är inte enbart tack vare att hårdvaran har blivit snabbare och effektivare. Algoritmerna som används för beräkningar har även de utvecklats i samma takt. Den ökade användningen av datorsimuleringar ställer höga krav på algoritmerna. Bra hårdvara är betydelselös om algoritmen som används är ineffektiv eller inte beräknar ett korrekt resultat.

Det finns en uppsjö av olika metoder för att lösa PDE:er. Några av de vanligaste är finita volymsmetoden, finita elementmetoden, och den som är huvudfokus i den här avhandlingen – finita differensmetoden. Vilken metod som än används så krävs det att metoden är;

1. Konsistent
2. Stabil
3. Effektiv

I allmänhet är två av de tre ovanstående kraven relativt lätta att åstadkomma. Att metoden är konsistent betyder att den faktiskt löser den PDE vi är intresserade av. Stabilitet betyder att störningar, t.ex. i data eller från diskretiserings- eller avrundningsfel, inte påverkar lösningen alltför mycket. Effektivitet betyder att metoden levererar en lösning inom rimlig tid. De flesta metoder som används är i praktiken är konsistenta. Däremot är långt ifrån alla metoder som används stabila och effektiva. En konsistent och stabil numerisk metod har ofta en låg noggrannhetsordning och är därmed ineffektiv, eftersom det krävs hög upplösning för ett noggrant resultat. En konsistent metod med hög noggrannhetsordning är ofta instabil.

Finita differensmetoder är i sitt grundutförande konsistenta och effektiva. Effektiviteten kommer av att det är lätt att åstadkomma hög noggrannhetsordning, samt att de lämpar sig väl för implementering på datorer. Ett stort problem är stabilitet. För att komma till rätta med stabilitetsproblemen har finita differensmetoder på partiell summationsform (eng. summation-by-parts, SBP) utvecklats. En SBP-operator är i grunden en central differensoperator som har modifierats för att vara ensidig vid ränderna. SBP-egenskapen i sig är tillräcklig för att varje linjärt välställt Cauchy-problem ska ha en stabil diskretisering.

För initial- och randvillkorsproblem (eng. initial boundary value problems, IBVP) är situationen lite mer komplicerad. De flesta PDE:er av fysikaliskt intresse, t.ex. Navier-Stokes ekvationer, kräver randvillkor för att vara väldefinierade. SBP-metodiken i sig har ingen hantering av randvillkor utan dessa måste läggas till separat. Den mest användbara metoden är att lägga till randvillkoren svagt, genom en så kallad SAT (eng. simultaneous approximation term). Tillsammans ger SBP-SAT ett ramverk för att konstruera konsistenta och stabila finita differensapproximationer av linjärt välställda IBVP, där noggrannhetsordningen inte är begränsad av stabilitetskrav.

Den här avhandlingen fokuserar på stabila och högre ordningens SBP-SAT-approximationer av olika IBVP som förekommer inom beräkningsfluiddynamik. I åtanke finns speciellt Navier-Stokes ekvationer samt multifysikproblem inklusive konjugerad värmeöverföring. Avhandlingen kan delas in i två delar. Den första delen består av artikel I–IV. I dessa utvecklas SBP-SAT-tekniken för kopplade problem samt randvillkorshantering för Navier–Stokes ekvationer. Det visas hur SBP-SAT används för att koppla ihop olika fysikmodeller och vilka egenskaper hos diskretiseringen som ändras vid kopplingen. Väggrandvillkor för Navier-Stokes ekvationer som leder till välställdhet för det kontinuerliga problemet härleds, tillsammans med stabilitet för det diskreta. Väggrandvillkoren är formulerade så att alla relevanta fysikaliska randvillkor, t.ex. no-slip, slip, isoterma och adiabatiska, finns representerade i en enhetlig formulering som är energistabil med skarpa energiuppskattningar.

I den andra delen, bestående av artikel V–VI, utvecklas SBP-SAT-tekniken för hantering av duala problem. Dualkonsistens relateras till stabilitet vilket resulterar i superkonvergenta linjära integralfunktionaler. Genom att samtidigt betrakta välställdhet hos det primära och duala problemet kan nya avancerade randvillkor härledas, vilka i en SBP-SAT-diskretisering direkt ger stabilitet och dualkonsistens, och därmed superkonvergenta funktionaler.

References

[1] I. Babuška and W. C. Rheinboldt. A-posteriori error estimates for the finite element method. *International Journal for Numerical Methods in Engineering*, 12(10):1597–1615, 1978.

[2] I. Babuška and W. C. Rheinboldt. Error estimates for adaptive finite element computations. *SIAM Journal on Numerical Analysis*, 15(4):736–754, 1978.

[3] M. H. Carpenter, D. Gottlieb, and S. 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] M. H. Carpenter, J. Nordström, and D. Gottlieb. A stable and conservative interface treatment of arbitrary spatial accuracy. *Journal of Computational Physics*, 148(2):341–365, 1999.

[5] M. H. Carpenter, J. Nordström, and D. Gottlieb. Revisiting and extending interface penalties for multi-domain summation-by-parts operators. *Journal of Scientific Computing*, 45(1-3):118–150, 2010.

[6] K. Eriksson and C. Johnson. An adaptive finite element method for linear elliptic problems. *Mathematics of Computation*, 50(182):361–383, 1988.

[7] K. J. Fidkowski and D. L. Darmofal. Review of output-based error estimation and mesh adaptation in computational fluid dynamics. *AIAA Journal*, 49(4):673–694, 2011.

[8] M. B. Giles. Stability analysis of numerical interface conditions in fluid-structure thermal analysis. *International Journal for Numerical Methods in Fluids*, 25:421–436, 1997.

[9] M. B. Giles, M. C. Duta, J.-D. Müller, and N. A. Pierce. Algorithm developments for discrete adjoint methods. *AIAA Journal*, 41(2):198–205, 2003.

[10] M. B. Giles, M. G. Larson, J. M. Levenstam, and E. Süli. Adaptive error control for finite element approximations of the lift and drag coefficients in viscous flow. Technical report, Report NA-97/06, Oxford University Computing Laboratory, 1997.

[11] M. B. Giles and N. A. Pierce. Adjoint equations in CFD: Duality, boundary conditions and solution behaviour. In *AIAA Paper 97-1850*, 1997.

[12] M. B. Giles and N. A. Pierce. On the properties of solutions of the adjoint Euler equations. *Numerical Methods for Fluid Dynamics VI ICFD*, pages 1–16, 1998.

[13] M. B. Giles and N. A. Pierce. An introduction to the adjoint approach to design. *Flow, Turbulence and Combustion*, 65:393–415, 2000.

[14] M. B. Giles and E. Süli. Adjoint methods for PDEs: A posteriori error analysis and postprocessing by duality. *Acta Numerica*, 11:145–236, 2002.

[15] B. Gustafsson, H.-O. Kreiss, and J. Oliger. *Time Dependent Problems and Difference Methods*. Wiley Interscience, 1995.

[16] J. Hadamard. *Lectures on Cauchy's problem in linear partial differential equations*. New Haven, 1923.

[17] W. D. Henshaw and K. K. Chand. A composite grid solver for conjugate heat transfer in fluid-structure systems. *Journal of Computational Physics*, 228(10):3708–3741, 2009.

[18] J. E. Hicken and D. W. Zingg. Superconvergent functional estimates from summation-by-parts finite-difference discretizations. *SIAM Journal on Scientific Computing*, 33(2):893–922, 2011.

[19] J. E. Hicken and D. W. Zingg. Summation-by-parts operators and high-order quadrature. *Journal of Computational and Applied Mathematics*, 237(1):111–125, 2013.

[20] X. Huan, J. E. Hicken, and D. W. Zingg. Interface and boundary schemes for high-order methods. In *the 39th AIAA Fluid Dynamics Conference, AIAA Paper No. 2009-3658*, San Antonio, USA, June 2009.

[21] A. Jameson. Aerodynamic design via control theory. *Journal of Scientific Computing*, 3:233–260, 1988.

[22] H.-O. Kreiss and J. Lorenz. *Initial-Boundary Value Problems and the Navier–Stokes Equations*. SIAM, 2004.

[23] H.-O. Kreiss and G. Scherer. Finite element and finite difference methods for hyperbolic partial differential equations. In *Mathematical Aspects of Finite Elements in Partial Differential Equations*. Academic Press, 1974.

[24] H.-O. Kreiss and G. Scherer. On the existence of energy estimates for difference approximations for hyperbolic systems. Technical report, Uppsala University, Division of Scientific Computing, 1977.

[25] H. O. Kreiss and L. Wu. On the stability definition of difference approximations for the initial boundary value problem. *Applied Numerical Mathematics*, 12(1–3):213–227, 1993.

[26] P. D. Lax and R. D. Richtmyer. Survey of the stability of linear finite difference equations. *Communications on Pure and Applied Mathematics*, 9(2):267–293, 1956.

[27] J. Lu. *An a posteriori error control framework for adaptive precision optimization using discontinuous Galerkin finite element method*. PhD thesis, Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, 2005.

[28] K. Mattsson. Boundary procedures for summation-by-parts operators. *Journal of Scientific Computing*, 18(1):133–153, 2003.

[29] K. Mattsson. Summation by parts operators for finite difference approximations of second-derivatives with variable coefficients. *Journal of Scientific Computing*, 51:650–682, 2012.

[30] K. Mattsson, F. Ham, and G. Iaccarino. Stable boundary treatment for the wave equation on second-order form. *Journal of Scientific Computing*, 41:366–383, 2009.

[31] K. Mattsson and J. Nordström. Summation by parts operators for finite difference approximations of second derivatives. *Journal of Computational Physics*, 199(2):503–540, 2004.

[32] K. Mattsson, M. Svärd, M. Carpenter, and J. Nordström. High-order accurate computations for unsteady aerodynamics. *Computers and Fluids*, 36(3):636–649, 2007.

[33] K. Mattsson, M. Svärd, and M. Shoeybi. Stable and accurate schemes for the

compressible Navier–Stokes equations. *Journal of Computational Physics*, 227:2293–2316, 2008.

[34] S. Nadarajah and A. Jameson. A comparison of the continuous and discrete adjoint approach to automatic aerodynamic optimization. In *AIAA 38th Aerospace Sciences Meeting and Exhibit, AIAA-2000-0667*, Reno, USA, Jan. 2000.

[35] S. Nadarajah and A. Jameson. Studies of the continuous and discrete adjoint approaches to viscous automatic aerodynamic shape optimization. In *AIAA 15th Computational Fluid Dynamics Conference, AIAA-2001-2530*, Anaheim, Canada, June 2001.

[36] J. Nordström and M. 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(2):621–645, 1999.

[37] J. Nordström and M. H. Carpenter. High-order finite difference methods, multidimensional linear problems, and curvilinear coordinates. *Journal of Computational Physics*, 173(1):149–174, 2001.

[38] J. Nordström, J. Gong, E. van der Weide, and M. Svärd. A stable and conservative high order multi-block method for the compressible Navier–Stokes equations. *Journal of Computational Physics*, 228(24):9020–9035, 2009.

[39] J. Nordström and M. Svärd. Well-posed boundary conditions for the Navier–Stokes equations. *SIAM Journal on Numerical Analysis*, 43(3):1231–1255, 2005.

[40] T. A. Oliver and D. L. Darmofal. Analysis of dual consistency for discontinuous Galerkin discretizations of source terms. *SIAM Journal on Numerical Analysis*, 47(5):3507–3525, 2009.

[41] P. Olsson. Summation by parts, projections, and stability. I. *Mathematics of Computation*, 64(211):1035–1065, 1995.

[42] P. Olsson. Summation by parts, projections, and stability. II. *Mathematics of Computation*, 64(212):1473–1493, 1995.

[43] F. Riesz and B. Szőkefalvi-Nagy. *Functional Analysis*. Dover Publications, 1990.

[44] W. Rudin. *Functional Analysis*. McGraw-Hill, 1991.

[45] M. Schäfer and I. Teschauer. Numerical simulation of coupled fluid-solid problems. *Computer Methods in Applied Mechanics and Engineering*, 190(28):3645–3667, 2001.

[46] B. Strand. Summation by parts for finite difference approximations for d/dx. *Journal of Computational Physics*, 110(1):47–67, 1994.

[47] J. C. Strikwerda. Initial boundary value problems for the method of lines. *Journal of Computational Physics*, 34(1):94–107, 1980.

[48] M. Svärd. On coordinate transformations for summation-by-parts operators. *Journal of Scientific Computing*, 20:29–42, 2004.

[49] M. Svärd, K. Mattsson, and J. Nordström. Steady-state computations using summation-by-parts operators. *Journal of Scientific Computing*, 24(1):79–95, 2005.

[50] M. Svärd and J. Nordström. On the order of accuracy for difference approximations of initial-boundary value problems. *Journal of Computational Physics*, 218(1):333–352, 2006.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1004

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology.

Distribution: publications.uu.se
urn:nbn:se:uu:diva-187204

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2013

Paper I

Journal of Computational Physics 229 (2010) 5440–5456

Contents lists available at ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp

A stable and high-order accurate conjugate heat transfer problem

Jens Lindström [a,*], Jan Nordström [a,b,c]

[a] Uppsala University, Department of Information Technology, 751 05, Uppsala, Sweden
[b] School of Mechanical, Industrial and Aeronautical Engineering, University of the Witvatersrand, PO WITS 2050, Johannesburg, South Africa
[c] FOI, The Swedish Defence Research Agency, Department of Aeronautics and Systems Integration, Stockholm, 164 90, Sweden

ARTICLE INFO

Article history:
Received 4 November 2009
Received in revised form 29 March 2010
Accepted 7 April 2010
Available online 13 April 2010

Keywords:
Conjugate heat transfer
Well-posedness
Stability
High-order accuracy
Summation-By-Parts
Weak boundary conditions

ABSTRACT

This paper analyzes well-posedness and stability of a conjugate heat transfer problem in one space dimension. We study a model problem for heat transfer between a fluid and a solid. The energy method is used to derive boundary and interface conditions that make the continuous problem well-posed and the semi-discrete problem stable. The numerical scheme is implemented using 2nd-, 3rd- and 4th-order finite difference operators on Summation-By-Parts (SBP) form. The boundary and interface conditions are implemented weakly. We investigate the spectrum of the spatial discretization to determine which type of coupling that gives attractive convergence properties. The rate of convergence is verified using the method of manufactured solutions.

1. Introduction

The coupling of fluid and heat equations is an area that has many interesting scientific and engineering applications. From the scientific side it is interesting to mathematically derive conditions to make the coupled system well-posed and compare with actual physics. The applications for conjugate heat transfer ranges between cooling of turbine blades, electronic components, nuclear reactors or spacecraft re-entry just to name a few. The particular application we are working towards here is a microscale satellite cold gas propulsion system with heat sources that will be used for controlling the flow rate [1]. See Fig. 1.

This paper is the first step of understanding the coupling procedure within our framework. The computational method that we are using is a finite difference method on Summation-By-Parts (SBP) form with the Simultaneous Approximation Term (SAT), a weak coupling at the fluid–solid interface. This method has been developed in many papers [2–7] and used for many difficult problems where it has proven to be robust [8–11]. The extensions to multiple dimensions is relatively straightforward once the one-dimensional case has been investigated. The difficulty in extending to multiple dimensions lies rather in a high performance implementation than in the theory.

The main idea of the SBP and SAT framework is that the difference operators should mimic integration by parts in the continuous case. This framework makes the discrete equations closely related to the PDE:s themselves. The difference operators are constructed such that they shift to one-sided close to the boundaries. This results in an energy estimate which gives stability for a well-posed Cauchy problem. The SAT method implements the boundary conditions weakly and an energy estimate, and hence stability, can be obtained for a well-posed initial boundary value problem.

* Corresponding author. Address: Division of Scientific Computing, Department of Information Technology, Uppsala University, Box 337, SE-751 05, Uppsala, Sweden. Tel.: +46 18 471 6253; fax: +46 18 523049/+46 18 511925.
E-mail address: jens.lindstrom@it.uu.se (J. Lindström).

0021-9991/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2010.04.010

Fig. 1. A micro machined nozzle with 3 heater coils positioned just before the nozzle throat. The nozzle throat is approximately 30 μm in a heat exchange chamber.

Since the operators shift to one-sided close to boundaries and interfaces there is no need to introduce ghost points or extrapolate values which in general causes stability issues. Once the scheme is correctly written and all coefficients determined the order of the scheme depends only on the order of the difference operators. In this paper we will present 2nd-, 3rd- and 4th-order operators and study their performance. The details of these operators can be found in for example [2,3,12].

2. The continuous problem

The equations we are studying in this paper are motivated by a gas flow in a long channel with heat sources. The channel is long compared to the height and hence the changes in the tangential direction are small in comparison to the changes in the normal direction, see Fig. 2.

The equations are an incompletely parabolic system of equations for the flow and the scalar heat equation for the heat transfer,

$$w_t + Aw_x = \varepsilon B w_{xx}, \quad -1 \leqslant x \leqslant 0 \tag{1}$$

and

$$T_t = kT_{xx}, \quad 0 \leqslant x \leqslant 1, \tag{2}$$

where

$$w = \begin{bmatrix} \rho \\ u \\ \mathcal{T} \end{bmatrix}, \quad A = \begin{bmatrix} a & b & 0 \\ b & a & c \\ 0 & c & a \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \alpha & 0 \\ 0 & 0 & \beta \end{bmatrix}. \tag{3}$$

We can view (1) as the Navier–Stokes equations linearized and symmetrized around a constant state. In that case we would have

$$a = \bar{u}, \quad b = \frac{\bar{c}}{\sqrt{\gamma}}, \quad c = \bar{c}\sqrt{\frac{\gamma - 1}{\gamma}}, \quad \alpha = \frac{\lambda + 2\mu}{\bar{\rho}}, \quad \beta = \frac{\gamma\mu}{Pr\bar{\rho}}, \tag{4}$$

where $\bar{u}, \bar{\rho}$ and $\bar{c}$ is the mean velocity, density and speed of sound. γ is the ratio of specific heats, Pr the Prandtl number and λ and μ are the second and dynamic viscosities, [8,13,14]. At this point the only assumption on the coefficients is that $\alpha, \beta > 0$.

Our main objective is to couple (1) and (2) at $x = 0$ and investigate which boundary and interface conditions that will lead to a well-posed coupled system.

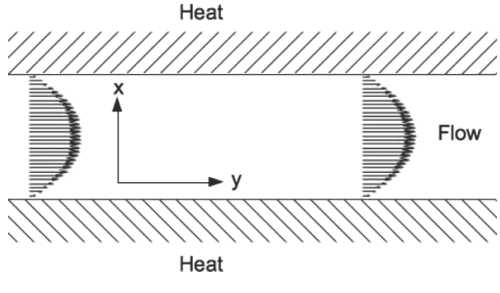


Fig. 2. By assuming an infinitely long channel with homogenicity in the tangential direction y we get an one-dimensional problem in the normal direction x for the conjugate heat transfer problem.

To simplify, we assume for the rest of the paper that $a > 0$. We are allowed to use three boundary conditions at $x = -1$, three interface conditions at $x = 0$ and one boundary condition at $x = 1$. See e.g. [8,9,13,15].

2.1. Boundary conditions at $x = -1$

The boundary and interface conditions will be derived using the energy method. Define the energy norm of w as

$$\|w\|^2 = \int_{-1}^{0} w^T w dx. \tag{5}$$

By multiplying (1) with w^T and integrating over the domain we get

$$\|w\|_t^2 = -w^T A w\big|_{-1}^{0} + 2\varepsilon w^T B w_x\big|_{-1}^{0} - 2\varepsilon \int_{-1}^{0} w_x^T B w_x dx. \tag{6}$$

Let

$$X = \frac{1}{\sqrt{2}d}\begin{bmatrix} -\sqrt{2}c & b & b \\ 0 & d & d \\ \sqrt{2}b & c & c \end{bmatrix}, \quad d = \sqrt{b^2 + c^2}, \tag{7}$$

be the diagonalizing matrix of A. We have $X^{-1} = X^T$ and $A = X\Lambda X^T$ where

$$\Lambda = \begin{bmatrix} a & 0 & 0 \\ 0 & a+d & 0 \\ 0 & 0 & a-d \end{bmatrix}, \tag{8}$$

contains the eigenvalues of A. Using these relations we can write (6) as

$$\|w\|_t^2 = (X^T w)^T \Lambda (X^T w) - 2\varepsilon w^T B w_x - 2\varepsilon \int_{-1}^{0} w_x^T B w_x dx, \tag{9}$$

where all boundary terms are evaluated at $x = -1$. We make the change of variables

$$X^T w = \frac{1}{\sqrt{2}d}\begin{bmatrix} -\sqrt{2}c\rho + \sqrt{2}b\mathcal{T} \\ b\rho + du + c\mathcal{T} \\ b\rho - du + c\mathcal{T} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}, \tag{10}$$

which are the characteristic variables for the hyperbolic part, cf. [13,15]. In order to bound the energy for the hyperbolic part we need to put boundary conditions on the characteristic variables that are related to the positive eigenvalues of A. If we assume that $a < d$ which corresponds to subsonic inflow, then A has two positive eigenvalues and we need to use two boundary conditions on the corresponding characteristic variables. Thus we need to impose the boundary conditions

$$c_1 = \frac{1}{\sqrt{2}d}\left(-\sqrt{2}c\rho + \sqrt{2}b\mathcal{T}\right) = f_1(t), \tag{11}$$

$$c_2 = \frac{1}{\sqrt{2}d}(b\rho + du + c\mathcal{T}) = f_2(t), \tag{12}$$

to bound the hyperbolic part.

We are allowed to use one more boundary conditions that will need to bound the parabolic term $-2\varepsilon w^T B w_x$. Assume $f_1 = f_2 = 0$. By taking linear combinations of (11) and (12) we can eliminate ρ and obtain

$$cu + d\mathcal{T} = 0. \tag{13}$$

The parabolic term is expanded using relation (13) to obtain

$$-2\varepsilon w^T B w_x = -2\varepsilon u\left(\alpha u_x - \frac{\beta c}{d}\mathcal{T}_x\right). \tag{14}$$

If we put

$$\alpha d u_x - \beta c \mathcal{T}_x = f_3(t), \tag{15}$$

as the final boundary condition for (1) at $x = -1$, then with $f_3 = 0$ the parabolic term (14) is zero and all the boundary terms are bounded.

Remark 2.1. The assumption of zero boundary data is necessary to obtain Eq. (15). If we could have bounded the left boundary terms with non-zero boundary data, it could lead to a strongly well-posed problem [16].

2.2. Boundary conditions at x = 1

At $x = 1$ we have the scalar heat equation. By applying the energy method we get

$$\|T\|_t^2 = 2kTT_x - 2k\|T_x\|^2, \tag{16}$$

from which it is easy to see that either

$$T = h_1(t), \quad T_x = h_2(t) \quad \text{or} \quad \alpha_1 T + \beta_1 T_x = h_3(t), \tag{17}$$

will result in an energy estimate (for suitable choices of the constants α_1 and β_1). In the rest of the paper and in the numerical experiments we have used $T = h_1(t)$.

2.3. Interface conditions at x = 0

At the interface we apply the energy method to both equations and add them together to get (when ignoring boundary terms)

$$\frac{d}{dt}\left(\|w\|^2 + \|T\|^2\right) = -w^T A w + 2\varepsilon w^T B w_x - 2kTT_x - 2\varepsilon \int_{-1}^{0} w_x^T B w_x \, dx - 2k \int_0^1 T_x^2 \, dx. \tag{18}$$

Since we are considering the interface as a solid wall which separates the fluid from the solid and since we want a continuous heat transfer we impose

$$u = 0, \quad \mathcal{T} = T. \tag{19}$$

Using the interface conditions (19), Eq. (18) reduces to

$$\frac{d}{dt}\left(\|w\|^2 + \|T\|^2\right) = -a(\rho^2 + \mathcal{T}^2) + 2\mathcal{T}(\beta\varepsilon\mathcal{T}_x - kT_x) - 2\varepsilon \int_{-1}^{0} w_x^T B w_x \, dx - 2k \int_0^1 T_x^2 \, dx \tag{20}$$

and we can easily see that if we impose

$$\beta\varepsilon\mathcal{T}_x - kT_x = 0, \tag{21}$$

as the final interface condition we get an energy estimate. Without (21), the interface can act as an unphysical heat source.

Using all these boundary and interface conditions we can conclude the following.

Proposition 2.1. *Eqs. (1) and (2) coupled at x = 0 are well-posed with boundary conditions (11), (12), (15) and (17) and interface conditions (19) and (21).*

Remark 2.2. Note that in arriving at Proposition 2.1 we have assumed that the data is identically zero. If we had been able to obtain an energy estimate for non-zero data the problem would have been strongly well-posed [16].

3. The semi-discrete problem

Eq. (1) is discretized on the single domain [-1,0] on a uniform grid of $M + 1$ grid points. The vector $\mathbf{w} = [w_0, w_1, \ldots, w_M]^T = [\rho_0, u_0, \mathcal{T}_0, \rho_1, u_1, \mathcal{T}_1, \ldots, \rho_M, u_M, \mathcal{T}_M]^T$ is the discrete approximation of w. The derivatives are approximated by the operators on SBP form

$$\mathbf{w}_x \approx \left(D_1^L \otimes I_3\right)\mathbf{w} = \left(P_L^{-1} Q_L \otimes I_3\right)\mathbf{w}, \tag{22}$$

$$\mathbf{w}_{xx} \approx \left(D_2^L \otimes I_3\right)\mathbf{w} = \left(P_L^{-1} Q_L \otimes I_3\right)^2 \mathbf{w}, \tag{23}$$

where P_L is a symmetric positive definite matrix and Q_L is an almost skew symmetric matrix satisfying $Q_L + Q_L^T = B_L = \mathrm{diag}(-1, 0, \ldots, 0, 1)$ [2,3]. I_3 is the 3×3 identity matrix. Eq. (2) is similarly discretized on a uniform grid of $N + 1$ grid points.

Remark 3.1. The approximation (23) has the drawback that the computational stencil is wide. This is however necessary for variable coefficients. Compact formulations that uses minimal bandwidth does however exist for constant coefficient problems [3].

In (22) and (23) we have introduced the Kronecker product, defined as

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix} \tag{24}$$

for the $m \times n$ and $p \times q$ matrices A and B respectively. It is a special case of a tensor product so it is bilinear and associative. Some of its important properties are

$$(A \otimes B)(C \otimes D) = (AC \otimes BD), \tag{25}$$
$$(A \otimes B)^{-1} = (A^{-1} \otimes B^{-1}), \tag{26}$$

if the usual matrix products and inverses are defined.

Given a partial differential equation,

$$\begin{aligned} &v_t = \mathcal{P}(x,t,v), \quad x \in \Omega,\ t \geqslant 0, \\ &v(x,0) = f(x), \quad x \in \Omega,\ t = 0, \\ &Lv = g(t), \quad x \in \partial\Omega,\ t \geqslant 0, \end{aligned} \tag{27}$$

the SAT method will be used to implement the boundary condition $Lv = g$ weakly. This means that $\mathbf{Lv} - \mathbf{g} = \mathcal{O}(h^p)$ in the discrete case. The discretization of (27) using the SAT method would schematically look like

$$\mathbf{v}_t = \mathscr{D}\mathbf{v} + (P^{-1}E \otimes \Sigma)(\mathbf{Lv} - \mathbf{g}), \tag{28}$$

where $\mathscr{D}$ is a discrete SBP approximation of $\mathcal{P}$ and $\mathbf{L}$ is a matrix that approximates the continuous operator L. E is a matrix which picks the correct boundary terms at the correct positions in space. Σ is an unknown matrix of the same size as the system of PDE:s to be determined for stability.

With these tools and the boundary and interface conditions derived in Proposition 2.1 we can discretize (1) and (2) using the SAT method as

$$\begin{aligned} \mathbf{w}_t = &-\left(D_1^L \otimes A\right)\mathbf{w} + \varepsilon\left(D_2^L \otimes B\right)\mathbf{w} + \left(P_L^{-1}E_0^L \otimes \Sigma_1^0\right)\left(X^T w_0 - g_1^0\right) + \left(P_L^{-1}E_0^L \otimes \Sigma_3^0\right)\left(\alpha d\left(D_1^L u\right)_0 - \beta c\left(D_1^L \mathcal{T}\right)_0 - g_3^0\right) \\ &+ \left(P_L^{-1}\left(D_1^L\right)^T E_0^L \otimes \Sigma_5^0\right)(cu_0 + d\mathcal{T}_0 - g_5^0) + \left(P_L^{-1}E_M^L \otimes \Sigma_1^M\right)(w_M - g_1^M) + \left(P_L^{-1}E_M^L \otimes \Sigma_2^M\right)(w_M - g_1^M) \\ &+ \left(P_L^{-1}E_M^L \otimes \Sigma_3^M\right)(\mathcal{T}_M - T_0) + \left(P_L^{-1}\left(D_1^L\right)^T E_M^L \otimes \Sigma_4^M\right)(\mathcal{T}_M - T_0) + \left(P_L^{-1}E_M^L \otimes \Sigma_5^M\right)\left(\beta\varepsilon\left(D_1^L \mathcal{T}\right)_M\right. \\ &\left. -k\left(D_1^R T\right)_0\right) - \left(P_L^{-1} \otimes I_3\right)\left(\widetilde{D}_L^T \widetilde{B}_L \widetilde{D}_L \otimes I_3\right), \end{aligned} \tag{29}$$

$$\begin{aligned} \mathbf{T}_t = &\, kD_2^R\mathbf{T} + \tau_1^0 P_R^{-1}E_0^R(T_0 - \mathcal{T}_M) + \tau_2^0 P_R^{-1}\left(D_1^R\right)^T E_0^R(T_0 - \mathcal{T}_M) + \tau_3^0 P_R^{-1}E_0^R\left(k\left(D_1^R T\right)_0 - \beta\varepsilon\left(D_1^L \mathcal{T}\right)_M\right) \\ &+ \tau_1^N P_R^{-1}E_N^R\left(T_N - h_1^N\right) - P_R^{-1}\widetilde{D}_R^T \widetilde{B}_R \widetilde{D}_R. \end{aligned} \tag{30}$$

The matrices $E_0^L = \mathrm{diag}(1,0,\ldots,0), E_M^L = \mathrm{diag}(0,\ldots,0,1)$ and $E_{0,N}^R$ similarly defined, are used to select boundary elements. The 3×3 matrices $\Sigma_i^{0,M}$ and coefficients $\tau_i^{0,N}$ are called penalty matrices and penalty coefficients which have to be determined for stability [2–4]. All $g_i^{0,M}$ and h_1^N are arbitrary boundary data, except for $g_5^0 = \frac{bg_1^0 + \sqrt{2}cg_3^0}{\sqrt{2}d}$ which was derived as a linear combination of the other boundary conditions.

Remark 3.2. In (29) we have $X^T w_0 - g_1^0 = [c_1 - f_1, c_2 - f_2, c_3 - f_3]^T$ where c_1, c_2 and c_3 are the characteristic variables. Moreover $w_M - g_1^M = [\rho_M - g_1, u_M - g_2, \mathcal{T}_M - g_3]^T$. The rest of the SAT boundary and interface terms are 3×1 vectors with the scalar values given on each row. The penalty matrices are constructed such that they select the correct entries and cancels the rest.

The terms $\widetilde{D}_{L,R}^T \widetilde{B}_{L,R} \widetilde{D}_{L,R}$ are artificial dissipation operators which reduce spurious oscillations. The matrices $\widetilde{D}_{L,R}$ are undivided forward or backward difference operators and $B_{L,R}$ are diagonal matrices which make the dissipation operator symmetric and determines the amount and location of the dissipation. In this case we have for 2nd-order the dissipation operators

$$\widetilde{D}_{L,R} = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -1 & 1 \\ 0 & \cdots & 0 & 0 & -1 \end{bmatrix}, \quad \widetilde{B}_{L,R} = \mathrm{diag}(\gamma_{L,R}, \gamma_{L,R}, \ldots, \gamma_{L,R}, 0)$$

$$\widetilde{D}_{L,R}^T \widetilde{B}_{L,R} \widetilde{D}_{L,R} = \gamma_{L,R} \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -1 & 2 & -1 \\ 0 & \cdots & 0 & 0 & -1 & 1 \end{bmatrix}, \tag{31}$$

where $\gamma_{L,R}$ is a positive parameter determining the amount of dissipation. These operators lead to an energy estimate and does not reduce the order of the scheme. An extensive study of these dissipation operators can be found in [12].

3.1. Stability conditions at $x = -1$

We will use the discrete analogue of the energy method to show that the scheme is stable. Define the discrete energy norm

$$\|\mathbf{w}\|_{P_L}^2 = \mathbf{w}^T (P_L \otimes I_3) \mathbf{w}, \tag{32}$$

where I_3 is the 3×3 identity matrix. Omit all terms which are not related to the left boundary, and multiply (29) with $\mathbf{w}^T(P \otimes I_3)$. Since D_1^L and D_2^L are on SBP form we obtain after some algebra

$$\begin{aligned} \frac{d}{dt}\|\mathbf{w}\|_{P_L}^2 = {} & w_0^T A w_0 - 2\varepsilon w_0^T B \left(D_1^L w\right)_0 - 2\varepsilon \left(D_1^L \mathbf{w}\right)^T (I_{N+1} \otimes B) \left(D_1^L \mathbf{w}\right) + 2 w_0^T \Sigma_1^0 \left(X^T w_0 - g_1^0\right) \\ & + 2 w_0^T \Sigma_3^0 \left(\alpha d \left(D_1^L u\right)_0 - \beta c \left(D_1^L \mathcal{T}\right)_0 - g_3^0\right) + 2 \left(D_1^L w\right)_0^T \Sigma_5^0 (c u_0 + d \mathcal{T}_0 - g_5^0). \end{aligned} \tag{33}$$

As in the continuous case we let $g_1^0 = g_3^0 = g_5^0 = 0$ and consider the hyperbolic and parabolic parts separately.

The hyperbolic part with the corresponding penalty term is

$$w_0^T A w_0 + 2 w_0^T \Sigma_1^0 X^T w_0. \tag{34}$$

By diagonalizing A and make a change of variables in the same way as in the continuous case we obtain that with

$$\Sigma_1^0 = \frac{1}{\sqrt{2}d} \begin{bmatrix} -\sqrt{2}c\sigma_1^0 & b\sigma_2^0 & 0 \\ 0 & d\sigma_2^0 & 0 \\ \sqrt{2}b\sigma_1^0 & c\sigma_2^0 & 0 \end{bmatrix}, \tag{35}$$

where

$$a + 2\sigma_1^0 \leqslant 0, \quad a + d + 2\sigma_2^0 \leqslant 0, \tag{36}$$

the hyperbolic part is bounded.

The parabolic part with the corresponding penalty terms is

$$-2\varepsilon w_0^T B \left(D_1^L w\right)_0 + 2 w_0^T \Sigma_3^0 \left(\alpha d \left(D_1^L u\right)_0 - \beta c \left(D_1^L \mathcal{T}\right)_0\right) + 2 \left(D_1^L w\right)_0^T \Sigma_5^0 (c u_0 + d \mathcal{T}_0) \tag{37}$$

and again we have to choose Σ_3^0 and Σ_5^0 such that (37) is negative semi-definite. Let

$$\Sigma_3^0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \varepsilon\sigma_3^0 & 0 \\ 0 & 0 & \varepsilon\sigma_4^0 \end{bmatrix}, \quad \Sigma_5^0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \varepsilon\sigma_5^0 & 0 \\ 0 & 0 & \varepsilon\sigma_6^0 \end{bmatrix}. \tag{38}$$

We formulate (37) as a quadratic form $\varepsilon v_0^T M_0 v_0$ with $v_0 = [u_0, (D_1^L u)_0, \mathcal{T}_0, (D_1^L \mathcal{T})_0]^T$ and

$$M_0 = \begin{bmatrix} 0 & -\alpha + \alpha d\sigma_3^0 + c\sigma_5^0 & 0 & -\beta c\sigma_3^0 + c\sigma_6^0 \\ -\alpha + \alpha d\sigma_3^0 + c\sigma_5^0 & 0 & \alpha d\sigma_4^0 + d\sigma_5^0 & 0 \\ 0 & \alpha d\sigma_4^0 + d\sigma_5^0 & 0 & -\beta - \beta c\sigma_4^0 + d\sigma_6^0 \\ -\beta c\sigma_3^0 + c\sigma_6^0 & 0 & -\beta - \beta c\sigma_4^0 + d\sigma_6^0 & 0 \end{bmatrix}. \tag{39}$$

In order for (37) to be negative semi-definite we need to choose the coefficients σ_i^0 such that $M_0 \leqslant 0$. Since all diagonal entries of M_0 is zero, all other entries must also be zero. This results in a system of equations with one parameter family of solutions

$$r \in \mathbb{R}, \quad \sigma_3^0 = \frac{1 + cr}{d}, \quad \sigma_4^0 = r, \quad \sigma_5^0 = -\alpha r, \quad \sigma_6^0 = \frac{\beta(1 + cr)}{d}. \tag{40}$$

The arbitrary parameter r will later be used in the analysis of the discrete spectrum when we study convergence and stiffness properties of the discretization. With these choices $M_0 = 0$ and we obtain an energy estimate and hence the left boundary is stable.

3.2. Stability conditions at x = 1

Consider the semi-discrete scheme (30) at $x = 1$, where all interface terms have been neglected,

$$\mathbf{T}_t = kD_2^R\mathbf{T} + \tau_1^N P_R^{-1} E_N^R\left(T_N - h_1^N\right). \tag{41}$$

By assuming $h_1^N = 0$ and multiplying with $\mathbf{T}^{\mathbf{T}}P_R$ we get (when ignoring interface terms)

$$\frac{d}{dt}\|\mathbf{T}\|_{P_R}^2 = 2kT_N\left(D_1^R T\right)_N + 2\tau_1^N T_M^2 - 2k\left(D_1^R\mathbf{T}\right)^T P_R\left(D_1^R\mathbf{T}\right). \tag{42}$$

Define p_N^R as the last entry on the diagonal of P_R, that is $p_N^R = P_R^{(N,N)}$. Then (42) is bounded by choosing

$$\tau_1^N \leqslant \frac{-k}{4p_N^R}. \tag{43}$$

This means that τ_1^N is proportional to $\frac{1}{\Delta x}$, and in particular we have $\frac{k}{4p_N^R} = \frac{k}{2\Delta x}, \frac{12k}{17\Delta x}, \frac{10800k}{13649\Delta x}$ for 2nd-, 3rd- and 4th-order operators respectively. This technique is discussed in e.g. [5,6].

3.3. Stability conditions at x = 0

At $x = 0$ we have the two interface schemes

$$\begin{aligned}\mathbf{w}_t = &-\left(D_1^L\otimes A\right)\mathbf{w} + \varepsilon\left(D_2^L\otimes B\right)\mathbf{w} + \left(P^{-1}E_M^L\otimes\Sigma_1^M\right)(w_M - g_1^M) + \left(P^{-1}E_M^L\otimes\Sigma_2^M\right)(w_M - g_1^M)\\ &+\left(P^{-1}E_M^L\otimes\Sigma_3^M\right)(\mathcal{T}_M - T_0) + \left(P^{-1}\left(D_1^L\right)^T E_M^L\otimes\Sigma_4^M\right)(\mathcal{T}_M - T_0)\\ &+\left(P^{-1}E_M^L\otimes\Sigma_5^M\right)\left(\beta\varepsilon\left(D_1^L\mathcal{T}\right)_M - k\left(D_1^R T\right)_0\right),\end{aligned} \tag{44}$$

$$\mathbf{T}_t = kD_2^R\mathbf{T} + \tau_1^0 P_R^{-1}E_0^R(T_0 - \mathcal{T}_M) + \tau_2^0 P_R^{-1}\left(D_1^R\right)^T E_0^R(T_0 - \mathcal{T}_M) + \tau_3^0 P_R^{-1}E_0^R\left(k\left(D_1^R T\right)_0 - \beta\varepsilon(D_1^L\mathcal{T})_M\right). \tag{45}$$

The penalty terms related to the outer boundaries are omitted.

A formulation which clearly shows the coupled system can be written

$$\begin{aligned}\begin{bmatrix}\mathbf{w}\\ \mathbf{T}\end{bmatrix}_t = &\begin{bmatrix}D_1^L\otimes(-A) & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\mathbf{w}\\ \mathbf{T}\end{bmatrix} + \begin{bmatrix}D_1^L\otimes\epsilon B & 0\\ 0 & kD_2^R\end{bmatrix}\begin{bmatrix}\mathbf{w}\\ \mathbf{T}\end{bmatrix} + \overline{P}^{-1}\underbrace{\begin{bmatrix}e_M^L\otimes\tilde{\Sigma}_3^M\\ -\tau_1^0 e_0^R\end{bmatrix}\begin{bmatrix}e_M^L\otimes f_3\\ -e_0^R\end{bmatrix}^T}_{J_1}\begin{bmatrix}\mathbf{w}\\ \mathbf{T}\end{bmatrix}\\ &+\overline{P}^{-1}\underbrace{\begin{bmatrix}(D_1^L)^T e_M^L\otimes\tilde{\Sigma}_4^M\\ -\tau_2^0(D_1^R)^T e_0^R\end{bmatrix}\begin{bmatrix}e_M^L\otimes f_3\\ -e_0^R\end{bmatrix}^T}_{J_2}\begin{bmatrix}\mathbf{w}\\ \mathbf{T}\end{bmatrix} + \overline{P}^{-1}\underbrace{\begin{bmatrix}e_M^L\otimes\tilde{\Sigma}_5^M\\ -\tau_3^0 e_0^R\end{bmatrix}\begin{bmatrix}\beta\epsilon(D_1^L\otimes I_3)^T(e_M^L\otimes f_3)\\ -k(D_1^R)^T e_0^R\end{bmatrix}^T}_{J_3}\begin{bmatrix}\mathbf{w}\\ \mathbf{T}\end{bmatrix},\end{aligned} \tag{46}$$

where

$$\overline{P}^{-1} = \begin{bmatrix}P_L^{-1}\otimes I_3 & 0\\ 0 & P_R^{-1}\end{bmatrix}, \quad \tilde{\Sigma}_i^M = [0,0,\sigma_i^M]^T, \quad f_3 = [0,0,1]^T. \tag{47}$$

The interface matrices J_i are sparse with entries only close to the interface. For 2nd-order difference operators they are

$$J_1 = \begin{bmatrix}0 & \cdots & \cdots & & 0\\ \vdots & & \vdots & \vdots & & \vdots\\ & \cdots & \sigma_3^M & -\sigma_3^M & \cdots & \\ & \cdots & -\tau_1^0 & \tau_1^0 & \cdots & \\ \vdots & & \vdots & \vdots & & \vdots\\ 0 & & \cdots & \cdots & & 0\end{bmatrix}, \tag{48}$$

$$J_2 = \begin{bmatrix} 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \cdots & -\frac{\sigma_4^M}{\Delta x_L} & \frac{\sigma_4^M}{\Delta x_L} & \cdots \\ \cdots & 0 & 0 & \cdots \\ \cdots & 0 & 0 & \cdots \\ \cdots & \frac{\sigma_4^M}{\Delta x_L} & -\frac{\sigma_4^M}{\Delta x_L} & \cdots \\ \cdots & -\frac{\tau_2^0}{\Delta x_R} & \frac{\tau_2^0}{\Delta x_L} & \cdots \\ \cdots & \frac{\tau_2^0}{\Delta x_R} & -\frac{\tau_2^0}{\Delta x_L} & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 \end{bmatrix}, \tag{49}$$

$$J_3 = \begin{bmatrix} 0 & & & \cdots & \cdots & & & & 0 \\ \vdots & & & \vdots & \vdots & & & & \vdots \\ \cdots & -\frac{\sigma_5^M \beta\epsilon}{\Delta x_L} & 0 & \frac{\sigma_5^M \beta\epsilon}{\Delta x_L} & -\frac{\sigma_5^M \beta\epsilon}{\Delta x_L} & \frac{\sigma_5^M \beta\epsilon}{\Delta x_L} & \cdots & \\ \cdots & -\frac{\tau_3^0 k}{\Delta x_R} & 0 & \frac{\tau_3^0 k}{\Delta x_R} & -\frac{\tau_3^0 k}{\Delta x_R} & \frac{\tau_3^0 k}{\Delta x_R} & \cdots & \\ \vdots & & & \vdots & \vdots & & & & \vdots \\ 0 & & & \cdots & \cdots & & & & 0 \end{bmatrix}. \tag{50}$$

By letting $g_1^M = 0$, applying the energy method to both equations and adding together we get (when ignoring the outer boundary terms)

$$\begin{aligned} \frac{d}{dt}\left(\|\mathbf{w}\|_{P_L}^2 + \|\mathbf{T}\|_{P_R}^2\right) &= -w_M^T A w_M + 2\varepsilon w_M^T B\left(D_1^L w\right)_M - 2\varepsilon\left(D_1^L \mathbf{w}\right)^T (I_N \otimes B)\left(D_1^L \mathbf{w}\right) + 2w_M^T \Sigma_1^M w_M + 2w_M^T \Sigma_2^M w_M \\ &\quad + 2w_M^T \Sigma_3^M (\mathcal{T}_M - T_0) + \left(D_1^L w\right)_N^T \Sigma_4^M (\mathcal{T}_M - T_0) + 2w_M^T \Sigma_5^M \left(\beta\varepsilon\left(D_1^L w\right)_M - k\left(D_1^R T\right)_0\right) \\ &\quad - 2kT_0\left(D_1^R T\right)_0 - 2k\left(D_1^R \mathbf{T}\right)^T P_R\left(D_1^R \mathbf{T}\right) + 2\tau_1^0 T_0(T_0 - \mathcal{T}_M) + 2\tau_2^0\left(D_1^R T\right)_0 (T_0 - \mathcal{T}_M) \\ &\quad + 2\tau_3^0 T_0\left(k\left(D_1^R T\right)_0 - \beta\varepsilon (D_1^L \mathcal{T})_M\right). \end{aligned} \tag{51}$$

As in the continuous case we have the hyperbolic part with the corresponding penalty term

$$-w_M^T A w_M + 2w_M^T \Sigma_1^M w_M = w_M^T \left(\underbrace{-A + 2\Sigma_1^M}_{M_H}\right) w_M, \tag{52}$$

which we want to bound by making M_H negative semi-definite. Note that A is symmetric by assumption. By choosing

$$\Sigma_1^M = \begin{bmatrix} 0 & \sigma_1^H & 0 \\ 0 & \sigma_2^H & 0 \\ 0 & \sigma_3^H & 0 \end{bmatrix}, \tag{53}$$

we can explicitly compute the eigenvalues of M_H and see that with

$$\sigma_1^H = \frac{b}{2}, \quad \sigma_2^H \leqslant 0, \quad \sigma_3^H = \frac{c}{2}, \tag{54}$$

we have $M_H \leqslant 0$. Note that Σ_1^M acts on u only.

The parabolic part is split into parts containing u and $\mathcal{T}$ separately. For the interface condition on u at $x = 0$ we get by expanding (51)

$$2\alpha\varepsilon u_M (D_1^L u)_M + 2w_M^T \Sigma_2^M w_M - 2\alpha\varepsilon (D_1^L \mathbf{u})^T P_L (D_1^L \mathbf{u}). \tag{55}$$

We choose

$$\Sigma_2^M = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \sigma_2^M & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{56}$$

and rewrite (55) as

$$2\alpha\varepsilon u_M (D_1^L u)_M + 2\sigma_2^M u_M^2 - 2\alpha\varepsilon \|D_1^L \mathbf{u}\|_{P_L}^2. \tag{57}$$

This expression is bounded by choosing

$$\sigma_2^M \leqslant \frac{-\alpha\varepsilon}{4p_M^L}, \tag{58}$$

where p_M^L is defined analogously to p_N^R in (43).

The remaining terms are used for coupling the two equations. Let the penalty matrices have the form

$$\Sigma_3^M = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \sigma_3^M \end{bmatrix}, \quad \Sigma_4^M = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \sigma_4^M \end{bmatrix}, \quad \Sigma_5^M = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \sigma_5^M \end{bmatrix} \tag{59}$$

and expand the remaining terms. This gives us the expression

$$\begin{aligned} &2\beta\varepsilon \mathcal{T}_M \left(D_1^L \mathcal{T}\right)_M - 2\beta\varepsilon \left(D_1^L \mathcal{T}\right)^T P_L \left(D_1^L \mathcal{T}\right) + 2\sigma_3^M \mathcal{T}_M (\mathcal{T}_M - T_0) + 2\sigma_4^M \left(D_1^L \mathcal{T}\right)_M (\mathcal{T}_M - T_0) \\ &\quad + 2\sigma_5^M \mathcal{T}_M \left(\beta\varepsilon \left(D_1^L \mathcal{T}\right)_M - k\left(D_1^R T\right)_0\right) - 2kT_0 \left(D_1^R T\right)_0 - 2k\left(D_1^R \mathbf{T}\right)^T P_R \left(D_1^R \mathbf{T}\right) + 2\tau_1^0 T_0 (T_0 - \mathcal{T}_M) \\ &\quad + 2\tau_2^0 \left(D_1^R T\right)_0 (T_0 - \mathcal{T}_M) + 2\tau_3^0 T_0 \left(k\left(D_1^R T\right)_0 - \beta\varepsilon \left(D_1^L \mathcal{T}\right)_M\right), \end{aligned} \tag{60}$$

which we need to bound by choosing appropriate penalty coefficients. Expression (60) can be written in matrix form as

$$v_I^T M_I v_I - 2\beta\varepsilon \left\| D_1^L \mathscr{T} \right\|_{P_L}^2 - 2k \left\| D_1^R \mathbf{T} \right\|_{P_R}^2, \tag{61}$$

where $v_I = [\mathcal{T}_M, (D_1^L \mathcal{T})_M, T_0, (D_1^R T)_0]^T$ and

$$M_I = \begin{bmatrix} 2\sigma_3^M & \beta\varepsilon + \alpha\varepsilon\sigma_5^M + \sigma_4^M & -(\sigma_3^M + \tau_1^0) & -(\beta k\sigma_5^M + \tau_2^0) \\ \beta\varepsilon + \alpha\varepsilon\sigma_5^M + \sigma_4^M & 0 & -(\sigma_4^M + \alpha\varepsilon\sigma_3^M) & 0 \\ -(\sigma_3^M + \tau_1^0) & -(\sigma_4^M + \alpha\varepsilon\sigma_3^M) & 2\tau_1^0 & -k + \beta k\tau_3^0 + \tau_2^0 \\ -(\beta k\sigma_5^M + \tau_2^0) & 0 & -k + \beta k\tau_3^0 + \tau_2^0 & 0 \end{bmatrix}. \tag{62}$$

In order for the coupling terms to be bounded we need $M_I \leqslant 0$. The columns which have zero on the diagonal must be canceled. This gives a system of equations with one parameter family of solutions

$$s \in \mathbb{R}, \quad \sigma_4^M = -\beta\varepsilon(1+s), \quad \sigma_5^M = s, \quad \tau_2^0 = -ks, \quad \tau_3^0 = 1 + s. \tag{63}$$

Using relations (63), M_I reduces to

$$M_I = \begin{bmatrix} 2\sigma_3^M & 0 & -(\sigma_3^M + \tau_1^0) & 0 \\ 0 & 0 & 0 & 0 \\ -(\sigma_3^M + \tau_1^0) & 0 & 2\tau_1^0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{64}$$

and by choosing

$$\sigma_3^M = \tau_1^0 \leqslant 0, \tag{65}$$

we have $M_I \leqslant 0$ and all coupling terms are bounded. The parameter s will be of particular interest in later sections since it determines the type of the coupling.

