Technische Universität München
Zentrum Mathematik

Modeling the anomalous heat transport in a tokamak plasma

Discontinuity in the derivative of the heat conductivity coefficient

Katya Dimova

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.) genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. Peter Rentrop
Prüfer der Dissertation:	
	1. Priv.-Doz. Dr. Rita Meyer-Spasche
	2. Univ.-Prof. Dr. Folkmar Bornemann
	3. Prof. Dr. Harold Weitzner, New York Univ. / USA (schriftliche Beurteilung)

Die Dissertation wurde am 19.01.2006 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 23.05.2006 angenommen.

Acknowledgements

Most of all I need to thank my advisor, Priv.-Doz. Dr. Rita-Meyer Spasche, for suggesting me the Ph.D work and for her guidance in the last four years. Next, I would like to thank Prof. Weitzner, director of Magnetofluid Dynamics Division at Courant Institute, for his expert opinion and advice, and acknowledge his input and help. I am grateful to Prof. Stefka Dimova, from the Sofia University, who was besides me whenever I need her with always helpful commentaries.

Special thanks to Prof. Pereverzev, who provide me with the interesting topic, for the helpful comments and discussions. I like to thank Dr. Brambilla and Prof. Lackner for their remarks in improving my thesis, especially from the physical point of view. I am grateful to Prof. Guenter for her understanding and support during my PhD. Thank to Prof. Olof Widlund, from Courant Institute, for the mathematical discussion we had during my three months visit at the New York University.

Last, but not least, I like to thank my family for their continuous support and love all this years. Special appreciation to my sister for her help on improving my English and wishes for success in her PhD in bio-statistics. Some of the special appreciations I address to my friends Monica, Zdenka, Dorina, Alina, Yu-ying and ..., I missed somebody and I am deeply sorry for this. Finally, a very special thanks to Dimitris not only for being besides me but also for the useful discussions we had.

Munich,
18.01.2006 Katya Dimova

Zusammenfassung

Einer der Hauptgründe, warum bei der Kernfusion bis jetzt noch keine Energie gewonnen werden konnte, ist der *anomale Wärmetransport* (anomalous heat transport) im Plasma des Tokamaks. Der Begriff “anomal” gründet in den unerwartet großen Koeffizienten, welche für den Energietransport verantwortlich sind und bei den Experimenten beobachtet werden. Dieser zusätzliche Transport tritt immer dann ein, wenn der Absolutwert des Temperaturgradienten einen bestimmten Schwellwert überschreitet.

Das mathematische Modell entspricht einer nicht-standard Wärmeleitungsgleichung mit einem bezüglich dem Wärmefluß unstetigen Temperaturkoeffizienten. Um den Bereich zu ermitteln, in dem sich der anomale Transport auszuwirken beginnt, entwickeln wir ein explizites *front tracking* Verfahren. Erreicht der Gradient der Lösung an einer Stelle (*front point*) einen vorgegebenen Schwellwert, so wird die Differentialgleichung an diese Stelle in zwei Teilprobleme aufgespalten. Wir zeigen, daß die Teilprobleme separat behandelt werden können, und daß ihre Lösungen sich zu einer C^1 -Funktion zusammensetzen lassen. Zur Bestimmung der Position der inneren Grenze (*front point*) verwenden wir eine zusätzliche gewöhnliche Differentialgleichung. Wir benutzen für die räumliche Diskretisierung eine Finite-Element-Methode. Für die Diskretisierung in der Zeit wenden wir einen Algorithmus an, der von uns entwickelt ist. Er basiert auf der Trapezregel, wobei der Defekt vierter Ordnung des Lobatto III A Schemas zur Optimierung der Anzahl der Newton-Iterationen verwendet wird. Desweiteren wird der Zeitschritt adaptiv angepaßt und der lokale und globale Fehler abgeschätzt. Die numerische Leistungsfähigkeit des Algorithmus wird an mehreren Beispielen demonstriert. Das *anomalous transport* Problem wird für einen großen Parameterbereich gelöst. Die hohe Genauigkeit des Algorithmus wird durch Vergleich des genäherten Gradienten mit der exakten Lösung im stationären Fall gezeigt. Darüber hinaus wird die Behandlung mehrerer *front points* vorgeführt. Dies entspricht in der Wärmeleitung einer turbulenten Region.

Abstract

One of the main problems in fusion research is the understanding of the dynamics governing the heat transport in a tokamak plasma. Because of unexpectedly large transport coefficients observed in experiments the transport is called "anomalous". This property is one of the main reasons why there is no energy produced by fusion yet.

Mathematically, the anomalous heat transport problem is modelled by a nonstandard heat equation, with a heat conductivity coefficient depending on the gradient of the solution in a piecewise differentiable way with a jump discontinuity. In order to detect precisely the region where the anomalous transport starts playing a role, we develop an explicit *front tracking technique*. The differential equation is split at the discontinuity points (front points) into sub-problems. We prove that each of the problems can be treated separately and that their solutions match continuously at the inner boundary. To find the position of the inner boundary (the front point), we solve an additional ordinary differential equation. Numerically, we use the finite element method for the space discretization. For the time discretization, we apply a newly developed algorithm. It is based on the trapezoidal rule, but uses the defect of the 4^{th} order Lobatto III A scheme for optimization of the number of the Newton iterations, adaptation of the time step and estimation of the local and global errors. The numerical capabilities of the algorithm are demonstrated on several numerical examples. The anomalous transport problem is solved and the parameter space is explored. The high accuracy of the algorithm is shown comparing the numerical and analytical results at the stationary state. Moreover, the treatment of multiple front points, which correspond to turbulence regime in the heat transport, is performed.

Contents

Chapter 1

Introduction

There is currently considerable interest in the modelling and numerical solution of reaction diffusion problems arising in many applications areas of the physical sciences. One important class, representative of the complexities which can arise, involves problems in plasma physics, particularly the 'anomalous' heat transport in a tokamak plasma. The resulting mathematical models are typically nonlinear reaction diffusion partial differential equations which pose a number of analytical and computational challenges.

The idea of fusion is to build a system by which light particles undergo fusion reactions with an energy gain, and then use that energy gain as part of the energy supply of an industrial economy. One possible way is through the creation and control of a thermonuclear plasma. Due to a high temperature the molecules of the gas are decomposed into atoms and the atoms are then decomposed into electrons and positively charged ions. The degree of ionization increases as the temperature rises. The ionized gas formed in this way is called high-temperature plasma. It consists of high number of positively charged heavy ions and negatively charged light electrons.

The most investigated and furthest advanced configuration for the magnetic cage of a fusion plasma is the tokamak. "Tokamak" is a generic name for axisymmetric, toroidal, magnetic confinement devices used to produce high temperature plasmas and to stably confine them by means of a strong magnetic field (see fig. 1.1). The basic idea of the tokamak is a torus of plasma, confined by a magnetic field. Toroidal and Poloidal Coils establish strong field along the torus, which is needed to keep the system stable against disturbances. A current flows along the torus to produce a magnetic field with helical force lines, and in this way confine the plasma. However, there are still some disturbances which lead to turbulence in the plasma. One of them is called "anomalous transport". It is a major difficulty on the road to realize controlled fusion.

From the above it is clear that the investigation of the mechanisms governing the heat transport is one of the main considerations in fusion research (for example [23], [37]). One simple description of it is the following one. The current

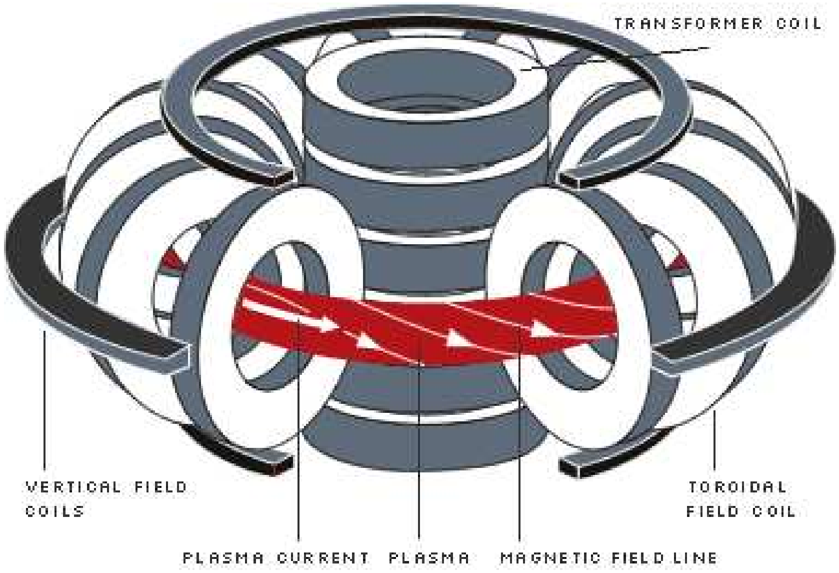

Figure 1.1: Tokamak configuration

in the plasma creates a magnetic field with helical lines forming magnetic surfaces. The magnetic surfaces are nested in each other (see fig 1.2). The charged

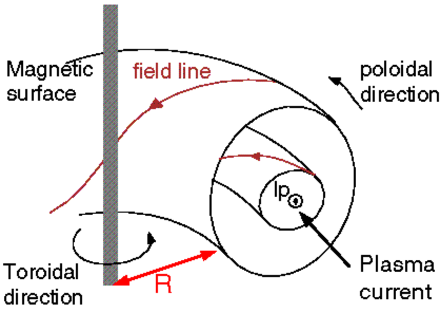

Figure 1.2: a cross-section of a tokamak

particles, in a homogeneous magnetic field, move around magnetic lines realizing a free gyration motion in a plane. To a first approximation, particles move along magnetic field lines with constant energy. Collisions, however, make them cross from a magnetic field line to another. This leads to energy transport across magnetic field surfaces (this fact is reflected in the one dimensionality of the model describing the heat transport in the tokamak plasma). Particles finally escape when they reach the walls of the tokamak. This is the so called classical transport theory. In reality, the magnetic field is not homogeneous, which gave rise to "neoclassical transport theory". As a consequence of this one anticipates transport which would be a factor of 10 times classical. The actually observed transport, however, is much greater. Because of this unexpected property, the transport of heat and particles in tokamaks is called ”anomalous”.

One explanation of this anomalous behaviour, in general terms, is that the plasma is subject to instabilities. It turns out that plasma is always quivering with ”fluctuations” in all its parameters: density, temperature and magnetic field. Our attention is focused on the anomalous transport caused by the temperature gradients. In the tokamak experiments it is observed a strong dependence of the heat conductivity coefficient on the gradient of the temperature, but only when it exceeds a particular critical value (fig 1.3).

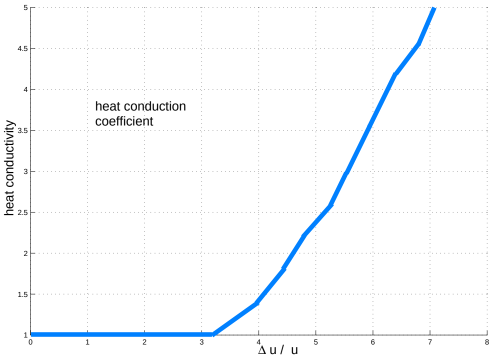

Figure 1.3: Measured heat conductivity coefficient against the temperature gradient

Since the model is still in an examination stage, there is a need of accurate numerical solution. This will give additional information on the heat transport in the plasma and improve our understanding of the anomalous transport. Mathematically, the anomalous heat transport problem is a non-standard problem, with a discontinuous dependence of the derivative of the heat conductivity coefficient on the temperature gradient.

Before a particular mathematical model is to be used, its correctness from mathematical point of view has to be proved. This includes verification of properties like existence, uniqueness and regularity of the solutions of the respective PDEs. Due to the complexity of such problems, they are solved numerically, and not so much is known about their exact solutions. The progress in computational capabilities allows more and more physics to be included into the mathematical models. Apart from a few situations where mathematical analysis can actually be applied, the numerical analysis of PDEs is the main tool to assess the modelling process for large number of physical problems. Successful adaptive methods lead to considerable savings in computational work. In fact, a posteriori error estimates can be used to judge the quality of a numerical approximation and to determine an adaptive strategy to improve the accuracy where needed. Our aim is to analyze and to design an adaptive algorithm which includes an a posteriori estimate of the numerical error, and solves nonlinear parabolic problems with solution dependent operators of type

$$u_t = \partial_x(D(t, x, u, |u_x|)u_x) + S(x).$$

Nowadays, there is a number of different adaptive techniques which numerically solve PDEs. One possibility is to discretize simultaneously in space and time using a discontinuous Galerkin method and to apply coupled space-time estimates (e.g. Eriksson, Johnson, Thomee [10]). The discretization, first in time then in space, known as Rothe's method, provides us with an other possibility for treating PDE. Interpreting the time-dependent PDE as an ODE in a Hilbert space, the temporal error can be estimated by classical ODE-procedures (e.g. Lang [27], Bornemann [2]). Another adaptive method is the method of Moving Finite Elements, which uses mesh points that automatically move in the space-time domain (e.g. Cao, Huang, Russell [4], Budd, Carretero-Gonzalez, Russell [3]). In this thesis, we use the method of lines (MOL) for discretizing the underlying problem (e.g. Schiesser [42], Thomee [47]). It consists of discretization in space, which leads to a transformation of the time-dependent PDE into a system of ODEs, that is solvable by an appropriate variable step-size time integrator.

It is well-known that differential operators give rise to infinite stiffness. Therefore, often an implicit discretization method coupled with a Newton-like iteration is applied to integrate in time. The classical Newton method is still the most widely used approach. It consists of an iterative procedure which goes on until the difference between two iterations becomes sufficiently small and/or the defect of the numerical method (a property of the approximate solution depending on the numerical method used). Another idea, which has grown to a class of linearly implicit methods like Rosenbrock methods [40], is that one Newton iteration should be enough to integrate stiff problems efficiently [6].

In this respect, we developed a new approach compromising between the just mentioned two through adaptation of the number of the Newton iterations. The idea is to minimize the defect of the differential equation rather than the defect of the difference scheme by reducing (within a given tolerance) the defect of a numerical method of higher order. This technique offers several advantages. It controls not only the number of the Newton iterations, but also provides an efficient computational estimate of the local numerical error. As is known, a fundamental property of the stable one-step integration method is that the global error consists of propagated and accumulated local truncation error. Thus, controlling the local errors of each individual time step, we control the global error. In addition, the local error estimate is used for an automatic adaptation of the time step.

This approach is incorporated into a numerical method, which we call AIM - Adaptive Implicit Method, for numerical treatment of nonlinear parabolic problems with solution-dependent operators, and it is part of this thesis. The method comprises the finite elements in space coupled with the implicit Trapezoidal rule in time. The defect of the 4^{th} order Lobatto III A method is used for adaptation of the number of the Newton iterations and estimation of the local error. We refer to this estimate as the Lobatto estimate. The detailed description of it is given in Chapter 2. A Lemma showing the connection between the local error and the defect of the 4^{th} order Lobatto III A scheme is proved. As a consequence, we show the relation between the defect of the 4^{th} order Lobatto III A scheme and the global error of the method used. This is derived at first for a single ODE and then for a system. Further, in the same chapter, we apply the Lobatto estimate of the local, respectively of the global error, to a quasilinear parabolic problem.

In Chapter 3 we consider a mathematical model describing the anomalous heat transport in a tokamak plasma. Because of the axisymmetry of the tokamak and the plane nature of the particle motion the model is a one dimensional radially symmetric quasilinear parabolic problem

$$\begin{aligned}
\frac{\partial u}{\partial t} &= \frac{1}{x^{d-1}}\frac{\partial}{\partial x}(x^{d-1}(D_0 + D_1 H(|u_x| - \bar{u})(|u_x| - \bar{u}))\frac{\partial u}{\partial x}) + S(t,x), \\
&\qquad\qquad \text{for} \quad x \in (0,1), \ t > 0 \\
u(0,x) &= u_0(x), \quad x \in [0,1] \\
\left.\frac{\partial u}{\partial x}\right|_{x=0} &= 0, \quad u(t,1) = u_1(t), \qquad t > 0,
\end{aligned}$$

where u represents typically the temperature, $D := (x^{d-1}(D_0 + D_1 H(|u_x| - \bar{u})(|u_x| - \bar{u}))$ is the heat conductivity coefficient, D_0, D_1, $\bar{u}$ are positive constants, and $S(t,x) \geq 0$ is a source which could be present or not. The function $H(x)$ is the Heaviside function which takes values 0 and 1. The parameter $\bar{u}$ represents the critical value for the temperature gradient. The natural value for the parameter d is 2 - cylindrical geometry, but for completeness we consider also $d = 1$ - slab geometry and $d = 3$ - spherical geometry. Finally, the flux is defined as the product of the heat conductivity coefficient and the temperature gradient.

In principle, this problem can be handled numerically by working with the equation in integral form w.r.t. space. This leads to conservative finite difference schemes and thus avoids assumptions about the differentiability of the solution. Instead of finite difference scheme, the Finite Element Method or the Boundary Element Method could be applied. In practice, however, this approach is seriously limited when a sharp interface is present. A high order numerical method may lead to numerical oscillations around the front position and that may couple into other parts of the solution. For lower order method, excessive numerical diffusion may rapidly destroy the sharpness of the front.

The jump in the first derivative of the flux suggests that the front tracking technique could be the proper way of treating this problem. Interface treatment commonly uses one of three basic schemes. These are explicit tracking method (Tryggvason 2005 [48], Glimm 2001 [16], Unverdii 1992 [49]), level set methods that propagate fronts by evolving a level set function whose zero set corresponds to the front location at a given time (Sethian 1999 [43], Osher 2000 [38]), and volume of fluid methods or interface reconstruction methods (Lopez 2004 [1], Benson 2002 [32]). Direct interface treatments (explicit front tracking or level sets) maintain an explicit representation of the interface, either as a geometric front, or as a level set function, and treat the individual fluid species as separate quantities (densities, temperatures, and tangential velocities) that are in equilibrium (pressure and normal velocity) across the interface. In the volume of fluid (VOF) method the interface evolution is described using a discrete function, F, whose value in each element of the computational mesh in single fluid problems is the fraction of the element occupied by the fluid. This volume fraction is a discretized version of a function, f, which is continuous everywhere except at the interface and satisfies a standard advection equation. At any time step, the interface is at first "reconstructed" at each element from the F distribution, and then it is advected solving the standard advection equation by means of using geometric considerations to compute volume fluxes through element boundaries.

For solving the anomalous transport problem, we preferred the explicit front tracking technique (FTT) which avoids completely some inaccuracies in tracking the interface encountered by the other methods. The differential equation is split into two subproblems at the point where the gradient of the solution reaches the threshold. We refer to the splitting point as a front point or interface. Each of the subproblems is solved separately through AIM (with linear and Hermitian finite elements) on a grid that partially varies from one time step to another. In this approach the majority of the grid elements do not vary. We have a fixed uniform underlying grid which is sufficient to represent the solution in smooth regions (away from the front position). Additional elements are introduced at the locations of the front through a subdivision of some regular elements into two subelements. By having this irregular sub-mesh which is moving together with the interface, we avoid the smearing and loss of accuracy which is inevitable when the front position falls within a grid element. At the same time, by keeping the underlying grid uniform, we avoid interpolation of the solution in the whole interval and interpolate only at the irregular grid. Employing the continuity of the flux over the interface, we derive an equation for the speed of the interface. This information is used for the choice of the sub-irregular grid at the next time step. The implementation of AIM and the Front Tracking Technique (FTT) is based mainly on the programming language **C**. The implementation of Hermitian basis is done in the programming language **FORTRAN**, and this module is incorporated in the **C** routine.

The description of the numerical technique, in Chapter 3, is preceded by the analytical investigation of the equivalence of the entire problem to the obtained two subproblems. The proof of this statement is based mainly on the work of Ladyzhenskaia [25]. Some exact solutions are found for a problem without source. A detailed convergence analysis of the front tracking technique is performed.

In Chapter 4, some computational results demonstrating the theoretically derived error estimate, and the analyzed numerical technique are presented. Some examples illustrating the efficiency of the Lobatto estimate are given. In this respect, the first example covers the behaviour of the Lobatto estimate for a system of ODEs, whereas the second one deals with a thermonuclear combustion model with electronic heat turbulence.

In a separate section are given some of the numerical results from the application of AIM and the Front Tracking Technique to the anomalous transport in the plasma. Verification of the developed front tracking technique for the anomalous transport is performed on one of the problems with analytical solution. Numerical investigation of the order of convergence is carried out for linear FEM, as well as for Hermitian elements. A comparison between FTT and the non front tracking is given based on the rate of energy conservation. In order to demonstrate the capabilities of FTT, a parameter investigation for physically relevant values is performed. The behaviour of the solution of the anomalous transport problem approaching steady state is examined. Finally, in the Appendixes, we give a collection of helpful definitions, propositions and theorems which are used throughout the thesis.

Chapter 2

Quasilinear parabolic problems. Error estimates

The mathematical model describing the anomalous heat transport consists of a quasilinear parabolic equation describing the propagation of the heat. For temperature gradient smaller than a certain threshold the evolution of the temperature is described by a linear heat conductivity equation. Above the threshold the heat diffuses in a nonlinear manner. Because of this switch between the two regimes, the problem possesses no classical solution, but it is rather “nicely” defined in each of the regimes. In order to take advantage of this “nicely” property of the problem we split it at the switch point. In this way we get two sub-problems, each of which is a well defined quasilinear parabolic problem that can be treated by the existing theory. In the second section, we proceed with the numerical discretization. The method used is the method of lines, i.e. the problem is first discretized in space by the finite element method (FEM) and then the time discretization is done with the help of a second order implicit scheme - trapezoidal rule. Implicit scheme leads to an implicit system of algebraic equations which we solve through the Newton method. It is standard to choose the number of the Newton iterations such that the difference between two iterations becomes small and/or the defect. Instead, we minimize the defect of the differential equation rather than the defect of the difference scheme. Using the same idea we succeed to estimate the local and global errors of approximation in a nonstandard way. This algorithm, initially developed for a system of ordinary differential equations, is naturally extended to quasilinear parabolic differential equations. This is given in the fifth section of this chapter.

2.1 Continuous problem. Notations

Now we proceed with some facts about quasilinear parabolic equations mostly based on the classical work of Ladyzhenskaia [25] back to 1968. Some more recent works in this field are the review-work of Lieberman 1996 [31], a sequence of papers from Lederman, Vazquez and Wolanski for the case of semilinear parabolic problems, for example [29] from 2001. In the light of analytic semigroup is the book of Alessandra Lunardi, [34], from 1995 which gives regularity conditions for existence and uniqueness of solutions for general nonlinear parabolic problems.

Let W be a real Banach space with norm $||.||$, and let $D \subset \mathbb{R}$ be an interval. As usual, $C^r(D)$ is the space of all functions $f : D \to W$ that are r times continuously differentiable, $L_2(D)$ is the space of all square integrable functions, and $W_2^r(D)$ is the Sobolev space with r times square integrable derivatives. The corresponding norms for the above spaces are

$$\begin{aligned} ||f||_\infty &:= \sup_D ||f|| && (2.1)\\ ||f|| &:= \left(\int_D f(x)^2 dx \right)^{1/2} && (2.2)\\ ||f||_r &:= (||f||^2 + \sum_{j=1}^{r} ||\mathcal{D}^j f||^2)^{1/2} && (2.3) \end{aligned}$$

where $\mathcal{D}^j f$ stands for the j^{th} generalized derivative of f. We make use of the Hölder spaces which are defined for $r \in \mathbb{N}$ and $\alpha \in (0,1)$ as follows

$$C^\alpha(D) = \left\{ f \in C(D), \text{bounded} : ||f||_{C^\alpha(D)} < +\infty \right\}$$

for

$$\begin{aligned} [f]_{C^\alpha(D)} &:= \sup_{t,s \in D, s<t} \frac{||f(t) - f(s)||}{(t-s)^\alpha} \\ ||f||_{C^\alpha(D)} &:= \sup_D ||f|| + [f]_{C^\alpha(D)} \end{aligned}$$

$$C^{r+\alpha}(D) = \left\{ f \in C^r(D), \text{bounded} : f^{(r)} \in C^\alpha(D) \right\}$$

for

$$||f||_{C^{r+\alpha}(D)} := ||f||_{C^r(D)} + [f^{(r)}]_{C^\alpha(D)}.$$

For functions $f : [a,b] \times D \to W$ we introduce the Hölder spaces $C^{\alpha,0}(D)$ and $C^{0,\alpha}(D)$ as

$$\begin{aligned} C^{\alpha,0}([a,b] \times D) \;=\; & \{ f \in C([a,b] \times D), \text{bounded} : f(.,x) \in C^\alpha([a,b]) \; \forall x \in D, \\ & ||f||_{C^{\alpha,0}} := \sup_{x \in D} ||f(.,x)||_{C^\alpha([a,b])} < +\infty \; \} \end{aligned}$$

and similarly

$$\begin{aligned} C^{0,\alpha}([a,b]\times D) \;&=\; \{f \in C([a,b]\times D), \text{bounded} : f(t,.) \in C^{\alpha}(D)\ \forall t \in [a,b], \\ &\qquad ||f||_{C^{0,\alpha}} := \sup_{t\in[a,b]} ||f(t,.)||_{C^{\alpha}(D)} < +\infty\,\}\,. \end{aligned}$$

Let $D \subset \mathbb{R}$ be an interval, and ∂D be the boundary of D. Let $\Omega_T = (0,T)\times D$, $T < \infty$ with boundary $\partial\Omega_T$. The notation $\bar{\Omega}_T$ stands for denoting the closure of Ω_T, so that $\bar{\Omega}_T = \Omega_T \cup \partial\Omega_T$.

We consider in $\bar{\Omega}_T$ a general one dimensional quasilinear parabolic problem with boundary condition of first kind that can be written as

$$\begin{aligned} \frac{\partial u(t,x)}{\partial t} &= a_{11}(t,x,u,|u_x|)u_{xx} + b(t,x), \qquad x \in \Omega_T, \qquad (2.4)\\ u(0,x) &= u_0(x), \qquad x \in \bar{D}, \\ u(t,x) &= \psi(x), \qquad t \geq 0,\ x \in \partial D. \end{aligned}$$

The required regularity conditions for the coefficients a_{11}, b and for the functions $u_0,\ \psi$ are

Regularity Conditions (RC 1.):
Let $M,\ M_1 > 0$ be constants, $\beta \in (0,1)$ and

a) $a_{11} = a_{11}(t,x,u,p) \in C^{\beta/2,\ \beta,\beta,\beta}(\bar{\Omega}_T \times \mathbb{R}^+ \times \mathbb{R}^+)$ *and* $a_{11}(t,x,u,0) \geq 0$ *,* *for* $(x,t) \in \bar{\Omega}_T,\ |u| \leq M$ *and* $0 \leq p \leq M_1$*,*

b) $b \in C^{\beta/2,\beta}(\bar{\Omega}_T)$ *and* $-b(t,x)u \leq |u|\Phi(|u|)$*,* *where* $|u| \leq M$ *and* $\Phi(\tau) > 0$ *is a nondecreasing function of* $\tau \geq 0$ *satisfying* $\int\limits_1^{\infty} \frac{d\tau}{\Phi(\tau)} = \infty$*,*

c)

$$\nu(1+|p|)^{m-2} \leq a_{11}(t,x,u,p) \leq \mu(1+|p|)^{m-2}, \qquad \nu,\mu,\mu_1 = const > 0$$

$$\left|\frac{\partial a_{11}}{\partial p}\right|(1+|p|)^3 + \left|\frac{\partial a_{11}}{\partial u}\right|(1+|p|)^2 + |b| \leq \mu_1(1+|p|)^m,$$

$$\left|\frac{\partial a_{11}}{\partial x}\right|(1+|p|)^2 + \left|\frac{\partial b}{\partial x}\right| \leq (\varepsilon + P(|p|))(1+|p|)^{m+1},$$

where $(t,x) \in \bar{\Omega}_T, \quad |u| \leq M, \quad m \in \mathbb{R}^+, \quad P(\rho) \geq 0$ *is continuous and* $P(\rho) \overset{\rho\to\infty}{\to} 0$*, and* $\quad 0 \leq \varepsilon = \varepsilon(M,\nu,\mu,\mu_1,\max\limits_{\rho\geq 0} P(\rho))$ *sufficiently small,*

d) $\psi(x), u_0(x) \in C^{2+\beta}(\bar{D})$ *and fulfil a compatibility condition of zeroth and first order (for definition of compatibility condition see Appendix B).*

Theorem 1. *([25, Theorem 4.1, Ch.VI Ladyzhenskaia])*
Under the regularity conditions RC1 there exists a unique solution of problem (2.4) from the class $C^{1+\beta/2,2+\beta}(\bar{\Omega}_T)$*, and has derivative* $u_{tx} \in L_2(\Omega_T)$*.*

Remark 1. *[25, Ch.VI, pp.561 Ladyzhenskaia])*
The problem in Theorem 1 is considered in a domain $\Omega_T = (0,T) \times D$*, with* D *fixed interval, but Theorem 1 is applicable, without any essential change, to domain of the form* $\{(t,x) : \quad \phi_1(t) \leq x \leq \phi_2(t)\}$ *with* $\phi_i'(t) \neq \infty$*.*

Definition 1. *We say that the quasilinear problem (2.4) possesses a* **classical solution** *if a function* u *exists such that all of the derivatives taking part in the quasilinear equation,* u_t*,* u_x, u_{xx}*, exist and are continuous, i.e.* $u \in C^{1,2}(\bar{\Omega}_T)$*, and* u *satisfies (2.4).*

If the coefficients making up the quasilinear parabolic problem are not smooth functions, then this problem does not have, in general, a classical solution. It is not excluded that it possesses a solution belonging to a Banach space that is wider than $C^{1,2}(\bar{\Omega}_T)$. The choice of this function space is dictated by the smoothness properties of the coefficients of the equation.

Let now $W_2^r(D)$ be the Sobolev space with inner product $(.,.)$ and the usual norm $||.||_r$ induced by it. Let $W(D)$ be the space

$$W(D) = \{v \in W_2^r(D) : \ v(x) = \psi(x), \ x \in \partial D\}.$$

Because we want to use separate space and time discretization, we introduce the notation $t \mapsto g(t) \in W(D)$. A weak formulation of problem (2.4) then reads

find $u(t) \in W(D)$ such that for all $v \in W(D)$

$$\begin{aligned} \partial_t(u(t), v) + a_0(u(t), v) &= (b(t,.), v), \ 0 < t \leq T, \qquad (2.5) \\ (u(0,.) - u_0(.), v) &= 0, \end{aligned}$$

where the symbol ∂_t stands for derivation with respect to t, and

$$a_0(u, v) = ((a_{11} \cdot v)_x, u_x) - (a_{11} \cdot v, u_x)|_{\partial D}.$$

Definition 2. *By a* **weak solution** *of the problem (2.4) we mean a bounded function* $u = u(t) \in W(D)$ *that satisfies the identity (2.5) for any function* $v \in W(D)$*. We call the weak solution admissible if the problem (2.4) has only one weak solution.*

2.2 Space and time discretization

Here, we consider the questions of space and time discretization. For this purpose we use the method of lines [42],[24], with the Finite Element Method for the space discretization coupled with a difference scheme for the time discretization.

We consider the weak formulation of problem (2.4), namely (2.5). Let W_h be a finite subspace of W. Then, the basic semidiscrete Galerkin finite element problem reads:

find $U_h(t) \in W_h$, such that for all $v \in W_h$ holds

$$\begin{aligned} \partial_t(U_h(t), v) + a_0(U_h(t), v) &= (b(t,x), v),\ 0 < t \leq T, \\ (U_h(0) - U_0, v) &= 0. \end{aligned} \tag{2.6}$$

The subspace W_h is finite and one can choose a finite basis in it. Then (2.6) can be transformed to a system of ordinary differential equations. We denote the right hand side by

$$f(U_h) := -a_0(U_h(t), v) + (b(t,x), v). \tag{2.7}$$

A full discretization may be realized by means of a stable time-stepping method. We concentrate our attention on the difference schemes from Lobatto III A family, and particularly the second order Lobatto III A (trapezoidal rule (TR)) and the fourth order. The Butcher's tableaux for the 2^{nd} and 4^{th} order Lobatto III A, respectively, 2^{nd} and 3^{th} stage, are given in Table 2.1 (the interpretation of Butcher's tableaux is given in Appendix C.2).

0	0	0
1	1/2	1/2
	1/2	1/2

0	0	0	0
1/2	5/24	1/3	-1/24
1	1/6	2/3	1/6
	1/6	2/3	1/6

Table 2.1: The 2^{nd} and 4^{th} order Lobatto III A methods

The explicit expression for TR applied to $y'(t) = f(t,y), \quad y(0) = y_0$ is

$$y_{j+1} = y_j + \frac{\tau}{2}(f(t_j + \tau, y_{j+1}) + f(t_j, y_j)),$$

where τ stands for the time step.

Regarding the linear stability of TR, i.e. the application of TR to $y'(t) = \lambda y$ (the famous Dahlquist equation) gives $y_{j+1} = R(\tau\lambda)y_j = \sum\limits_{k=0}^{j}\{R(\tau\lambda)\}^k y_0$, where $R(z) = \frac{1+z/2}{1-z/2}$. The Dahlquist equation is stable if $\lambda < 0$ which corresponds to $|R(z)| \leq 1$ for the difference scheme. The function $R(z)$ is referred to as stability function and the set $\{z \in \mathbb{C} : |R(z)| \leq 1\}$ as stability region. For nonlinear problems, λ is replaced by the eigenvalues of the Jacobian matrix. The stability function is considered as approximation to the exponential function $e^z = R(z) + C \cdot z^{p+1} + \mathcal{O}(z^{p+2})$ where p is the order of the difference method (for TR $p = 2$). Depending on how much $R(z)$ resembles e^z we consider different kinds of stability.

The trapezoidal rule is only A-stable. This implies that the stability region coincides exactly with the negative half-plane $\{z \in \mathbb{C} : \mathcal{R}e(z) \leq 0\}$, but the stability function $|R(z)|$ tends to 1 rather than zero as $\mathcal{R}e(z) \to -\infty$ (i.e. it is not L-stable) [19, pp.41]. That may cause an oscillatory behaviour of the numerical solution of stiff differential equations. The oscillations with the trapezoidal rule can be avoided by using a smaller step size in the transient phase; once the smooth region is reached larger step sizes can be used [24, pp.36]. Nevertheless, TR (as well as the whole family Lobatto IIIA) gives asymptotically exact results for $z \to \infty$, $z = \lambda\tau$ for λ being the Lipschitz constant and τ the time step, i.e. it is stiffly accurate [19, pp. 242]. This can be seen by looking at the behaviour of the local and global error for stiff problems - see Table 2.2.

stage,	local error	global error	
		constant τ	variable τ
s odd	$z^{-1}\tau^{s+1}$	$z^{-1}\tau^{s}$	$z^{-1}\tau^{s}$
s even	$z^{-1}\tau^{s+1}$	$z^{-1}\tau^{s+1}$	$z^{-1}\tau^{s}$

Table 2.2: Order reduction for Lobatto III A family, when $\tau \to 0$ and $z = \tau\lambda \to \infty$

From the same table, one can notice as well that there is no order reduction. The stiffly accurate property makes it suitable for solving differential-algebraic-equations of index 1, [19, pp. 408].

The trapezoidal rule has often been referred to as being symmetric or time-reversible and is therefore good for Hamiltonian systems. However, it is well-known that the trapezoidal rule is not symplectic (area preserving) but it is related to a symplectic method, namely the Midpoint Rule, through a coordinate transformation. The latter implies that both numerical solutions have the same long-time behaviour [20, ch VI.7.4]. Petzold et al. in [39] reported that TR fails for large step sizes in solving highly oscillatory systems such as Hamiltonian and mechanical systems containing strong potentials which force the motion to be close to a smooth manifold. They proposed to use a high order scheme, like the s-stage ($s \geq 3$) Lobatto IIIA methods. This is also supported by Faou at al. in [11], where sufficient conditions are derived for energy conservation with non-symplectic methods. For completeness, we gave all of the above notes although our problem is neither Hamiltonian nor highly oscillatory.