Using all the above we can thus conclude.

Proposition 3.1. *The schemes (29) and (30) coupled at $x = 0$ are stable using the SAT boundary and interface treatment with penalty coefficients given by (35), (40), (43), (54), (58), (63) and (65).*

Remark 3.3. As in the continuous case we have assumed the boundary data to be identically zero. If we would have obtained an energy estimate with non-zero data the coupled schemes would have been strongly stable [16].

4. Order of convergence

The order of convergence is studied by the method of manufactured solutions. The time step ($\Delta t = 10^{-5}$) for all computations is chosen such that the scheme with 4th-order operators is well below the stability limit with 256 grid points in each

subdomain, and we integrate in time until $t = 0.1$ using the classical 4th-order Runge–Kutta method. This ensures that the time integration errors are negligible compared to the spatial discretization error. We use the functions

$$\begin{aligned}
&\rho(x,t) = \cos(2\pi x - t) + \sin(2\pi x - t), \quad u(x,t) = x + \cos(2\pi x - t),\\
&\mathcal{T}(x,t) = \frac{1}{\varepsilon}\sin(2\pi x)e^{-\kappa t}, \quad T(x,t) = \frac{1}{k}\sin(2\pi x)e^{-\kappa t}, \quad \kappa = 0.1,
\end{aligned} \tag{66}$$

which inserted into (1) and (2) gives a modified system of equations with additional forcing functions

$$\begin{aligned}
&w_t + Aw_x = \epsilon B w_{xx} + F,\\
&T_t = kT_x x + G,
\end{aligned} \tag{67}$$

where $F = [F_1, F_2, F_3]^T$ and

$$\begin{aligned}
F_1 &= (1 - 2\pi(a+b))\sin(2\pi x - t) + (-1 + 2\pi a)\cos(2\pi x - t) + b,\\
F_2 &= (1 - 2\pi(a+b))\sin(2\pi x - t) + 2\pi(b + 2\pi\epsilon\alpha)\cos(2\pi x - t) + \frac{2\pi c}{\epsilon}\cos(2\pi x)e^{-\kappa t} + a,\\
F_3 &= -2\pi c\sin(2\pi x - t) + \left(-\frac{\kappa}{\epsilon} + 4\pi^2\beta\right)\sin(2\pi x)e^{-\kappa t} + \frac{2\pi a}{\epsilon}\cos(2\pi x)e^{-\kappa t} + c,\\
G &= \left(-\frac{\kappa}{k} + 4\pi^2\right)\sin(2\pi x)e^{-\kappa t}.
\end{aligned} \tag{68}$$

The functions (66) are analytic solutions to the modified system (67) and they satisfy the interface conditions in a non-trivial way. Using (66) we create exact initial- and time dependent boundary data where needed. The penalty parameters have been chosen with equality sign where there are inequalities, $r = -1/2c$ and $s = -1/2$. The rate of convergence is obtained as

$$q_j^i = \log_{10}\left(\frac{\left\|u_{j-1}^i - v_{j-1}^i\right\|}{\left\|u_j^i - v_j^i\right\|}\right) \Big/ \log_{10}\left(\frac{h_j}{h_{j-1}}\right) \tag{69}$$

where q_j^i denotes the convergence rate for either of the variables $i = \rho, u, \mathcal{T}, T$ at mesh refinement level j. u_j^i is the exact analytic solution for either of the variables i at mesh refinement level j and v_j^i is the discrete solution. The ratio h_j/h_{j-1} is the ratio between the number of grid points at each refinement level. The coefficients in (1) and (2) have been chosen as

$$a = 0.5, \quad b = \frac{1}{\sqrt{\gamma}}, \quad c = \sqrt{\frac{\gamma - 1}{\gamma}}, \quad \gamma = 1.4, \quad \alpha = \beta = 1, \quad \varepsilon = 0.1, \quad k = 1 \tag{70}$$

and the results are seen in Table 1.

The rates of convergence in Table 1 agree with the theoretically expected results [3,6]. The convergence in this case can be improved by using a second derivative difference operator on compact form (if the solution of the coupled problem is proven

Table 1
Order of convergence.

$M = N$	2nd-order	3rd-order	4th-order
	ρ	ρ	ρ
32	1.5397	3.3169	3.9166
64	1.8835	3.3032	4.1544
128	1.9808	3.1561	4.1998
256	1.9934	3.0453	4.1291
	u	u	u
32	2.0177	3.3919	5.7397
64	2.0123	3.2439	4.0481
128	2.0018	3.1309	3.5984
256	2.0024	3.0619	3.8251
	$\mathcal{T}$	$\mathcal{T}$	$\mathcal{T}$
32	1.9774	2.8456	4.3129
64	1.9868	2.9676	4.7098
128	1.9920	2.9973	4.8654
256	1.9959	3.0023	4.9148
	T	T	T
32	1.9260	3.0821	4.2883
64	1.9529	3.0257	4.5497
128	1.9751	3.0152	4.3572
256	1.9873	3.0088	4.0936

to be pointwise bounded and the penalty coefficients are chosen correctly) [17]. This case is not considered in this paper since we are aiming for the compressible Navier–Stokes equations where the diffusive terms have variable coefficients. For this type of problem the theory for the compact formulation is not yet satisfactory and work remains to be done.

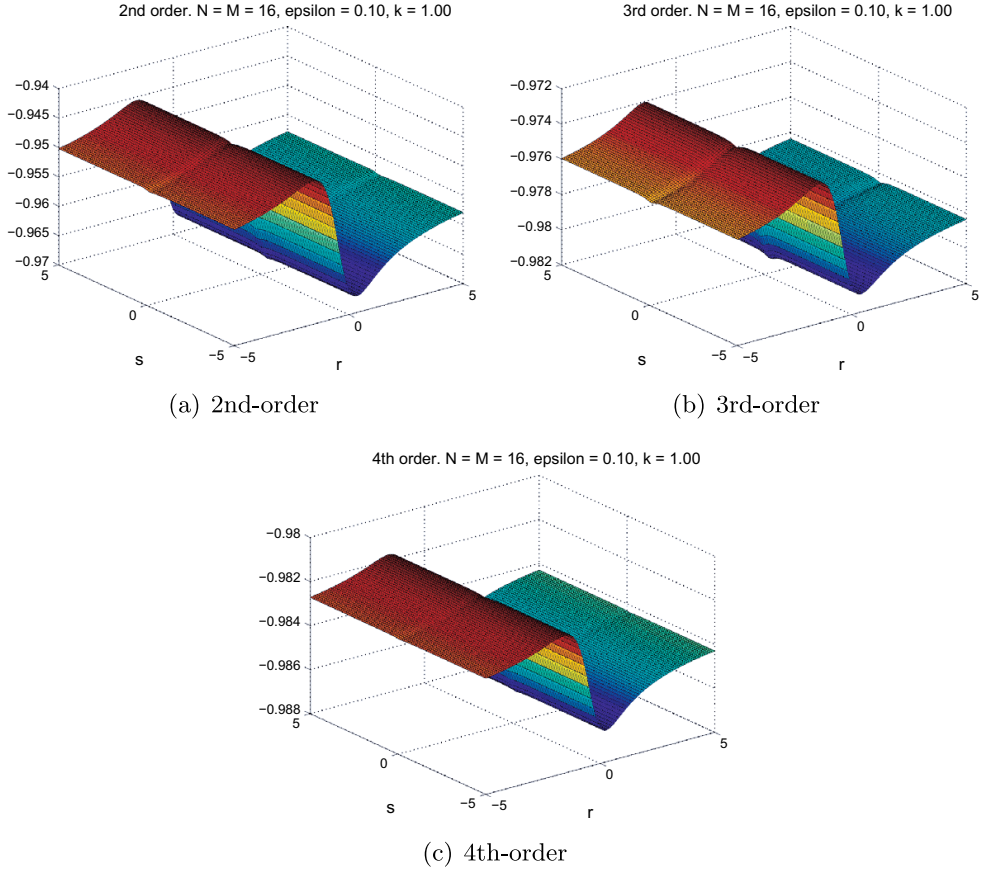
5. Spectral analysis and convergence to steady-state

When doing flow computations one is often interested in reaching the steady-state solution fast. From (29) and (30) we can see that we can write the fully coupled scheme as

$$\frac{d\mathbf{v}}{dt} = H\mathbf{v} + F \tag{71}$$

where the entire spatial discretization has been collected in the matrix H and F contains the boundary data. There are mainly two ways of enhancing convergence to steady-state. One is to make a spatial discretization which has negative real parts of the eigenvalues with as large magnitude as possible. That will optimize the convergence to steady-state for the ODE system (71) [18–20]. The second is to advance in time with as large time step as possible. For an explicit time integration method, the time step is limited by the eigenvalue with largest modulus.

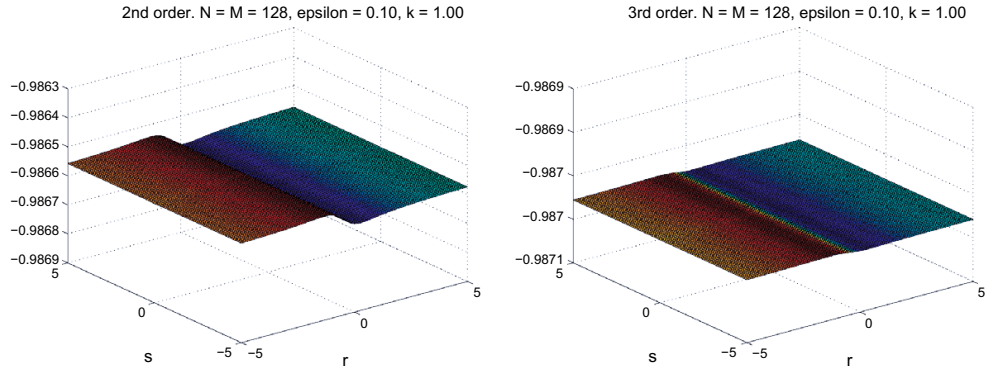
The scheme and penalty parameters are independent of the order of accuracy of the difference operators and hence we can study the spectrum of H for different orders. The first thing to be noticed is that there are two undetermined parameters r and s coming from the left boundary (40) and the interface (63). Theoretically any choice of these parameters lead to a stable scheme. With a too large magnitude they will make the problem stiff and a smaller time step is needed. Within a decent range it is interesting to see how the spectrum of H changes as a function of these parameters.



(a) 2nd-order (b) 3rd-order

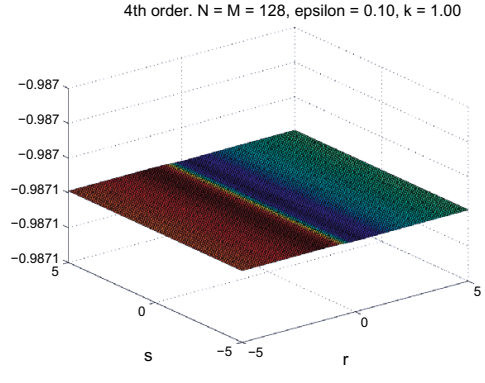
(c) 4th-order

Fig. 3. Minimum real part of the eigenvalues of the spatial discretization as a function of the boundary and interface parameters r and s for $M = N = 16$ grid points. Note that the surfaces become flatter with higher orders due to the improved convergence.



(a) 2nd-order

(b) 3rd-order



(c) 4th-order

Fig. 4. Minimum real part of the eigenvalues of the spatial discretization as a function of the boundary and interface parameters r and s for $M = N = 128$ grid points.

In Fig. 3 the minimum real part of the spectrum of H is plotted as a function of r and s for $M = N = 16$. Since the scheme is stable all real parts are negative.[1]

As the mesh is refined the dependence of the boundary and interface parameter disappears and the minimum real part of the eigenvalues converge to the same value for all choices and all orders of accuracy, see Fig. 4.

To see the convergence of the spectrum we compute the minimum real part of the eigenvalues of the spatial discretization for an increasing number of grid points. The boundary and interface parameter have been chosen as $r = -0.4$ and $s = -0.5$ for all orders and number of grid points. The choice $r = -0.4$ makes the penalty coefficients at the left boundary to be of approximately the same magnitude. All choices of r with a magnitude of order one lead to approximately the same results. The results are shown in Table 2 and Fig. 5 where we can see that the minimum real part of the spectrum of the discretization converges for all orders as they should.

The parameter s in (63) is of particular interest. In the figures and tables below we have chosen $\sigma_3^M = \tau_1^0 = 0$ and hence the coupling depends only on s. By choosing $s = 0$, Dirichlet conditions for continuity of temperature are given to the fluid domain and Neumann conditions for continuity of heat flux to the solid domain. By choosing $s = -1$ we get the reversed order. By choosing s such that no terms are canceled in (44) and (45) we get a mixed type of interface conditions.

As can be seen from Fig. 3 there are variations depending on the choice of r and s for a coarse mesh. Since we are interested in the properties of the discretization depending on the coupling, we fix $r = -0.4$ and compute the minimum real part of the spectrum as a function of s. The result can be seen in Table 3.

[1] Minimum will refer to the minimum modulus of the real part of the spectrum. It is the eigenvalue with negative real part closest to zero which will be of our interest.

Table 2
Minimum real part of the spectrum of the spatial discretization.

$M = N$	Minimum real part of the spectrum		
	2nd-order	3rd-order	4th-order
16	−0.95933	−0.97811	−0.98496
32	−0.97933	−0.98540	−0.98666
64	−0.98510	−0.98681	−0.98701
128	−0.98658	−0.98703	−0.98706
256	−0.98694	−0.98706	−0.98706

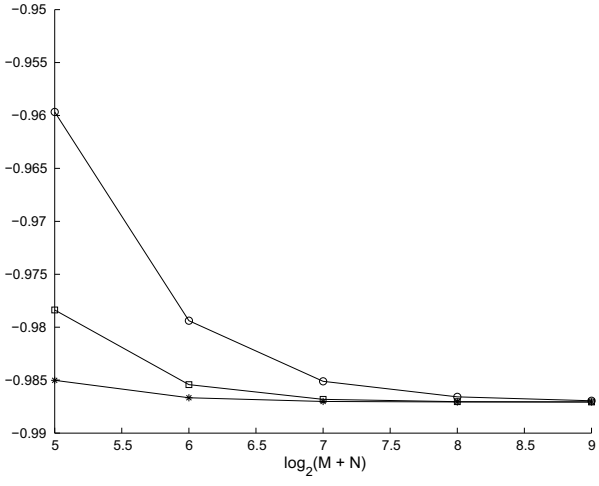


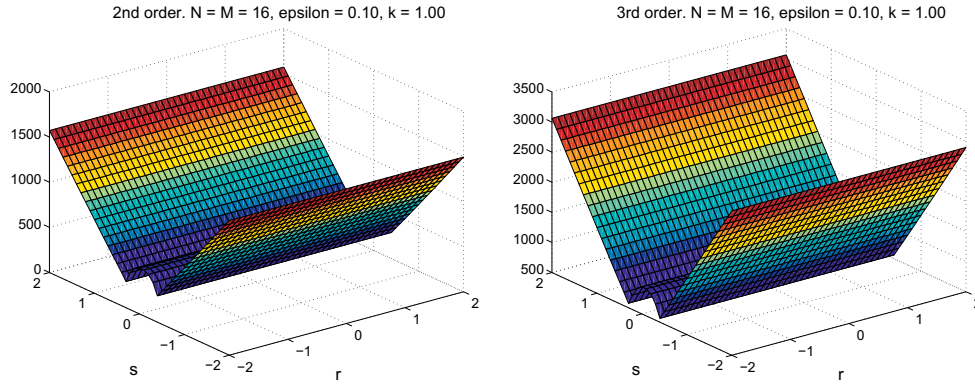
Fig. 5. Convergence of the minimum real part of the discrete spectrum for 2nd- (circle), 3rd- (square) and 4th-order (star) spatial discretization.

Table 3
The value of s which give minimal real part of the spectrum is shown in the upper part. The lower part includes a comparison with the case $s = -1$.

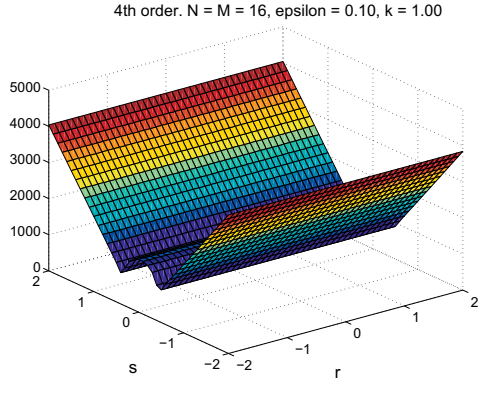
$M = N$	2nd-order		3rd-order		4th-order	
	s	min $\Re(\lambda)$	s	min $\Re(\lambda)$	s	min $\Re(\lambda)$
16	0.0	−0.97367	0.0	−0.97837	−0.1	−0.98502
32	0.0	−0.98310	0.0	−0.98542	0.0	−0.98667
64	0.0	−0.98600	0.0	−0.98681	0.0	−0.98701
128	0.0	−0.98679	0.0	−0.98703	0.0	−0.98706
16	−1.0	−0.97117	−1.0	−0.97806	−1.0	−0.98495
32	−1.0	−0.98259	−1.0	−0.98540	−1.0	−0.98666
64	−1.0	−0.98589	−1.0	−0.98681	−1.0	−0.98701
128	−1.0	−0.98676	−1.0	−0.98703	−1.0	−0.98706

Interface procedures for the heat equation have been considered before by e.g. Giles [21], Roe et al. [22] and recently by Henshaw and Chand [23]. Giles demonstrates a method where giving Dirichlet conditions for continuity of temperature to the fluid domain and Neumann conditions for continuity of heat flux to the solid domain is necessary for preserving stability, but that the time step restriction for certain discretizations and diffusion coefficients is more severe than in each of the subdomains. Roe et al. utilizes a different discretization and is able to circumvent this restriction by deriving a set of interface equations from the interface conditions that improve the stability characteristics and also preserve the accuracy of the scheme. Henshaw and Chand considers many different interface procedures and prove both stability and second order accuracy independent of the diffusive properties in contrast to the results in [21]. They also state that more attractive convergence results might be obtained by considering a mixed type of interface conditions.

As can be seen from Table 3 the choice $s = 0$ maximizes the real part of the spectrum and hence improves the convergence. It is also clear that the difference between the results for $s = 0$ and $s = -1$ are small. We investigated the intermediate values



(a) 2nd-order (b) 3rd-order



(c) 4th-order

Fig. 6. Maximum absolute value of the eigenvalues of the spatial discretization as a function of the boundary and interface parameters r and s for $M = N = 16$ grid points.

Table 4
The values of s which gives minimum largest modulo of the spectrum is shown in the upper part. The lower parts includes a comparison with the cases $s = 0$ and $s = -1$.

$M = N$	2nd-order		3rd-order		4th-order	
	s	max $\lvert\lambda\rvert$	s	max $\lvert\lambda\rvert$	s	max $\lvert\lambda\rvert$
16	0.27	264.37212	0.26	499.93406	−0.47	835.66403
32	0.28	1048.74365	0.26	1956.96051	−0.47	3326.02711
64	0.28	4131.04022	0.26	7730.02818	−0.47	13290.63251
128	0.28	16385.84328	0.26	30875.85671	−0.47	53149.05486
16	0.0	506.90795	0.0	947.57621	0.0	1464.10848
32	0.0	2024.77672	0.0	3788.48085	0.0	5848.25739
64	0.0	8096.41699	0.0	15152.37150	0.0	23385.03672
128	0.0	32383.30564	0.0	60608.47526	0.0	93532.47649
16	−1.0	736.69526	−1.0	1427.61237	−1.0	1889.15048
32	−1.0	2941.57122	−1.0	5703.70657	−1.0	7549.01304
64	−1.0	11756.87139	−1.0	22802.26105	−1.0	30181.19951
128	−1.0	47009.65698	−1.0	91184.83169	−1.0	120695.41482

as well and not much difference in min $\Re(\lambda)$ was found. From this point of view, the choice $s = 0$ is preferable. However we shall see that when regarding the time step, it is not.

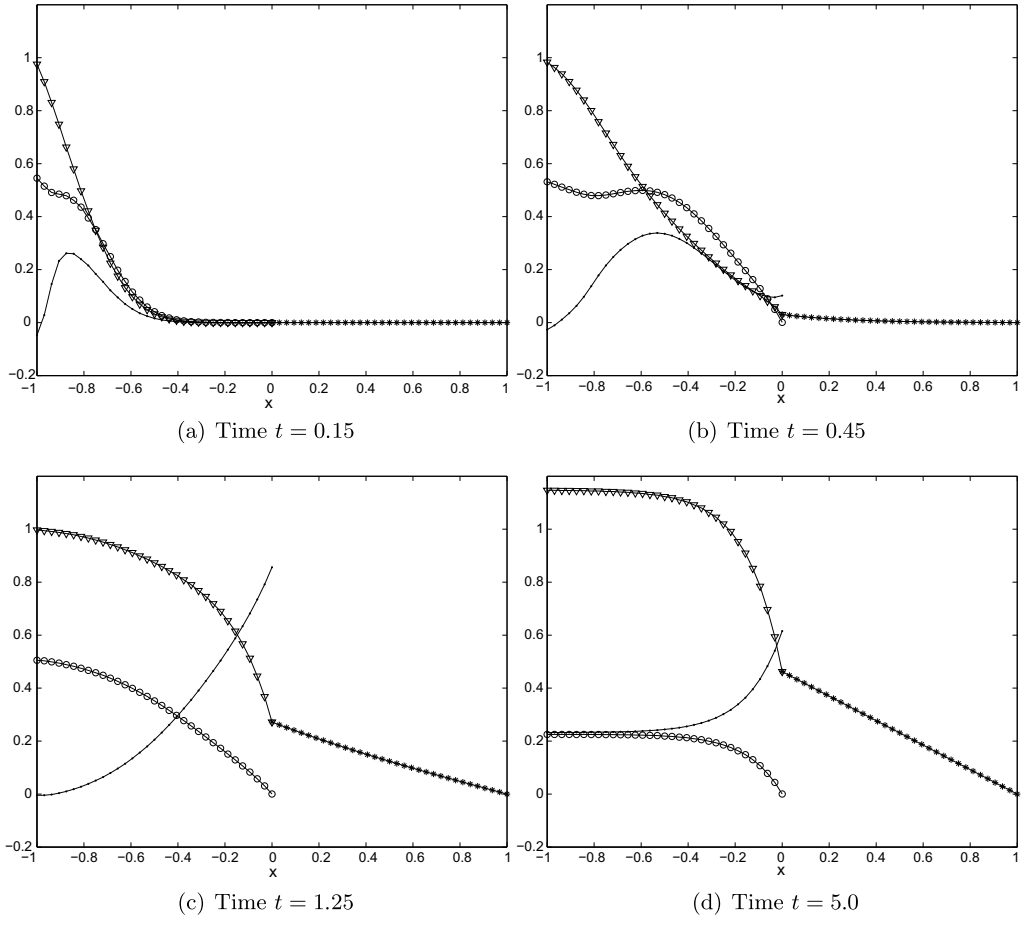
Regarding the issue of stiffness and the time step we can perform the same procedure as above but instead compute the maximum modulus of the spectrum as a function of r and s. The results for $M = N = 16$ grid points are shown in Fig. 6.

Clearly the stiffless is strongly influenced by s related to the interface coupling s but not by r relating to the left boundary condition. As before we fix $r = -0.4$ and compute the maximum modulus of the spectrum as a function of s. The result is seen in Table 4 together with a comparison with the extremal cases $s = 0$ and $s = -1$. As can be seen in Table 4 the stiffness can be reduced by choosing a mixed type of interface condition, and hence bigger time steps can be used. Compared to the extremal values $s = 0$ and $s = -1$ the optimal choices of s allows one to take almost twice as big time step and maintain stability for an explicit time integration method. This result is discussed in [23] for the heat equation and we can now verify it for this more general problem.

When performing computations of (1) and (2) on separate domains given standard boundary conditions, it was seen that the time step restriction for the coupled problem is the same as that in the worst of the subdomain problems when the optimal value of s was used. However, when a non-optimal value of s is used, the time step restriction for the coupled problems will be more severe than in that of the worst subdomain problems.

6. Two applications

An example of a solution, where the coefficients are given by (70) is given in Fig. 7. We start with zero initial data and at time $t = 0$ we let $\rho = 0$,$u = 0.5$ and $\mathcal{T} = 1$ at the left boundary while $T = 0$ at the right boundary and $u = 0$ at the interface. The



(a) Time $t = 0.15$ (b) Time $t = 0.45$

(c) Time $t = 1.25$ (d) Time $t = 5.0$

Fig. 7. ρ (solid), u (circle), $\mathcal{T}$ (triangle), T (star). A sequence of solutions for different times using $M = N = 32$ grid points and 3rd-order operators. The last figure shows the steady-state solution.

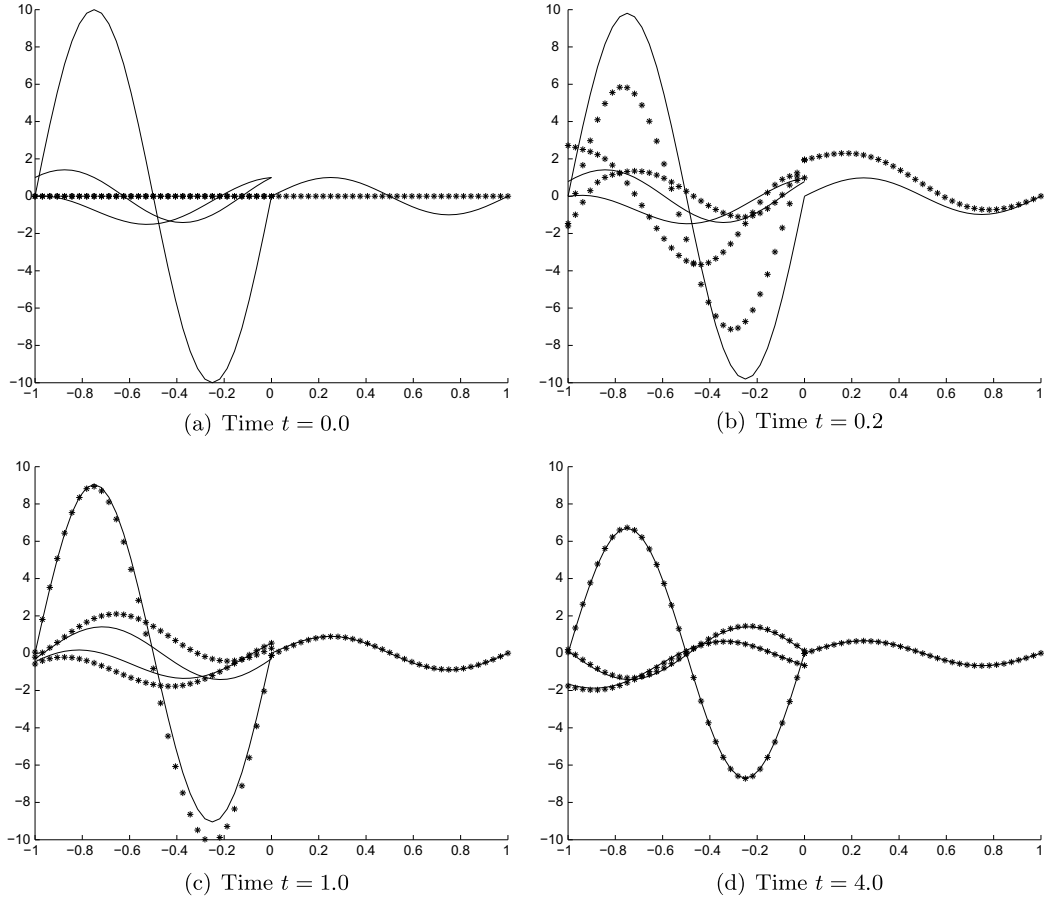


Fig. 8. Computed (star), analytic (solid). A sequence of computed vs. analytic solutions with wrong initial data for different times using $M = N = 32$ grid points and 3rd-order operators.

values at the left boundary are transformed into data for the characteristic boundary conditions. We can see how the influences from the left boundary travel across the domain and reaches the interface. No external data is created for $\mathcal{T}$, T, $\mathcal{T}_x$ or T_x but the weak interface conditions (19) and (21) together with the fully coupled formulation (46) make sure that the temperature is continuous across the interface and that the heat fluxes are equal up to the order of the scheme.

To obtain a correct solution, it is not necessary to initialize with correct data. By using the functions a(66) we can initiate the computation with zero data and investigate whether or not the computed solution converges to the analytic solution with time. Fig. 8 clearly shows that it indeed does.

7. Summary and conclusions

An incompletely parabolic system of equations is coupled with the heat equation in one space dimension. The energy method is used to derive well-posed boundary and interface conditions. The equations are discretized using finite differences on Summation-by-Parts form where the boundary and interface conditions are weakly imposed using the Simultaneous Approximation Term. The penalty matrices and coefficients are determined such that we can prove that the coupled scheme is stable.

The interface conditions are derived such that we can study different interface conditions as a function of one parameter. By looking at the spectrum of the spatial discretization as a function of the interface parameter, it can be seen that there are only minor differences between the minimum real part of the spectrum for different coupling techniques. However when

giving a mixed type of interface condition the stiffness is greatly reduced and an almost twice as big time step can be used while maintaining stability for an explicit time integration method.

The rate of convergence is verified by the method of manufactured solutions and the result is consistent with the theory within the SBP framework. The derived numerical schemes are independent of the order of accuracy and higher-order accuracy is easily obtained by using difference operators of higher-orders. Two examples where the system is solved using 3rd-order operators are shown and it can be seen that the correct interface conditions are obtained.

References

[1] Jens Lindström, Johan Bejhed, and Jan Nordström. Measurements and numerical modelling of orifice flow in microchannels, in: The 41st AIAA Thermophysics Conference, AIAA Paper No. 2009-4098, San Antonio, USA, 22–25 June, 2009.
[2] Bo Strand, Summation by parts for finite difference approximations for d/dx, Journal of Computational Physics 110 (1) (1994) 47–67.
[3] Ken Mattsson, Jan Nordström, Summation by parts operators for finite difference approximations of second derivatives, Journal of Computational Physics 199 (2) (2004) 503–540.
[4] Ken Mattsson, Boundary procedures for summation-by-parts operators, Journal of Scientific Computing 18 (1) (2003) 133–153.
[5] Mark H. Carpenter, Jan Nordström, David Gottlieb, A stable and conservative interface treatment of arbitrary spatial accuracy, Journal of Computational Physics 148 (2) (1999) 341–365.
[6] Jing Gong, Jan Nordström, Stable, Accurate and Efficient Interface Procedures for Viscous Problems, Report, Uppsala University, Disciplinary Domain of Science and Technology, Mathematics and Computer Science, Department of Information Technology, Numerical Analysis, Department of Information Technology, Uppsala University, Uppsala, 2006.
[7] X. Huan, J.E. Hicken, D.W. Zingg, Interface and boundary schemes for high-order methods, in: The 39th AIAA Fluid Dynamics Conference, AIAA Paper No. 2009-3658, San Antonio, USA, 22–25 June, 2009.
[8] Magnus Svärd, Mark H. Carpenter, Jan Nordström, A stable high-order finite difference scheme for the compressible Navier–Stokes equations, far-field boundary conditions, Journal of Computational Physics 225 (1) (2007) 1020–1038.
[9] Magnus Svärd, Jan Nordström, A stable high-order finite difference scheme for the compressible Navier–Stokes equations: no-slip wall boundary conditions, Journal of Computational Physics 227 (10) (2008) 4805–4824.
[10] Ken Mattsson, Magnus Svärd, Mark Carpenter, Jan Nordström, High-order accurate computations for unsteady aerodynamics, Computers and Fluids 36 (3) (2007) 636–649.
[11] Jan Nordström, Jing Gong, Edwin van der Weide, Magnus Svärd, A stable and conservative high order multi-block method for the compressible Navier–Stokes equations, Journal of Computational Physics 228 (24) (2009) 9020–9035.
[12] Ken Mattsson, Magnus Svärd, Jan Nordström, Stable and accurate artificial dissipation, Journal of Scientific Computing 21 (1) (2004) 57–79.
[13] Jan Nordström, Magnus Svärd, Well-posed boundary conditions for the Navier–Stokes equations, SIAM Journal of Numerical Analysis 43 (3) (2005) 1231–1255.
[14] Saul Abarbanel, David Gottlieb, Optimal time splitting for two- and three-dimensional Navier–Stokes equations with mixed derivatives, Journal of Computational Physics 41 (1) (1981) 1–33.
[15] Jan Nordström, The use of characteristic boundary conditions for the Navier–Stokes equations, Computers and Fluids 24 (5) (1995) 609–623.
[16] Bertil Gustafsson, Heinz-Otto Kreiss, Joseph Oliger, Time Dependent Problems and Difference Methods, Wiley Interscience, 1995.
[17] Magnus Svärd, Jan Nordström, On the order of accuracy for difference approximations of initial-boundary value problems, Journal of Computational Physics 218 (1) (2006) 333–352.
[18] Magnus Svärd, Ken Mattsson, Jan Nordström, Steady-state computations using summation-by-parts operators, Journal of Scientific Computing 24 (1) (2005) 79–95.
[19] Jan Nordström, The influence of open boundary conditions on the convergence to steady state for the Navier–Stokes equations, Journal of Computational Physics 85 (1) (1989) 210–244.
[20] Björn Engquist, Bertil Gustafsson, Steady state computations for wave propagation problems, Mathematics of Computation 49 (179) (1987) 39–64.
[21] M.B. Giles, Stability analysis of numerical interface conditions in fluid-structure thermal analysis, International Journal for Numerical Methods in Fluids 25 (1997) 421–436. August.
[22] B. Roe, R. Jaiman, A. Haselbacher, P.H. Geubelle, Combined interface boundary condition method for coupled thermal simulations, International Journal for Numerical Methods in Fluids 57 (2008) 329–354. May.
[23] William D. Henshaw, Kyle K. Chand, A composite grid solver for conjugate heat transfer in fluid-structure systems, Journal of Computational Physics 228 (10) (2009) 3708–3741.

Paper II

Applied Numerical Mathematics 62 (2012) 1620–1638

Contents lists available at SciVerse ScienceDirect

Applied Numerical Mathematics

www.elsevier.com/locate/apnum

Spectral analysis of the continuous and discretized heat and advection equation on single and multiple domains

Jens Berg [a,*], Jan Nordström [b]

[a] Division of Scientific Computing, Department of Information Technology, Uppsala University, Box 337, SE-751 05, Uppsala, Sweden
[b] Linköping University, Department of Mathematics, SE-581 83, Linköping, Sweden

ARTICLE INFO

Article history:
Received 6 December 2010
Received in revised form 29 August 2011
Accepted 11 May 2012
Available online 15 May 2012

Keywords:
Spectral analysis
Eigenvalues
Diffusion
Advection
Summation-By-Parts
Weak boundary conditions
Weak interface conditions

ABSTRACT

In this paper we study the heat and advection equation in single and multiple domains. The equations are discretized using a second order accurate finite difference method on Summation-By-Parts form with weak boundary and interface conditions. We derive analytic expressions for the spectrum of the continuous problem and for their corresponding discretization matrices.

It is shown how the spectrum of the single domain operator is contained in the multi domain operator spectrum when artificial interfaces are introduced. The interface treatments are posed as a function of one parameter, and the impact on the spectrum and discretization error is investigated as a function of this parameter. Finally we briefly discuss the generalization to higher order accurate schemes.

1. Introduction

When performing large scale computations in scientific computations involving partial differential equations (PDEs), there is often a need to divide the computational domain into smaller subdomains. This is done either to allow more flexible geometry handing for structured methods or to obtain sufficient resolution by distributing the computations in the subdomains on parallel computers. Independently of which PDE (Navier–Stokes, Euler, Maxwell, Schrödinger, wave, ...) that is being solved, one would like to construct the interfaces between the subdomains in such a way that certain properties of the discretization is preserved, or even improved, for example accuracy, stability, conservation, convergence and stiffness.

Stable and accurate interface treatments are required in many applications, for example fluid–structure interaction [17], conjugate heat transfer [12,9], computational fluid dynamics [18,10] and computational quantum dynamics [15] to mention a few. From the mathematical point of view, an interface is purely artificial and has no influence on the solution. However when introducing interfaces in a computational domain, the numerical scheme is modified and one has to make sure that these modifications does not destroy the solution.

The focus in this paper is a finite difference method on Summation-By-Parts form together with the Simultaneous Approximation Term (SAT) for imposing the boundary and interface conditions weakly. The equations we consider are the heat equation and advection equation in one space dimension.

The SBP and SAT method has been used for many applications in fluid dynamics since it has the benefit of being provable energy stable when the correct boundary and interface conditions are imposed for the PDE [19,21,24,1,2].

* Corresponding author. Tel.: +46 18 471 6253; faxes: +46 18 523049, +46 18 511925.
E-mail address: jens.berg@it.uu.se (J. Berg).

0168-9274/$36.00
http://dx.doi.org/10.1016/j.apnum.2012.05.002

Here we investigate the details of the diffusion and advection operators by considering the one-dimensional heat and advection equation on single and multiple domains. The boundary and interface conditions are imposed weakly using the SAT technique and the equations are discretized using a second order accurate SBP operator. There are SBP operators accurate of order 2, 3, 4 and 5 derived in for example [14,20] and we stress that the stability analysis given in this paper holds for any order of accuracy. The second order operators was chosen since it allows us to derive analytical results regarding certain spectral properties of the operators.

The analysis is performed using the Laplace transform method [11,8,7]. Since the SBP and SAT discretization is a method of lines, time is kept continuous and only space is discretized. Hence the Laplace transform turns the numerical scheme into a system of ordinary differential equations (ODE) in transformed space. This ODE is an eigenvalue problem and the solution determines the spectral properties of the spatial discretization.

2. Single domain spectral analysis of the heat equation

In order to compare the effects on the spectrum when introducing an artificial interface we shall begin by decomposing the heat equation on a single domain both continuously and discretized. This allows us to isolate expressions stemming from the boundaries only and separate them from the interface part.

2.1. Continuous case

Consider the heat equation on $-1 \leqslant x \leqslant 1$,

$$\begin{aligned} &u_t = u_{xx}, \\ &u(x,0) = f(x), \\ &u(-1,t) = g_1(t), \\ &u(1,t) = g_2(t) \end{aligned} \tag{1}$$

where the notation u_ξ denotes the partial derivative of u with respect to the variable ξ where ξ is either the space or time variable x or t respectively. To analyze (1) we introduce the Laplace transform

$$\hat{u} = \mathcal{L}u = \int_0^\infty e^{-st} u \, dt \tag{2}$$

which is defined for locally integrable functions on $[0,\infty)$ where the real part of s has to be sufficiently large [11,8]. The basic property that we are going to use is that it transforms differentiation with respect to the time variable to multiplication with the complex number s. Hence a time-dependent PDE in Laplace transformed space is an ordinary differential equation (ODE) which we can solve. Finding analytically the inverse transformation is in general a very difficult problem but that is not our interest here.

We shall use the Laplace transform to determine the spectrum of (1). Assume that $g_1 = g_2 = 0$ and take the Laplace transform of (1). The initial condition is omitted since it does not enter in the spectral analysis. We get an ODE in transformed space,

$$\begin{aligned} &s\hat{u} = \hat{u}_{xx}, \\ &\hat{u}(-1,s) = 0, \\ &\hat{u}(1,s) = 0, \end{aligned} \tag{3}$$

which is an eigenvalue problem for the second derivative operator. By the ansatz $\hat{u} = e^{kx}$ we can determine that the general solution to (3) is

$$\hat{u} = c_1 e^{\sqrt{s}x} + c_2 e^{-\sqrt{s}x}. \tag{4}$$

By applying the boundary conditions we obtain

$$c_1 e^{-\sqrt{s}} + c_2 e^{\sqrt{s}} = 0, \tag{5}$$

$$c_1 e^{\sqrt{s}} + c_2 e^{-\sqrt{s}} = 0 \tag{6}$$

which we write in matrix form as

$$\underbrace{\begin{bmatrix} e^{-\sqrt{s}} & e^{\sqrt{s}} \\ e^{\sqrt{s}} & e^{-\sqrt{s}} \end{bmatrix}}_{E(s)} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{7}$$

Eq. (7) will have a non-trivial solution when the coefficient matrix $E(s)$ is singular. We hence seek the values of s such that the determinant is zero. We have

$$\det\big(E(s)\big) = -2\sinh(2\sqrt{s}) \tag{8}$$

which is zero for

$$s = -\frac{\pi^2 n^2}{4}, \quad n \in \mathbb{N}. \tag{9}$$

This infinite sequence of values is thus the spectrum of (1). Note that $s = 0$ is not considered a solution since then we have a double root and $\hat{u} = c_1 + c_2 x$. From the boundary conditions we get that $\hat{u} \equiv 0$ and hence $u \equiv 0$, which is trivial.

2.2. Discrete case

To discretize (1) we use a second order accurate finite difference operator on SBP form,

$$u_{xx} \approx D_2 v \tag{10}$$

where $v = [v_0, v_1, \ldots, v_N]^T$ is the discrete grid function and the mesh is uniform with $N+1$ grid points. The exact form of the operator D_2 is, see [14,2],

$$D_2 = P^{-1}(-A + BD) = \frac{1}{\Delta x^2}\begin{bmatrix} 0 & 0 & 0 & 0 & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \tag{11}$$

where

$$P = \Delta x \begin{bmatrix} \frac{1}{2} & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \\ 0 & \cdots & 0 & 0 & \frac{1}{2} \end{bmatrix}, \qquad A = \frac{1}{\Delta x}\begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 1 \end{bmatrix},$$

$$B = \begin{bmatrix} -1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 0 \\ 0 & \cdots & 0 & 0 & 1 \end{bmatrix}, \qquad D = \frac{1}{\Delta x}\begin{bmatrix} -1 & 1 & \cdots & 0 & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & -\frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & \cdots & -1 & 1 \end{bmatrix}. \tag{12}$$

Note that (11) has zeros on the top and bottom row and is hence inconsistent at the boundaries. This does however not affect the global accuracy because of the SAT implementation of the weak boundary conditions [14,6,23].

The entire scheme for (1) can be written as

$$v_t = D_2 v + \sigma_1 P^{-1} D^T e_0 (v_0 - g_1) + \sigma_2 P^{-1} D^T e_N (v_N - g_2) \tag{13}$$

where P is the positive symmetric matrix in (12) which defines a discrete norm by $\|w\|^2 = w^T P w$. The vectors $e_{0,N}$ are zero vectors except for the first and last position respectively, which is one. The two parameters $\sigma_{1,2}$ will be determined such that the scheme is stable in the P-norm [20,13].

Note that we only discretize in space and keep time continuous. The discrete norm is hence a function of time and the stability of the scheme will depend upon whether or not we can derive a bounded estimate for the time rate of change of the discrete norm.

2.2.1. Stability

We use the energy method to determine the coefficients $\sigma_{1,2}$ such that the scheme is stable. The stability of the scheme ensures that all eigenvalues of the complete difference operator, including the boundary conditions, have non-positive real parts.

By multiplying (13) by $v^T P$ and adding the transpose to itself we obtain

$$\|v\|_t^2 = 2(\sigma_1 - 1)v_0(Dv)_0 + 2(\sigma_2 + 1)v_N(Dv)_N - v^T\left(A + A^T\right)v. \tag{14}$$

It is clear that the scheme is stable if we choose

$$\sigma_1 = 1, \qquad \sigma_2 = -1 \tag{15}$$

since A is symmetric and positive semi-definite which can be seen from (12). Hence $A + A^T \geqslant 0$ and the last term in (14) is dissipative.

2.2.2. Complete eigenspectrum

Consider (13) again with homogeneous boundary conditions. Since we have kept time continuous we can take the Laplace transform of the entire scheme and after rearranging we get

$$\underbrace{\left(sI - D_2 - \sigma_1 P^{-1} D^T E_0 - \sigma_2 P^{-1} D^T E_N\right)}_{M} \hat{v} = 0 \tag{16}$$

where I is the $N+1$-dimensional identity operator and $E_{0,N}$ are zero matrices except for the $(0,0)$ and (N,N) positions respectively which is one. To determine the complete eigenspectrum of the discrete operator M we start by considering the difference scheme for an internal point. The internal scheme is the standard central finite difference scheme and hence

$$\frac{\partial}{\partial t} v_i = \frac{v_{i-1} - 2v_i + v_{i+1}}{\Delta x^2}. \tag{17}$$

By taking the Laplace transform of (17) we obtain a recurrence relation

$$s\hat{v}_i = \frac{\hat{v}_{i-1} - 2\hat{v}_i + \hat{v}_{i+1}}{\Delta x^2} \tag{18}$$

for which we can obtain the general solution by the ansatz $\hat{v}_i = \sigma\kappa^i$. The ansatz yields the second order equation

$$\kappa^2 - (\tilde{s} + 2)\kappa + 1 = 0 \tag{19}$$

with the two solutions

$$\kappa_{+,-} = \frac{\tilde{s}+2}{2} \pm \sqrt{\left(\frac{\tilde{s}+2}{2}\right)^2 - 1} \tag{20}$$

where $\tilde{s} = s\Delta x^2$. Hence the general solution to (18) is

$$\hat{v}_i = c_1\kappa_+^i + c_2\kappa_-^i \tag{21}$$

where i is a grid point index on the left-hand side and the corresponding power on the right-hand side.

To obtain the eigenspectrum of M we consider the boundary points. The scheme is modified at grid points x_0, x_1, x_{N-1} and x_N and the corresponding equations are after substituting (15)

$$\begin{aligned}
&(\tilde{s}+2)\hat{v}_0 = 0,\\
&-2\hat{v}_0 + (\tilde{s}+2)\hat{v}_1 - \hat{v}_2 = 0,\\
&-\hat{v}_{N-2} + (\tilde{s}+2)\hat{v}_{N-1} - 2\hat{v}_N = 0,\\
&(\tilde{s}+2)\hat{v}_N = 0.
\end{aligned} \tag{22}$$

If we assume that the ansatz (21) is valid at grid points x_i, $i = 1, \ldots, N-1$ we get by substituting (21) into (22) the square matrix equation

$$\underbrace{\begin{bmatrix}
\tilde{s}+2 & 0 & 0 & 0\\
-2 & ((\tilde{s}+2)-\kappa_+)\kappa_+ & ((\tilde{s}+2)-\kappa_-)\kappa_- & 0\\
0 & ((\tilde{s}+2)\kappa_+ - 1)\kappa_+^{N-2} & ((\tilde{s}+2)\kappa_- - 1)\kappa_-^{N-2} & -2\\
0 & 0 & 0 & \tilde{s}+2
\end{bmatrix}}_{E(s,\kappa)}
\begin{bmatrix} \hat{v}_0\\ c_1\\ c_2\\ \hat{v}_N \end{bmatrix} = \begin{bmatrix} 0\\0\\0\\0 \end{bmatrix}. \tag{23}$$

Eq. (23) will have a non-trivial solution for the values of $\tilde{s}$ which makes $E(s,\kappa)$ singular. Thus we seek the values of $\tilde{s}$ for which $\det(E(s,\kappa))=0$. These values of $\tilde{s}$ constitute the spectrum of M [8]. The determinant of $E(s,\kappa)$ is

$$\det\big(E(s,\kappa)\big)=(\tilde{s}+2)^2\big(\kappa_-^N-\kappa_+^N\big) \tag{24}$$

and we can see that the spectrum contains the points for which

$$\tilde{s}=-2,\qquad \kappa_+^N=\kappa_-^N. \tag{25}$$

In the second case we have a binomial equation for the complex number $\tilde{s}$. To solve it we write $\kappa_+=ae^{i\theta}$ and $\kappa_-=be^{i\phi}$ in polar form where $a,b=|\kappa_{+,-}|$ and $\phi,\psi=\arg(\kappa_{+,-})$. After identifying a,b and ϕ,ψ we can determine that

$$\tilde{s}=2\left((-1)^k\cos\left(\frac{\pi k}{N}\right)-1\right),\quad k=1,\dots,N-1. \tag{26}$$

This solution method of binomial equations in complex variables can be found in any standard textbook in complex analysis. Thus we have found $N+1$ values of $\tilde{s}$ which gives non-trivial solutions to (23) and hence they constitute the entire spectrum of M.

Remark 2.1. One has to be careful with double roots. From (20) and (24) a possible solution would be when

$$\left(\frac{\tilde{s}+2}{2}\right)^2-1=0 \tag{27}$$

or equivalently $\tilde{s}=-4$ or $\tilde{s}=0$. These are however false roots. A proof is given in Appendix A. Interesting is though that all eigenvalues of M are contained between $\tilde{s}=-4$ and $\tilde{s}=0$.