The underlying problem, the anomalous heat transport in the tokamak plasma, is energy conserving and not oscillatory. This makes trapezoidal rule acceptable for it. With the strategy we choose the time step for getting a certain accuracy (see Ch.3, section 3.5.2), we have not encountered oscillation due to too large time step. This is natural since we apply front tracking technique with an explicit tracking of the front (see Ch. 3). There, as is going to be discussed in Ch. 3, section 3.5.3, we solve a differential equation for the position of the front using an explicit scheme. The stability of this explicit scheme requires some constrains on the size of the time step.

After the utilization of the trapezoidal rule, one obtains an implicit system of algebraic equations, which could be solved by a Newton-like iteration. The procedure described up to now is a standard way of dealing with PDEs. The new idea in the algorithm developed by us consists in a new way to adapt the number of Newton iterations. The substantial role is played by the defect of the 4^{th} order Lobatto III A scheme computed at the approximate solution from the trapezoidal rule. We attempt to reduce this defect, with respect to a given tolerance, and in this way to minimize the defect of the differential equation. According to the magnitude of the defect of the 4^{th} order Lobatto III A scheme we decide if the Newton procedure is stopped. In addition, we use the defect of the 4^{th} order Lobatto III A scheme as a local error estimate. The price we pay for this is one additional calculation of the r.h.s function.

From all we said above it is clear that the trapezoidal rule is not really recommended for oscillatory or mechanical stiff problems. Still the trapezoidal rule gives satisfactory results for “mildly” stiff problems, like the one obtained after the space discretization of PDEs. The simpleness of the application of TR and its stability properties make it favourable for solving large systems of ODEs. Further on, in Ch.4, we demonstrate the behaviour of the time integrator of AIM (adapted trapezoidal rule) in solving some ODEs. For some of the considered problems the time integrator of AIM is better (in terms of computational effort) than Lobatto IIIA 4^{th} order, but not always. Particularly for the anomalous heat transport problem it shows an advantage.

In the next section we give a precise definition of the concept “defect”, as well as the theoretical base for our statements. We first develop them for the case of ordinary differential equations. The generalization to partial differential equations is straight forward and is subject of the last section of this chapter.

2.3 Lobatto estimate. Local and global errors

The general formulation of an initial value problem (IVP) for an autonomous system of ODEs is:

$$\begin{aligned} \frac{d\mathbf{u}(t)}{dt} &= \mathbf{f}(\mathbf{u}), \qquad t \in (0, T] \\ \mathbf{u}(0) &= \mathbf{u_0}, \end{aligned} \tag{2.8}$$

where $\mathbf{f}: \mathbb{R}^n \to \mathbb{R}^n$ is a sufficiently smooth function, and there exists a unique solution $\mathbf{u}$ bounded in a domain $D \subset \mathbb{R}^n$. The vector $\mathbf{U}_j \in \mathbb{R}^n$ approximates the exact solution $\mathbf{u}_j := \mathbf{u}(t_j)$ at $t = t_j := j \cdot \tau$ for $j = 0, 1, 2 \ldots$, and τ is the time step.

Here, in this section we use a subscript for denoting the time discretization. Later on when the space discretization is also involved we mark the time discretization with a superscript and the space discretization with a subscript.

For simplicity the step size τ is taken constant in these considerations. A general formulation of one step difference scheme is

$$\mathbf{U}_{j+1} = \mathbf{F}(\tau, \mathbf{U}_j, \mathbf{U}_{j+1}), \qquad \mathbf{U}_0 = \mathbf{u}(\mathbf{0}), \tag{2.9}$$

where $\mathbf{F}$ is a vector function depending on the time step τ, on the solution at the previous time step $\mathbf{U}_j$, and possibly on the solution at the current time step $\mathbf{U}_{j+1}$.

Definition 3. *Under* **<u>local error</u>** *(truncation error)* T_{j+1} *we understand the error which is done in a single step, that is*

$$T_{j+1} := \mathbf{u}_{j+1} - \mathbf{F}(\tau, \mathbf{u}_j, \mathbf{U}_{j+1}). \tag{2.10}$$

Definition 4. *The* **<u>global error</u>** e_{j+1} *of the difference scheme could be defined as:*

$$e_{j+1} := \mathbf{u}_{j+1} - \mathbf{U}_{j+1} = \mathbf{u}_{j+1} - \mathbf{F}(\tau, \mathbf{U}_j, \mathbf{U}_{j+1}), \quad \mathbf{U}_0 = \mathbf{u}(0). \tag{2.11}$$

We use two more concepts - the defect and the residual of the difference scheme.

Definition 5. *The* **<u>defect</u>** *of the difference scheme is:*

$$def(\tau, \mathbf{U}_j, \mathbf{X}) := \mathbf{X} - \mathbf{F}(\tau, \mathbf{U}_j, \mathbf{X}) \quad \textit{for } \mathbf{X} \in \mathbb{R}^n. \tag{2.12}$$

It vanishes for a solution obtained by solving the scheme exactly.

Definition 6. *The* **<u>residual</u>** $\tilde{T}_{j+1}$ *of the scheme is*

$$\tilde{T}_{j+1} := \mathbf{u}_{j+1} - \mathbf{F}(\tau, \mathbf{u}_j, \mathbf{u}_{j+1}) = def(\tau, \mathbf{u}_j, \mathbf{u}_{j+1}). \tag{2.13}$$

For an explicit scheme, the residual coincides with the truncation error, i.e. $T_{j+1} = \tilde{T}_{j+1}$.

Definition 7. *[45]*
The difference scheme (2.9) *applied to ordinary differential equations with right-hand side which is sufficiently differentiable is said to be of* **<u>order</u>** <u>p</u>, *a positive integer number, if*

i) for all functions $\mathbf{f} \in C^\infty(\mathbb{R}^n, \mathbb{R}^n)$ *and for all* j

$$\lim_{\tau \to 0} \frac{||T_{j+1}||}{\tau^{p+1}} < \infty,$$

ii) and there exist $\mathbf{f} \in C^\infty(\mathbb{R}^n, \mathbb{R}^n)$ *and* j *such that*

$$\lim_{\tau \to 0} \frac{||T_{j+1}||}{\tau^{p+2}} = \infty.$$

We are interested in the difference schemes from the Lobatto III A family, and especially the 2nd (trapezoidal rule) and the 4th order. They are defined as follows:

- the second order Lobatto III A -Trapezoidal rule:
  - scheme:
    $$\mathbf{U}_{j+1} = \mathbf{U}_j + \frac{\tau}{2}(\mathbf{f}(\mathbf{U}_{j+1}) + \mathbf{f}(\mathbf{U}_j)),$$
  - defect:
    $$def(\tau, \mathbf{U}_j, \mathbf{X}) := \mathbf{X} - \mathbf{U}_j - \frac{\tau}{2}(\mathbf{f}(\mathbf{X}) + \mathbf{f}(\mathbf{U}_j)), \qquad \text{for } X \in \mathbb{R}^n$$
  - local error:
    $$T^{tr}_{j+1} := T_{j+1} = \mathbf{u}_{j+1} - \mathbf{u}_j - \frac{\tau}{2}(\mathbf{f}(\mathbf{u}_j) + \mathbf{f}(\mathbf{U}_{j+1})),$$
  - residual:
    $$\tilde{T}^{tr}_{j+1} := \tilde{T}_{j+1} = \mathbf{u}_{j+1} - \mathbf{u}_j - \frac{\tau}{2}(\mathbf{f}(\mathbf{u}_j) + \mathbf{f}(\mathbf{u}_{j+1})),$$
  - connection between T^{tr}_{j+1} and $\tilde{T}^{tr}_{j+1}$:
    $$\begin{aligned} \tilde{T}^{tr}_{j+1} &= T^{tr}_{j+1} - \frac{\tau}{2}(\mathbf{f}(\mathbf{u}_{j+1}) - \mathbf{f}(\mathbf{U}_{j+1})) \\ &= \left(I + \frac{\tau}{2} \sum_{k=1}^{\infty} \frac{(-1)^k \mathbf{f}^{(k)}(\mathbf{u}_{j+1})}{k!} (T^{tr}_{j+1})^{k-1} \right) T^{tr}_{j+1}. \end{aligned}$$

    The last expression is simply a result from Taylor expansion of $\mathbf{f}(\mathbf{U}_{j+1})$ at $\mathbf{u}_{j+1}$.
- the fourth order Lobatto III A scheme
  - scheme:
    $$\mathbf{U}_{j+1} = \mathbf{U}_j + \frac{\tau}{6}(\mathbf{f}(\mathbf{U}_{j+1}) + \mathbf{f}(\mathbf{U}_j)) + \frac{2\tau}{3}\mathbf{f}\left(\frac{\mathbf{U}_{j+1} + \mathbf{U}_j}{2} + \frac{\tau}{8}(\mathbf{f}(\mathbf{U}_j) - \mathbf{f}(\mathbf{U}_{j+1})) \right),$$

– defect:

$$\begin{aligned} L_j(\tau,\mathbf{U}_j,\mathbf{X}) : &= def(\tau,\mathbf{U}_j,\mathbf{X}) \qquad (2.14) \\ &= \mathbf{X}-\mathbf{U}_j-\frac{\tau}{6}(\mathbf{f}(\mathbf{X})+\mathbf{f}(\mathbf{U}_j)) \\ &\quad -\frac{2\tau}{3}\mathbf{f}\left(\frac{\mathbf{X}+\mathbf{U}_j}{2}+\frac{\tau}{8}(\mathbf{f}(\mathbf{U}_j)-\mathbf{f}(\mathbf{X}))\right), \qquad \text{for } \mathbf{X}\in\mathbb{R}^n, \end{aligned}$$

– local error:

$$\begin{aligned} T^{lob}_{j+1} : &= T_{j+1} \\ &= \mathbf{u}_{j+1}-\mathbf{u}_j-\frac{\tau}{6}(\mathbf{f}(\mathbf{U}_{j+1})+\mathbf{f}(\mathbf{u}_j)) \\ &\quad -\frac{2\tau}{3}\mathbf{f}\left(\frac{\mathbf{U}_{j+1}+\mathbf{u}_j}{2}+\frac{\tau}{8}(\mathbf{f}(\mathbf{u}_j)-\mathbf{f}(\mathbf{U}_{j+1}))\right), \end{aligned}$$

– residual:

$$\begin{aligned} \tilde{T}^{lob}_{j+1} : &= \tilde{T}_{j+1} \\ &= \mathbf{u}_{j+1}-\mathbf{u}_j-\frac{\tau}{6}(\mathbf{f}(\mathbf{u}_{j+1})+\mathbf{f}(\mathbf{u}_j)) \\ &\quad -\frac{2\tau}{3}\mathbf{f}\left(\frac{\mathbf{u}_{j+1}+\mathbf{u}_j}{2}+\frac{\tau}{8}(\mathbf{f}(\mathbf{u}_j)-\mathbf{f}(\mathbf{u}_{j+1}))\right). \end{aligned}$$

2.3.1 The case $n = 1$

Here, we prove a Lemma, giving a connection between the truncation error and the defect of Lobatto III A for the case $n = 1$.

Lemma 1. :
Let U_1 be an approximate solution, at $t = t_1$, of the initial value problem (2.8) computed through a difference scheme of type (2.9) of order $1 \le p \le 4$. Then, there exists the following connection between the truncation error, T_1, of the difference scheme and the defect of the 4th order Lobatto III A, $L_0(\tau, U_0, U_1)$,

$$L_0(\tau,U_0,U_1) = T_1(-1+\frac{\tau}{2}f'(U_0)) + \mathcal{O}(\tau^{\min(5,p+3)}).$$

Proof: We prove this Lemma using the package Mathematica. However, we give in Appendix C.1 an alternative way of proving the following a bit weaker, statement

$$L_0(\tau,U_0,U_1) = T_1(-1+\frac{\tau}{2}f'(U_0)) + \mathcal{O}(\tau^4). \qquad (2.15)$$

♢

One consequence of Lemma 1 is that the defect of the 4^{th} order Lobatto III A scheme can be used as an approximation of the truncation error of lower order method. The following Theorem is giving an estimate of the global error.

Theorem 2. :
Let us consider IVP (2.8), with sufficiently smooth r.h.s function and let $\{U_j\}_{j=0}^m$ be an approximation of the solution of (2.8), obtained through the trapezoidal rule. The following expressions for the error e_{j+1} hold

A)

$$\left(1-\frac{\tau}{2}\sum_{k=1}^{\infty}\frac{1}{k!}f^{(k)}(U_{j+1})e_{j+1}^{k-1}\right)\ e_{j+1}=\tilde{T}_{j+1}^{tr}+\left(1+\frac{\tau}{2}\sum_{k=1}^{\infty}\frac{1}{k!}f^{(k)}(U_j)e_j^{k-1}\right)\ e_j,$$

B)

$$\left(1-\frac{\tau}{2}f'(U_{j+1})\right)\ e_{j+1}=-L_j(\tau,U_j,U_{j+1})+\tilde{T}_{j+1}^{lob}+\left(1+\frac{\tau}{2}f'(U_j)\right)e_j+\mathcal{O}(\tau^5).$$

Remark 2. *A) is a well known expression for the error reformulated using the residual of the considered scheme. Our contribution is stated in B), where the defect of the 4^{th} order Lobatto III A scheme is used.*

Proof:

A) Let us consider the global error at t_{j+1}:

$$e_{j+1}=u_{j+1}-U_{j+1}=\tilde{T}_{j+1}^{tr}+u_j+\frac{\tau}{2}(f(u_j)+f(u_{j+1}))-U_{j+1}.$$

By adding and subtracting U_j we get

$$e_{j+1}=\tilde{T}_{j+1}^{tr}+(u_j-U_j)+U_j+\frac{\tau}{2}(f(u_j)+f(u_{j+1}))-U_{j+1}.$$

Now, using the Taylor expansion of $f(u_j)$ at U_j, and of $f(u_{j+1})$ at U_{j+1}, we obtain

$$\begin{aligned}e_{j+1}= &\ \tilde{T}_{j+1}^{tr}+e_j+U_j-U_{j+1}+\tfrac{\tau}{2}\sum_{k=0}^{\infty}\tfrac{1}{k!}\left(f^k(U_j)e_j^k+f^k(U_{j+1})e_{j+1}^k\right)\\ = &\ \tilde{T}_{j+1}^{tr}+\left(1+\sum_{k=1}^{\infty}\tfrac{1}{k!}f^k(U_j)e_j^{k-1}\right)e_j+\tfrac{\tau}{2}\left(\sum_{k=1}^{\infty}\tfrac{1}{k!}f^k(U_{j+1})e_{j+1}^{k-1}\right)e_{j+1}\\ &-\underbrace{(\ U_{j+1}-U_j-\frac{\tau}{2}f(U_j)-\frac{\tau}{2}f(U_{j+1})\ )}_{0}.\end{aligned}$$

B) Let us consider the difference between $L_j(\tau, \mathbf{U}_j, U_{j+1})$ and $\tilde{T}^{lob}_{j+1}$, i.e. $I := L_j(\tau, U_j, U_{j+1}) - \tilde{T}^{lob}_{j+1}$ and make use of the Taylor expansion:

$$\begin{aligned}
I &= (U_{j+1} - u_{j+1}) - (U_j - u_j) - \tfrac{\tau}{6}(f(U_{j+1}) - f(u_{j+1})) - \tfrac{\tau}{6}(f(U_j) - f(u_j)) \\
&\quad -\tfrac{2\tau}{3}\left(f(\tfrac{U_{j+1}+U_j}{2} + \tau\tfrac{f(U_j)-f(U_{j+1})}{8}) - f(\tfrac{u_{j+1}+u_j}{2} + \tau\tfrac{f(u_j)-f(u_{j+1})}{8})\right) \\
&= -e_{j+1} + e_j - \tfrac{\tau}{6}(-\ f'(U_{j+1})e_{j+1} + \tfrac{1}{2}f''(U_{j+1})e^2_{j+1} - \mathcal{O}(e^3_{j+1})\ ) \\
&\quad -\tfrac{\tau}{6}(\ -f'(U_j)e_j + \tfrac{1}{2}f''(U_j)e^2_j - \mathcal{O}(e^3_j)\ ) \\
&\quad -\tfrac{2\tau}{3}f'(U_{j+1})\left(-\tfrac{e_j+e_{j+1}}{2} + \tfrac{\tau}{8}(f(U_j) - f(u_j) + f(u_{j+1}) - f(U_{j+1}))\right) \\
&= -\left(\ 1 - \tfrac{\tau}{2}f'(U_{j+1}) + \tfrac{\tau}{12}f''(U_{j+1})e_{j+1} + \tau\mathcal{O}(e^2_{j+1})\ \right) e_{j+1} \\
&\quad +\left(\ 1 + \tfrac{\tau}{2}f'(U_j) - \tfrac{\tau}{12}f''(U_j)e_j + \tau\mathcal{O}(e^2_j)\ \right) e_j.
\end{aligned}$$

This proves the statement.

$\diamondsuit$

Theorem 2, B, is used in the numerical calculation as an approximation of the global error

$$\left(1 - \frac{\tau}{2}f'(U_{j+1})\right)\ e_{j+1} \approx -L_j(\tau, U_j, U_{j+1}) + \tilde{T}^{lob}_{j+1} + \left(1 + \frac{\tau}{2}f'(U_j)\right)e_j.$$

Here, the terms of order five and higher are neglected.

In the next section, we consider the case of a system of ODEs and we show similar expressions for the error.

2.3.2 The case of $n > 1$

The following considerations are extensions of Lemma 1 and Theorem 2, B for the case $n > 1$. The stress is laid on the case of stiff systems of ODEs. The contents of this section follows the Finckenstein paper [12], which presents convergence results for the method of lines (difference scheme coupled with trapezoidal rule) applied to parabolic systems. However, the truncation error in [12] is not replaced by the defect of Lobatto as we have done it here.

Assumption 1. :
Assume that there exists a constant C such that :

$$w^t \left(\frac{\mathcal{D}\boldsymbol{f}(\boldsymbol{u}) + \mathcal{D}\boldsymbol{f}(\boldsymbol{u})^t}{2} \right) w \leq C \quad \textit{for any} \quad \boldsymbol{u} \in D \quad \textit{and any } \boldsymbol{w} \in D : \ \boldsymbol{w}^t \boldsymbol{w} = 1,$$

where $\mathcal{D}f$ is the Jacobian matrix of $\boldsymbol{f}$.

Definition 8 (Logarithmic matrix norm)**. :**
For a given matrix $A \in \mathbb{R}^{n \times n}$ the **logarithmic norm***, $\mu[A]$, is defined by*

$$\mu[A] := \lim_{\delta \to 0+} \frac{||I + \delta A|| - 1}{\delta},$$

where $I \in \mathbb{R}^{n \times n}$ stands for the identity matrix.

For an inner product norm, the logarithmic matrix norm of A is the smallest possible one-side Lipschitz constant for the matrix A (see also Appendix A).
We use the notations: A^t - the transpose of the matrix A, and $\lambda_{max}[A]$ - the largest eigenvalue of A.
One important characteristic of the logarithmic norm is as follows:

Remark 3. *Let us consider the logarithmic norm of $\mathcal{D}f$. It follows from the first property of the logarithmic matrix norm, Appendix A.1 that*

$$\mu[\mathcal{D}f] = \max_{||\boldsymbol{w}||=1} < \mathcal{D}\boldsymbol{f}(\boldsymbol{u})\boldsymbol{w}, \boldsymbol{w} >, \qquad \textit{any } \boldsymbol{u} \in D,$$

then

$$< \mathcal{D}\boldsymbol{f}(\boldsymbol{u})\boldsymbol{w}, \boldsymbol{w} > \, = \, < \frac{\mathcal{D}\boldsymbol{f}(\boldsymbol{u}) + \mathcal{D}\boldsymbol{f}(\boldsymbol{u})^t}{2} \boldsymbol{w}, \boldsymbol{w} > \, \leq \, \lambda_{max}[\frac{\mathcal{D}\boldsymbol{f}(\boldsymbol{u}) + \mathcal{D}\boldsymbol{f}(\boldsymbol{u})^t}{2}].$$

Definition 9. *Let us assign to the constant C from Assumption 1 the following value*

$$C := \max_{\boldsymbol{u} \in D} \mu[\mathcal{D}\boldsymbol{f}(\boldsymbol{u})] = \max_{\boldsymbol{u} \in D} \lambda_{max}[\frac{\mathcal{D}\boldsymbol{f}(\boldsymbol{u}) + \mathcal{D}\boldsymbol{f}(\boldsymbol{u})^t}{2}]. \tag{2.16}$$

We use, throughout this chapter, one more constant: C_{max}.

Definition 10. *We denote the classical Lipschitz constant for the function $\boldsymbol{f}(\boldsymbol{u})$ as*

$$C_{max} := \max_{\boldsymbol{u} \in D} ||\mathcal{D}\boldsymbol{f}(\boldsymbol{u})||. \tag{2.17}$$

Now, let us define a vector function $\mathbf{L} : D \times D \to \mathbb{R}^n$

$$\mathbf{L}(\tau, \mathbf{w}, \mathbf{v}) := \mathbf{v} - \mathbf{w} - \frac{\tau}{6}(\mathbf{f}(\mathbf{v}) + \mathbf{f}(\mathbf{w})) - \frac{2\tau}{3}\mathbf{f}\left(\frac{\mathbf{v} + \mathbf{w}}{2} + \frac{\tau}{8}(\mathbf{f}(\mathbf{w}) - \mathbf{f}(\mathbf{v}))\right), \tag{2.18}$$

for any $\mathbf{v}, \ \mathbf{w} \in D$. Let $\{\mathbf{U}_j\}_{j=1,..}$, $\mathbf{U}_0 = \mathbf{u}_0$ be an approximation of the solution of (2.8) and $\mathbf{U}_j \in D$, for $j = 0, \cdots$.

Remark 4. :

- $\boldsymbol{L}(\tau, \boldsymbol{u}_j, \boldsymbol{u}_{j+1})$ *is actually the residual of the* 4^{th} *order Lobatto III A scheme, i.e.* $\boldsymbol{L}(\tau, \boldsymbol{u}_j, \boldsymbol{u}_{j+1}) = \mathcal{O}(\tau^5)$

- $\boldsymbol{L}(\tau, \boldsymbol{U}_j, \boldsymbol{U}_{j+1})$ *is the defect of the* 4^{th} *order Lobatto III A scheme.*

Our goal is to prove that $\mathbf{L}(\tau, \mathbf{U}_j, \mathbf{U}_{j+1})$, which can be computed easier than the usual truncation error of the difference scheme used, yields a good approximation of the local respectively, the global error. This is shown in the following theorem.

Theorem 3. :
Let the r.h.s. function $f(\mathbf{u})$ *in (*2.8*) be such that Assumption 1 holds with a constant* C *defined by (2.16). Moreover, let* $\{\boldsymbol{U}_j\}_j \in D$ *be an approximation of the solution of (*2.8*) obtained through trapezoidal rule (TR). Then the global error of approximation permits the estimate*

a)

$$||\boldsymbol{e}_{j+1}|| \leq d(\tau)(j+1), \quad \textit{for } C \leq 0 \tag{2.19}$$

b)

$$||\boldsymbol{e}_{j+1}|| \leq \frac{d(\tau)(j+1)e^{K_1 T}}{\sqrt{1 - \frac{\tau^2 C_{max}^2}{6}} - \frac{\tau}{2} C_{max}}, \quad \textit{for } C > 0,\ \tau \leq \sqrt{\frac{12}{5C_{max}^2}}, \tag{2.20}$$

where
$d(\tau) = \max\limits_{0 \leq k \leq j+1} ||\boldsymbol{L}(\tau, \boldsymbol{u}_k, \boldsymbol{u}_{k+1}) - \boldsymbol{L}(\tau, \boldsymbol{U}_k, \boldsymbol{U}_{k+1})||$, $K_1 = \frac{C}{1 - C\tau - \frac{\tau^2}{6} C_{max}^2}$ *and* C_{max} *is defined by (2.17). If one approximates the solution of (*2.8*) through another method, different from TR, additional restriction on the time step may play a role.*

Let us shortly discuss this theorem. In the numerical calculation one can neglect $\mathbf{L}(\tau, \mathbf{u}_k, \mathbf{u}_{k+1})$ and uses only the defect of the 4^{th} order Lobatto III A scheme to estimate the global error. We refer to this estimate as the **Lobatto estimate**. The quality of this estimate is investigated on examples in Sections 4.1 and 4.2. The expense of this error estimate is an additional calculation of the function **f**.

- **Preparation for proving Theorem 3**:

In order to get an expression for the error $\mathbf{e}_{j+1} = \mathbf{u}_{j+1} - \mathbf{U}_{j+1}$, we consider

$$\mathbf{L}(\tau, \mathbf{u}_j, \mathbf{u}_{j+1}) - \mathbf{L}(\tau, \mathbf{U}_j, \mathbf{U}_{j+1}) = \mathbf{e}_{j+1} - \mathbf{e}_j - \tfrac{\tau}{6}(\mathbf{f}(\mathbf{U}_{j+1}) - \mathbf{f}(\mathbf{u}_{j+1})) \\ -\tfrac{\tau}{6}(\mathbf{f}(\mathbf{U}_j) - \mathbf{f}(\mathbf{u}_j)) - \tfrac{2\tau}{3}(\mathbf{f}(\mathbf{Z}_j) - \mathbf{f}(\mathbf{z}_j)), \tag{2.21}$$

where we denote

$$\mathbf{Z}_j = \frac{\mathbf{U}_{j+1} + \mathbf{U}_j}{2} + \frac{\tau}{8}(\mathbf{f}(\mathbf{U}_j) - \mathbf{f}(\mathbf{U}_{j+1})),\ \mathbf{z}_j = \frac{\mathbf{u}_{j+1} + \mathbf{u}_j}{2} + \frac{\tau}{8}(\mathbf{f}(\mathbf{u}_j) - \mathbf{f}(\mathbf{u}_{j+1})). \tag{2.22}$$

We assume that $\mathbf{Z}_j,\ \mathbf{z}_j$ are in D. Let us have a closer look at this assumption. We take the r.h.s function of (2.8) to be a linear function i.e. $f(\mathbf{u}) = \alpha\mathbf{u}, \quad \alpha \in \mathbb{R}^{n\times n},\ \alpha = diag(\alpha_1, \cdots, \alpha_n)$ and $\alpha_{max} = \max\limits_{1\le i\le n} \alpha_i$.

i) First, we suppose that $\{\mathbf{U}_j\}_j$ is obtained by applying the explicit Euler, i.e.

$$\mathbf{U}_{j+1} = \mathbf{U}_j + \tau\mathbf{f}(\mathbf{U_j}) = \mathbf{U}_j + \tau\alpha\mathbf{U}_j.$$

Let us have a look at $\mathbf{Z}_j$

$$\mathbf{Z}_j = \frac{\mathbf{U}_{j+1} + \mathbf{U}_j}{2} + \frac{\tau}{8}(\mathbf{f}(\mathbf{U_j}) - \mathbf{f}(\mathbf{U_{j+1}})) = \begin{cases} \mathbf{U}_j + \frac{\tau\alpha}{8}(5\mathbf{U}_j - \mathbf{U}_{j+1}) \\ \\ \mathbf{U}_{j+1} - \frac{\tau\alpha}{8}(3\mathbf{U}_j + \mathbf{U}_{j+1}). \end{cases}$$

It is guaranteed that $\mathbf{Z}_j \in D$ for all j, if $\mathbf{Z}_j$ is between $\mathbf{U}_j$ and $\mathbf{U}_{j+1}$. We consider two cases $\alpha_i < 0,\ i = 1, \cdots, n$ and $\alpha > 0,\ i = 1, \cdots, n$.

- The case $\alpha < 0,\ i = 1, \cdots, n$: decreasing exact solution $\mathbf{u}_{j+1} \le \mathbf{u}_j$. This is true for the approximation $\mathbf{U}_{j+1} \le \mathbf{U}_j$ if $|1 + \tau\alpha| < 1$, i.e. $\tau \le \frac{2}{-\alpha_{max}}$.
- The case $\alpha > 0,\ i = 1, \cdots, n$: increasing exact solution. The condition-$\mathbf{U}_j \le \mathbf{Z}_j \le \mathbf{U}_{j+1}$ gives $\tau \le \frac{4}{\alpha_{max}}$.

In the case of explicit Euler method we need to constrain the time step.

ii) Now, let us consider the trapezoidal rule:

$$\mathbf{U}_{j+1} = \mathbf{U}_j + \frac{\tau}{2}(\mathbf{f}(\mathbf{U_j}) + \mathbf{f}(\mathbf{U_{j+1}}))$$

and again $\mathbf{Z}_j$:

$$\mathbf{Z}_j = \frac{\mathbf{U}_{j+1} + \mathbf{U}_j}{2} + \frac{\tau}{8}(\mathbf{f}(\mathbf{U_j}) - \mathbf{f}(\mathbf{U_{j+1}})) = \begin{cases} \mathbf{U}_j + \frac{\tau}{8}(3\mathbf{f}(\mathbf{U_j}) + \mathbf{f}(\mathbf{U_{j+1}})) \\ \\ \mathbf{U}_{j+1} - \frac{\tau}{8}(\mathbf{f}(\mathbf{U_j}) + 3\mathbf{f}(\mathbf{U_{j+1}})). \end{cases}$$

For $\mathbf{f}(\mathbf{u}) \ge 0$ without constraints on the time step it is fulfilled that $\mathbf{U}_j \le \mathbf{Z}_j \le \mathbf{U}_{j+1}$. The same is true for the case $\mathbf{f}(\mathbf{u}) < 0$.

If the discretization is carried out through the explicit Euler method then an additional restriction on the time step is needed to assure that $\mathbf{z}_j$ and $\mathbf{Z}_j \in D$. The trapezoidal rule provides this condition without any restriction on the time step.

We define a function $\Gamma : D \times D \to \mathbb{R}^n$, such that

$$\Gamma(\mathbf{v}, \mathbf{w}) := \int_0^1 \mathcal{D}\mathbf{f}(\mathbf{w} + \sigma(\mathbf{v} - \mathbf{w}))d\sigma, \tag{2.23}$$

with $\mathcal{D}\mathbf{f}$ the Jacobian matrix of $\mathbf{f}$. One can verify that

$$\Gamma(\mathbf{v}, \mathbf{w})(\mathbf{v} - \mathbf{w}) = \mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{w}). \tag{2.24}$$

By applying (2.24) to (2.21) we obtain that

$$\begin{aligned}\mathbf{L}(\tau, \mathbf{u}_j, \mathbf{u}_{j+1}) - L(\tau, \mathbf{U}_j, \mathbf{U}_{j+1}) \quad &= \quad \mathbf{e}_{j+1} - \mathbf{e}_j - \frac{\tau}{6}(\Gamma(\mathbf{U}_{j+1}, \mathbf{u}_{j+1})\mathbf{e}_{j+1} - \Gamma(\mathbf{U}_j, \mathbf{u}_j)\mathbf{e}_j) \\ &\quad -\frac{2\tau}{3}\Gamma(\mathbf{Z}_j, \mathbf{z}_j)(\mathbf{Z}_j - \mathbf{z}_j).\end{aligned} \tag{2.25}$$

For the term $\mathbf{Z}_j - \mathbf{z}_j$ we obtain

$$\begin{aligned}\mathbf{Z}_j - \mathbf{z}_j = \; &\tfrac{\mathbf{U}_{j+1} - \mathbf{u}_{j+1}}{2} + \tfrac{\mathbf{U}_j - \mathbf{u}_j}{2} + \tfrac{\tau}{8}(\mathbf{f}(\mathbf{U}_j) - \mathbf{f}(\mathbf{u}_j)) - \tfrac{\tau}{8}(\mathbf{f}(\mathbf{U}_{j+1}) - \mathbf{f}(\mathbf{u}_{j+1})) \\ = \; &\tfrac{\mathbf{e}_{j+1} + \mathbf{e}_j}{2} + \tfrac{\tau}{8}(\;\Gamma(\mathbf{U}_j, \mathbf{u}_j)\;\mathbf{e}_j - \Gamma(\mathbf{U}_{j+1}, \mathbf{u}_{j+1})\;\mathbf{e}_{j+1}\;).\end{aligned}$$

Finally for (2.25) we get

$$\begin{aligned}\mathbf{L}(\tau, \mathbf{u}_j, \mathbf{u}_{j+1}) - \mathbf{L}(\tau, \mathbf{U}_j, \mathbf{U}_{j+1}) \quad = \quad &\Bigg(I + \frac{\tau^2}{12}\Gamma(\mathbf{Z}_j, \mathbf{z}_j)\Gamma(\mathbf{U}_{j+1}, \mathbf{u}_{j+1}) \\ &-\frac{\tau}{6}\Gamma(\mathbf{U}_{j+1}, \mathbf{u}_{j+1}) - \frac{2\tau}{6}\Gamma(\mathbf{Z}_j, \mathbf{z}_j)\Bigg)\,\mathbf{e}_{j+1} \\ &-\Bigg(I + \frac{\tau^2}{12}\Gamma(\mathbf{Z}_j, \mathbf{z}_j)\Gamma(\mathbf{U}_j, \mathbf{u}_j) \\ &+\frac{\tau}{6}(\Gamma(\mathbf{U}_j, \mathbf{u}_j) + \frac{2\tau}{6}\Gamma(\mathbf{Z}_j, \mathbf{z}_j)\Bigg)\,\mathbf{e}_j.\end{aligned} \tag{2.26}$$

Let us make the notations

$$\begin{aligned}A_{k,l} &:= I + \tfrac{\tau^2}{12}\;\Gamma(\mathbf{Z}_k, \mathbf{z}_k)\;\Gamma(\mathbf{U}_l, \mathbf{u}_l), \\ B_{k,l} &:= 2\;\Gamma(\mathbf{Z}_k, \mathbf{z}_k) + \Gamma(\mathbf{U}_l, \mathbf{u}_l).\end{aligned} \tag{2.27}$$

We use the notations (2.27) for (2.26) (for the term in front of $\mathbf{e}_{j+1}$ with $k = j$ and $l = j + 1$, and for the term in front of $\mathbf{e}_j$ with $k = l = j$)

$$\mathbf{L}(\tau, \mathbf{U}_j, \mathbf{U}_{j+1}) - \mathbf{L}(\tau, \mathbf{u}_j, \mathbf{u}_{j+1}) = (A_{j,j+1} - \frac{\tau}{6}B_{j,j+1})\mathbf{e}_{j+1} - (A_{j,j} + \frac{\tau}{6}B_{j,j})\mathbf{e}_j.$$

Using the following transformation of the error

$$\tilde{\mathbf{e}}_{k,l} = (A_{k,l} - \frac{\tau}{6}B_{k,l})\mathbf{e}_l \tag{2.28}$$

we get

$$\tilde{\mathbf{e}}_{j,j+1} = (A_{j,j} + \frac{\tau}{6}B_{j,j}) \left(A_{j,j} - \frac{\tau}{6}B_{j,j}\right)^{-1} \tilde{\mathbf{e}}_{j,j} + L(\tau, \mathbf{u}_j, \mathbf{u}_{j+1}) - L(\tau, \mathbf{U}_j, \mathbf{U}_{j+1}). \tag{2.29}$$

The L_2 norm of the error can be bounded by

$$\begin{aligned} ||\mathbf{e}_{j+1}|| &\leq ||(A_{j,j+1} - \tfrac{\tau}{6}B_{j,j+1})^{-1}|| \; ||\tilde{\mathbf{e}}_{j,j+1}|| \\ &\leq ||(A_{j,j+1} - \tfrac{\tau}{6}B_{j,j+1})^{-1}|| \; ||(A_{j,j} + \tfrac{\tau}{6}B_{j,j}) \left(A_{j,j} - \tfrac{\tau}{6}B_{j,j}\right)^{-1}|| \; ||\tilde{\mathbf{e}}_{j,j}|| \\ &+ ||(A_{j,j+1} - \tfrac{\tau}{6}B_{j,j+1})^{-1}|| \; ||L(\tau, \mathbf{u}_j, \mathbf{u}_{j+1}) - L(\tau, \mathbf{U}_j, \mathbf{U}_{j+1})||. \end{aligned} \tag{2.30}$$

In the next sub-section we give the proof of Theorem 3, for this purpose at first some Lemmas are proved.