2.2.3. Convergence of eigenvalues

To see how the eigenvalues of the discretization matrix converge to the eigenvalues of the continuous PDE, we let (9) be denoted by μ_n. We rescale (26) with Δx^2 and denote it $\tilde{\lambda}_k$. Since $\Delta x=\frac{2}{N}$ we can rewrite $\tilde{\lambda}_k$ as

$$\lambda_n=\frac{N^2}{2}\left(\cos\left(\frac{\pi n}{N}\right)-1\right),\quad n=1,\dots,N-1 \tag{28}$$

which generates the same sequence as (26), but it is monotonically decreasing. This allows us to compare μ_n and λ_n elementwise.

By assuming that $n<N$ we can Taylor expand (28) around zero and simplify to get

$$\lambda_n=\mu_n+O\left(\frac{n^4}{N^2}\right). \tag{29}$$

We can see that for $n\ll\sqrt{N}$, the eigenvalues are well approximated while for larger values of n, they will start to diverge. This is the typical situation. When the resolution is increased, more eigenvalues will be converged but even more eigenvalues that are not converging will be created.

Remark 2.2. Note that since $N\sim 1/\Delta x$, Eq. (29) ensures asymptotically second order convergence of all eigenvalues.

3. Multi domain spectral analysis of the heat equation

In this section we shall use the knowledge obtained in the previous section to determine spectral properties when an artificial interface has been introduced in the domain. Our goal is to determine how the introduction of an interface influences the spectrum of both the continuous and discrete equations. Moreover we want to design the interface treatment in such a way that the resulting difference operator is similar to, or maybe even better than, the single domain operator.

3.1. The continuous case

Consider now two heat equations coupled over an interface at $x=0$ with homogeneous boundary conditions

$$\begin{aligned}
&u_t=u_{xx},\quad -1\leqslant x\leqslant 0,\\
&v_t=v_{xx},\quad 0\leqslant x\leqslant 1,\\
&u(-1,t)=0,\\
&v(1,t)=0,
\end{aligned}$$

$$u(0,t) - v(0,t) = 0,$$
$$u_x(0,t) - v_x(0,t) = 0. \tag{30}$$

We take the Laplace transform again as before and obtain the general solutions for $\hat{u}$ and $\hat{v}$ as

$$\hat{u} = c_1 e^{\sqrt{s}x} + c_2 e^{-\sqrt{s}x},$$
$$\hat{v} = c_3 e^{\sqrt{s}x} + c_4 e^{-\sqrt{s}x}. \tag{31}$$

By applying the boundary and interface conditions we get the matrix equation

$$\underbrace{\begin{bmatrix} e^{-\sqrt{s}} & e^{\sqrt{s}} & 0 & 0 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 0 & 0 & e^{\sqrt{s}} & e^{-\sqrt{s}} \end{bmatrix}}_{E(s)} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{32}$$

with a non-trivial solution when $\det(E(s)) = 0$. A direct computation of the determinant shows that $\det(E(s)) = 4\sinh(2\sqrt{s})$ and hence the spectrum remains unchanged by the introduction of an interface. This is of course all in order since the interface is purely artificial. However when we discretize (30) we modify the scheme at the interface and we can expect that this modification will influence the eigenvalues of the complete difference operator.

3.2. The discrete case

In order to proceed we assume that there are equally many grid points in each subdomain and that the grid spacing is the same. This means that we can apply the same operators in both domains which will simplify the notation and algebra.

With a slight abuse of notation we now let u and v denote the discrete grid functions and both having $N+1$ components. Thus there are in total $2N+2$ grid points in the domain since the interface point occurs twice, and the resolution is twice as high as in the single domain case.

By using the SBP and SAT technique we can discretize (30) as

$$u_t = D_2 u + \sigma_1 P^{-1} D^T E_0 u + \sigma_2 P^{-1} D^T e_N (u_N - v_0) + \sigma_3 P^{-1} e_N \big((Du)_N - (Dv)_0\big),$$
$$v_t = D_2 v + \tau_1 P^{-1} D^T E_N v + \tau_2 P^{-1} D^T e_0 (v_0 - u_N) + \tau_3 P^{-1} e_0 \big((Dv)_0 - (Du)_N\big). \tag{33}$$

The unknown penalty parameters $\sigma_{1,2,3}$ and $\tau_{1,2,3}$ has again to be determined for stability.

3.2.1. Stability

To determine the unknown parameters $\sigma_{1,2,3}$ and $\tau_{1,2,3}$ we multiply the first equation in (33) with $u^T P$ and the second with $v^T P$. We add the transposes of the resulting expressions to themselves to get

$$\begin{aligned} \|u\|_t^2 &= -2u_0(Du)_0 + 2u_N(Du)_N - u^T\big(A + A^T\big)u \\ &\quad + 2\sigma_1(Du)_0 u_0 + 2\sigma_2(Du)_N(u_N - v_0) + 2\sigma_3 u_M\big((Du)_N - (Dv)_0\big), \\ \|v\|_t^2 &= -2v_0(Dv)_0 + 2v_N(Dv)_N - v^T\big(A + A^T\big)v \\ &\quad + 2\tau_1(Dv)_N v_N + 2\tau_2(Dv)_0(v_0 - u_N) + 2\tau_3 v_0\big((Dv)_0 - (Du)_N\big). \end{aligned} \tag{34}$$

By adding both expressions in (34) we can write the result as

$$\begin{aligned} \|u\|_t^2 + \|v\|_t^2 &= 2(\sigma_1 - 1)u_0(Du)_0 + 2(\tau_1 + 1)v_N(Dv)_N \\ &\quad + q^T H q - u^T\big(A + A^T\big)u - v^T\big(A + A^T\big)v \end{aligned} \tag{35}$$

where $q = [u_N, (Du)_N, v_0, (Dv)_0]^T$ and

$$H = \begin{bmatrix} 0 & 1+\sigma_2+\sigma_3 & 0 & -(\tau_2+\tau_3) \\ 1+\sigma_2+\sigma_3 & 0 & -(\tau_2+\tau_3) & 0 \\ 0 & -(\sigma_2+\tau_3) & 0 & -1+\tau_2+\tau_3 \\ -(\sigma_3+\tau_2) & 0 & -1+\tau_2+\tau_3 & 0 \end{bmatrix}. \tag{36}$$

In order to bound (35) we have to choose

$$\sigma_1 = 1, \qquad \tau_1 = -1 \tag{37}$$

as in the single domain case, and we have to choose the rest of the penalty parameters such that $H \leqslant 0$. This is easily accomplished by noting that the diagonal of H consists of zeros only, and hence by the Gershgorin theorem we need to put all remaining entires to zero to ensure the semi-definiteness of H. This gives us a one-parameter family of solutions

$$r \in \mathbb{R}, \qquad \sigma_2 = -(1+r), \qquad \sigma_3 = r, \qquad \tau_2 = -r, \qquad \tau_3 = 1+r. \tag{38}$$

Thus all penalty parameters have been determined and the scheme is stable.

Worth noting is that the parameter r determines how the equations are coupled. For $r = 0$ two of the penalty parameters in (38) disappear and renders the scheme one-sided coupled in the sense that the left domain receives a solution value from the right domain and gives the value of its gradient to the right domain. For $r = -1$ the situation is reversed and for other values of r, the scheme is fully coupled. Note that the scheme is stable for all choices of r. We shall investigate the influence of the interface parameter in later sections. More details can also be found in [12].

3.2.2. Eigenspectrum

The scheme (33) with a second order accurate difference operator makes eight grid points (two at each boundary and four at the interface) stray from a standard central finite difference scheme. This is a significant modification and we can expect that there will be a global impact depending on these modifications. A direct way of investigating this is by considering the change on the spectrum due to the modifications.

We take the Laplace transform of (33) and consider the difference equations at the modified boundary and interface points. We get after substituting (37) and (38) into (33) that

$$\begin{aligned}
&(\tilde{s}+2)\hat{u}_0 = 0,\\
&-2\hat{u}_0 + (\tilde{s}+2)\hat{u}_1 - \hat{u}_2 = 0,\\
&-\hat{u}_{N-2} + (\tilde{s}+2)\hat{u}_{N-1} - (2+r)\hat{u}_N + (1+r)\hat{v}_0 = 0,\\
&2r\hat{u}_{N-1} + (\tilde{s}+2)\hat{u}_N - 2(1+2r)\hat{v}_0 + 2r\hat{v}_1 = 0,\\
&-2(1+r)\hat{u}_{N-1} + 2(1+2r)\hat{u}_N + (\tilde{s}+2)\hat{v}_0 - 2(1+r)\hat{v}_1 = 0,\\
&-r\hat{u}_N - (1-r)\hat{v}_0 + (\tilde{s}+2)\hat{v}_1 - \hat{v}_2 = 0,\\
&-\hat{v}_{N-2} + (\tilde{s}+2)\hat{v}_{N-1} - 2\hat{v}_N = 0,\\
&(\tilde{s}+2)\hat{v}_N = 0.
\end{aligned} \tag{39}$$

From the internal schemes we have similarly as before that

$$\begin{aligned}
\hat{u}_i &= c_1\kappa_+^i + c_2\kappa_-^i,\\
\hat{v}_j &= c_3\kappa_+^j + c_4\kappa_-^j
\end{aligned} \tag{40}$$

where $\kappa_{+,-}$ are the same as in (20) and $i, j = 1, \ldots, N-1$. By substituting (40) into (39) we get the matrix equation $E(r,s,\kappa)w = 0$ for the unknowns

$$w = [\hat{u}_0, c_1, c_2, \hat{u}_N, \hat{v}_0, c_3, c_4, \hat{v}_N]^T \tag{41}$$

where

$$E(r,s,\kappa) = \begin{bmatrix}
\tilde{s}+2 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
-2 & 1 & 1 & 0 & 0 & 0 & 0 & 0\\
0 & e_{3,2} & e_{3,3} & e_{3,4} & e_{3,5} & e_{3,6} & e_{3,7} & 0\\
0 & e_{4,2} & e_{4,3} & e_{4,4} & e_{4,5} & e_{4,6} & e_{4,7} & 0\\
0 & e_{5,2} & e_{5,3} & e_{5,4} & e_{5,5} & e_{5,6} & e_{5,7} & 0\\
0 & e_{6,2} & e_{6,3} & e_{6,4} & e_{6,5} & e_{6,6} & e_{6,7} & 0\\
0 & 0 & 0 & 0 & 0 & \kappa_+^N & \kappa_-^N & -2\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \tilde{s}+2
\end{bmatrix} \tag{42}$$

with coefficients $e_{i,j}$ given by

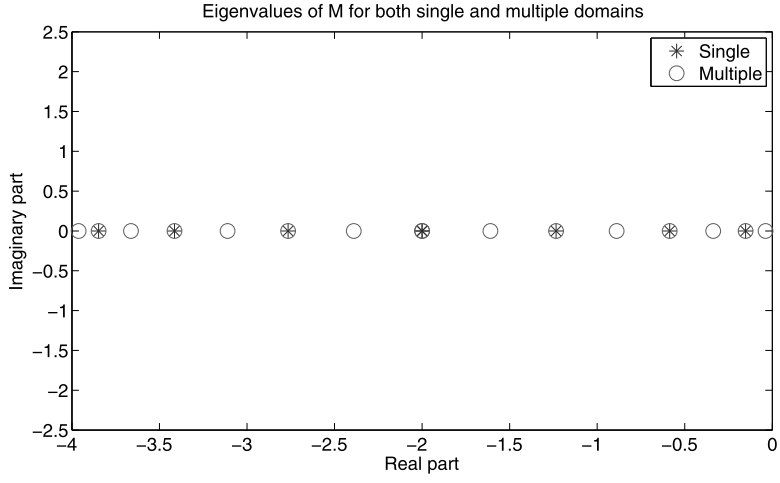


Fig. 1. Eigenvalues of the single domain operator with 9 grid points and the multi domain operator with $9+9$ grid points scaled with Δx^2. The single domain operator spectrum is always contained in the multi domain operator spectrum independent of the interface parameter r. There is a triple root at $\tilde{s}=-2$.

$$
\begin{aligned}
&e_{3,2}=\kappa_+^N, && e_{3,3}=\kappa_-^N, && e_{3,4}=-(2+r),\\
&e_{3,5}=1+r, && e_{3,6}=0, && e_{3,7}=0,\\
&e_{4,2}=2r\kappa_+^{N-1}, && e_{4,3}=2r\kappa_-^{N-1}, && e_{4,4}=\tilde{s}+2,\\
&e_{4,5}=-2(1+2r), && e_{4,6}=2r\kappa_+, && e_{4,7}=2r\kappa_-,\\
&e_{5,2}=-2(1+r)\kappa_+^{N-1}, && e_{5,3}=-2(1+r)\kappa_-^{N-1}, && e_{5,4}=2(1+2r),\\
&e_{5,5}=\tilde{s}+2, && e_{5,6}=-2(1+r)\kappa_+, && e_{5,7}=-2(1+r)\kappa_-,\\
&e_{6,2}=0, && e_{6,3}=0, && e_{6,4}=-r,\\
&e_{6,5}=-(1-r), && e_{6,6}=1, && e_{6,7}=1.
\end{aligned}
\tag{43}
$$

As before we obtain the spectrum by computing all values of $\tilde{s}$ such that $\det(E(r,s,\kappa))=0$. It is easy to see by expanding the determinant by the first and last row that

$$\det\big(E(r,s,\kappa)\big)=-(\tilde{s}+2)^2\det\big(\tilde{E}(r,s,\kappa)\big) \tag{44}$$

where $\tilde{E}(r,s,\kappa)$ is the inner 6×6 matrix. The determinant of $\tilde{E}(r,s,\kappa)$ is somewhat more complicated but by expanding it further and factorizing we get

$$\det\big(E(r,s,\kappa)\big)=(\tilde{s}+2)\big(\kappa_-^N-\kappa_+^N\big)f(r,s,\kappa). \tag{45}$$

We can see that the two first factors in (45) are exactly (24). Thus the spectrum from the single domain operator is contained in the multi domain operator spectrum. This is visualized in Fig. 1. The last factor $f(r,s,\kappa)$ is given explicitly by

$$
\begin{aligned}
f(r,s,\kappa)=&\big(16r^2+16r+\tilde{s}^2+4\tilde{s}+8\big)\big(\kappa_-^N-\kappa_+^N\big)\\
&+2\big(8r^3+12r^2+2r\tilde{s}+8r+\tilde{s}+2\big)\big(\kappa_+^N\kappa_--\kappa_+\kappa_-^N\big)\\
&+2\big(2r^2\tilde{s}-4r^2+4r\tilde{s}-4r+\tilde{s}-2\big)\big(\kappa_-^{N-1}-\kappa_-^{N-1}\big).
\end{aligned}
\tag{46}
$$

A closed form for the zeros of (46) have not been found. However, we can numerically compute the zeros.

3.3. Influence of the type of coupling

The type of coupling depends on the interface parameter r in (38) and by varying it, the spectral properties are modified. The interface parameter can be considered as a weight between Dirichlet and Neumann conditions. When $r=0$ or $r=-1$, some of the terms in (38) are canceled and renders the scheme one-sided coupled in the sense that one domain gives its value to the other domain and receives the value of the gradient. Since the extremal values are $r=-1,0$ one might expect

that something interesting happens when $r = -\frac{1}{2}$, that is when the two equations are coupled symmetrically. The case $r = -\frac{1}{2}$ will be denoted as the symmetric coupling and all other cases as unsymmetric coupling.

By considering the equations that are modified at the interface,

$$\begin{aligned}
\frac{\partial}{\partial t}u_{N-1} &= \frac{u_{N-2} - 2u_{N-1} + (2+r)u_N - (1+r)v_0}{\Delta x^2},\\
\frac{\partial}{\partial t}u_N &= \frac{-2ru_{N-1} - 2u_N + 2(1+2r)v_0 - 2rv_1}{\Delta x^2},\\
\frac{\partial}{\partial t}v_0 &= \frac{2(1+r)u_{N-1} - 2(1+2r)u_N - 2v_0 + 2(1+r)v_1}{\Delta x^2},\\
\frac{\partial}{\partial t}v_1 &= \frac{ru_N + (1-r)v_0 - 2v_1 + v_2}{\Delta x^2},
\end{aligned} \tag{47}$$

we can easily see how the difference scheme is modified due to the choice of r.

By taking an exact solution $w(x,t)$ to (30) we can by Taylor expanding (47) determine the accuracy. To simplify the notation we drop the indices and expand all equations around $x_j = x_*$. We get

$$\begin{aligned}
\frac{\partial}{\partial t}w(x_*,t) &= w_{xx}(x_*,t) + O\left(\Delta x^2\right),\\
\frac{\partial}{\partial t}w(x_*,t) &= -2rw_{xx}(x_*,t) + O\left(\Delta x^2\right),\\
\frac{\partial}{\partial t}w(x_*,t) &= 2(1+r)w_{xx}(x_*,t) + O\left(\Delta x^2\right),\\
\frac{\partial}{\partial t}w(x_*,t) &= w_{xx}(x_*,t) + O\left(\Delta x^2\right)
\end{aligned} \tag{48}$$

for the corresponding equations in (47). We can now easily see that we obtain the second order accurate second derivative only for $r = -\frac{1}{2}$. Even though some of the above equations correspond to inconsistent approximations of the second derivative, the global accuracy of the operator remain unchanged [14,13,23].

3.3.1. Stiffness and convergence to steady-state

To see how the stiffness is affected by the interface treatment we plot the largest absolute value of the eigenvalues of the discretization matrix as a function of r in Fig. 2. We can see that with increasing magnitude of r, the discretization become more stiff as expected. More unexpected is that the stiffness is slightly reduced below that of the uncoupled equations by choosing an unsymmetric coupling. This is contrary to the result in [12]. However in that paper a wide operator was used which have a different set of eigenvalues.

It is beneficial for the rate of convergence to steady-state with a discretization which have its real parts of the spectrum bounded away from zero as far as possible [22,4,16,3]. In Fig. 3 we show the real part of the spectrum closest to zero as a function of r.

We have used 33 grid points for the single domain in both Fig. 2 and Fig. 3, hence the coupled domains have $17 + 17$ grid points in total. The computation of the rightmost lying eigenvalue in Fig. 3 is resolved and the variation with r is small. For a coarse mesh the convergence to steady-state can be slightly improved by having an unsymmetric coupling. This is again contrary to the result in [12].

3.3.2. Error and convergence analysis

We will use the method of manufactured solutions to study the error as a function of the interface parameter r. Any function $v \in C^2$ is a solution to

$$\begin{aligned}
&u_t = u_{xx} + F(x,t), \quad -1 \leqslant x \leqslant 1,\\
&u(x,0) = v(x,0),\\
&u(-1,t) = v(-1,t),\\
&u(1,t) = v(1,t)
\end{aligned} \tag{49}$$

where the forcing function $F(x,t)$ has been chosen appropriately. In this particular case we choose

$$v(x,t) = \frac{\sin(2\pi x - t) + \sin(t)}{4} \tag{50}$$

which satisfies (49) with homogeneous boundary conditions and

$$F(x,t) = \frac{\cos(t) - \cos(2\pi x - t) + \pi^2 \sin(2\pi x - t)}{4}. \tag{51}$$

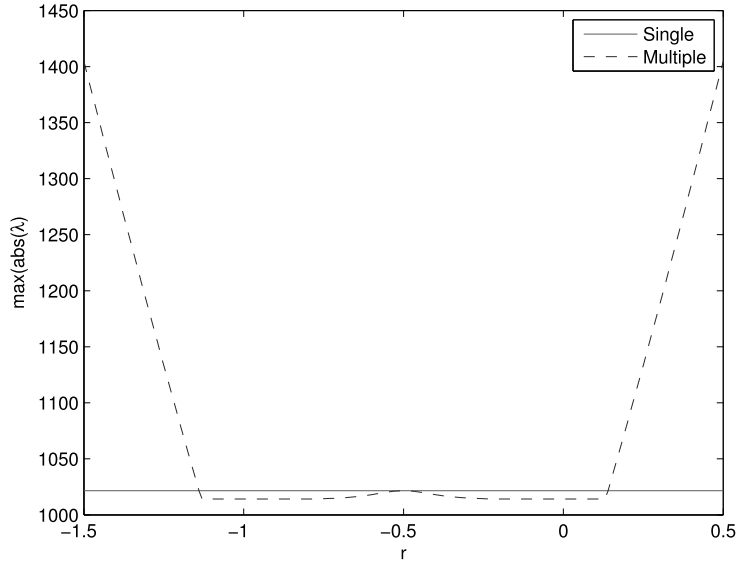


Fig. 2. max(abs(λ)) as a function of r. The single domain operator is using 33 grid points and the multi domain operator is using $17 + 17$ grid points.

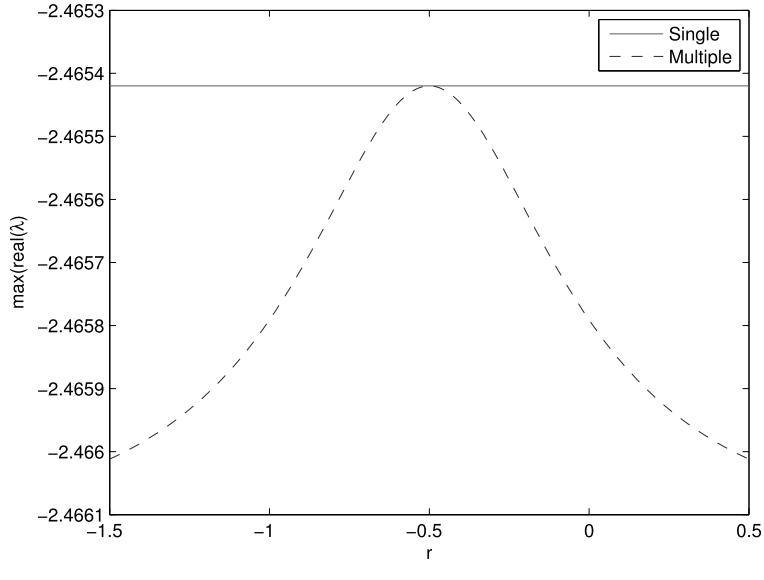


Fig. 3. max($\mathbb{R}$(λ)) as a function of r. The single domain operator is using 33 grid points and the multi domain operator is using $17 + 17$ grid points.

The spatial discretization and stability conditions are the same as before and we use the classical 4th-order Runge–Kutta time integration scheme to solve a system of the form

$$\frac{\partial \psi}{\partial t} = M\psi + F. \tag{52}$$

All spatial discretization, including boundary and interface conditions, is included in M and F is the above forcing function in discrete vector form. Thus we have an analytical solution which we can use to study the errors. In [10] it is stated that the errors can be reduced depending on the interface coupling for a hyperbolic problem and we will investigate if a similar effect exist for a parabolic problem.

Table 1
Error and convergence results using N grid points in each subdomain.

N	$r=-1$			$r=0$			$r=-1/2$		
	l_∞	l_2	q_2	l_∞	l_2	q_2	l_∞	l_2	q_2
8	0.1932	0.1301		0.1938	0.1302		0.1457	0.1270	
16	0.0504	0.0331	1.9742	0.0505	0.0331	1.9745	0.0377	0.0318	1.9979
32	0.0128	0.0083	1.9902	0.0128	0.0083	1.9903	0.0096	0.0079	2.0022
64	0.0032	0.0021	1.9953	0.0032	0.0021	1.9953	0.0024	0.0020	2.0013
128	0.0008	0.0005	1.9978	0.0008	0.0005	1.9978	0.0006	0.0005	2.0008
256	0.0002	0.0001	1.9989	0.0002	0.0001	1.9989	0.0002	0.0001	2.0004

In Table 1 we summarize the result. The solution is taken at time $t=\frac{\pi}{2}$. We show the errors in the l_∞ and l_2 norms for different resolutions together with the rate of convergence q_2 in the l_2-norm for the interesting values of r. We can see that the errors and order of convergence are only slightly better when using the symmetric coupling.

4. Single domain spectral analysis of the advection equation

We shall perform an analogous analysis for the advection equation to see if similar results hold for the advection operator.

4.1. Continuous case

Consider the advection equation in one domain,

$$\begin{aligned} &u_t + u_x = 0, \quad -1 \leqslant x \leqslant 1, \\ &u(-1,t) = g(t), \\ &u(x,0) = f(x). \end{aligned} \tag{53}$$

Eq. (53) is significantly different from (1) due to the directionality of the spatial operator. In this case there is one signal traveling from left to right and hence only one boundary condition is needed at $x=-1$. To obtain the spectrum we take the Laplace transform of (53) and proceed as before. We get

$$s\hat{u} + \hat{u}_x = 0 \tag{54}$$

which has the characteristic equation

$$\kappa + s = 0 \tag{55}$$

and thus the general solution of (54) is

$$\hat{u} = ce^{-sx}. \tag{56}$$

If we apply the boundary condition with $g=0$ we get $c=0$ and thus $\hat{u}=0$. Hence there is no continuous spectrum of (53) since there are no values of s such that $ce^s=0$ for $c\neq 0$.

4.2. Discrete case

We discretize (53) using the SBP and SAT technique on a uniform mesh of $N+1$ grid points

$$u_t + P^{-1}Q\,v = \sigma P^{-1}(v_0 - g)e_0 \tag{57}$$

where P and e_0 are as before and

$$Q = \frac{1}{2}\begin{bmatrix} -1 & 1 & 0 & 0 & \cdots & 0 \\ -1 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -1 & 0 & 1 \\ 0 & \cdots & 0 & 0 & -1 & 1 \end{bmatrix}, \qquad P^{-1}Q = \frac{1}{2\Delta x}\begin{bmatrix} -2 & 2 & 0 & 0 & \cdots & 0 \\ -1 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 & -1 \\ 0 & \cdots & 0 & 0 & -2 & 2 \end{bmatrix}. \tag{58}$$

Note that $Q+Q^T = \mathrm{diag}(-1,0,\ldots,0,1)$ which is used to select the boundary terms in the energy estimate. By applying the energy method to (57) with $g=0$ we get

$$\|v\|_t^2 = (1+2\sigma)v_0^2 - v_N^2 \tag{59}$$

which is bounded for $\sigma \leqslant -\frac{1}{2}$ and hence the scheme is stable.

To determine the spectrum we Laplace transform (57) (with $g=0$) and rewrite as

$$\big(sI + P^{-1}Q - \sigma P^{-1}E_0\big)\hat{v} = 0. \tag{60}$$

From the internal scheme we have

$$s\hat{v}_i + \frac{1}{2\Delta x}(\hat{v}_{i+1} - \hat{v}_{i-1}) = 0 \tag{61}$$

or equivalently

$$\hat{v}_{i+1} + 2\tilde{s}\hat{v}_i - \hat{v}_{i-1} = 0 \tag{62}$$

with $\tilde{s} = s\Delta x$. The characteristic equation is $\kappa^2 + 2\tilde{s}\kappa - 1 = 0$ which has solutions

$$\kappa_{+,-} = -\tilde{s} \pm \sqrt{\tilde{s}^2+1}. \tag{63}$$

Thus the general solution of (62) is

$$\hat{v}_i = c_1\kappa_+^i + c_2\kappa_-^i. \tag{64}$$

The first and last equation in (60) are modified and we can use them to write a matrix equation for the unknowns $c_{1,2}$. The equations are

$$\begin{aligned} &(\tilde{s} - 1 - 2\sigma)\hat{v}_0 + \hat{v}_1 = 0, \\ &-v_{N-1} + (\tilde{s}+1)\hat{v}_N = 0 \end{aligned} \tag{65}$$

and by inserting the general solution (64) into (65) we get the matrix equation $E(s,\kappa)c = 0$ where

$$E(s,\kappa) = \begin{bmatrix} \tilde{s} - 1 - s\sigma + \kappa_+ & \tilde{s} - 1 - s\sigma + \kappa_- \\ (\tilde{s}+1)\kappa_+^N - \kappa_+^{N-1} & (\tilde{s}+1)\kappa_-^N - \kappa_-^{N-1} \end{bmatrix}. \tag{66}$$

The spectrum consists as before of the singular points of $E(s,\kappa)$. A direct computation of the determinant of $E(s,\kappa)$ gives that

$$\begin{aligned} \det\big(E(s,\kappa)\big) = {} & \kappa_-^N\big(\sqrt{\tilde{s}^2+1} - 1 - 2\sigma\big)\big(1 + \sqrt{\tilde{s}^2+1}\big) \\ & + \kappa_+^N\big(\sqrt{\tilde{s}^2+1} + 1 + 2\sigma\big)\big(1 - \sqrt{\tilde{s}^2+1}\big). \end{aligned} \tag{67}$$

A closed form expression for the zeros of (67) have not been found. We can however compute the eigenvalues numerically. We will return to (67) when we consider the spectrum of the coupled problem.

5. Multi domain spectral analysis of the advection equation

We introduce again an artificial interface at $x=0$ for the advection equation to study how the spectral properties of the continuous and discrete operators are modified.

5.1. Continuous case

Consider now

$$\begin{aligned} &u_t + u_x = 0, \quad -1 \leqslant x \leqslant 0, \\ &v_t + v_x = 0, \quad 0 \leqslant x \leqslant 1, \\ &u(-1,t) = g(t), \\ &v(0,t) = u(0,t). \end{aligned} \tag{68}$$

The spectrum is again obtained by Laplace transforming (68) and applying the boundary and interface conditions. The general solutions to the Laplace transformed equations are $\hat{u} = c_1 e^{-sx}$ and $\hat{v} = c_2 e^{-sx}$. The boundary and interface conditions imply that $c_1 = 0$ and $c_2 = c_1$, and hence there is no spectrum as expected.

5.2. Discrete case

One form of the SBP and SAT discretization of (68) is

$$\begin{aligned} u_t + P^{-1} Q u &= \sigma P^{-1}(u_0 - g)e_0, \\ v_t + P^{-1} Q v &= \tau P^{-1}(v_0 - u_N)e_0 \end{aligned} \tag{69}$$

where u, v now denote the discrete grid functions. Both domains have equidistant grid spacing and equal number of grid points to allow for the same difference operators in both domains.

5.2.1. Conservation and stability

When constructing an interface for equations with advection it is important that the scheme is not only stable, but also conservative [18,1,2]. Let $\Phi(x)$ be a smooth testfunction and let $\phi = [\Phi(x_0), \ldots, \Phi(x_N)]^T$. We multiply both equations in (69) with $\phi^T P$ respectively. By using the SBP property of Q and adding the two equations we can shift the differentiation onto ϕ and get

$$\phi^T P u_t + \phi^T P v_t - \phi_0 u_0 + \phi_N v_N - (Q\phi)^T u - (Q\phi)^T v - \sigma\phi_0(u_0 - g) = \phi_i(\tau + 1)(v_0 - u_N) \tag{70}$$

where we have used $\phi_0 = \Phi(-1)$, $\phi_N = \Phi(1)$ and $\phi_i = \Phi(0)$ to denote the boundary and interface points. Conservation requires that the right-hand side of (70) is zero, and hence we need to put $\tau = -1$ to cancel the remaining terms. With this choice we thus have a conservative interface treatment.

To determine the stability condition we proceed with the energy method as before and multiply both equations in (69) with $u^T P$ and $v^T P$ respectively. By assuming that $g = 0$ we get

$$\|u\|_t^2 + \|v\|_t^2 = (1 + 2\sigma)u_0^2 + (1 + \tau)v_0^2 - v_N^2 - (u_N + \tau v_0)^2 \tag{71}$$

and we can see that the scheme is stable if we chose $\sigma \leqslant -\frac{1}{2}$ and $\tau \leqslant -1$. Thus the interface treatment is both stable and conservative with $\tau = -1$.

5.2.2. Eigenspectrum

We Laplace transform (69) and get the general solution from the internal schemes as before,

$$\begin{aligned} \hat{u}_i &= c_1\kappa_+^i + c_2\kappa_-^i, \\ \hat{v}_i &= c_3\kappa_+^i + c_4\kappa_-^i, \end{aligned} \tag{72}$$

where $\kappa_{+,-} = -\tilde{s} \pm \sqrt{\tilde{s}^2 + 1}$. The scheme at the boundaries and interfaces are different from the internal scheme and their corresponding equations are

$$\begin{aligned} &(\tilde{s} - 1 - 2\sigma)\hat{u}_0 + \hat{u}_1 = 0, \\ &(\tilde{s} + 1)\hat{u}_N - \hat{u}_{N-1} = 0, \\ &2\tau\hat{u}_N + (\tilde{s} - 1 - 2\tau)\hat{v}_0 + \hat{v}_1 = 0, \\ &(\tilde{s} + 1)\hat{v}_N - \hat{v}_{N-1} = 0. \end{aligned} \tag{73}$$

By inserting the general solutions into (73) we get again the matrix equation $E(s, \kappa)c = 0$ for the unknowns $c = [c_1, \ldots, c_4]^T$ where

$$E(s,\kappa) = \begin{bmatrix} \tilde{s} - 1 - 2\sigma + \kappa_+ & \tilde{s} - 1 - 2\sigma + \kappa_- & 0 & 0 \\ (\tilde{s} + 1)\kappa_+^N - \kappa_+^{N-1} & (\tilde{s} + 1)\kappa_-^N - \kappa_-^{N-1} & 0 & 0 \\ 2\tau\kappa_+^N & 2\tau\kappa_-^N & \tilde{s} - 1 - 2\tau + \kappa_+ & \tilde{s} - 1 - 2\tau + \kappa_- \\ 0 & 0 & (\tilde{s} + 1)\kappa_+^N - \kappa_+^{N-1} & (\tilde{s} + 1)\kappa_-^N - \kappa_-^{N-1} \end{bmatrix}. \tag{74}$$

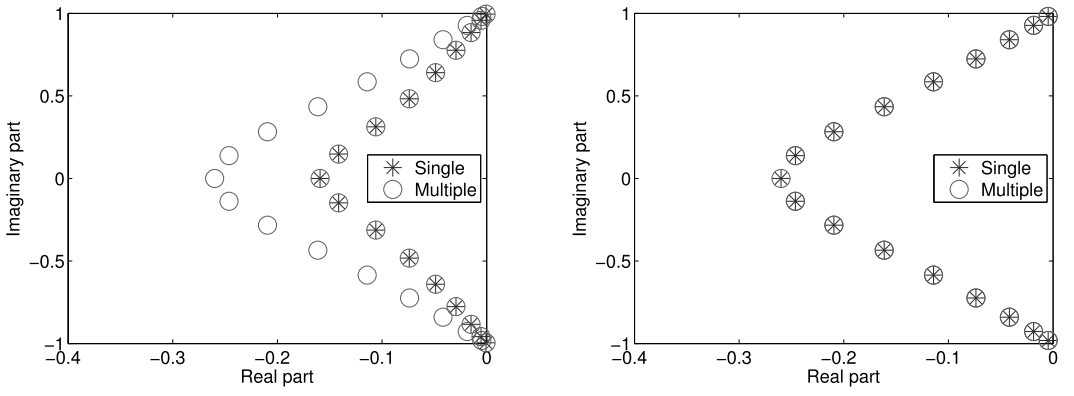
The spectrum is obtained for the singular values of $E(s, \kappa)$. By expanding the determinant of $E(s, \kappa)$ and factorizing we get

$$\det\big(E(s,\kappa)\big) = f(\sigma, s)g(\tau, s) \tag{75}$$

where

$$f(\sigma, s) = \kappa_-^N\big(\sqrt{\tilde{s}^2 + 1} - 1 - 2\sigma\big)\big(1 + \sqrt{\tilde{s}^2 + 1}\big) + \kappa_+^N\big(\sqrt{\tilde{s}^2 + 1} + 1 + 2\sigma\big)\big(1 - \sqrt{\tilde{s}^2 + 1}\big) \tag{76}$$

is exactly (67). The second factor is



(a) $\sigma = -\frac{1}{2}$ (b) $\sigma = -1$

Fig. 4. Spectrum of both the single and multi domain operator, scaled with Δx, using 17 grid points for the single domain operator and $17 + 17$ for the multi domain operator.

$$\begin{aligned} g(\tau, s) = {} & \big(\tilde{s}^2 - 1 - 2\tau\tilde{s} - 2\tau + \tilde{s}\kappa_+ + \kappa_+\big)\kappa_-^N \\ & - \big(\tilde{s}^2 - 1 - 2\tau\tilde{s} - 2\tau + \tilde{s}\kappa_- + \kappa_-\big)\kappa_+^N \\ & - (\tilde{s} - 1 - 2\tau + \kappa_+)\kappa_-^{N-1} + (\tilde{s} - 1 - 2\tau + \kappa_-)\kappa^{N-1}. \end{aligned} \tag{77}$$

We can see that the single domain operator spectrum is again contained in the multi domain operator spectrum, which is visualized in Fig. 4(a) and Fig. 4(b) for $\sigma = -\frac{1}{2}$ and $\sigma = -1$ respectively. In the second case, which is fully upwinded, we can see that the spectrum is identical for the single and multi domain operators since all eigenvalues of the multi domain operator are double eigenvalues.

5.3. Extending the interface treatment

In the previous section we discussed one of many different schemes for the advection equation coupled over an interface. The scheme was based on the boundary and interface conditions for the continuous PDE. The interface condition $v = u$ is of course identical to $u = v$ in the continuous sense, but this is not true in the discrete setting with weak interface conditions. We can hence modify the interface treatment by adding one additional term corresponding to $u = v$ in the discrete setting.

Consider the scheme (69) again but without the outer boundary term and with one additional term added,

$$\begin{aligned} u_t + P^{-1}Q\,u &= \gamma P^{-1}(u_N - v_0)e_N, \\ v_t + P^{-1}Q\,v &= \tau P^{-1}(v_0 - u_N)e_0. \end{aligned} \tag{78}$$

The stability and conservation criteria can be found in e.g. [5] so we just state the result here as a proposition,

Proposition 5.1. *The interface scheme* (78) *is stable and conservative for*

$$\gamma = \frac{1-\theta}{2}, \qquad \tau = -\frac{1+\theta}{2}, \tag{79}$$

where $\theta \geqslant 0$ *is a free parameter.*

The energy estimate of (78) is

$$\|u\|_t^2 + \|v\|_t^2 = -\theta\big(u_N^2 - v_0^2\big) \tag{80}$$

when ignoring the outer boundary terms. Note that $\theta = 1$ gives (69) which is fully upwinded while $\theta = 0$ gives minimal dissipation. By Taylor expanding (78) it can easily be seen that the formal accuracy is independent of the choice of θ. We did a convergence study and verified that the solution converges with second order accuracy independently of the choice of parameters.

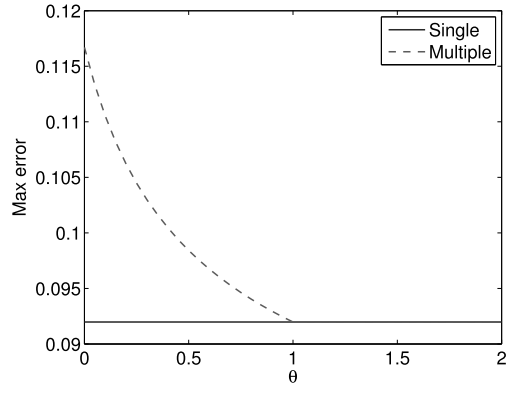


Fig. 5. l_∞-error of the single and multi domain operator as a function of θ using 33 and $17+17$ grid points respectively.

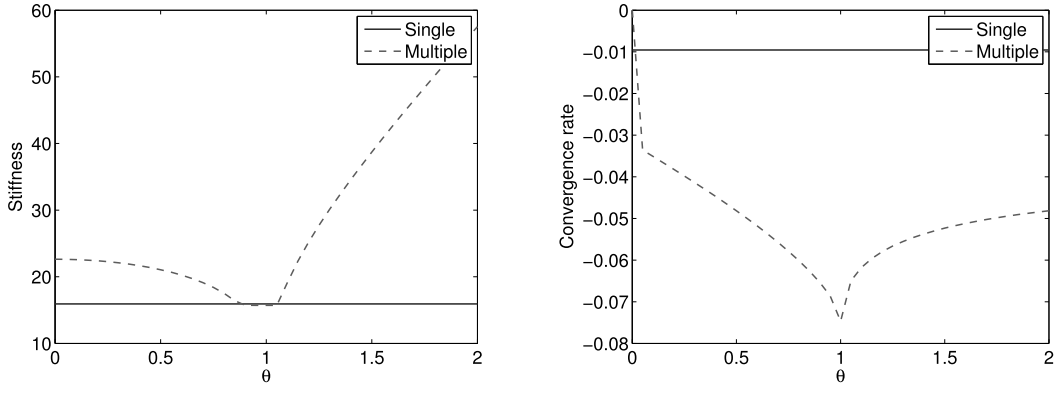


Fig. 6. Stiffness and rate of convergence as a function of θ.

5.3.1. Errors, stiffness and convergence

In the case of advection there are two free parameters compared to the diffusion case where there is only one. One parameter for the outer boundary $-\infty \leqslant \sigma \leqslant -\frac{1}{2}$ and one parameter for the interface $0 \leqslant \theta \leqslant \infty$. Since we are interested only in the interface treatment we let $\sigma = -1$ be fixed and consider the stiffness, rate of convergence and error as a function of θ.

In [10] the quasi-one-dimensional Euler equations were used with an interface treatment corresponding to $\theta = 1$ to study the errors. Their convergence study showed that the errors were small and do not increase with the number of subdomains. We continue with a more detailed investigation by posing the errors as functions of the interface treatment.

We use the manufactured solution $u(x,t) = \sin(2\pi(x-t))$ to study the errors as a function of θ. Using this solution we construct initial and boundary data to use in the error analysis. The maximum error is shown in Fig. 5. For $\theta \geqslant 1$, the maximum errors are indistinguishable to machine precision. Compared to the minimal dissipative case, the maximum error is approximately 20 percent smaller.

The stiffness and rate of convergence are shown in Fig. 6 where 33 grid points are used for the single domain and $17+17$ for the multi domain. We can see from Figs. 6(a) and 6(b) that it is possible to maintain and even improve the stiffness and rate of convergence when $\theta = 1$ which is the fully upwinded scheme.

From Fig. 6(b) we can see that the maximum real part of the spectrum is reduced by approximately a factor seven when $\theta = 1$, which is when the interface is fully upwinded. We can visualize this effect by performing a steady-state computation and measure the errors in the steady-state solution.

We consider the initial data given by

$$f(x) = e^{-100(x-x_0)^2} \tag{81}$$

Table 2
Error and convergence results for 1, 2, 4 and 8 domains.

# domains	1		2		4		8	
$2/\Delta x$	l_2	q_2	l_2	q_2	l_2	q_2	l_2	q_2
32	1.32e−01		1.25e−01		1.09e−01		9.76e−02	
64	3.90e−02	1.7556	3.80e−02	1.7194	3.58e−02	1.6015	3.26e−02	1.5819
128	2.77e−03	3.8131	2.76e−03	3.7850	2.66e−03	3.7493	1.06e−03	4.9386
256	6.65e−04	2.0596	6.64e−04	2.0558	5.75e−04	2.2107	4.55e−05	4.5470
512	1.66e−04	2.0055	1.65e−04	2.0041	1.26e−04	2.1852	5.12e−06	3.1510
1024	4.14e−05	2.0014	4.13e−05	2.0007	3.00e−05	2.0748	8.50e−07	2.5900
2048	1.03e−05	2.0003	1.03e−05	2.0000	7.37e−06	2.0237	1.58e−07	2.4291

Table 3
The time at which the l_2-norm of the solution is less than 10^{-16} for 1, 2, 4 and 8 domains with upwinded interfaces.

# domains	1	2	4	8
t	6021.6	912.0	146.5	28.8

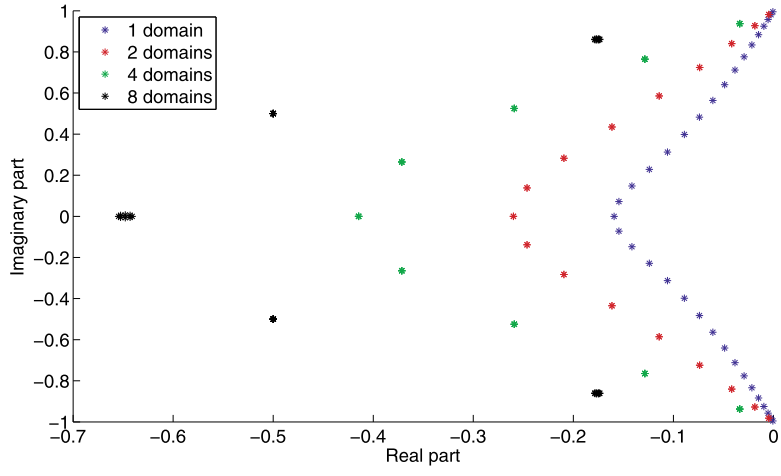


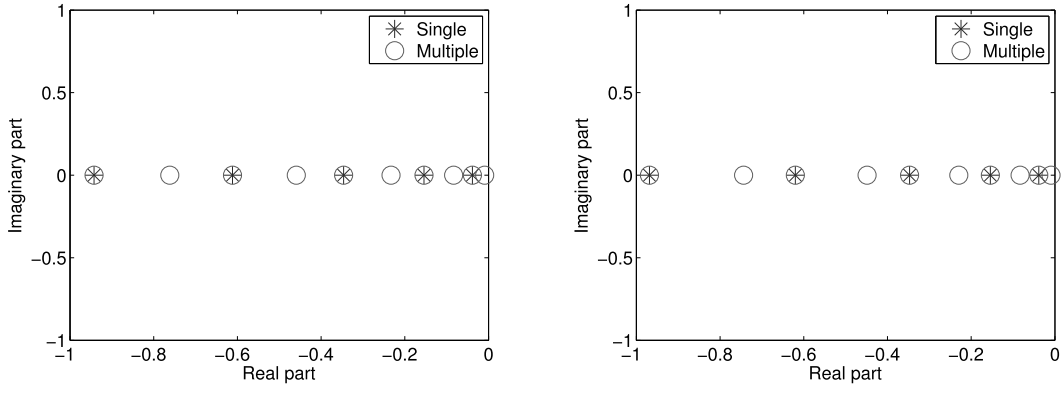
Fig. 7. Spectra of the 1-, 2-, 4- and 8-domain operator scaled with Δx for $\theta = 1$. Note that the eigenvalues of the 8-domain operator are clustered.

where $x_0 = -\frac{1}{2}$. The disturbance is transported out of the right boundary and the exact steady-state solution is identically zero. At time $t = 2$ when the initial disturbance have left the computational domain we measure the errors and rate of convergence for 1, 2, 4 and 8 domains with $\theta = 1$. The number of grid points in each subdomain is chosen such that the resolution is the same for all number of subdomains. The results are seen in Table 2. As we can see from Table 2, when using 8 subdomains, the steady-state errors are significantly smaller compared to the single or two domain case. For high resolutions the error is two orders of magnitude smaller and the rate of convergence is still higher.

In Table 3 it is shown how long it is needed to compute in time until the l_2-norm of the solution is less than 10^{-16} which is considered to be the steady-state solution. We can see that there is a huge increase in gain to reach steady-state when more upwinded interfaces are introduced. Note that the time to reach steady-state for one and two domains differ by almost a factor seven which is what we can expect from Fig. 6(b).

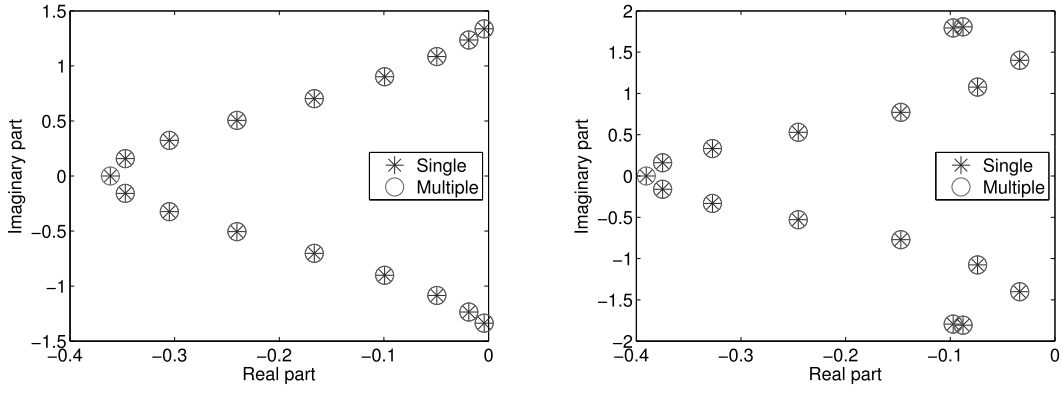
The spectra of all cases is seen in Fig. 7. We can see that the real part of the eigenvalues are shifted further to the left when more upwinded interfaces are added. Hence the accelerated convergence rate to steady-state [22]. Note that as more upwinded interfaces are added, the multiplicity of the eigenvalues increase and hence there will be less distinct eigenvalues.

Remark 5.1. Independently of the number of interfaces we can pose the complete scheme as $\tilde{P} w_t = \tilde{Q}\, w + F$ where $\tilde{P}$ is the norm and all differentiation is collected in $\tilde{Q}$. The minimally dissipative interface treatment with $\theta = 0$ renders $\tilde{Q}$ completely skew-symmetric except at the boundary points.



(a) 3rd-order (b) 4rd-order

Fig. 8. Spectra of the 3rd- and 4th-order accurate diffusion operators with $r = -\frac{1}{2}$.



(a) 3rd-order (b) 4th-order

Fig. 9. Spectra of the 3rd- and 4th-order accurate advection operators with $\sigma = -1$.

6. Higher order accurate approximations

The previous analysis was performed on the second order accurate operators since it was possible to derive analytical results. In this section we briefly show numerical results for the 3rd- and 4th-order accurate SBP operators. Details on the operators can be found in [14,20].

The stability and conservation criteria is independent of the order of accuracy. The schemes and stability conditions for the heat and advection equation are thus identical except that the difference operators have been replaced by 3rd- and 4th-order accurate operators.

The corresponding determinants of (42) and (74) for the higher order operators are not possible to compute and factor analytically. We can however compute the spectrum numerically. In Fig. 8 (diffusion) and Fig. 9 (advection) we show the analogues of Figs. 1 and 4(b) which shows that a similar factorization appears to exist even in the higher order cases. In Fig. 8, only a part of the spectrum is shown but the trend is continued throughout the spectrum. In Fig. 9 all eigenvalues of the multi domain operator are double eigenvalues. In both figures we have used 17 grid points for the single domain and $17 + 17$ for the multi domain operators.