2.3.3 Proof of the Error Estimate

In this section we assume that the discretization of problem (2.8) is carried out so that $Z_j,\ z_j \in D$ ($Z_j,\ z_j$ are defined by (2.22)). That is the case if the discretization is performed by the trapezoidal rule. For the explicit Euler this is true with some restrictions on the time step.

In the following Lemmas, we bound the terms $||(A_{j,j+1} - \frac{\tau}{6}B_{j,j+1})^{-1}||$ and $||(A_{j,j} + \frac{\tau}{6}B_{j,j}) \left(A_{j,j} - \frac{\tau}{6}B_{j,j}\right)^{-1}||$. The proof of Theorem 3 is given at the end of this sub-section. Auxiliary concepts for positive definite matrix are given in Appendix A.2 and are used throughout this section.

Lemma 2. :
If Assumption 1 is fulfilled with a constant C defined by (2.16) then

$$1) \quad \boldsymbol{w}^t(\Gamma(\boldsymbol{v}, \boldsymbol{y}) + \Gamma(\boldsymbol{v}, \boldsymbol{y})^t)\boldsymbol{w} \leq 2C,$$

$$2) \quad \boldsymbol{w}^t(\Gamma(\boldsymbol{x}, \boldsymbol{y})\Gamma(\boldsymbol{v}, \boldsymbol{z})^t)\boldsymbol{w} \leq C_{max}^2,$$

for $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v} \in D$, $\boldsymbol{w}^t\boldsymbol{w} = 1$, $\Gamma(.,.)$ defined by (2.23) and C_{max} defined by (2.17).

Proof:

1)

$$\begin{aligned}
\mathbf{w}^t(\Gamma(\mathbf{v},\mathbf{y})+\Gamma(\mathbf{v},\mathbf{y})^t)\mathbf{w} = &\ \mathbf{w}^t(\int\limits_0^1 \mathcal{D}\mathbf{f}(\mathbf{y}+\sigma(\mathbf{v}-\mathbf{y}))+\mathcal{D}\mathbf{f}(\mathbf{y}+\sigma(\mathbf{v}-\mathbf{y}))^t d\sigma)\mathbf{w} \\
\leq &\ 2\lambda_{max}[\int\limits_0^1 \tfrac{\mathcal{D}\mathbf{f}(\mathbf{y}+\sigma(\mathbf{v}-\mathbf{y}))+\mathcal{D}\mathbf{f}(\mathbf{y}+\sigma(\mathbf{v}-\mathbf{y}))^t}{2} d\sigma] \\
\leq &\ 2\int\limits_0^1 \lambda_{max}[\tfrac{\mathcal{D}\mathbf{f}(\mathbf{y}+\sigma(\mathbf{v}-\mathbf{y}))+\mathcal{D}\mathbf{f}(\mathbf{y}+\sigma(\mathbf{v}-\mathbf{y}))^t}{2})]d\sigma \\
\leq &\ 2\int\limits_0^1 C d\sigma = 2C.
\end{aligned}$$

2) Let C_1 be the following constant

$$C_1 := \max_{\mathbf{w}^t\mathbf{w}=1} \max_{\mathbf{v},\mathbf{y}\in D} \sqrt{\mathbf{w}^t(\ \Gamma(\mathbf{v},\mathbf{y})\ \Gamma(\mathbf{v},\mathbf{y})^t)\ \mathbf{w}} \geq 0.$$

Using the Cauchy-Schwartz inequality we get the estimate

$$\begin{aligned}
\mathbf{w}^t\left(\Gamma(\mathbf{x},\mathbf{y})\Gamma(\mathbf{v},\mathbf{z})^t\right)\mathbf{w} \leq &\ \sqrt{\mathbf{w}^t\ \Gamma(\mathbf{x},\mathbf{y})\ \Gamma(\mathbf{x},\mathbf{y})^t\ \mathbf{w}}\sqrt{\mathbf{w}^t\ \Gamma(\mathbf{v},\mathbf{z})\ \Gamma(\mathbf{v},\mathbf{z})^t\ \mathbf{w}} \\
\leq &\ C_1^2.
\end{aligned}$$

Since

$$\begin{aligned}
\sqrt{\mathbf{w}^t(\ \Gamma(\mathbf{x},\mathbf{y})\ \Gamma(\mathbf{x},\mathbf{y})^t)\ \mathbf{w}} = &\ ||\mathbf{w}^t\int_0^1 \mathcal{D}\mathbf{f}(\mathbf{y}+\sigma(\mathbf{x}-\mathbf{y}))d\sigma\ || \\
\leq &\ \int_0^1 ||\mathcal{D}\mathbf{f}(\mathbf{y}+\sigma(\mathbf{x}-\mathbf{y}))||d\sigma \leq \max_{\mathbf{u}\in D}||\mathcal{D}\mathbf{f}(\mathbf{u})||,
\end{aligned}$$

the constant C_1 can be bounded by C_{max}, which shows the validity of the statement.

$\diamondsuit$

Remark 5. *If $C < 0$, then $\boldsymbol{w}^t(\Gamma(\boldsymbol{x},\boldsymbol{y})\Gamma(\boldsymbol{v},\boldsymbol{z}))\boldsymbol{w} \geq 0$.*

Remark 5 follows from Proposition 3 in Appendix A.

Lemma 3. :
If Assumption 1 holds with a constant C defined by (2.16) and in addition it is satisfied that

a) $\boldsymbol{w}^t(\Gamma(\boldsymbol{x},\boldsymbol{y})+\Gamma(\boldsymbol{x},\boldsymbol{y})^t)\boldsymbol{w} \leq 2C$,

b) $\boldsymbol{w}^t(\Gamma(\boldsymbol{x},\boldsymbol{y})\Gamma(\boldsymbol{v},\boldsymbol{z})^t)\boldsymbol{w} \leq C^2_{max}$,

c) $\boldsymbol{w}^t(\frac{\Gamma(\boldsymbol{x},\boldsymbol{y})+\Gamma(\boldsymbol{x},\boldsymbol{y})^t}{2})(\frac{\Gamma(\boldsymbol{v},\boldsymbol{z})+\Gamma(\boldsymbol{v},\boldsymbol{z})^t}{2})\boldsymbol{w} \geq 0$,

then it is fulfilled that

$$1) \quad \boldsymbol{w}^t(B_{k,l}+B^t_{k,l})\boldsymbol{w} \leq 6C,$$

$$2) \quad \boldsymbol{w}^t(\ B_{k,l}B^t_{k,l}\ )\boldsymbol{w} \leq 9C^2_{max},$$

$$3) \quad \boldsymbol{w}^t(\ A_{k,l}A^t_{k,l}\ )\boldsymbol{w} \geq 1-\frac{\tau^2C^2_{max}}{6},$$

for $\boldsymbol{x},\boldsymbol{y},\boldsymbol{z},\boldsymbol{v} \in D$ *and* $\boldsymbol{w}^t\boldsymbol{w}=1$.
In case of $C<0$ *3) can be replaced by a sharper result*

$$3') \quad \boldsymbol{w}^t(\ A_{k,l}A^t_{k,l}\ )\boldsymbol{w} \geq 1.$$

Proof:

1) We have

$$\mathbf{w}^t(\ B_{k,l}+B^t_{k,l}\ )\mathbf{w} = \mathbf{w}^t\ (\ \Gamma(\mathbf{U}_l,\mathbf{u}_l)+\Gamma(\mathbf{U}_l,\mathbf{u}_l)^t+2\ (\Gamma(\mathbf{Z}_k,\mathbf{z}_k)+\Gamma(\mathbf{Z}_k,\mathbf{z}_k)^t\ ))\ \mathbf{w},$$

and because of the assumptions a) we obtain

$$\mathbf{w}^t(B_{k,l}+B^t_{k,l})\mathbf{w} \leq 2C+4C=6C.$$

2) Using the definition of $B_{k,l}$, (2.27), and writing the corresponding expression in terms of $\Gamma(.,.)$ we get

$$\begin{aligned}\mathbf{w}^t\ (B_{k,l}B^t_{k,l})\ \mathbf{w} &= \mathbf{w}^t\ \Gamma(\mathbf{U}_l,\mathbf{u}_l)\Gamma(\mathbf{U}_l,\mathbf{u}_l)^t\ \mathbf{w}+4\mathbf{w}^t\ \Gamma(\mathbf{U}_l,\mathbf{u}_l)\Gamma(\mathbf{Z}_k,\mathbf{z}_k)^t\ \mathbf{w} \\ &\quad +4\mathbf{w}^t\ \Gamma(\mathbf{Z}_k,\mathbf{z}_k)\Gamma(\mathbf{Z}_k,\mathbf{z}_k)^t\ \mathbf{w} \leq 9C^2_{max}.\end{aligned}$$

3) Using the definition of $A_{k,l}$ we get

$$\begin{aligned}\mathbf{w}^t(A_{k,l}A^t_{k,l})\mathbf{w} &= 1+\frac{\tau^2}{12}\mathbf{w}^t(\ \Gamma(\mathbf{Z}_k,\mathbf{z}_k)\Gamma(\mathbf{U}_l,\mathbf{u}_l)+\Gamma(\mathbf{U}_l,\mathbf{u}_l)^t\Gamma(\mathbf{Z}_k,\mathbf{z}_k)^t\ )\mathbf{w} \\ &\quad +\left(\frac{\tau^2}{12}\right)^2\mathbf{w}^t\ \Gamma(\mathbf{Z}_k,\mathbf{z}_k)\Gamma(\mathbf{U}_l,\mathbf{u}_l)(\Gamma(\mathbf{Z}_k,\mathbf{z}_k)\Gamma(\mathbf{U}_l,\mathbf{u}_l))^t\ \mathbf{w}.\end{aligned} \tag{2.31}$$

By adding and subtracting in (2.31) the term
$\frac{\tau^2}{12}\mathbf{w}^t(\ \Gamma(\mathbf{Z}_k,\mathbf{z}_k)\Gamma(\mathbf{Z}_k,\mathbf{z}_k)^t + \Gamma(\mathbf{U}_l,\mathbf{u}_l)^t\Gamma(\mathbf{U}_l,\mathbf{u}_l)\ )\mathbf{w}$, we obtain that

$$\begin{aligned}
\mathbf{w}^t(A_{k,l}A_{k,l}^t)\mathbf{w} \geq\ & 1 + \overbrace{\frac{\tau^2}{12}\mathbf{w}^t\left(\Gamma(\mathbf{Z}_k,\mathbf{z}_k) + \Gamma(\mathbf{U}_l,\mathbf{u}_l)^t\right)\left(\Gamma(\mathbf{Z}_k,\mathbf{z}_k)^t + \Gamma(\mathbf{U}_l,\mathbf{u}_l)\right)\mathbf{w}}^{\geq 0} \\
& -\tfrac{\tau^2}{12}(\ \mathbf{w}^t\ \Gamma(\mathbf{U}_l,\mathbf{u}_l)\Gamma(\mathbf{U}_l,\mathbf{u}_l)^t\ \mathbf{w} + \mathbf{w}^t\ \Gamma(\mathbf{Z}_k,\mathbf{z}_k)\Gamma(\mathbf{Z}_k,\mathbf{z}_k)^t\ \mathbf{w}) \\
\geq\ & 1 - \tfrac{C_{max}^2\tau^2}{6}.
\end{aligned}$$

3’) We consider the special case $C < 0$.
We prove 3’) taking into account that

$$\left(\frac{\tau^2}{12}\right)^2 \mathbf{w}^t\left(\Gamma(\mathbf{Z}_k,\mathbf{z}_k)\Gamma(\mathbf{U}_l,\mathbf{u}_l)\right)\left(\Gamma(\mathbf{Z}_k,\mathbf{z}_k)\Gamma(\mathbf{U}_l,\mathbf{u}_l)\right)^t\ \mathbf{w} > 0,$$

and applying Remark 5 to $\frac{\tau^2}{12}\mathbf{w}^t(\ \Gamma(\mathbf{Z}_k,\mathbf{z}_k)\Gamma(\mathbf{U}_l,\mathbf{u}_l)\ )\mathbf{w}$.

♢

Proposition 1. :
If the Jacobian matrix of the right hand side function in (2.8) takes the form $\mathcal{D}\boldsymbol{f}(\boldsymbol{u}) = K\mathcal{D}\boldsymbol{G}(\boldsymbol{u})$, where $K \in \mathbb{R}^{n\times n}$ is a constant matrix, and $\mathcal{D}\boldsymbol{G}(\boldsymbol{u}) \in \mathbb{R}^{n\times n}$ is a diagonal matrix with only positive or only negative elements, then it holds

$$\boldsymbol{w}^t(A_{k,l}A_{k,l}^t)\boldsymbol{w} + \frac{\tau^2}{36}\boldsymbol{w}^t(\ B_{k,l}B_{k,l}^t\ )\boldsymbol{w} \geq 1.$$

Proof:
Since $\mathbf{w}^t(\ B_{k,l}B_{k,l}^t\ )\mathbf{w} \geq 0$ we have to verify that $\mathbf{w}^t(A_{k,l}A_{k,l}^t)\mathbf{w} \geq 1$. Let $\mathcal{D}\mathbf{f}(\mathbf{u}) = K\ \mathcal{D}\mathbf{G}(\mathbf{u})$ satisfies the assumption of the proposition. Hence, the matrix $\Gamma(\mathbf{v},\mathbf{y})$ can be represented as

$$\Gamma(\mathbf{v},\mathbf{y}) \ = \int_0^1 K\mathcal{D}\mathbf{G}(\mathbf{y}+\sigma(\mathbf{v}-\mathbf{y}))d\sigma = K\int_0^1 \mathcal{D}\mathbf{G}(\mathbf{y}+\sigma(\mathbf{v}-\mathbf{y}))d\sigma$$

where $\int_0^1 \mathcal{D}\mathbf{G}(\mathbf{y}+\sigma(\mathbf{v}-\mathbf{y}))d\sigma$ is a diagonal matrix. Now, we have

$$\mathbf{w}^t(\Gamma(\mathbf{Z}_k,\mathbf{z}_k)\Gamma(\mathbf{U}_l,\mathbf{u}_l))\mathbf{w} \ = \ \mathbf{w}^t(KK^t\tilde{D}(\mathbf{Z}_k,\mathbf{z}_k,\mathbf{U}_l,\mathbf{u}_l))\mathbf{w},$$

where $\tilde{\mathbf{D}}(\mathbf{Z}_k,\mathbf{z}_k,\mathbf{U}_l,\mathbf{u}_l) := \int_0^1 \mathcal{D}\mathbf{G}(\mathbf{z}_k+\sigma(\mathbf{Z}_k-\mathbf{z}_k))d\sigma \int_0^1 \mathcal{D}\mathbf{G}(\mathbf{u}_l+\sigma(\mathbf{U}_l-\mathbf{u}_l))d\sigma$.
The matrix $\tilde{\mathbf{D}}(\mathbf{Z}_k,\mathbf{z}_k,\mathbf{U}_l,\mathbf{u}_l)$ is a diagonal matrix with positive elements and

therefore it is positive definite. Apparently, matrix KK^t is positive definite as well. Hence, we can say the same for $KK^t\tilde{D}(\mathbf{Z}_k, \mathbf{z}_k, \mathbf{U}_l, \mathbf{u}_l)$. Thus, we conclude that

$$\mathbf{w}^t \ (\ \Gamma(\mathbf{Z}_k, \mathbf{z}_k)\Gamma(\mathbf{U}_l, \mathbf{u}_l) \ ) \ \mathbf{w} \geq 0.$$

If we looking back at $\mathbf{w}^t(A_{k,l}A_{k,l}^t)\mathbf{w}$, we obtain the desired inequality.

♢

Lemma 4. :
If the constant C defined by (2.16) is such that $C < 0$ and $\boldsymbol{w}^t\boldsymbol{w} = 1$, then $\boldsymbol{w}^t(\ B_{k,l}A_{k,l}^t \ )\boldsymbol{w} \leq 0$.

Proof
If $C < 0$, then $A_{k,l}$ is positive definite as a sum of two positive definite matrices, and $B_{k,l}$ is negative definite according to Lemma 3. Let us rewrite our statement as $\mathbf{w}^t(\ B_{k,l}A_{k,l}^t \ )\mathbf{w} = -\mathbf{w}^t(\ (-B_{k,l})A_{k,l}^t \ )\mathbf{w}$ and examine the matrix $(-B_{k,l})A_{k,l}^t$. The matrices $(-B_{k,l})$ and $A_{k,l}$ are positive definite and according to proposition 3 from Appendix A the product $(-B_{k,l})A_{k,l}^t$ is also positive definite.

♢

Lemma 5. :
If Assumption 1 is fulfilled with a constant C defined by (2.16), then

1)

$$||(A_{k,l} + \frac{\tau}{6}B_{k,l})\left(A_{k,l} - \frac{\tau}{6}B_{k,l}\right)^{-1}|| \leq 1 + \tau \ K_1 \tag{2.32}$$

2)

$$||\left(A_{k,l} - \frac{\tau}{6}B_{k,l}\right)^{-1}|| \leq 1 + \tau K_2, \tag{2.33}$$

for $K_1 = K_1(\tau)$ and $K_2 = K_2(\tau)$ such that

- *If $C \leq 0$, then $-\tau^{-1} \leq K_i \leq 0$ for $i = 1, 2$ and all $\tau > 0$,*
- *If $C > 0$, then $K_1, K_2 > 0$ for*

$$\tau \leq \min(\frac{-3C + \sqrt{9C^2 + 6C_{max}^2}}{C_{max}^2}, \sqrt{\frac{12}{5C_{max}^2}}) = \sqrt{\frac{12}{5C_{max}^2}}.$$

Proof:

1) Let $P := [(A^t_{k,l} - \frac{\tau}{6}B^t_{k,l})(A_{k,l} - \frac{\tau}{6}B_{k,l})]^{-1}\ (A^t_{k,l} + \frac{\tau}{6}B^t_{k,l})(A_{k,l} + \frac{\tau}{6}B_{k,l})$. The matrix P is symmetric and positive definite. If we denote by

   $Q := (A_{k,l} + \frac{\tau}{6}B_{k,l})\left(A_{k,l} - \frac{\tau}{6}B_{k,l}\right)^{-1}$, then it follows that Q^tQ and P are similar $(P = (A_{k,l} - \frac{\tau}{6}B_{k,l})^{-1}Q^tQ(A_{k,l} - \frac{\tau}{6}B_{k,l}))$. Therefore, for the spectral radius ρ of both matrices P and QQ^t, is valid

$$\rho(\ P\ ) = \rho(\ QQ^t\ ) = ||Q||^2.$$

   By setting $P\mathbf{w} = \mu_1\mathbf{w},\ \mu_1 \geq 0,\ \mathbf{w}^t\mathbf{w} = 1$, we get

$$\begin{aligned}
\mu_1 &= \frac{\mathbf{w}^t(A^t_{k,l} + \frac{\tau}{6}B^t_{k,l})(A_{k,l} + \frac{\tau}{6}B_{k,l})\mathbf{w}}{\mathbf{w}^t(A^t_{k,l} - \frac{\tau}{6}B^t_{k,l})(A_{k,l} - \frac{\tau}{6}B_{k,l})\mathbf{w}} \\
&= 1 + \frac{\frac{2\tau}{6}\mathbf{w}^t(B^t_{k,l} + B_{k,l})\mathbf{w}}{\mathbf{w}^t(A^t_{k,l}A_{k,l} - \frac{\tau}{6}(B^t_{k,l} + B_{k,l}) + \left(\frac{\tau}{6}\right)^2 B^t_{k,l}B_{k,l})\mathbf{w}} \\
&\leq 1 + \frac{\frac{2\tau}{6}6C}{\mathbf{w}^t(A^t_{k,l}A_{k,l})\mathbf{w} - \frac{\tau}{6}6C + \left(\frac{\tau}{6}\right)^2 \mathbf{w}^t(B^t_{k,l}B_{k,l})\mathbf{w}} = 1 + 2\tau K_1.
\end{aligned}$$

   - The case $C \leq 0$.

     The constant K_1 is non-positive, and therefore $\mu_1 \leq 1$ without any restrictions on the time step.

   - The case $C > 0$.

     The constant $K_1 > 0$. Using Lemma 3 we bound K_1

$$0 < K_1 \leq \frac{C}{1 - C\tau - \frac{\tau^2}{6}C^2_{max}}.$$

     This is justified for the time step in the range

$$\tau \in (0, \frac{-3C + \sqrt{9C^2 + 6C^2_{max}}}{C^2_{max}}).$$

2) Let us set $P := (A^t_{k,l} - \frac{\tau}{6}B^t_{k,l})^{-1}(A_{k,l} - \frac{\tau}{6}B_{k,l})^{-1}$, and $Q := (A_{k,l} - \frac{\tau}{6}B_{k,l})^{-1}$. We have that $P = QQ^t$ is symmetric and positive semi-definite. Therefore, $\rho(P) \geq 0$, and $||Q|| = \sqrt{\rho(QQ^t)} = \sqrt{\rho(P)}$. Again, we take $P\mathbf{w} = \mu_2\mathbf{w},\ \mu_2 > 0,\ \mathbf{w}^t\mathbf{w} = 1$. Then, we get

$$\begin{aligned}
\mu_2 &= \frac{1}{\mathbf{w}^t((A_{k,l} - \frac{\tau}{6}B_{k,l})(A^t_{k,l} - \frac{\tau}{6}B^t_{k,l}))\mathbf{w}} \\
&= \frac{1}{\mathbf{w}^t(A_{k,l}A^t_{k,l})\mathbf{w} - 2\frac{\tau}{6}\mathbf{w}^t(B_{k,l}A^t_{k,l})\mathbf{w} + \left(\frac{\tau}{6}\right)^2 \mathbf{w}^t(B_{k,l}B^t_{k,l})\mathbf{w}}.
\end{aligned}$$

- The case $C \leq 0$.
According to Lemma 4, $\mathbf{w}^t(B_{k,l}A^t_{k,l})\mathbf{w} \leq 0$, and because of $\mathbf{w}^t(A_{k,l}A^t_{k,l} + \frac{\tau^2}{36}B_{k,l}B^t_{k,l})\mathbf{w} > 1$, it holds that $0 \leq \mu_2 \leq 1$ respectively, $||Q|| \leq 1$.

- The case $C > 0$.
We have that

$$\mu_2 \quad \leq \quad \frac{1}{\left(\sqrt{\mathbf{w}^t(A_{k,l}A^t_{k,l})\mathbf{w}} - \frac{\tau}{6}\sqrt{\mathbf{w}^t(B_{k,l}B^t_{k,l})\mathbf{w}}\right)^2}.$$

Then, for the spectral norm of the matrix Q, it is fulfilled that

$$||Q|| = \sqrt{\mu_2} = \frac{1}{|\sqrt{\mathbf{w}^t(A_{k,l}A^t_{k,l})\mathbf{w}} - \frac{\tau}{6}\sqrt{\mathbf{w}^t(B_{k,l}B^t_{k,l})\mathbf{w}}|}.$$

In terms of τ and C_{max} one can make the following estimate

$$||Q|| \leq \frac{1}{\sqrt{1 - \frac{\tau^2 C^2_{max}}{6}} - \frac{\tau}{6}\sqrt{9C^2_{max}}} = 1 + \tau K_2,$$

where $K_2 > 0$ and $\tau \leq \sqrt{\frac{12}{5C^2_{max}}}$.

This ends the proof of this Lemma.

◇

Now we are ready to prove Theorem 3.

Proof of Theorem 3

Applying Lemmas 2,3,5 and the inequality (2.30), we get:

$$\begin{aligned} ||\tilde{e}_{j,j+1}|| &\leq ||(A_{j,j} + \frac{\tau}{6}B_{j,j})\left(A_{j,j} - \frac{\tau}{6}B_{j,j}\right)^{-1}||\ ||\tilde{e}_{j,j}|| \\ &\quad +||L(\tau, u_j, u_{j+1}) - L(\tau, U_j, U_{j+1})|| \\ &\leq (1 + \tau K_1)||\tilde{e}_{j,j}|| + ||L(\tau, u_j, u_{j+1}) - L(\tau, U_j, U_{j+1})|| \\ &\leq \sum_{k=0}^{j}(1 + \tau K_1)^{j-k}||L(\tau, u_k, u_{k+1}) - L(\tau, U_k, U_{k+1})||. \end{aligned}$$

Making use of (2.28), we obtain for the error e_{j+1}

$$\begin{aligned} ||e_{j+1}|| &\leq ||(A_{j,j+1} - \tfrac{\tau}{6}B_{j,j+1})^{-1}|| \left(\sum_{k=0}^{j}(1 + \tau K_1)^{j-k}||L(\tau, u_k, u_{k+1}) - L(\tau, U_k, U_{k+1})||\right) \\ &\leq (1 + \tau K_2)\left(\textstyle\sum_{k=0}^{j}(1 + \tau K_1)^{j-k}||L(\tau, u_k, u_{k+1}) - L(\tau, U_k, U_{k+1})||\right). \end{aligned} \tag{2.34}$$

- For $C \leq 0$, according to Lemma 5, the constants $K_1,\ K_2 \geq 0$ are nonpositive, which give
$$||e_{j+1}|| \leq \frac{(1+K_1\tau)^{j+1}-1}{K_1\tau} d(\tau) \leq (j+1)d(\tau).$$
- For $C > 0$ we use that $\sum_{k=0}^{j}(1+\tau K_1)^{j-k} \leq (j+1)e^{TK_1}$ and $1+\tau K_2 =: \frac{1}{\sqrt{1-\frac{\tau^2 C_{max}^2}{6}}-\frac{\tau}{2}C_{max}}$ to obtain
$$||e_{j+1}|| \leq \frac{d(\tau)(j+1)e^{K_1 T}}{\sqrt{1-\frac{\tau^2 C_{max}^2}{6}}-\frac{\tau}{2}C_{max}}.$$
The last is valid for $\tau \leq \sqrt{\frac{12}{5C_{max}^2}}$.

♢

2.4 The case of quasilinear parabolic equations

In this section, we derive an estimate of the error of approximation in the case of quasilinear parabolic problems. First, we consider the error coming from the time discretization (time error). Here, we make use of the results we have obtained in the previous section, for the case of system of ODEs. We prove a theorem and some auxiliary statements for a special case of quasilinear parabolic equations. In order to get an estimate of the global error of approximation, we incorporate the estimate of the error coming from the space discretization (space error) into the already derived estimate of the time error. Since the procedure for obtaining the estimate of the space error is a standard one, we do not pay a lot of attention on it here.

We consider the following class of quasilinear parabolic problems in $\Omega_T = (0,T)\times(a,b)$

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial}{\partial x}(D(u)\frac{\partial u}{\partial x}) + S(x), \quad (t,x)\in\Omega_T, \qquad (2.35)\\ u(0,x) &= u_0(x), \quad x\in[a,b] \\ \mathcal{B}_a u|_{x=a} &= 0, \quad \mathcal{B}_b u|_{x=b} = 0, \qquad t\in[0,T] \end{aligned}$$

where $\mathcal{B}_a u = \alpha_1 u - \alpha_2 u_x$, $\mathcal{B}_b u = \beta_1 u + \beta_2 u_x$ and $\alpha_i,\ \beta_i$ are nonnegative and $\alpha_1 + \alpha_2 > 0,\ \beta_1+\beta_2 > 0$. Here the function $D(u) \geq 0$ and $D \in C^2(0,\infty)\bigcap C([0,\infty))$.

By setting

$$G(u) := \int_{u_0}^{u} D(\eta)d\eta$$

and by taking the derivative of $G(u)$ with respect to x we obtain the following equation

$$\frac{\partial G(u)}{\partial x} = \frac{\partial G}{\partial u}\frac{\partial u}{\partial x} = D(u)\frac{\partial u}{\partial x}.$$

We rewrite (2.35) as

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 G(u)}{\partial x^2} + S(x), \qquad (t,x) \in \Omega_T \qquad (2.36)\\ u(0,x) &= u_0(x), \qquad x \in [a,b]\\ \mathcal{B}_a u|_{x=a} &= 0, \quad \mathcal{B}_b u|_{x=b} = 0, \qquad t \in [0,T]. \end{aligned}$$

Let W be an appropriately chosen space of functions with scalar product $(.,.)_{\Omega_T}$ and norm $||.||_W$. If we look for a classical solution, then $W \equiv C^{1,2}(\bar{\Omega}_T)$. If we look for a weak solution, then W consists of possible less times continuously differentiable functions depending on the differentiability of the coefficients making up the parabolic equation.

Now, let W_h be the finite element subspace of W, where we seek for an approximate solution of (2.36). Let us divide the interval $[a,b]$ by $\{x_i\}_{i=0}^n$, and let $\{\phi_i(x)\}_{i=0}^n$ be the Lagrangian finite element basis in W_h (i.e. $\phi_i(x_j) = \delta_{ij}$). Now, let us multiply (2.36) by a test function $v \in W_h$ and integrate in $[a,b]$ with respect to x. Using integration by parts we get the weak Galerkin form of the problem

$$\int_a^b \partial_t u(t,x)v(x)dx = -\int_a^b G_x v'(x)dx + \int_a^b S(x)v(x)dx. \qquad (2.37)$$

Each function $p \in W_h$ can be represented as $p(x) = \sum\limits_{i=1}^{n} p_i\phi_i(x)$. We identify v in (2.37) with the basis functions $\phi_j(x)$ and we get a system of n ODEs. In this way, the time-dependant PDE is transformed into the following system of ODEs

$$\begin{aligned} \dot{\mathbf{U}} &= -\frac{1}{h^2}\tilde{M}^{-1}K\bar{G}(\mathbf{U}) + \bar{\mathbf{S}}, \qquad (2.38)\\ \mathbf{U}(0) &= \mathbf{U}^0, \end{aligned}$$

where $\mathbf{U} = (U(t,x_0),\cdots,U(t,x_n))^T \in D_T \subseteq \mathbb{R}^n$ is the semi-discrete solution, $\tilde{M}^{-1}$ and K are constant matrices, $\bar{G}(U) = (G(U(t,x_0)),\cdots,G(U(t,x_n)))^T$, and $\bar{\mathbf{S}} = (S(x_0),\cdots,S(x_n))^T$ is the source vector.

In this section and throughout the end of the thesis, unlike in the previous sections, the subscript denotes space integration whereas the superscript denotes time integration.

We assign the right hand side of (2.38) to **f(U)**, i.e.

$$\mathbf{f}(\mathbf{U}) := -\frac{1}{h^2}\tilde{M}^{-1}K\mathbf{G}(\mathbf{U}) + \bar{S}. \tag{2.39}$$

- **Incorporation of the spatial error into the defect of Lobatto III A scheme**

  The terms describing the spatial error in the r.h.s. function **f(U)** are actually neglected. Taking them into account **f(U)** reads

  $$-\frac{1}{h^2}\tilde{M}^{-1}K\mathbf{G}(\mathbf{U}_h) + \bar{S} + 2^p h^p \cdot Const =: g(U_h),$$

  where p is the order of convergence of the used FEM. Here, we emphasize the dependency on the space step size h through a subscript h. The term $2^p h^p \cdot Const$ is referred to as the spatial error. Let us write the defect of Lobatto III A scheme for the system of ODEs with r.h.s $g(\mathbf{U}_h) = f(\mathbf{U}_h) + 2^p h^p \cdot Const$ and for the ODEs with r.h.s. $f(\mathbf{U}_h)$, and subtract them

  $$\text{Lobatto for } g(\mathbf{U}_h) = \text{Lobatto for } f(\mathbf{U}_h) - \tau \cdot Const \cdot h^p. \tag{2.40}$$

  In order to calculate the last term in (2.40) we may use the idea of Richardson [18]. First, we compute the PDE with spatial step size h to obtain a solution U_h. Then, we compute with "big" spatial step $2h$ to obtain a solution U_{2h}, i.e.

  $$\begin{aligned} \frac{d\mathbf{U}_h}{dt} &= f(\mathbf{U}_h) + 2^p h^p \cdot Const && (2.41)\\ \frac{d\mathbf{U}_{2h}}{dt} &= f(\mathbf{U}_{2h}) + 2^p (2h)^p \cdot Const. && (2.42) \end{aligned}$$

  Now, applying the trapezoidal rule to both semidiscrete equations (2.41) (2.42), for $t = t_1$ (the upper index denotes the number of the time step)

  $$\begin{aligned} \mathbf{U}_h^1 &= \mathbf{U}_h^0 + \frac{\tau}{2}(f(\mathbf{U}_h^0) + f(\mathbf{U}_h^1)) + \tau 2^p h^p \cdot Const \\ \mathbf{U}_{2h}^1 &= \mathbf{U}_{2h}^0 + \frac{\tau}{2}(f(\mathbf{U}_{2h}^0) + f(\mathbf{U}_{2h}^1)) + \tau 2^p (2h)^p \cdot Const, \end{aligned}$$

  and taking their difference at the common grid points, we get that

  $$\tau \cdot Const \cdot h^p = \frac{1}{2^{2p} - 2^p}(U_{2h}^1 - U_h^1 + \frac{\tau}{2}(f(\mathbf{U}_{2h}^1) - f(\mathbf{U}_h^1))).$$

> An example, where the spatial error is included into the Lobatto time estimate, is given in Chapter 4.
>
> The expression (2.40) gives the way to incorporate the spatial error into the defect of Lobatto III A scheme. The latter is used to estimate the approximation error. Theorem 3, from the previous section, gives such an estimate but now the estimated error includes the spatial and the time error.

Taking the specific form of the r.h.s. function **f(U)**, defined by (2.39), Theorem 3 could be formulated in the following way.

Theorem 4. :
*Let the r.h.s. function $f(\mathbf{u})$ defined by (2.39) be such that Assumption 1 holds with a constant C defined by (2.16) and $C > 0$. Let **f(U)** be such that its Jacobian matrix is given by $\mathcal{D}\boldsymbol{f(U)} = K\mathcal{D}G(\boldsymbol{U})$, where $K \in \mathbb{R}^{n\times n}$ is a constant matrix, $\mathcal{D}G(\boldsymbol{U}) \in \mathbb{R}^{n\times n}$ is a diagonal matrix with only positive or only negative elements. Moreover, let $\{U_j\}_j$, $\mathbf{U}_j \in D_T$ be an approximation of the solution of (2.38) obtained through trapezoidal rule (TR). Then, the global error of approximation satisfies*

$$||\boldsymbol{e}_{j+1}|| \leq (j+1)\frac{e^{K_1 T}}{1-\frac{\tau}{2}C_{max}}d(\tau),$$

for $\tau \leq \min(\frac{1}{C}, \frac{2}{C_{max}})$, $K_1 = \frac{C}{1-C\tau}$, and C_{max} defined by (2.17).

Proof:
With the help of Proposition 1 we reconsider Lemma 5 and Theorem 3.
In Lemma 5, the constant K_1 could be set to

$$K_1 := \frac{C}{1-C\tau}, \qquad \text{for } \tau \in [0, \frac{1}{C}].$$

We can assign to the constant K_2

$$K_2 := \frac{C_{max}}{2-\tau C_{max}}, \qquad \text{for } \tau \in [0, \frac{2}{C_{max}}].$$

Inserting the new values of K_1 and K_2 in Theorem 3 we obtain for the error

$$||\mathbf{e}_{j+1}|| \leq (j+1)\frac{e^{K_1 T}}{1-\frac{\tau}{2}C_{max}}d(\tau), \qquad \text{for } \tau \leq \min(\frac{1}{C}, \frac{2}{C_{max}}).$$

◇

The next proposition, and in particular the corollary after it, handles the constant C and shows a way to calculate it.

Proposition 2. : *Let $G, A : D_T \subseteq \mathbb{R}^n \to \mathbb{R}^{n\times n}$ and let $\mathbf{U_{min}} = \min\limits_{\mathbf{U}\in\mathbf{D_T}} \mathbf{U}$ and $\mathbf{U_{max}} = \max\limits_{\mathbf{U}\in\mathbf{D_T}} \mathbf{U}$ (we allow $\mathbf{U_{min}} = -\infty$ or $\mathbf{U_{max}} = \infty$). The matrix $G(\boldsymbol{U})$ is a positive definite diagonal matrix, i.e. $G(\boldsymbol{U}) = diag(G_1(\boldsymbol{U}), \cdots, G_n(\boldsymbol{U}))$, $G_i(\boldsymbol{U}) \geq 0$, $i = 1, \cdots, n$, and let $\mathcal{D}G_i(\boldsymbol{U}) \geq 0$. Let the matrix A be $A(\boldsymbol{U}) = KG(\boldsymbol{U})$ with $K \in \mathbb{R}^{n\times n}$ constant matrix.*

a) If K is a negative definite constant matrix, then

$$\max_{\boldsymbol{U}\in[\boldsymbol{U}_{min},\boldsymbol{U}_{max}]} \mu[A(\boldsymbol{U})] = \mu[A(\boldsymbol{U}_{min})].$$

b) If K is a positive definite constant matrix, then

$$\max_{\boldsymbol{U}\in[\boldsymbol{U}_{min},\boldsymbol{U}_{max}]} \mu[A(\boldsymbol{U})] = \mu[A(\boldsymbol{U}_{max})].$$

c) if K is an indefinite constant matrix, then

$$\max_{\boldsymbol{U}\in[\boldsymbol{U}_{min},\boldsymbol{U}_{max}]} \mu[A(\boldsymbol{U})] = \mu[A(\boldsymbol{U}_{max})].$$

Proof:

a) Function $G(\mathbf{U})$ is increasing, and therefore for any $\mathbf{V}, \mathbf{W} \in D_T$, such that $\mathbf{V} \leq \mathbf{W}$, follows $G(\mathbf{V}) \leq G(\mathbf{W})$. Now, let us consider

$$\begin{aligned} \mathbf{w}^t A(\mathbf{V})\mathbf{w} - \mathbf{w}^t A(\mathbf{W})\mathbf{w} &= \mathbf{w}^t(KG(\mathbf{V}) - KG(\mathbf{W}))\mathbf{w} \\ &= \mathbf{w}^t\ K(G(\mathbf{V}) - G(\mathbf{W}))\mathbf{w} \\ &= \mathbf{w}^t(-K)(G(\mathbf{W}) - G(\mathbf{V}))\mathbf{w}. \qquad (2.43) \end{aligned}$$

Both matrices, $(-K)$ and $(G(W)-G(V))$, are positive definite (the second one as a diagonal matrix with positive elements). Hence, the same is true for their product

$$\mathbf{w}^t A(\mathbf{V})\mathbf{w} - \mathbf{w}^t A(\mathbf{W})\mathbf{w} \geq 0, \qquad \forall \mathbf{w} \in D_T : \quad \mathbf{w}^t\mathbf{w} = 1.$$

For the logarithmic matrix norms of $A(\mathbf{V})$, and of $A(\mathbf{W})$, we obtain

$$\mu[A(\mathbf{V})] \geq \mu[A(\mathbf{W})].$$

From here we conclude that

$$\mu[A(\mathbf{U}_{min})] \geq \mu[A(\mathbf{U})], \qquad \forall \mathbf{U} \in D_T.$$

b) Now, we consider the case when K is positive definite. Looking back at (2.43) we note that

$$\mathbf{w}^t A(\mathbf{V})\mathbf{w} - \mathbf{w}^t A(\mathbf{W})\mathbf{w} \leq 0,$$

therefore, it holds

$$\mu[A(\mathbf{V})] \leq \mu[A(\mathbf{W})], \qquad \text{for any } \mathbf{V} \leq \mathbf{W}.$$

Because $G(\mathbf{U})$ is an increasing function $G(\mathbf{U}_{max}) = \max\limits_{\mathbf{U}} G(\mathbf{U})$, and for $\mathbf{U}_{max}$ is fulfilled

$$\mu[A(\mathbf{U}_{max})] \geq \mu[A(\mathbf{U})], \qquad \forall \mathbf{U} \in D_T.$$

c) In this case, matrix K is indefinite i.e. $\mu[K] > 0$. The function $G(\mathbf{U})$ is positive and increasing therefore $G(\mathbf{U}_{min}) \leq G(\mathbf{U}) \leq G(\mathbf{U}_{max})$ for $\mathbf{U} \in D$. Without loss of generality, we assume that there are a_{min} and a_{max}, positive constants, such that $0 < a_{min} \leq a_{max}$ and $G(\mathbf{U}_{min}) = a_{min}I$, $G(\mathbf{U}_{max}) = a_{max}I$. Therefore, $A(\mathbf{U}_{min}) = KG(\mathbf{U}_{min}) = a_{min}K$ and $A(\mathbf{U}_{max}) = KG(\mathbf{U}_{max}) = a_{max}K$. In this case using the 4^{th} property of the logarithmic matrix norm, Appendix A, we obtain that $\mu[A(\mathbf{U}_{min})] = a_{min}\mu[K] \leq a_{max}\mu[K] = \mu[A(\mathbf{U}_{max})]$.
Hence,

$$\mu[A(\mathbf{U}_{min})] \leq \mu[A(\mathbf{U})] \leq \mu[A(\mathbf{U}_{max})], \qquad \forall \mathbf{U} \in D.$$

$\diamondsuit$

Corollary 1. :

- *For function $f(\boldsymbol{U})$ with the Jacobian matrix $\mathcal{D}f(\boldsymbol{U})$, such that $\mu[\mathcal{D}f(\boldsymbol{U})] < 0$, the constant C can be computed as the maximal eigenvalue of the symmetrical part of the Jacobian matrix calculated at the initial solution $\boldsymbol{U}^0$, that is*

$$C = \lambda_{max}\left[\frac{\mathcal{D}f(\boldsymbol{U}^0) + \mathcal{D}f(\boldsymbol{U}^0)^t}{2}\right].$$

- *For $\mathcal{D}f(\boldsymbol{U})$, which satisfies $\mu[\mathcal{D}f(\boldsymbol{U})] > 0$, the constant C is*

$$C = \lambda_{max}\left[\frac{\mathcal{D}f(\boldsymbol{U}_{max}) + \mathcal{D}f(\boldsymbol{U}_{max})^t}{2}\right],$$

where $\boldsymbol{U}_{max}$ is the argument in which the Jacobian matrix gets its maximum in D.

Let us look at C

$$C = \max_{\mathbf{U}\in D} \lambda_{max} \left[\frac{\mathcal{D}f(\mathbf{U}) + \mathcal{D}f(\mathbf{U})^t}{2}\right] = \max_{\mathbf{U}\in D} \lambda_{max} \left[-\frac{\tilde{M}^{-1}K\mathcal{D}\bar{G}(\mathbf{U}) + (\tilde{M}^{-1}K\mathcal{D}\bar{G}(\mathbf{U}))^t}{2h^2}\right].$$

For the case of $C > 0$, the Jacobian matrix fits in the requirements of Theorem 4. Hence $\tau \leq \min(\frac{1}{C}, \frac{2}{C_{max}})$, which leads to a restriction on the ratio of the time and space step

$$\tau \leq h^2 \min(\frac{1}{\max\limits_{\mathbf{U}\in D} \mu[-\tilde{M}^{-1}K\mathcal{D}\bar{G}(\mathbf{U})]}, \frac{2}{\max\limits_{\mathbf{U}\in D} ||\tilde{M}^{-1}K\mathcal{D}\bar{G}(\mathbf{U})||}),$$

that is

$$\frac{\tau}{h^2} \leq \min(\frac{1}{\max \mu[-\tilde{M}^{-1}K\mathcal{D}\bar{G}(\mathbf{U}_{max})]}, \frac{2}{\max ||\tilde{M}^{-1}K\mathcal{D}\bar{G}(\mathbf{U}_{max})||}).$$

We finish this chapter with a discussion of the question whether we overdo using the defect of the 4^{th} order Lobatto III A scheme. At first glance, it seems that one would do the same work as if one used the 4^{th} order method, but this is not the case. The 4^{th} order method requires one additional calculation of the Jacobian matrix. This calculation takes a substantial part in the numerical effort. Let us stress also to the fact that the time integrator of AIM (the Trapezoidal rule plus the defect of the 4^{th} order Lobatto III A scheme) provides with an efficient tool for adapting the time step, controlling the number of the Newton iterations and an error estimate. In Ch.4, it is given several examples testing the time integrator of AIM against the 4^{th} order Lobatto scheme. The first one performs well but not always better than the 4^{th} order Lobatto III A scheme. However, for the anomalous heat transport problem the time integrator of AIM gives approximately 1.7 times better results.

Chapter 3

The Anomalous Transport

In this chapter we consider in details the topic of this thesis - the anomalous heat transport in a tokamak plasma. We start with some very brief introductory notes about the physics behind the heat transport in the plasma and the energy equation describing it. In the second section, we formulate the mathematical problem as a result of the dimensionless physical problem. Furthermore, in section 3.3, we introduce the idea of the Front Tracking Technique (FTT). For this purpose we define the concepts - *solution* of the considered problem and *front point*. In the same section we investigate the question of smoothness of the solution, and some additional considerations about the point we call front point. In section 3.4, we derive some exact solutions for a simpler version of the underlying problem. Later on, in section 3.5, we proceed to the numerical treatment - space and time discretization. Here we make use of the results of Ch.2. Finally, in section 3.6 we give a convergence analysis of the applied FTT.

3.1 Physical considerations

To produce energy through thermonuclear controlled fusion is still a challenging goal for the fusion community.

Fusion of deuterium and tritium occurs in appreciable amount if the plasma temperature gets over 10 millions K. At such temperature the fuel is fully ionized. The electrostatic charge of the nuclear ions is neutralized by the presence of an equal number of electrons and the resulting neutral gas is called plasma.

In the last 50 years many different techniques were investigated in order to approach the fusion reaction. Among them, the most advanced concept is the so - called Tokamak, realized for the first time by L. Artsimovitch in 1952. The Tokamak is a toroidal system which confines the plasma by a magnetic field [51]. The dominant magnetic field is the toroidal one, produced by external coils. However, this field alone does not allow the confinement of the plasma and an additional poloidal magnetic field is necessary for an equilibrium. This additional magnetic field is produced by a large toroidal current in the plasma and by outer poloidal field coils. The combination of the toroidal and the poloidal fields results in magnetic field lines with helical trajectories around the torus lying on the so called magnetic surfaces.

Due to the magnetic field a charged particle executes a gyrating along a magnetic line. The particles collide with each other and the collisions cause a displacement of their orbits. These displacements are random and so the particles diffuse across the magnetic field until they reach the edge of the plasma. When a particle makes a collisional step it takes its energy with it. In this way it is realized a diffusive transfer of heat in radial direction. Due to this fact and the axisymmetry of the considered geometry the natural dimension for the mathematical problem is one.

In tokamak experiments it is found that the transport coefficients are much larger then the anticipated from collisions. The name “anomalous” was given because the process was not understood. Moreover, the actual transport was much greater than expected, especially for electrons – electron particle and heat fluxes are one- to ten-thousand times their classical levels. This is one of the reasons fusion in plasmas confined by a magnetic field has not been achieved yet.

A particular class of drift instabilities which has been recently proposed as the most probable candidate to explain the anomalous ion thermal transport in tokamaks is the ion temperature gradient (ITG) driven drift mode [21].

For better understanding, let us start from the simplest form of the energy equation

$$n_j \frac{\partial u_j}{\partial t} = -\frac{1}{r}\frac{\partial q_j}{\partial r} + S_j = \frac{1}{r}\frac{\partial}{\partial r}\left(r n_j \chi_j^{PB} \frac{\partial u_j}{\partial r} \right) + S_j, \quad r \in [0, a],\ t > 0,$$

where the labels refer to electrons and ions ($j = e, i$), n_j and u_j represent the particle density and temperature, respectively, and q_j is the heat flux. The function S_j represents a source, and it could be decomposed to

$$S_j = S_{Qj}^{in} - S_{Qj}^{out},$$

where S_{Qj}^{in} summarizes the positive power sources and S_{Qj}^{out} the negative. In the literature, χ_j^{PB} is often referred to as a power balance diffusivity. It is in principle a function of the local plasma parameters as ∇n_j, ∇u_j, u_j, the safety factor q, Z^{eff}, the magnetic shear $\hat{s}$, the plasma collisionality, and the ratio u_e/u_i. The typical value for the upper bound of r, a (called minor radius) is between 0.5m (for ASDEX - the tokamak device in Garching, Germany) and 2m (for ITER - the future international tokamak in France). For the diffusivity we have the range $0.01 m^2/s \leq \chi_j^{PB} \leq 10^2 m^2/s$. The source S_j can be expressed as the ratio between the total power $\mathcal{P}$ and the considered volume, i.e. $S_j = \frac{\mathcal{P}}{2\pi^2 R a^2}$. Here, R is the major radius and it is between 1.5m (ASDEX) and 6m (ITER) and the total power $\mathcal{P}$ is in the range [1MW, 100MW].

Dimensionless variables are defined by introducing the length of the minor radius a and by measuring the time in units of $1/t_0$; t_0 is typically 0.1s:

$$\bar{r} = \frac{r}{a}, \quad \bar{t} = \frac{t}{t_0}.$$

The equation can be rewritten in terms of dimensionless variables

$$\bar{n}_j \frac{\partial \bar{u}_j}{\partial \bar{t}} = \frac{1}{\bar{r}} \frac{\bar{r}\partial}{\partial \bar{r}}(\bar{r}\bar{n}_j \bar{\chi}_j^{PB} \frac{\partial \bar{u}_j}{\partial \bar{r}}) + \bar{S}_j, \quad \bar{r} \in [0,1],\ \bar{t} > 0.$$

The dimensionless variable $\bar{\chi}_j^{PB}$ is staying in the same range $[0.01, 10^2]$ and the source $\bar{S}_j \approx \mathcal{P}$. From now on we consider the dimensionless problem and we skip the bars in the notations.

Two are the most often considered scenario about the heat flux. In the first one, significant contributions to the heat flux are driven by the temperature. The power balance diffusivity χ^{PB} does not represent anymore the thermal diffusivity and exhibits a complicated dependence on the temperature gradient. In this case the problem under consideration takes the form

$$\begin{aligned}
\frac{\partial u}{\partial t} &= \frac{1}{x^{d-1}} \frac{\partial}{\partial x}(x^{d-1} u^{\sigma} \frac{\partial u}{\partial x}) + u^{\beta}, \qquad x \in (0,L),\ t \in (0,t^*), \\
(x^{d-1} u^{\sigma} \frac{\partial u}{\partial x})\Big|_{x=0} &= u(t,L) = 0, \quad t \in [0,t^*], \\
u(0,x) &= g(x), \qquad x \in [0,L].
\end{aligned}$$

The set of parameters $\sigma = 1.5,\ \beta = 5/2,\ d = 1$ represents the typical diffusion in a Tokamak fusion plasma [30]. The second scenario for the heat flux is the subject of this thesis and is described in the next section.

3.2 Mathematical model

Difficulties occur, if the model incorporates the additional transport, the anomalous transport, triggered by temperature gradient above a certain threshold. Such scenario can be formulated as the following mathematical problem, that is subject of this chapter

$$P : \begin{cases}
\frac{\partial u}{\partial t} = \frac{1}{x^{d-1}} \frac{\partial}{\partial x}(x^{d-1}(D_0 + D_1 H(|u_x| - \bar{u})(|u_x| - \bar{u}))\frac{\partial u}{\partial x}) + S(x), \\
\qquad \text{for} \quad (t,x) \in \Omega_T = (0,T) \times (0,1) \\
u(0,x) = u_0(x), \quad x \in [0,1] \\
u_x(t,0) = 0, \quad t \in [0,T], \\
u(t,1) = 0, \quad t \in [0,T],
\end{cases} \tag{3.1}$$

where

d - defines the geometry, $d = 1$ - slab geometry, $d = 2$ cylindrical and $d = 3$ spherical,

$D_0,\ D_1 = Const \geq 0$,

$H(|u_x| - \bar{u})$ - the Heaviside function,

$S(x) \geq 0$ is a source function. Meaningful for the anomalous heat transport problem is

$$S(x) = S_0 e^{-\frac{(x-x_0)^2}{\delta^2}}, \tag{3.2}$$

$x_0 \in [0, 1]$, $\delta > 0,\ S_0 = Const \geq 0$.

The parameter $\bar{u}$ is the threshold value for the temperature gradient and it is considered to be constant. The Heaviside function is defined as

$$H(x) := \begin{cases} 0, & x \leq 0 \\ 1, & x > 0. \end{cases}$$

The heat flux is defined as

$$q(t, x) := x^{d-1}(D_0 + D_1 H(|u_x| - \bar{u})(|u_x| - \bar{u}))u_x. \tag{3.3}$$

We consider two strategies for solving the anomalous transport problem.

- The first one is the AIM strategy. This is treating the whole problem, using the method of lines [42], [47]. That is the finite element method (FEM) transforms the problem P to a system of ODEs, which is solved through the trapezoidal rule using the Lobatto estimate.

- The second strategy is using an explicit front tracking technique. Briefly said, the problem P is split into two subproblems P1 and P2 at the point where the gradient reaches the critical value $\bar{u}$. We refer to this point as a front point or an interface. On each side of the interface we apply the AIM approach. In addition an equation tracking the position of the front is considered. This strategy we call Front Tracking Technique (FTT) and this chapter is devoted to it.

3.3 Front Tracking

We start this section with a definition of solution of problem (3.1) and the point we refer to as a front point or interface. In what follows we consider the case $d = 1$, but the extension to $d > 1$ is straightforward.

Definition 11. *We say that a function $u : B \subset \mathbb{R}^2 \to \mathbb{R}$, $\bar{\Omega}_T \subset B$ is a* **solution of problem P** *if $u \in C^{1+\alpha/2,1+\alpha}(\bar{\Omega}_T)$ for some $\alpha \in (0,1)$, u satisfies problem P a.e. and u_{xx} is defined and piece-wise continuous in $\bar{\Omega}_T$.*

Remark 6. *If u is a solution of problem P then $u_x \in C^{1+\alpha/2,\alpha}(\bar{\Omega}_T)$, but in addition u_{xx} is piece-wise continuous in $\bar{\Omega}_T$, therefore u_x is even Lipschitz continuous in $\bar{\Omega}_T$ with respect to x.*

Definition 12. *Let $u = u(t,x)$ be a solution of problem P, $\bar{u} > 0$ given. $x_F \in (0,1)$ is called* **(non-degenerate) front point at** *t if $|u_x(t,x_F)| = \bar{u}$ and if both $\lim\limits_{x \to x_F^-} u_{xx}(t,x) \neq 0$ and $\lim\limits_{x \to x_F^+} u_{xx}(t,x) \neq 0$. A point x_F is called* **degenerate front point** *if $|u_x(t,x_F)| = \bar{u}$ and $\lim\limits_{x \to x_F^-} u_{xx}(t,x) = 0$ and/or $\lim\limits_{x \to x_F^+} u_{xx}(t,x) = 0$.*

Remark 7 (Degenerate front points)**.** *At a degenerate front point in anomalous transport problems $|u_x|$ might cross the line $\bar{u}$ at a saddle point or it might touch the line $\bar{u}$ in a local minimum or maximum.*

- *The case that $|u_x|$ crosses the line $\bar{u}$ at a saddle point was never observed in our anomalous transport studies. It thus has not been investigated and is not considered here.*

- *The case of touching of the line $\bar{u}$ at $\bar{x}$ at a local maximum or minimum, without crossing, is possible and does occur in anomalous transport problems. It is important only if anomalous transport sets in or ceases to happen at $\bar{x}$. In the first case it gives rise to two additional front points for larger t, in the other case a pair of front points disappears in $\bar{x}$. Both cases are shown to happen in the example leading to Ch.4, subsection 4.3.2, 'Multiple front points'.*

- *What about $|u_x| = \bar{u}$ in a closed subinterval $[\bar{x}^I, \bar{x}^{II}] \subset (0,1)$ with or without crossing of the line $\bar{u}$ before and afterwards? In this case $u_{xx}(t,x) \equiv 0$ in $[\bar{x}^I, \bar{x}^{II}]$ and problem P reduces locally to the ordinary initial value problem*

  $$\frac{du}{dt} = S(x) \quad \text{in } [\bar{x}^I, \bar{x}^{II}], \quad u(0,x) = u_0(x),$$

  depending on a parameter x. It can be integrated analytically as long as an x-interval with $u_{xx}(t,x) \equiv 0$ exists. Special sources $S(x)$ and initial conditions $u_0(x)$ will allow such solutions. The sources relevant to the anomalous transport problem will not allow such x-intervals to persist. Though the case of Turing bifurcations *[36] is mathematically interesting, we will not enter this field here since irrelevant to anomalous transport.*

Note that there are two possible cases: $u_x(t, x_F) = \bar{u}$ or $u_x(t, x_F) = -\bar{u}$ for some $x_F \in (0, 1)$.

If the initial function u_0 possesses a non-degenerate front point then this point exists at least in a small $t-$interval $[0, T^*)$.

Theorem 5. *Let u be a solution of problem P and let $\int\limits_0^1 u_{tt}(t,x)dx$ be bounded. If there exists a non-degenerate front point $x_{F,0}$ at $t_0 = 0$ with $|u_x(0, x_{F,0})| = \bar{u}$, then there are an interval $[0, T^*)$ and a C^1-function $x_F(t)$ on $[0, T^*)$ such that $x_F(t)$ is a non-degenerate front point for every $t \in [0, T^*)$, satisfying*

$$u_x(t, x_F(t)) = \bar{u}\ sgn(u_x(t, x_{F,0})) \qquad \textit{and} \qquad x_F(0) = x_{F,0}. \tag{3.4}$$

The velocity of the front point is given by

$$\dot{x}_F(t) = -\frac{u_{xt}(t, x_F)}{u_{xx}(t, x_F)}. \tag{3.5}$$

Proof

Assume that there exists a non-degenerate front point $x_{F,0}$ at $t_0 = 0$ with $|u_x(0, x_{F,0})| = \bar{u}$. Without loss of generality we consider the case

$$u_x(0, x_{F,0}) = \bar{u}. \tag{3.6}$$

From our assumption follows $\lim_{x \to x_F^-} u_{xx}(t,x) \neq 0$ and $\lim_{x \to x_F^+} u_{xx}(t,x) \neq 0$ and $u_x \in C^{1+\alpha/2, \alpha}(\bar{\Omega}_T)$. It also follows that u_x is Lipschitz continuous w.r.t. x (Remark 6) and thus that u_x is differentiable a.e. w.r.t. x (Theorem of Rademacher [46]). We cannot apply the classical implicit function theorem [45, p. 658] which would require u_x to be continuously differentiable w.r.t. all variables, but we can apply its generalization to a.e. differentiable functions: Clarke's Theorem [46] (Theorem 14, Appendix C.3). The generalized Jacobian (C.4), for our particular case, has the form

$$\partial f := \mathrm{conv}\{(u_{xt}(t, x_F^-), u_{xx}(t, x_F^-))^t, (u_{xt}(t, x_F^+), u_{xx}(t, x_F^+))^t\}.$$

Then $\pi_x \partial f$ consists of all $\beta \in \mathbb{R}$ such that for some $\gamma \in \mathbb{R}$ the vector $(\gamma, \beta)^t \in \partial f$. Since we assumed $\lim\limits_{x \to x_F^-} u_{xx}(t,x) \neq 0$ and $\lim\limits_{x \to x_F^+} u_{xx}(t,x) \neq 0$ the condition '$\pi_x \partial f$ has maximal rank' is fulfilled and we can apply Theorem 14. Thus there exists a one-sided open neighbourhood $[0, T^*)$ of 0 and a function $x_F : [0, T^*) \to \mathbb{R}$ such that x_F is locally Lipschitz in $[0, T^*)$, $x_F(0) = x_{F,0}$ and $u_x(t, x_F(t)) = \bar{u}$.

In order to avoid working with the implicit equation (3.4) defining the interface we derive an equation for the speed of the front point. To this end we compute the flux, defined by (3.3), at the front point

$$q(t, x_F) = D_0 u_x(t, x_F) = D_0\ \bar{u} \cdot sgn(u_x(t, x_F))$$

and take the derivative with respect to the time. We get

$$\dot{x}_F(t) = -\frac{q_t(t, x_F)}{q_x(t, x_F)}. \tag{3.7}$$

Equation (3.7) is equivalent to eq. (3.5). Integrating the problem P with respect to x in the interval $[0, x]$ we obtain

$$D_0 u_x + D_1 H(|u_x| - \bar{u})(|u_x| - \bar{u})u_x + s \quad = \quad 0,$$

where $s = \int\limits_0^x S(\xi) - u_t(t, \xi) d\xi$. Differentiating with respect to the time t we get

$$(D_0 + D_1 H(|u_x| - \bar{u})(2|u_x| - \bar{u}))u_{xt} + s_t = 0. \tag{3.8}$$

The value of u_{xt} at the front point is

$$\lim_{x \to x_F{}^-} u_{xt}(t, x) = -\frac{s_t(t, x_F)}{D_0}, \qquad \lim_{x \to x_F{}^+} u_{xt}(t, x) = -\frac{s_t(t, x_F)}{D_0 + D_1 \bar{u}}.$$

Taking into account the value of u_{xx} at x_F

$$\begin{aligned} \lim_{x \to x_F{}^-} u_{xx}(t, x) &= -\frac{u_t(t, x_F) - S(x_F)}{D_0}, \\ \lim_{x \to x_F{}^+} u_{xx}(t, x) &= -\frac{u_t(t, x_F) - S(x_F)}{D_0 + D_1 \bar{u}}, \end{aligned}$$

we finally get that

$$\dot{x}_F = \frac{-s_t(t, x_F)}{u_t(t, x_F) - S(x_F)},$$

which implies that $\dot{x}_F$ is continuous.

The time of existence, T^*, of the solution of (3.7) depends on the maximum of $\dot{x}_F$ (Peano's existence theorem [22, pp. 10]: $T^* = \min(t, \frac{1}{\max|\dot{x}_F|})$).

$\diamondsuit$

We make the following assumptions:

Assumption 2. :

- *the initial function, u_0, and the source, S, belong to $C^{2+\alpha}([0, 1])$, i.e. $u_0, S \in C^{2+\alpha}([0, 1])$, and fulfil the compatibility condition of zeroth and first order (see Appendix B).*
- *for the initial function u_0 there exists a non-degenerate front point, i.e. $|u_x(0, x_F(0))| = \bar{u}$,*

- *for the source function, it holds*

$$\left|\frac{dS}{dx}\right| \leq (\varepsilon + P(|u_x|))(1+|u_x|)^4,$$

where $P(\rho) \geq 0$ *is continuous,* $P(\rho) \overset{\rho\to\infty}{\to} 0$*, and* $\varepsilon = \varepsilon(M, \nu, \mu, \mu_1, \max_{\rho\geq 0} P(\rho))$*,* $\varepsilon \geq 0$ *sufficiently small.*

If the second assumption is not fulfilled for $t = 0$ but for $t = t_0$ we simply transform $t' = t - t_0$. We split the problem P into two subproblems P1 and P2 defined as

$$P1 : \begin{cases} \frac{\partial u}{\partial t} = \frac{\partial q}{\partial x} + S(x) = a(|u_x|)u_{xx} + S(x), \quad 0 < x < x_F(t),\ t > 0, \\ u(0,x) = u_0(x), \quad 0 \leq x \leq x_F(t), \\ u_x(t,0) = 0, \quad t \geq 0, \\ u_x(t, x_F(t)) = u_{0,x}(x_F(0)),\ |u_{0,x}(x_F(0))| = \bar{u}, \quad t \geq 0, \end{cases}$$

and

$$P2 : \begin{cases} \frac{\partial u}{\partial t} = \frac{\partial q}{\partial x} + S(x) = a(|u_x|)u_{xx} + S(x), \quad x_F(t) < x < 1,\ t > 0 \\ u(0,x) = u_0(x), \quad x_F(t) \leq x \leq 1, \\ u_x(t, x_F(t)) = u_{0,x}(x_F(0)),\ |u_{0,x}(x_F(0))| = \bar{u}, \quad t \geq 0, \\ u(t,1) = 0, \quad t \geq 0, \end{cases}$$

with flux defined by

$$q(t,x) = D(t,x,u_x)u_x = \begin{cases} D_0 u_x, & |u_x| \leq \bar{u}, \\ (D_0 + D_1(|u_x| - \bar{u}))u_x, & |u_x| \geq \bar{u}, \end{cases} \tag{3.9}$$

or for the non divergence representation

$$a(|u_x|) = \begin{cases} D_0, & |u_x| \leq \bar{u}, \\ D_0 + D_1(2|u_x| - \bar{u}), & |u_x| \geq \bar{u}. \end{cases}$$

Theorem 6. *Let* $x_F(t) \in C^1([0,T])$ *be given and* $\dot{x}_F(t) \neq \infty$*. Then P1 and P2 possess unique classical solutions* $u^-(t,x)$ *and* $u^+(t,x)$*, respectively. Moreover the function*

$$u(t,x) := \begin{cases} u^-(t,x), & x \in [0, x_F],\ t > 0 \\ u^+(t,x), & x \in [x_F, 1],\ t > 0 \end{cases} \tag{3.10}$$

is a solution of problem P.

Proof

First we show that P1 and P2 possess classical solutions.

Problem P1 is linear and according to Assumptions 2 the initial function and the source fulfil the conditions of Theorem 9, Appendix B. Now taking into account Remark 11, Appendix B, we conclude that P1 has a unique solution from the class $C^{1+\alpha/2,2+\alpha}([0,T]\times[0,x_F(t)])$. For problem P2, we make use of Theorem 1 and Remark 1. Problem P2 has mixed boundary conditions, but since Ω_T is a cylinder we do apply Theorem 1, provided that the regularity conditions RC1 are fulfilled. For P2, it holds that

$$a_{11}(t,x,u,p)=(D_0+D_1(2p-\bar{u})),\quad b(x)=S(x).$$

We verify the RC 1:

a) The condition $a_{11}\geq 0$ is equivalent to $p\geq\frac{\bar{u}}{2}-\frac{D_0}{2D_1}$ and it is fulfilled since $p\geq\bar{u}\geq 0$. Moreover it holds that $a_{11}=a_{11}(t,x,u,p)\in C^{\alpha/2,\ \alpha,\alpha,\alpha}(\bar{\Omega}_T\times\mathbb{R}^+\times\mathbb{R}^+)$

b) It is fulfilled that $b=S(x)\geq 0$, $S\in C^{1+\alpha/2,2+\alpha}(\bar{\Omega}_T)$ and for the function $\Phi(\tau)$ we can take $\Phi\equiv 0$.

c) We take $m=3$, then

$$\frac{D_0-\bar{u}D_1}{2(1+M_1)}(1+p)\leq a_{11}(t,x,u,p)\leq 2D_1|D_0-D_1\bar{u}|(1+p)$$

$$|2D_1|(1+p)^3+|b|\leq(2D_1+|b|)(1+p)^3,\ \text{i.e. } \mu_1=(2D_1+|b|)$$

The last inequality in c) follows from Assumption 2.

d) It follows from Assumption 2.

Let u^- be a solution of P1 and u^+ be a solution of P2, then arises the question whether $u^-(t,x_F(t))=u^+(t,x_F(t))$.

We consider the following equation

$$u^\varepsilon_t(t,x)\quad=\quad\tilde{a}(|u_x|,\varepsilon)u^\varepsilon_{xx}(t,x)+S(x),\tag{3.11}$$

where

$$\tilde{a}(v,\varepsilon)=D_0+D_1(\frac{1}{2}+\frac{1}{\pi}\arctan\frac{v-\bar{u}}{\varepsilon})(2v-\bar{u}).$$

For ε going to 0, $\tilde{a}(v,\varepsilon)\to a(v)$.

We can represent $\tilde{a}$ as $\tilde{a}(v,\varepsilon)=a(v)+f(v,\varepsilon)$, where

$$f(v,\varepsilon)=\begin{cases}D_1(2v-\bar{u})(\frac{1}{2}+\frac{1}{\pi}\arctan\frac{v-\bar{u}}{\varepsilon}), & v<\bar{u},\\ D_1(2v-\bar{u})(-\frac{1}{2}+\frac{1}{\pi}\arctan\frac{v-\bar{u}}{\varepsilon}), & v>\bar{u},\end{cases}$$

and

$$f'_v(v,\varepsilon) = \begin{cases} D_1(1+\frac{2}{\pi}\arctan\frac{v-\bar{u}}{\varepsilon}) + \frac{D_1(2v-\bar{u})}{\pi}\frac{\varepsilon^2}{\varepsilon^2+(v-\bar{u})^2} & v<\bar{u} \\ \\ D_1(-1+\frac{2}{\pi}\arctan\frac{v-\bar{u}}{\varepsilon}) + \frac{D_1(2v-\bar{u})}{\pi}\frac{\varepsilon^2}{\varepsilon^2+(v-\bar{u})^2} & v>\bar{u}. \end{cases}$$

Both f and f'_v go to zero for $\varepsilon \to 0$ and $v \neq \bar{u}$. Note that the function f is uniformly continuous in ε since it is defined and continuous for any ε including large ε and $f \overset{\varepsilon\to\pm\infty}{\longrightarrow} \frac{D_1(2v-\bar{u})}{2}$.

We solve

$$\begin{aligned} u^\varepsilon_t &= \tilde{a}(|u_x|,\varepsilon)u^\varepsilon_{xx} + S(x), \quad 0<x<x_F,\ t>0 \qquad (3.12)\\ u^\varepsilon(0,x) &= u_0(x), \qquad 0\le x\le x_F,\\ u^\varepsilon_x(t,0) &= 0, \qquad t\ge 0,\\ |u^\varepsilon_x(t,x_F)| &= \bar{u}, \qquad t\ge 0, \end{aligned}$$

and

$$\begin{aligned} u^\varepsilon_t &= \tilde{a}(|u_x|,\varepsilon)u^\varepsilon_{xx} + S(x), \quad x_F<x<1,\ t>0 \qquad (3.13)\\ u^\varepsilon(0,x) &= u_0(x), \qquad x_F\le x\le 1,\\ u^\varepsilon(t,1) &= 0, \qquad t\ge 0,\\ |u^\varepsilon_x(t,x_F)| &= \bar{u}, \qquad t\ge 0. \end{aligned}$$

Each of these problems can be transformed such that $x \in [0,1]$ (we have done this in more details, later on in the proof, for equations (3.14) and (3.16)). In this way, the function x_F enters in the main equation. The coefficient $\tilde{a}(v,\varepsilon)$ is Holder continuous in v with a constant α, and according to [25, Ch.IV, Th.5.3] (see appendix B, Th.9), problem (3.12) has a unique solution in the class $C^{1+\alpha/2,2+\alpha}([0,T]\times[0,x_F])$. Problem (3.13) has mixed boundary conditions and Theorem 9 is not directly applicable. However, Theorem 10 combined with Theorem 11, given in appendix B (corresponding to Th.5.1 and Th.12.1 from [25]), assures that the mixed boundary problem (3.13) has a unique solution in $C^{1+\alpha/2,2+\alpha}([0,T]\times[x_F,1])$, for $0<\alpha<1$, provided that $\tilde{a}, S \in C^{\alpha/2,\alpha}(\Omega_T)$.