7. Conclusions

7.1. Diffusion

In the single domain case, a closed form expression for all eigenvalues of the discretization matrix, including the boundary conditions, was found. We showed how the eigenvalues of the discretization matrix converged to the eigenvalues of the continuous equation. For the multiple domain case we showed how the spectrum of the single domain operator is contained in the multi domain operator spectrum independent of the interface treatment. Numerical experiments indicate that this inclusion generalizes to higher order accurate operators.

The stiffness and rate of convergence were not significantly effected by the choice of interface treatment. We used a manufactured solution to study the errors. When the symmetric coupling was used, the maximum error of the multi domain case reduced to the level of the single domain case. Compared to the unsymmetric coupling, the maximum errors were reduced by almost 35 percent when the symmetric coupling was used.

7.2. Advection

For the advection equation we showed that the spectrum of the single domain operator is contained in the multi domain operator spectrum independent of the interface treatment similarly to the diffusion case. A numerical computation of the spectrum indicated that the result carries over to higher order accurate operators.

The stiffness showed only minor differences depending on the interface treatment. The rate of convergence to steady-state was significantly improved when adding one upwinded interface. By adding more upwinded interfaces we could dramatically decrease the error in the steady-state calculation.

We used an exact solution to study the errors as a function of the interface treatment. We showed that it is possible to bring down the maximum errors to the level of the single domain case by using the upwinded coupling. The maximum error was about 20 percent smaller when using a fully upwinded coupling compared to the minimal dissipative coupling.

Appendix A. Double roots

When determining the solutions to the recurrence relation from the Laplace transformed scheme in the interior, one has to be careful with double roots of the characteristic equation. Due to the ansatz, false roots might be introduced and it is necessary to confirm whether or not these roots belong to the spectrum.

The characteristic equation (19) has double roots for $\tilde{s} = -4$ and $\tilde{s} = 0$. The solutions are

$$\kappa = -1, \qquad \kappa = 1 \tag{A.1}$$

respectively. The general solution to the recurrence relation is then

$$\hat{v}_i = (c_1 + c_2 i)\kappa^i. \tag{A.2}$$

We assume that the general solution (A.2) is valid for $i = 1, \ldots, N-1$ and insert into the modified boundary equations to get the matrix equation $E(s,\kappa)c = 0$ for the unknowns $c = [\hat{v}_0, c_1, c_2, \hat{v}_N]^T$ where

$$E(s,\kappa) = \begin{bmatrix} \tilde{s}+2 & 0 & 0 & 0 \\ -2 & ((\tilde{s}+2)-\kappa)\kappa & ((\tilde{s}+2)-2\kappa)\kappa & 0 \\ 0 & ((\tilde{s}+2)\kappa-1)\kappa^{N-2} & ((\tilde{s}+2)(N-1)\kappa-(N-2))\kappa^{N-2} & -2 \\ 0 & 0 & 0 & \tilde{s}+2 \end{bmatrix}. \tag{A.3}$$

By inserting $s = -4$ and $\kappa = -1$ into (A.3) we get $\det(E(s,\kappa)) = 4N(-1)^N \neq 0$. By inserting $\tilde{s} = 0$ and $\kappa = 1$ into (A.3) we get $\det(E(s,\kappa)) = 4N \neq 0$. Hence neither $\tilde{s} = -4$ nor $\tilde{s} = 0$ is a part of the spectrum.

References

[1] M.H. Carpenter, J. Nordström, D. Gottlieb, A stable and conservative interface treatment of arbitrary spatial accuracy, Journal of Computational Physics 148 (2) (1999) 341–365, http://dx.doi.org/10.1006/jcph.1998.6114.
[2] M.H. Carpenter, J. Nordström, D. Gottlieb, Revisiting and extending interface penalties for multi-domain summation-by-parts operators, Journal of Scientific Computing 45 (1–3) (2010) 118–150, http://dx.doi.org/10.1007/s10915-009-9301-5.
[3] P. Eliasson, S. Eriksson, J. Nordström, The influence of weak and strong solid wall boundary conditions on the convergence to steady-state of the Navier–Stokes equations, in: Proc. 19th AIAA CFD Conference, AIAA, 2009 (no. 2009-3551 in Conference Proceeding Series).
[4] B. Engquist, B. Gustafsson, Steady state computations for wave propagation problems, Mathematics of Computation 49 (179) (1987) 39–64.
[5] S. Eriksson, Q. Abbas, J. Nordström, A stable and conservative method for locally adapting the design order of finite difference schemes, Journal of Computational Physics 230 (11) (2011) 4216–4231 (special issue High Order Methods for CFD Problems), http://dx.doi.org/10.1016/j.jcp.2010.11.020.
[6] S. Eriksson, J. Nordström, Analysis of the order of accuracy for node-centered finite volume schemes, Applied Numerical Mathematics 59 (10) (2009) 2659–2676.
[7] B. Gustafsson, H.-O. Kreiss, J. Oliger, Time Dependent Problems and Difference Methods, Wiley Interscience, 1995.

[8] B. Gustafsson, H.-O. Kreiss, A. Sundström, Stability theory of difference approximations for mixed initial boundary value problems. II, Mathematics of Computation 26 (119) (1972) 649–686.
[9] W.D. Henshaw, K.K. Chand, A composite grid solver for conjugate heat transfer in fluid–structure systems, Journal of Computational Physics 228 (10) (2009) 3708–3741, http://dx.doi.org/10.1016/j.jcp.2009.02.007.
[10] X. Huan, J.E. Hicken, D.W. Zingg, Interface and boundary schemes for high-order methods, in: The 39th AIAA Fluid Dynamics Conference, San Antonio, USA, 2009 (AIAA Paper No. 2009-3658).
[11] H.-O. Kreiss, Stability theory for difference approximations of mixed initial boundary value problems. I, Mathematics of Computation 22 (104) (1968) 703–714.
[12] J. Lindström, J. Nordström, A stable and high-order accurate conjugate heat transfer problem, Journal of Computational Physics 229 (14) (2010) 5440–5456, http://dx.doi.org/10.1016/j.jcp.2010.04.010.
[13] K. Mattsson, Boundary procedures for summation-by-parts operators, Journal of Scientific Computing 18 (1) (2003) 133–153, http://dx.doi.org/10.1023/A:1020342429644.
[14] K. Mattsson, J. Nordström, Summation by parts operators for finite difference approximations of second derivatives, Journal of Computational Physics 199 (2) (2004) 503–540, http://dx.doi.org/10.1016/j.jcp.2004.03.001.
[15] A. Nissen, G. Kreiss, M. Gerritsen, Stability at nonconforming grid interfaces for a high order discretization of the Schrödinger equation, Tech. Rep. 2011-017, Uppsala University, Division of Scientific Computing, 2011.
[16] J. Nordström, The influence of open boundary conditions on the convergence to steady state for the Navier–Stokes equations, Journal of Computational Physics 85 (1) (1989) 210–244, http://dx.doi.org/10.1016/0021-9991(89)90205-2.
[17] J. Nordström, S. Eriksson, Fluid structure interaction problems: the necessity of a well posed, stable and accurate formulation, Communications in Computational Physics 8 (2010) 1111–1138.
[18] J. Nordström, J. Gong, E. van der Weide, M. Svärd, A stable and conservative high order multi-block method for the compressible Navier–Stokes equations, Journal of Computational Physics 228 (24) (2009) 9020–9035, http://dx.doi.org/10.1016/j.jcp.2009.09.005.
[19] J. Nordström, M. Svärd, Well-posed boundary conditions for the Navier–Stokes equations, SIAM Journal on Numerical Analysis 43 (3) (2005) 1231–1255, http://dx.doi.org/10.1137/040604972.
[20] B. Strand, Summation by parts for finite difference approximations for d/dx, Journal of Computational Physics 110 (1) (1994) 47–67, http://dx.doi.org/10.1006/jcph.1994.1005.
[21] M. Svärd, M.H. Carpenter, J. Nordström, A stable high-order finite difference scheme for the compressible Navier–Stokes equations, far-field boundary conditions, Journal of Computational Physics 225 (1) (2007) 1020–1038, http://dx.doi.org/10.1016/j.jcp.2007.01.023.
[22] M. Svärd, K. Mattsson, J. Nordström, Steady-state computations using summation-by-parts operators, Journal of Scientific Computing 24 (1) (2005) 79–95, http://dx.doi.org/10.1007/s10915-004-4788-2.
[23] M. Svärd, J. Nordström, On the order of accuracy for difference approximations of initial-boundary value problems, Journal of Computational Physics 218 (1) (2006) 333–352, http://dx.doi.org/10.1016/j.jcp.2006.02.014.
[24] M. Svärd, J. Nordström, A stable high-order finite difference scheme for the compressible Navier–Stokes equations: No-slip wall boundary conditions, Journal of Computational Physics 227 (10) (2008) 4805–4824, http://dx.doi.org/10.1016/j.jcp.2007.12.028.

Paper III

Journal of Computational Physics 230 (2011) 7519–7532

Contents lists available at ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp

Stable Robin solid wall boundary conditions for the Navier–Stokes equations

Jens Berg [a,*], Jan Nordström [b]

[a] Uppsala University, Department of Information Technology, SE-751 05 Uppsala, Sweden
[b] Linköping University, Department of Mathematics, SE-581 83 Linköping, Sweden

ARTICLE INFO

Article history:
Received 15 April 2011
Received in revised form 21 June 2011
Accepted 24 June 2011
Available online 5 July 2011

Keywords:
Navier–Stokes
Robin boundary conditions
Well-posedness
Stability
High order accuracy
Summation-By-Parts

ABSTRACT

In this paper we prove stability of Robin solid wall boundary conditions for the compressible Navier–Stokes equations. Applications include the no-slip boundary conditions with prescribed temperature or temperature gradient and the first order slip-flow boundary conditions. The formulation is uniform and the transitions between different boundary conditions are done by a change of parameters. We give different sharp energy estimates depending on the choice of parameters.

The discretization is done using finite differences on Summation-By-Parts form with weak boundary conditions using the Simultaneous Approximation Term. We verify convergence by the method of manufactured solutions and show computations of flows ranging from no-slip to almost full slip.

1. Introduction

There has recently been a development of stable boundary [1,2] and interface [3] conditions of a specific form for the compressible Navier–Stokes equations. This paper extends the result in [2] to more general solid wall boundary conditions and includes sharp energy estimates. While [2] deals only with the no-slip boundary conditions, we will provide a uniform formulation which includes the no-slip boundary conditions with prescribed temperature or temperature gradient and slip-flow boundary conditions or any combination thereof.

The tools that we will use to obtain a uniform formulation together with proof of stability are finite difference approximations on Summation-By-Parts (SBP) form together with the Simultaneous Approximation term. This method has the benefit of being stable by construction for any linear well-posed Cauchy problem [4,5] and the robustness has been shown in a wide range of applications [5–9].

The first derivative is approximated by $u_x \approx Dv = P^{-1}Qv$, where v is the discrete grid function, D is the differentiation matrix, $P = P^T > 0$ defines a norm by $\|v\|^2 = v^T P v$ and Q has the SBP property $Q + Q^T = B = [-1, 0, \ldots, 0, 1]^T$, see [10,11] for details about these operators.

There exist operators accurate of order 2, 4, 6 and 8 and the stability analysis does not depend on the order of accuracy of the operators. We will pose our equations on conservative form and hence we do not need an operator approximating the second derivative. Operators approximating the second derivative with constant coefficients are derived in [11] and have recently been developed for variable coefficients problems in [12].

* Corresponding author. Address: Division of Scientific Computing, Department of Information Technology, Uppsala University, Box 337, SE-751 05 Uppsala, Sweden. Tel.: +46 18 471 6253; fax: +46 18 523049/511925.
E-mail address: jens.berg@it.uu.se (J. Berg).

0021-9991/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2011.06.027

The boundary conditions will be imposed weakly using the Simultaneous Approximation Term (SAT). The SAT term is added to the right-hand-side of the discretized equations as a penalty term which forces the equation towards the boundary conditions. Together the SBP and SAT technique provide a tool for creating stable approximations for well-posed initial-boundary value problems. The relation between weak and strong boundary conditions in terms of accuracy is discussed in [13].

2. The Navier–Stokes equations

2.1. Continuous case

We consider the two-dimensional Navier–Stokes equations on conservative form

$$q_t + F_x + G_y = 0, \tag{1}$$

where

$$F = F^I - \varepsilon F^V, \quad G = G^I - \varepsilon G^V. \tag{2}$$

The superscript I denotes the inviscid part of the fluxes and V the viscous part. The components of the solution vector are $q = [\rho, \rho u, \rho v, e]^T$ which are the density, x- and y-directional momentum, respectively and energy. The components of the fluxes are given by

$$\begin{aligned} F^I &= [\rho u, p + \rho u^2, \rho u v, u(p+e)]^T, \\ G^I &= [\rho v, \rho u v, p + \rho v^2, v(p+e)]^T, \\ F^V &= [0, \tau_{xx}, \tau_{xy}, u\tau_{xx} + v\tau_{xy} - Q_x]^T, \\ G^V &= [0, \tau_{xy}, \tau_{yy}, u\tau_{yx} + v\tau_{yy} - Q_y]^T, \end{aligned} \tag{3}$$

where p is the pressure, Pr the Prandtl number, γ the ratio of specific heat and $Q = -\kappa T$ is the thermal conductivity times the temperature according to Fourier's law. The stress tensors are given by

$$\tau_{xx} = 2\mu\frac{\partial u}{\partial x} + \lambda\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right), \quad \tau_{yy} = 2\mu\frac{\partial v}{\partial y} + \lambda\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right), \quad \tau_{xy} = \tau_{yx} = \mu\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right), \tag{4}$$

where μ and λ are the dynamic and second viscosity, respectively.

All the equations above have been non-dimensionalized as

$$u = \frac{u^*}{c^*_\infty}, \quad v = \frac{v^*}{c^*_\infty}, \quad \rho = \frac{\rho^*}{\rho^*_\infty}, \quad T = \frac{T^*}{T^*_\infty}, \quad p = \frac{p^*}{\rho^*_\infty (c^*_\infty)^2}, \quad e = \frac{e^*}{\rho^*_\infty (c^*_\infty)^2}, \quad \lambda = \frac{\lambda^*}{\mu^*_\infty}, \quad \mu = \frac{\mu^*}{\mu^*_\infty}, \tag{5}$$

where the $*$-superscript denotes a dimensional variable and the ∞-subscript the freestream value. In (2) we have $\varepsilon = \frac{Ma}{Re}$ where Ma is the Mach-number and $Re = \frac{\rho^*_\infty u^*_\infty L^*_\infty}{\mu^*_{\text{inf ty}}}$ is the Reynolds-number with L^*_∞ being a characteristic length scale.

The equations as stated in (1) is a highly non-linear system of equations. The well-posedness and stability conditions that will be derived in this paper will be based on a linear symmetric formulation.

We freeze the coefficients at some constant state $\bar{w} = [\bar{\rho}, \bar{u}, \bar{v}, \bar{p}]^T$ and linearize as $\tilde{w} = \bar{w} + w'$ where $w' = [\rho, u, v, p]^T$ is a perturbation around the constant state $\bar{w}$. This yields an equation with constant matrix coefficients. Next we transform to primitive variables and use the parabolic symmetrizer derived in [14] to get the linear, constant coefficient and symmetric system

$$w_t + Aw_x + Bw_y = \varepsilon\Big((C_{11}w_x + C_{12}w_y)_x + (C_{21}w_x + C_{22}w_y)_y\Big), \tag{6}$$

where the symmetrized variables are

$$w = \left[\frac{\bar{c}}{\sqrt{\gamma}\bar{\rho}}\rho, u, v, \frac{1}{\bar{c}\sqrt{\gamma(\gamma-1)}}T\right]^T. \tag{7}$$

All matrix coefficients can be found in [14] but we restate them in Appendix A for convenience.

We will use the energy method to determine the well-posedness of (6). The energy norm in which we will derive our estimates is defined by

$$\|w\|^2 = \int_\Omega w^T w \, d\Omega. \tag{8}$$

By multiplying (6) with w^T, integrating over Ω and using the Gauss–Green theorem for higher-dimensional integration by parts we obtain

$$\|w\|_t^2 = -\oint_{\partial\Omega} w^T(Aw - 2\varepsilon(C_{11}w_x + C_{12}w_y),\quad Bw - 2\varepsilon(C_{21}w_x + C_{22}w_y)) \cdot n\,ds - 2\varepsilon\int_\Omega \begin{bmatrix} w_x \\ w_y \end{bmatrix}^T \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix} d\Omega, \tag{9}$$

where the last term in (9) can be seen to be dissipative by computing the eigenvalues of the matrix in the quadratic form [1,2,4].

To simplify we let the domain of interest be the unit square $0 \leqslant x, y \leqslant 1$ and we consider only the south boundary at $y = 0$. Eq. (9) is then simplified to

$$\|w\|_t^2 = \int_0^1 w^T(Bw - 2\varepsilon(C_{21}w_x + C_{22}w_y))|_{y=0}dx - 2\varepsilon\int_\Omega \begin{bmatrix} w_x \\ w_y \end{bmatrix}^T \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix} d\Omega. \tag{10}$$

The east, west and north boundaries are omitted and we consider the south boundary as a solid wall.

A solid wall requires 3 boundary conditions [2,4]. Since we do not want any penetration through the wall we require that

$$v(x,0,t) = 0. \tag{11}$$

A Robin boundary condition does not apply to the v-velocity since it is not a well-posed boundary condition for the Euler equations. When inserting (11) into (10) and considering only the south boundary at $y = 0$ we get

$$\|w\|_t^2 \leqslant -2\varepsilon\int_0^1 \left(\frac{\mu}{\bar\rho}uu_y + \frac{\gamma\mu}{\bar\rho\bar c^2\gamma(\gamma-1)Pr}TT_y\right)dx. \tag{12}$$

Note that the dissipative term has been omitted and the equality has been replaced by an inequality.

We are allowed to use 2 more boundary conditions. The boundary conditions we consider are the Robin conditions

$$\alpha u - \beta u_y = g_1,\quad \phi T - \psi T_y = g_2, \tag{13}$$

where any combination of α, β, ϕ and ψ are allowed as long as no boundary condition is removed. This allows us to study all physically relevant boundary conditions in one uniform formulation. In particular we can include the standard no-slip boundary conditions with prescribed temperature or temperature gradient and the first order slip-flow boundary conditions.

Remark 2.1. Note that if $u(x,0,t) \neq 0$ then we need to use that $v(x,0,t) = 0$ imply $v_x(x,0,t) = 0$ to obtain (12). As we shall see later, the relation $v_x(x,0,t) = 0$ must be explicitly included in the discrete case in order to obtain stability.

Depending on how we chose α, β, ϕ and ψ in (13) we obtain different energy estimates. Assume that $g_{1,2} = 0$. If we restrict ourselves to the case where β, $\psi \neq 0$ and insert (13) into (12) we obtain the energy estimate

$$\|w\|_t^2 \leqslant -2\varepsilon\int_0^1 \left(\frac{\mu}{\bar\rho}\frac{\alpha}{\beta}u^2 + \frac{\gamma\mu}{\bar\rho\bar c^2\gamma(\gamma-1)Pr}\frac{\phi}{\psi}T^2\right)dx. \tag{14}$$

We can see that the energy is bounded if

$$\alpha\beta \geqslant 0,\quad \phi\psi \geqslant 0. \tag{15}$$

We can now let α, $\phi \to 0$ and obtain the Neumann boundary conditions which have the energy estimate $\|w\|_t^2 \leqslant 0$. By restricting ourselves to the case where α, $\phi \neq 0$ we get the energy estimate

$$\|w\|_t^2 \leqslant -2\varepsilon\int_0^1 \left(\frac{\mu}{\bar\rho}\frac{\beta}{\alpha}u_y^2 + \frac{\gamma\mu}{\bar\rho\bar c^2\gamma(\gamma-1)Pr}\frac{\psi}{\phi}T_y^2\right)dx, \tag{16}$$

which gives an energy estimate if (15) hold. If we let β, $\psi \to 0$ we recover the standard no-slip boundary conditions which have the energy estimate $\|w\|_t^2 \leqslant 0$. Compared to the Robin boundary conditions (13), the no-slip boundary conditions are less damping than if we keep α, β, ϕ and ψ non-zero.

2.2. Discrete case

To extend the SBP and SAT technique to systems in higher dimensions it is convenient to introduce the Kronecker product, which is defined for arbitrary matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times q}$ by

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}. \tag{17}$$

As a special case of a tensor product, the Kronecker product is bilinear and associative, and one can prove the mixed product property $(A \otimes B)(C \otimes D) = (AC \otimes BD)$ if the usual matrix products are defined. For inversion and transposing we have

$$(A \otimes B)^{-1,T} = (A^{-1,T} \otimes B^{-1,T}), \tag{18}$$

if the usual inverse exist. The mixed product property is particularly useful since it allows the operators to operate in each coordinate direction independently of each other.

Let the domain $0 \leqslant x, y \leqslant 1$ be discretized by $M+1$ and $N+1$ equidistant grid points respectively. We define the following operators:

$$D_x = P_x^{-1} Q_x, \quad D_y = P_y^{-1} Q_y, \quad Q_{x,y} + Q_{x,y}^T = B_{x,y} = \mathrm{diag}(-1, 0, \ldots, 0, 1), \tag{19}$$

where $P_{x,y}$ is symmetric and positive definite. In this paper a diagonal $P_{x,y}$ is used but there are more general forms available [10,11]. The details for the second order case are found in Appendix B.

The extension to the two-dimensional domain is done using the Kronecker product. The following matrices will be used:

$$\begin{aligned}
&\overline{D}_x = (D_x \otimes I_y \otimes I_4) \quad \overline{D}_y = (I_x \otimes D_y \otimes I_4) \quad \overline{P}_x = (P_x \otimes I_y \otimes I_4),\\
&\overline{P}_y = (I_x \otimes P_y \otimes I_4) \quad \overline{P} = (P_x \otimes P_y \otimes I_4) \quad \overline{B}_x = (B_x \otimes I_y \otimes I_4),\\
&\overline{B}_y = (I_y \otimes B_y \otimes I_4) \quad \overline{C}_{11} = (I_x \otimes I_y \otimes C_{11}) \quad \overline{C}_{12} = (I_x \otimes I_y \otimes C_{12}),\\
&\overline{C}_{21} = (I_x \otimes I_y \otimes C_{21}) \quad \overline{C}_{22} = (I_x \otimes I_y \otimes C_{22}) \quad \overline{E}_0 = (I_x \otimes E_0 \otimes I_4),
\end{aligned} \tag{20}$$

where $E_0 = \mathrm{diag}(1,0,\ldots,0)$. The solution vector is aligned as $w = [w_0, \ldots, w_{M\times N}]^T = [\rho_0, (\rho u)_0, (\rho v)_0, e_0, \ldots, \rho_{M\times N}, (\rho u)_{M\times N}, (\rho v)_{M\times N}, e_{M\times N}]^T$.

With the definitions given in (20) we can discretize (1) as

$$w_t + \overline{D}_x \mathbf{F} + \overline{D}_y \mathbf{G} = 0, \tag{21}$$

where w, $\mathbf{F}$ and $\mathbf{G}$ are the discrete grid function and fluxes. In order to analyze (21) we need to use the linear, symmetric formulation (6). After linearizing, freezing the coefficients and transforming to symmetric variables, we apply the energy method to (21) by multiplying with $w^T\overline{P}$ and using the SBP properties of the operators. For a thorough derivation, see [1,4]. The result is

$$\|w\|_t^2 + w^T \overline{B}_x \overline{P}_y \left(\mathbf{F}_s^I - 2\varepsilon \mathbf{F}_s^V\right) + w^T \overline{B}_y \overline{P}_x \left(\mathbf{G}_s^I - 2\varepsilon \mathbf{G}_s^V\right) + 2\varepsilon \begin{bmatrix} \overline{D}_x w \\ \overline{D}_y w \end{bmatrix}^T \begin{bmatrix} \overline{P} & 0 \\ 0 & \overline{P} \end{bmatrix} \begin{bmatrix} \overline{C}_{11} & \overline{C}_{12} \\ \overline{C}_{21} & \overline{C}_{22} \end{bmatrix} \begin{bmatrix} \overline{D}_x w \\ \overline{D}_y w \end{bmatrix} = 0, \tag{22}$$

where the norm is defined by $\|w\|^2 = w^T\overline{P}w$ and

$$\begin{aligned}
&\mathbf{F}_s^I = \overline{A} w, \quad \mathbf{F}_s^V = \overline{C}_{11} w_x + \overline{C}_{12} w_y,\\
&\mathbf{G}_s^I = \overline{B} w, \quad \mathbf{G}_s^V = \overline{C}_{21} w_x + \overline{C}_{22} w_y
\end{aligned} \tag{23}$$

with $\overline{A} = (I_x \otimes I_y \otimes A)$ and $\overline{B} = (I_x \otimes I_y \otimes B)$. The last term in (22) is dissipative and we need to construct a SAT which bounds the indefinite boundary terms.

To simplify we consider only the terms related to the south boundary at $y = 0$. Eq. (22) becomes

$$\|w\|_t^2 - w^T \overline{P}_x \overline{E}_0 \left(\mathbf{G}_s^I - 2\varepsilon \mathbf{G}_s^V\right) + 2\varepsilon \begin{bmatrix} \overline{D}_x w \\ \overline{D}_y w \end{bmatrix}^T \begin{bmatrix} \overline{P} & 0 \\ 0 & \overline{P} \end{bmatrix} \begin{bmatrix} \overline{C}_{11} & \overline{C}_{12} \\ \overline{C}_{21} & \overline{C}_{22} \end{bmatrix} \begin{bmatrix} \overline{D}_x w \\ \overline{D}_y w \end{bmatrix} = 0. \tag{24}$$

Denote the last term in (24) by DI and expand the fluxes according to the definitions in (23). Eq. (24) then simplifies to

$$\|w\|_t^2 - \underbrace{w^T \overline{P}_x \overline{E}_0 \overline{B} w + 2\varepsilon w^T \overline{P}_x \overline{E}_0 (\overline{C}_{21} w_x + \overline{C}_{22} w_y)}_{\mathrm{BT}} + DI = 0. \tag{25}$$

Based on (25) we will construct a SAT which we add to the right-hand-side of (21) that will bound the indefinite terms and implement the correct boundary conditions.

Remember that the boundary conditions being imposed are

$$\alpha u - \beta u_y = g_1, \quad \phi T - \psi T_y = g_2, \quad v = g_3, \tag{26}$$

where g_3 will be set to zero at a solid wall. In order to obtain stability we also need to include the discrete version of

$$v_x = \frac{\partial g_3}{\partial x}, \tag{27}$$

which does not automatically follow from (26) as it does in the continuous case.

Due to the different forms of the boundary conditions we split the SAT into 5 different terms. One term for the inviscid part and one additional term for each condition in (26) and (27). The SAT we will use is

$$\begin{aligned}
\mathbb{S} = {} & \overline{P}_y^{-1} \overline{E}_0 \overline{\Sigma} (w - g^I) + \varepsilon \sigma_2 \overline{P}_y^{-1} \overline{E}_0 (\alpha \overline{H}_2 w - \beta \overline{D}_y \overline{H}_2 w - g_1) + \varepsilon \sigma_3 \overline{P}_y^{-1} \overline{E}_0 (\overline{H}_3 w - g_3) + \varepsilon \overline{P}_y^{-1} \overline{E}_0 \overline{\Theta} \left(\overline{D}_x w - \frac{\partial g_3}{\partial x}\right)\\
& + \varepsilon \sigma_4 \overline{P}_y^{-1} \overline{E}_0 (\phi \overline{H}_4 w - \psi \overline{D}_y \overline{H}_4 w - g_2),
\end{aligned} \tag{28}$$

where $\overline{H}_i = (I_x \otimes I_y \otimes H_i)$ and H_i are 4×4 matrices that have the only non-zero element 1 at the (i,i) position on the diagonal. We have $\overline{\Sigma} = (I_x \otimes I_y \otimes \Sigma)$ where Σ is an undetermined 4×4 matrix that will be determined for stability. $\overline{\Theta} = (I_x \otimes I_y \otimes \Theta)$ where Θ is a 4×4 penalty matrix that acts on v only and hence has the structure

$$\Theta = \begin{bmatrix} 0 & 0 & \theta_1 & 0 \\ 0 & 0 & \theta_2 & 0 \\ 0 & 0 & \theta_3 & 0 \\ 0 & 0 & \theta_4 & 0 \end{bmatrix}. \tag{29}$$

Both $\overline{\Sigma}$ and $\overline{\Theta}$ will be determined for stability.

The first row in (28) is used to bound the inviscid part and the three last rows are scaled with ε and enforces each of the boundary conditions in (26) and (27). This construction will ensure that the solution converges to that of the Euler equations as $\varepsilon \to 0$. The Robin boundary conditions does not apply to the Euler equations. Hence as $\varepsilon \to 0$, the viscous terms mush vanish and leave $v = 0$ as the only boundary condition for the Euler equations at a solid wall.

By considering zero boundary data and carrying (28) through the derivations in the energy estimate it will appear on the right-hand-side of (24) as

$$\begin{aligned} 2w^T\overline{P}\mathbb{S} = {} & 2w^T\overline{P}_x\overline{E}_0\overline{\Sigma}w + 2\varepsilon\sigma_2 w^T\overline{P}_x\overline{E}_0(\alpha\overline{H}_2 w - \beta\overline{D}_y\overline{H}_2 w) + 2\varepsilon\sigma_3 w^T\overline{P}_x\overline{E}_0\overline{H}_3 w + 2\varepsilon\overline{P}_x\overline{E}_0\overline{D}_x\overline{\Theta}w \\ & + 2\varepsilon\sigma_4 w^T\overline{P}_x\overline{E}_0(\phi\overline{H}_4 w - \psi\overline{D}_y\overline{H}_4 w). \end{aligned} \tag{30}$$

By moving all terms to the right hand side we get

$$\|w\|_t^2 = \mathrm{BT} + \mathrm{SAT} - \mathrm{DI} \tag{31}$$

and we have to choose the coefficients in (28) such that $\|w\|_t^2 \leqslant 0$. In order to proceed we split the BT and SAT into inviscid and viscous parts respectively.

By considering only the inviscid terms we have

$$\|w\|_t^2 = w^T\overline{P}_x\overline{E}_0\overline{B}w + 2w^T\overline{P}_x\overline{E}_0\overline{\Sigma}w \tag{32}$$

$$= w^T\overline{P}_x\overline{E}_0(\overline{B} + 2\overline{\Sigma})w \tag{33}$$

and we have to choose $\overline{\Sigma}$ such that $\overline{B} + 2\overline{\Sigma} \leqslant 0$. Since the Kronecker product preserves positive semi-definiteness it is sufficient to determine the 4×4 matrix Σ such that

$$B + 2\Sigma \leqslant 0. \tag{34}$$

This is easily accomplished by diagonalizing $B = X\Lambda X^T$ and rewriting (34) as

$$\Lambda^+ + \Lambda^- + 2\Sigma_c \leqslant 0, \tag{35}$$

where $\Lambda^{+,-}$ holds the positive and non-positive eigenvalues of B respectively and $\Sigma_c = X^T\Sigma X$. We have $\Lambda^+ = \mathrm{diag}(0,0,\bar{c},0)$ and hence we construct $\Sigma_c = \mathrm{diag}(0,0,\sigma,0)$ with $\sigma \leqslant -\frac{\bar{c}}{2}$. By transforming back we get

$$\Sigma = X\Sigma_c X^T = \frac{\sigma}{2\gamma}\begin{bmatrix} 1 & 0 & \sqrt{\gamma} & \sqrt{\gamma(\gamma-1)} \\ 0 & 0 & 0 & 0 \\ \sqrt{\gamma} & 0 & \gamma & \sqrt{\gamma(\gamma-1)} \\ \sqrt{\gamma(\gamma-1)} & 0 & \sqrt{\gamma(\gamma-1)} & \gamma-1 \end{bmatrix}. \tag{36}$$

With Σ given by (36) the inviscid boundary terms are bounded and implements the wall normal velocity boundary condition for the Euler equations.

Remark 2.2. The transformation from conservative to primitive to symmetric to characteristic variables and back is used only for the purpose of analysis. In a code the transformation from conservative to characteristic variables and back can be done directly by following the transformations given in [15].

By considering only the viscous terms we have

$$\begin{aligned} \|w\|_t^2 = {} & -2\varepsilon w^T\overline{P}_x\overline{E}_0\left(\overline{D}_x\overline{C}_{21}w + \overline{D}_y\overline{C}_{22}w\right) + 2\varepsilon\sigma_2 w^T\overline{P}_x\overline{E}_0(\alpha\overline{H}_2 w - \beta\overline{D}_y\overline{H}_2 w) + 2\varepsilon\sigma_3 w^T\overline{P}_x\overline{E}_0(\overline{H}_3 w) + 2\varepsilon\overline{P}_x\overline{E}_0\overline{D}_x\overline{\Theta}w \\ & + 2\varepsilon\sigma_4 w^T\overline{P}_x\overline{E}_0(\phi\overline{H}_4 w - \psi\overline{D}_y\overline{H}_4 w) - DI, \end{aligned} \tag{37}$$

which can be written as a quadratic form

$$\|w\|_t^2 = -\varepsilon W^T\overline{P}_0\overline{M}_0 W - DI, \tag{38}$$

where

$$W = \begin{bmatrix} w \\ \overline{D}_x w \\ \overline{D}_y w \end{bmatrix}, \quad \overline{P}_0 = \begin{bmatrix} \overline{P}_x \overline{E}_0 & 0 & 0 \\ 0 & \overline{P}_x \overline{E}_0 & 0 \\ 0 & 0 & \overline{P}_x \overline{E}_0 \end{bmatrix}, \quad \overline{M}_0 = \begin{bmatrix} \bar{m}_1 & \bar{m}_2 & \bar{m}_3 \\ \bar{m}_2^T & 0 & 0 \\ \bar{m}_3 & 0 & 0 \end{bmatrix}, \tag{39}$$

where $\bar{m}_3$ and $\overline{M}_0$ are symmetric and

$$\begin{aligned} \bar{m}_1 &= -2(I_x \otimes I_y \otimes \sigma_2 \alpha H_2 + \sigma_3 H_3 + \sigma_4 \phi H_4), \\ \bar{m}_2 &= (I_x \otimes I_y \otimes C_{21} - \Theta), \\ \bar{m}_3 &= (I_x \otimes I_y \otimes C_{22} + \sigma_2 \beta H_2 + \sigma_4 \psi H_4). \end{aligned} \tag{40}$$

In order to stabilize the viscous terms we need to choose our coefficients $\sigma_{2,3,4}$ and Θ such that $\overline{M}_0 \geqslant 0$. Note that only the positive semi-definiteness of $\overline{M}_0$ is required since $\overline{P}_0$ is positive semi-definite and commutes with $\overline{M}_0$. Hence if $\overline{M}_0$ is positive semi-definite, so is the product $\overline{P}_0 \overline{M}_0$.

Unfortunately though, there is no choice of $\sigma_{2,3,4}$ and Θ such that $\bar{m}_1 = \bar{m}_2 = \bar{m}_3 = 0$ which would give $\overline{M}_0 = 0$. Hence in the current form we will always end up with an indefinite $\overline{M}_0$.

To remedy this fact we can use a part of the dissipation term DI in (38),

$$DI = 2\varepsilon \begin{bmatrix} \overline{D}_x w \\ \overline{D}_y w \end{bmatrix}^T \begin{bmatrix} \overline{P} & 0 \\ 0 & \overline{P} \end{bmatrix} \begin{bmatrix} \overline{C}_{11} & \overline{C}_{12} \\ \overline{C}_{21} & \overline{C}_{22} \end{bmatrix} \begin{bmatrix} \overline{D}_x w \\ \overline{D}_y w \end{bmatrix}. \tag{41}$$

The matrix $\overline{P}$ can be rewritten as

$$\overline{P} = (P_x \otimes P_y \otimes I_4) = \left(P_x \otimes \widetilde{P}_y + rE_0 \otimes I_4\right) = \underbrace{\left(P_x \otimes \widetilde{P}_y \otimes I_4\right)}_{\widetilde{P}} + r\overline{P}_x \overline{E}_0, \tag{42}$$

where r is small enough such that $P_y(1,1) - r \geqslant 0$ [16,17]. If we choose r such that strict inequality holds, the remainder $\widetilde{P}$ is still a full norm. Note that r is proportional to Δy. The dissipation term can thus be rewritten as

$$DI = 2\varepsilon \begin{bmatrix} \overline{D}_x w \\ \overline{D}_y w \end{bmatrix}^T \begin{bmatrix} \widetilde{P} & 0 \\ 0 & \widetilde{P} \end{bmatrix} \begin{bmatrix} \overline{C}_{11} & \overline{C}_{12} \\ \overline{C}_{21} & \overline{C}_{22} \end{bmatrix} \begin{bmatrix} \overline{D}_x w \\ \overline{D}_y w \end{bmatrix} + 2\varepsilon r \begin{bmatrix} \overline{D}_x w \\ \overline{D}_y w \end{bmatrix}^T \begin{bmatrix} \overline{P}_x \overline{E}_0 & 0 \\ 0 & \overline{P}_x \overline{E}_0 \end{bmatrix} \begin{bmatrix} \overline{C}_{11} & \overline{C}_{12} \\ \overline{C}_{21} & \overline{C}_{22} \end{bmatrix} \begin{bmatrix} \overline{D}_x w \\ \overline{D}_y w \end{bmatrix}. \tag{43}$$

The second term in (43) can be used to fill in the empty 2×2 bottom block in $\overline{M}_0$ to obtain

$$\overline{M} = \begin{bmatrix} \bar{m}_1 & \bar{m}_2 & \bar{m}_3 \\ \bar{m}_2^T & 2r\overline{C}_{11} & 2r\overline{C}_{12} \\ \bar{m}_3 & 2r\overline{C}_{21} & 2r\overline{C}_{22} \end{bmatrix}. \tag{44}$$

To determine positive semi-definiteness of $\overline{M}$ it is sufficient to only consider the reduced matrix

$$M = \begin{bmatrix} -2(\sigma_2 \alpha H_2 + \sigma_3 H_3 + \sigma_4 \phi H_4) & C_{21} - \Theta & C_{22} + \sigma_2 \beta H_2 + \sigma_4 \psi H_4 \\ C_{21} - \Theta^T & 2rC_{11} & 2rC_{12} \\ C_{22} + \sigma_2 \beta H_2 + \sigma_4 \psi H_4 & 2rC_{21} & 2rC_{22} \end{bmatrix}, \tag{45}$$

where we have removed the Kronecker products. This can be done since the Kronecker product is permutation similar, i.e. there exist a permutation matrix Y such that for arbitrary square matrices A and B we have $A \otimes B = Y^T(B \otimes A)Y$. Hence we can rewrite (38) as

$$\|w\|_t^2 = -\varepsilon(YW)^T(M \otimes P_0)YW - \widetilde{DI}, \tag{46}$$

where $P_0 = P_x \otimes E_0$ is positive semi-definite.

In order to proceed we chose

$$\Theta = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{\lambda+\mu}{2\rho} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{47}$$

The matrix M in (45) is of size 12×12 but with the 1st, 5th and 9th row and column being zero. We can hence remove these rows and columns and condense (45) into the 9×9 matrix

$$\widetilde{M} = \begin{bmatrix} -2\left(\sigma_2 \alpha \widetilde{H}_2 + \sigma_3 \widetilde{H}_3 + \sigma_4 \phi \widetilde{H}_4\right) & \widetilde{C}_{21} - \tilde{\Theta} & \widetilde{C}_{22} + \sigma 2 \beta \widetilde{H}_2 + \sigma_4 \psi \widetilde{H}_4 \\ \widetilde{C}_{21} - \tilde{\Theta}^T & 2r\widetilde{C}_{11} & 2r\widetilde{C}_{12} \\ \widetilde{C}_{22} + \sigma_2 \alpha \widetilde{H}_2 + \sigma_4 \psi \widetilde{H}_4 & 2r\widetilde{C}_{21} & 2r\widetilde{C}_{22} \end{bmatrix}. \tag{48}$$

By defining the matrices

$$\begin{aligned}
\widetilde{A} &= \sigma_2\alpha\widetilde{H}_2 + \sigma_3\widetilde{H}_3 + \sigma_4\phi\widetilde{H}_4,\\
\widetilde{B} &= \widetilde{C}_{22} + \sigma_2\beta\widetilde{H}_2 + \sigma_4\psi\widetilde{H}_4,\\
\widetilde{C} &= \begin{bmatrix} \widetilde{C}_{11} & \widetilde{C}_{12} \\ \widetilde{C}_{21} & \widetilde{C}_{22} \end{bmatrix},\\
\widetilde{J} &= \begin{bmatrix} \widetilde{C}_{21} - \tilde{\Theta} & \widetilde{B} \end{bmatrix},
\end{aligned} \tag{49}$$

we can rewrite (48) as

$$\widetilde{M} = \begin{bmatrix} -2\widetilde{A} & \widetilde{J} \\ \widetilde{J}^T & 2r\widetilde{C} \end{bmatrix}, \tag{50}$$

which can be rotated into block-diagonal form. The rotation matrix is defined by

$$\widetilde{S} = \begin{bmatrix} I_3 & -\frac{1}{2r}\widetilde{J}\widetilde{C}^{-1} \\ 0_{6\times 3} & I_6 \end{bmatrix}, \tag{51}$$

where $0_{p\times q}$ is a zero matrix of size indicated by the subscript. Note that $\widetilde{C}^{-1}$ is well-defined since we have removed the zero rows and columns. Using (51) we can rotate (50) by

$$\widetilde{S}\widetilde{M}\widetilde{S}^T = \begin{bmatrix} -2\widetilde{A} - \frac{1}{2r}\widetilde{J}\widetilde{C}^{-1}\widetilde{J}^T & 0_{3\times 6} \\ 0_{6\times 3} & 2r\widetilde{C} \end{bmatrix} \tag{52}$$

and it is clear that a sufficient condition for positive semi-definiteness is that the Schur complement of $2r\widetilde{C}$ in $\widetilde{M}$ satisfies

$$Q = -2\widetilde{A} - \frac{1}{2r}\widetilde{J}\widetilde{C}^{-1}\widetilde{J}^T \geqslant 0. \tag{53}$$

Eq. (53) leads to the main result of this paper which is

Theorem 2.3. *The scheme for the compressible Navier–Stokes equations*

$$w_t + \overline{D}_x\mathbf{F} + \overline{D}_y\mathbf{G} = \mathbb{S} \tag{54}$$

with Robin boundary conditions given in (26) and (27), *where* $\mathbb{S}$ *is given by* (28), *can be made stable for all choices of* α, β, ϕ *and* ψ *using* (36) and (47) *and appropriate choices of* $\sigma_{2,3,4}$.

Proof. The inviscid part that implements the wall normal velocity boundary condition for the Euler equations is bounded using (36). Using (47), the matrix Q in (53) is a 3×3 diagonal matrix

$$Q = \begin{bmatrix} \lambda_1(\sigma_2) & 0 & 0 \\ 0 & \lambda_2(\sigma_3) & 0 \\ 0 & 0 & \lambda_3(\sigma_4) \end{bmatrix}, \tag{55}$$

where the diagonal entries are given by

$$\begin{aligned}
\lambda_1(\sigma_2) &= -2\sigma_2\alpha - \frac{2\mu(\mu+\sigma_2\beta\bar{\rho})^2}{r(\lambda+3\mu)(\mu-\lambda)\bar{\rho}},\\
\lambda_2(\sigma_3) &= -2\sigma_3 - \frac{1}{2}\frac{\lambda+2\mu}{r\bar{\rho}},\\
\lambda_3(\sigma_4) &= -2\sigma_4\phi - \frac{1}{2}\frac{(\gamma\mu+\sigma_4\psi Pr\bar{\rho})^2}{r\gamma\mu Pr\bar{\rho}}.
\end{aligned} \tag{56}$$

For any choice of α, β, ϕ and ψ such that no boundary condition is removed and (15) holds, it is possible to determine $\sigma_{2,3,4}$ such that $\lambda_{1,2,3} \geqslant 0$. The actual values of $\sigma_{2,3,4}$ are determined once the choices of α, β, ϕ and ψ has been made. □

The standard no-slip boundary conditions with prescribed temperature

$$u = 0, \quad v = 0, \quad T = T_w, \tag{57}$$

where T_w is the wall temperature follows as a corollary.

Corollary 2.4. *The standard no-slip boundary conditions with prescribed temperature given by*

$$u = 0, \quad v = 0, \quad T = T_w \tag{58}$$

are stable using (36) and (47) *and*

$$\begin{aligned} \sigma_2 &\leqslant -\frac{\mu^3}{r(\lambda+3\mu)(\mu-\lambda)\bar{\rho}}, \\ \sigma_3 &\leqslant -\frac{1}{4}\frac{\lambda+2\mu}{r\bar{\rho}}, \\ \sigma_4 &\leqslant -\frac{1}{4r}\frac{\gamma\mu}{Pr\bar{\rho}}. \end{aligned} \tag{59}$$

Proof. The no-slip boundary conditions with prescribed temperature, which are thoroughly discussed in [2], are obtained by putting

$$\alpha = 1, \quad \beta = 0, \quad \phi = 1, \quad \psi = 0, \tag{60}$$

in which case (56) reduces to

$$\begin{aligned} \lambda_1(\sigma_2) &= -2\sigma_2 - \frac{2\mu^3}{r(\lambda+3\mu)(\lambda-\mu)\bar{\rho}}, \\ \lambda_2(\sigma_3) &= -2\sigma_3 - \frac{1}{2}\frac{\lambda+2\mu}{r\bar{\rho}}, \\ \lambda_3(\sigma_4) &= -2\sigma_4 - \frac{1}{2}\frac{\gamma\mu}{rPr\bar{\rho}}. \end{aligned} \tag{61}$$

By demanding

$$\lambda_i \geqslant 0, \quad i = 1,2,3, \tag{62}$$

we obtain (59). □

Note that the estimates (59) are sharp since there are no approximations or embeddings involved in the derivation of (53) as in contrast to the result in [2]. The results in [2] are obtained in this setting by having

$$\Theta = 0_{4\times 4} \tag{63}$$

and taking

$$\sigma_{1,2,3} = \sigma \leqslant -\frac{1}{4r}\lambda_{\max}, \tag{64}$$

where $\lambda_{\max}$ is the maximum eigenvalue of $\widetilde{J}\widetilde{C}^{-1}\widetilde{J}^T$. Since the system becomes stiffer with increasing magnitude of the coefficients it is desirable with sharp estimates to minimize the magnitudes. If we compare (59) and (64) we get

$$\begin{aligned} \frac{\sigma_2}{\sigma} &= \frac{4\mu^2 Pr}{\gamma(\lambda+3\mu)(\mu-\lambda)}, \\ \frac{\sigma_3}{\sigma} &= \frac{(\lambda+2\mu)Pr}{\gamma\mu}, \\ \frac{\sigma_4}{\sigma} &= 1. \end{aligned} \tag{65}$$

With some reasonable numerical values, $\rho = 1$, $\gamma = 1.4$, $Pr = 0.72$, $\mu = 1$ and $\lambda = -\frac{2}{3}\mu$, the ratios become

$$\frac{\sigma_2}{\sigma} \approx 0.53, \quad \frac{\sigma_3}{\sigma} \approx 0.69, \quad \frac{\sigma_4}{\sigma} = 1, \tag{66}$$

which is an improvement for the velocity components.

The proof of stability using (63) and (64) does not extend to the case where $\beta \neq 0$ in which case $\Theta \neq 0_{4\times 4}$ is required.

For the adiabatic solid wall boundary conditions we have

Corollary 2.5. *The adiabatic boundary conditions*

$$u = 0, \quad v = 0, \quad T_y = 0, \tag{67}$$

are stable using (36) and (47) *and*

$$\begin{aligned}\sigma_2 &\leqslant -\frac{\mu^3}{r(\lambda+3\mu)(\mu-\lambda)\bar{\rho}},\\ \sigma_3 &\leqslant -\frac{1}{4}\frac{\lambda+2\mu}{r\bar{\rho}},\\ \sigma_4 &= -\frac{\gamma\mu}{Pr\bar{\rho}}.\end{aligned} \tag{68}$$

Proof. The adiabatic boundary conditions are obtained by having

$$\alpha = 1, \quad \beta = 0, \quad \phi = 0, \quad \psi = 1, \tag{69}$$

in which case (56) reduces to

$$\begin{aligned}\lambda_1(\sigma_2) &= -2\sigma_2 - \frac{2\mu^3}{r(\lambda+3\mu)(\lambda-\mu)\bar{\rho}},\\ \lambda_2(\sigma_3) &= -2\sigma_3 - \frac{1}{2}\frac{\lambda+2\mu}{r\bar{\rho}},\\ \lambda_3(\sigma_4) &= -\frac{1}{2}\frac{(\gamma\mu+\sigma_4 Pr\bar{\rho})^2}{r\gamma\mu Pr\bar{\rho}}.\end{aligned} \tag{70}$$

By demanding

$$\lambda_i \geqslant 0, \quad i = 1,2,3, \tag{71}$$

we obtain (68). □

Remember that r is proportional to Δy. As the mesh is refined, the penalty coefficients will increase in magnitude and make the discretization stiffer. If $\beta, \psi \neq 0$ we can cancel the $1/r$ dependence in $\sigma_{2,4}$ by choosing

$$\sigma_2 = -\frac{1}{\beta}\frac{\mu}{\bar{\rho}}, \quad \sigma_4 = -\frac{1}{\psi}\frac{\gamma\mu}{Pr\bar{\rho}}, \tag{72}$$

in which case (56) reduces to

$$\begin{aligned}\lambda_1 &= \frac{2\mu}{\bar{\rho}}\frac{\alpha}{\beta},\\ \lambda_2(\sigma_3) &= -2\sigma_3 - \frac{1}{2}\frac{\lambda+2\mu}{r\bar{\rho}},\\ \lambda_3 &= \frac{2\gamma\mu}{Pr\bar{\rho}}\frac{\phi}{\psi}.\end{aligned} \tag{73}$$

It is easy to see from (73) that the continuous well-posedness conditions (15) are required in order for $\lambda_{1,3} \geqslant 0$. The $1/r$ dependence in σ_3 is not possible to remove unless a different form of the SAT is used.