Now, let us consider the difference between the solutions of (3.12) and P1, $w^- := u^\varepsilon - u^-$, and the corresponding differential equation fulfilled by it,

$$\begin{aligned} w^-_t &= a(|u_x|)w^-_{xx} + f(|u_x|,\varepsilon)u^\varepsilon_{xx}, \quad 0<x<x_F(t), \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (3.14)\\ w^-(0,x) &= 0, \quad w^-_x(t,0) = |w^-_x(t,x_F)| = 0. \end{aligned}$$

We map the interval $[0,x_F]$ to $[0,1]$ through $x \mapsto \xi = \frac{x}{x_F}$. In terms of this new variable the problem reads

$$\begin{aligned} w^-_t &= \frac{1}{x_F(t)^2}(D_0 w^-_{\xi\xi} + f(\frac{|u_\xi|}{x_F},\varepsilon)u^\varepsilon_{\xi\xi}), \quad 0<\xi<1,\ t>0 \qquad (3.15)\\ w^-(0,\xi) &= 0, \quad w^-_\xi(t,0) = |w^-_\xi(t,1)| = 0. \end{aligned}$$

Again according to Ladyzhenskaia [25], problem (3.15) possesses a unique solution if the coefficients making up the problem belong to the class $C^{\alpha/2,\alpha}$. For the coefficient in front of $w^-_{\xi\xi}$ this is true because of the continuity of $x_F(t)$. In order to prove that $\frac{1}{x_F(t)^2}f(\frac{|u_\xi|}{x_F},\varepsilon)u^\varepsilon_{\xi\xi}$ belongs to the class $C^{\alpha/2,\alpha}$, we need that $u^\varepsilon_{\xi\xi\xi}$ and $u^\varepsilon_{\xi\xi t}$ exist and are continuous. To argue for this we use the fact that the solution of (3.12) belongs to the class $C^{(3+\alpha)/2,3+\alpha}$ since the coefficients making up the equation possess a greater smoothness.

Because problem (3.15) has a unique solution and $f(v,\varepsilon) \overset{\varepsilon\to 0}{\rightarrow} 0$, uniformly, follows that the solution of (3.15) goes to the zero solution for $\varepsilon \to 0$, i.e. $u^\varepsilon(t,x) \overset{\varepsilon\to 0}{\rightarrow} u^-(t,x)$.

Similarly, we proceed with the interval $[x_F,1]$. We define a function $w^+(t,x)$ in $[0,T]\times[x_F,1]$, such that $w^+ = u^\varepsilon(t,x) - u^+(t,x)$ and it fulfils the problem

$$\begin{aligned} w^+_t &= a(|u_x|)w^+_{xx} + f(|u_x|,\varepsilon)u^\varepsilon_{xx}, \quad x_F < x < 1, \\ w^+(0,x) = 0,& \qquad w^+_x(t,x_F) = w^+(t,1) = 0. \end{aligned} \tag{3.16}$$

We transform this problem into $[0,1]$, through $\xi = \frac{x-x_F}{1-x_F}$, and obtain a linear parabolic problem

$$\begin{aligned} w^+_t &= \frac{1}{(1-x_F)^2}\left(a(|\frac{u_\xi}{(1-x_F)}|)w^+_{\xi\xi} + f(|\frac{u_\xi}{1-x_F}|,\varepsilon)u^\varepsilon_{\xi\xi}\right), \quad 0<\xi<1 \\ w^+(0,\xi) &= 0, \quad w^+_\xi(t,0) = w^+(t,1) = 0. \end{aligned} \tag{3.17}$$

We use similar arguments as in the previous case. According to Ladyzhenskaia problem (3.17) possesses a unique solution if $\frac{1}{(1-x_F)^2}a(|\frac{u_\xi}{(1-x_F)}|)$ and $\frac{1}{(1-x_F)^2}f(|\frac{u_\xi}{1-x_F}|,\varepsilon)u^\varepsilon_{\xi\xi}$ belong to $C^{\alpha/2,\alpha}$. For the latter we use the same arguments as in the previous case. The Hölder continuity of the term in front of $w^+_{\xi\xi}$ follows from the boundedness of $u_{\xi\xi}$, $u_{\xi t}$, and a_v. In this way we get that problem (3.16) possesses unique solution and $w^+(t,x) \overset{\varepsilon\to 0}{\rightarrow} 0$.

Now, let $g : [0,T] \to \mathbb{R}$ and

$$g(t) = u^\varepsilon(t,x_F(t)^-) - u^\varepsilon(t,x_F(t)^+) \tag{3.18}$$

By taking the derivative of (3.18) with respect to t we obtain

$$\frac{dg(t)}{dt} = u^\varepsilon_t(t,x_F(t)^-) + \bar{u}\dot{x}^-_F - u^\varepsilon_t(t,x_F(t)^+) - \bar{u}\dot{x}^+_F.$$

Because of $\dot{x}^-_F = \dot{x}^+_F$ and the continuity of u^ε_t on the interface we get

$$\frac{dg(t)}{dt} = 0.$$

In addition, $g(0) = 0$ leads to $g(t) \equiv 0$, i.e, $u^\varepsilon(t, x_F(t)^-) = u^\varepsilon(t, x_F(t)^+)$. The same is true for $u(t, x_F(t)^-) = u(t, x_F(t)^+)$, for $t \in [0, T]$.

We now show that the function u defined by (3.10) is a solution of problem P according to Definition 11. The functions u and u_x are a continuous functions since $u^- \in C^{1+\alpha/2,2+\alpha}([0,T] \times [0, x_F))$ and $u^+ \in C^{1+\alpha/2,2+\alpha}([0,T] \times (x_F, 1])$ and for every fixed $t \in [0,T]$ it holds that $u^-(t, x_F(t)) = u^+(t, x_F(t)) = u(t, x_F(t))$ and $u_x^-(t, x_F(t)) = u_x^+(t, x_F(t)) = u_x(t, x_F(t)) = u_{0,x}(x_F(0))$. Furthermore, u_{xx} is continuous everywhere except at the front point $x_F(t)$. Now all we need in addition is to show that u satisfies problem P. Indeed, that is the case, because for every fixed $t \in [0,T]$, $u(t,x) \equiv u^-(t,x)$ for $x \in [0, x_F(t)]$ and $u^-(t,x)$ is a solution of P1, respectively P in that interval. Similarly, in $[x_F(t), 1]$ for every fixed $t \in [0,T]$ it holds that $u(t,x) \equiv u^+(t,x)$ and $u^+(t,x)$ is a solution of P2, respectively P in the corresponding interval.
$\diamondsuit$

We finish this section with some comments on sub- and super- solutions for problem P based on the comparison theorem of Nagumo type (see Appendix B).

Lemma 6. *Let us consider problem P with source $S(x)$ defined by (3.2) and with initial function $0 \leq u_0 \leq \bar{u}x$. Then, functions $u_{sub} \equiv 0$ and $u_{sup} = \bar{u}x + S_0 t$ are sub- and super- solutions for problem P.*

Proof:
Let us rewrite the problem P as

$$\begin{aligned} u_t &= Lu + S(x), \quad (t,x) \in \Omega_T, \\ u(0,x) &= u_0(x), \quad x \in [0,1], \\ u_x(t,0) = 0, & \qquad u(t,1) = 0, \quad t \geq 0, \end{aligned}$$

where

$$L := \frac{\partial q(t,x)}{\partial x} = \frac{\partial}{\partial x}\left((D_0 + D_1 H(|u_x| - \bar{u})(|u_x| - \bar{u}))u_x\right).$$

Now, let us define by $Z = \{y \in C(\bar{\Omega}_T) : \ y_t, y_x, q(t,x), \mathcal{L}y \in C\}$, and let us notice that $u_{sub}, \ u_{sup} \in Z$. In addition, it holds for u_{sub} that:

- $\partial_t u_{sub} - Lu_{sub} - S(x) \leq 0$,
- $u_{sub}(0,x) \leq u_0$ and $\partial_x u_{sub}(t,0) = 0, \ u_{sub}(t,1) = 0$

and for u_{sup} that

- $\partial_t u_{sup} - Lu_{sup} - S(x) = S_0 - S(x) \geq 0$,
- $u_{sup}(0,x) = \bar{u}x \geq u_0$ and $\partial_x u_{sup}(t,0) = S_0 t \geq 0, \ u_{sup}(t,1) = S_0 t + \bar{u} \geq 0$.

Therefore, according to Definition 18, AppendixB u_{sub} and u_{sup} are sub- and super-solutions for problem P, respectively.

$\diamondsuit$

As a consequence of the comparison theorem of Nagumo type follows that if there is a solution u of problem P with initial date $0 \leq u(0,x) \leq x\bar{u}$ and source $S(x)$ defined by (3.2) then u is the only solution of P such that

$$0 \leq u(t,x) \leq x\bar{u} + tS_0 \text{ for } (t,x) \in \bar{\Omega}_T.$$

3.4 Exact solutions for some special cases

In this section, we consider the equation

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left((1 + D_1 H(u_x - \bar{u})(u_x - \bar{u}))\frac{\partial u}{\partial x}\right), \quad 0 < x < 1, \ t > 0. \tag{3.19}$$

This is the main equation of the problem P, where :

- $S \equiv 0$,
- $D_0 = 1$,
- and initial condition u_0 such that $\frac{\partial u_0}{\partial x}(x_F(0)) = \bar{u}$.

Using the front tracking idea (the idea of separating the problem into two subproblems) allows us to derive analytic solutions of (3.19). Equation (3.19) decomposes into two sub-equations $(\tilde{P}1)$ and $(\tilde{P}2)$:

$$u_t = \begin{cases} u_{xx}, & u_x \leq \bar{u} \qquad (\tilde{P}1) \\ \\ (1 + 2D_1 u_x - D_1\bar{u})u_{xx}, & u_x \geq \bar{u}. \qquad (\tilde{P}2) \end{cases}$$

We have found two families of exact solutions for given D_1 and $\bar{u}$. We obtained the first one matching, on the interface, the solution of $(\tilde{P}1)$ to a polynomial solution in x and t of $(\tilde{P}2)$. The second family of solutions we obtained in a similar way. We connected, on the interface, the self-similar solution of $(\tilde{P}1)$ to a polynomial solution in x of $(\tilde{P}2)$. Now, we describe in details the derivation of the two families of solutions.

Lemma 7. *The function*

$$u(t,x) = \begin{cases} \frac{D_1\bar{u}-1}{2D_1}(x + \frac{A-\bar{u}}{C}) + \frac{D_1\bar{u}+1}{4CD_1^2}(e^{2D_1(Cx+2D_1C^2t+A-\bar{u})} - 1) + \frac{\bar{u}^2-A^2}{2C}, & u_x \leq \bar{u} \\ \\ \frac{C}{2}x^2 + 2C^3D_1t^2 + 2C^2D_1xt + Ax + C(1 - D_1\bar{u} + 2AD_1)t, & u_x \geq \bar{u} \end{cases} \tag{3.20}$$

defines a family of solutions of equation (3.19), for A and C free parameters and $A + 2C^2D_1t \leq \bar{u} \leq C + A + 2C^2D_1t$.

Proof
By calculation.

◇

We have constructed the function u defined by (3.20) in the following way. We start with looking for a solution of $(\tilde{P}2)$ of the type

$$u(t,x) = Ax + Bt + C\frac{x^2}{2} + Dxt + E\frac{t^2}{2}$$

Inserting this expression into $(\tilde{P}2)$, we obtain

$$\begin{aligned} B &= (1-\bar{u}D_1)C + 2ACD_1 \\ D &= 2C^2D_1 \\ E &= 2DCD_1. \end{aligned} \tag{3.21}$$

Then the solution reads

$$u(t,x) = Ax + Bt + \frac{C}{2}(x + \frac{D}{C}t)^2 + \frac{t^2}{2}(-\frac{D^2}{C} + E)$$

with derivative $u_x = A + Cx + Dt$. We now look for a solution of $(\tilde{P}1)$ on the other side of the interface. At the interface x_F, $A + Cx_F + Dt = \bar{u}$, the solution of $(\tilde{P}1)$ should fulfil

$$u(t,x_F) = Ax_F + (1-\bar{u}D_1)Ct + 2ACD_1t + \frac{1}{2C}(\bar{u}-A)^2.$$

We make the following transformation

$$\begin{aligned} x \mapsto \xi &= x + \frac{D}{C}t + \frac{A}{C} - \frac{\bar{u}}{C} \\ t \mapsto t' &= t \\ u \mapsto w &= u + \frac{A^2}{2C} - \frac{\bar{u}^2}{2C}. \end{aligned}$$

In the new coordinates problem $\tilde{P}1$ reads

$$\begin{aligned} w_{t'} &= w_{\xi\xi} - \frac{D}{C}w_\xi \\ w_\xi(t',0) &= \bar{u} \\ w(t',0) &= \alpha t'. \end{aligned} \tag{3.22}$$

where $\alpha = -\frac{AD}{C} + B$. The solution of (3.22) is given by (we omit the prime)

$$w = \bar{u}(\xi - \frac{D}{C}t) + (\alpha + \frac{\bar{u}D}{C})t + g(\xi),$$

where the function $g(\xi)$ satisfies

$$\begin{aligned} g(0) &= 0 \\ g'(0) &= 0 \\ g''(\xi) &= (\alpha + \frac{\bar{u}D}{C}) + \frac{D}{C}g'(\xi). \end{aligned} \tag{3.23}$$

The solution of the latter problem is

$$g(\xi) = (e^{\frac{D}{C}\xi} - 1 - \frac{D}{C}\xi)(B + \frac{(\bar{u}-A)D}{C})\frac{C^2}{D^2}. \tag{3.24}$$

The bounds for the threshold value $\bar{u}$ come from the requirement $0 \leq x_F \leq 1$, i.e. $A + 2C^2 D_1 t \leq \bar{u} \leq C + A + 2C^2 D_1 t$.

Another special solution we find using similar technique.

Lemma 8. *The second family of solutions of equation (3.19) is given by*

$$u(t,x) = \begin{cases} (K-t)^\alpha f(\frac{x^2}{K-t}), & u_x \leq \bar{u} \\ \frac{x^3}{36D_1(K-t)} + \frac{x(D_1\bar{u}-1)}{2D_1}, & u_x \geq \bar{u}, \end{cases} \tag{3.25}$$

where $K = const > 0$ and $\alpha = const > 0$ are free parameters, and $\bar{u} \leq \frac{1-6(K-t)}{6D_1(K-t)}$. The function $f(.)$ is a solution of the Confluent Hypergeometric equation and $f(.)$ is defined by

$$f\,(\xi) = b_1\,{}_1F_1(-\alpha, 1/2, \xi) + b_2 U(-\alpha, 1/2, \xi), \qquad \xi = \frac{x^2}{K-t},\ \xi_F = \frac{x_F^2}{K-t}$$

where

$${}_1F_1(a,c,\xi) = \frac{\Gamma(c)}{\Gamma(c-a)\Gamma(a)} \int\limits_0^1 e^{\xi t} t^{a-1}(1-t)^{c-a-1}dt \tag{3.26}$$

and

$$U(a,c,\xi) = \frac{1}{\Gamma(a)} \int\limits_0^\infty e^{-\xi t} t^{a-1}(1+t)^{c-a-1}dt, \tag{3.27}$$

the coefficients b_1 and b_2 are defined by the conditions

$$\begin{aligned} f(\xi_F) &= (K-t)^{1/2-\alpha}\frac{(2D_1\bar{u}-1)\sqrt{6(D_1\bar{u}+1)}}{3D_1} \\ f'(\xi_F) &= \frac{\bar{u}}{2(K-t)^{\alpha-1/2}\sqrt{6(D_1\bar{u}+1)}}. \end{aligned}$$

Proof
By computation.

$\diamondsuit$

The way one can get this solution is the following one. We start looking for a solution of $(\tilde{P}2)$ of the type

$$u(t,x) \quad = \quad f_0(t) + x f_1(t) + \frac{x^2}{2} f_2(t) + \frac{x^3}{3} f_3(t). \tag{3.28}$$

One possible choice for the functions $f_0,\ f_1, f_2$ and f_3 that satisfy $(\tilde{P}2)$ is

$$f_3(t) = \frac{1}{12D_1(K-t)}, \quad f_2(t) = 0, \quad f_1(t) = \frac{D_1\bar{u} - 1}{2D_1}, \quad f_0(t) = 0,$$

for K positive constant. We rewrite equation (3.28) as

$$u(t,x) = \frac{x}{3}\left(\frac{x^2}{12D_1(K-t)}\right) + \frac{x(D_1\bar{u} - 1)}{2D_1}.$$

At the front point x_F the derivative of u takes the value $\bar{u}$, that leads to

$$\frac{x_F^2}{12D_1(K-t)} = \frac{D_1\bar{u} + 1}{2D_1}. \tag{3.29}$$

The solution at the front point takes the value

$$u(t,x_F) = \frac{x_F}{3D_1}(2D_1\bar{u} - 1).$$

We want to match, at the front point, this solution to the solution of the heat equation $(\tilde{P}1)$, i.e.

$$\begin{aligned} u_t &= u_{xx} \\ u_x(t,x_F) &= \bar{u}, \\ u(t,x_F) &= \frac{x_F}{3D_1}(2D_1\bar{u} - 1). \end{aligned} \tag{3.30}$$

The self similar form of the solution of the heat equation (3.30) takes the form

$$u = (K-t)^\alpha f\left(\frac{x^2}{K-t}\right). \tag{3.31}$$

Putting $\xi := \frac{x^2}{K-t}, \quad \xi_F := \frac{x_F^2}{K-t}$ and substituting (3.31) in (3.30) gives us a second order differential equation for $f(\xi)$

$$4\xi f'' + (2-\xi)f'(\xi) + \alpha f = 0 \tag{3.32}$$

with initial conditions

$$\begin{aligned} f(\xi_F) &= (K-t)^{1/2-\alpha}\frac{(2D_1\bar{u}-1)\sqrt{6(D_1\bar{u}+1)}}{3D_1} \qquad (3.33) \\ f'(\xi_F) &= \frac{\bar{u}}{2(K-t)^{\alpha-1/2}\sqrt{6(D_1\bar{u}+1)}}. \end{aligned}$$

We change the independent variable $\xi = 4\eta$. In this way (3.32) becomes

$$\eta f'' + (\frac{1}{2} - \eta)f' + \alpha f = 0. \qquad (3.34)$$

This is the Confluent Hypergeometric equation

$$xy'' + (c-x)y' - ay = 0$$

for $c = 1/2$ and $a = -\alpha$. The general solution of (3.34) is given by

$$f\,(\eta) = b_1\ {}_1F_1(-\alpha, 1/2, \eta) + b_2 U(-\alpha, 1/2, \eta)$$

where ${}_1F_1(a, c, \eta)$ and $U(a, c, \eta)$ are defined by (3.26) and (3.27), respectively.

The first family of solutions is used in the numerical computations for verifying the accuracy of the numerical method.

3.5 Space and time discretization of the anomalous transport problem

In this section we discuss the discretization of the two subproblems P1 and P2. At first, the problem Pi, $i = 1, 2$ is discretized in space through the standard Galerkin method [47]. Then in sections 3.5.2 - 3.5.4 we discuss the time discretization, advancing of the front point in time and up-dating of the spatial grid.

3.5.1 Space discretization

In order to treat both subproblems, P1 and P2 together, we introduce the Galerkin method in a general interval $D \equiv [a, b]$, where it should be understood as $[0, x_F]$ for P1 and as $[x_F, 1]$ for P2. Additional comments are given when there are differences depending on the considered subproblem. The function space W is chosen to be the Sobolev space $W \equiv W_2^2(D)$ for P1, and $W = \{v \in W_2^2(D) : \quad v(1) = 0\}$ for P2.

We are looking for an approximate solution of

$$\begin{aligned} \partial_t u(t,x) &= \partial_x q(t,x) + S(x), \quad a < x < b \qquad (3.35) \\ u(0,x) &= u_0(x), \quad a \le x \le b, \\ &+ \ \text{boundary conditions} \end{aligned}$$

in a finite dimensional subspace W_h of W. Let $\{x_i\}_{i=0}^n$ be a division of the interval $[a,b]$, and let $\{\phi_i(x)\}_{i=0}^n$ be the Lagrangian finite element basis in W_h (i.e. $\phi_i(x_j) = \delta_{ij}$). Now, let us multiply (3.35) by a test function $v \in W_h$ and integrate in $[a,b]$. Using integration by parts we get the weak Galerkin form of the problem:

$$\begin{aligned}\int_a^b \partial_t u(t,x)v(x)dx &= -\int_a^b q(t,x)v'(x)dx + q(t,b)v(b) - q(t,a)v(a) \\ &+ \int_a^b S(x)v(x)dx. \end{aligned} \tag{3.36}$$

Each function $p(x)$ from W_h could be presented as $p(x) = \sum\limits_{i=1}^{n} p_i\phi_i(x)$. Now, by identifying v in (3.36) with the basis functions $\phi_j(x)$ and using the above indicated representation for $\partial_t u(t,x)$, u_x, $S(x)$ we get a system of n ODEs. The j^{th} equation has the form

$$\begin{aligned}\sum_i \partial_t u_i(t) \int_a^b \phi_i(x)\phi_j(x)dx &= -\sum_i u_i(t) \int_a^b D(u,u_x)\phi_i'(x)\phi_j'(x)dx \\ &+q(t,b)\underbrace{\phi_n(b)}_{=1} - q(t,a)\underbrace{\phi_0(a)}_{=1} \\ &+\sum_i S_i \int_a^b \phi_i(x)\phi_j(x)dx.\end{aligned}$$

The whole system could be written in a matrix form

$$\begin{aligned} M_h\partial_t U_h &= -K_h(U_h)U_h + M_h\tilde{S}_h, \quad 0 < t \leq T, \qquad (3.37) \\ U_h(0) &= U_0. \end{aligned}$$

The vector $U(t) = (U(t,x_0),\cdots,U(t,x_n))^T$ is the semi-discrete solution of our problem, $\tilde{S}_h = (S(x_0),\cdots,S(x_n))^T$ is the source computed at the mesh. The matrix M_h is a constant block diagonal mass matrix. The matrix $K_h(U_h)$ is a block diagonal stiffness-like matrix. In the case of P1, it is the usual constant stiffness matrix whereas, in the case of P2, it is a solution dependent matrix. Throughout this chapter we omit the index h.

We use the Galerkin method with linear and Hermitian basis functions. Here, we list separately both of the cases, giving the specific details concerning each of them.

Linear finite elements

Let us first consider the case when the finite dimensional space W_h is the space of continuous functions in $[0,1]$ which reduce to linear functions in each of the subintervals $e_i := [x_i, x_{i+1}],\ i = 0 \cdots n$. Every function v of W_h can be presented as a linear combination of linear polynomials

$$v(x) = \sum_{i=0}^{n} (\ v_i \phi(\xi_i) + v_{i+1}\phi(-1 + \xi_i)\ ),$$

where ξ_i is the following normalized coordinate on the element i

$$\xi_i = \begin{cases} \frac{x-x_i}{x_i - x_{i-1}}, & x_{i-1} \le x \le x_i, \\ \frac{x-x_i}{x_{i+1}-x_i}, & x_i \le x \le x_{i+1}, \end{cases}$$

and $\phi(\xi)$ is the normalized basis function

$$\phi(\xi) = \begin{cases} \xi, & -1 \le \xi \le 0, \\ 1-\xi, & 0 \le \xi \le 1. \end{cases}$$

The mass matrix M consists of elements $\{m_{ij}\}$ given by

$$\begin{aligned} m_{i,i-1} &= \int_{e_i} \phi_{i-1}(x)\phi_i(x)dx, \\ m_{i,i} &= \int_{e_i} \phi_i^2(x)dx + \int_{e_{i+1}} \phi_i^2(x)dx, \\ m_{i,i+1} &= \int_{e_{i+1}} \phi_{i+1}(x)\phi_i(x)dx. \end{aligned}$$

In the case of P1, the stiffness matrix $K_h(U_h)$ is the usual constant stiffness matrix with elements

$$\begin{aligned} k_{i,i-1} &= \int_{e_i} \phi'_{i-1}(x)\phi'_i(x)dx, \\ k_{i,i} &= \int_{e_i} (\phi'_i(x))^2dx + \int_{e_{i+1}} (\phi'_i(x))^2dx, \\ k_{i,i+1} &= \int_{e_{i+1}} \phi'_{i+1}(x)\phi'_i(x)dx. \end{aligned}$$

In the case of P2, it is a solution dependent matrix

$$\begin{aligned} k_{i,i-1} &= \int_{e_i} D(U, U_x)\phi'_{i-1}(x)\phi'_i(x)dx \\ k_{i,i} &= \int_{e_i} D(U, U_x)(\phi'_i(x))^2dx + \int_{e_{i+1}} D(U, U_x)(\phi'_i(x))^2dx \\ k_{i,i+1} &= \int_{e_{i+1}} D(U, U_x)\phi'_{i+1}(x)\phi'_i(x)dx. \end{aligned}$$

In order to reduce the computational work, two modifications are performed.

i.) Lumping of the mass matrix M:

In (3.37), the mass matrix is substituted by the so called lump mass matrix $\tilde{M}$ to avoid the inversion of M. The matrix $\tilde{M}$ is a diagonal matrix with elements which are the sums of the entries of M at each row, i.e.

$$\tilde{m}_{i,i} = m_{i,i-1} + m_{i,i} + m_{i,i+1}, \qquad \forall i = 0, \cdots, n.$$

ii.) Reducing the number of computations of the stiffness matrix:

The diffusion coefficient could be presented as $D(U, U_x) = D_0 + D_1(U_x - \bar{u}) = \sum_l (D_0 + D_1 U_x - D_1 \bar{u})|_{x=x_l} \phi_l$. Now, at each row i in K, the elements that are different from zero are calculated as a product of a row-vector D_i and a 3×3 matrix K_i.

The vector D_i has three elements $\{(D_0 + D_1(U_x - \bar{u}))|_{x=x_j} \phi_j\}_j, \quad j = i-1, i, i+1$. The matrix K_i has the form

$$K_i = \begin{pmatrix} \int_{e_i} \phi_{i-1}\phi'_{i-1}\phi'_i dx & \int_{e_i} \phi_{i-1}\phi'_i\phi'_i dx & 0 \\ \int_{e_i} \phi_i\phi'_{i-1}\phi'_i dx & \int_{e_i} \phi_i\phi'_i\phi'_i dx + \int_{e_{i+1}} \phi_i\phi'_i\phi'_i dx & \int_{e_{i+1}} \phi_i\phi'_{i+1}\phi'_i dx \\ 0 & \int_{e_{i+1}} \phi_{i+1}\phi'_i\phi'_i dx & \int_{e_{i+1}} \phi_i\phi'_{i+1}\phi'_i dx \end{pmatrix}$$

The procedure described above can be sketched as

$$(D_{i-1}\ , D_i\ , D_{i+1}) \overbrace{\begin{pmatrix} *, *, 0 \\ , *, * \\ 0, *, * \end{pmatrix}}^{K_i} = (k_{i-1,i}\ , k_{i,i}\ , k_{i,i+1})$$

Using this strategy we compute the matrices K_i once, and only the vectors D_i are calculated at each time step.

Remark 8. *The application of Lobatto estimate (Theorem 3, Theorem 4) to the case of linear FEM:*

We saw that if we discretize problem P through the FEM and lump the mass matrix we obtain a system of ODEs with the following right hand side

$$F(\boldsymbol{U}) := -\tilde{M}^{-1} K(\boldsymbol{U})\boldsymbol{U} + \tilde{S} = -\frac{1}{h^2} K(\boldsymbol{U}) + \tilde{S},$$

where

$$K(\boldsymbol{U}) = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & -1 & 2 + \frac{U_3 - U_2}{h} - \bar{u} & -1 - \frac{U_3 - U_2}{h} + \bar{u} & 0 & 0 \\ 0 & 0 & -1 - \frac{U_3 - U_2}{h} + \bar{u} & 2 + \frac{U_3 - U_2}{h} + \frac{U_4 - U_3}{h} - 2\bar{u} & -1 - \frac{U_4 - U_3}{h} + \bar{u} & 0 \end{pmatrix}$$

Then, for the Jacobian matrix holds

$$\mathcal{D}F(\boldsymbol{U}) = -\frac{1}{h^2}(K(\boldsymbol{U}) + \mathcal{D}K(\boldsymbol{U}) \cdot \boldsymbol{U}),$$

where

$$\mathcal{D}K(\boldsymbol{U}) \cdot \boldsymbol{U} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{U_3-U_2}{h} & -\frac{U_3-U_2}{h} & 0 & 0 \\ 0 & 0 & -\frac{U_3-U_2}{h} & \frac{U_3-U_2}{h} + \frac{U_4-U_3}{h} & -\frac{U_4-U_3}{h} & 0 \end{pmatrix}.$$

The matrix $\mathcal{D}F$ is symmetric and negative definite. Therefore, the logarithmic matrix norm is negative, and we have the case $C \leq 0$. This implies that in the numerical calculation the estimate from Theorem 3,a) is justified.

Hermitian finite elements

Let W_h be the finite dimensional space of continuous functions on $[0, 1]$ which reduce to cubic functions in each of the subintervals $[x_i, x_{i+1}]$. Usage of Hermitian basis looks quite natural if we keep in mind that they assure continuous derivative everywhere. A function $v \in W_h$ can be represented as

$$\begin{aligned} v(x) &= \sum_{i=0}^{n}(v_i\phi^0(\xi_i) + v_{i+1}\phi^0(-1+\xi_k)) \\ &\quad + \sum_{i=0}^{n}(v'_i\phi^1(\xi_i) + v'_{i+1}\phi^1(-1+\xi_i)). \end{aligned}$$

Here, ξ_i is the normalized coordinate on the element i

$$\xi_i = \begin{cases} \frac{x-x_i}{x_{i+1}-x_i}, & x_i \leq x \leq x_{i+1}, \\ \frac{x-x_i}{x_i-x_{i-1}}, & x_{i-1} \leq x \leq x_i, \end{cases}$$

and ϕ^0 and ϕ^1 are the Hermitian basis

$$\phi^0(\xi_i) = \psi_0(\xi_i), \quad -1 \leq \xi_i \leq 1$$

$$\phi^1(\xi_i) = \begin{cases} (x_{i+1} - x_i)\psi_1(\xi_i), & 0 \leq \xi_i \leq 1, \\ (x_i - x_{i-1})\psi_1(\xi_i), & -1 \leq \xi_i \leq 0, \end{cases}$$

where ψ_0 and ψ_1 are the Hermitian interpolating polynomials defined on the master interval $[-1, 1]$ as

$$\begin{aligned} \psi_0(\xi) &= (|\xi| - 1)^2(2|\xi| + 1) \\ \psi_1(\xi) &= (|\xi| - 1)^2\xi. \end{aligned} \tag{3.38}$$

More information about Hermitian elements can be found in [44, pp. 56].

3.5.2 Time discretization

Now, we discuss the time discretization of the anomalous transport problem. We multiply (3.37) by the inverse lumped mass matrix. This leads to the following system of ODEs

$$\dot{U} = -\tilde{M}^{-1}K(U)U + \tilde{S} =: f(U). \tag{3.39}$$

Full discretization of P1/P2 is realized by the trapezoidal rule. It reads for the particular equation (3.39)

$$U^{j+1} = U^j - \frac{\tau_j}{2}\tilde{M}^{-1}(K(U^{j+1})U^{j+1} + K(U^j)U^j) + \tau_j\tilde{S},$$

where $U^j = U(t_j)$ and $\tau_j = t_{j+1} - t_j$. The obtained nonlinear system is solved iteratively using the Newton method. The value of the increment U^{j+1} can be computed as a solution of $G(X) = 0$, where the function G is defined by

$$G(X) := X - U^j + \frac{\tau_j}{2}\tilde{M}^{-1}(K(X)X + K(U^j)U^j) - \tau_j\tilde{S}.$$

The iterative Newton method states that $U^{j+1} = \lim_k X^{k+1}$, where

$$\begin{array}{rcl} \mathcal{D}G(X^k)(X^{k+1} - X^k) & = & -G(X^k),\ k = 0, 1, \cdots \\ X^0 & = & U^j, \end{array}$$

for $\mathcal{D}G(X^k)$ the Jacobian matrix of G at the previous iteration X^k. The iteration procedure is truncated in one of the following two cases:

i) if the defect of Lobatto III A scheme, $L_j(\tau, U_j, X^k)$, becomes smaller than a prescribed tolerance (Tol).

   The defect of Lobatto III A is defined as (see Ch.2, (2.14))

$$\begin{array}{rcl} L_j(\tau, U_j, X^k) : & = & X^k - U^j - \frac{\tau}{6}(f(X^k) + f(U^j)) \\ & & -\frac{2\tau}{3} f\left(\frac{X^k + U^j}{2} + \frac{\tau}{8}(f(U^j) - f(X^k))\right). \end{array}$$

ii) if $L_j(\tau, U_j, X^{k+1}) \geq L_j(\tau, U_j, X^k)$. In this case a new time step is suggested and the iterative process is repeated.

The new time step is defined by a standard formula for adaptation of the time step [18, pp. 166]

$$\tau_{new} = \min(\tau_{max}, \max(\tau_{min}, 0.8\left(\frac{Tol}{err}\right)^{1/3})) \cdot \tau_{old}, \tag{3.40}$$

where τ_{max}, τ_{min}, and Tol are prescribed, and err is the local error. The modification, which we introduce, is the use of the defect of Lobatto III A scheme (2.14) for err. Formula (3.40) is also used in case of successful exit from the Newton method, that is, $L_j(\tau, U_j, X^k) \leq 10^{-2} \cdot Tol$, in order to obtain a new larger time step.

We refer to the time integrator described above as the time integrator of **AIM** (Adaptive Implicit Method). It uses the defect of Lobatto scheme (2.14) in adaptation of the time step and in optimization of the number of iterations in the Newton method. More about the time integrator of AIM is given in Ch. 2.3.

3.5.3 Computing the front

After the numerical solutions of problems P1 and P2 are advanced in time the new position of the front point should be updated, as well as the grid.

We start with the time discretization of the differential equation (3.7) which gives the position of the interface.

As we have seen in Theorem 5, $\dot{x_F}$ is continuous. Let us have a closer look at q_x

$$q_x = \begin{cases} D_0 u_{xx}, & |u_x(t,x)| < \bar{u}, \\ \\ (D_0 + D_1(2|u_x| - \bar{u}))u_{xx}, & |u_x(t,x)| > \bar{u}. \end{cases}$$

We express q_t as

$$q_t = \begin{cases} D_0 u_{xt} = D_0(u_t)_x = D_0(q_x + S)_x = D_0(q_{xx} + S_x), & |u_x| < \bar{u} \\ \\ (D_0 + D_1(2|u_x| - \bar{u}))u_{xt} = (D_0 + D_1(2|u_x| - \bar{u}))(q_{xx} + S_x), & |u_x| > \bar{u}. \end{cases}$$

This is possible provided that u_{tx} is continuous in each of the intervals $[0, x_F]$ and $[x_F, 1]$. Since the partial derivatives of first and second order of the coefficients of the two subproblems P1 and P2 possess higher smoothness, we obtain higher smoothness for the solutions as well (see [25, pp. 456]).