Remark 2.6. For the north boundary at $y = 1$, the conditions in Theorem 2.3 and its corollaries apply without modifications. However, the Robin boundary conditions (26) are replaced by

$$\alpha u + \beta u_y = g_1, \quad \phi T + \psi T_y = g_2, \quad v = g_3. \tag{74}$$

3. Numerical results

The stability theory developed in the previous section does not depend on the order of accuracy of the numerical scheme. In order to verify that the scheme attains its design order we will use the method of manufactured solutions.

Any sufficiently smooth function $H(x,y,t)$ is a solution to the modified Navier–Stokes equations

$$q_t + F_x + G_y = R(x,y,t), \tag{75}$$

where the forcing function $R(x,y,t)$ has to be appropriately chosen depending on $H(x,y,t)$. By the principle of Duhamel [18], the inhomogeneous Eq. (75) is well-posed if the homogeneous Eq. (1) is [18]. The boundary conditions remain unchanged and we can use the manufactured solution $H(x,y,t)$ to create the initial and boundary data. Thus we have an analytic solution to (75) which can be used to test the order of accuracy of the computational scheme.

Table 1
Order of accuracy.

# Grid points	32×32	64×64	128×128	256×256
2nd-order				
ρ	1.5034	1.7651	1.9666	1.9752
ρu	1.8596	2.0101	2.0594	2.0402
ρv	1.8540	2.0216	2.0643	2.0163
e	1.4702	1.8064	2.0253	1.9725
4th-order				
ρ	2.1722	2.3449	2.7873	2.7933
ρu	2.5558	2.6331	2.7796	2.6916
ρv	2.5409	2.5925	2.8033	2.7474
e	2.1944	2.4076	2.7627	2.7141
6th-order				
ρ	3.4865	3.6814	3.8377	3.8038
ρu	3.7509	3.7349	3.9500	3.9811
ρv	3.6202	3.8639	4.0055	4.0174
e	3.4023	4.0812	4.1063	3.9139

In this particular case we specify

$$\begin{aligned} \rho(x,y,t) &= e^{-\nu(\sin(\xi\pi x - t))^2 - \nu(\cos(\xi\pi y - t))^2}, \\ u(x,y,t) &= \cos(\xi\pi(x+y) - t), \\ v(x,y,t) &= \sin(\xi\pi(x+y) - t), \\ p(x,y,t) &= e^{-\nu(\sin(\xi\pi(x-y) - t))^2}, \end{aligned} \tag{76}$$

where ν and ξ can be used to tune the amplitude and frequency of the solution. In this case we have chosen $\nu = \xi = 0.1$. Using (76) we specify $H(x,y,t)$ as

$$H(x,y,t) = \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ e \end{bmatrix}, \quad e = \frac{p}{\gamma - 1} + \frac{1}{2}\rho(u^2 + v^2), \tag{77}$$

where $\gamma = 1.4$.

The scheme for (75) is

$$w_t + \overline{D}_x \mathbf{F} + \overline{D}_y \mathbf{G} = R(x,y,t) + \mathbb{S} \tag{78}$$

and in order to obtain a higher order accurate scheme, the difference operators $\overline{D}_{x,y}$ are simply replaced with operators of the desired order of accuracy. The penalty coefficients in Theorem 2.3 remain unchanged. The forcing function $R(x,y,t)$ is too tedious to write in text but can be computed using a symbolic software such as Maple®.

The scheme (78) was implemented using SBP operators of order 2, 4 and 6 which gives a global accuracy of 2, 3 and 4 [10,19]. The result can be seen in Table 1. The order of accuracy is independent of the choices of α, β, ϕ and ψ and in Table 1 the no-slip with prescribed temperature, using $\alpha = 1$, $\beta = 0$, $\phi = 1$ and $\psi = 0$, is seen.

4. Applications

An application of the Robin boundary condition is the slip-flow boundary conditions used for moderate Knudsen numbers (Kn) in micro fluid flows. The slip-flow boundary conditions extends the use of the Navier–Stokes equations to the slip-flow regime where $10^{-3} \leqslant Kn \leqslant 10^{-1}$ [20].

Computations in the slip-flow regime corresponds to having $\alpha = 1$, $\phi = 1$, $\psi = 0$ and $\beta = Kn$ which gives a first order slip-flow boundary condition. Stability is shown in

Corollary 4.1. *The first order slip-flow boundary conditions*

$$u = (Kn)u_y, \quad v = 0, \quad T = T_w \tag{79}$$

are stable using (36), (47) *and*

$$\begin{aligned} \sigma_2 &= -\frac{\mu}{(Kn)\bar{\rho}}, \\ \sigma_3 &\leqslant -\frac{1}{4}\frac{\lambda + 2\mu}{r\bar{\rho}}, \\ \sigma_4 &\leqslant -\frac{1}{4r}\frac{\gamma\mu}{Pr\bar{\rho}}. \end{aligned} \tag{80}$$

Proof. The slip-flow boundary conditions are obtained by

$$\alpha = 1, \quad \beta = Kn, \quad \phi = 1, \quad \psi = 0, \tag{81}$$

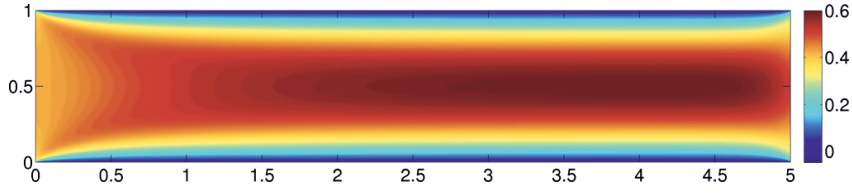
in which case (56) reduces to

$$\begin{aligned} \lambda_1(\sigma_2) &= -2\sigma_2 - \frac{2\mu(\mu+\sigma_2(Kn)\bar{\rho})^2}{r(\lambda+3\mu)(\lambda-\mu)\bar{\rho}}, \\ \lambda_2(\sigma_3) &= -2\sigma_3 - \frac{1}{2}\frac{\lambda+2\mu}{r\bar{\rho}}, \\ \lambda_3(\sigma_4) &= -2\sigma_4 - \frac{1}{2}\frac{\gamma\mu}{rPr\bar{\rho}}. \end{aligned} \tag{82}$$

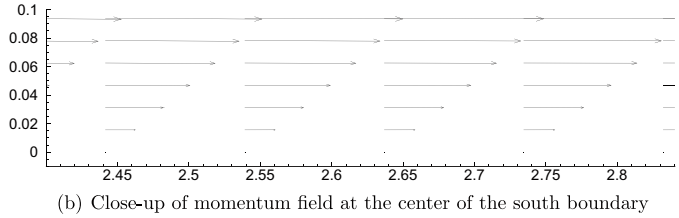
By demanding

$$\lambda_i \geqslant 0, \quad i = 1,2,3, \tag{83}$$

we obtain (80). □

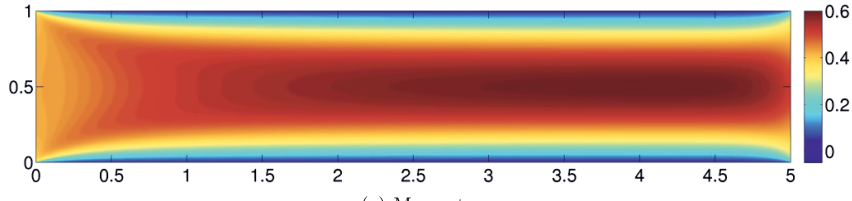


(a) Momentum

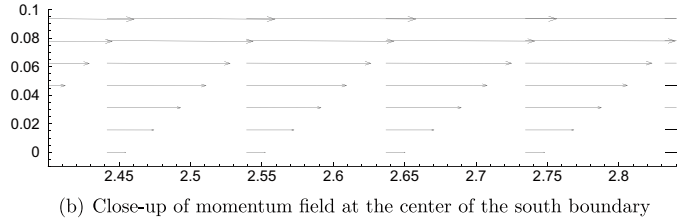


(b) Close-up of momentum field at the center of the south boundary

Fig. 1. $\beta = 0.0$, corresponding to no-slip.

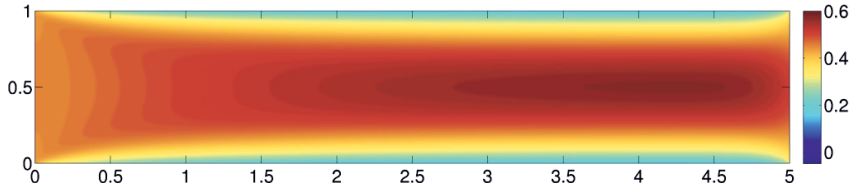


(a) Momentum

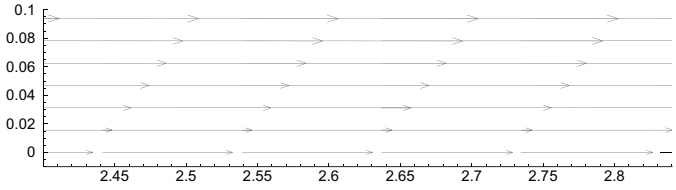


(b) Close-up of momentum field at the center of the south boundary

Fig. 2. $\beta = 0.01$, corresponding to moderate slip.

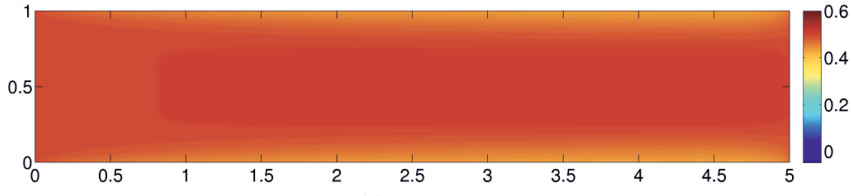


(a) Momentum

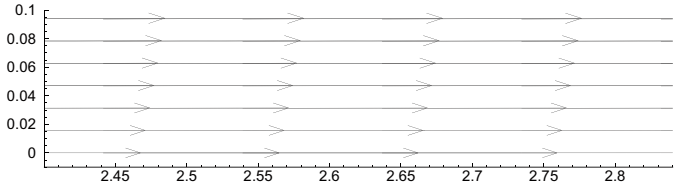


(b) Close-up of momentum field at the center of the south boundary

Fig. 3. $\beta = 0.1$, corresponding to large slip.



(a) Momentum



(b) Close-up of momentum field at the center of the south boundary

Fig. 4. $\beta = 1.0$, corresponding to almost full slip.

Figs. 1–4 shows the flow field from no-slip ($\beta = 0$) to almost full slip ($\beta = 1$). In the computations we have used the domain $0 \leqslant x \leqslant 5$, $0 \leqslant y \leqslant 1$ with 512×128 grid points. The Mach number is 0.5 and the Reynolds number is 100. All scales are normalized with respect to the no-slip case.

The inflow and outflow boundary conditions are implemented as described in [1] which means that there is a severe mismatch between the boundary conditions and the boundary data at the corners. However because of the weak boundary treatment the computations remain stable.

5. Summary and conclusions

We have proved stability for Robin solid wall boundary conditions for the compressible Navier–Stokes equations using a finite difference method on Summation-By-Parts (SBP) form with weak boundary conditions using the Simultaneous Approximation Term (SAT).

The formulation of the SAT allows for easy change between common boundary conditions such as the no-slip with prescribed temperature or temperature gradient and slip-flow or any combination thereof.

The energy estimates were derived without using approximations or embeddings which yields sharp estimates in contrast to previous results.

The accuracy of the numerical scheme was tested using a manufactured solution. The computational scheme was verified to attain 2nd-, 3rd- and 4th-order of accuracy which are the design orders of the SBP scheme.

We did computations of flows in a rectangular domain when the solid wall boundary conditions were changed from no-slip to substantial slip by a simple variation of one parameter.

Acknowledgments

The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project p2010056. The reviewer of the manuscript is greatly acknowledged for improving the quality of the final article.

Appendix A. Matrix coefficients

The matrix coefficients in (6) are given by

$$A=\begin{bmatrix} \bar{u} & \frac{\bar{c}}{\sqrt{\gamma}} & 0 & 0 \\ \frac{\bar{c}}{\sqrt{\gamma}} & \bar{u} & 0 & \bar{c}\sqrt{\frac{\gamma-1}{\gamma}} \\ 0 & 0 & \bar{u} & 0 \\ 0 & \bar{c}\sqrt{\frac{\gamma-1}{\gamma}} & 0 & \bar{u} \end{bmatrix}, \quad B=\begin{bmatrix} \bar{v} & 0 & \frac{\bar{c}}{\sqrt{\gamma}} & 0 \\ 0 & \bar{v} & 0 & 0 \\ \frac{\bar{c}}{\sqrt{\gamma}} & 0 & \bar{v} & \bar{c}\sqrt{\frac{\gamma-1}{\gamma}} \\ 0 & 0 & \bar{c}\sqrt{\frac{\gamma-1}{\gamma}} & \bar{v} \end{bmatrix},$$

$$C_{11}=\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{\lambda+2\mu}{\bar{\rho}} & 0 & 0 \\ 0 & 0 & \frac{\mu}{\bar{\rho}} & 0 \\ 0 & 0 & 0 & \frac{\gamma\mu}{Pr\bar{\rho}} \end{bmatrix}, \quad C_{22}=\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{\mu}{\bar{\rho}} & 0 & 0 \\ 0 & 0 & \frac{\lambda+2\mu}{\bar{\rho}} & 0 \\ 0 & 0 & 0 & \frac{\gamma\mu}{Pr\bar{\rho}} \end{bmatrix}, \tag{A.1}$$

$$C_{12}=C_{21}=\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{\lambda+\mu}{2\bar{\rho}} & 0 \\ 0 & \frac{\lambda+\mu}{2\bar{\rho}} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Appendix B. SBP operators

In the second order case the SBP operators are explicitly given by

$$D_\xi = P_\xi^{-1} Q_\xi, \tag{B.1}$$

where ξ is either x or y and

$$P_\xi=\frac{1}{\Delta\xi}\begin{bmatrix} \frac{1}{2} & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & 1 & 0 \\ 0 & \dots & 0 & 0 & \frac{1}{2} \end{bmatrix}, \quad Q_\xi=\frac{1}{2}\begin{bmatrix} -1 & 1 & 0 & 0 & \dots & 0 & 0 \\ -1 & 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & -1 & 0 & 1 & 0 \\ 0 & \dots & 0 & 0 & -1 & 0 & 1 \\ 0 & \dots & 0 & 0 & 0 & -1 & 1 \end{bmatrix},$$

$$D_\xi=\frac{1}{2\Delta\xi}\begin{bmatrix} -2 & 2 & 0 & 0 & \dots & 0 & 0 \\ -1 & 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & -1 & 0 & 1 & 0 \\ 0 & \dots & 0 & 0 & -1 & 0 & 1 \\ 0 & \dots & 0 & 0 & 0 & -2 & 2 \end{bmatrix}. \tag{B.2}$$

For SBP operators of higher order accuracy we refer the reader to [10,11].

References

[1] Magnus Svärd, Mark H. Carpenter, Jan Nordström, A stable high-order finite difference scheme for the compressible Navier–Stokes equations, far-field boundary conditions, Journal of Computational Physics 225 (1) (2007) 1020–1038.
[2] Magnus Svärd, Jan Nordström, A stable high-order finite difference scheme for the compressible Navier–Stokes equations: no-slip wall boundary conditions, Journal of Computational Physics 227 (10) (2008) 4805–4824.
[3] Jan Nordström, Jing Gong, Edwin van der Weide, Magnus Svärd, A stable and conservative high order multi-block method for the compressible Navier–Stokes equations, Journal of Computational Physics 228 (24) (2009) 9020–9035.
[4] Jan Nordström, Magnus Svärd, Well-posed boundary conditions for the Navier–Stokes equations, SIAM Journal on Numerical Analysis 43 (3) (2005) 1231–1255.
[5] Ken Mattsson, Magnus Svärd, Mohammad Shoeybi, Stable and accurate schemes for the compressible Navier–Stokes equations, Journal of Computational Physics 227 (2008) 2293–2316.
[6] Magnus Svärd, Ken Mattsson, Jan Nordström, Steady-state computations using summation-by-parts operators, Journal of Scientific Computing 24 (1) (2005) 79–95.
[7] Ken Mattsson, Magnus Svärd, Mark Carpenter, Jan Nordström, High-order accurate computations for unsteady aerodynamics, Computers and Fluids 36 (3) (2007) 636–649.
[8] X. Huan, J.E. Hicken, D.W. Zingg, Interface and boundary schemes for high-order methods, in: The 39th AIAA Fluid Dynamics Conference, AIAA Paper No. 2009-3658, San Antonio, USA, 22–25 June 2009.
[9] Jan Nordström, Sofia Eriksson, Craig Law, Jing Gong, Shock and vortex calculations using a very high order accurate Euler and Navier–Stokes solver, Journal of Mechanics and MEMS 1 (1) (2009) 19–26.
[10] Bo Strand, Summation by parts for finite difference approximations for d/dx, Journal of Computational Physics 110 (1) (1994) 47–67.
[11] Ken Mattsson, Jan Nordström, Summation by parts operators for finite difference approximations of second derivatives, Journal of Computational Physics 199 (2) (2004) 503–540.
[12] Ken Mattsson, Summation by parts operators for finite difference approximations of second-derivatives with variable coefficients, Technical Report 2010-023, Uppsala University, Division of Scientific Computing, 2010.
[13] Qaisar Abbas, Jan Nordström, Weak versus strong no-slip boundary conditions for the Navier–Stokes equations, Engineering Applications of Computational Fluid Mechanics 4 (2010) 29–38.
[14] Saul Abarbanel, David Gottlieb, Optimal time splitting for two- and three-dimensional Navier–Stokes equations with mixed derivatives, Journal of Computational Physics 41 (1) (1981) 1–33.
[15] T.H. Pulliam, D.S. Chaussee, A diagonal form of an implicit approximate-factorization algorithm, Journal of Computational Physics 39 (2) (1981) 347–363.
[16] Mark H. Carpenter, Jan Nordström, David Gottlieb, A stable and conservative interface treatment of arbitrary spatial accuracy, Journal of Computational Physics 148 (2) (1999) 341–365.
[17] Mark H. Carpenter, Jan Nordström, David Gottlieb, Revisiting and extending interface penalties for multi-domain summation-by-parts operators, Journal of Scientific Computing 45 (1–3) (2010) 118–150.
[18] Bertil Gustafsson, Heinz-Otto Kreiss, Joseph Oliger, Time Dependent Problems and Difference Methods, Wiley Interscience, 1995.
[19] Magnus Svärd, Jan Nordström, On the order of accuracy for difference approximations of initial-boundary value problems, Journal of Computational Physics 218 (1) (2006) 333–352.
[20] Mohamed Gad el Hak, The fluid mechanics of microdevices—the freeman scholar lecture, Journal of Fluids Engineering 121 (1) (1999) 5–33.

Paper IV

Conjugate Heat Transfer for the Unsteady Compressible Navier-Stokes Equations Using a Multi-block Coupling

Jan Nordström[†*]

Department of Mathematics, Linköping University, SE-581 83 Linköping, Sweden

Jens Berg[†]

Department of Information Technology, Uppsala University, SE-751 05, Uppsala, Sweden

Abstract

This paper deals with conjugate heat transfer problems for the time-dependent compressible Navier-Stokes equations. One way to model conjugate heat transfer is to couple the Navier-Stokes equations in the fluid with the heat equation in the solid. This requires two different physics solvers. Another way is to let the Navier-Stokes equations govern the heat transfer in both the solid and in the fluid. This simplifies calculations since the same physics solver can be used everywhere.

We show by energy estimates that the continuous problem is well-posed when imposing continuity of temperature and heat fluxes by using a modified L^2-equivalent norm. The equations are discretized using finite difference on summation-by-parts form with boundary- and interface conditions imposed weakly by the simultaneous approximation term. It is proven that the scheme is energy stable in the modified norm for any order of accuracy.

We also show what is required for obtaining the same solution as when the unsteady compressible Navier-Stokes equations are coupled to the heat equation. The differences between the two coupling techniques are discussed theoretically as well as studied numerically, and it is shown that they are indeed small.

Keywords: Conjugate heat transfer, Navier-Stokes, compressible, unsteady, heat equation, finite difference, summation-by-parts, weak interface conditions, weak multi-block conditions, stability, high order accuracy

*Corresponding author: Jan Nordström, E-mail: jan.nordstrom@liu.se

[†]Parts of this work were completed while the authors visited the Centre for Turbulence Research at Stanford University

1. Introduction

Heat transfer is an important factor in many fluid dynamics applications. Flows are often confined within some material with heat transfer properties. Whenever there is a temperature difference between the fluid and the confining solid, heat will be transferred and change the flow properties in a non-trivial way. This interaction and heat exchange is referred to as the conjugate heat transfer problem [1, 2, 3, 4]. Examples of application areas include cooling of turbine blades and nuclear reactors, atmospheric reentry of spacecrafts and gas propulsion micro thrusters for precise satellite navigation.

Conjugate heat transfer problems have been computed using a variety of methods. For stationary problems, methods include the finite volume method [5], the finite element method [6, 7] and the Semi-Implicit Method for Pressure-Linked Equations (SIMPLE) [8]. For unsteady problems, overlapping grids [3] and finite difference methods [1] have been used. The interface conditions have been imposed either strongly, weakly or by a mixture of both.

There are many ways in which conjugate heat transfer problems can be analyzed and computed. Giles [1] considered the simplified case of two coupled heat equations and performed a stability analysis which put restrictions on how to chose the interface conditions. Henshaw and Chand [3] performed numerical simulations of incompressible, temperature dependent fluids with the Boussinesq approximation coupled with the heat equation. The stability analysis was restricted to the case of two coupled heat equations. Stability and second order accuracy for the coupled model problem was proven, together with a numerical accuracy study of the full coupled problem showing second order accuracy, as expected. In [7] a steady, compressible fluid with heat transfer properties is considered and it is stated that accuracy is a key element in computational heat transfer. The authors develop an adaptive strategy with error estimators, showing at most second order accuracy.

When reviewing the literature on conjugate heat transfer problems, one can conclude that for incompressible problems, the heat transfer part is either modeled by the heat equation, or by using the incompressible Navier-Stokes equations also in the solid region. The latter strategy is possible since the energy equation in the incompressible Navier-Stokes equations decouples from the continuity- and momentum equations. In the compressible flow case, the situation is different and more complicated. Two major differences exist. Firstly, the energy equation *does not decouple* from the continuity- and momentum equations. Secondly, for compressible fluids, steady problems are mostly considered since the stability of the coupling becomes an issue.

The numerical methodology presented in this paper is based on a finite difference on Summation-By-Parts (SBP) form with the Simultaneous Approximation Term (SAT) for imposing the boundary and interface conditions weakly. The SBP-SAT method has been used for a variety of problems and has proven to be robust [9, 10, 11, 12, 13, 14, 15]. The SBP finite difference operators were originally constructed by Kreiss and Schearer [16] with the purpose of constructing an energy stable finite difference method [17]. Together with the weak imposition of boundary [18] and interface [19] conditions, the SBP-SAT provides a method for constructing energy stable schemes for well-posed initial-boundary value problems [20]. There are SBP operators based on diagonal norms for the first [21] and second [22, 23] derivative accurate of order 2, 3, 4 and 5, and the stability analysis we will present is independent of the order of accuracy.

From an implementational point of view, coupling the compressible Navier-Stokes equations to the heat equation is complicated as different solvers are required in the fluid and solid domains. With two different solvers, two different codes, are required and data has to be transferred between them by using possibly a third code [24].

A less complicated method would be to only use the Navier-Stokes equations everywhere and modify an already existing multi-block coupling [12] such that heat is transferred between the fluid and solid domains. In the blocks marked as solids, it is possible to construct initial and boundary conditions such that the velocities and density gradients are small. The difference between the energy component of the compressible Navier-Stokes equations and the heat equation should then also be small.

We will show how to scale and choose the coefficients of the energy part of the Navier-Stokes equations, such that it is as similar to the heat equation as possible. Numerical simulations of heat transfer in solids are performed to show the similarities, and differences, of the temperature distributions obtained by the Navier-Stokes equations and the heat equation. We will *not* overwrite, or strongly force, the velocities in the Navier-Stokes equations to zero in each time integration stage since that would ruin the stability of the numerical method that we use. Instead, the velocities will be enforced weakly at the boundaries and interfaces only.

In the previous literature, a mathematical investigation of the interface conditions in terms of well-posedness of the continuous equations, stability of the resulting numerical scheme and high order accuracy has not been performed to our knowledge. We shall in this paper hence focus on the mathematical treatment of the fluid-solid interface rather than computing physically relevant scenarios.

2. The compressible Navier-Stokes equations

The two-dimensional compressible Navier-Stokes equations in dimensional, conservative form are

$$q_t + F_x + G_y = 0 \tag{1}$$

where the conserved variables, $q = [\rho, \rho u, \rho v, e]^T$, are the density, x- and y-directional momentum and energy, respectively. The energy is given by

$$e = c_V \rho T + \frac{1}{2}\rho(u^2 + v^2), \tag{2}$$

where c_V is the specific heat capacity under constant volume and T is the temperature. Furthermore, we have $F = F^I - F^V$ and $G = G^I - G^V$, where the superscript I denotes the inviscid part of the flux and V the viscous part. The components of the flux vectors are given by

$$\begin{aligned} F^I &= [\rho u, p + \rho u^2, \rho uv, u(p+e)]^T, \\ G^I &= [\rho v, \rho uv, p + \rho v^2, v(p+e)]^T, \\ F^V &= [0, \tau_{xx}, \tau_{xy}, u\tau_{xx} + v\tau_{xy} + \kappa T_x]^T, \\ G^V &= [0, \tau_{xy}, \tau_{yy}, u\tau_{yx} + v\tau_{yy} + \kappa T_y]^T, \end{aligned} \tag{3}$$

where we have the pressure p and the thermal conductivity coefficient κ. The stress tensor is given by

$$\tau_{xx} = 2\mu\frac{\partial u}{\partial x} + \lambda\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right), \tau_{yy} = 2\mu\frac{\partial v}{\partial y} + \lambda\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right), \tau_{xy} = \tau_{yx} = \mu\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right), \tag{4}$$

where μ and λ are the dynamic and second viscosity, respectively. To close the system we need to include an equation of state, for example the ideal gas law

$$p = \rho R T. \tag{5}$$

Here $R = c_P - c_V$ is the specific gas constant and c_P the specific heat capacity under constant pressure. Both c_P and c_V are considered constants in this paper.

Since the aim is to model heat transfer in a solid using the Navier-Stokes equations, we study the equations with vanishing velocities. If we let $u = v = 0$, all the convective terms and viscous stresses are zero and by using (2) and (5), equation (1)

reduces to

$$\begin{aligned} \rho_t &= 0 \\ p_x &= 0 \\ p_y &= 0 \\ T_t &= \frac{1}{c_V \rho} \left((\kappa T_x)_x + (\kappa T_y)_y \right). \end{aligned} \tag{6}$$

The last equation is similar, but not identical, to the variable coefficient heat equation.

For ease of comparison with the heat equation we transform to non-dimensional form as follows (note the slight abuse of notation since we let the dimensional and non-dimensional variables have the same notation. Hereafter, all quantities are non-dimensional):

$$u = \frac{u^*}{c^*_\infty}, \qquad v = \frac{v^*}{c^*_\infty}, \qquad \rho = \frac{\rho^*}{\rho^*_\infty}, \qquad T = \frac{T^*}{T^*_\infty}, \tag{7}$$

$$p = \frac{p^*}{\rho^*_\infty (c^*_\infty)^2}, \qquad e = \frac{e^*}{\rho^*_\infty (c^*_\infty)^2}, \qquad \lambda = \frac{\lambda^*}{\mu^*_\infty}, \qquad \mu = \frac{\mu^*}{\mu^*_\infty}, \tag{8}$$

$$c_P = \frac{c^*_P}{c^*_{P\infty}}, \qquad c_V = \frac{c^*_V}{c^*_{P\infty}}, \qquad R = \frac{R^*}{c^*_{P\infty}}, \qquad \kappa = \frac{\kappa^*}{\kappa^*_\infty}, \tag{9}$$

$$x = \frac{x^*}{L^*_\infty}, \qquad y = \frac{y^*}{L^*_\infty}, \qquad t = \frac{c^*_\infty}{L^*_\infty} t^*, \tag{10}$$

where the $*$-superscript denotes a dimensional variable and the ∞-subscript the reference value. L^*_∞ is a characteristic length scale and c^*_∞ is the reference speed of sound. The equation of state (5) becomes in non-dimensional form

$$\gamma p = \rho T. \tag{11}$$

and the energy equation can be written as

$$e = \frac{p}{\gamma - 1} + \frac{1}{2} \rho (u^2 + v^2). \tag{12}$$

By using (7)-(10), the last equation in (6) becomes

$$T_t = \frac{1}{Pe_c} \frac{1}{c_V \rho} \left((\kappa T_x)_x + (\kappa T_y)_y \right) \tag{13}$$

where

$$Pe_c = \frac{c^*_\infty L^*_\infty}{\alpha^*_\infty}, \quad \alpha^*_\infty = \frac{\kappa^*_\infty}{\rho^*_\infty c^*_{P\infty}} \tag{14}$$

are the Péclet number based on the reference speed of sound and the thermal diffusivity, respectively.

3. Similarity conditions

Since the fluid is compressible, the density in (6) is non-constant and the energy component in the Navier-Stokes equations will differ from the constant coefficient heat equation. We can however quantify in which way the equations differ and which terms that have to be minimized in order for the two equations to be as similar as possible. The heat equation, non-dimensionalized using (7)-(10), can be written as

$$\tilde{T}_t = \frac{1}{Pe_c}\frac{1}{c_s\rho_s}\left(\left(\kappa_s\tilde{T}_x\right)_x + \left(\kappa_s\tilde{T}_y\right)_y\right) \tag{15}$$

where Pe_c is defined in (14) and c_s, ρ_s, κ_s are the specific heat capacity, density and thermal conductivity of the solid, respectively. In this case, all coefficients are constant but rewritten in a form which resembles (13).

In order to compare (13) and (15), we define $\beta = Pe_c\rho c_V, \beta_s = Pe_c\rho_s c_s$ and rewrite (13) and (15) as

$$\beta T_t = (\kappa T_x)_x + (\kappa T_y)_y\,, \tag{16}$$

$$\beta\tilde{T}_t = \frac{\beta}{\beta_s}\left(\left(\kappa_s\tilde{T}_x\right)_x + \left(\kappa_s\tilde{T}_y\right)_y\right). \tag{17}$$

Note that β_s is constant for the solid. Furthermore, since $\beta > 0$ and (6) yields $\frac{\partial\beta}{\partial t} = 0$, we can estimate the difference $T - \tilde{T}$ in the β-norm defined by

$$||T - \tilde{T}||^2_\beta = \int_\Omega \left(T - \tilde{T}\right)^2 \beta d\Omega \tag{18}$$

where Ω is the computational domain. By subtracting (17) from (16), multiplying

with $T - \tilde{T}$ and integrating over Ω we obtain

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}||T-\tilde{T}||_\beta^2 = &-\int_\Omega \left(\kappa\nabla T\cdot\nabla T + \frac{\kappa_s}{\beta_s}\beta\nabla\tilde{T}\cdot\nabla\tilde{T}\right)\\ &+\oint_{\partial\Omega}\left(T-\tilde{T}\right)\left(\kappa\nabla T - \frac{\kappa_s}{\beta_s}\beta\nabla\tilde{T}\right)\cdot n ds\\ &+\int_\Omega \frac{\kappa_s}{\beta_s}\left(T-\tilde{T}\right)\nabla\beta\cdot\nabla\tilde{T}d\Omega + \int_\Omega\left(\kappa+\frac{\kappa_s}{\beta_s}\beta\right)\nabla T\cdot\nabla\tilde{T}d\Omega.\end{aligned} \tag{19}$$

In order to obtain as similar temperature distributions from the heat equation and Navier-Stokes equation as possible, the right-hand-side of (19) has to be less than or equal to zero. Note that we specify the same boundary data for T and $\tilde{T}$, in which case the boundary integral is zero. By further assuming that $\nabla\beta = 0$ we can rewrite (19) as the quadratic form

$$\frac{d}{dt}||T-\tilde{T}||_\beta^2 = -\int_\Omega \begin{bmatrix}\nabla T\\ \nabla\tilde{T}\end{bmatrix}^T \begin{bmatrix} 2\kappa & -\left(\kappa+\frac{\kappa_s}{\beta_s}\beta\right)\\ -\left(\kappa+\frac{\kappa_s}{\beta_s}\beta\right) & 2\frac{\kappa_s}{\beta_s}\beta\end{bmatrix}\begin{bmatrix}\nabla T\\ \nabla\tilde{T}\end{bmatrix}. \tag{20}$$

By computing the eigenvalues of the matrix in (20) and requiring that they be non-negative, we can conclude that we need $\kappa - \frac{\kappa_s}{\beta_s}\beta = 0$. Thus, if the relations

$$\frac{\kappa}{\beta} - \frac{\kappa_s}{\beta_s} = 0, \quad \nabla\beta = 0 \tag{21}$$

hold, then

$$\frac{d}{dt}||T-\tilde{T}||_\beta^2 \leq 0 \tag{22}$$

and the Navier-Stokes equations and the heat equation produces the exact same solution for the temperature if given identical initial data.

Remark 1. The heat equation and energy component in the Navier-Stokes equations produces exactly the same results only if the relations in (21) hold. In a numerical simulation, the initial, and boundary, data are chosen such that (21) holds exactly to begin with. Because of the weak imposition of the boundary and interface conditions, the relations will no longer hold as time passes. Small variations in the velocities at the boundaries and interfaces will produce small variations in the density which

propagate into the domain. These deviations are however very small and the effects are studied in later sections.

4. SBP-SAT discretization

In the basic formulation, the first derivative is approximated by an operator on SBP form

$$u_x \approx Dv = P^{-1}Qv, \tag{23}$$

where v is the discrete grid function approximating u. The matrix P is symmetric, positive definite and defines a discrete norm by $||v||^2 = v^T P v$. In this paper, we consider diagonal norms only. The matrix Q is almost skew-symmetric and satisfies the SBP property $Q + Q^T = \text{diag}[-1, 0, \ldots, 0, 1]$. There are SBP operators based on diagonal norms with 2nd, 3rd, 4th and 5th order accuracy, and the stability analysis does not depend on the order of the operators [21, 25]. The second derivative is approximated either using the first derivative twice, i.e.

$$u_{xx} \approx D^2 v = (P^{-1}Q)^2 v. \tag{24}$$

or a compact formulation with minimal bandwidth [22, 23]. In the conservative formulation of the Navier-Stokes equations, the second derivative operator is not used.

In order to extend the operators to higher dimensions, it is convenient to introduce the Kronecker product. For arbitrary matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times q}$, the Kronecker product is defined as

$$A \otimes B = \begin{bmatrix} a_{1,1}B & \ldots & a_{1,m}B \\ \vdots & \ddots & \vdots \\ a_{n,1}B & \ldots & a_{m,n}B \end{bmatrix}. \tag{25}$$

The Kronecker product is bilinear, associative and obeys the mixed product property

$$(A \otimes B)(C \otimes D) = (AC \otimes BD) \tag{26}$$

if the usual matrix products are defined. For inversion and transposing we have

$$(A \otimes B)^{-1,T} = A^{-1,T} \otimes B^{-1,T} \tag{27}$$

if the usual matrix inverse is defined. The Kronecker product is not commutative in

general, but for square matrices A and B there is a permutation matrix R such that

$$A \otimes B = R^T (B \otimes A) R. \tag{28}$$

Let $P_{x,y}$, $Q_{x,y}$ and $D_{x,y}$ denote the difference operators in the coordinate direction indicated by the subscript. The extension to multiple dimensions is done by using the Kronecker product as follows:

$$\begin{aligned} \bar{P}_x &= P_x \otimes I_y, \quad \bar{Q}_x = Q_x \otimes I_y, \\ \bar{P}_y &= I_x \otimes P_y, \quad \bar{Q}_y = I_x \otimes Q_y, \\ \bar{D}_x &= D_x \otimes I_y, \quad \bar{D}_y = I_x \otimes D_y. \end{aligned} \tag{29}$$

Due to the mixed product property (26), the operators commute in different coordinate directions and hence differentiation can be performed in each coordinate direction independently. The norm is defined by

$$||u||^2 = u^T \bar{P} u \tag{30}$$

where $\bar{P} = \bar{P}_x \bar{P}_y = P_x \otimes P_y$.

5. Temperature coupling of the Navier-Stokes equations

The compressible Navier-Stokes equations in two space dimensions requires three boundary conditions at a solid wall [20]. Since we are aiming for modelling heat transfer in a solid using (1), both the tangential and normal velocities are zero. The third condition is used to couple the temperature in the fluid and solid domain.

We consider the Navier-Stokes equations in the two domains $\Omega_1 = [0,1] \times [0,1]$ and $\Omega_2 = [0,1] \times [-1,0]$ with an interface at $y = 0$. Denote the solution in Ω_1 by $q = [\rho, \rho u, \rho v, e]$ and in Ω_2 by $\tilde{q} = [\tilde{\rho}, \tilde{\rho}\tilde{u}, \tilde{\rho}\tilde{v}, \tilde{e}]$.

The interface will be considered as a solid wall and hence we impose no-slip interface conditions for the velocities

$$\begin{aligned} u &= 0, \quad v = 0, \\ \tilde{u} &= 0, \quad \tilde{v} = 0. \end{aligned} \tag{31}$$

More general interface conditions can be imposed by considering Robin conditions as described in [26].

To couple the temperature of the two equations we will use continuity of temperature and heat fluxes,

$$T = \tilde{T}, \quad \kappa_1 T_y = \kappa_2 \tilde{T}_y. \tag{32}$$

For the purpose of analysis, we consider the linearized, frozen coefficient and symmetric Navier-Stokes equations

$$\begin{aligned} w_t + (A_1 w)_x + (A_2 w)_y &= \varepsilon \left((A_{11} w_x + A_{12} w_y)_y + (A_{21} w_x + A_{22} w_y)_y \right), \\ \tilde{w}_t + (B_1 \tilde{w})_x + (B_2 \tilde{w})_y &= \varepsilon \left((B_{11} \tilde{w}_x + B_{12} \tilde{w}_y)_y + (B_{21} \tilde{w}_x + B_{22} \tilde{w}_y)_y \right), \end{aligned} \tag{33}$$

where $\epsilon = \frac{Ma}{Re}$, Re is the Reynolds number and Ma is the Mach number. The coefficient matrices can be found in [20, 27]. The symmetrized variables are

$$w = \left[\frac{\bar{c}}{\sqrt{\gamma}\bar{\rho}} \rho, u, v, \frac{1}{\bar{c}\sqrt{\gamma(\gamma-1)}} T \right]^T, \tag{34}$$

where an overbar denotes the constant state which we have linearized around. More details can be found in [28, 20, 27]. This procedure is motivated by the principle of linearization and localization [29]. Note that the linerarization around $u = v = 0$, and hence $\bar{u} = \bar{v} = 0$, is exact at the interface due to the interface conditions. The well-posedness of (33) with the conditions (31) and (32) are shown in

Proposition 1. *The coupled compressible Navier-Stokes equations are well-posed using the interface conditions (31) and (32).*

Proof. The energy estimates of w and $\tilde{w}$ will be derived in the L^2-equivalent norms

$$||w||^2_{H_1} = \int_{\Omega_1} w^T H_1 w d\Omega, \quad ||\tilde{w}||^2_{H_2} = \int_{\Omega_2} \tilde{w}^T H_2 \tilde{w} d\Omega \tag{35}$$

where

$$H_{1,2} = \text{diag}[1, 1, 1, \delta_{1,2}], \quad \delta_{1,2} > 0 \tag{36}$$

are to be determined. We apply the energy method and consider only the terms at the interface $y = 0$. We get by using the conditions (31) that

$$\frac{d}{dt} \left(||w||^2_{H_1} + ||\tilde{w}||^2_{H_2} \right) \leq -2\varepsilon \int_0^1 \left(\frac{\delta_1 \bar{\mu}_1}{\bar{\rho}_1 \bar{c}_1^2 (\gamma_1 - 1) Pr_1} T T_y - \frac{\delta_2 \bar{\mu}_2}{\bar{\rho}_2 \bar{c}_2^2 (\gamma_2 - 1) Pr_2} \tilde{T} \tilde{T}_y \right) dx, \tag{37}$$

where the bar denotes the state around which we have linearized and the subscript 1 or 2 refer to values from the corresponding subdomain Ω_1 or Ω_2. By requiring

continuity of temperature ($T = \tilde{T}$) equation (37) reduces to

$$\frac{d}{dt}\left(||w||_{H_1}^2 + ||\tilde{w}||_{H_2}^2\right) \leq -2\varepsilon \int_0^1 T\left(\frac{\delta_1 \bar{\kappa}_1}{\bar{\rho}_1 \bar{c}_1^2(\gamma_1 - 1)c_{P_1}} T_y - \frac{\delta_2 \bar{\kappa}_2}{\bar{\rho}_2 \bar{c}_2^2(\gamma_2 - 1)c_{P_2}} \tilde{T}_y\right) dx. \tag{38}$$

In order to obtain an energy estimate by using continuity of the heat fluxes, we need to choose the weights

$$\delta_1 = \bar{\rho}_1 \bar{c}_1^2(\gamma_1 - 1)c_{P_1}, \quad \delta_2 = \bar{\rho}_2 \bar{c}_2^2(\gamma_2 - 1)c_{P_2} \tag{39}$$

since then

$$\frac{d}{dt}\left(||w||_{H_1}^2 + ||\tilde{w}||_{H_2}^2\right) \leq -2\varepsilon \int_0^1 T\left(\bar{\kappa}_1 T_y - \bar{\kappa}_2 \tilde{T}_y\right) dx = 0. \tag{40}$$

Hence the interface conditions (32) gives an energy estimate and no unbounded energy growth can occur. □

Remark 2. The physical interface conditions (32) requires an estimate in a different norm than the standard L^2-norm. The norm defined by the (positive) weights in (39) is, however, only a scaling of the L^2-norm and they are hence equivalent. From a mathematical point of view, any interface condition which give positive weights will result in a well-posed coupling.

5.1. The discrete problem and stability

In [12], a stable and conservative multi-block coupling of the Navier-Stokes equations was developed. The coupling was done by considering continuity of all quantities and of the fluxes with the purpose of being able to handle different coordinate transforms in different blocks. In our case, the velocities are uncoupled and the equations are coupled only by continuity of temperature and heat fluxes. This enable us to compute conjugate heat problems by modifying the interface conditions for the multi-block coupling.

We consider again the formulation (33) and discretize using SBP-SAT for imposing the interface conditions (31) and (32) weakly. We let for simplicity the subdomains be discretized by equally many uniformly distributed gridpoints which allow us to use the same difference operators in both subdomains. We stress that the subdomains can have different discretizations [12, 30], this assumption merely simplifies the notation and avoids the use of too many subscripts.

We discretize (33) using the SBP-SAT technique as

$$\begin{aligned} \mathbf{w}_t + \hat{D}_x \mathbf{F} + \hat{D}_y \mathbf{G} &= \mathbb{S}, \\ \tilde{\mathbf{w}}_t + \hat{D}_x \tilde{\mathbf{F}} + \hat{D}_y \tilde{\mathbf{G}} &= \tilde{\mathbb{S}}, \end{aligned} \tag{41}$$

where the discrete fluxes are given by

$$\begin{aligned} \mathbf{F} &= \hat{A}_1 \mathbf{w} - \varepsilon \left(\hat{A}_{11} \hat{D}_x \mathbf{w} + \hat{A}_{12} \hat{D}_y \mathbf{w} \right), \\ \mathbf{G} &= \hat{A}_2 \mathbf{w} - \varepsilon \left(\hat{A}_{21} \hat{D}_x \mathbf{w} + \hat{A}_{22} \hat{D}_y \mathbf{w} \right), \\ \tilde{\mathbf{F}} &= \hat{B}_1 \tilde{\mathbf{w}} - \varepsilon \left(\hat{B}_{11} \hat{D}_x \tilde{\mathbf{w}} + \hat{B}_{12} \hat{D}_y \tilde{\mathbf{w}} \right), \\ \tilde{\mathbf{G}} &= \hat{B}_2 \tilde{\mathbf{w}} - \varepsilon \left(\hat{B}_{21} \hat{D}_x \tilde{\mathbf{w}} + \hat{B}_{22} \tilde{D}_y \tilde{\mathbf{w}} \right). \end{aligned} \tag{42}$$

The hat notation denotes that the matrix has been extended to the entire system as

$$\begin{aligned} \hat{D}_x &= D_x \otimes I_y \otimes I_4, \quad \hat{D}_y = I_x \otimes D_y \otimes I_4, \\ \hat{A}_\xi &= I_x \otimes I_y \otimes A_\xi, \quad \hat{B}_\xi = I_x \otimes I_y \otimes B_\xi, \end{aligned} \tag{43}$$

where ξ is a generic index ranging over the indicies which occur in (42).

The SATs imposing the interface conditions (31) and (32) can be written as

$$\begin{aligned} \mathbb{S} = {} & \hat{P}_y^{-1} \hat{E}_{x,y_N} \hat{\Sigma}_1 \left(\mathbf{w} - g^I \right) + \varepsilon \sigma_2 \hat{P}_y^{-1} \hat{E}_{x,y_N} \left(\hat{H}_2 \mathbf{w} - g_1 \right) \\ & + \varepsilon \sigma_3 \hat{P}_y^{-1} \hat{E}_{x,y_N} \left(\hat{H}_3 \mathbf{w} - g_2 \right) + \varepsilon \hat{P}_y^{-1} \hat{E}_{x,y_N} \hat{\Theta}_1 \left(\hat{H}_3 \hat{D}_x \mathbf{w} - \frac{\partial g_2}{\partial x} \right) \\ & + \varepsilon \sigma_4 \hat{P}_y^{-1} \hat{E}_{x,y_N} \left(\hat{I}_1^T \mathbf{w} - \hat{I}_2^T \tilde{\mathbf{w}} \right) \\ & + \varepsilon \sigma_5 \hat{P}_y^{-1} \hat{D}_y^T \hat{E}_{x,y_N} \left(\hat{I}_1^T \mathbf{w} - \hat{I}_2^T \tilde{\mathbf{w}} \right) \\ & + \varepsilon \sigma_6 \hat{P}_y^{-1} \hat{E}_{x,y_N} \left(\bar{\kappa}_1 \hat{I}_1^T \hat{D}_y \mathbf{w} - \bar{\kappa}_2 \hat{I}_2^T \hat{D}_y \tilde{\mathbf{w}} \right) \end{aligned} \tag{44}$$

and

$$\begin{aligned}\tilde{\mathbb{S}} = &\hat{P}_y^{-1}\hat{E}_{x,y_0}\hat{\Sigma}_2\left(\tilde{\mathbf{w}}-\tilde{g}^I\right)+\varepsilon\tilde{\sigma}_2\hat{P}_y^{-1}\hat{E}_{x,y_0}\left(\hat{H}_2\tilde{\mathbf{w}}-\tilde{g}_1\right)\\&+\varepsilon\tilde{\sigma}_3\hat{P}_y^{-1}\hat{E}_{x,y_0}\left(\hat{H}_3\tilde{\mathbf{w}}-\tilde{g}_2\right)+\varepsilon\hat{P}_y^{-1}\hat{E}_{x,y_0}\hat{\Theta}_2\left(\hat{H}_3\hat{D}_x\tilde{\mathbf{w}}-\frac{\partial\tilde{g}_2}{\partial x}\right)\\&+\varepsilon\tilde{\sigma}_4\hat{P}_y^{-1}\hat{E}_{x,y_0}\left(\hat{I}_2^T\tilde{\mathbf{w}}-\hat{I}_1^T\mathbf{w}\right)\\&+\varepsilon\tilde{\sigma}_5\hat{P}_y^{-1}\hat{D}_y^T\hat{E}_{x,y_0}\left(\hat{I}_2^T\tilde{\mathbf{w}}-\hat{I}_1^T\mathbf{w}\right)\\&+\varepsilon\tilde{\sigma}_6\hat{P}_y^{-1}\hat{E}_{x,y_0}\left(\bar{\kappa}_2\hat{I}_2^T\hat{D}_y\tilde{\mathbf{w}}-\bar{\kappa}_1\hat{I}_1^T\hat{D}_y\mathbf{w}\right).\end{aligned}\tag{45}$$

Here $\hat{P}=\bar{P}\otimes I_4$, $\hat{E}_{x,y_0}=\bar{E}_{x,y_0}\otimes I_4$, $\hat{H}_j=I_x\otimes I_y\otimes H_j$ and H_j is a 4×4 matrix with the only non-zero element 1 at the (j,j)th position on the diagonal and the operators $\hat{I}_{1,2}$ selects the interface elements. The penalty matrices $\hat{\Sigma}_{1,2}=I_x\otimes I_y\otimes\Sigma_{1,2}$, $\hat{\Theta}_{1,2}=I_x\otimes I_y\otimes\Theta_{1,2}$, and the penalty coefficients $\sigma_{2,\ldots,6}$ and $\tilde{\sigma}_{2,\ldots,6}$ has to be determined such that the scheme is stable.

Remark 3. The terms which involve $\hat{\Theta}_{1,2}$ originate from the fact that the boundary condition $v=0$ implies that $v_x=0$, which is used to obtain an energy estimate in the continuous case. The terms hence represent the artificial boundary condition $v_x=0$ which is needed to obtain an energy estimate in the discrete case.

Remember that in the energy estimates for the continuous coupling, a non-standard L^2-equivalent norm was used. The same modification to the norms has to be done in the discrete case. Thus, the discrete energy estimates will be derived in the norms

$$||\mathbf{w}||_{\hat{J}_1}^2=\mathbf{w}^T\hat{P}\hat{J}_1\mathbf{w},\quad ||\tilde{\mathbf{w}}||_{\hat{J}_2}^2=\tilde{\mathbf{w}}^T\hat{P}\hat{J}_2\tilde{\mathbf{w}},\tag{46}$$

where,

$$\hat{J}_1=I_x\otimes I_y\otimes H_1,\quad \hat{J}_2=I_x\otimes I_y\otimes H_2,\tag{47}$$

and the matrices $H_{1,2}$ are defined in (36) with the weights given in (39). Note that $\hat{P}\hat{J}_{1,2}=\hat{J}_{1,2}\hat{P}$.

By applying the energy method to (41) and adding up we get

$$\frac{d}{dt}||\mathbf{w}||_{\hat{J}_1}^2+\frac{d}{dt}||\tilde{\mathbf{w}}||_{\hat{J}_2}^2+DI=IT\tag{48}$$

where the dissipation term, DI, is given by

$$\begin{aligned} DI = 2\varepsilon \begin{bmatrix} \hat{D}_x\mathbf{w} \\ \hat{D}_y\mathbf{w} \end{bmatrix}^T \begin{bmatrix} \hat{P}\hat{J}_1 & 0 \\ 0 & \hat{P}\hat{J}_1 \end{bmatrix} \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} \\ \hat{A}_{21} & \hat{A}_{22} \end{bmatrix} \begin{bmatrix} \hat{D}_x\mathbf{w} \\ \hat{D}_y\mathbf{w} \end{bmatrix} \\ + 2\varepsilon \begin{bmatrix} \hat{D}_x\tilde{\mathbf{w}} \\ \hat{D}_y\tilde{\mathbf{w}} \end{bmatrix}^T \begin{bmatrix} \hat{P}\hat{J}_2 & 0 \\ 0 & \hat{P}\hat{J}_2 \end{bmatrix} \begin{bmatrix} \hat{B}_{11} & \hat{B}_{12} \\ \hat{B}_{21} & \hat{B}_{22} \end{bmatrix} \begin{bmatrix} \hat{D}_x\tilde{\mathbf{w}} \\ \hat{D}_y\tilde{\mathbf{w}} \end{bmatrix}. \end{aligned} \tag{49}$$

The interface terms can be split into three parts as $IT = IT_1 + IT_2 + IT_3$ where IT_1 are the inviscid terms, IT_2 the velocity terms and IT_3 the coupling terms related to the temperature.

In [26] it is shown how to choose $\Sigma_{1,2}$, $\Theta_{1,2}$, $\sigma_{2,3}$ and $\tilde{\sigma}_{2,3}$, with small modifications, such that the inviscid and velocity terms are bounded. Here we focus on the coupling terms. With appropriate choices of $\Sigma_{1,2}$, $\Theta_{1,2}$, $\sigma_{2,3}$ and $\tilde{\sigma}_{2,3}$ as described in [26] we get

$$\frac{d}{dt}||\mathbf{w}||^2_{H_1} + \frac{d}{dt}||\tilde{\mathbf{w}}||^2_{H_2} + DI \leq IT_3, \tag{50}$$

where IT_3 can be written as the quadratic form

$$IT_3 = -\varepsilon(R\xi)^T \left(\mathbf{P}_x \otimes M\right) R\xi. \tag{51}$$

To obtain (51), we have used the permutation similarity property of the Kronecker product, R is a permutation matrix and $\xi = [\mathbf{T}_i, \tilde{\mathbf{T}}_i, (D_y\mathbf{T})_i, (D_y\tilde{\mathbf{T}})_i]^T$ where the subscript i denotes the values at the interface. Note that we do not need the specific form of R, it is sufficient to know that such a matrix exists. Furthermore, we have

$$\mathbf{P}_x = \mathrm{diag}[\delta_1 P_x, \delta_2 P_x, \delta_1 P_x, \delta_2 P_x], \tag{52}$$

with $\delta_{1,2}$ from (39), and

$$M = \begin{bmatrix} -2\sigma_4 & \sigma_4 + \tilde{\sigma}_4 & \bar{\kappa}_1\gamma_1 - \sigma_5 - \bar{\kappa}_1\sigma_6 & \bar{\kappa}_2\sigma_6 + \tilde{\sigma}_5 \\ \sigma_4 + \tilde{\sigma}_4 & -2\tilde{\sigma}_4 & \sigma_5 + \bar{\kappa}_1\tilde{\sigma}_6 & -\bar{\kappa}_2\gamma_1 - \tilde{\sigma}_5 - \bar{\kappa}_2\tilde{\sigma}_6 \\ \bar{\kappa}_1\gamma_1 - \sigma_5 - \bar{\kappa}_1\sigma_6 & \sigma_5 + \bar{\kappa}_1\tilde{\sigma}_6 & 0 & 0 \\ \bar{\kappa}_2\sigma_6 + \tilde{\sigma}_5 & -\bar{\kappa}_2\gamma_1 - \tilde{\sigma}_5 - \bar{\kappa}_2\tilde{\sigma}_6 & 0 & 0 \end{bmatrix}. \tag{53}$$

Since $\mathbf{P}_x$ is positive definite and the Kronecker product preserves positive definiteness, the necessary requirement for (50) to be bounded is that the penalty coefficients are chosen such that $M \geq 0$. The penalty coefficients are given in

Theorem 1. *The coupling terms IT_3 in (50) are bounded using*

$$\tilde{\sigma}_4 = \sigma_4 \leq 0, \quad \sigma_5 = -\bar{\kappa}_1 r, \quad \sigma_6 = \gamma + r, \quad \tilde{\sigma}_5 = -\bar{\kappa}_2(\gamma_1 + r), \quad \tilde{\sigma}_6 = r, \quad r \in \mathbb{R} \tag{54}$$

and hence the scheme (41) is stable.

Proof. With the choices of penalty coefficients given in Proposition 1, the matrix M in (53) reduces to

$$M = 2\sigma_4 \begin{bmatrix} -1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{55}$$

with eigenvalues $\lambda_{1,2,3} = 0$ and $\lambda_4 = -4\sigma_4$. Hence if $\sigma_4 \leq 0$ we have $M \geq 0$. □

6. Numerical results

To verify the numerical scheme we use what is often called the method of manufactured solutions [4, 31]. We chose the solution and use that to compute a right-hand-side forcing function, initial- and boundary data. According to the principle of Duhamel [32], the number or form of the boundary conditions does not change due to the addition of the forcing function. We can hence test the convergence of the scheme towards this analytical solution. The interface conditions (32) are of course not satisfied in general by this solution and we need to modify them by adding a right-hand-side.

We use the manufactured solution

$$\begin{aligned} \rho(x, y, t) &= 1 + \eta \sin(\theta\pi(x - y) - t)^2 \\ u(x, y, t) &= \eta \cos(\theta\pi(x + y) - t) \\ v(x, y, t) &= \eta \sin(\theta\pi(x - y) - t) \\ p(x, y, t) &= 1 + \eta \cos(\theta\pi(x + y) - t)^2, \end{aligned} \tag{56}$$

with different values of η, θ in the fluid and solid domains, to generate the solution. The energy and temperature can be computed using (11) and (12). Since the stability of the scheme is independent of the order of accuracy, the difference operators is the only thing which have to be changed in order to achieve higher, or lower, accuracy. The rate of convergence, Q, is computed as

$$Q^{(j)} = \frac{1}{\log\left(\frac{N_{i+1}}{N_i}\right)} \log\left(\frac{E_i^{(j)}}{E_{i+1}^{(j)}}\right) \tag{57}$$

for each of the conserved varables $q^{(j)}, j = 1, 2, 3, 4$. We have used the same number of grid points, N, in both coordinate directions for both the fluid and solid domain. N_k denotes the number of gridpoints at refinement level k and $E_k^{(j)}$ is the L_2-error between the computed and exact solution for each conserved variable. The time integration is done with the classical 4th-order Runge-Kutta method until time $t = 0.1$ using 1000 time steps.

In Table 1 we list the convergence results for the conserved variables for both the fluid and solid domains. As we can see from Table 1 we can achieve 5th-order accuracy by simply replacing the difference operators. No other modifications to the scheme is necessary.

Table 1: Convergence results for the conjugate heat transfer problem

	2nd-order			3rd-order		
N	32/64	64/96	96/128	32/64	64/96	96/128
ρ	1.8367	1.8931	2.0133	2.6222	3.0699	3.4795
ρu	2.0824	2.0803	2.1187	2.9846	3.0748	3.1927
ρv	2.0503	2.0549	2.0922	3.4222	3.7512	3.4199
e	1.8174	1.9065	1.9963	2.4639	2.7749	3.0523
$\tilde{\rho}$	1.8933	1.8533	1.9628	2.5761	2.9791	3.5767
$\tilde{\rho}\tilde{u}$	2.0544	2.0803	2.0992	3.1094	3.0374	3.2732
$\tilde{\rho}\tilde{v}$	1.9411	2.0190	2.0894	3.3928	3.7465	3.3628
$\tilde{e}$	1.9483	1.9151	1.9409	2.9451	2.8399	3.2560
	4th-order			5th-order		
N	32/64	64/96	96/128	32/64	64/96	96/128
ρ	3.9662	4.1381	4.1138	4.4824	5.2584	5.5131
ρu	4.4531	4.3640	4.2799	4.6819	4.7521	4.6733
ρv	4.3175	4.0918	4.0284	4.9824	4.9257	4.7839
e	3.9757	4.1723	4.0957	4.3760	4.6227	4.7207
$\tilde{\rho}$	3.9935	4.3902	4.5538	4.4421	5.1497	5.5388
$\tilde{\rho}\tilde{u}$	4.2072	4.3159	4.4366	4.9665	4.9739	4.9512
$\tilde{\rho}\tilde{v}$	4.3672	4.3331	4.3212	5.1007	5.1370	4.9087
$\tilde{e}$	3.9025	4.3178	4.4091	4.8746	4.8573	4.9518

6.1. Comparison of the different approaches to the conjugate heat transfer problem

When the heat transfer in the solid is governed by the compressible Navier-Stokes equations, one does not solve the same conjugate heat transfer problem as when the heat transfer is governed by the heat equation. This is because the relations in (21)

holds only approximately as time passes. The exchange of heat between the fluid and solid domains will affect the temperature and hence also the density, because of the equation of state, and introduce small density variations in the solid domain. We can numerically solve the conjugate heat transfer problem in both ways and determine the difference between the two methods. Note that we do not overwrite, or enforce, the velocities to zero inside the solid domain. The velocities are weakly enforced to zero at the boundaries and interfaces only.

Let NS-NS denote the case when the heat transfer is governed by the compressible Navier-Stokes equations and NS-HT the case where the heat transfer is governed by the heat equation. The well-posedness and stability of NS-HT coupling is proven in the appendix. The initial and boundary data are chosen such that NS-NS and NS-HT have identical solutions initially, and we study the differences of the two methods over time.

To quantify the difference between the two methods, NS-NS and NS-HT, we compute two representative cases. The computational domain is $\Omega = \Omega_1 \cup \Omega_2$ where $\Omega_1 = [0,1]\times[0,1]$ and $\Omega_2 = [0,1]\times[-1,0]$. All computations are done using 3rd-order accurate SBP operators and the time integration is done using the classical 4th-order Runge-Kutta method.

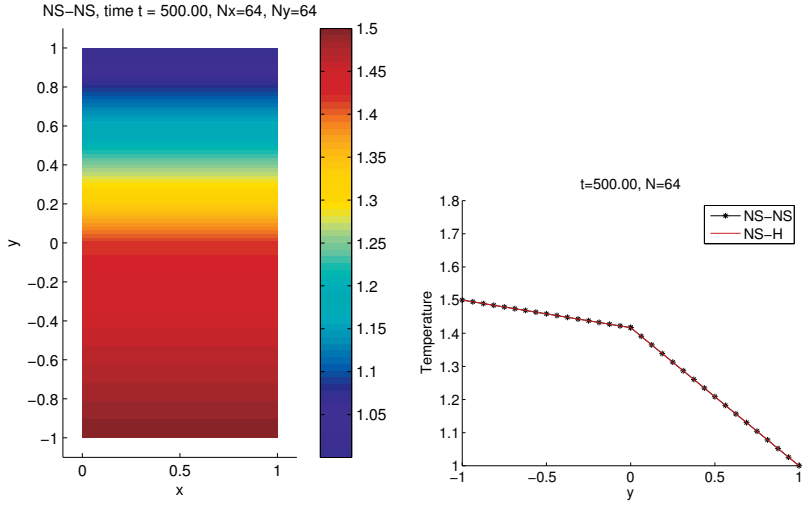
In the first case, the computations are initialized with zero velocities everywhere and temperature $T = 1$ in both subdomains. In the x-direction we have chosen periodic boundary conditions. At $y = -1$ we specify $T = 1.5$ and at $y = 1$ we have $T = 1$. For the Navier-Stokes equations we have no-slip solid walls as described in [26] for the velocities. These choices of boundary conditions renders the solution to be homogeneous in the x-direction.

Under the assumption of identically zero velocities and periodicity in the x-direction, the exact steady-state solution can be obtained as

$$\begin{aligned} T &= -\frac{k}{2(k+1)}y + \frac{3k+2}{2(k+1)}, \\ \tilde{T} &= -\frac{1}{2(k+1)}y + \frac{3k+2}{2(k+1)}, \end{aligned} \tag{58}$$

where $k = \kappa_2/\kappa_1$ is the ratio of the steady-state thermal conductivities. We can see from (58) that the only occurring material parameter is the ratio between the thermal conductivity coefficients. Neither the density nor the thermal diffusivity has any effect on the steady-state solution. The larger the ratio of the thermal conductivities is, the stiffer the problem becomes. In the calculations below, we have chosen the parameters such that $k = 5$.

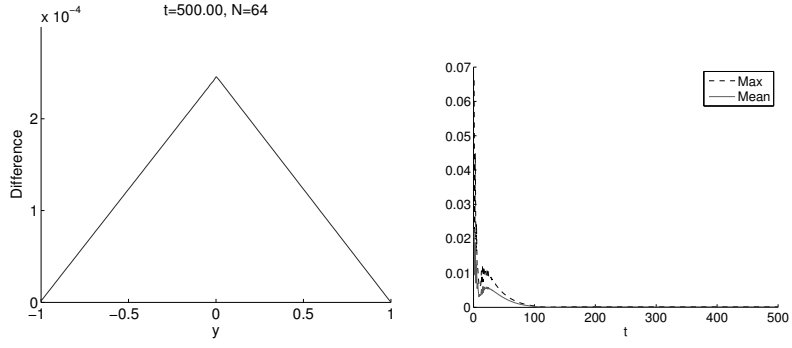
The temperature distribution at time $t = 500$, which is the steady-state solution, is seen in Figure 1 when using 65 grid points in each coordinate direction and subdomain. In Figure 2 we show an intersection of the absolute difference along the line $-1 \leq y \leq 1$ at $x = 0.5$ together with the time-evolution of the l_∞- and l_2-differences. In Figure 2(b) we can see that the large initial discontinuity gives differences in the beginning of the computation. As the velocities are damped over time, the difference decreases rapidly towards zero.



(a) Temperature distribution from NS-NS

(b) Intersection along y at $x = 0.5$ of the temperature distribution for NS-NS and NS-HT

Figure 1: Temperatures at time $t = 500$ from NS-NS and NS-HT using 65 grid points in each coordinate direction and subdomain



(a) Intersection along y at $x = 0.5$ of the absolute difference in temperature distribution between NS-NS and NS-HT

(b) l_∞- and l_2-difference in time

Figure 2: Temperature intersection and time differences for NS-NS and NS-HT using 65 grid points in each coordinate direction and subdomain

In Table 2 we list the results for different number of grid points.

Table 2: Difference between NS-NS and NS-HT at time $t = 500$

	Difference		
N	l_∞	l_2	Interface
32	1.1514e-03	6.8992e-04	1.1514e-03
64	2.4612e-04	1.4491e-04	2.4612e-04
128	4.3440e-05	2.5329e-05	4.3440e-05

As we can see from Table 2, the differences are very small. Even for the coarsest mesh, the relative maximum and interface differences are less than 0.1% while the relative l_2-difference is approximately 0.05%. Note that the differences are decreasing with the resolution. The steady-state solutions will become identical as the mesh is further refined.

Next, we consider an unsteady problem. The boundary data at the south boundary is perturbed by the time-dependent perturbation

$$f(x,t) = 1 + 0.25 * \sin(t) * \sin(\pi x) \tag{59}$$

and hence there will be no steady-state solution. In the x-direction in the solid domain, we have changed from periodic boundary conditions to solid wall boundary

conditions with prescribed temperature $T = 1$. This is a more realistic way to enclose the solid domain, and it has the additional benefit of damping the induced velocities in the Navier-Stokes equations.

The results can be seen in Figure (3). We plot the l_∞- and l_2-difference as a function of time. As we can see, the difference does not approach zero but remains bounded and small. The relative mean difference is less than 0.5% while the maximum difference is less than 1.5%. Thus, despite the rather large variation in the boundary data, NS-NS and NS-HT produces very similar solutions.

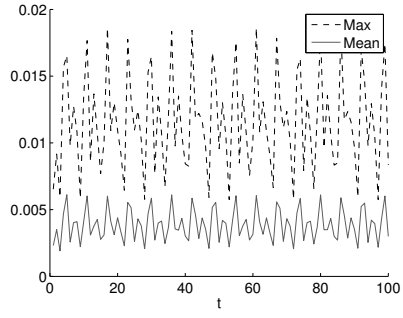
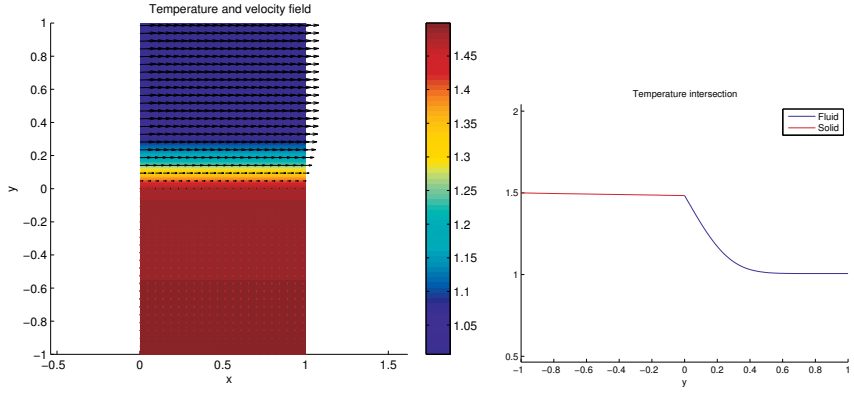


Figure 3: l_∞- and l_2-difference in time between NS-HS and NS-HT for an unsteady problem

In a CFD computation, the part of the domain which is solid is in general small compared to the fluid domain, for example when computing the flow field around an airfoil or aircraft. Despite the Navier-Stokes equations being significantly more expensive to solve, the overall additional cost of solving the Navier-Stokes equations also in the solid is in general limited.

6.2. A numerical example of conjugate heat transfer

As a final computational example, we consider the coupling of a flow over a slab of material for which the ratio of the thermal conductivities is 100. The initial temperature condition is $T = 1$ in the fluid domain and $\tilde{T} = 1.5$ in the solid domain. The boundary conditions are periodic in the x-direction. At the south boundary, $y = -1$, in the solid domain we let $\tilde{T} = 1.5$ and at the north boundary, $y = 1$, in the fluid domain, there is a Mach 0.5 free-stream boundary condition with $T = 1$, as described in [28]. Figure 4 shows a snapshot of the solution at time $t = 2.5$. The velocity components in the solid domain are zero to machine precision and the heat transfer in the solid is exclusively driven by diffusion.



(a) Temperature distribution and velocity profile

(b) Intersection of the temperature distribution along $x = 0.5$.

Figure 4: Temperature and velocity profiles for a flow past a slab of material using 65x65 grid points in both domains

7. Conclusions

We have proven that a conjugate heat transfer coupling of the compressible Navier-Stokes equations is well-posed when a modified norm is used. The equations were discretized using a finite difference method on summation-by-parts form with boundary- and interface conditions imposed weakly by the simultaneous approximation term. It was shown that a modified discrete norm was needed in order to prove energy stability of the scheme. The stability is independent of the order of accuracy, and it was shown that we can achieve all orders of accuracy by simply using higher order accurate SBP operators.

We showed that the difference when the heat transfer is governed by the heat equation, compared to the compressible Navier-Stokes equations, is small. The steady-state solutions differed by less than 0.005% as the mesh was refined while a perturbed, unsteady solution differed by less than 0.5% on average.

There are many multi-block codes for the compressible Navier-Stokes equations available. To implement conjugate heat transfer is significantly easier with the method of modifying the interface conditions, rather than coupling to a different physics solver for the heat transfer part. While the Navier-Stokes equations are more expensive to solve, usually only a small part of the computational domain is solid and the heat transfer is computed at a low additional cost.

Acknowledgments

The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project p2010056.

Appendix A. Coupling of the compressible Navier-Stokes equations with the heat equation

In [4], a model problem for conjugate heat transfer was considered. The equations were one-dimensional, linear and symmetric. In this appendix we extend the work to the two-dimensional compressible, non-linear Navier-Stokes equations coupled with the heat equation in two space dimensions. The well-posedness of the coupling is shown in

Proposition 2. *The compressible Navier-Stokes equations coupled with the heat equation, is well-posed with the interface conditions*

$$T = \tilde{T}, \quad \kappa T_y = \kappa_s \tilde{T}_y \tag{A.1}$$

for the temperature, and the no-slip[1] *conditions*

$$u = 0, \quad v = 0 \tag{A.2}$$

for the velocities.

Proof. Consider the heat equation (15) and the Navier-Stokes equations in the constant, linear, symmetric formulation. The estimates of w and $\tilde{T}$ will be derived in the L^2-equivalent norms

$$||w||^2_{J_1} = \int_{\Omega_1} w^T J_1 w d\Omega_1, \quad ||\tilde{T}||^2_{\nu_2} = \int_{\Omega_2} \tilde{T}^2 \nu_2 d\Omega_2 \tag{A.3}$$

where $J_1 = \text{diag}[1, 1, 1, \nu_1]$ and $\nu_{1,2} > 0$ are to be determined.

Remember that the symmetrized variables for the Navier-Stokes equations are

$$w = \left[\frac{\bar{c}}{\sqrt{\gamma}\bar{\rho}} \rho, u, v, \frac{1}{\bar{c}\sqrt{\gamma(\gamma-1)}} T \right]^T. \tag{A.4}$$

[1] See [26] for more general conditions.

and note that there is a scaling coefficient in the temperature component. To simplify the analysis, we rescale (15) by multiplying the equation with $\frac{1}{\bar{c}\sqrt{\gamma(\gamma-1}}$. To apply the energy method, we rewrite the speed of sound based Péclet number Pe_c in (14) as

$$Pe_c = \frac{Pr \cdot Re}{Ma} = \frac{Pr}{\varepsilon} \tag{A.5}$$

where Pr is the Prandtl number. Then (15) becomes

$$\frac{\tilde{T}_t}{\bar{c}\sqrt{\gamma(\gamma-1)}} = \frac{\varepsilon\kappa_s}{Pr\bar{c}\sqrt{\gamma(\gamma-1)}\rho_s c_s}\left(\tilde{T}_{xx} + \tilde{T}_{yy}\right). \tag{A.6}$$

By applying the energy method to each equation and adding the results we obtain

$$\frac{d}{dt}\left(||w||^2_{J_1} + \frac{1}{\bar{c}^2\gamma(\gamma-1)}||\tilde{T}||^2_{\nu_2}\right) \leq \frac{-2\varepsilon}{\bar{c}^2\gamma(\gamma-1)Pr}\int_0^1 \left(\frac{\nu_1\gamma\mu}{\bar{\rho}}TT_y - \frac{\nu_2\kappa_s}{\rho_s c_s}\tilde{T}\tilde{T}_y\right)dx. \tag{A.7}$$

If we choose

$$\nu_1 = \frac{\bar{\kappa}\bar{\rho}}{\gamma\mu}, \quad \nu_2 = \rho_s c_s \tag{A.8}$$

and apply the interface conditions (A.1) we get

$$\frac{d}{dt}\left(||w||^2_{J_1} + \frac{1}{\bar{c}^2\gamma(\gamma-1)}||\tilde{T}||^2_{\nu_2}\right) \leq \frac{-2\varepsilon}{\bar{c}^2\gamma(\gamma-1)Pr}\int_0^1 T\left(\bar{\kappa}T_y - \kappa_s\tilde{T}_y\right)dx = 0 \tag{A.9}$$

and hence the conditions (A.1) does not contribute to unbounded energy growth. □

Note again that the application of the physical interface conditions (A.1) requires the use of a non-standard norm in the energy estimates. All quantities involved in the weights $\nu_{1,2}$ are, however, always positive and they will hence always define a norm.

The discretization of the coupled system is analogous to that which is presented in [4], and extended to multiple dimensions as described before. We hence only present the numerical scheme and the choice of interface penalty coefficients such that the scheme is stable.

An SBP-SAT discretization of the Navier-Stokes equations coupled with the heat

equation is given by, when only considering the interface terms,

$$\begin{aligned} \mathbf{w}_t + \left(\bar{D}_x \otimes I_4\right)\mathbf{F} + \left(\bar{D}_y \otimes I_4\right)\mathbf{G} &= \mathbb{S}, \\ \tilde{\mathbf{T}}_t - \left(\bar{D}_x^2\tilde{\mathbf{T}} + \bar{D}_y^2\tilde{\mathbf{T}}\right) &= \tilde{\mathbb{S}}. \end{aligned} \tag{A.10}$$

The penalty terms are given by

$$\begin{aligned} \mathbb{S} = &\left(\bar{P}_y^{-1}\bar{E}_{x,y_N} \otimes \bar{\Sigma}_1\right)\left(\mathbf{w} - g^I\right) \\ &+ \varepsilon\sigma_2\left(\bar{P}_y^{-1}\bar{E}_{x,y_N} \otimes I_4\right)\left(\bar{H}_2\mathbf{w} - g_1\right) \\ &+ \varepsilon\sigma_3\left(\bar{P}_y^{-1}\bar{E}_{x,y_N} \otimes I_4\right)\left(\bar{H}_3\mathbf{w} - g_2\right) \\ &+ \varepsilon\left(\bar{P}_y^{-1}\bar{E}_{x,y_N} \otimes I_4\right)\bar{\Theta}_1\left(\bar{H}_3\left(\bar{D}_x \otimes I_4\right)\mathbf{w} - \frac{\partial g_2}{\partial x}\right) \\ &+ \varepsilon\left(\bar{P}_y^{-1}\bar{E}_{x,y_N} \otimes \Sigma_4\right)\left(\bar{I}_1^T\mathbf{w} - \bar{I}_2^T(\tilde{\mathbf{T}} \otimes \mathbf{e_4})\right) \\ &+ \varepsilon\left(\bar{P}_y^{-1}\bar{D}_y^T\bar{E}_{x,y_N} \otimes \Sigma_5\right)\left(\bar{I}_1^T\mathbf{w} - \bar{I}_2^T(\tilde{\mathbf{T}} \otimes \mathbf{e_4})\right) \\ &+ \varepsilon\left(\bar{P}_y^{-1}\bar{E}_{x,y_N} \otimes \Sigma_6\right)\left(\bar{\kappa}\bar{I}_1^T\left(\bar{D}_y \otimes I_4\right)\mathbf{w} - \kappa_s\bar{I}_2^T(\bar{D}_y\tilde{\mathbf{T}} \otimes \mathbf{e_4})\right), \end{aligned} \tag{A.11}$$

where $\Sigma_{4,5,6} = \mathrm{diag}[0,0,0,\sigma_{4,5,6}]$ and the term involving $\bar{\Theta}_1$ is explained in Remark 3. The SAT for the heat equation is given by

$$\begin{aligned} \tilde{\mathbb{S}} = &\varepsilon\tau_4\bar{P}_y^{-1}\bar{E}_{x,y_N}\left(\tilde{\mathbf{T}} - \mathbf{T}\right) + \varepsilon\tau_5\bar{P}_y^{-1}\bar{D}_y^T\bar{E}_{x,y_N}\left(\tilde{\mathbf{T}} - \mathbf{T}\right) \\ &+ \varepsilon\tau_6\bar{P}_y^{-1}\bar{E}_{x,y_N}\left(\kappa_s\bar{D}_y\tilde{\mathbf{T}} - \bar{\kappa}\bar{D}_y\mathbf{T}\right) \end{aligned} \tag{A.12}$$

and the choices of penalty parameters such that the coupled scheme is stable is given in

Theorem 2. *The scheme (A.10) for coupling the Navier-Stokes equations with the heat equation is stable with the SATs given by (A.11), (A.12) where the penalty coefficients for the coupling terms are given by*

$$\begin{aligned} &r \in \mathbb{R}, \\ &\sigma_4 = \tau_4 \leq 0, \quad \sigma_5 = -\kappa_s r, \quad \sigma_6 = \frac{-1 + rPr}{Pr}, \quad \tau_5 = -\frac{\bar{\kappa}\left(-1 + rPr\right)}{Pr}, \quad \tau_6 = r. \end{aligned} \tag{A.13}$$

Proof. We apply the energy method, using the modified discrete norms,

$$||\mathbf{w}||^2_{J_1} = \mathbf{w}^T(\bar{P} \otimes J_1)w, \quad ||\tilde{\mathbf{T}}||^2_{\nu_2} = \nu_2 \tilde{\mathbf{T}}^T \bar{P} \mathbf{T}, \tag{A.14}$$

where $J_1 = \text{diag}[1, 1, 1, \nu_1]$ and $\nu_{1,2}$ are given in (A.8). Using appropriate penalty terms for the inviscid part and the velocity components of the Navier-Stokes equation, see [33, 26], we obtain the energy estimate

$$\frac{d}{dt}||\mathbf{w}||^2_{J_1} + \frac{d}{dt}||\tilde{\mathbf{T}}||^2_{\nu_2} \leq 0 \tag{A.15}$$

when using the penalty coefficients given in (A.13). □

References

[1] M. B. Giles. Stability analysis of numerical interface conditions in fluid-structure thermal analysis. *International Journal for Numerical Methods in Fluids*, 25:421–436, August 1997.

[2] B. Roe, R. Jaiman, A. Haselbacher, and P. H. Geubelle. Combined interface boundary condition method for coupled thermal simulations. *International Journal for Numerical Methods in Fluids*, 57:329–354, May 2008.

[3] William D. Henshaw and Kyle K. Chand. A composite grid solver for conjugate heat transfer in fluid-structure systems. *Journal of Computational Physics*, 228(10):3708–3741, 2009.

[4] Jens Lindström and Jan Nordström. A stable and high-order accurate conjugate heat transfer problem. *Journal of Computational Physics*, 229(14):5440–5456, 2010.

[5] Michael Schäfer and Ilka Teschauer. Numerical simulation of coupled fluid-solid problems. *Computer Methods in Applied Mechanics and Engineering*, 190(28):3645–3667, 2001.

[6] Niphon Wansophark, Atipong Malatip, and Pramote Dechaumphai. Streamline upwind finite element method for conjugate heat transfer problems. *Acta Mechanica Sinica*, 21:436–443, 2005.

[7] É. Turgeon, D. Pelletier, and F. Ilinca. Compressible heat transfer computations by an adaptive finite element method. *International Journal of Thermal Sciences*, 41(8):721–736, 2002.

[8] Xi Chen and Peng Han. A note on the solution of conjugate heat transfer problems using simple-like algorithms. *International Journal of Heat and Fluid Flow*, 21(4):463–467, 2000.

[9] Magnus Svärd, Ken Mattsson, and Jan Nordström. Steady-state computations using summation-by-parts operators. *Journal of Scientific Computing*, 24(1):79–95, 2005.

[10] Ken Mattsson, Magnus Svärd, Mark Carpenter, and Jan Nordström. High-order accurate computations for unsteady aerodynamics. *Computers and Fluids*, 36(3):636–649, 2007.

[11] X. Huan, Jason E. Hicken, and David W. Zingg. Interface and boundary schemes for high-order methods. In *the 39th AIAA Fluid Dynamics Conference, AIAA Paper No. 2009-3658*, San Antonio, USA, 22-25 June 2009.

[12] Jan Nordström, Jing Gong, Edwin van der Weide, and Magnus Svärd. A stable and conservative high order multi-block method for the compressible Navier-Stokes equations. *Journal of Computational Physics*, 228(24):9020–9035, 2009.

[13] Ken Mattsson, Magnus Svärd, and Mohammad Shoeybi. Stable and accurate schemes for the compressible Navier-Stokes equations. *Journal of Computational Physics*, 227:2293–2316, February 2008.

[14] Jan Nordström, Sofia Eriksson, Craig Law, and Jing Gong. Shock and vortex calculations using a very high order accurate Euler and Navier-Stokes solver. *Journal of Mechanics and MEMS*, 1(1):19–26, 2009.

[15] Jan Nordström, Frank Ham, Mohammad Shoeybi, Edwin van der Weide, Magnus Svärd, Ken Mattsson, Gianluca Iaccarino, and Jing Gong. A hybrid method for unsteady inviscid fluid flow. *Computers & Fluids*, 38:875–882, 2009.

[16] Heinz-Otto Kreiss and Godela Scherer. Finite element and finite difference methods for hyperbolic partial differential equations. In *Mathematical Aspects of Finite Elements in Partial Differential Equations*, number 33 in Publ. Math. Res. Center Univ. Wisconsin, pages 195–212. Academic Press, 1974.

[17] Heinz-Otto Kreiss and Godela Scherer. On the existence of energy estimates for difference approximations for hyperbolic systems. Technical report, Uppsala University, Division of Scientific Computing, 1977.

[18] 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.

[19] Mark H. Carpenter, Jan Nordström, and David Gottlieb. A stable and conservative interface treatment of arbitrary spatial accuracy. *Journal of Computational Physics*, 148(2):341–365, 1999.

[20] Jan Nordström and Magnus Svärd. Well-posed boundary conditions for the Navier-Stokes equations. *SIAM Journal on Numerical Analysis*, 43(3):1231–1255, 2005.

[21] Bo Strand. Summation by parts for finite difference approximations for d/dx. *Journal of Computational Physics*, 110(1):47–67, 1994.

[22] Ken Mattsson and Jan Nordström. Summation by parts operators for finite difference approximations of second derivatives. *Journal of Computational Physics*, 199(2):503–540, 2004.

[23] Ken Mattsson. Summation by parts operators for finite difference approximations of second-derivatives with variable coefficients. *Journal of Scientific Computing*, pages 1–33, 2011.

[24] Jorg U. Schlüter, Xiaohua Wu, Edwin van der Weide, S. Hahn, Juan J. Alonso, and Heinz Pitsch. Multi-code simulations: A generalized coupling approach. In *the 17th AIAA CFD Conference, AIAA-2005-4997*, Toronto, Canada, June 2005.

[25] Magnus Svärd and Jan Nordström. On the order of accuracy for difference approximations of initial-boundary value problems. *Journal of Computational Physics*, 218(1):333–352, 2006.

[26] Jens Berg and Jan Nordström. Stable Robin solid wall boundary conditions for the Navier-Stokes equations. *Journal of Computational Physics*, 230(19):7519–7532, 2011.

[27] Saul Abarbanel and David Gottlieb. Optimal time splitting for two- and three-dimensional Navier-Stokes equations with mixed derivatives. *Journal of Computational Physics*, 41(1):1–33, 1981.

[28] Magnus Svärd, Mark H. Carpenter, and Jan Nordström. A stable high-order finite difference scheme for the compressible Navier-Stokes equations, far-field boundary conditions. *Journal of Computational Physics*, 225(1):1020–1038, 2007.

[29] Heinz-Otto Kreiss and Jens Lorenz. *Initial-Boundary Value Problems and the Navier-Stokes Equations*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 2004.

[30] Ken Mattsson and Mark H. Carpenter. Stable and accurate interpolation operators for high-order multiblock finite difference methods. *SIAM Journal on Scientific Computing*, 32(4):2298–2320, 2010.

[31] Lee Shunn, Frank Ham, and Parviz Moin. Verification of variable-density flow solvers using manufactured solutions. *Journal of Computational Physics*, 231(9):3801–3827, 2012.

[32] Bertil Gustafsson, Heinz-Otto Kreiss, and Joseph Oliger. *Time Dependent Problems and Difference Methods*. Wiley Interscience, 1995.

[33] Magnus Svärd and Jan Nordström. A stable high-order finite difference scheme for the compressible Navier-Stokes equations: No-slip wall boundary conditions. *Journal of Computational Physics*, 227(10):4805–4824, 2008.

Paper V

Journal of Computational Physics 231 (2012) 6846–6860

Contents lists available at SciVerse ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp

Superconvergent functional output for time-dependent problems using finite differences on summation-by-parts form

Jens Berg [a,*], Jan Nordström [b]

[a] Uppsala University, Department of Information Technology, SE-751 05 Uppsala, Sweden
[b] Linköping University, Department of Mathematics, SE-581 83 Linköping, Sweden

ARTICLE INFO

Article history:
Received 13 February 2012
Received in revised form 14 June 2012
Accepted 19 June 2012
Available online 6 July 2012

Keywords:
High order finite differences
Summation-by-parts
Superconvergence
Time-dependent functional output
Dual consistency
Stability

ABSTRACT

Finite difference operators satisfying the summation-by-parts (SBP) rules can be used to obtain high order accurate, energy stable schemes for time-dependent partial differential equations, when the boundary conditions are imposed weakly by the simultaneous approximation term (SAT).

In general, an SBP-SAT discretization is accurate of order $p+1$ with an internal accuracy of $2p$ and a boundary accuracy of p. Despite this, it is shown in this paper that any linear functional computed from the time-dependent solution, will be accurate of order $2p$ when the boundary terms are imposed in a stable and dual consistent way.

The method does not involve the solution of the dual equations, and superconvergent functionals are obtained at no extra computational cost. Four representative model problems are analyzed in terms of convergence and errors, and it is shown in a systematic way how to derive schemes which gives superconvergent functional outputs.

1. Introduction

When numerically computing solutions to equations in computational fluid dynamics (CFD), accurate solutions to the equations themselves might not be the primary target. Typically, functionals computed from the solution, such as the lift and drag coefficients, are of equal or even larger interest.

Already in the late 1990s, Giles et al. realized the importance of duality to enhance the computation of functionals in CFD applications [1–6]. Since then, duality and adjoint equations have been vastly studied in the context of finite element methods (FEM) [2] and more recently using discontinuous Galerkin (DG) methods [7–10], finite volume methods (FVM) [11] and spectral difference methods [12].

One can separate three distinct uses of the adjoint equations; adaptive mesh refinement [13], error analysis [14] and optimal design problems [15,16]. The success of duality based approaches to, in particular, adaptive mesh refinement and error estimation, has made the study of duality somewhat restricted to unstructured methods such as FEM, DG and FVM.

Recently, however, it was shown by Hicken and Zingg [17,18] that the adjoint equations can be used for finite difference (FD) methods to raise the order of accuracy of linear functionals computed from the FD solution. The technique was based on using FD operators on summation-by-parts (SBP) form [19,20] together with the simultaneous approximation term (SAT) for imposing boundary conditions weakly [21]. It was shown that when discretizing the equations in a dual consistent [9,17] way, the order of accuracy of the output functional was higher than the FD solution itself. This superconvergent behaviour

* Corresponding author. Address: Division of Scientific Computing, Department of Information Technology, Uppsala University, Box 337, SE-751 05 Uppsala, Sweden. Tel.: +46 18 471 6253; fax: +46 18 523049/511925.
E-mail address: jens.berg@it.uu.se (J. Berg).

0021-9991/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jcp.2012.06.032

was seen already in [3] for FEM and in [7] for DG, but it had not been previously proven for finite difference schemes. Some work on solution superconvergence for FD-based methods, using mimetic operators, can be seen in i.e. [22].

So far, most applications of the adjoint equations deal with steady-state problems, including the recent results presented in [17]. The reason is that the adjoint equation has limited use for realistic (non-linear) time-dependent problems since it runs backwards in time [23]. Hence to actually solve the adjoint time-dependent equation, the full time history of the primal equation has to be stored [24]. For large scale problems, this quickly becomes unfeasible [25,26]. Some work has been done in the time-dependent setting [25,23], in particular for adaptive error control [24,11,27] and optimization [26,12].

What is to be presented in this paper is the extension of [17] to unsteady problems for computing superconvergent time-dependent linear functionals. By superconvergence, we mean that the order of convergence of the output functional is higher than the design order of accuracy of the scheme. We will address two problems which usually occurs when attempting to use duality for time-dependent functional computations;

- The discrete adjoint equations does not approximate the continuous adjoint equations, i.e. the scheme is dual inconsistent
- If the scheme is dual consistent, it is unstable

The SBP discretization together with the SAT technique is highly suitable for addressing the above issues since the scheme allows for a multitude of parameters which can be chosen such that the scheme is both dual consistent and stable. These two features will result in a superconvergent time-dependent functional output.

2. SBP-SAT discretizations

Summation-by-parts finite difference operators were originally constructed by Kreiss and Scherer [28] in the 70's as a means for constructing energy stable [29] finite difference approximations. The operators are constructed such that they are automatically stable for linearly well-posed Cauchy problems. Together with the SAT procedure introduced by Carpenter et al. [21], the SBP-SAT technique provides a method of constructing energy stable and high order accurate finite difference schemes for any linearly well-posed initial-boundary value problem. Since then, the technique has been widely used and proven robust for a variety of problem. See for example [30–37] and references therein.

The SBP operators can be defined as follows.

Definition 1. A matrix D is called a first derivative *SBP operator* if D can be written as

$$D = P^{-1}Q, \tag{1}$$

where P defines a norm by $||u||^2 = u^T P u$ and Q satisfies

$$Q + Q^T = \text{diag}[-1, 0, \ldots, 0, 1]. \tag{2}$$

In this paper, only diagonal matrices P will be used. In that case, D consist of a $2p$-order accurate central difference approximation in the interior while at the boundaries, the accuracy reduces to a p-order one-sided difference. The global accuracy can then be shown to be $p + 1$ [32].

By using non-diagonal matrices P as norms in the SBP definition, it is possible to raise both the boundary and global order of accuracy. For a block-diagonal P, the boundary stencil can be chosen to be $2p - 1$ order accurate which increases the global accuracy to $2p$ [19,32,38,39]. There are, however, drawbacks with a non-diagonal matrix P. In many cases, the equations are non-linear or have variable coefficients and energy stability can only be proven if P commutes with diagonal matrices. Unless P is carefully constructed to fit each problem under consideration, a diagonal P is the only alternative.

For many realistic problems, the boundary of the domain is non-smooth and the domain has to be split into blocks, where a curvilinear coordinate transformation is applied in each block. If the matrix P is not diagonal, energy stability cannot be shown in general since P is required to commute with the (diagonal) Jacobian matrix of the coordinate transformation [35,40–42].

When computing linear functionals, however, we can recover the loss compared to the accuracy from a non-diagonal P, while keeping the simplicity and flexibility of a diagonal P. It is hence always possible to prove energy stability, and keeping the full order of accuracy.

Currently there exist diagonal norm SBP operators for the first derivative accurate of order 2, 3, 4 and 5. The second derivative can be approximated using either the first derivative twice which results in a wide finite difference stencil, or a compact operator as described in [20,43]. In this paper, we will rewrite the equations in a form which does not require the application of a second derivative operator.

A first order hyperbolic PDE, for example the advection equation on $0 \leqslant x \leqslant 1$,

$$\begin{aligned} &u_t + a u_x = 0, \\ &u(0,t) = d_1(t), \\ &u(x,0) = d_2(x), \end{aligned} \tag{3}$$

with $a > 0$, can be approximated on an equidistant grid with $N + 1$ gridpoints as

$$\frac{d}{dt}u_h + aDu_h = 0, \tag{4}$$

where u_h is the discrete gridfunction approximating u. However, since the continuous PDE (3) needs to be supplied with a boundary condition at the inflow boundary, the scheme (4) has to be modified. The imposition of the boundary condition is done weakly using SAT as

$$\frac{d}{dt}u_h + aDu_h = \sigma P^{-1}(e_0^T u_h - d_1)e_0, \tag{5}$$

where $e_0 = [1, 0, \ldots, 0]$ and $d_1 = d_1(t)$ is the time-dependent boundary data. The coefficient σ is a parameter which has to be determined such that the scheme is stable in the P-norm.

2.1. The energy method

To prove well-posedness of the continuous Eq. (3) and stability of the numerical scheme (5), the energy metod in continuous and discrete form is used. We multiply (3) with u and integrate by parts over the spatial domain to obtain (when assuming $d_1 = 0$)

$$||u||_t^2 = -au^2(1,t). \tag{6}$$

It is clear that the growth rate of energy is bounded and hence we say that (3) is well-posed.[1]

In the discrete case we multiply (5) with $u_h^T P$ and use the SBP properties of the operator to obtain

$$||u_h||_t^2 = (a + 2\sigma)u_h^2(x_0) - au_h^2(x_N). \tag{7}$$

It is clear that an energy estimate is obtained for

$$\sigma \leqslant -\frac{a}{2} \tag{8}$$

and for $\sigma = -\frac{a}{2}$ we have exactly (6).

We can see that the parameter σ is allowed to vary in a semi-infinite range for which the scheme is stable. Any additional requirement we place on the scheme, for example dual consistency, has to be within a subset of values allowed by the energy estimate. This flexibility together with the ability to mimic integration by parts is what makes the SBP-SAT method suitable for treating adjoint problems.

Remark 2.1. Note that the assumption $d_1 = 0$ merely simplifies the analysis. Boundary and initial data can be included, in which case the problem is called strongly well-posed. If the boundary and initial data is included in the discrete case, and an energy estimate is obtained, the problem is called strongly stable [45].

3. Adjoint problems and dual consistency

There are various ways of obtaining the adjoint equations. Most common is to consider a PDE subject to a set of control parameters and a functional output of interest, and in various ways taking derivatives of the functional with respect to the control parameters [1,27]. The adjoint equation can then be seen as a sensitivity equation for the primal PDE, and is sometimes referred to as the sensitivity equation. In this work we will adopt the notation in [17] and derive the adjoint equation by posing the SBP-SAT method in a variational framework similar to the one used in FEM.

The order of convergence is measured in space, not in time. To obtain a superconvergent time-dependent linear functional output, it is sufficient to consider the steady equations and discretize them in a dual consistent way which does not violate any stability conditions for the unsteady equations.

We shall use the following notations regarding the inner products. The continuous inner product is defined as

$$(f, g) = \int_\Omega fg d\Omega \tag{9}$$

and the corresponding discrete inner product is defined as

$$(f_h, g_h)_h = f_h^T P g_h, \tag{10}$$

where f_h, g_h are projections of f, g onto a grid, and P is the matrix (and integration operator) used to define a norm in the definition of the SBP operator. The subscript h will be omitted for known functions if the meaning is clear from the context.

Before we begin, we need to define what is meant by the continuous dual problem, discrete dual problem and dual consistency. Let L be a linear differential operator and consider the (steady) equation

[1] Existence of solutions is not formally considered in this context. Existence is motivated by the fact that a minimal number of boundary conditions is used to obtain an energy estimate [44].

$$Lu - f = 0, \quad \forall x \in \Omega, \tag{11}$$

subject to homogeneous boundary conditions. Let

$$J(u) = (g, u) \tag{12}$$

be a linear functional output of interest. We obtain the adjoint equation by seeking ϕ in some appropriate function space, such that

$$J(u) = (\phi, f). \tag{13}$$

A formal computation gives

$$J(u) = (g, u) - (\phi, Lu - f) = (\phi, f) - (L^*\phi - g, u) \tag{14}$$

and hence the adjoint equation is given by

$$L^*\phi - g = 0, \tag{15}$$

where L^* is the formal adjoint of L. Note that L^* is abstractly defined, and finding an exact expression for the dual operator is in general a non-trivial task. In the case of linear differential operators, the adjoint operator is obtained by integration by parts.

Remark 3.1. In this paper, we consider homogeneous boundary and initial conditions. This is only for the purpose of analysis. The dual problem depends only on the form of the boundary conditions, but not on the particular boundary or initial data. In computations, the boundary and initial data can be non-zero.

The boundary conditions for the adjoint equation are obtained by considering the boundary terms resulting from the integration by parts procedure. After applying the homogeneous boundary conditions for the primal PDE, the dual boundary conditions are defined as the minimal set of homogeneous conditions such that all boundary terms vanish.

Definition 2. The *continuous dual problem* is given by

$$L^*\phi = g \tag{16}$$

subject to the dual boundary conditions.

The same reasoning can be applied in the discrete setting. Let

$$L_h u_h - f = 0 \tag{17}$$

be a discretization of (11), including the homogeneous boundary conditions. Then

$$J_h(u_h) = (g, u_h)_h \tag{18}$$

is an approximation of (12). We obtain the discrete dual problem by seeking ϕ_h such that

$$J_h(u_h) = (\phi_h, f)_h. \tag{19}$$

The same formal computation as before gives

$$J_h(u_h) = (g, u_h)_h - (\phi_h, L_h u_h - f)_h = (\phi_h, f)_h - (P^{-1}L_h^T P\phi_h - g, u_h)_h \tag{20}$$

and we have

Definition 3. The *discrete dual problem* is given by

$$P^{-1}L_h^T P\phi_h - g = 0. \tag{21}$$

Remark 3.2. In an SBP-SAT setting, the difference operator L_h can be written as

$$L_h = P^{-1}\widetilde{L}_h \tag{22}$$

and the discrete dual problem reduces to

$$P^{-1}\widetilde{L}_h^T\phi_h = g. \tag{23}$$

Finally, by using (16) and (21) we make the definition of dual consistency.

Definition 4. A discretization is called *dual consistent* if (21) is a consistent approximation of (16).

So far, we have been concerned with steady problems only. Since we are interested in unsteady problems, we need to define what is meant by dual consistency in this context. Consider an unsteady problem

$$u_t + Lu - f = 0, \quad t > 0, \ \forall x \in \Omega, \tag{24}$$

subject to homogeneous boundary and initial conditions. By seeking ϕ such that

$$\int_0^T J(u)dt = \int_0^T (\phi, f)dt \tag{25}$$

we obtain

$$\begin{aligned} &\int_0^T J(u)dt = \int_0^T J(u)dt - \int_0^T (\phi, u_t + Lu - f)dt \\ &\int_0^T (\phi_t - L^*\phi + g, u)dt + \int_0^T (\phi, f)dt. \end{aligned} \tag{26}$$

The time-dependent dual problem thus becomes

$$-\phi_t + L^*\phi = g \tag{27}$$

subject to the dual boundary conditions. A homogeneous initial condition for the dual problem is placed at time $t = T$ which removes the boundary term from the partial time integration.

The discrete procedure can be formulated analogously. Let

$$\frac{d}{dt}u_h + L_h u_h - f = 0 \tag{28}$$

be a semi-discretization of (24), including the boundary conditions. We then have the following definition regarding dual consistency of time-dependent problems,

Definition 5. The semi-discretization (28) is called *spatially dual consistent* if the corresponding steady problem is dual consistent.