At the front point it holds

$$q_t(t, x_F(t)) = \begin{cases} D_0 \lim\limits_{\varepsilon \to 0}(q_{xx}(t, x_F - \varepsilon) + S_x(x_F - \varepsilon)), \\ \\ (D_0 + D_1\bar{u}) \lim\limits_{\varepsilon \to 0}(q_{xx}(t, x_F + \varepsilon) + S_x(x_F + \varepsilon)). \end{cases} \tag{3.41}$$

In this way, equation (3.7) can be rewritten as

$$\dot{x}_F = -\frac{q_{xx}(t, x_F) + S_x(x_F)}{u_{xx}(t, x_F)} =: g_F(t, x_F(t), u_{xx}(t, x_F(t)), q_{xx}(t, x_F(t))). \tag{3.42}$$

Let us discretize equation (3.42) using one step scheme

$$x_F^{j+1} = x_F^j + F(\tau, x_F^j, x_F^{j+1}), \tag{3.43}$$

where $F(\tau, x_F^j, x_F^{j+1})$ is a function defining a scheme of order p (we keep the notations from Ch.2). The local error of approximation takes the form

$$\begin{aligned}
|x_F(t_{j+1}) - x_F^{j+1}| &= g_F(t_j, x_F(t_j), u_{xx}(t_j, x_F(t_j)), q_{xx}(t_j, x_F(t_j)))\tau + \mathcal{O}(\tau^2) \\
&\quad -F(\tau, x_F(t_j), x_F^{j+1}) \\
&= \mathcal{O}(\tau^{p+1}).
\end{aligned}$$

To actually get a local error of size $\mathcal{O}(\tau^{p+1})$ not only u^{j+1}, x_F^j have to be known exactly but also u_{xx} and q_{xx}, at least at x_F^j. Since u_{xx} and q_{xx} always have to be approximated the total local error is of size $\hat{C}(h)\tau + \mathcal{O}(\tau^{p+1})$.

Remark 9. *The local error of approximation of a difference scheme of type (3.43) applied to the interface equation (3.42) is of size $\hat{C}(h)\tau$, where $\hat{C}$ is a constant depending on the spatial grid.*

Let us carry out the discretization of (3.42) through the explicit Euler method which leads to

$$x_F^{j+1} = x_F^j + \tau_j g_F^j, \tag{3.44}$$

where x_F^j and g_F^j approximate $x_F(t_j)$ and $g_F(t_j, x_F^j, u_{xx}(t_j, x_F^j), q_{xx}(t_j, x_F^j))$, respectively, and τ_j is the time step. The subscript F marks computation at the front point.

The discretization of the r.h.s. of (3.42) is carried out in agreement with the used finite element basis. The flux at the front point can be written as

$$q(t, x_F) \approx \begin{cases} D_0[u_x]_F, & \text{on the left side of } x_F, \\ \\ (D_0 + D_1([u_x]_F - \bar{u}))[u_x]_F, & \text{on the right side of } x_F, \end{cases}$$

where $[u_x]_F$ stands for the approximation of u_x at x_F. For linear elements $[u_x]_F$ is

$$[u_x]_F \approx \begin{cases} \frac{u_F - u_{F-1}}{h_{P1}}, & \text{on the left side of } x_F, \\ \\ \frac{u_{F+1} - u_F}{h_{P2}}, & \text{on the right side of } x_F, \end{cases}$$

where h_{P1} and h_{P2} are the spatial steps for the problems P1 and P2, respectively. For Hermitian elements, an approximation of $u_x(., x_i),\ i = 1, \cdots, n$ is available at each time t_j. This is due to the fact that we compute not only the function u itself but also its derivative at each grid point x_i.

Now using the following central differences for u_{xx} and q_{xx}

$$u_{xx} = \frac{2}{h_{P1} + h_{P2}}\left([u_x]_{F+1} - [u_x]_{F-1}\right) - u_{xxx}\frac{h_{P2}^2 - h_{P1}^2}{3(h_{P1} + h_{P2})} - u_x^{(IV)}\frac{h_{P2}^3 + h_{P1}^3}{12(h_{P1} + h_{P2})} + \cdots,$$

$$q_{xx} = \frac{2}{h_{P1}+h_{P2}}\left(\frac{q_{F+1}-q_F}{h_{P2}} - \frac{q_F-q_{F-1}}{h_{P1}}\right) - q_{xxx}\frac{h_{P2}^2-h_{P1}^2}{3(h_{P1}+h_{P2})} - q_x^{(IV)}\frac{h_{P2}^3+h_{P1}^3}{12(h_{P1}+h_{P2})} + \cdots$$

and neglecting the higher order terms we get for the new position of the front the expression

$$x_F^{j+1} = x_F^j - \tau_j\left(\frac{\frac{q_{F+1}-q_F}{h_{P2}} - \frac{q_F-q_{F-1}}{h_{P1}} + (h_{P1}+h_{P2})S_x(x_F)}{2([u_x]_{F+1}-[u_x]_{F-1})}\right). \tag{3.45}$$

This approximation has local error

$$\begin{aligned} |x_F(t_{j+1}) - x_F^{j+1}| &\leq \frac{\tau_j}{12}\left|\frac{4(h_{P2}-h_{P1})C1 + ((h_{P2}-h_{P1})^2 + h_{P1}h_{P2})C_2 + \cdots}{(u_F)_{xx}^2 + C_3(h_{P2}-h_{P1}) + ((h_{P2}-h_{P1})^2 + h_{P1}h_{P2})C_4 + \cdots}\right| \\ &\leq \tau_j h_-\left(\frac{4|C_1| + h_-|C_2|}{12(u_F)_{xx}^2 - 4h_-|C_3| + \mathcal{O}(h_-^2)}\right) + \mathcal{O}(\tau_j^2), \end{aligned}$$

where $h_- = \max(h_{P1}, h_{P2})$, $C_1 = ((q_F)_{xx} - S_F')(u_F)_{xxx} - (q_F)_{xxx}(u_F)_{xx}$, $C_2 = ((q_F)_{xx} - S_F')(u_F)_x^{(IV)} - (q_F)_x^{(IV)}(u_F)_{xxx}$, $C_3 = \frac{4}{3}(u_F)_{xx}(u_F)_{xxx}$, $C_4 = \frac{1}{12}(u_F)_{xx}(u_F)_x^{(IV)}$. Let us denote by

$$\hat{C}(h_{P1}, h_{P2}) := \frac{1}{12}\left|\frac{4(h_{P2}-h_{P1})C1 + ((h_{P2}-h_{P1})^2 + h_{P1}h_{P2})C_2 + \cdots}{(u_F)_{xx}^2 + C_3(h_{P2}-h_{P1}) + ((h_{P2}-h_{P1})^2 + h_{P1}h_{P2})C_4 + \cdots}\right|$$

i.e.

$$|x_F(t_{j+1}) - x_F^{j+1}| \leq \tau_j\hat{C}(h_{P1}, h_{P2}). \tag{3.46}$$

A proper choice of the spatial steps h_{P1} and h_{P2} can increase the order of convergence of (3.45), for instance

$$|x_F(t_j) - x_F^j| = \begin{cases} \mathcal{O}(\tau_j^2), & \text{for } h_{P2} = h_{P1} + \tau_j, \\ \mathcal{O}(\tau h_{P1}^2), & \text{for } h_{P2} = h_{P1}. \end{cases} \tag{3.47}$$

Remark 9 for the case of explicit Euler method states

Remark 10. *The local error of approximation of the explicit Euler method applied to the interface equation (3.42) is given by (3.46). For $h_{P2} = h_{P1}$ the local error has the size τh_{P1}^2.*

Hence,

$$|\dot{x}_F(t_j) - g_F^j| \leq Const\ h_{P1}^2, \tag{3.48}$$

provided that $h_{P2} = h_{P1}$.

The utilization of explicit scheme for the interface equation imposes some restrictions on the time step. We compute an approximation of the derivative of the r.h.s. function of (3.42)

$$\lambda^j = \left| \frac{g_F^j - g_F^{j-1}}{x_F^j - x_F^{j-1}} \right|.$$

If $|x_F^{j+1} - x_F^j|$ is greater than the size of the spatial step we change the size of the predicted time step using the PI step size control [19, pp. 33]. The suggested new time step is

$$\tau_{j+1} = \max\left(0.2, \min\left(\left(\frac{Tol}{\lambda^j \tau_j x_F^j}\right)^{0.5} \left(\frac{\lambda^{j-1}\tau_{j-1}x_F^{j-1}}{Tol}\right)^{0.3}, 0.8\right)\right)\tau_j.$$

At the same time the value of the tolerance Tol is halved, i.e. $Tol = Tol/2$.

Trapezoidal rule used for the time discretization of problem P, itself, does not require restriction on the size of the time step due to stability reasons. However, the explicit scheme used for the front equation does. Depending on Tol (the magnitude of the desired accuracy for solving the PDE) the stability requirement of the difference scheme for front equation may become severe.

3.5.4 Updating the grid

The front point is performed as a double grid point. Each of the subproblems is solved on a grid that partially varies from one time step to another. The majority of grid elements do not vary. We have a fixed, uniform, underlying grid that is sufficient to represent the solution in smooth regions (away from the front position). A finer grid around the location of the front is formed in a narrow *band* that has width of 4 regular steps. If a presence of a front point is identified then the *band* of finer grid is introduced through a subdivision of some regular elements. There are always four local uniform steps at each side of the interface.

Thus we have:

- away of the front - regular uniform mesh with spatial step h,
- around the front - a *band* with width $4h$,
- in the *band* - local steps: h_{P1} on the left side of the front and h_{P2} on the right side of the front and $4h_{P1} + 4h_{P2} = 4h$.

At each time level, the new position of the front is computed by (3.45). Changes in the position of the front inside the *band* change the size of the local steps h_{P1} and h_{P2}. In order to avoid too small local steps we move the *band* with one regular spatial step forward or backward whenever $\min(h_{P1}, h_{P2}) \leq 0.4h$.

Advantage of this strategy is that we interpolate the solution only inside of the *band*. Linear and spline interpolations are integrated in the numerical code.

3.6 Convergence analysis

Here, we discuss the convergence of the finite element approximation, described in the previous sections. We extend the convergence analysis performed by Thomee in [47] to the case where the diffusion coefficient depends on the gradient of the solution. Furthermore, we include the effect of the motion of the front. Since we have two subproblems, P1 and P2, we investigate their convergence separately.

Let us at first discuss the convergence of P2. We map the interval $[x_F(t), 1]$ to $[0, 1]$, i.e.

$$x \mapsto x' = \frac{x - x_F}{1 - x_F}.$$

In the new variable P2 reads

$$\begin{cases} u_t = \nabla(D(|\nabla u|)\nabla u) + S(t,x), \quad 0 < x' < 1, \ t > 0 \\ u(0,x') = u_0(x'), \quad 0 \le x' \le 1 \\ |\nabla u(t,0)| = (1 - x_F(t))\bar{u}, \quad t > 0, \\ u(t,1) = 0, \quad t > 0, \end{cases}$$

where

$$D(|\nabla u|) = \frac{1}{(1 - x_F(t))^2}(D_0 + D_1(\frac{|\nabla u|}{1 - x_F(t)} - \bar{u})),$$

$$S(t,x') = S_0 \exp(-\frac{(x'(1 - x_F(t)) + x_F(t) - x_0)^2}{\delta^2}).$$

We omit the prime in the following considerations. We denote by $||.||$ the norm in L_2 and by $||.||_r$ that in the Sobolev space $H^r = W_2^r([0,1])$ so that for the real-valued functions v

$$||v|| = ||v||_{L_2} = \left(\int_0^1 v^2 dx \right)^{1/2},$$

and for r a positive integer

$$||v||_r = ||v||_{H^r} = \left(\sum_{\alpha \le r} ||\frac{\partial^\alpha v}{\partial x^\alpha}||^2 \right)^{1/2}.$$

We consider space $W = \{v \in W_2^r([0,1]), \ v(1) = 0\}$ and divide the interval $[0,1]$ into $0 = x_0 < x_1 < \cdots < x_n = 1$ with $h = x_i - x_{i-1}$. Let W_h be the corresponding finite dimensional subspace of W consisting of continuous functions on $[0,1]$ which reduce to piecewise polynomials of degree k in each of the subintervals $[x_i, x_{i+1}]$ and vanish at $x = 1$.

Assumption 3. *We assume that*

- *the solution of P1 exists and it is in* $W = H^r$*,*
- *the solution of P2 exists and it is in* $W = \{v \in W_2^r([0,1]),\ v(1) = 0\}$*,*
- *there exists a positive constant* A *such that*

$$0 < D_0 \leq \frac{D_0}{(1-x_F)^2} \leq D(|\nabla u|) \leq A,$$

- *there exists* $a \in (0,1)$ *such that* $a \leq x_F(t) < 1$*, for any* $t \geq 0$
- *for small space step* h *and* $1 \leq r \leq k+1$

$$\inf_{\xi \in W_h} ||v - \xi|| + h||\nabla(v - \xi)|| \leq Ch^r ||v||_r, \tag{3.49}$$

for any $v \in W$*.*

Because of the smoothness property of the solution u we should restrict to the case $r = 2$.

Under assumption (3.49) functions from W and their gradients may be approximated by functions from W_h with an optimal order of $\mathcal{O}(h^{k+1})$ and $\mathcal{O}(h^k)$, respectively ($k = 1$ for linear elements, $k = 2$ for quadratic, $k = 3$ for cubic). This optimal order could be achieved if the solution is smooth enough, that is, for example 4^{th} order for cubic (Hermitian) elements only if $u \in H^4$.

We may then pose the semidiscrete problem to find $u_h : \bar{J} \longrightarrow W_h$, for J interval of time, such that:

$$\begin{aligned} (\partial_t u_h, \chi) &+ (D(|\nabla u_h|)\nabla u_h, \chi) \\ &= (S(t), \chi) + \frac{D_0}{(1-x_F)} s\bar{u}\chi(0), \quad \chi \in W_h, t \in J, \\ u_h(0) &= u_{0h}, \end{aligned} \tag{3.50}$$

where u_{0h} is an approximation of u_0 in W_h, and s is the sign of $u_x(t, x_F)$. The solution, $u_h(x,t)$, can be written as $u_h(x,t) = \sum\limits_{j=1}^{n} \alpha_j(t)\phi_j(x)$, where $\{\phi_j\}_{j=1}^n$ is the standard nodal basis. Equation (3.50) may be written as

$$\begin{aligned} \sum_{j=1}^{n} \alpha_j'(t)(\phi_j, \phi_k) &+ \sum_{j=1}^{n} \alpha_j(t)(D(|\sum_{l=1}^{n} \alpha_l \nabla\phi_l|)\nabla\phi_j, \nabla\phi_k) \\ &= (S(t), \phi_k), \quad k = 1, \cdots, n. \end{aligned} \tag{3.51}$$

Now, we set $\alpha(t) = (\alpha_1(t), \cdots, \alpha_n(t))^T$, and denote by $M = \{m_{ij}\}$ and $K(\alpha) = \{k_{ij}(\alpha)\}$ the matrices with elements

$$m_{ij} = (\phi_i, \phi_j) \text{ and } \; k_{ij} = (D(|\sum_{l=1}^{n} \alpha_l \nabla \phi_l|)\nabla \phi_i, \nabla \phi_j),$$

and by $\tilde{S}(t) = (S_1(t), \cdots, S_n(t))$ the vector with elements $S_i(t) = (S(t), \phi_i)$. Then, system (3.51) can be written in matrix form as

$$M\alpha' + K(\alpha)\alpha = \tilde{S}(t).$$

The last element of the vector $\tilde{S}$ is corrected with the boundary condition $\frac{D_0}{(1-x_F)} s\bar{u}$. To solve the obtained system of ODEs, we use the Trapezoidal rule, which in this case reads

$$\begin{aligned}(\bar{\partial} U^j, \chi) \;&+\; \frac{1}{2}(D(\nabla U^j)\nabla U^j, \nabla \chi) \\ &+\; \frac{1}{2}(D(\nabla U^{j-1})\nabla U^{j-1}, \nabla \chi) \\ &=\; (\frac{S^j + S^{j-1}}{2}, \chi) + D_0 s\bar{u}\chi(0)(\frac{1}{2(1-x_F^j)} + \frac{1}{2(1-x_F^{j-1})}),\end{aligned} \tag{3.52}$$

for all $\chi \in W_h$, $t_j \in J$ and $U^0 = u_0$. Here U^j, S^j and x_F^j are approximations of $\alpha(t_j)$, $S(t_j)$, $x_F(t_j)$, respectively. The notation $\bar{\partial} U^j$ stands for $\bar{\partial} U^j = (U^j - U^{j-1})/\tau$, where τ is the time step. In matrix form equation (3.52) can be written as

$$(M + \frac{\tau}{2}K(U^j))U^j = (M - \frac{\tau}{2}K(U^{j-1}))U^{j-1} + \frac{\tau}{2}(\tilde{S}^j + \tilde{S}^{j-1}). \tag{3.53}$$

In order to show that there exists a solution of equation (3.53), we rewrite it as

$$\begin{aligned}(G_h(\chi), \chi) \;:=\;& 2(\chi - U^{j-1}, \chi) + \tau(D(\nabla \chi)\nabla \chi, \nabla \chi) \\ &+ \tau(D(\nabla U^{j-1})\nabla U^{j-1}, \nabla \chi) - \tau(S^j + S^{j+1}, \chi) = 0,\end{aligned}$$

where $G_h : W_h \to W_h$ is continuous. As a consequence of the Brouwer fixed point theorem, the equation $G_h(X) = 0$ has a solution $X \in B_q = \{\chi \in W_h : \; ||\chi|| \leq q\}$ if $(G_h(\chi), \chi) > 0$ for $||\chi|| = q$. Then,

$$\begin{aligned}(G_h(\chi), \chi) \;&=\; 2(\chi - U^{j-1}, \chi) + \tau(D(\nabla \chi)\nabla \chi, \nabla \chi) \\ &\quad + \tau(D(\nabla U^{j-1})\nabla U^{j-1}, \nabla \chi) - \tau(S^j + S^{j+1}, \chi) \\ &\geq\; ||\chi||^2 - ||U^{j-1}||^2 + \tau D_0 ||\nabla \chi||^2 \\ &\quad + \tau D_0 (\nabla U^{j-1}, \nabla \chi) - \tau S_0 ||\chi||^2 \\ &\geq\; ||\chi||^2 - ||U^{j-1}||^2 + \tau D_0 (1 + \bar{u})||\nabla \chi||^2 - \tau S_0 ||\chi||^2,\end{aligned}$$

which is positive if $||\chi||$ is large enough and $\tau \leq \tau_0(D_0, D_1, \bar{u}, S_0)$.
Now, we consider the error of the fully discrete problem written as a sum of two terms:

$$U^j - u^j = (U^j - \tilde{U}^j) + (\tilde{U}^j - u^j) =: \theta^j + \rho^j, \tag{3.54}$$

with $u^j = u(t_j)$ and $\tilde{U}^j := \tilde{u}_h(t_j)$ where $\tilde{u}_h$ is the so call elliptic projection in W_h of the exact solution $u(t)$. We use the projection $\tilde{u}_h = \tilde{u}_h(t)$ defined by

$$(D(|\nabla u|)\nabla(\tilde{u}_h(t) - u(t)), \chi) = 0, \ \forall \chi \in W_h. \tag{3.55}$$

Let us remark for later reference that, because the family W_h is based on a family of quasiuniform triangulation $\mathcal{T}_h$, and W_h consists of piecewise polynomials of degree k, holds the *inverse inequality*

$$||\nabla \chi|| \leq Ch^{-1}||\chi||, \quad \forall \chi \in W_h. \tag{3.56}$$

We begin with the following auxiliary results.

Lemma 9. *Let $b = b(x)$ be a smooth function in $[0,1]$ with $D_0 \leq b(x) \leq A$, for $x \in [0,1]$ and $A = const > 0$. Assume that $w \in W$ and let w_h be defined by*

$$(b\nabla(w_h - w), \nabla \chi) = 0, \ \forall \chi \in W_h.$$

Then,

$$||\nabla(w_h - w)|| \leq C_1 h^{r-1}||w||_r \tag{3.57}$$

and

$$||w_h - w|| \leq C_0 h^r ||w||_r. \tag{3.58}$$

The constants C_1 and C_0 depend on D_0, D_1, $\bar{u}$.

Proof
For $\chi \in W_h$, we have

$$\begin{aligned} D_0||\nabla(w_h - w)||^2 &\leq (b\nabla(w_h - w), \nabla(w_h - w)) \\ &= (b\nabla(w_h - w), \nabla(\chi - w)) \\ &\leq A||\nabla(w_h - w)||||\nabla(\chi - w)||, \end{aligned}$$

and for $\chi = I_h w$, the standard interpolation of w, we get

$$||\nabla(w_h - w)|| \ \leq \ (A/D_0)||\nabla(I_h w - w)|| \leq C_1 h^{r-1}||w||_r. \tag{3.59}$$

For the L_2 norm we proceed by a duality argument. Let φ be arbitrary in L_2, and take ψ as the solution of

$$\begin{aligned} -\nabla \cdot (b\nabla\psi) \equiv -\nabla b \cdot \nabla\psi - b\Delta\psi = \varphi, \\ \psi(0) = \psi'(0) = 0. \end{aligned} \tag{3.60}$$

Since $||\psi|| \leq C||\nabla\psi||$ it follows

$$\begin{aligned} D_0||\nabla\psi||^2 &\leq (b\nabla\psi,\nabla\psi) = -(\nabla\cdot(b\nabla\psi),\psi) \\ &\leq (\varphi,\psi) \leq C||\varphi||||\nabla\psi||, \end{aligned} \tag{3.61}$$

so that $||\nabla\psi|| \leq C||\varphi||$. Using the *elliptic regularity inequality* ($b(x) > 0$) and for ∇b bounded, one obtains

$$||\psi||_2 \leq C||\Delta\psi|| \leq C||b\Delta\psi|| = C||\varphi + \nabla b\nabla\psi|| \leq C||\varphi||. \tag{3.62}$$

Making use of (3.59) together with (3.49) with $r = 2$ we get

$$\begin{aligned} (w_h - w,\varphi) &= (b\nabla(w_h - w),\nabla\psi) \leq (b\nabla(w_h - w),\nabla(\psi - \mathcal{I}_h\psi)) \\ \\ &\leq A||\nabla(w_h - w)||||\nabla(\psi - \mathcal{I}_h\psi)|| \leq Ch^{r-1}||w||_r h||\psi||_2 \\ &\leq Ch^r||w||_r||\varphi||. \end{aligned}$$

♢

Lemma 10. *With $\tilde{u}_h$ defined by (3.55) and $\rho = \tilde{u}_h - u$ we have under the appropriate regularity on u*

$$\begin{aligned} ||\rho(t)|| + h||\nabla\rho(t)|| \leq C(u)h^r, &\quad \textit{for } t \in J, \\ ||\rho_t(t)|| + h||\nabla\rho_t(t)|| \leq C(u)h^r, &\quad \textit{for } t \in J, \end{aligned} \tag{3.63}$$

where $C(u)$ is independent of $t \in J$.

Proof

The first estimate follows from Lemma 9 with $b(x) = D(|\nabla u|)$, where we assume that $\nabla b = D's\Delta u$ is bounded (s is the sign of ∇u).

By differentiation (3.55) with respect to t we have

$$(D'(|\nabla u|)|\nabla u_t|\nabla\rho,\nabla\chi) + (D(|\nabla u|)\rho_t,\nabla\chi) = 0, \quad \forall\chi \in W_h.$$

Assuming $D(|\nabla u|)$ and $D'(|\nabla u|)$ uniformly bounded, we get

$$\begin{aligned} D_0||\nabla\rho_t||^2 &\leq (D(|\nabla u|)\nabla\rho_t,\nabla\rho_t) \\ &= (D(|\nabla u|)\nabla\rho_t,\nabla(\chi - u_t)) + (D(|\nabla u|)\nabla\rho_t,\nabla(\tilde{u}_{t,h} - \chi)) \\ &= (D(|\nabla u|)\nabla\rho_t,\nabla(\chi - u_t)) + (\frac{D_1|\nabla u_t|}{(1-x_F)^2}\nabla\rho,\nabla(\chi - \tilde{u}_{t,h}) \\ &\leq C(||\nabla\rho_t||||\nabla(\chi - u_t)|| + ||\nabla\rho||||\nabla(\chi - \tilde{u}_{t,h})||), \end{aligned}$$

and setting $\chi = \mathcal{I}_h u_t$ we obtain

$$\begin{aligned} D_0||\nabla\rho_t||^2 &\leq Ch^{r-1}||u_t||_r||\nabla\rho_t|| + ||\nabla\rho||(Ch^{r-1}||u_t||_r + ||\nabla\rho_t||) \\ &\leq ||\nabla\rho_t||^2 + C(||\nabla\rho||^2 + h^{2(r-1)}||u_t||_r^2). \end{aligned}$$

Taking into account the estimate for $\nabla\rho$ we obtain the second inequality $||\nabla\rho_t|| \leq C(u)h^{r-1}$.
For the L_2 estimate we use again the duality argument and we follow exactly the proof of Lemma 9 to get

$$|(\rho_t, \varphi)| \leq C(||\nabla\rho_t||h||\psi||_2 + ||\nabla\rho||h||\psi||_2 + ||\rho||||\psi||_2)$$

whence, by the already shown estimates for ρ, $\nabla\rho$ and $\nabla\rho_t$ we conclude

$$|(\rho_t, \varphi)| \leq C(u)h^r||\psi||_2 \leq C(u)||\varphi||.$$

♢

Lemma 11. *For $\tilde{u}_h$ defined as above, we have*

$$||\nabla\tilde{u}_h||_\infty \leq C(u).$$

Proof
Using that $\nabla\chi$ is constant on each subinterval

$$||\nabla\chi||_\infty \leq Ch^{-1}||\nabla\chi||, \quad \text{for } \chi \in W_h, \tag{3.64}$$

together with the result of Lemma 10 and the estimate for $\mathcal{I}_h u$ we have

$$\begin{aligned} ||\nabla(\tilde{u}_h - \mathcal{I}_h u)||_\infty &\leq Ch^{-1}||\nabla(\tilde{u}_h - \mathcal{I}_h u)|| \\ &\leq Ch^{-1}(||\nabla\rho|| + ||\nabla(\mathcal{I}_h u - u)||) \leq C(u). \end{aligned}$$

Since $||\mathcal{I}_h u||_\infty \leq C||\nabla u||_\infty$ the result follows.

♢

Theorem 7. *Let U^j and $u(t_j)$ be solutions of (3.52) and P2 at $t = t_j$, respectively. Then, under the appropriate regularity assumptions for u, we have, for small τ,*

$$||u(t_j) - U^j|| \leq ||u_0 - u_{0,h}|| + C(\tau^2 + h^r + \tau h^2 + h^{r-1}), \quad \textit{for } t_j \in J.$$

Proof
As before we write

$$U^j - u(t_j) = (U^j - \tilde{U}^j) + (\tilde{U}^j - u(t_j)) = \theta^j + \rho^j.$$

It remains to bound θ^j. We have for $\chi \in W_h$

$$(\bar{\partial}\theta^j, \chi) \quad + \quad \frac{1}{2}(D(|\nabla U^j|)\nabla\theta^j, \nabla\chi) + \frac{1}{2}(D(|\nabla U^{j-1}|)\nabla\theta^{j-1}, \nabla\chi) = \tag{3.65}$$

$$\begin{aligned}
&= (\bar{\partial}U^j,\chi) + \frac{1}{2}(D(|\nabla U^j|)\nabla U^j,\nabla\chi) + \frac{1}{2}(D(|\nabla U^{j-1}|)\nabla U^{j-1},\nabla\chi) \\
&\quad -(\bar{\partial}\tilde{U}^j,\chi) - \frac{1}{2}(D(|\nabla U^j|)\nabla\tilde{U}^j,\nabla\chi) - \frac{1}{2}(D(|\nabla U^{j-1}|)\nabla\tilde{U}^{j-1},\nabla\chi) \\
&= (\frac{S^j+S^{j-1}}{2},\chi) + \frac{D_0 s\bar{u}\chi(0)}{2(1-x_F^j)} + \frac{D_0 s\bar{u}\chi(0)}{2(1-x_F^{j-1})} \pm (\frac{S(t_j)+S(t_{j-1})}{2},\chi) \\
&\quad \pm\frac{D_0 s\bar{u}\chi(0)}{2}\left(\frac{1}{1-x_F(t_{j-1})}+\frac{1}{1-x_F(t_j)}\right) - (\bar{\partial}\tilde{U}^j \pm \frac{u_t(t_j)+u_t(t_{j-1})}{2},\chi) \\
&\quad -\frac{1}{2}(D(|\nabla u(t_j)|)\nabla\tilde{U}^j,\nabla\chi) - \frac{1}{2}(D(|\nabla u(t_{j-1})|)\nabla\tilde{U}^{j-1},\nabla\chi) \\
&\quad -\sum_{k=j-1}^{j}(\frac{D(|\nabla U^k|)-D(|\nabla u(t_k)|)}{2}\nabla\tilde{U}^k,\nabla\chi) \\
&= -(\bar{\partial}\tilde{U}^j - \frac{u_t(t_j)+u_t(t_{j-1})}{2},\chi) + (\frac{S^j-S(t_j)}{2}+\frac{S^{j-1}-S(t_{j-1})}{2},\chi) \\
&\quad +\sum_{k=j-1}^{j}\frac{D_0 s\bar{u}\chi(0)}{2}(\frac{x_F(t_k)-x_F^k}{(1-x_F^k)(1-x_F(t_k))}) \\
&\quad -\sum_{k=j-1}^{j}\frac{1}{2}(D(|\nabla u(t_k)|)(\nabla\tilde{U}^k-u(t_k)),\nabla\chi) \\
&\quad -\sum_{k=j-1}^{j}(\frac{D(|\nabla U^k|)-D(|\nabla u(t_k)|)}{2}\nabla\tilde{U}^k,\nabla\chi) \\
&= (\frac{u_t(t_j)+u_t(t_{j-1})}{2}-\bar{\partial}u(t_j),\chi) - (\bar{\partial}\rho^j,\chi) \\
&\quad +\sum_{k=j-1}^{j}(\frac{S^k-S(t_k)}{2},\chi) - \sum_{k=j-1}^{j}(\frac{D(|\nabla u(t_k)|)-D(|\nabla U^k|)}{2}\nabla\tilde{U}^k,\nabla\chi) \\
&\quad +\frac{D_0 s\bar{u}\chi(0)}{2}\sum_{k=j-1}^{j}\frac{x_F(t_k)-x_F^k}{(1-x_F^k)(1-x_F(t_k))}.
\end{aligned}$$

We set $\chi = \frac{\theta^j+\theta^{j-1}}{2}$. Let us consider separately some of the terms taking part in the above estimate. First, we start with the terms where the diffusion coefficient takes part, for this purpose we look at

$$|D(|\nabla u(t_j)|) - D(|\nabla U^j|)| \leq C_0|x_F(t_j)-x_F^j| + C_1|\nabla u(t_j)-\nabla U^j|,$$

where $C_0 = C_0(\bar{u},D_0,D_1,\nabla u)$ is a constant, as well as C_1. In the light of the above estimate and using that $||\nabla\tilde{U}^j||_\infty \leq C(u)$ we have

$$|(\frac{D(|\nabla u(t_j)|)-D(|\nabla U^j|)}{2}\nabla\tilde{U}^j,\nabla\frac{\theta^j+\theta^{j-1}}{2})| \quad\leq\quad C_1(|\nabla\frac{u(t_j)-U^j}{2}|\nabla\tilde{U}^j,\nabla\frac{\theta^j+\theta^{j-1}}{2})$$

$$\begin{aligned} &+C_0|x_F(t_j)-x_F^j|(\nabla\tilde{U}^j,\nabla\frac{\theta^j+\theta^{j-1}}{2}) \\ \leq\ & C_1||\nabla\frac{u(t_j)-U^j}{2}||||\nabla\frac{\theta^j+\theta^{j-1}}{2}|| \\ &+C_0|x_F(t_j)-x_F^j|||\nabla\frac{\theta^j+\theta^{j-1}}{2}||. \end{aligned}$$

The terms where the source plays role we treat as follows

$$|\tilde{S}^j-\tilde{S}(t_j)|\leq CS_0|x_F(t_j)-x_F^j|.$$

Let us go back to the main equation (3.65) and substitute the above estimates

$$\begin{aligned} \bar{\partial}||\theta^j||^2 \leq\ & \frac{C_1}{2}(||\nabla\frac{u^j-U^j}{2}||^2+||\nabla\frac{u^{j-1}-U^{j-1}}{2}||^2)+(D_0\bar{u})^2C|x_F(t_j)-x_F^j|^2 \\ &+(D_0\bar{u})^2C|x_F(t_{j-1})-x_F^{j-1}|^2 \\ &+\frac{1}{2}||\frac{u_t^j+u_t^{j-1}}{2}-\bar{\partial}u^j-\bar{\partial}\rho^j||^2+||\frac{\theta^j+\theta^{j-1}}{2}||^2 \\ &+CS_0(|x_F(t_j)-x_F^j|+|x_F(t_{j-1})-x_F^{j-1}|)||\frac{\theta^j+\theta^{j-1}}{2}|| \\ \leq\ & C(||\nabla\theta^j+\nabla\rho^j||^2+||\nabla\theta^{j-1}+\nabla\rho^{j-1}||^2) \\ &+2||\frac{\theta^j+\theta^{j-1}}{2}||^2+(D_0\bar{u})^2C|x_F(t_j)-x_F^j|^2 \\ &+(D_0\bar{u})^2C|x_F(t_{j-1})-x_F^{j-1}|^2+\frac{1}{2}||\frac{u_t^j+u_t^{j-1}}{2}-\bar{\partial}u^j-\bar{\partial}\rho^j||^2 \\ &+CS_0^2(|x_F(t_j)-x_F^j|^2+|x_F(t_{j-1})-x_F^{j-1}|^2). \end{aligned}$$

In order to estimate $||\nabla\theta^j||$ we apply the inverse inequality (3.56) for $\chi=\theta\in W_h$. Denoting with

$$\begin{aligned} R^j =\ & ||\frac{u_t^j+u_t^{j-1}}{2}-\bar{\partial}u^j||^2+||\bar{\partial}\rho^j||^2+C(||\nabla\rho^j||^2+||\nabla\rho^{j-1}||^2) \\ +\ & C((D_0\bar{u})^2+S_0^2)(|x_F(t_j)-x_F^j|^2+|x_F(t_{j-1})-x_F^{j-1}|^2), \end{aligned} \tag{3.66}$$

we get

$$(1-\tau C)||\theta^j||^2 \leq \tau R^j+(1+\tau C)||\theta^{j-1}||^2,$$

or for small τ, by repeated application we obtain

$$\begin{aligned} ||\theta^j||^2 \leq\ & (1+2\tau C)||\theta^{j-1}||^2+\tau R^j \\ \leq\ & (1+2\tau C)^j||\theta^0||^2+\tau\sum_{l=0}^{j}(1+2\tau C)^{j-l}R^l \\ \leq\ & C||\theta^0||^2+C\tau\sum_{l=0}^{j}R^l. \end{aligned}$$

For the terms in R^l are fulfilled

$$||\bar{\partial}\rho^j|| = ||\frac{1}{\tau}\int\limits_{t_{j-1}}^{t_j} \rho_t ds|| \leq C(u)h^r;$$

$$\begin{aligned}
\left(u(t_j) - u(t_{j-1}) - \tau\frac{u_t^j + u_t^{j-1}}{2}\right) &= \\
&= \int\limits_{t_{j-1}}^{t_j} ((s-t_j)(s-t_{j-1}) + (s-t_{j-1/2})^2)u_{ttt}ds \\
&\leq \tau^2(\int\limits_{t_{j-1}}^{t_j} u_{ttt}ds),
\end{aligned}$$

and

$$||\bar{\partial}u^j - \frac{u_t^j + u_t^{j-1}}{2}|| \leq \tau^2||u_{ttt}||_{L_2\cap L_2}.$$

For the front, in case of equidistant mesh, holds (3.47), i.e.

$$|x_F(t_j) - x_F^j| \leq C\tau h^2.$$

Since $||\nabla\rho||$ takes part in R^j, the order of convergence, with respect to x, drops with one unit

$$\begin{aligned}
\tau\sum_{l=1}^{j} R^l &= \tau\sum_{l=1}^{j}\{||\frac{u_t^l + u_t^{l-1}}{2} - \bar{\partial}u^l||^2 + ||\bar{\partial}\rho^l||^2 + C(||\nabla\rho^l||^2 + ||\nabla\rho^{l-1}||^2) \\
&+ C((D_0\bar{u})^2 + S_0^2)(|x_F(t_l) - x_F^l|^2 + |x_F(t_{l-1}) - x_F^{l-1}|^2)\} \\
&\leq C(\tau^2 + h^r + h^{r-1} + h^2\tau)^2.
\end{aligned}$$

The final estimate for the error is

$$||u(t_j) - U^j|| \leq ||u_0 - u_{0,h}|| + C(\tau^2 + h^r + \tau h^2 + h^{r-1}).$$

$\diamondsuit$

Now, we concentrate our attention to the first problem P1. The transformation

$$x \mapsto x' = \frac{x}{x_F}$$

maps the interval $[0, x_F(t)]$ to $[0,1]$ and brings us to

$$\begin{cases} u_t = D(t)\Delta u + S(t,x'), \quad 0 < x' < 1, \ t > 0 \\ u(0,x') = u_0(x'), \quad 0 \leq x' \leq 1 \\ \nabla u(t,0) = 0, \quad t > 0 \\ |\nabla u(t,1)| = \bar{u} x_F(t), \qquad t > 0. \end{cases}$$

where $D(t) = \frac{D_0}{x_F(t)^2}$. We skip the prime in the notations.