Note that a stable and consistent discretization of the primal PDE does not imply spatial dual consistency.

To prove the main result of this paper, we need Corollary 1 from [46], which states that P is a $2p$-order accurate quadrature. For our purpose, we can restate the result as.

Lemma 3.1. *Let P be the norm-matrix of an SBP discretization with 2p-order internal accuracy. Then for $u \in C^{2p}$ we have*

$$J_h(u) = J(u) + O(h^{2p}). \tag{29}$$

Using Lemma 3.1 we can prove the main result of this paper which is.

Theorem 3.2. *Let*

$$\frac{d}{dt}u_h + L_h u_h = f \tag{30}$$

be a stable and spatial dual consistent SBP-SAT discretization of the continuous problem

$$u_t + Lu = f. \tag{31}$$

Then the linear functional

$$J_h(u_h) = g^T P u_h \tag{32}$$

is a 2p-order accurate approximation of

$$J(u) = \int_\Omega g^T u d\Omega. \tag{33}$$

Proof. By using the results in [46] together with the definition of the discrete dual problem, we can add and subtract terms to relate the the continuous functional to the discrete as

$$\begin{aligned} J(u) &= J_h(u) + O(h^{2p}) = g^T P u_h + g^T P(u - u_h) + O(h^{2p}) = g^T P u_h + g^T P(u - u_h) - \phi_h^T P(L_h u_h - f) + O(h^{2p}) \\ &= J_h(u_h) + g^T P(u - u_h) - \phi_h^T P L_h(u - u_h) - \phi^T P f + \phi_h^T P L_h u + O(h^{2p}) \\ &= J_h(u_h) - (u - u_h)^T P(P^{-1}L_h^T P\phi_h - g) + \phi^T P(L_h u - f) + O(h^{2p}) = J_h(u_h) + \phi^T P(L_h u - f) + O(h^{2p}), \end{aligned} \tag{34}$$

where the last error term is of order h^{2p} [46]. We can hence conclude that

$$J(u) = J_h(u_h) + O(h^{2p}). \qquad \square \tag{35}$$

4. Derivation of stable and spatially dual consistent schemes

Based on Theorem 3.2, we will derive stable and spatially dual consistent schemes for four time-depedent model problems in a systematic way. The equations we consider are the advection equation, the heat equation, the viscous Burgers' equation and an incompletely parabolic system of equations. We will see that a stable and spatial dual consistent discretization produces superconvergent time-dependent linear functionals.

4.1. The advection equation

Consider (3) again together with a linear functional output of interest. We let the boundary condition be homogeneous, add a forcing function and ignore the initial condition,

$$\begin{aligned} &u_t + au_x = f, \\ &u(0,t) = 0, \\ &J(u) = (g,u). \end{aligned} \tag{36}$$

Note that $J(u)$ is a time-dependent functional. The adjoint equation is obtained by letting $u_t = 0$ and finding ϕ such that $J(u) = (\phi, f)$. We get

$$J(u) = J(u) - \int_0^1 \phi(au_x - f)dx = -a\phi(1,t)u(1,t) - \int_0^1 (g + a\phi_x)udx + (v,f) \tag{37}$$

and hence the steady adjoint problem is given by

$$\begin{aligned} &-a\phi_x = g, \\ &\phi(1,t) = 0. \end{aligned} \tag{38}$$

Note that the sign has changed and the adjoint boundary condition is located at the opposite boundary compared to the primal problem.

Eq. (36) is discretized as before,

$$\frac{d}{dt}u_h + aP^{-1}Qu_h = f + \sigma P^{-1}(e_0^T u_h - \mathbf{0})e_0, \tag{39}$$

where $\mathbf{0}$ is the boundary data. We know from the preceding energy estimate (7) that the scheme is stable if $\sigma \leqslant -\frac{a}{2}$. The addition of the forcing function does not change the number or form of the boundary conditions and can be assumed to be zero in an energy estimate according to the principle of Duhamel [45]. To determine spatial dual consistency, we let $\frac{d}{dt}u_h = 0$ and rewrite (39) as

$$L_h u_h = Pf, \tag{40}$$

where

$$L_h = aQ - \sigma E_0 \tag{41}$$

and $E_0 = e_0^T e_0 = \mathrm{diag}[1,0,\ldots,0]$. According to the definition of dual consistency,

$$L_h^T \phi_h = Pg \tag{42}$$

has to be a consistent approximation of the adjoint Eq. (38). By using the SBP property of Q, we expand (42) as

$$-aP^{-1}Q\phi_h = g - aP^{-1}E_N\phi_h + (\sigma + a)P^{-1}E_0\phi_h \tag{43}$$

which is a consistent approximation of (38) only if

$$\sigma = -a. \tag{44}$$

For any other value of σ, the numerical scheme would impose a boundary condition at $x = 0$ which does not exist in the adjoint equation. We can also see that $\sigma = -a$ does not violate the stability condition given by the energy estimate. Thus the scheme is both stable and spatially dual consistent.

Remark 4.1. Note that the parameter σ is allowed to vary in a semi-infinite range from the stability requirements, while spatial dual consistency requires a unique value.

4.2. The heat equation

The heat equation on $0 \leqslant x \leqslant 1$ with homogeneous Dirichlet boundary conditions is given by

$$\begin{aligned} &u_t = \alpha u_{xx} + f, \\ &u(0,t) = 0, \\ &u(1,t) = 0, \\ &J(u) = (g,u). \end{aligned} \tag{45}$$

The initial condition is omitted since the derivation of the dual problem depends only on the equation and the form of the boundary conditions. In the computations, however, an initial condition has to be supplied. In order to derive a stable and spatially dual consistent scheme, (45) has to be rewritten as a first order system in the same way as in the local discontinuous Galerkin (LDG) method [47]. It has been shown that the LDG method has interesting superconvergent features not only for functionals, but also for the solution itself [7,30,48]. We hence adapt the LDG formulation and rewrite (45) as

$$\begin{aligned} &u_t = \sqrt{\alpha} v_x + f, \\ &v = \sqrt{\alpha} u_x, \\ &u(0,t) = 0, \\ &u(1,t) = 0, \\ &J(u) = (g,u). \end{aligned} \tag{46}$$

To obtain the dual problem, we let $u_t = 0$ and write (46) as

$$Aw + Bw_x = F, \tag{47}$$

where $w = [u, v]^T$, $F = [f, 0]^T$ and

$$A = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & -\sqrt{\alpha} \\ -\sqrt{\alpha} & 0 \end{bmatrix}. \tag{48}$$

Let now $G = [g, 0]^T$, $\theta = [\phi, \psi]^T$ and find θ such that

$$J(w) = (\theta, F). \tag{49}$$

Note that

$$J(w) = (G, w) = (g, u) \tag{50}$$

and we are still computing the functional of interest from the primal problem. This gives us the adjoint problem by computing

$$J(w) = J(w) - \int_0^1 \theta^T (Aw + Bw_x - F) dx = \int_0^1 w^T (G - A\theta + B\theta_x) dx - \left[\theta^T B w\right]_0^1 + (\theta, F). \tag{51}$$

The adjoint equation is thus given by

$$A\theta - B\theta_x = G \tag{52}$$

and the adjoint boundary conditions are the minimal number of conditions such that $\left[\theta^T B w\right]_0^1 = 0$. After applying the homogeneous boundary conditions for the primal problem, we get the adjoint problem on component form

$$\begin{aligned} &\sqrt{\alpha}\psi_x = g, \\ &\psi + \sqrt{\alpha}\phi_x = 0, \\ &\phi(0,t) = 0, \\ &\phi(1,t) = 0. \end{aligned} \tag{53}$$

The primal PDE on LDG form (46) is discretized as

$$\begin{aligned} &\frac{d}{dt} u_h = \sqrt{\alpha} P^{-1} Q v_h + f + \sigma_L P^{-1} (e_0^T u_h - \mathbf{0}) e_0 + \sigma_R P^{-1} (e_N^T u_h - \mathbf{0}) e_N, \\ &v_h = \sqrt{\alpha} P^{-1} Q u_h + \tau_L P^{-1} (e_0^T u_h - \mathbf{0}) e_0 + \tau_R P^{-1} (e_N^T u_h - \mathbf{0}) e_N. \end{aligned} \tag{54}$$

By multiplying the first equation by $u_h^T P$, the second by $v_h^T P$ and adding the results we get

$$\frac{1}{2}\frac{d}{dt}||u_h||^2 + ||v_h||^2 = (\tau_L - \sqrt{\alpha}) u_h^T E_0 v_h + (\tau_R + \sqrt{\alpha}) u_h^T E_N v_h + \sigma_L u_h^T E_0 u_h + \sigma_R u_h^T E_N u_h \tag{55}$$

and the scheme is clearly stable if

$$\tau_L = \sqrt{\alpha}, \quad \tau_R = -\sqrt{\alpha}, \quad \sigma_L \leqslant 0, \quad \sigma_R \leqslant 0. \tag{56}$$

To determine spatial dual consistency we again let $u_t = 0$ and rewrite (54), using (56), as

$$L_h w_h = \widetilde{F}, \tag{57}$$

where $w_h = [u_h, v_h]^T$, $\widetilde{F} = [Pf, 0]^T$ and

$$L_h = \begin{bmatrix} -\sigma_L E_0 - \sigma_R E_N & -\sqrt{\alpha}Q \\ -\sqrt{\alpha}Q - \sqrt{\alpha}E_0 + \sqrt{\alpha}E_N & P \end{bmatrix}. \tag{58}$$

The discrete dual problem is given by

$$L_h^T \theta_h = \widetilde{G}, \tag{59}$$

where $\theta_h = [\phi_h, \psi_h]^T$, $\widetilde{G} = [Pg, 0]^T$, and it has to be a consistent approximation of (53) without violating the stability conditions (56). By using the SBP properties of the operators we expand (59) and write it in component form as

$$\begin{aligned} &\sqrt{\alpha}P^{-1}Q\psi_h = g + \sigma_L P^{-1}E_0\phi_h + \sigma_R P^{-1}E_N\phi_h \\ &\psi_h + \sqrt{\alpha}P^{-1}Q\phi_h = -\sqrt{\alpha}P^{-1}E_0\phi_h + \sqrt{\alpha}P^{-1}E_N\phi_h \end{aligned} \tag{60}$$

which exactly approximates (53), including the dual boundary conditions. Note that there are no restrictions on $\sigma_{L,R}$ for dual consistency.

Remark 4.2. Note that the stability requirements are sufficient for spatial dual consistency, in contrast to the pure advection case.

Remark 4.3. The LDG form can be transformed back to second order form, see also [30], in which case the scheme becomes

$$\frac{d}{dt}u_h = \alpha(P^{-1}Q)^2 u_h + f + (\sigma_L I + \alpha P^{-1}Q)P^{-1}(e_0^T u_h - \mathbf{0})e_0 + (\sigma_R I - \alpha P^{-1}Q)P^{-1}(e_N^T u_h - \mathbf{0})e_N, \tag{61}$$

where I is the identity matrix of size $N+1$. Note that we get back the wide second derivative operator, possibly suggesting that dual consistency requires a second derivative operator which can be factorized into the product of two first derivative operators.

4.3. The viscous Burgers' equation

The viscous Burgers' equation, together with a linear functional of interest, with homogeneous Dirichlet boundary conditions on $0 \leqslant x \leqslant 1$ is given on conservative form by

$$\begin{aligned} &u_t + \left(\frac{u^2}{2}\right)_x = \varepsilon u_{xx} + f, \\ &u(0,t) = 0, \\ &u(1,t) = 0, \\ &J(u) = (g,u). \end{aligned} \tag{62}$$

Since (62) is a non-linear equation, the present theory cannot directly be applied. The viscous Burger's equation have regular solutions due to the viscosity term, and the behavior of the solution is not far from that of a linear problem. In the absence of a general method for non-linear analysis, a linear analysis is used. In the presence of shocks, for more complicated equations, it is not clear what meaning a linear analysis have.

We linearize (62) around a constant state $u = a$ to obtain the linear equation,

$$\begin{aligned} &u_t + au_x = \varepsilon u_{xx} + f, \\ &u(0,t) = 0, \\ &u(1,t) = 0, \\ &J(u) = (g,u), \end{aligned} \tag{63}$$

which is usually referred to as the advection–diffusion equation.

Since (62) contains second derivatives, we introduce the auxiliary variable $v = \sqrt{\varepsilon}u_x$ and rewrite the steady (linear) problem as

$$Aw + Bw_x = F, \tag{64}$$

where $w = [u, v]^T$, $F = [f, 0]^T$ and

$$A = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} a & -\sqrt{\varepsilon} \\ -\sqrt{\varepsilon} & 0 \end{bmatrix}. \tag{65}$$

To find the adjoint equation, we define $G = [g, 0]^T$ and seek $\theta = [\phi, \psi]^T$ such that

$$J(w) = (\theta, F) \tag{66}$$

as before. Integration by parts leads to

$$J(w) = J(w) - \int_0^1 \theta^T (Aw + Bw_x - F)dx = \int_0^1 w^T (G - A\theta + B\theta_x)dx - \left[\theta^T B w\right]_0^1 + (\theta, F) \tag{67}$$

and hence the adjoint equation is given on component form as

$$\begin{aligned} &-a\phi_x + \sqrt{\varepsilon}\psi_x = g, \\ &\phi + \sqrt{\varepsilon}\phi_x = 0, \\ &\phi(0,t) = 0, \\ &\phi(1,t) = 0. \end{aligned} \tag{68}$$

The stability analysis will also be performed on the linearized equations. The time-dependent equation on LDG form is discretized as

$$\begin{aligned} &\frac{d}{dt}u_h + aP^{-1}Qu_h = \sqrt{\varepsilon}P^{-1}Q\,v_h + \sigma_L P^{-1}(e_0^T u_h - \mathbf{0})e_0 + \sigma_R P^{-1}(e_N^T u_h - \mathbf{0})e_N + f, \\ &v_h = \sqrt{\varepsilon}P^{-1}Qu_h + \tau_L P^{-1}(e_0^T u_h - \mathbf{0})e_0 + \tau_R P^{-1}(e_N^T u_h - \mathbf{0})e_N \end{aligned} \tag{69}$$

and the coefficients $\sigma_{L,R}$ and $\tau_{L,R}$ has to be determined such that the scheme is stable. By multiplying the first equation in (69) by $u_h^T P$ and the second by $v_h^T P$, we obtain by adding the results

$$\frac{d}{dt}||u_h||^2 + 2||v_h||^2 = (2\sigma_L + a)u_h^T E_0 u_h + (2\sigma_R - a)u_h^T E_N u_h + 2(\tau_L - \sqrt{\varepsilon})v_h^T E_0 u_h + 2(\tau_R + \sqrt{\varepsilon})v_h^T E_N u_h. \tag{70}$$

We can see that (70) is stable if we chose

$$\sigma_L \leqslant -\frac{a}{2}, \quad \sigma_R \leqslant \frac{a}{2}, \quad \tau_L = \sqrt{\varepsilon}, \quad \tau_R = -\sqrt{\varepsilon}. \tag{71}$$

To determine if the scheme is spatially dual consistent, we let $u_t = 0$ and rewrite (69), using (71), as

$$L_h w_h = \widetilde{F}, \tag{72}$$

where $w_h = [u_h, v_h]^T$, $\widetilde{F} = [Pf, 0]^T$ and

$$L_h = \begin{bmatrix} aQ + \sigma_L E_0 + \sigma_R E_N & -\sqrt{\varepsilon} \\ -\sqrt{\varepsilon}Q - \sqrt{\varepsilon}E_0 + \sqrt{\varepsilon}E_N & P \end{bmatrix}. \tag{73}$$

The discrete dual problem is then given by

$$L_h^T \theta_h = \widetilde{G}, \tag{74}$$

where $\theta_h = [\phi_h, \psi_h]^T$ and $\widetilde{G} = [Pg, 0]^T$, which has to be a consistent approximation of (68) without violating the stability conditions (71). By expanding (74), we can write it in component form as

$$\begin{aligned} &-aP^{-1}Q\phi_h + \sqrt{\varepsilon}P^{-1}Q\psi_h = -(\sigma_L - a)P^{-1}E_0\phi_h - (\sigma_R + a)P^{-1}E_N\phi_h + g \\ &\psi_h + \sqrt{\varepsilon}P^{-1}Q\phi_h = -\sqrt{\varepsilon}P^{-1}E_0\phi_h + \sqrt{\varepsilon}P^{-1}E_N\phi_h \end{aligned} \tag{75}$$

which can be seen to be a consistent approximation of (74) without violating any of the stability conditions in (71). Hence the scheme (69) is both a stable and spatially dual consistent approximation of the linearized equation.

When performing the computations, however, we use the nonlinear LDG formulation

$$\begin{aligned} &\frac{d}{dt}u_h + P^{-1}Q\left(\frac{u_h^2}{2}\right) = \sqrt{\varepsilon}P^{-1}Q\,v_h + \sigma_L P^{-1}(e_0^T u_h - \mathbf{0})e_0 + \sigma_R P^{-1}(e_N^T u_h - \mathbf{0})e_N + f, \\ &v_h = \sqrt{\varepsilon}P^{-1}Qu_h + \tau_L P^{-1}(e_0^T u_h - \mathbf{0})e_0 + \tau_R P^{-1}(e_N^T u_h - \mathbf{0})e_N, \end{aligned} \tag{76}$$

where every occurence of the mean flow coefficient, a, in the SAT is replaced by u to obtain a nonlinear SAT. This procedure is motivated by the linearization and localization principle, see [49] for details.

Remark 4.4. Note again that stability is sufficient for spatial dual consistency and no extra conditions have to be placed on the SAT coefficients. The coefficients $\sigma_{L,R}$ are still allowed to vary in a semi-infinite range.

4.4. An incompletely parabolic system

In this section we consider the incompletely parabolic system

$$\begin{aligned} &U_t + AU_x = BU_{xx} + F, \\ &J(U) = (G, U), \end{aligned} \tag{77}$$

where $U = [p, u], F = [f_1, f_2]^T, G = [g_1, g_2]^T$ and

$$A = \begin{bmatrix} \bar{u} & \bar{c} \\ \bar{c} & \bar{u} \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 0 & \varepsilon \end{bmatrix}. \tag{78}$$

Eq. (77) can be thought of as the symmetrized [50] Navier–Stokes equations linearized around the mean velocity $\bar{u} > 0$ and speed of sound $\bar{c}$. We shall assume a linearization around a subsonic flow field, that is $\bar{u} < \bar{c}$. In this case, (77) requires two boundary conditions at the inflow boundary and one at the outflow. For the purpose of analysis, we will use the homogeneous Dirichlet conditions

$$u(0,t) = 0, \quad p(0,t) = 0, \quad u(1,t) = 0. \tag{79}$$

To obtain the adjoint equations, we let $u_t = p_t = 0$ and rewrite (77) in LDG form as

$$\overline{A}w + \overline{B}w_x = \overline{F}, \tag{80}$$

where $w = [p, u, v]^T, \overline{F} = [f_1, f_2, 0]^T, v = \sqrt{\varepsilon}u_x$ and

$$\overline{A} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \overline{B} = \begin{bmatrix} \bar{u} & \bar{c} & 0 \\ \bar{c} & \bar{u} & -\sqrt{\varepsilon} \\ 0 & -\sqrt{\varepsilon} & 0 \end{bmatrix}. \tag{81}$$

The adjoint equations are now found by seeking $\theta = [\phi, \psi, v]^T$ such that

$$J(w) = (\theta, \overline{F}). \tag{82}$$

Integration by parts gives

$$J(w) = J(w) - \int_0^1 \theta^T(\overline{A}w + \overline{B}w_x - \overline{F})dx = \int_0^1 w^T(\overline{G} - \overline{A}\theta + \overline{B}\theta_x)dx - \left[\theta^T\overline{B}w\right]_0^1 + (\theta, \overline{F}), \tag{83}$$

where $\overline{G} = [g_1, g_2, 0]^T$. The adjoint problem is hence given on component form as

$$\begin{aligned} &-\bar{u}\phi_x - \bar{c}\psi_x = g_1, \\ &-\bar{c}\phi_x - \bar{u}\psi_x + \sqrt{\varepsilon}v_x = g_2, \\ &v + \sqrt{\varepsilon}\psi_x = 0, \\ &\psi(0,t) = 0, \\ &\phi(1,t) = 0, \\ &\psi(1,t) = 0. \end{aligned} \tag{84}$$

Note that the dual problem has one boundary condition at $x = 0$ and two at $x = 1$, in contrast to the primal problem for which the situation is reversed.

The time-dependent problem (80) is discretized as

$$\begin{aligned} &\frac{d}{dt}p_h + \bar{u}P^{-1}Qp_h + \bar{c}P^{-1}Qu_h = \sigma_1 P^{-1}(e_0^T p_h - \mathbf{0})e_0 + \sigma_2 P^{-1}(e_0^T u_h - \mathbf{0})e_0 + \sigma_3 P^{-1}(e_N^T u_h - \mathbf{0})e_N, \\ &\frac{d}{dt}u_h + \bar{c}P^{-1}Qp_h + \bar{u}P^{-1}Qu_h - \sqrt{\varepsilon}P^{-1}Q\,v_h = \tau_1 P^{-1}(e_0^T p_h - \mathbf{0})e_0 + \tau_2 P^{-1}(e_0^T u_h - \mathbf{0})e_0 + \tau_3 P^{-1}(e_N^T u_h - \mathbf{0})e_N, \\ &v_h - \sqrt{\varepsilon}P^{-1}Qu_h = \gamma_1 P^{-1}(e_0^T p_h - \mathbf{0})e_0 + \gamma_2 P^{-1}(e_0^T u_h - \mathbf{0})e_0 + \gamma_3 P^{-1}(e_N^T u_h - \mathbf{0})e_N \end{aligned} \tag{85}$$

and the coefficients $\sigma_{1,2,3}, \tau_{1,2,3}$ and $\gamma_{1,2,3}$ has to be determined such that the scheme is stable. By applying the energy method to each of the equations and adding them, we can write the result as

$$\frac{d}{dt}||p_h||^2 + \frac{d}{dt}||u_h||^2 + 2||v_h||^2 = w_h^T M_0 w + w_h M_N w_h, \tag{86}$$

where $w_h = [p_h, u_h, v_h]$ and

$$M_0 = \begin{bmatrix} (\bar{u}+2\sigma_1)E_0 & (\bar{c}+\sigma_2+\tau_1)E_0 & \gamma_1 E_0 \\ (\bar{c}+\sigma_2+\tau_1)E_0 & (\bar{u}+2\tau_2)E_0 & (\gamma_2-\sqrt{\varepsilon})E_0 \\ \gamma_1 E_0 & (\gamma_2-\sqrt{\varepsilon})E_0 & 0 \end{bmatrix},$$
$$M_N = \begin{bmatrix} -\bar{u}E_N & (-\bar{c}+\sigma_3)E_N & 0 \\ (-\bar{c}+\sigma_3)E_N & (-\bar{u}+2\tau_3)E_N & (\gamma_3+\sqrt{\varepsilon})E_N \\ 0 & (\gamma_3+\sqrt{\varepsilon})E_N & 0 \end{bmatrix}. \tag{87}$$

To simplify (86), we introduce the Kronecker product, which is defined for arbitrary matrices X and Y by

$$X \otimes Y = \begin{bmatrix} x_{11}Y & x_{12}Y & \dots & x_{1n}Y \\ x_{21}Y & x_{22}Y & \dots & x_{2n}Y \\ \vdots & \ddots & \ddots & \vdots \\ x_{m1}Y & x_{m2}Y & \dots & x_{mn}Y \end{bmatrix}. \tag{88}$$

The Kronecker product is bilinear, associative and satisfies the mixed product property

$$(X_1 \otimes Y_1)(X_2 \otimes Y_2) = (X_1X_2 \otimes Y_1Y_2) \tag{89}$$

if the usual matrix products are defined. For inversion and transposing we have

$$(X \otimes Y)^{-1,T} = (X^{-1,T} \otimes Y^{-1,T}) \tag{90}$$

if the usual matrix inverses are defined.

Using the Kronecker product, we can factorize (86) as

$$\frac{d}{dt}||p_h||^2 + \frac{d}{dt}||u_h||^2 + 2||v_h||^2 = w_h^T(m_0 \otimes E_0)w_h + w_h^T(m_N \otimes E_N)w_h, \tag{91}$$

where $m_{0,N}$ are the smaller submatrices

$$m_0 = \begin{bmatrix} \bar{u}+2\sigma_1 & \bar{c}+\sigma_2+\tau_1 & \gamma_1 \\ \bar{c}+\sigma_2+\tau_1 & \bar{u}+2\tau_2 & \gamma_2-\sqrt{\varepsilon} \\ \gamma_1 & \gamma_2-\sqrt{\varepsilon} & 0 \end{bmatrix}, \tag{92}$$

$$m_N = \begin{bmatrix} -\bar{u} & -\bar{c}+\sigma_3 & 0 \\ -\bar{c}+\sigma_3 & -\bar{u}+2\tau_3 & \gamma_3+\sqrt{\varepsilon} \\ 0 & \gamma_3+\sqrt{\varepsilon} & 0 \end{bmatrix}. \tag{93}$$

Since $E_0, E_N \geqslant 0$, we obtain a stable scheme is the coefficients are chosen such that $m_0, m_N \leqslant 0$. The coefficients are given in

Proposition 4.1. *The scheme (85) is stable using*

$$\sigma_1 \leqslant -\frac{\bar{u}}{2}, \quad \bar{c}+\sigma_2+\tau_1 = 0, \quad \tau_2 \leqslant -\frac{\bar{u}}{2}, \quad \gamma_1 = 0, \quad \gamma_2 = \sqrt{\varepsilon} \tag{94}$$

for the coefficients in (92) and

$$\sigma_3 = \bar{c}, \quad \gamma_3 = -\sqrt{\varepsilon}, \quad \tau_3 \leqslant \frac{\bar{u}}{2} \tag{95}$$

for the coefficients in (93).

Proof. By inserting the coefficients (94) and (95) into the scheme (85), the energy estimate (91) reduces to

$$\frac{d}{dt}||p_h||^2 + \frac{d}{dt}||u_h||^2 + 2||v_h||^2 \leqslant 0. \qquad \square \tag{96}$$

To determine the spatial dual consistency of (85), we let $p_t = u_t = 0$ and rewrite as

$$L_h w_h = \widetilde{F}, \tag{97}$$

where $\widetilde{F} = [Pf_1, Pf_2, 0]^T$ and

$$L_h = \begin{bmatrix} \bar{u}Q - \sigma_1 E_0 & \bar{c}Q - \sigma_2 E_0 - \bar{c}E_N & 0 \\ \bar{c}Q - \tau_1 E_0 & \bar{u}Q - \tau_2 E_0 - \tau_3 E_N & -\sqrt{\varepsilon}Q \\ 0 & -\sqrt{\varepsilon}Q - \sqrt{\varepsilon}E_0 + \sqrt{\varepsilon}E_N & P \end{bmatrix}. \tag{98}$$

The discrete dual problem is then given by

$$L_h^T \theta_h = \widetilde{G}, \tag{99}$$

where $\theta_h = [\phi_h, \psi_h, v_h]^T$, $\widetilde{G} = [Pg_1, Pg_2, 0]^T$, and it has to be a consistent approximation of (84) without violating the stability conditions. By expanding (99), using (94) and (95), we get

$$\begin{aligned}
&-\bar{u}P^{-1}Q\phi_h - \bar{c}P^{-1}Q\psi_h = (\sigma_1 + \bar{u})P^{-1}E_0\phi_h + (\tau_1 + \bar{c})P^{-1}E_0\psi_h - \bar{u}P^{-1}E_N\phi_h - \bar{c}P^{-1}E_N\psi_h + g_1,\\
&-\bar{c}P^{-1}Q\phi_h - \bar{u}P^{-1}Q\psi_h + \sqrt{\varepsilon}P^{-1}Qv_h = (\sigma_2 + \bar{c})P^{-1}E_0\phi_h + (\tau_2 + \bar{u})P^{-1}E_0\psi_h + (\tau_3 - \bar{u})P^{-1}E_N\psi_h + g_2,\\
&\sqrt{\varepsilon}P^{-1}Q\psi_h + v = -\sqrt{\varepsilon}P^{-1}E_0\psi_h + \sqrt{\varepsilon}P^{-1}E_N\psi_h.
\end{aligned} \tag{100}$$

Remember that the boundary conditions in the dual Eq. (84) are different from those of the primal equation. This puts restrictions on the coefficients in order to obtain a consistent approximation of the dual problem. The coefficients are given in

Proposition 4.2. *The scheme (*85*) is stable and spatially dual consistent with (*94*), (*95*) and the choices*

$$\sigma_1 = -\bar{u}, \quad \sigma_2 = -\bar{c}. \tag{101}$$

Proof. The choice (101) cancels the terms in (100) for which additional erroneous boundary conditions would be imposed for the dual problem. Note that $\sigma_2 = -\bar{c}$ implies

$$\tau_1 = 0. \tag{102}$$

The choice of coefficients given in (101) and (102) does not violate the stability conditions given in (94) and (95). □

Remark 4.5. Note that only the coefficients at the inflow boundary are uniquely determined by the spatial dual consistency requirements. For the outflow boundary, the conditions for stability are sufficient.

Remark 4.6. The requirements for spatial dual consistency has always constituted a subset of the stability requirements. We have hence been able to construct schemes which are both energy stable and spatially dual consistent. The energy analysis for stability typically renders some coefficients in the SAT to be semi-bounded, while the additional requirement of spatial dual consistency fixes some coefficients to unique values in the semi-bounded region.

5. Numerical results

A forcing function have been chosen in all cases such that an analytical solution is known, and the rate of convergence and errors are computed with respect to the analytical solution. The analytical solution is smooth for all times, even for the viscous Burger's equation. This is known as the method of manufactured solutions [51]. Note that the boundary and initial data are constructed from the analytical solution and are hence the conditions are no longer homogeneous.

The time integration is performed until time $t = 10$ using the classical 4th-order Runge–Kutta method with timestep $\Delta t = 2 \times 10^{-6}$, to ensure that the time integration errors are negligible. In each time step we perform a mesh refinement from 32 to 160 gridpoints, in steps of 16, and compute the rate of convergence for both the solution and the functional. In this way, the rate of convergence can be computed as a function of time.

We compare the new schemes with standard SBP-SAT schemes which impose the Dirichlet boundary conditions traditionally without respect to the dual problem. The solutions to all problems were verified to converge with the design order of accuracy. In Tables 1 and 2 we summarize the time-average rates of functional convergence for the dual consistent and dual inconsistent cases, respectively.

The advection equation, heat equation, viscours Burger's equation and the incompletely parabolic system of equations are representatives for the hyperbolic, parabolic, nonlinear and mixed type of partial differential equations. Despite them being different in nature, the results regarding the functional convergence are consistent. A spatially dual consistent SBP-SAT discretization gives rise to time-dependent superconvergent linear functional output.

Table 1
Time-average rates of the functional convergence for the dual consistent discretization.

Accuracy	Advection	Heat	Burger's	System $(J(p), J(u))$
3rd	4.14808	4.0073	4.19861	4.27252, 4.18926
4th	6.9023	6.86841	6.36518	6.61803, 6.53875
5th	6.99999	8.83809	8.61754	8.76432, 8.67103

Table 2
Time-average rates of the functional convergence for the dual inconsistent discretization.

Accuracy	Advection	Heat	Burger's	System $(J(p), J(u))$
3rd	3.06438	4.17441	3.93663	2.71162, 3.68422
4th	4.13107	5.22073	5.08856	3.41406, 3.72249
5th	4.64093	5.42542	5.60646	4.53447, 4.25429

Table 3
Average errors using $N = 32$ grid points.

Accuracy	Solution for p			Functional for p		
	Consistent	Wide	Compact	Consistent	Wide	Compact
3rd	2.0446e−03	2.0571e−03	1.6012e−03	5.0140e−05	2.8720e−04	5.6833e−04
4th	1.8328e−03	1.3131e−03	1.2423e−03	2.3244e−05	4.0409e−04	8.4830e−04
5th	1.1855e−02	6.9236e−03	6.9241e−03	1.2150e−05	1.2854e−03	3.1519e−03
	Solution for u			Functional for u		
3rd	5.0395e−03	1.0337e−03	4.2541e−04	1.0125e−04	5.1268e−04	4.6830e−04
4th	2.1250e−03	1.0265e−03	4.1681e−04	1.6691e−05	3.1254e−04	3.6392e−04
5th	1.5030e−02	1.1059e−02	3.9369e−03	9.7499e−06	6.1595e−04	5.2289e−03

We stress that the method presented does not require any knowledge about the solution of the adjoint equations. Spatial dual consistency is a property of the discretization based upon knowledge of the form of the adjoint equation and its boundary conditions. Superconvergent functionals are thus obtained at no extra computational cost.

The superconvergence of the functional ensures that for sufficiently high resolutions, the dual consistent discretization will outperform a spatially dual inconsistent discretization. Most realistic simulations are, however, marginally or under-resolved and it is desirable that the higher order accuracy does not come at the cost of large error constants which ruin computations on a coarse mesh.

The errors in the solution and in the linear functionals were computed for a coarse mesh. The solution and functional errors were computed as a function of time for the coarsest grid level, $N = 32$ grid points. We consider only the incompletely parabolic case to reduce the number of tables. The results were verified to be analogous for the other cases. We have also included an inconsistent scheme with a more accurate compact discretization of the second derivative as described in [20,52]. The errors are summarized in Table 3, where we present the average error over time for both the solutions and the functionals. From Table 3, we can see that the dual consistent discretization is somewhat less accurate in computing the solution, but much more accurate in computing the functionals. The 5th-order accurate spatially dual consistent discretization is already at 32 gridpoints 2 orders of magnitude more accurate than the spatially dual inconsistent discretization.

6. Summary and conclusions

We have defined and derived spatially dual consistent discretizations based on finite difference operators satisfying the summation-by-parts properties. The boundary conditions were imposed weakly using the simultaneous approximation term. We have presented derivations of spatial dual consistency in a general way and applied the technique to four representative equations; the advection equation, the heat equation, the viscous Burgers' equation and an incompletely parabolic system of equations.

In the cases we considered, the requirements for spatial dual consistency conform with the stability requirements. It was hence always possible to derive schemes which are both energy stable and spatially dual consistent for the cases we have considered, despite all model problems being of different type.

It was shown for all considered cases that a spatial dual consistent discretization produced superconvergent linear functionals computed from the solution. By superconvergece we mean that the solution is accurate of order $p + 1$ (or $p + 2$ under certain conditions), while the linear functional is computed with $2p$-order accuracy.

We have computed the errors in both the solution and in the linear functionals for a coarse mesh to ensure that the superconvegence does not come at the cost of large error constants. It was seen that the solution computed from the spatially dual consistent scheme was somewhat less accurate, while the functional could be two orders of magnitude more accurate already on a coarse grid.

The superconvergence does not require any knowledge about the solution of the adjoint equations. By considering only the form of the adjoint equation and its boundary conditions, it is a matter of choosing the SAT such that the scheme becomes stable and spatial dual consistent. Superconvergent functional outputs can thus be computed at no extra computational cost compared to a standard discretization.

Acknowledgments

Most of the computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project p2010056.

This work has partly been carried out within the IDIHOM project which is supported by the European Commission under contract No. ACP0-GA-2010–265780.

References

[1] Michael B. Giles, Niles A. Pierce. Adjoint equations in CFD: duality, boundary conditions and solution behaviour. In AIAA Paper 97–1850, 1997.
[2] Michael B. Giles, Mats G. Larson, J. Mårten Levenstam, Endre Süli, Adaptive error control for finite element approximations of the lift and drag coefficients in viscous flow. Technical report, Report NA-97/06, Oxford University Computing Laboratory, 1997.
[3] Niles A. Pierce, Michael B. Giles, Adjoint recovery of superconvergent functionals from PDE approximations, SIAM Review 42 (2) (2000) 247–264.
[4] Michael B. Giles, Niles A. Pierce, Superconvergent lift estimates through adjoint error analysis. Innovative Methods for Numerical Solutions of Partial Differential Equations, 2001.
[5] Niles A. Pierce, Michael B. Giles, Adjoint and defect error bounding and correction for functional estimates, Journal of Computational Physics 200 (November 2004) 769–794.
[6] Michael B. Giles and Niles A. Pierce. On the properties of solutions of the adjoint Euler equations. Numerical Methods for Fluid Dynamics VI ICFD, pages 1–16, 1998.
[7] Yingda Cheng, Chi-Wang Shu, Superconvergence of discontinuous galerkin and local discontinuous galerkin schemes for linear hyperbolic and convection-diffusion equations in one space dimension, SIAM Journal on Numerical Analysis 47 (6) (2010) 4044–4072.
[8] Krzysztof J. Fidkowski, Output error estimation strategies for discontinuous galerkin discretizations of unsteady convection-dominated flows, International Journal for Numerical Methods in Engineering 88 (12) (2011) 1297–1322.
[9] James Lu. An a posteriori error control framework for adaptive precision optimization using discontinuous Galerkin finite element method. PhD thesis, Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, 2005.
[10] Todd A. Oliver, David L. Darmofal, Analysis of dual consistency for discontinuous galerkin discretizations of source terms, SIAM Journal on Numerical Analysis 47 (5) (2009) 3507–3525.
[11] Karthik Mani, Dimitri J. Mavriplis, Error estimation and adaptation for functional outputs in time-dependent flow problems, Journal of Computational Physics 229 (January 2010) 415–440.
[12] Kui Ou and Antony Jameson. Unsteady adjoint method for the optimal control of advection and Burgers' equations using high order spectral difference method. In 49th AIAA Aerospace Sciences Meeting, 2011.
[13] Krzysztof J. Fidkowski, David L. Darmofal, Review of output-based error estimation and mesh adaptation in computational fluid dynamics, AIAA Journal 49 (4) (2011) 673–694.
[14] Michael B. Giles, Endre Süli, Adjoint methods for PDEs: a posteriori error analysis and postprocessing by duality, Acta Numerica 11 (2002) 145–236.
[15] Antony Jameson, Aerodynamic design via control theory, Journal of Scientific Computing 3 (September 1988) 233–260.
[16] Michael B. Giles, Niles A. Pierce, An introduction to the adjoint approach to design, Flow, Turbulence and Combustion 65 (2000) 393–415.
[17] Jason E. Hicken, David W. Zingg, Superconvergent functional estimates from summation-by-parts finite-difference discretizations, SIAM Journal on Scientific Computing 33 (2) (2011) 893–922.
[18] Jason E. Hicken, Output error estimation for summation-by-parts finite-difference schemes, Journal of Computational Physics 231 (9) (2012) 3828–3848.
[19] Bo Strand, Summation by parts for finite difference approximations for d/dx, Journal of Computational Physics 110 (1) (1994) 47–67.
[20] Ken Mattsson, Jan Nordström, Summation by parts operators for finite difference approximations of second derivatives, Journal of Computational Physics 199 (2) (2004) 503–540.
[21] Mark H. Carpenter, David Gottlieb, 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) (1994) 220–236.
[22] M. Berndt, K. Lipnikov, M. Shashkov, M.F. Wheeler, I. Yotov, Superconvergence of the velocity in mimetic finite difference methods on quadrilaterals, SIAM Journal on Numerical Analysis 43 (4) (2005) 1728–1749.
[23] Qiqi Wang, Parviz Moin, Gianluca Iaccarino, Minimal repetition dynamic checkpointing algorithm for unsteady adjoint calculation, SIAM Journal on Scientific Computing 31 (4) (2009) 2549–2567.
[24] Steven M. Kast, Krzysztof J. Fidkowski, and Philip L. Roe. An unsteady entropy adjoint approach for adaptive solution of the shallow-water equations. In AIAA Paper Number 2011–3694, 2011.
[25] Qiqi Wang, David Gleich, Amin Saberi, Nasrollah Etemadi, Parviz Moin, A monte carlo method for solving unsteady adjoint equations, Journal of Computational Physics 227 (12) (2008) 6184–6205.
[26] Nail K. Yamaleev, Boris Diskin, Eric J. Nielsen, Local-in-time adjoint-based method for design optimization of unsteady flows, Journal of Computational Physics 229 (July 2010) 5394–5407.
[27] Shengtai Li, Linda Petzold, Adjoint sensitivity analysis for time-dependent partial differential equations with adaptive mesh refinement, Journal of Computational Physics 198 (1) (2004) 310–325.
[28] Heinz-Otto Kreiss and Godela Scherer. Finite element and finite difference methods for hyperbolic partial differential equations. In Mathematical Aspects of Finite Elements in Partial Differential Equations, number 33 in Publ. Math. Res. Center Univ. Wisconsin, pages 195–212. Academic Press, 1974.
[29] Heinz-Otto Kreiss and Godela Scherer. On the existence of energy estimates for difference approximations for hyperbolic systems. Technical report, Uppsala University, Division of Scientific Computing, 1977.
[30] Mark H. Carpenter, Jan Nordström, David Gottlieb, Revisiting and extending interface penalties for multi-domain summation-by-parts operators, Journal of Scientific Computing 45 (1–3) (2010) 118–150.
[31] Jan Nordström, Jing Gong, Edwin van der Weide, Magnus Svärd, A stable and conservative high order multi-block method for the compressible Navier-Stokes equations, Journal of Computational Physics 228 (24) (2009) 9020–9035.
[32] Magnus Svärd, Jan Nordström, On the order of accuracy for difference approximations of initial-boundary value problems, Journal of Computational Physics 218 (1) (2006) 333–352.
[33] Ken Mattsson, Magnus Svärd, Mark Carpenter, Jan Nordström, High-order accurate computations for unsteady aerodynamics, Computers and Fluids 36 (3) (2007) 636–649.
[34] Magnus Svärd, Jan Nordström, A stable high-order finite difference scheme for the compressible Navier-Stokes equations: No-slip wall boundary conditions, Journal of Computational Physics 227 (10) (2008) 4805–4824.
[35] Magnus Svärd, Ken Mattsson, Jan Nordström, Steady-state computations using summation-by-parts operators, Journal of Scientific Computing 24 (1) (2005) 79–95.
[36] X. Huan, Jason E. Hicken, and David W. Zingg. Interface and boundary schemes for high-order methods. In the 39th AIAA Fluid Dynamics Conference, AIAA Paper No. 2009–3658, San Antonio, USA, 22–25 June 2009.

[37] Ken Mattsson, Magnus Svärd, Mohammad Shoeybi, Stable and accurate schemes for the compressible Navier-Stokes equations, Journal of Computational Physics 227 (February 2008) 2293–2316.
[38] Pelle Olsson, Summation by parts, projections, and stability, I. Mathematics of Computation 64 (211) (1995) 1035–1065.
[39] Pelle Olsson, Summation by parts, projections, and stability, II. Mathematics of Computation 64 (212) (1995) 1473–1493.
[40] Jan Nordström, Mark H. Carpenter, High-order finite difference methods, multidimensional linear problems, and curvilinear coordinates, Journal of Computational Physics 173 (1) (2001) 149–174.
[41] Jan Nordström, Conservative finite difference formulations, variable coefficients, energy estimates and artificial dissipation, Journal of Scientific Computing 29 (2006) 375–404.
[42] Jeremy E. Kozdon, Eric M. Dunham, and Jan Nordström. Simulation of Dynamic Earthquake Ruptures in Complex Geometries Using High-Order Finite Difference Methods. Technical Report LiTH-MAT-R, 2012:2, Linköping University, Department of Mathematics, Sweden, 2011.
[43] Ken Mattsson. Summation by parts operators for finite difference approximations of second-derivatives with variable coefficients. Journal of Scientific Computing, pages 1–33, 2011.
[44] Jan Nordström, Magnus Svärd, Well-posed boundary conditions for the Navier–Stokes equations, SIAM Journal on Numerical Analysis 43 (3) (2005) 1231–1255.
[45] Bertil Gustafsson, Heinz-Otto Kreiss, and Joseph Oliger. Time Dependent Problems and Difference Methods. Wiley Interscience, 1995.
[46] Jason E. Hicken and David W. Zingg. Summation-by-parts operators and high-order quadrature. ArXiv e-prints, March 2011.
[47] Bernardo Cockburn, Chi-Wang Shu, The local discontinuous galerkin method for time-dependent convection-diffusion systems, SIAM Journal on Numerical Analysis 35 (October 1998) 2440–2463.
[48] Jing Gong, Jan Nordström, Interface procedures for finite difference approximations of the advection-diffusion equation, Journal of Computational and Applied Mathematics 236 (5) (2011) 602–620.
[49] Heinz-Otto Kreiss and Jens Lorenz. Initial-Boundary Value Problems and the Navier-Stokes Equations. SIAM, 2004.
[50] Saul Abarbanel, David Gottlieb, Optimal time splitting for two- and three-dimensional Navier-Stokes equations with mixed derivatives, Journal of Computational Physics 41 (1) (1981) 1–33.
[51] Jens Lindström, Jan Nordström, A stable and high-order accurate conjugate heat transfer problem, Journal of Computational Physics 229 (14) (2010) 5440–5456.
[52] Mark H. Carpenter, Jan Nordström, David Gottlieb, A stable and conservative interface treatment of arbitrary spatial accuracy, Journal of Computational Physics 148 (2) (1999) 341–365.

Paper VI

On the impact of boundary conditions on dual consistent finite difference discretizations

Jens Berg*

Uppsala University, Department of Information Technology, SE-751 05, Uppsala, Sweden

Jan Nordström

Linköping University, Department of Mathematics, SE-581 83, Linköping, Sweden

Abstract

In this paper we derive well-posed boundary conditions for a linear incompletely parabolic system of equations, which can be viewed as a model problem for the compressible Navier–Stokes equations. We show a general procedure for the construction of the boundary conditions such that both the primal and dual equations are well-posed. The form of the boundary conditions is chosen such that reduction to first order form with its complications can be avoided.

The primal equation is discretized using finite difference operators on summation-by-parts form with weak boundary conditions. It is shown that the discretization can be made energy stable, and that energy stability is sufficient for dual consistency. Since reduction to first order form can be avoided, the discretization is significantly simpler compared to a discretization using Dirichlet boundary conditions.

We compare the new boundary conditions with standard Dirichlet boundary conditions in terms of rate of convergence, errors and discrete spectra. It is shown that the scheme with the new boundary conditions is not only far simpler, but also has smaller errors, error bounded properties, and highly optimizable eigenvalues, while maintaining all desirable properties of a dual consistent discretization.

Keywords: High order finite differences; Summation-by-parts; Superconvergence;

*Corresponding author: Jens Berg
Address: Division of Scientific Computing, Department of Information Technology, Uppsala University, Box 337, SE-751 05, Uppsala, Sweden
Phone: +46 18 - 471 6253
Fax: +46 18 523049, +46 18 511925
E-mail: jens.berg@it.uu.se

Boundary conditions; Dual consistency; Stability

1. Introduction

Functionals can represent the lift or drag on an aircraft, energy or any other scalar quantity computed from the solution to a partial differential equation (PDE). In many engineering applications, high order accurate functionals are often of greater interest than accurate solutions of the equations themselves. Whenever there is a functional involved, the concept of duality becomes important. The solution of a PDE resides in some function space, and the set of all bounded linear functionals on that space is called its dual space. Knowledge of the functional of interest can thus be obtained by studying the dual space. This is the main topic in functional analysis and references can be found in any standard textbook.

The dual equations need, as the primal equations, to be supplied with the correct boundary conditions in order for a solution to exist and be unique. Like the dual differential operator, the dual boundary conditions depend on the primal problem and it is a non-trivial task to construct boundary conditions such that both the primal and dual problems are well-posed. The most common choice is Dirichlet boundary conditions since the analysis of the continuous equations is simplified. However, it is well-known that Dirichlet boundary conditions cause large reflections and reduces the accuracy and stability properties of a numerical scheme [18]. Other kinds of boundary conditions are beneficial for a discretization, but complicate the analysis of the dual equations. Here, we shall study boundary conditions of far-field type which have been shown to be beneficial for discretizations [19, 23, 21].

In numerical analysis, and in particular for computational fluid dynamics problems, duality has been exploited for optimal control problems [27, 15, 8], error estimation [26, 34, 35, 12, 7] and convergence acceleration [9, 25, 13, 4]. An extensive summary of the use of adjoint problems can be found in [10], and more recently in [6] with focus on error estimation and adaptive mesh refinement.

In this paper we will use a finite difference method on summation by parts (SBP) form with boundary conditions weakly imposed by the simultaneous approximation term (SAT). There has recently been a development of the quadrature properties of SBP-SAT discretizations. The base for an SBP operator is a norm matrix, denoted by P. The norm matrix is an integration operator for equidistant grid points. The integration properties of P has been studied in [14] and it was shown that P is a high-order accurate quadrature rule which extends the Gregory formulas. In [13], the discretization of steady problems were considered. The authors showed that certain SBP-SAT discretizations, so called dual consistent discretizations, led to

superconvergent functionals if the same P was used in the discretization as in the functional evaluation. The theory of dual consistency and superconvergence was extended to time-dependent problems in [4]. Several problems were analyzed and it was shown that dual consistency and stability implies superconvergence of linear functionals.

In order to avoid additional theoretical difficulties in [4], Dirichlet boundary conditions for both the primal and dual problems were used. The Dirichlet boundary conditions ensured that both problems were well-posed without additional efforts. In an Euler or Navier–Stokes calculation, however, Dirichlet boundary conditions are rarely used at far-field boundaries. Unless exact boundary data is known, Dirichlet boundary conditions cause reflections which pollute the solution.

In this paper we will investigate the potential gain, or loss, when replacing the Dirichlet boundary conditions with more sophisticated ones. The aim of the paper is to derive well-posed boundary conditions for both the primal and dual problems such that the complications of having Dirichlet boundary conditions are removed, while maintaining all desirable properties of a dual consistent discretization and sophisticated boundary conditions.

2. Preliminaries

In this paper, we will consider time-dependent partial differential equations of the form

$$\begin{aligned} u_t + L(u) &= f, \\ J(u) &= (g, u), \end{aligned} \tag{1}$$

where $J(u)$ is a linear functional output of interest with $g = g(x, t)$ being an arbitrary weight function. L can be either linear or non-linear and u can represent either a scalar or vector valued function. Detailed descriptions can be found in [4] but for convenience, we summarize the main preliminaries here.

The inner product is the standard L^2-inner product

$$(u, v) = \int_\Omega u^T v d\Omega.$$

In [4], the concept of spatial dual consistency was introduced to avoid treating the full time-dependent dual equations when discretizing using a method of lines. The concept is motivated by the following. To find the dual problem, we follow the

notation in [4, 13], and seek a function θ in some appropriate function space, such that

$$\int_0^T J(u)dt = \int_0^T (\theta, f)dt.$$

A formal computation (assume L linear and u, θ to have compact support in space) gives

$$\begin{aligned}\int_0^T J(u)dt &= \int_0^T J(u)dt - \int_0^T (\theta, u_t + Lu - f)dt \\ &= \int_0^T (\theta_t - L^*\theta + g, u)dt - \int_\Omega [\theta u]_0^T d\Omega + \int_0^T (\theta, f)dt,\end{aligned}$$

where L^* is the formal adjoint, or dual, operator associated with L under the inner product such that $(\theta, Lu) = (L^*\theta, u)$. By having homogeneous initial conditions for the primal problem, we obtain the time-dependent dual problem as

$$-\theta_t + L^*\theta = g, \tag{2}$$

where we have to put an initial condition for the dual problem at time $t = T$. The time transformation $\tau = T - t$ transforms (2) to

$$\theta_\tau + L^*\theta = g$$

with an initial condition at $\tau = 0$. A discretization which simultaneously approximates the spatial primal and dual operator consistently, is called spatially dual consistent and produces superconvergent time-dependent linear functionals if the scheme for the primal problem is stable [4].

A difference operator for the first derivative is said to be on SBP form if it can be written as $D_1 = P^{-1}Q$. P defines a norm by $||u_h||^2 = u_h^T P u_h$ and Q satisfies the SBP property $Q + Q^T = E_N - E_0$, where

$$E_N = \text{diag}[0, \ldots, 0, 1], \quad E_0 = \text{diag}[1, 0, \ldots, 0].$$

The second derivative operator can be constructed either by applying the first derivative twice, i.e. $D_2 = (P^{-1}Q)^2$ which results in a wide operator, or a compact operator with minimal bandwidth of the form

$$D_2 = P^{-1}(-H + (E_N - E_0)S)$$

as described in [5, 17, 16]. In this paper, we consider only diagonal [28] norms and wide second derivative operators. The diagonal norm is flexible for realistic simulations as the resulting schemes can be shown to be energy stable under curvilinear coordinate transforms. This does not hold for non-diagonal norms [20, 22, 30, 32].

A first derivative SBP operator is essentially a $2s$-order accurate central finite difference operator which has been modified close to the boundaries such that it becomes one-sided. Together with the diagonal norm, the boundary closure is accurate of order s making the SBP operator accurate of order $s+1$ in general [28]. For problems with second derivatives, the compact operator can be modified with higher order accurate boundary closures to gain one extra order of accuracy [17, 33].

A discretization of the primal problem (1) can be written as

$$\frac{d}{dt}u_h + L_h u_h = f,$$

where u_h is the discrete approximation of u and L_h is a discrete approximation of L, including the boundary conditions. The discrete inner product in an SBP setting is defined by

$$(u_h, v_h)_h = u_h^T P v_h$$

and hence the discrete adjoint operator can be computed, according to the definition

$$(v_h, L_h u_h)_h = (L_h^* v_h, u_h)_h,$$

as

$$L_h^* = P^{-1} L_h^T P.$$

The proof that a stable and spatially dual consistent SBP scheme produces superconvergent linear functionals is presented in [4]. The proof is based on the fact that the mass matrix, P, in the norm is a $2s$-order accurate integration operator [14, 13].

The procedure for constructing stable schemes which produce superconvergent linear functionals can now be summarized as follows;

1. Determine boundary conditions such that both the primal and dual problems are well-posed
2. Discretize the primal problem and ensure stability
3. Compute L_h^* and choose the remaining parameters (if any) such that the continuous adjoint L^* is consistently approximated together with the dual boundary conditions

Note that a stable and consistent discretization of the primal problem does not imply that the dual problem is consistently approximated.

3. Linear incompletely parabolic system

We shall study the linear incompletely parabolic system of equations on $0 \leq x \leq 1$ given by

$$U_t + AU_x = BU_{xx}, \tag{3}$$

where $U = [p, u]^T$ and

$$A = \begin{bmatrix} \bar{u} & \bar{c} \\ \bar{c} & \bar{u} \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 0 & \varepsilon \end{bmatrix},$$

together with a linear functional of interest,

$$J(U) = \int_0^1 G^T U dx. \tag{4}$$

Equation (3) can be viewed as the symmetrized [2], one-dimensional Navier–Stokes equations, linearized around a flow field with constant velocity $\bar{u} > 0$ and speed of sound $\bar{c} > 0$. In this case, we assume $\bar{u} < \bar{c}$ and it can be shown [24, 29] that (3) requires two boundary conditions at the inflow boundary, $x = 0$, and one at the outflow boundary at $x = 1$, in both the subsonic and supersonic case. Even though this is a model problem, we denote the case $\bar{u} < \bar{c}$ as subsonic, and supersonic if $\bar{u} > \bar{c}$.

3.1. Dirichlet boundary conditions

In [4], equation (3) was supplied with the Dirichlet boundary conditions

$$p(0,t) = 0, \quad u(0,t) = 0, \quad u(1,t) = 0.$$

An energy estimate results in

$$||U||_t^2 + ||BU_x||^2 \leq 0, \tag{5}$$

where we used the notation $||U||_t^2 = \frac{d}{dt}(||U||^2)$. Note that the boundary conditions cancel all boundary terms completely, and does not give any additional damping of the energy, $||U||^2$.

The spatial dual operator was obtained by reducing (3) to a first order system by introducing the auxiliary variable $v = \sqrt{\varepsilon} u_x$. There are several drawbacks with this technique. The most obvious one is that it results in a larger system of equations which complicates the analysis. The drawbacks of the first order form also carries over to the discretization. In the discretization of the first order form, there are nine unknown penalty parameters in the SAT procedure which have to be determined for stability and dual consistency [4]. This makes extensions to larger system in multiple dimensions complicated.

3.2. Flux based boundary conditions

The new boundary conditions we consider are of the form

$$H_{L,R}U \mp BU_x = G_{L,R}, \tag{6}$$

where $H_{L,R}$ will be determined for well-posedness of both the primal and dual problems.

There are many different forms of the matrices $H_{L,R}$ in (6) which give well-posed inflow or outflow boundary conditions. The typical way to determine the structure of $H_{L,R}$ is to diagonalize the hyperbolic part of the equation and consider the ingoing or outgoing characteristics. This method will provide an energy estimate with optimal damping properties [19]. However, the dual problem associated with the linear functional (4) will most likely be ill-posed. Well-posedness of the primal problem does not imply well-posedness of the dual problem.

Since we are only interested in the spatial dual operator, it is sufficient to consider the steady, inhomogeneous, problem

$$AU_x - BU_{xx} = F.$$

In this case, the differential operator L is given by

$$L = A\frac{\partial}{\partial x} - B\frac{\partial^2}{\partial x^2}$$

and we seek $\theta = [\phi, \psi]^T$ such that $J(U) = (\theta, F)$. Integration by parts gives

$$\begin{aligned} J(U) &= (G, U) - (\theta, LU - F) \\ &= (G - L^*\theta, U) - [\theta^T AU - \theta^T BU_x + \theta_x^T BU]_0^1 + (\theta, F), \end{aligned}$$

where $L^*\theta = -A\theta_x - B\theta_{xx}$ and hence the dual operator is given by

$$L^* = -A\frac{\partial}{\partial x} - B\frac{\partial^2}{\partial x^2}. \tag{7}$$

To determine the boundary conditions for the dual problem, we have to find a minimal set of conditions such that

$$[\theta^T AU - \theta^T BU_x + \theta_x^T BU]_0^1 = 0 \tag{8}$$

after the homogeneous boundary conditions for the primal problem have been applied. This is what put restrictions on the matrices $H_{L,R}$ in (6). Not only does the boundary terms have to vanish, they will also have to satisfy the correct number and form for the dual equation. Wrong choice of boundary conditions for the primal problem will cause the dual problem to be ill-posed [24, 29].

3.2.1. Left boundary conditions

To simplify the analysis, we assume the left boundary condition to be homogeneous,

$$H_L U - BU_x = 0. \tag{9}$$

By considering only the terms at $x = 0$, we can write (8) as

$$\theta^T(AU - BU_x) + \theta_x^T BU = U^T((A - H_L^T)\theta + B\theta_x)$$

after having applied the homogeneous boundary conditions (9) for the primal equation. The boundary conditions for the dual equation are thus given by

$$(A - H_L^T)\theta + B\theta_x = 0. \tag{10}$$

The form of the matrix H_L can now be determined. Since the left boundary is an inflow boundary for the primal equation, but an outflow boundary for the dual equation, only one boundary condition is allowed for the dual equation. The boundary conditions (10) hence have to be of rank one and thus it is required that

$$A - H_L^T = \begin{bmatrix} 0 & 0 \\ \alpha_L & \beta_L \end{bmatrix} \tag{11}$$

or equivalently

$$H_L = \begin{bmatrix} \bar{u} & \bar{c} - \alpha_L \\ \bar{c} & \bar{u} - \beta_L \end{bmatrix}. \tag{12}$$

Any other form of H_L would impose too many boundary conditions for the dual equation and it would not be well-posed. The coefficients α_L, β_L have to be chosen such that we obtain an energy estimate for both the time-dependent primal and dual problems.

We can now turn our attention back to the primal equation. The primal equation needs two boundary conditions at $x = 0$, and hence H_L is required to have a non-zero top row. This requirement is automatically fulfilled since $\bar{u} > 0$ by assumption. To determine the coefficients α_L, β_L we apply the energy method to (3) and consider only the left boundary terms. We get

$$||U||_t^2 = U^T AU - U^T BU_x - U_x^T BU. \tag{13}$$

By applying the homogeneous boundary conditions (9), we can write (13) as

$$||U||_t^2 = -U^T M_L U,$$

where

$$M_L = -A + H_L + H_L^T = \begin{bmatrix} \bar{u} & \bar{c} - \alpha_L \\ \bar{c} - \alpha_L & \bar{u} - 2\beta_L \end{bmatrix} \tag{14}$$

and we have to choose the coefficients α_L, β_L such that $M_L \geq 0$. There are several strategies for how to choose the parameters α_L and β_L such that $M_L \geq 0$. The most general is to compute the eigenvalues of M_L and determine the parameters such that all eigenvalues are positive. In this simple 2×2 case, the eigenvalues can be directly computed as

$$\mu_{1,2}^{(L)} = \bar{u} - \beta_L \pm \sqrt{(\bar{u} - \beta_L)^2 - \bar{u}(\bar{u} - 2\beta_L) + (\bar{c} - \alpha_L)^2}$$

and it is required that

$$(\bar{c} - \alpha_L)^2 - \bar{u}(\bar{u} - 2\beta_L) \leq 0.$$

For larger systems and more complicated equations, this approach might not be possible as the eigenvalues are not analytically available. A simpler strategy is proposed in

Proposition 3.1. *The primal equation (3) is well-posed with the left boundary conditions given in (9), where H_L is defined in (12) and the parameters α_L, β_L satisfy*

$$\alpha_L = \bar{c}, \quad \beta_L \leq \frac{\bar{u}}{2}. \tag{15}$$

Proof. The primal problem requires two boundary conditions at $x = 0$. Hence the top row of H_L needs to be non-zero. We can see from (12) that this is always the case since $\bar{u} > 0$ by assumption. By inserting the values in (15) into (14), we get

$$M_L = \begin{bmatrix} \bar{u} & 0 \\ 0 & \bar{u} - 2\beta_L \end{bmatrix}$$

which is diagonal with non-negative diagonal entries. □

Note that the strategy in proposition 3.1 is to i) cancel the off-diagonal terms and ii) ensure that the remaining diagonal terms have the correct sign.

A third option is to determine α_L and β_L such that the boundary conditions converge uniformly to a well-posed set of boundary conditions for the hyperbolic system

$$U_t + AU_x = 0$$

as $\varepsilon \to 0$. The third option will be discussed later, and for now we consider the choices in (15).

It is not only the primal equation which has to be well-posed. The time-dependent dual equation with its dual boundary conditions need also be well-posed with the conditions given in (15). By introducing the time transformation $\tau = T - t$, we can write the time-dependent dual equation as

$$\theta_\tau - A\theta_x = B\theta_{xx}. \tag{16}$$

The well-posedness of the dual problem is proven in

Proposition 3.2. *The time-dependent dual problem (16) is well-posed with the dual boundary conditions (10) where H_L is defined in (12) with the parameters given in (15).*

Proof. The dual problem is only allowed to have one boundary condition at $x = 0$. It is thus required that the top row of $A - H_L^T$ is zero. By construction of H_L, this is the case as can be seen in (11). By applying the energy method to (16), and only considering the terms at $x = 0$, we obtain

$$\begin{aligned} ||\theta||_\tau^2 &= -\theta^T A\theta - \theta^T B\theta_x - \theta_x^T B\theta \\ &= -\theta^T(-A + H_L + H_L^T)\theta \\ &= -\theta^T M_L \theta \end{aligned}$$

after applying the boundary conditions (10). The semi-definiteness of M_L, with the choices (15), were already proven in the energy estimate of the primal equation. □

To summarize, the left homogeneous boundary conditions for the primal problem are given in (9) and for the dual problem in (10), where H_L is defined in (12) with the coefficients given in (15).

3.2.2. Right boundary conditions

The right boundary, $x = 1$, is an outflow boundary for the primal problem and hence only one boundary condition can be used, while we have two variables in the system. This immediately puts restrictions on the homogeneous primal boundary condition

$$H_R U + BU_x = 0 \tag{17}$$

in such a way that H_R is required to have the form

$$H_R = \begin{bmatrix} 0 & 0 \\ \alpha_R & \beta_R \end{bmatrix}. \tag{18}$$

For any other form of H_R, too many boundary conditions are placed at the outflow boundary and the primal problem is ill-posed. The coefficients α_R, β_R have to be determined such that both the time-dependent primal and dual problems are well-posed.

Once the form of H_R in (18) has been determined, we must make sure that the correct number of boundary conditions are imposed on the dual problem. To determine the boundary conditions for the dual problem, we restrict (8) to the terms at $x = 1$. After applying the homogeneous primal boundary conditions (17) we obtain

$$\theta^T(AU - BU_x) + \theta_x^T BU = U^T((A + H_R^T)\theta + B\theta_x)$$

and hence the boundary conditions for the dual problem are given by

$$(A + H_R^T)\theta + B\theta_x = 0, \tag{19}$$

where

$$A + H_R^T = \begin{bmatrix} \bar{u} & \bar{c} + \alpha_R \\ \bar{c} & \bar{u} + \beta_R \end{bmatrix}. \tag{20}$$

Since the dual problem requires two boundary conditions at $x = 1$, it is required that the top row of $A + H_R^T$ is non-zero. We can see that this requirement is automatically fulfilled since $\bar{u} > 0$ by assumption.

The coefficients α_R and β_R can now be determined such that we obtain energy estimates for both the primal and dual equations. The energy method applied to the time-dependent dual problem (16), with the homogeneous dual boundary conditions in (19), gives

$$||\theta||_\tau^2 = -\theta^T M_R \theta,$$

where

$$M_R = A + H_R + H_R^T = \begin{bmatrix} \bar{u} & \bar{c} + \alpha_R \\ \bar{c} + \alpha_R & \bar{u} + 2\beta_R \end{bmatrix}. \tag{21}$$

As this is a 2×2 system, we can directly compute the eigenvalues of the symmetric matrix M_R as

$$\mu_{1,2}^{(R)} = \bar{u} + \beta_R \pm \sqrt{(\bar{u} + \beta_R)^2 - \bar{u}(\bar{u} + 2\beta_R) + (\bar{c} + \alpha_R)^2}$$

and see that they are both non-negative if we choose α_R and β_R such that

$$(\bar{c} + \alpha_R)^2 - \bar{u}(\bar{u} + 2\beta_R) \leq 0.$$

Again, in more realistic situations the eigenvalues might not be analytically computable and we use the same strategy as before — to cancel the off-diagonal elements and ensure that the remaining diagonal terms have the correct sign. The values of α_R and β_R with this strategy are given in

Proposition 3.3. *The time-dependent dual problem (16) is well-posed with the right dual boundary conditions in (19) where the parameters in H_R satisfy*

$$\alpha_R = -\bar{c}, \quad \beta_R \geq -\frac{\bar{u}}{2}. \tag{22}$$

Proof. The dual problem requires two boundary conditions at $x = 1$ and hence the top row of $A + H_R^T$ must be non-zero. That this is always the case can be seen in (20) since $\bar{u} > 0$ by assumption. By substituting the relations in (22) into (21), we obtain

$$M_R = \begin{bmatrix} \bar{u} & 0 \\ 0 & \bar{u} + 2\beta \end{bmatrix}$$

which is diagonal with non-negative diagonal entries. □

The well-posedness of the primal problem is given in

Proposition 3.4. *The primal problem (3) is well-posed with the right boundary conditions in (17), where the parameters in H_R are given in (22).*

Proof. The primal problem requires one boundary condition at $x = 1$ which is fulfilled by the construction of H_R in (18). As before, the energy method applied to the time-dependent primal problem gives

$$\begin{aligned} ||U||_t^2 &= -U^T(A + H_R + H_R^T)U \\ &= -U^T M_R U, \end{aligned}$$

where semi-definiteness of M_R is already proven from the energy estimate of the dual equation. □

To summarize, the right homogeneous boundary conditions for the primal problem are given in (17) and for the dual problem in (19), where H_R is defined in (18) with the coefficients given in (22).

Remark 3.1. Note how it is the problem which requires the least number of boundary conditions which sets restrictions on the form of the boundary conditions. When also considering well-posedness of the dual problem, it can be used to reduce the number of unknown parameters in the boundary conditions of the primal problem.

Remark 3.2. The energy estimate, when considering all terms, for the primal problem is given by

$$||U||_t^2 + ||BU_x||^2 = -U^T M_L U - U^T M_R U,$$

and for the dual problem by

$$||\theta||_\tau^2 + ||B\theta_x||^2 = -\theta^T M_L \theta - \theta^T M_R \theta.$$

Both matrices M_L and M_R have at least one positive eigenvalue and hence the boundary conditions contribute to damping of the energy. In the energy estimate (5) with Dirichlet boundary conditions, no additional damping is obtained.

3.3. Convergence to the hyperbolic system

As was discussed previously, the parameters $\alpha_{L,R}$ and $\beta_{L,R}$ can be chosen such that they converge to well-posed boundary conditions of the hyperbolic system

$$U_t + AU_x = 0 \tag{23}$$

as $\varepsilon \to 0$. The energy method applied to (23) results in

$$||U||_t^2 = -[U^T A U]_0^1.$$

Since A is symmetric there is an orthonormal matrix X which diagonalizes A as $A = X\Lambda X^T$, where $\Lambda = \text{diag}[\bar{u} + \bar{c}, \bar{u} - \bar{c}]$ contains the eigenvalues of A. The energy estimate can hence be rewritten as

$$||U||_t^2 = [(X^T U)^T \Lambda (X^T U)]_0^1$$

and we can see that one boundary condition is required on each boundary, since by assumption we have $\bar{u} < \bar{c}$. Hence, as $\varepsilon \to 0$ the number of boundary conditions change from 2 to 1 on the left boundary for the primal problem. For the dual problem, the number of boundary condition change from 2 to 1 on the right boundary. As a consequence it is required that the matrices H_L in (12) and $A + H_R^T$ in (20) both have rank 1 and non-zero top rows, and that energy estimates can be obtained. The choices which fulfills these requirements are given in

Proposition 3.5. *Let*

$$\alpha_L = \bar{c} - \bar{u}, \quad \beta_L = \bar{u} - \bar{c}, \quad \alpha_R = \bar{u} - \bar{c}, \quad \beta_R = \bar{c} - \bar{u}. \tag{24}$$

Then the boundary conditions

$$H_L U - BU_x = 0,$$
$$H_R U + BU_x = 0,$$

constitute a well-posed set of boundary conditions for the incompletely parabolic system of equations (3) and its dual equations (16), and converge to a well-posed set of boundary conditions for the hyperbolic system (23) and its dual as $\varepsilon \to 0$.

Proof. On the left boundary it is required that H_L has rank 1 and non-zero top row. In this case, two boundary conditions will be imposed if $\varepsilon \neq 0$ and one linearly independent condition if $\varepsilon = 0$. On the right boundary it is required that $A + H_R^T$ has rank 1 and non-zero top row. Thus the dual equations will have two conditions if $\varepsilon \neq 0$ and one linearly independent condition if $\varepsilon = 0$. Inserting the values of $\alpha_{L,R}$ and $\beta_{L,R}$ in (24) gives

$$H_L = \begin{bmatrix} \bar{u} & \bar{u} \\ \bar{c} & \bar{c} \end{bmatrix}, \quad A + H_R^T = \begin{bmatrix} \bar{u} & \bar{u} \\ \bar{c} & \bar{c} \end{bmatrix}$$

and we can see that the requirements are fulfilled. The energy estimates of both the primal and dual equations, independently of ε, are of the form

$$||\xi||_t^2 + ||B\xi_x||^2 = -\xi^T M_L \xi - \xi^T M_R \xi,$$

where $M_{L,R}$ are as before. The 2×2 matrices M_L and M_R have the same eigenvalues which are given by

$$\lambda_{1,2} = \bar{c} \pm \sqrt{\bar{c}^2 - 2\bar{u}(\bar{c} - \bar{u})} \geq 0,$$

since $\bar{u} < \bar{c}$ by assumption. Hence energy estimates are obtained for all cases. □

Remark 3.3. The choices in (24) make the boundary conditions for both the primal and dual equations relate to the characteristics of A by

$$\begin{aligned} H_L &= A + H_R^T = X_L^T \Lambda^+ X_L, \\ -H_R &= A - H_L^T = X_R^T \Lambda^- X_R, \end{aligned}$$

where $X_{L,R}$ are the normalized eigenvector matrices and

$$\Lambda^{\pm} = \frac{\Lambda \pm |\Lambda|}{2}$$

contains the positive and negative eigenvalues of A, respectively. See for example [18, 31].

3.4. Discretization, stability and spatial dual consistency

To discretize systems of equations, it is convenient to introduce the Kronecker product which is defined for arbitrary matrices C, D as

$$C \otimes D = \begin{bmatrix} C_{11}D & C_{12}D & \cdots & C_{1n}D \\ C_{21}D & C_{22}D & \cdots & C_{2n}D \\ \vdots & \ddots & \ddots & \vdots \\ C_{n1}D & C_{n2}D & \cdots & C_{nn}D \end{bmatrix}.$$

For the matrix inverse and transpose we have

$$(C \otimes D)^{-1,T} = C^{-1,T} \otimes D^{-1,T}$$

if the usual matrix inverses are defined. Furthermore, a useful property which will be extensively used, is the mixed product property,

$$(C \otimes D)(\tilde{C} \otimes \tilde{D}) = C\tilde{C} \otimes D\tilde{D},$$

if the usual matrix products are defined.

Using the Kronecker product, equation (3) with the boundary conditions (6) can be discretized using the SBP-SAT technique as

$$\begin{aligned} \frac{d}{dt}U_h + (D_1 \otimes A)U_h = (D_2 \otimes B)U_h \\ + (P^{-1}E_0 \otimes \Sigma_L)((I_{N+1} \otimes H_L)U_h - (D_1 \otimes B)U_h - G_L) \\ + (P^{-1}E_N \otimes \Sigma_R)((I_{N+1} \otimes H_R)U_h + (D_1 \otimes B)U_h - G_R), \end{aligned} \tag{25}$$

where $\Sigma_{L,R}$ are 2×2 matrices which have to be determined for stability. The second derivative is approximated using the wide operator, $D_2 = D_1 D_1 = (P^{-1}Q)^2$. The matrices $\Sigma_{L,R}$ are given in

Proposition 3.6. *The scheme (25) is energy stable by choosing*

$$\Sigma_L = \Sigma_R = -I. \tag{26}$$

Proof. We let $G_L = G_R = 0$ and apply the energy method to (25). By using the SBP properties of the operators, we obtain

$$\begin{aligned} ||U_h||_t^2 + 2||(D_1 \otimes B)U_h||^2 = U_h^T(E_0 \otimes (A + \Sigma_L H_L + H_L^T \Sigma_L^T))U_h \\ - U_h^T(E_N \otimes (A - \Sigma_R H_R - H_R^T \Sigma_R^T))U_h \\ - 2U_h^T(E_0 D_1 \otimes (B + \Sigma_L B))U_h \\ + 2U_h^T(E_N D_1 \otimes (B + \Sigma_R B))U_h. \end{aligned} \tag{27}$$

By choosing $\Sigma_L = \Sigma_R = -I$, equation (27) simplifies to

$$\begin{aligned} ||U_h||_t^2 + 2||(D_1 \otimes B)U_h||^2 = - U_h^T(E_0 \otimes (-A + H_L + H_L^T))U_h \\ - U_h^T(E_N \otimes (A + H_R + H_R^T))U_h, \end{aligned}$$

where, by construction in the continuous case,

$$-A + H_L + H_L^T \geq 0, \quad A + H_R + H_R^T \geq 0$$

according to propositions 3.1 and 3.3. Since the Kronecker product preserves positive (semi) definiteness, we have

$$\begin{aligned}||U_h||_t^2 + 2||(D_1 \otimes B)U_h||^2 = &- U_h^T(E_0 \otimes (-A + H_L + H_L^T))U_h \\ &- U_h^T(E_N \otimes (A + H_R + H_R^T))U_h \leq 0\end{aligned}$$

and the scheme is energy stable. □

The choices of penalty matrices in (26) renders the scheme not only energy stable, but also spatially dual consistent. To prove this we must show that the discrete adjoint operator L_h^* consistently approximates the continuous adjoint L^*, including the dual boundary conditions (10) and (19). This is done in

Proposition 3.7. *The scheme (25) is spatially dual consistent with the choices of $\Sigma_{L,R}$ given in (26).*

Proof. For $G_{L,R} = 0$, we can write the scheme (25) as

$$\frac{d}{dt}U_h + L_h U_h = 0, \tag{28}$$

where

$$\begin{aligned}L_h = &(D_1 \otimes A) - (D_2 \otimes B) \\ &+ (P^{-1}E_0 \otimes I_2)((I_{N+1} \otimes H_L) - (D_1 \otimes B)) \\ &+ (P^{-1}E_N \otimes I_2)((I_{N+1} \otimes H_R) + (D_1 \otimes B)).\end{aligned}$$

The discrete dual operator is defined by

$$L_h^* = (P \otimes I_2)^{-1}L_h^T(P \otimes I_2) \tag{29}$$

and a straight forward calculation shows that

$$\begin{aligned}L_h^* = &-(D_1 \otimes A) - (D_2 \otimes B) \\ &- (P^{-1}E_0 \otimes I_2)((I_{N+1} \otimes (A - H_L^T)) + (D_1 \otimes B)) \\ &+ (P^{-1}E_N \otimes I_2)((I_{N+1} \otimes (A + H_R^T)) + (D_1 \otimes B))\end{aligned}$$

which is a consistent approximation of (7) including the dual boundary conditions (10) and (19). The scheme is hence spatially dual consistent. □

Not only is the scheme dual consistent, the discrete dual scheme is also stable which is shown in

Proposition 3.8. *The discrete dual scheme*

$$\frac{d}{d\tau}\theta_h + L_h^*\theta_h = 0, \tag{30}$$

is stable with L_h^* given in (29).

Proof. The energy estimate of (30) is given by

$$\begin{aligned} ||\theta_h||_t^2 + 2||(D_1 \otimes B)\theta_h||^2 = &- \theta_h^T(E_0 \otimes (-A + H_L + H_L^T))\theta_h \\ &- \theta_h^T(E_N \otimes (A + H_R + H_R^T))\theta_h \leq 0, \end{aligned}$$

which is again stable by construction of the continuous boundary conditions. The discretization of the primal problem (28) is hence simultaneously a stable discretization of the dual problem (30). □

Remark 3.4. To obtain a stable and dual consistent scheme with the flux based boundary conditions, only one penalty parameter at each boundary is required. From the stability analysis both of them are determined uniquely as the identity matrix, which is also sufficient for spatial dual consistency. For Dirichlet boundary conditions, nine non-trivial penalty parameters were required, and the stability requirements were not sufficient for dual consistency.

4. Convergence and errors

A forcing function has been chosen such that an analytical solution is known, and the rates of convergence and errors are computed with respect to the analytical solution. This is known as the method of manufactured solutions. The solution in this case is given by

$$\begin{aligned} p(x,t) &= (\arctan(x) - \delta\cos(\alpha x - t) + 1)e^{-x^2}, \\ u(x,t) &= (\arctan(x) + \delta\sin(\alpha x - t) + 1)e^{-x^2}, \end{aligned}$$

and the functionals by

$$\begin{aligned} J(p) &= 1 + \frac{\pi}{4} - \frac{\log(2)}{2} + \frac{\delta(\sin(t-\alpha) - \sin(t))}{\alpha}, \\ J(u) &= 1 + \frac{\pi}{4} - \frac{\log(2)}{2} + \frac{\delta(\cos(t) - \cos(t-\alpha))}{\alpha}. \end{aligned}$$

Typical values of the parameters are

$$\bar{u} = 0.5, \quad \bar{c} = 1, \quad \varepsilon = 10^{-2}, \quad \delta = 0.1, \quad \alpha = 5\pi.$$

	Flux 1				Dirichlet			
N	64	96	128	160	64	96	128	160
3rd-order								
p	3.0530	3.0426	3.0377	3.0345	3.5301	3.3552	3.2357	3.1667
u	2.8870	2.9740	2.9973	3.0061	3.6191	3.5355	3.4484	3.3812
$J(p)$	2.5584	3.9209	4.2617	4.4285	3.9162	4.1626	4.2958	4.3643
$J(u)$	4.2536	4.2841	4.3698	4.4192	3.7781	3.5831	4.2249	4.5709
4th-order								
p	3.9728	4.2644	4.3975	4.4581	3.6293	4.3273	4.4387	4.4715
u	5.0736	4.6958	4.5338	4.4499	4.6610	5.0024	4.9220	4.7676
$J(p)$	2.2407	4.5614	5.5758	5.9743	4.0073	5.1505	5.4943	5.6757
$J(u)$	4.4065	5.6051	6.0071	6.2345	3.9695	4.3917	5.2755	5.5775
5th-order								
p	5.4375	5.2515	5.0597	4.9655	5.2057	5.7541	5.7028	5.5475
u	4.1043	5.0985	5.2485	5.2911	5.7376	5.1068	4.8237	4.8244
$J(p)$	4.9722	7.4842	7.8269	8.1507	6.0856	7.1729	7.7521	8.2018
$J(u)$	6.4334	7.1878	7.7660	8.1503	5.9507	6.9309	7.7971	8.2529

Table 1: Convergence rates for the variables and functionals using both flux based and Dirichlet boundary conditions

In Table 1 we present the numerical results regarding the order of convergence of the solution and functionals using both flux based and Dirichlet boundary conditions. The time integration is performed until time $t = 0.2$ with the classical 4th-order Runge-Kutta method using 1000 time steps. The convergence rates for the flux based boundary conditions (6) are not sensitive to the choice of the continuous parameters, and in the numerical examples that follow, we have chosen the marginal values

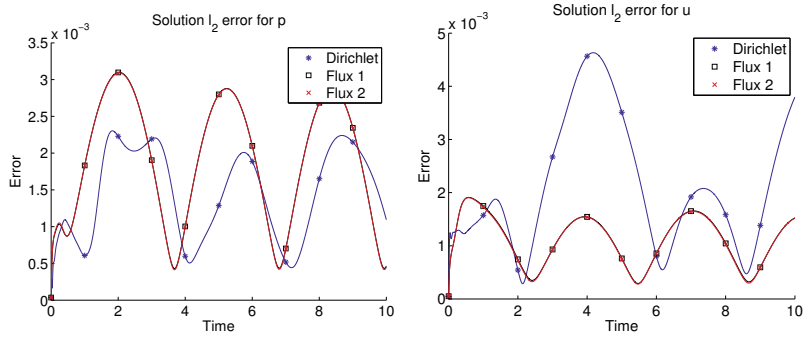
$$\alpha_L = \bar{c}, \quad \beta_L = \bar{u}/2, \quad \alpha_R = -\bar{c}, \quad \beta_R = -\bar{u}/2. \tag{31}$$

The choices in (24) give very similar results and are excluded.

As can be seen from Table 1, both schemes results in superconvergent functionals with similar rates of convergence. The error in the solutions and functionals as a function of time for 3rd-order and $N = 32$ grid points can be seen in Figure 1 and 2, respectively. We denote the values in (31) by "Flux 1" and the characteristic values in (24) by "Flux 2".

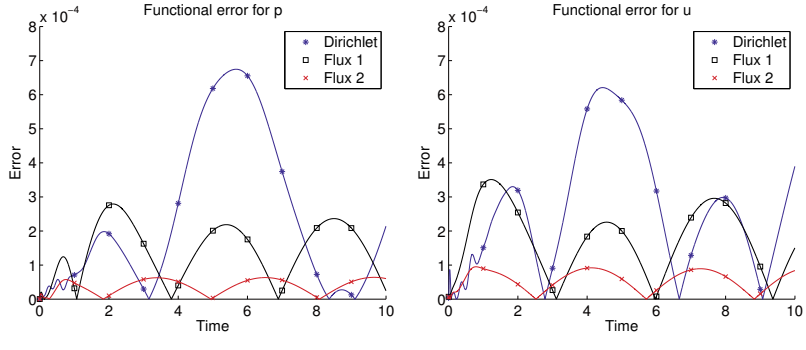
As we can see from Figure 1 and 2, in particular the characteristic flux based scheme (Flux 2) has significantly smaller functional errors.

In [21, 1] it was shown that schemes based on characteristic boundary conditions for hyperbolic problems, have non-growing errors in time – so called error bounded



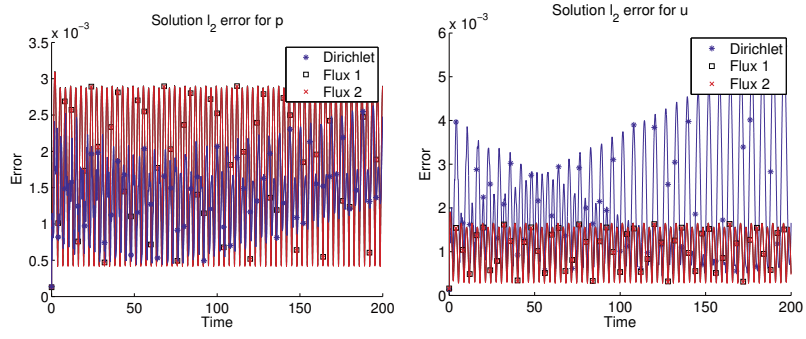
(a) Solution l_2 error for p (b) Solution l_2 error for u

Figure 1: Solution l_2 errors as a function of time using $N = 32$ grid points and 3rd-order for both the Dirichlet and the flux based boundary conditions



(a) Functional error for $J(p)$ (b) Functional error for $J(u)$

Figure 2: Functional errors as a function of time using $N = 32$ grid points and 3rd-order for both the Dirichlet and the flux based boundary conditions



(a) Solution l_2 error for p

(b) Solution l_2 error for u

Figure 3: Solution l_2 errors as a function of time using $N = 32$ grid points and 3rd-order for both the Dirichlet and the flux based boundary conditions using $\varepsilon = 10^{-6}$

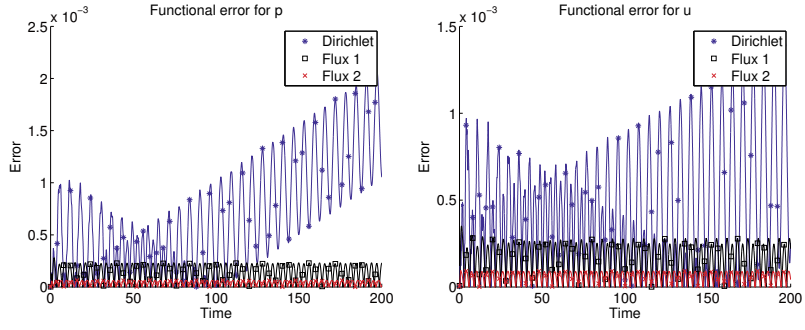
schemes. We can hence expect the discretization with Dirichlet boundary condition to have linearly growing errors in time for $\varepsilon << 1$, while the flux based boundary conditions are error bounded. Indeed, in Figures 3 and 4 we plot the error as a function of time for large times using $\varepsilon = 10^{-6}$. The Dirichlet boundary conditions give linearly growing errors while the flux based boundary conditions are error bounded.

With increasing ε, the discretization using the Dirichlet boundary conditions also becomes error bounded due to the parabolic effects. Since the characteristic choices in (24) converges uniformly to well-posed and stable boundary conditions for the hyperbolic system, we have an error bounded scheme for all values of ε, including $\varepsilon = 0$.

5. Spectral analysis

A consistent numerical scheme should have eigenvalues which converge to the continuous eigenvalues as the mesh is refined [23, 3]. The continuous spectrum of a PDE is obtained by Laplace transforms in time to reduce the PDE to an ordinary differential equation, and solve the corresponding Sturm-Liouville problem. Let

$$\hat{U} = \int_0^\infty e^{st} U dt$$



(a) Functional error for $J(p)$ (b) Functional error for $J(u)$

Figure 4: Functional errors as a function of time using $N = 32$ grid points and 3rd-order for both the Dirichlet and the flux based boundary conditions using $\varepsilon = 10^{-6}$

denote the Laplace transform of U. By ignoring the initial condition, as usual, and Laplace transforming (3), we obtain

$$s\hat{U} - A\hat{U}_x = B\hat{U}_{xx} \tag{32}$$

and the ansatz $\hat{U} = e^{kx}\Psi$ turns (32) into the eigenvalue problem

$$(sI + kA - k^2B)\Psi = 0.$$

The general solution to (32), assuming distinct roots, can be written as

$$\hat{U} = \sum_{i=1}^{3} \sigma_i e^{k_i}\Psi_i,$$

where k_i are the roots of the polynomial $\det(sI+kA-k^2B)$ and $\Psi_i = [\Psi_i(1), \Psi_i(2)]^T$ are the corresponding eigenvectors. Once the k_i and Ψ_i have been determined, the system of equations for determining σ_i can be written as

$$E(s)\sigma = 0$$

after the homogeneous boundary conditions have been applied. The continuous spectrum is then given by the values of s such that $\det(E(s)) = 0$. See [11] for

further details on the proceedure. For the Dirichlet boundary conditions we obtain the 3×3 matrix

$$E_D(s) = \begin{bmatrix} \Psi_1(1) & \Psi_2(1) & \Psi_3(1) \\ \Psi_1(2) & \Psi_2(2) & \Psi_3(2) \\ e^{k_1}\Psi_1(2) & e^{k_2}\Psi_2(2) & e^{k_3}\Psi_3(2) \end{bmatrix}.$$

The flux based boundary conditions yield the 4×3 matrix

$$E_F(s) = \begin{bmatrix} (H_L - k_1 B)\Psi_1 & (H_L - k_2 B)\Psi_2 & (H_L - k_3 B)\Psi_3 \\ (H_R + k_1 B)e^{k_1}\Psi_1 & (H_R + k_2 B)e^{k_2}\Psi_2 & (H_R + k_3 B)e^{k_3}\Psi_3 \end{bmatrix},$$

where the 3rd row is zero and hence $E_F(s)$ is condensed to a 3×3 matrix. Analytical solutions to $\det(E(s)) = 0$ can in general not be obtained due to the algebraic complexity. For scalar equations, analytical results are available in [3], while for more complicated equations, numerical methods have to be used. See [23] for more details.

Once $E_{D,F}(s)$ has been computed, eigenvalues s from the discrete spectrum can be inserted into the matrices to see whether or not a discrete eigenvalue actually belongs to the spectrum of the PDE. This technique can be used to verify convergence of discrete eigenvalues with certain properties. The semi-discretization of a linear system of equations can be written as

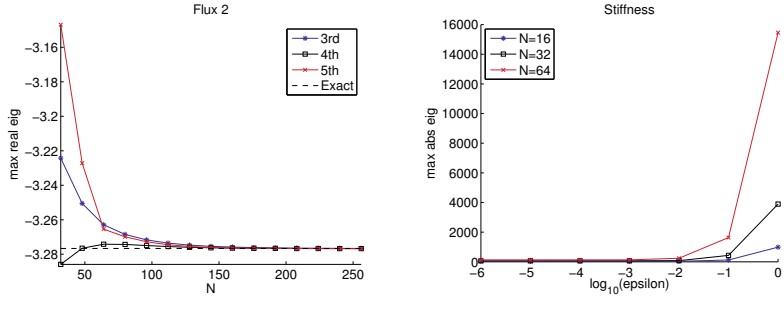
$$\frac{d}{dt} u_h = K u_h + f,$$

where the entire spatial discretization, including the boundary conditions, are included in the matrix K. The discrete spectrum can be modified and tuned for certain purposes depending on the boundary treatment.

The maximum real part and absolute value of the spectra are important since the first determines the convergence rate to steady-state [18, 23] while the second is a measure of the stiffness of the system. Both of these depend on the value of ε, which can be viewed as the viscosity. To see the effects, we show the maximum real part and absolute value in Table 2. For small values of ε, the characteristic flux based conditions, Flux 2, has significantly smaller real part of the spectrum. The difference is about a factor of 50–100 compared to Flux 1, and between 3 and 8 compared to Dirichlet. The spectrum of Flux 1 can, however, easily be shifted to the left in the complex plane by having strict inequalities in (24). The maximum absolute values are similar for small values of ε, while larger values forces the flux based conditions towards a much stiffer discretization.

	Maximum real part			Maximum absolute value		
ε	Dirichlet	Flux 1	Flux 2	Dirichlet	Flux 1	Flux 2
10^{-6}	-0.412	-0.029	-1.515	34.4	34.4	32.1
10^{-5}	-0.412	-0.029	-1.517	34.4	34.4	32.1
10^{-4}	-0.412	-0.029	-1.539	34.4	34.4	32.1
10^{-3}	-0.412	-0.029	-1.753	34.4	34.5	32.1
0.01	-0.415	-0.029	-3.158	34.5	34.9	32.1
0.1	-0.463	-0.030	-1.492	35.7	85.0	121.0
1	-0.537	-0.027	-0.498	45.6	961.3	987.0

Table 2: Maximum real part and absolute values of the spectras using 3rd order and $N = 16$



(a) Maximum real part, all orders of accuracy (b) Maximum absolute value, 3rd-order

Figure 5: Convergence of maximum real part and absolute value for the Flux 2 boundary conditions

The continuous eigenvalue with real part closest to zero can be computed from the determinants of $E_{D,F}(s)$. As an example, we show the convergence of Flux 2 towards this eigenvalue in Figure 5(a). In Figure 5(b) we show the increase in stiffness for different values of ε as the mesh is refined.

6. Conclusions

New flux based boundary conditions for a linear incompletely parabolic system of equations have been derived. The boundary conditions are constructed such that both the primal and dual problems are well-posed. Depending on parameter variations in the new boundary conditions, choices can be made to either provide well-posedness independently of sub- or supersonic conditions, or such that convergence to the hyperbolic system is ensured for the subsonic case.

The numerical scheme based on the new boundary conditions can be constructed to be both energy stable and dual consistent. Compared to a discretization using standard Dirichlet boundary conditions, the new scheme is significantly simpler since reduction to first order form, and additional penalty parameters, can be avoided.

The solutions and functionals computed using the flux based boundary conditions are more accurate and have better damping properties. Long time computations showed that the new scheme can provide error boundedness independently of the amount of viscosity, even in the hyperbolic limit.

Eigenvalue computations showed that the maximum real part of the discrete spectrum converges to the analytical value. The flux based boundary conditions can provide smaller real parts than a discretization with Dirichlet boundary conditions which is beneficial for steady-state computations. The stiffness is, however, increased for large viscosity.

The analysis of this model problem will be applied to the more complicated compressible Navier–Stokes equations in future work.

References

[1] S. Abarbanel, A. Ditkowski, and B. Gustafsson. On error bounds of finite difference approximations to partial differential equations — temporal behavior and rate of convergence. *Journal of Scientific Computing*, 15:79–116, 2000.

[2] S. Abarbanel and D. Gottlieb. Optimal time splitting for two- and three-dimensional Navier–Stokes equations with mixed derivatives. *Journal of Computational Physics*, 41(1):1–33, 1981.

[3] J. Berg and J. Nordström. Spectral analysis of the continuous and discretized heat and advection equation on single and multiple domains. *Applied Numerical Mathematics*, 62(11):1620–1638, 2012.

[4] J. Berg and J. Nordström. Superconvergent functional output for time-dependent problems using finite differences on summation-by-parts form. *Journal of Computational Physics*, 231(20):6846–6860, 2012.

[5] M. H. Carpenter, J. Nordström, and D. Gottlieb. A stable and conservative interface treatment of arbitrary spatial accuracy. *Journal of Computational Physics*, 148(2):341–365, 1999.

[6] K. J. Fidkowski and D. L. Darmofal. Review of output-based error estimation and mesh adaptation in computational fluid dynamics. *AIAA Journal*, 49(4):673–694, 2011.

[7] M. B. Giles, M. G. Larson, J. M. Levenstam, and E. Süli. Adaptive error control for finite element approximations of the lift and drag coefficients in viscous flow. Technical report, Report NA-97/06, Oxford University Computing Laboratory, 1997.

[8] M. B. Giles and N. A. Pierce. An introduction to the adjoint approach to design. *Flow, Turbulence and Combustion*, 65:393–415, 2000.

[9] M. B. Giles and N. A. Pierce. Superconvergent lift estimates through adjoint error analysis. *Innovative Methods for Numerical Solutions of Partial Differential Equations*, 2001.

[10] M. B. Giles and E. Süli. Adjoint methods for PDEs: A posteriori error analysis and postprocessing by duality. *Acta Numerica*, 11:145–236, 2002.

[11] B. Gustafsson, H.-O. Kreiss, and J. Oliger. *Time Dependent Problems and Difference Methods*. Wiley Interscience, 1995.

[12] J. E. Hicken. Output error estimation for summation-by-parts finite-difference schemes. *Journal of Computational Physics*, 231(9):3828–3848, 2012.

[13] J. E. Hicken and D. W. Zingg. Superconvergent functional estimates from summation-by-parts finite-difference discretizations. *SIAM Journal on Scientific Computing*, 33(2):893–922, 2011.

[14] J. E. Hicken and D. W. Zingg. Summation-by-parts operators and high-order quadrature. *Journal of Computational and Applied Mathematics*, 237(1):111–125, 2013.

[15] A. Jameson. Aerodynamic design via control theory. *Journal of Scientific Computing*, 3:233–260, September 1988.

[16] K. Mattsson. Summation by parts operators for finite difference approximations of second-derivatives with variable coefficients. *Journal of Scientific Computing*, pages 1–33, 2011.

[17] K. Mattsson and J. Nordström. Summation by parts operators for finite difference approximations of second derivatives. *Journal of Computational Physics*, 199(2):503–540, 2004.

[18] J. Nordström. The influence of open boundary conditions on the convergence to steady state for the Navier–Stokes equations. *Journal of Computational Physics*, 85(1):210–244, 1989.

[19] J. Nordström. The use of characteristic boundary conditions for the Navier–Stokes equations. *Computers & Fluids*, 24(5):609–623, 1995.

[20] J. Nordström. Conservative finite difference formulations, variable coefficients, energy estimates and artificial dissipation. *Journal of Scientific Computing*, 29:375–404, 2006.

[21] J. Nordström. Error bounded schemes for time-dependent hyperbolic problems. *SIAM Journal on Scientific Computing*, 30(1):46–59, 2008.

[22] J. Nordström and M. H. Carpenter. High-order finite difference methods, multidimensional linear problems, and curvilinear coordinates. *Journal of Computational Physics*, 173(1):149–174, 2001.

[23] J. Nordström, S. Eriksson, and P. Eliasson. Weak and strong wall boundary procedures and convergence to steady-state of the Navier–Stokes equations. *Journal of Computational Physics*, 231(14):4867–4884, 2012.

[24] J. Nordström and M. Svärd. Well-posed boundary conditions for the Navier–Stokes equations. *SIAM Journal on Numerical Analysis*, 43(3):1231–1255, 2005.

[25] N. A. Pierce and M. B. Giles. Adjoint recovery of superconvergent functionals from PDE approximations. *SIAM Review*, 42(2):247–264, 2000.

[26] N. A. Pierce and M. B. Giles. Adjoint and defect error bounding and correction for functional estimates. *Journal of Computational Physics*, 200:769–794, November 2004.

[27] O. Pironneau. On optimum design in fluid mechanics. *Journal of Fluid Mechanics*, 64(01):97–110, 1974.

[28] B. Strand. Summation by parts for finite difference approximations for d/dx. *Journal of Computational Physics*, 110(1):47–67, 1994.

[29] J. C. Strikwerda. Initial boundary value problems for incompletely parabolic systems. *Communications on Pure and Applied Mathematics*, 30(6):797–822, 1977.

[30] M. Svärd. On coordinate transformations for summation-by-parts operators. *Journal of Scientific Computing*, 20:29–42, 2004.

[31] M. Svärd, M. H. Carpenter, and J. Nordström. A stable high-order finite difference scheme for the compressible Navier–Stokes equations, far-field boundary conditions. *Journal of Computational Physics*, 225(1):1020–1038, 2007.

[32] M. Svärd, K. Mattsson, and J. Nordström. Steady-state computations using summation-by-parts operators. *Journal of Scientific Computing*, 24(1):79–95, 2005.

[33] M. Svärd and J. Nordström. On the order of accuracy for difference approximations of initial-boundary value problems. *Journal of Computational Physics*, 218(1):333–352, 2006.

[34] D. A. Venditti and D. L. Darmofal. Adjoint error estimation and grid adaptation for functional outputs: Application to quasi-one-dimensional flow. *Journal of Computational Physics*, 164(1):204–227, 2000.

[35] D. A. Venditti and D. L. Darmofal. Anisotropic grid adaptation for functional outputs: application to two-dimensional viscous flows. *Journal of Computational Physics*, 187(1):22–46, 2003.