Here, W_h is the corresponding finite dimensional subspace of H^r, $r = 2$, of continuous functions on $[0,1]$ which reduce to piecewise polynomials of degree k in each of the subintervals $[x_i, x_{i+1}]$. The semidiscrete problem reads:

find $u_h : \bar{J} \longrightarrow W_h$ such that

$$\begin{aligned} (\partial_t u_h, \chi) + (D(t)\nabla u_h, \nabla\chi) &= (S(t),\chi) - \frac{D_0}{x_F(t)} s \ \bar{u}\chi(1), \quad \chi \in W_h, t \in J, \\ u_h(0) &= u_{0h}, \end{aligned} \tag{3.67}$$

where $s = \pm 1$ is the sign of $u_x(t,1)$, and u_{0h} is an approximation of u_0 in W_h. Representing the solution as $u_h(x,t) = \sum_{j=1}^{n} \alpha_j(t)\phi_j(x)$, where $\{\phi_j\}_{j=1}^n$ is the standard nodal basis, equation (3.67) may be written as

$$\sum_{j=1}^{n} \alpha_j'(t)(\phi_j,\phi_k) + D(t)\sum_{j=1}^{n} \alpha_j(t)(\nabla\phi_j, \nabla\phi_k) = (S,\phi_k) - \frac{D_0 s}{x_F} \ \bar{u}\phi_n(1), \quad k = 1, \cdots, n.$$

Denoting, again, $\alpha(t) = (\alpha_1(t), \cdots, \alpha_n(t))^T$, and introducing the standard mass and stiff matrices, $M = \{m_{ij}\}$ and $K(\alpha) = \{k_{ij}(\alpha)\}$, and the vector $\tilde{S}(t) = (S_1(t), \cdots, S_n(t))$, where $S_i(t) = (S(t), \phi_i)$, the system may be written in matrix form

$$M\alpha' + D(t)K\alpha = \tilde{S}(t).$$

The last element of the vector $\tilde{S}$ is corrected with the boundary condition $-\frac{D_0}{x_F}\bar{u}$. We solve the so obtained system of ODEs using the Trapezoidal rule, which in this case reads

$$\begin{aligned} (\bar{\partial} U^j, \chi) \ &+ \ \frac{D^j}{2}(\nabla U^j, \nabla\chi) + \frac{D^{j-1}}{2}(\nabla U^{j-1}, \nabla\chi) \\ &= \ (\frac{S^j + S^{j-1}}{2}, \chi) - D_0 \bar{u} s \chi(1)(\frac{1}{2x_F^j} + \frac{1}{2x_F^{j-1}}), \end{aligned} \tag{3.68}$$

for $\forall \chi \in W_h, \ t_j \in J$ with $U^0 = u_0$. Again U^j, S^j approximate $\alpha(t_j)$, $S(t_j)$, τ is the time step, $D^j = \frac{D_0}{(x_F^j)^2}$, with x_F^j an approximation of $x_F(t_j)$, and $\bar{\partial} U^j =$

$(U^j - U^{j-1})/\tau$. In matrix form equation (3.68) may be written as

$$(M + \frac{\tau D^j}{2}K)\alpha^j = (M - \frac{\tau D^{j-1}}{2}K)\alpha^{j-1} + \frac{\tau}{2}(\tilde{S}^j + \tilde{S}^{j-1}).$$

We consider the error of the fully discrete problem written as a sum of two terms:

$$U^j - u(t_j) = (U^j - R_h u(t_j)) + (R_h u(t_j) - u(t_j)) =: \theta^j + \rho^j, \tag{3.69}$$

where $R_h u(t)$ is the elliptic projection of the exact solution $u(t)$ onto W_h, so that

$$(\nabla(R_h u - u(t)), \nabla\chi) = 0, \ \forall\chi \in W_h. \tag{3.70}$$

Lemma 12. *Assume that (3.49) holds. Then we have for the solution u of P1*

$$||\rho|| + h||\nabla\rho|| \leq Ch^r(||u_0||_r + \int_0^t ||u_t(\tau)||_r)d\tau.$$

Proof

The proof follows from (3.49), together with

$$||\tilde{u}_h - u|| + h||\nabla(\tilde{u}_h - u)|| \leq Ch^r||u||_r,$$

and the fact that $u(t) = u(0) + \int_0^t u_t(\tau)d\tau$.

♢

Theorem 8. *Let U^j and $u(t_j)$ be solutions at $t = t_j$ of (3.68) and P1, respectively. Then,*

$$||U^j - u(t_j)|| \leq Ch^r||u_0||_r + C(h^r + \tau^2 + h^2\tau).$$

Proof

Let us consider

$$\begin{aligned} I : &= (\bar{\partial}\theta^j, \chi) + (\frac{D^j\nabla\theta^j + D^{j-1}\nabla\theta^{j-1}}{2}, \nabla\chi) \\ &= (\bar{\partial}U^j, \chi) + (\nabla\frac{D^jU^j + D^{j-1}U^{j-1}}{2}, \nabla\chi) \\ &\quad -(\bar{\partial}R_h u(t_j), \chi) - (\nabla\frac{D^jR_h u(t_j) + D^{j-1}R_h u(t_{j-1})}{2}, \nabla\chi). \end{aligned}$$

With the help of (3.70), we obtain

$$I = (\frac{S^j + S^{j-1}}{2}, \chi) - \frac{\chi(1)D_0 s\bar{u}}{2}(\frac{1}{x_F^j} + \frac{1}{x_F^{j-1}}) - (\frac{D^j\nabla u(t_j) + D^{j-1}\nabla u(t_{j-1})}{2}, \nabla\chi)$$

$$
\begin{aligned}
&-(\bar{\partial} R_h u(t_j),\chi) \pm (\frac{u_t(t_j)+u_t(t_{j-1})}{2},\chi) \\
= \;& \frac{1}{2}(u_t(t_j)+u_t(t_{j-1}),\chi) - (R_h\bar{\partial}u(t_j),\chi) + \sum_{k=j-1}^{j}(\frac{S^k - S(t_k)}{2},\chi) \\
& - \sum_{k=j-1}^{j}(\frac{D^k - D(t_k)}{2}\nabla u(t_k),\nabla\chi) - \frac{\chi(1)D_0 s\bar{u}}{2}\sum_{k=j-1}^{j}(\frac{1}{x_F^k} - \frac{1}{x_F(t_k)}) \\
=:\;& (w^j,\chi) - v_1 - v_2,
\end{aligned}
$$

where $w^j := w_1^j + w_2^j + w_3^j$, $w_1^j := (R_h - I)\bar{\partial}u(t_j)$, $w_2^j := (\bar{\partial}u(t_j) - \frac{u_t^j + u_t^{j-1}}{2})$, $w_3^j := \frac{S^j - S(t_j)}{2} + \frac{S^{j-1} - S(t_{j-1})}{2}$, $v_1 := \sum\limits_{k=j-1}^{j}(\frac{D^k - D(t_k)}{2}\nabla u(t_k),\nabla\chi)$, and $v_2 := \frac{\chi(1)D_0 s\bar{u}}{2}\sum\limits_{k=j-1}^{j}(\frac{1}{x_F^k} - \frac{1}{x_F(t_k)})$. Choosing $\chi = \frac{\theta^j + \theta^{j-1}}{2}$ and using the fact that $v_1 + v_2$ may be estimated by

$$
\begin{aligned}
v_1 + v_2 &\le (|\frac{D^j - D(t_j)}{2}|||\Delta u(t_j)|| + |\frac{D^{j-1} - D(t_{j-1})}{2}|||\Delta u(t_{j-1})||)||\chi|| \\
&\le \frac{CD_0}{a^4}(|x_F(t_j) - x_F^j|| + |x_F(t_{j-1}) - x_F^{j-1}|)||\chi||,
\end{aligned}
$$

we get

$$
\begin{aligned}
||\theta^j||^2 - ||\theta^{j-1}||^2 &\le \tau(||w^j|| + C\frac{D_0}{a^4}\sum_{k=j-1}^{j}|x_F(t_k) - x_F^k|)(||\theta^j|| + ||\theta^{j-1}||); \\
||\theta^j|| - ||\theta^{j-1}|| &\le \tau(||w^j|| + \frac{D_0}{a^4}\sum_{k=j-1}^{j}||\Delta u(t_k)|||x_F(t_k) - x_F^k|),
\end{aligned}
$$

or

$$
\begin{aligned}
||\theta^j|| \le\;& ||\theta^0|| + \tau\sum_{k=1}^{j}||w^k|| + \frac{2D_0\tau}{a^4}\sum_{k=1}^{j-1}||\Delta u(t_k)|||x_F(t_k) - x_F^k| \\
& + \frac{D_0\tau}{a^4}(||\Delta u(t_j)|||x_F(t_j) - x_F^j|) + ||\Delta u(t_0)|||x_F(t_0) - x_F^0|).
\end{aligned}
$$

Let us express w_1^j as

$$
w_1^j = (R_h - I)\bar{\partial}u(t_j) = \frac{1}{\tau}(R_h - I)\int\limits_{t_{j-1}}^{t_j} u_t ds = \tau^{-1}\int\limits_{t_{j-1}}^{t_j}(R_h - I)u_t ds,
$$

and estimate its L_2 norm by

$$\tau \sum_{l=1}^{j} ||w_1^l|| \leq \sum_{l=1}^{j} \int_{t_{j-1}}^{t_j} Ch^r ||u_t|| ds = Ch^r \int_0^{t_j} ||u_t|| ds.$$

Now, let consider the second term w_2

$$\begin{aligned} ||\tau w_2^j|| &= ||u(t_j) - u(t_{j-1}) - \tau \frac{u_t^j + u_t^{j-1}}{2}|| \\ &= || \int_{t_{j-1}}^{t_j} ((s-t_j)(s-t_{j-1}) + (s-t_{j-1/2})^2) u_{ttt}(s) ds|| \\ &\leq \tau^2 \int_{t_{j-1}}^{t_j} ||u_{ttt}|| ds. \end{aligned}$$

The third one, w_3^j, we bound by

$$w_3^j = \frac{S^j - S(t_j)}{2} + \frac{S^{j-1} - S(t_{j-1})}{2} \leq S_0 C \sum_{k=j-1}^{j} (x_F(t_k) - x_F^k), \qquad (3.71)$$

and because of (3.47), i.e.

$$|x_F(t_j) - x_F^j| \leq C\tau h^2, \qquad (3.72)$$

we get $\tau ||w_3^j|| \leq CS_0 \tau^2 h^2$. Here, the constant C depends on u, δ, x_0.
The combination of the above estimates gives

$$\begin{aligned} ||\theta^j|| &\leq ||\theta^0|| + Ch^r \int_0^{t_j} ||u_t|| ds + CS_0 \tau h^2 + \tau^2 \int_0^{t_j} ||u_{ttt}|| ds + \frac{D_0 C \tau h^2}{a^4} \\ &\leq ||\theta^0|| + C(h^r \int_0^{t_j} ||u_t|| ds + S_0 \tau h^2 + \tau^2 \int_0^{t_j} ||u_{ttt}|| ds + \tau h^2 \frac{D_0}{a^4}). \end{aligned}$$

Finally, we prove that

$$\begin{aligned} ||U^j - u(t_j)|| &\leq Ch^r ||u_0||_2 \\ &+ C \left(\tau^2 \int_0^{t_j} ||u_{ttt}|| ds + h^r \int_0^{t_j} ||u_t|| ds + \tau h^2 (S_0 + \frac{D_0}{a^4}) \right). \end{aligned}$$

$\diamondsuit$

Let us summarize, there is an order reduction in space due to the dependence of the diffusion coefficient on the gradient of the solution. This is valid for both linear and hermitian elements. Actually, this is not essential for hermitian elements since, for the particular problem we have, they can assure only second order of convergence - result of the insufficient smoothness of the solution $u \in H^{r=2}$ (the standard estimate for hermitian elements is fourth order).

Chapter 4

Numerical Results

We present some computational results to demonstrate that the theoretically derived error estimate, as well as the analyzed numerical technique are useful in the numerical practice. The first two subsections deal mainly with the adequacy and efficiency of the Lobatto estimate. The third subsection demonstrates the application of a non Front Tracking and the Front Tracking Technique to the anomalous heat transport in the plasma.

4.1 The adequacy of AIM

Before discussing the question whether the time integrator of AIM suggested by us is overheaded or not, we introduce the Incomplete Lobatto IIIA 4^{th} order method. The latter uses an approximate Jacobian in the Newton iteration. Since we consider implicit schemes we have to solve a general nonlinear algebraic system of equations. The Newton iteration for Lobatto IIIA 4^{th} order method is defined as:
Let G be the defect of Lobatto 4^{th} order (2.14). Then

$$G(X) = X - U_j - \frac{\tau}{6}(f(X) + f(U_j)) - \frac{2\tau}{3} f\left(\frac{X + U_j}{2} + \frac{\tau}{8}(f(U_j) - f(X))\right)$$

and the iteration procedure is

$$\begin{aligned} X^0 &= U_j \\ G'(X^k)(X^{k+1} - X^k) &= -G(X^k), k = 0, 1, .., n \\ U_{j+1} &= X^{n+1} \end{aligned}$$

with Jacobian matrix

$$G'(X) = I - \frac{\tau}{6} f'(X) - \frac{2\tau}{3} f'(\frac{X + U_j}{2} + \frac{\tau}{8}(f(U_j) - f(X)))(\frac{1}{2}I - \frac{\tau}{8} f'(X)).$$

For the *incomplete Lobatto IIIA 4^{th} order method* the Jacobian matrix is

$$G'(X) = I - \frac{\tau}{6} f'(X). \tag{4.1}$$

Now, we demonstrate the performance of the time integrator of AIM and the 4^{th} order Lobatto IIIA scheme - the full (abbr. by Lobatto) and the incomplete version (abbr by Incompl. Lobatto) on a set of problems. In the following list of examples the first two problems deal with blow-up solutions, the third one is an example of stiff problem taken from [19, pp. 2]. The forth example is from [24, pp. 26] - again an example for stiff systems. The last problem is the anomalous heat transport in plasmas.

In the comparisons are monitored:

AbsErr - the absolute error

RelErr - the relative error

NFE - the number of function evaluations

NJE - the number of Jacobian evaluations

NTS - the number of time steps.

Example 1.

$$\dot{u} = u^2 - u, \quad u(0) = 2, \quad 0 \leq t \leq 0.6$$

exact solution

$$u(t) = \frac{1}{1 - (1 - \frac{1}{u(0)})e^t}.$$

The results of the comparison, given in Table 4.1, show the advantage of the time integrator of AIM over the full 4^{th} order Lobatto IIIA scheme.

method	NFE	NJE	NTS	RelErr
AIM	107	49	44	1.67348e-4
Lobatto	309	239	69	1.28705e-4

Table 4.1: example 1

Example 2.

$$\dot{u} = u^2, \quad u(0) = 1, \quad 0 \leq t \leq 0.99$$

exact solution $u(t) = \frac{u(0)}{1-u(0)t}$.

The results of the comparison are given in Table 4.2. It is not a surprise that they show advantage of the time integrator of AIM over the 4^{th} order Lobatto due to the exactness of the approximate solution produced by the trapezoidal rule for this example [35],[15].

method	NFE	NJE	NTS	RelErr	Tol
AIM	273	112	87	1.3903e-14	1.e-4
Lobatto	847	655	183	3.3565e-3	1.e-9

Table 4.2: example 2

Example 3.

$$\dot{u} = -2000(u - cos(t)), \quad u(0) = 1, \quad 0 \leq t \leq 0.6$$

method	NFE	NJE	NTS	AbsErr
AIM	241	117	113	7.416e-4
Lobatto	287	203	60	7.385e-4

Table 4.3: example 3

The results of the comparison are given in Table 4.3. The time integrator of AIM is only just better than the 4^{th} order Lobatto IIIA method.

Example 4.

$$\begin{aligned} \dot{u}_1 &= -k_1 u_1 + k_2 u_2, \quad u_1(0) = 0.1 \\ \dot{u}_2 &= k_1 u_1 - k_2 u_2, \quad u_2(0) = 0.9 \end{aligned}$$

where $k_1 = 1,\ k_2 = 100$. *The exact solution fulfils*

$$\begin{aligned} u_1 &= \frac{k_2}{k_1 + k_2}(u_1(0) + u_2(0)) + \frac{e^{-(k_1+k_2)t}}{k_1 + k_2}(k_1 u_1(0) - k_2 u_2(0)) \\ u_2 &= \frac{k_1}{k_1 + k_2}(u_1(0) + u_2(0)) - \frac{e^{-(k_1+k_2)t}}{k_1 + k_2}(k_1 u_1(0) - k_2 u_2(0)) \end{aligned}$$

method	NFE	NJE	NTS	AbsErr
AIM	363	180	103	2.127e-3
Lobatto	201	143	55	2.387e-3
Incompl. Lobatto	87	44	21	2.650e-3

Table 4.4: example 4

The results of the comparison are given in Table 4.4. Apparently the 4^{th} order Lobatto IIIA gives better results than AIM, but the incomplete Lobatto 4^{th} has drastic advantage.

Example 5. *the anomalous heat transport with parameters:*

$$\textit{grid points } = 100, \ \textit{linear finite elements}$$
$$D_0 = 1, \ D_1 = 10, \ \bar{u} = 1, \ S_0 = 3, \ x_0 = \delta^2 = 0.5, \ Tol = 1.e - 6.$$

The results are listed in Table 4.5

method	NFE	NJE	NTS	cpu time
AIM	16410	5450	2472	1m8.529s
Incompl. Lobatto	29366	9376	4271	2m7.093s

Table 4.5: example 5

From the above examples it is clear that for some problems the time integrator of AIM is better than Lobatto IIIA, 4^{th} order but not always. Particularly for the anomalous transport problem it shows an advantage (AIM is approximately 1.7 times better than the incomplete 4^{th} order Lobatto IIIA method).

4.2 Quality of the Lobatto Estimator

In order to illustrate the accuracy of the Lobatto estimate, we consider two examples. The first one is a system of ODEs, which has a positive one-side Lipschitz constant C (defined by (2.16), Ch 2). The second example is an example of a parabolic differential equation modelling heat transport in the magnetically confined plasma of thermonuclear fusion experiments.

Example 6.

We consider the following system of ODEs

$$\begin{aligned} \frac{d}{dt}\begin{pmatrix} v(t) \\ w(t) \end{pmatrix} &= \begin{pmatrix} -6 & 1 \\ & \\ 8 & 4 \end{pmatrix}\begin{pmatrix} v(t) \\ w(t) \end{pmatrix}, \quad t \in (0,1] \\ (v(0), w(0))^t &= (0.5, 0.5)^t. \end{aligned}$$

Since this is a linear system of ODEs, the exact solution $u := (v, w)$ is known. By $\{U_j = (V_j, W_j)\}_j$ we denote the approximate solution, obtained by AIM and by τ the time step. The constants C and C_{max}, defined by (2.16) and (2.17) respectively, are evaluated and as a result the values $C = 5.726812$, $C_{max} = 10.3668669$ are assessed. The quality of the estimators is usually [26] measured by the effectivity index Θ defined as

$$\Theta = \frac{\text{Estimated Error}}{\text{Exact Error}}. \tag{4.2}$$

At first, we demonstrate the efficiency of the defect of the Lobatto III A scheme L_j, given by (2.14), to approximate the local error of trapezoidal rule given by

$$T_j^{tr} := (T_{1,j}^{tr}, T_{2,j}^{tr}) = (V_j - v(t_j), W_j - w(t_j)).$$

Table 4.6 shows the exact relative error computed at $t = \tau$

$$RelErr(\tau) = (\frac{|V(\tau) - v(\tau)|}{|v(\tau)|}, \frac{|W(\tau) - w(\tau)|}{|w(\tau)|})$$

and the effectivity index computed at $t = \tau$

$$\Theta = (\frac{L_0(\tau, v(0), V_1)}{|T_{1,1}^{tr}|}, \frac{L_0(\tau, w(0), W_1)}{|T_{2,1}^{tr}|}).$$

The results indicate that the defect of the Lobatto III A scheme (2.14) produces excellent approximation of the local error of TR.

$t = \tau$	RelErr		Θ	
	$v(t)$	$w(t)$	$v(t)$	$w(t)$
0.01	8.90535e-4	8.90746e-4	1.038	1.005
0.005	2.22503e-4	2.22555e-4	1.019	1.002
0.0025	5.56176e-5	5.56307e-5	1.009	1.001

Table 4.6: The Relative Error and the effectivity index

As a second step, we compare the Lobatto estimate (2.20) to the Kraaijevanger estimate for the system of ODEs considered above.
Kraaijevanger derived an estimate of the global error obtained from the discretization of an ODE through the trapezoidal rule. The exact formulation of this result is given for instance in [19, pp. 25]. Here, we rewrite it using our notations.

The global error, $e_{j+1} = (e_{1,j+1}, e_{2,j+1}) : \quad ||e_{j+1}|| = ||U_{j+1} - u_{j+1}||$, of the trapezoidal rule with fixed time step τ permits for $\tau \cdot C \le \alpha < 2$ the Kraaijevanger estimate

$$||e_{j+1}|| \le \begin{cases} \frac{1}{6(2-C\tau)} \max\limits_{t\in[t_0,t_{j+1}]} ||u^{(3)}(t)||\ (j+1)\tau^3 = \frac{j+1}{1-\frac{C\tau}{2}} \max\limits_{0\le k\le j+1} ||T_k^{tr}||, \\ \hfill C \le 0 \\ \frac{e^{2CT}}{12} \max\limits_{t\in[t_0,t_{j+1}]} ||u^{(3)}(t)||\ (j+1)\tau^3 = (j+1)e^{2CT} \max\limits_{0\le k\le j+1} ||T_k^{tr}||, \\ \hfill C > 0. \end{cases}$$

As is shown, Theorem 3, the Lobatto estimate is:

$$||e_{j+1}|| \le \begin{cases} \frac{(1-C\tau)^{j+1}-1}{-C\tau(1-C\tau)^j} d(\tau) \le (j+1)d(\tau), & \text{no restriction on } \tau,\ C \le 0 \\ (j+1)\frac{e^{K_1T}}{\sqrt{1-\frac{(C_{max}\tau)^2}{6}-\frac{\tau}{2}C_{max}}} d(\tau), & \text{for } \tau C_{max} \le \sqrt{\frac{12}{5}},\ C > 0 \end{cases}$$

$d(\tau) = \max\limits_{0\leq k\leq j+1} ||L(\tau, u_{k+1}, u_k) - L(\tau, U_{k+1}, U_k)|| = \max\limits_{0\leq k\leq j+1} ||T_k^{tr}(1 - \frac{\tau f'(U_k)}{2}) + \mathcal{O}(\tau^5)||,$
$K_1 = \frac{C}{1 - C\tau - \frac{(C_{max}\tau)^2}{6}}$

Here, we list the results we obtained for the Lobatto and the Kraaijevanger estimates for the fixed time step $\tau = 0.001$.

- Results for the Kraaijevanger estimate
  - restriction on the step size: $\tau \leq \frac{2}{C} = 0.3492$,
  - exact truncation error: $\max\limits_{0\leq j\leq m} ||T_j^{tr}|| = 8.3471e-7$
  - estimate of the error of approximation:

$$\begin{aligned}(e_{1,m}, e_{2,m}) &\leq (j+1)e^{2C}\ (\max_j |T_{1,j}^{tr}|, \max_j |T_{2,j}^{tr}|) \\ &= (j+1)\ (0.0011, 0.0787)\end{aligned}$$

- Results for the Lobatto estimate
  - restriction on the step size: $\tau \leq \sqrt{\frac{12}{5C_{max}^2}} = 0.14943$
  - defect of Lobatto: $\max\limits_{0\leq j\leq m} ||L(\tau, U_{j+1}, U_j)|| = 8.3628e-7$
  - estimate of the error of approximation: $K_1 = 5.7599$

$$\begin{aligned}(e_{1,m}, e_{2,m}) &\leq (j+1)1.01e^{K_1}(\max_j |L(\tau, V_j, V_{j+1})|, \max_j |L(\tau, W_j, W_{j+1})|) \\ &= (j+1)\ (2.47e-5, 2.65e-4)\end{aligned}$$

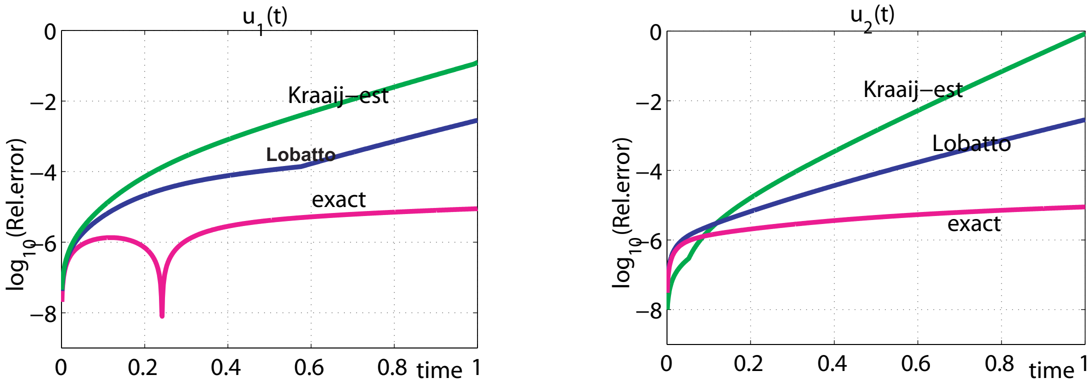

Figure 4.1: Comparison of the Kraaijevanger and the Lobatto estimates for u_1 - left and u_2- right: error in log_{10} -scale vs. t, red exact error, blue Lobatto estimate, and green Kraaijevanger estimate

As it could be seen from the comparison, as well as from Fig. 4.1, both estimates bound the exact error for most t. Apparently, the Lobatto estimate gives rigorous results. At the same time, it imposes stronger restriction on the time step than the Kraaijevanger estimate. However, the Lobatto estimate is always applicable, due to its dependence only on the defect of the Lobatto III A scheme.

Let illustrate through an example the Lobatto estimate in the case of parabolic equation.

Example 7. *The application of the Lobatto estimate (2.19) to a parabolic equation*

In the basic fluid model of plasma physics, the energy transport and continuity equations have the form of generalized reaction-diffusion equations, with non-linear source and loss terms. Similar to those equations occur in many fields of science, in physics as well as in chemistry or biology. The essential physical processes of diffusion and reaction occurring simultaneously lead to special solutions, which do not exist if only one kind of process contributes to the dynamics of the system. Particularly interesting phenomena occur when the non-linear reaction term drives solutions of large amplitude. In such situation the solution may exhibit an exploding instability.

Here, we consider a model describing heat transport in the magnetically confined plasma of thermonuclear fusion experiments:

$$\begin{aligned}
\frac{\partial u}{\partial t} &= \frac{\partial}{\partial x}(u^\sigma u_x) + u^\beta \; x \in \mathbb{R}, \; t > 0 \\
u(0,x) &= u_0(x), \; x \in \mathbb{R} \\
u^\sigma(t,0)u_x(t,0) &= 0, \; u(t,R) = 0 \; t \geq 0 \qquad (4.3)
\end{aligned}$$

Here, $u(t,x)$ is the particle temperature. For a typical fusion plasma, the value of the parameter $\sigma = 3/2$ represents diffusion caused by drift wave turbulence [52]. The parameter $\beta = 5/2$ models the effects of alpha-particle heating of a fusion plasma. Similar problems are considered in [28],[53],[8]. The solution of this problem possesses several interesting properties: blow-up in a finite time T_0, localization in the space, singularities of the moving boundaries [41],[9].

As is known for $\sigma > 0$ and $\beta > 1$, Eq. (4.3) has unbounded self-similar solution $u(t,x) = (1 - \frac{t}{T_0})^{-1/(\beta-1)}\theta(\xi)$ where $\xi = \frac{x}{(1-t/T_0)^m}$, $m = \frac{\beta-\sigma-1}{2(\beta-1)}$. Here, T_0 is the blow-up time and the function $\theta(\xi)$ is such that u satisfies (4.3). In the case of $\sigma + 1 = \beta$, the exact solution is known [9]

$$\theta(\xi) = \begin{cases} \left(\frac{2\beta}{\beta+1} cos^2 \frac{\pi\xi}{2R}\right)^{1/\sigma} & |\xi| \leq R = \frac{\pi\beta^{1/2}}{\sigma} \\ 0 & elsewhere. \end{cases}$$

For this particular case, the solution is localized in the interval $[0, R]$ and each point of the solution tends to infinity in a finite time (called regional blow up).

As we mentioned, we consider the case of $\sigma = 3/2$, $\beta = 5/2$. Here, the blow up time is $T_0 = 1/(\beta - 1) = 0.666...$, and the interval of localization is $[0, 3.31153]$. The initial data are chosen from the exact solution. The discretization is done using AIM - linear FEM combined with the modified trapezoidal rule. Because of the localization of the solution and its explosion at each point, no spatial adaptation is needed. The computation of the solution is performed until the time step becomes less than $1.e - 16$. The reached time, $\tilde{T_0}$, is considered as an approximation of the blow-up time.

The evolution of the solution in logarithmic scale is shown on the l.h.s of fig. 4.2. In fig. 4.2, on the r.h.s, are depicted the absolute error $AbsErr$, the estimate

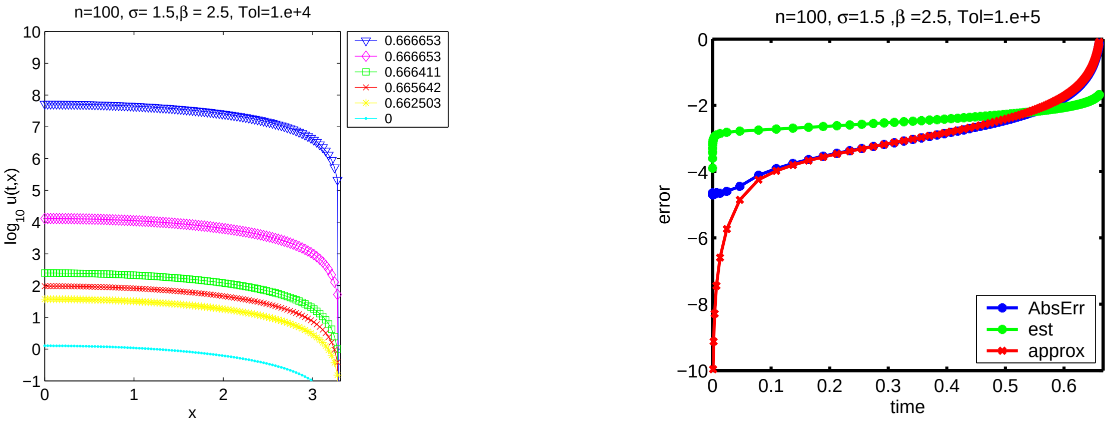

Figure 4.2: The evolution of the solution in logarithmic scale (left); estimation of the absolute error in logarithmic scale (right)

est of the absolute error and an approximation of it, $approx$. The approximation $||approx_{j+1}||$ at the $j + 1$th time step is computed as

$$\begin{aligned}\left(I - \frac{\tau}{2}\mathcal{D}f(U_{j+1})\right)\ approx_{j+1} &= -(L_j(\tau, U_j, U_{j+1}) - \tau h^2 Const) \\ &\quad + \left(I + \frac{\tau}{2}\mathcal{D}f(U_j)\right) approx_j\end{aligned}$$

(see Ch.2: Theorem 2 and “Incorporation of the spatial error into the defect of Lobatto III A scheme”), whereas the error is estimated by

$$||AbsErr_{j+1}|| \le est := (j + 1)Tol\sqrt{n}. \tag{4.4}$$

Here, Tol is a prescribed tolerance, n is the number of the grid points and j is the number of the current time step. Let us note, that for this problem the constant C (from the Lobatto estimate) is positive (the derivative of the source is positive and $\sigma + 1 = \beta$). Therefore, est in (4.4) may not give the correct growth of the error. The correct bound of the error is given in Theorem 3, b), but its computation is more complicated than est. Figure 4.2, on the r.h.s, shows that

(4.4) justifies to be used for times away from the blow-up time ($T_0 = 0.666...$). On the contrary, the approximation *approx* follows the error for times quite close to the blow-up time. As in the previous example, we test the quality of the defect of the Lobatto III A scheme (2.14) as an estimate of the local error by monitoring the effectivity index Θ, defined in (4.2). The results are listed in Table 4.7.

τ	n	Θ
0.0319	100	1.07
0.0240	200	1.01
0.00767	700	1.00

Table 4.7: The Effectivity Index

As in example 6, the defect of the Lobatto III A scheme (2.14) gives an excellent estimate of the local error.

To examine the order of convergence, we have computed the solution $u(t, x)$ on three embedded grids with respectively n, $2n$ and $4n$ points, for $n = 25$. In Table 4.8 are presented the exact L_2 relative error $RelErr$ at two different times, together with the numerical order of convergence q_{num} and the numerical blow up time $\tilde{T_0}$. For the numerical order of convergence is used the following formula

n	$RelErr(0.36)$,	q_{num}	$RelErr(0.66)$,	q_{num}	$\tilde{T_0}$
25	4.4449e-4,	-	7.8420e-3,	-	0.666586
50	1.1590e-4,	1.93	2.4164e-3,	1.69	0.666633
100	2.9757e-5,	1.96	6.5465e-4,	1.88	0.666659

Table 4.8: The relative error and the approximation of the blow-up time

$$q_{num} = \log_2 \frac{||U\_AbsErr||_n}{||U\_AbsErr||_{2n}}. \tag{4.5}$$

The relative error is computed as

$$RelErr(t) = \frac{||u(t,.) - u_{ex}(t,.)||}{||u_{ex}(t,.)||}.$$

The theoretically expected order of convergence is 2. One can see that it decreases approaching the blow-up time. However, the blow-up time $T_0 = 0.666...$ is approximated well - between $99.97\% - 99.99\%$.

4.3 Numerical results for the anomalous transport problem

Here in this section, we present the numerical results from the application of the Front Tracking Technique to the anomalous heat transport problem.

The implemented mathematical model is

$$P : \begin{cases} \frac{\partial u}{\partial t} & = & \frac{1}{x^{d-1}}\frac{\partial}{\partial x}(x^{d-1}(D_0 + D_1 H(|u_x| - \bar{u})(|u_x| - \bar{u}))\frac{\partial u}{\partial x}) + S(x), \\ & & \text{for } (t,x) \in (0,T) \times (0,1) \\ u(0,x) & = & u_0(x), \quad x \in [0,1] \\ u_x(t,0) & = & 0, \quad t \in [0,T] \\ u(t,1) & = & 0, \quad t \in [0,T] \end{cases}$$

where D_0, D_1 are positive constants; $\bar{u}$ is the threshold parameter for the gradient $|u_x|$; $d = 1, 2, 3$ and $S(x)$ is the source. The considered source has the form $S(x) = S_0 \exp(-\frac{(x-x_0)^2}{\delta^2})$, where $0 < x_0 < 1$, δ and S_0 are positive constants describing, respectively, the position of the peak, the width and the strength of the source.

The front tracking technique used consists of separation of the problem P into subproblems at the point(s) where the gradient reaches the critical values $\pm\bar{u}$, application of the AIM algorithm to each of the subproblems, and tracking the front point(s). The algorithm AIM uses the method of lines, that is, the FEM - for the space discretisation and the modified trapezoidal rule - for the time discretisation.

The numerical investigation proceeds as follows:

- Verification of the accuracy of the methods on a problem with exact solution and investigation of the order of convergence of the proposed methods using nested meshes;

- Examination of the parameter space. Investigation of the solution approaching stationary state.

4.3.1 Verification of the accuracy of the methods

Regarding the accuracy and the convergence of the numerical method applied, we monitor several characteristics. We keep track of the absolute/relative error of the numerical solution (which is only available for problems with analytical solutions), of the absolute/relative error of the gradient of the solution, and the numerically observed temporal order of convergence defined by (4.5). From physical point of view one of the most important features of the approximate solution is that it has to conserve the energy of the underlying problem. To measure this property of the numerical solution, we define I_Error, in case of $d = 1$, as

$$I\_Error := \int_{x_i}^{x_k} ((t_n - t_{n-1})S(\xi) - u(t_n, \xi) + u(t_{n-1}, \xi))d\xi + (t_n - t_{n-1})(q(x_k) - q(x_i)).$$

It can be obtained by integrating the problem P with respect to x and t in the intervals $[x_i, x_k]$ and $[t_{n-1}, t_n]$, respectively and by subtracting the r.h.s. from the l.h.s. of the obtained expression. The both ends of the interval $[x_i, x_k]$ are grid points. One can consider the integral in the whole interval $[0, 1]$ provided that the solution does not have irregularities. Because of the front point(s), we calculate I_Error between the front point and each end of the interval $[0, 1]$, or in case of multiple front points also in between the front points. This estimate of the error is essential to judge the quality of the approximation, especially for the cases when the exact solution is unknown.

As was shown in Ch.3, some exact solutions can be found in the case of $S(x) \equiv 0, \quad D_0 = 1$. That is, we consider the following problem

$$\begin{aligned} u_t &= \partial_x((1 + D_1 H(|u_x| - \bar{u})(|u_x| - \bar{u}))u_x), \quad 0 < x < 1, \ t > 0, \quad (4.6) \\ u_x(t, 0) &= u_1(t), \quad u(t, 1) = u_2(t), \\ u(0, x) &= u_0(x). \end{aligned}$$

Particularly, we make use of the exact solution (3.20) given in Lemma 7, with values for the free parameters $A = 2, \ C = 1, \ \bar{u} = 3$ and for the beginning $D_1 = 1$. The boundary and the initial conditions are taken from the exact solution

$$u(t, x) = \begin{cases} \frac{1+\bar{u}}{4} \exp^{2x+4t+4-2\bar{u}} + x\frac{(\bar{u}-1)}{2} + \frac{5\bar{u}-13}{4}, & u_x \leq \bar{u}, \\ \frac{x^2}{2} + 2t^2 + 2tx + 2x + 5t - \bar{u}t, & u_x \geq \bar{u}. \end{cases}$$

We have applied, in addition to the FTT, the algorithm AIM on the whole problem P, without making use of the front tracking idea. We refer to this strategy as AIM or as non FTT.

In Fig.4.3, on the l.h.s, is depicted the solution at three different times $t = 0.04, \ t = 0.25,$ and $t = 0.4525$. The magenta curve corresponds to the solution of P1, whereas the blue one - to the solution of P2. The plot on the right hand side shows the behaviour of the corresponding gradient.

- *Investigation of the order of convergence*

  To investigate the order of convergence, we computed the solution over three embedded meshes with $n = 50, \ 100, \ 200$ grid points, at $t = 0.25$. In tables 4.9 and 4.10 are given the values of:

  – the absolute error U_AbsErr ,

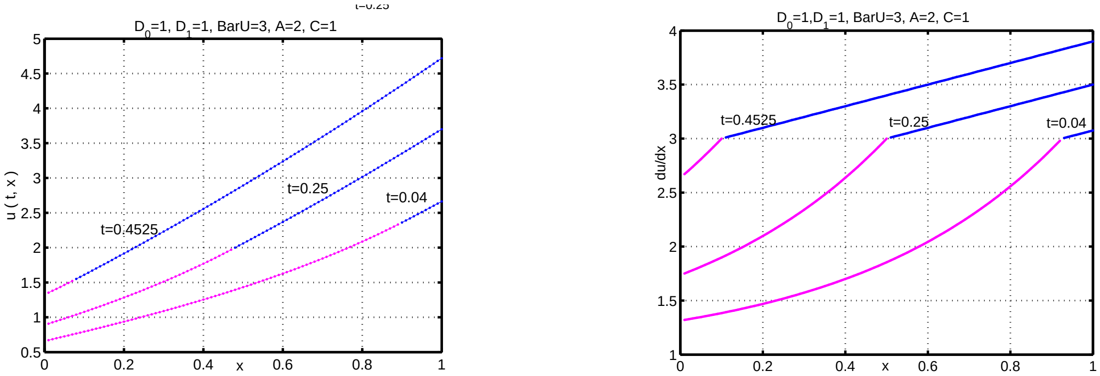

Figure 4.3: The exact solution (left) and its gradient (right) for time $0.04, 0.25, 0.45$

- the absolute error for the gradient of the solution U_x_AbsErr,
- the numerical order of convergence q_{num},

for the non FTT - AIM (table 4.9) and FTT (table4.10) computed with linear FEM.

n	I_Error		U_AbsErr	U_x_AbsErr	q_{num}
	$[0, x_F]$	$[x_F, 1]$			
50	1.6865e-05	6.9977e-06	3.1010e-05	3.9179e-02	
100	3.3960e-06	1.4625e-06	8.6241e-06	1.9815e-02	1.846
200	6.7359e-06	6.9742e-07	2.5450e-06	9.9789e-03	1.760

Table 4.9: AIM, linear FEs

n	x_F_Err	I_Error		U_AbsErr	U_x_AbsErr	q_{num}
		$[0, x_F]$	$[x_F, 1]$			
50	1.4295e-04	4.1731e-06	1.4425e-06	1.8514e-04	3.7441e-02	
100	2.2988e-05	2.8878e-06	4.4175e-07	3.7930e-05	1.9365e-02	2.287
200	5.8467e-06	2.9104e-06	8.4310e-07	7.1166e-06	9.8398e-03	2.414

Table 4.10: FTT, linear FEs

For AIM, the order of convergence is less than two (≈ 1.8), whereas the front tracking technique (FTT) shows order even higher than two.

The order reduction for FTT, due to the dependence of the diffusion coefficient on the derivative of the solution, as predicted in Ch. 3, is not observed for this concrete example.

To complete the accuracy analysis, we give the error in the computation of the position of the front, x_F_Err, in table 4.10, second column. The exact position of the front satisfies the equation $x_F(t) = 1 - 2t$. It could be observed second order of convergence.

- *Energy conservation*

  For the front tracking technique, I_Error is computed separately in each of the sub-intervals $[0, x_F(t)]$ and $[x_F(t), 1]$, and the results are given in Table 4.10. In order to make fair comparison between the AIM and FTT, the value I_Error obtained using the AIM is computed also separately at each side of the front. The values of I_Error are not satisfactory small for both techniques, i.e. they do not fully conserve the energy of the problem.

  In order to conserve the energy to higher accuracy, we use Hermitian finite elements instead of linear. Again, the solution is computed onto nested meshes with $n = 50,\ 100,\ 200$. The absolute error for U and U_x, together with I_Error are listed in tables 4.11, 4.12.

n	I_Err		U_AbsErr	U_x_AbsErr	q_{num}
	$[0, x_F]$	$[x_F, 1]$			
50	1.6913e-08	6.7423e-09	2.4059e-04	4.9327e-03	
100	1.9578e-08	8.8236e-09	5.1605e-05	2.0267e-03	2.221
200	2.1746e-08	1.0035e-08	1.6749e-05	2.0679e-03	1.623

Table 4.11: AIM, hermitian FEs

n	x_F_Err	I_Err		U_AbsErr	U_x_AbsErr	q_{num}
		$[0, x_F]$	$[x_F, 1]$			
50	2.0901e-5	2.0625e-08	7.9174e-09	9.0466e-05	2.2439e-04	
100	4.4649e-6	9.2928e-09	3.4329e-09	2.4110e-05	6.3776e-05	1.9077
200	9.951e-07	2.6817e-10	1.0273e-10	6.4572e-06	1.8045e-05	1.9007

Table 4.12: FTT, hermitian FEs

  Apparently, the Hermitian elements do not increase the order of convergence neither for AIM nor for the front tracking technique. This is a consequence of the discontinuity of the second derivative of the solution u at the interface. However, the energy is conserved with two magnitudes better. From now on, in the computations is used the FTT with Hermitian Finite Elements.

- *Front Tracking against non Front Tracking Technique*

  In Fig.4.4 are presented U_AbsErr and U_x_AbsErr for FTT and non FTT, respectively. Clearly the error for the non FTT is mostly due to the front

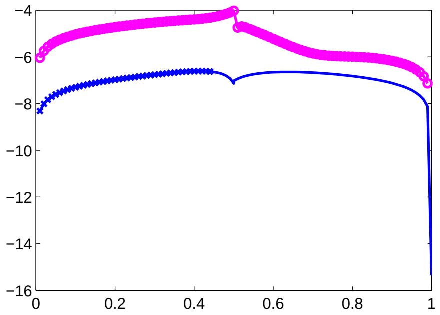

Figure 4.4: The Absolute error for u and u_x for FTT and non FTT

point. The use of the front tracking strategy reduces the error of approximation for both the solution and the gradient of the solution.

In Table 4.13 we give the results from the comparison of the front tracking and the non front tracking technique for hermitian elements. In the comparison the numerical error is kept fixed and the time needed for the numerical computation (the cpu-time) and the number of the grid points n are monitored. The numerical error, in this case, is computed in H_2 norm, i.e.

$$Err = ||u_{app} - u_{ex}||_{L_2} + ||(u_x)_{app} - (u_x)_{ex}||_{L_2}.$$

Method	Err	n	cpu(s)
AIM, her	0.00537	150	284
FTT, her	0.00592	100	134

Table 4.13: AIM, hermitian FEs

The computations are performed on a IBM-SP system, which runs under AIX5, contains 6 Power4 CPUs, 1.3GHz and 12GB of memory.

The comparison between the FTT and non FTT shows that the non FTT needs one and a half times more grid points and two times more cpu-time than FTT.

- *Smoothing out the conductivity coefficient*
  Here we consider a modification of the problem (4.6), where the conductivity coefficient is replaced by a smoother function. This is achieved by approximating the Heaviside function by *arctan* depending on a small parameter ε, that is

$$H_\varepsilon := (\frac{1}{2} + \frac{1}{\pi} arctan \frac{|u_x| - \bar{u}}{\varepsilon}) \overset{\varepsilon \to 0}{\to} H(|u_x| - \bar{u}).$$

We have solved the problem (4.6), with $H_\varepsilon(.)$ instead of $H(.)$, using AIM for values of the parameters $D_1 = 1,\ A = 2,\ C = 1, \bar{u} = 3,\ n = 100$ and varying the value of $\varepsilon \in [1.e-7, 1.e-2]$ - see Fig.4.5. Apparently, small

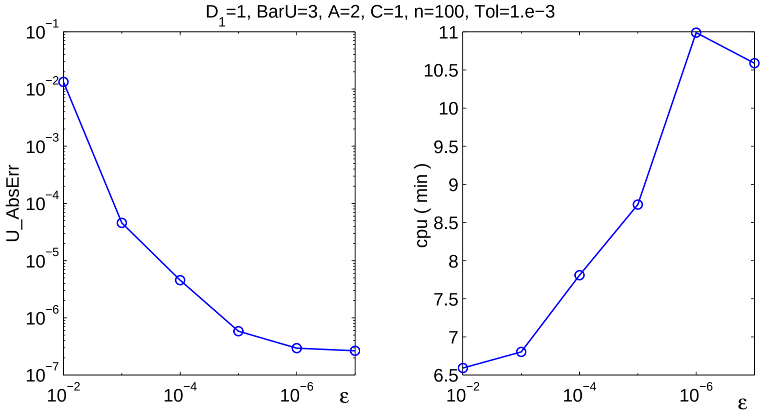

Figure 4.5: Smoothing out the conductivity coefficient

ε gives better approximation of the solution of (4.6). However, there is no significant improvement in the accuracy for $\varepsilon \leq 10^{-5}$. Not surprisingly the cpu - time increases whenever the value of ε decreases- Fig.4.5, r.h.s.

We compare the performance of the FTT and AIM. The latter is applied to problem (4.6) with the smoothed conductivity coefficient. Variation of the value of D_1 is carried out. In the computation it is used space step $h = 0.01$ for $t \leq 0.25$ and again the parameters A and C are set to $A = 2,\ C = 1$. Since the value of D_1 changes, the value of the critical gradient and the initial time t_0 change as well - see Table 4.14.

D_1	$\bar{u}$	t_0
1	3	0.04
10	7.5	0.229
20	12.5	0.2395

Table 4.14: The values of D_1 and $\bar{u}$

We monitor the cpu-time of each of the algorithms used to compute the solution for a certain absolute error. We give, in Table 4.15, for $D_1 = 10,\ 20$ an interval for the absolute error in u and u_x, as well as for the corresponding cpu-time (measured in seconds).

Graphically we presented the results from the comparison in Fig. 4.6. FTT is much faster than the non front tracking technique. Nevertheless, it is worth to mention that the non FTT, applied to the smoothed problem for the case $D_1 = 1$, gives quite satisfactory results for surprisingly large value of Tol, as large as $Tol = 0.1$.

Method	D_1	u_AbsErr	du_AbsErr	cpu(s)
AIM	1	4.5660e-5	2.6523e-4	408
FTT	1	6.5592e-5	2.4929e-4	77
AIM	10	[2.11e-5, 2.42e-4]	[2.56e-3, 2.81e-3]	[2940, 3900]
FTT	10	[2.42e-5, 3.05e-4]	[4.01e-4, 3.32e-2]	[251, 468]
AIM	20	[8.36e-5, 1.57e-4]	[2.02e-2, 8.59e-2]	[900, 50400]
FTT	20	4.5631e-5	2.9823e-3	870

Table 4.15:

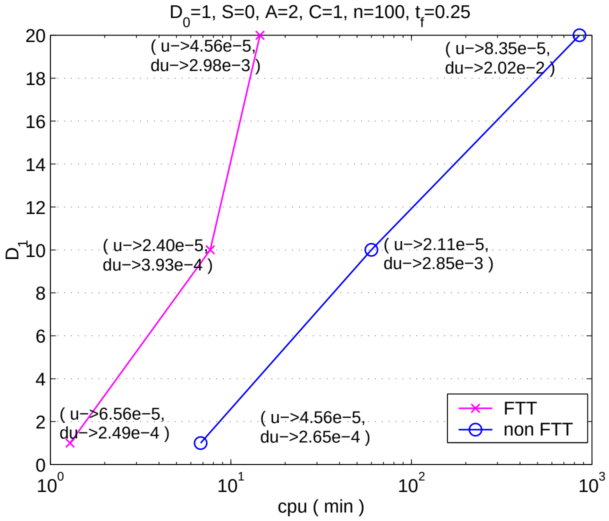

Figure 4.6: Smoothing out the conductivity coefficient

4.3.2 Examination of the parameter space

In the center of our attention is the problem P, with values of the parameters that are meaningful for the anomalous heat transport in the plasma. At first, we consider, in more details, the case $d = 1$.

- *Behaviour of the temperature as a function of D_1*

We investigate the behaviour of the temperature as a function of D_1, especially at $D_1 \gg 1$. First we do this analytically and then numerically. Let us look at the integral form of the problem P2 and write it as

$$|u_x|u_x + (\frac{D_0}{D_1} - \bar{u})u_x + \frac{1}{D_1}s = 0,$$

where $s = \int\limits_0^x S(\xi) - u_t(t,\xi)d\xi$. Now, D_1 going to ∞ leads to $|u_x| \to \bar{u}$. This behaviour of the analytic solution is mimicked by the numerical solution.

In Fig. 4.7, on the left hand side, is depicted the temperature, $u(0.088, x)$, for different values of D_1 - between 1 and 100. In the same figure, on the right hand side, is presented the corresponding temperature gradient. The values for the other parameters are

$$\bar{u} = 1.28,\ D_0 = 1,\ S_0 = 3,\ x_0 = 0.5,\ \delta = \sqrt{0.5},\ u_0(x) = \cos(x\frac{\pi}{2}) \tag{4.7}$$

The growth of D_1 keeps the gradient almost constant and brings it close to the critical value, $\bar{u}$.

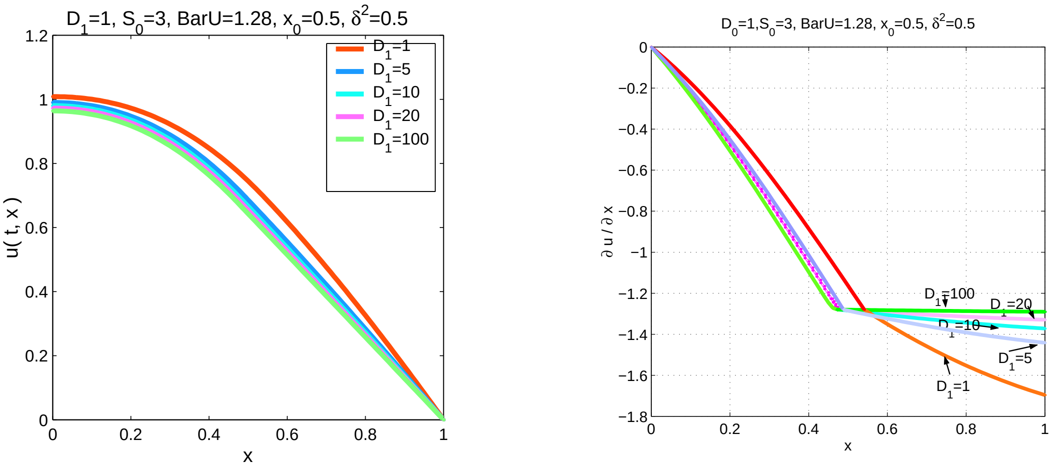

Figure 4.7: The behaviour of the solution (left) and its gradient (right) varying D_1

- *Approaching Steady State*

Let us monitor the process of approaching the stationary state ($u_t = 0$). From the subproblem P1 we get the following expression for the gradient

$$u_x(t, x) = -\frac{1}{D_0} \int\limits_0^x S(\xi) d\xi, \quad |u_x| \leq \bar{u}. \tag{4.8}$$

Since we consider non negative source, $S(x) \geq 0$, from (4.8) follows that $u_x \leq 0$ at the steady state.

For the subproblem P2, i.e. the case $-u_x \geq \bar{u}$, we obtain a quadratic equation

$$-D_1 u_x^2 + (D_0 - D_1 \bar{u}) u_x + \int\limits_0^x S(\xi) d\xi = 0. \tag{4.9}$$

The non positive solution of (4.9) is

$$u_x = \frac{(D_0 - D_1\bar{u}) - \sqrt{(D_0 - D_1\bar{u})^2 + 4D_1 \int\limits_0^x S(\xi)d\xi}}{2D_1}. \tag{4.10}$$

From $-u_x \geq \bar{u}$ we obtain a *condition for existence of a front point at the stationary state*

$$\bar{u} \leq \frac{1}{D_0} \int\limits_0^{x_F} S(\xi)d\xi \leq \frac{1}{D_0} \int\limits_0^{1} S(\xi)d\xi.$$

The fact that $u_x \leq 0$ at the stationary state assures that the temperature, at large t at least, behaves monotonically provided that the initial condition is a monotonic function.

The stationary position of the front point can be obtained from (4.8), for $x = x_F$, i.e.

$$D_0\bar{u} = \int\limits_0^{x_F} S(\xi)d\xi. \tag{4.11}$$

Let us suppose that there are more than one front point then

$$D_0\bar{u} = \int\limits_0^{x_F^{II}} S(\xi)d\xi = \overbrace{\int\limits_0^{x_F^{I}} S(\xi)d\xi}^{D_0\bar{u}} + \int\limits_{x_F^{I}}^{x_F^{II}} S(\xi)d\xi,$$

therefore

$$\int\limits_{x_F^{I}}^{x_F^{II}} S(\xi)d\xi = 0.$$

This means that either $S(x) \equiv 0$ for $x \in [x_F^I, x_F^{II}]$ or the interval $x \in [x_F^I, x_F^{II}]$ has zero length. The source, considered by us, is $S(x) = S_0 \exp(-(x - x_0)^2/\delta^2) > 0$ for $x \neq x_0$, therefore in the stationary state only one front point is possible.

Apparently, the value of D_1 does not have influence on the stationary state of the interface. On the other hand, D_1 affects the speed of reaching the stationary state of x_F. The larger the value of D_1, the faster x_F goes to its stationary state. These two features of the solution could be seen in Fig. 4.8, where the motion of the front is presented. The left plot depicts the

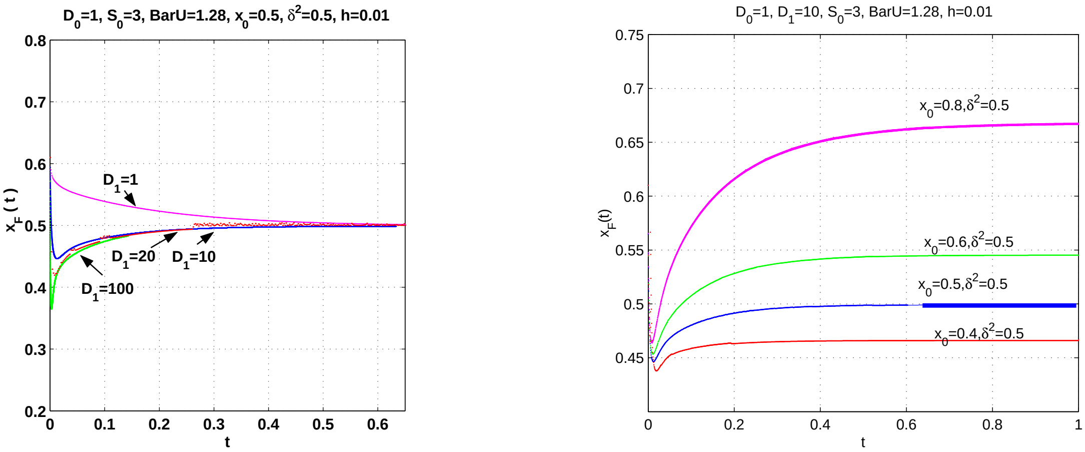

Figure 4.8: The evolution of the front point, varying D_1 - on the left; varying x_0 - on the right

interface evolution for different values of D_1. The right one presents the stationary states of the front obtained varying the values of x_0. Variation in x_0 changes the stationary state of the front point.

One can use the analytic forms for the gradient at the stationary state, (4.8) and (4.10), to verify the correctness of the numerical evolution of the temperature gradient. Such a comparison is performed in Fig. 4.9, where the high accuracy of the FTT shows up. The values of the parameters D_0, $\bar{u}$, S_0, x_0, δ are given by (4.7), and $D_1 = 10$.

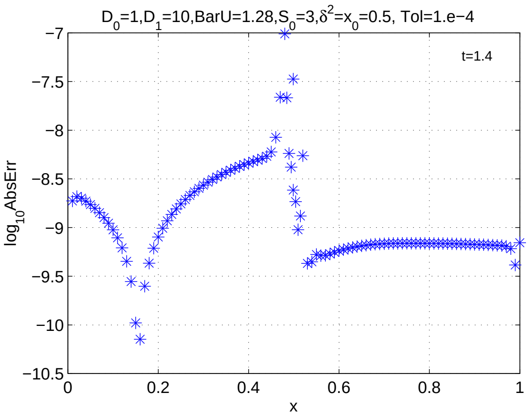

Figure 4.9: The error of the solution's gradient at the stationary state

- *Pulse Source*
  The Pulse Source Method is a powerful method used in the plasma experiments to determine the value of the parameters. The idea is to perturb

the stationary temperature and to measure the differences in the temperature behaviour. This is modelled by an additional time-dependent source $\phi(t)S_1\exp(-\frac{(x-x_1)^2}{\delta_1^2})$ introduced in problem P. The solution of P at the stationary state is taken as an initial condition. The considered model is

$$\begin{aligned}\frac{\partial u}{\partial t} &= \frac{1}{x^{d-1}}\frac{\partial}{\partial x}(x^{d-1}(D_0+D_1H(|u_x|-\bar{u})(|u_x|-\bar{u}))\frac{\partial u}{\partial x})\\ &\quad +S_0\exp(-\frac{(x-x_0)^2}{\delta^2})+\phi(t)S_1\exp(-\frac{(x-x_1)^2}{\delta_1^2}),\\ u(0,x) &= \text{stationary solution},\\ u_x(t,0) &= 0,\ u(t,1)=0.\end{aligned}$$

In Fig. 4.10, on the left hand side, is presented the evolution of the position of the front point. We have chosen the strength of the pulse source to be the same as in the non-time dependent source, i.e. $S_1 = S_0$, as well as $x_1 = x_0$, $\delta_1 = \delta$. The function reflecting the time dependence, ϕ, is taken to be the nonnegative function $\phi(t) = \max(0, \cos(t\pi\frac{40}{3S_0}))$ (in fig.4.10 the blue curve). Driven by the time dependent source the front point moves toward the origin. Within the interval where ϕ is set to zero the temperature tries to restore its stationary state, respectively the front moves also in direction its stationary (initial) position. The trajectory of the front point is nearly periodic.

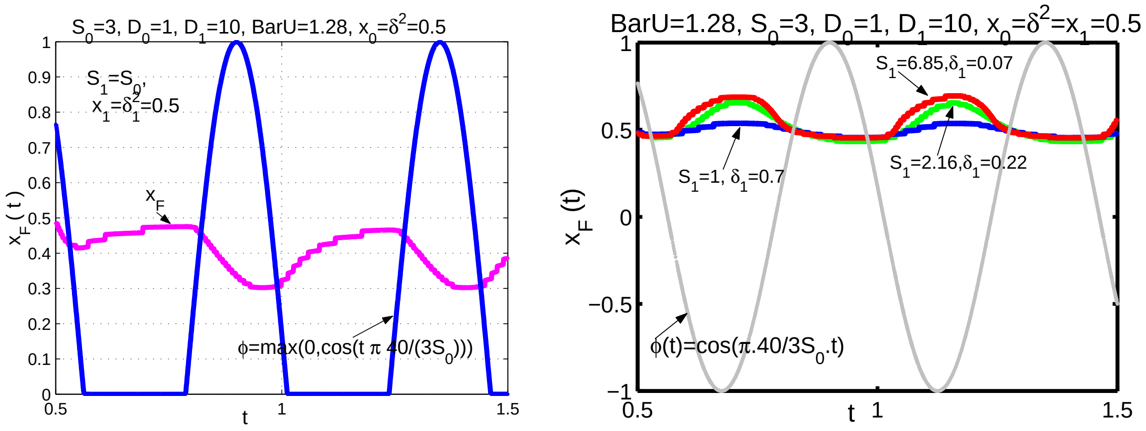

Figure 4.10: The evolution of the front point for different sources

We want to investigate the behaviour of the temperature varying δ_1 - Fig. 4.10, r.h.s. The parameter δ_1 describes the broadness of the pulse source, $\delta_1 \to 0$ corresponds to point pulse source. We kept the power of the pulse source constant, $\int\limits_0^1 \phi(t) S_1 \exp(-\frac{(x-x_1)^2}{\delta_1^2})dx = Const$, and observe the motion of the front point. The function ϕ, considered by us, is $\phi(t) = cos(t\pi\frac{40}{3}S_0)$ (the gray curve in fig.4.10, r.h.s). The amplitude in the motion of the front increases when δ_1 decreases. The tendency is to enlarge the amplitude in the front point motion for a point pulse source.

- *Multiple front points*

As mentioned above, in the stationary state there is at most one front point; during the evolution of the process, nevertheless, additional front points may appear for a while and disappear before reaching the steady state. Such situation reflects the appearing and disappearing of turbulence during the heat transport. This happens, for example, for the following set of parameters

$$D_0 = 1,\ \ D_1 = 10, S_0 = 90, \delta^2 = 0.05,\ \ x_0 = 0.4,\ \ \bar{u} = 1 \ll \frac{1}{2}\int\limits_0^1 S(x)dx.$$

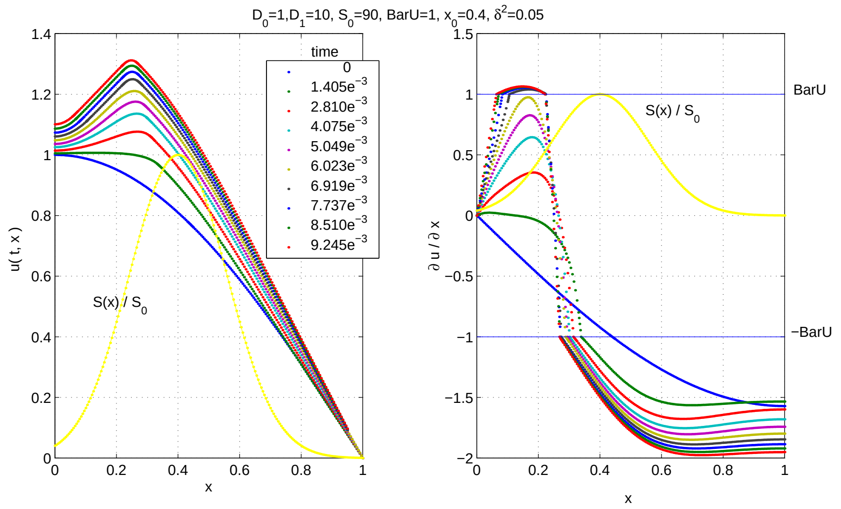

Figure 4.11: The evolution of the solution (l.h.s.) and the gradient (r.h.s.), away from the stationary state

The source defines a characteristic time shorter than the characteristic time defined by the heat conductivity. In this way the strong source drives the temperature to increase rapidly - Fig 4.11. As a result, close to the peak of the source, the temperature gradient becomes positive and reaches the critical value $+\bar{u}$. The heat transport switches to the anomalous in a new region, i.e additional front points appear. The left plot in Fig. 4.11 presents the evolution of the temperature for $t \in [0, 0.01]$, whereas the right shows the corresponding values for the gradient. The yellow curve depicts the re-scaled source $S(x)$. The characteristic time defined by the strength of the source is around $1/S_0 \approx 0.01$. After that time, actually at around $t \approx 0.025$, the heat starts to diffuse - see Fig. 4.13. This causes decay in the temperature gradient - the two front points coalesce and disappeared. On the steady state there is only one front point. In Fig 4.12 is shown the motion of the front points.

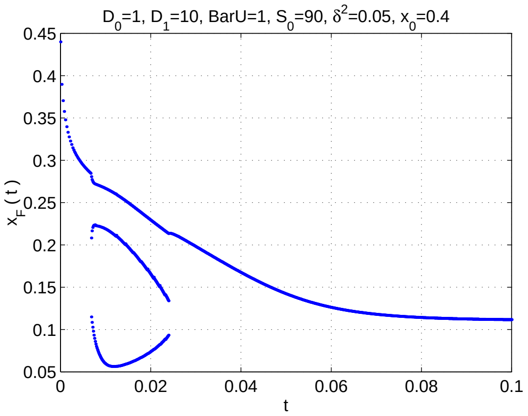

Figure 4.12: The evolution of the front point(s)

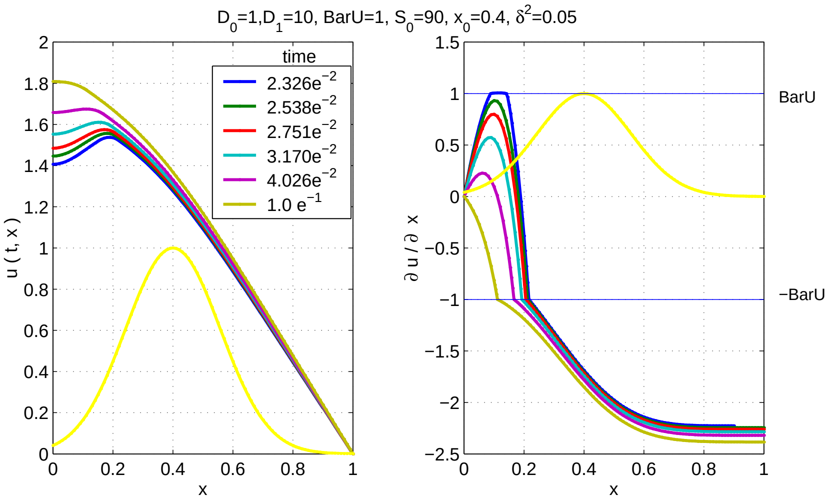

Figure 4.13: The evolution of the solution (left) and the gradient (right) approaching the stationary state

As we have discussed in Ch.3, we may have non-degenerate and degenerate front points - and here we have an example of both of them. Here, the degenerate front point is the point at which the temperature gradient touches the line $\bar{u}$ - at $t \approx 0.01$. This degenerate point gives birth to two non-degenerate front points - see Fig 4.12.

Actually the curve in Fig 4.12 that describes the motion of this additional front points should be a closed curve. Because of resolutions reasons we can not, numerically, track the whole trajectory. We solve the problem P1 till the gradient u_x crosses twice the critical value $\bar{u}$ at x_F^I and x_F^{II}, where the distance between them is 4 times the space step, i.e. $x_F^{II} - x_F^I \geq 4h$. Then we carry on with solving P2 in the interval $[x_F^I, x_F^{II}]$.

Turbulence in the heat transport could appear for specific combinations of the parameter values. It is most likely to happen for strong source (large value for S_0) and threshold value much smaller than half of the source power ($\bar{u} \ll \frac{1}{2D_0} \int\limits_0^1 S(x)dx$). The position of the source, x_0, also plays a role, a source with a position of the peak closer to the right boundary is more favourable to cause turbulence, as well as a source with more narrow width. Since D_1 defines the speed of the heat diffusion, a low value of it together with proper values for the other parameters could bring the process to a situation when appears and disappears turbulence in the heat transport.

Let us consider the question how the presence of front points influences the energy conservation. We have computed I_Err at $t = 0.02326$ and at $t = 0.1$. The first time, $t = 0.02326$, corresponds to the blue line from fig 4.13. There are three front points: $x_F^I = 0.088$, $x_F^{II} = 0.140$ and $x_F^{III} = 0.216$. The second time, $t = 0.1$, corresponds to the ochre line in the same figure, when the solution is close to its stationary state - the only front point is $x_F = 0.1116$. In Table 4.16 are listed the values of I_Err in each of the subintervals between the front points.

t	$[0, x_F^I]$	$[x_F^I, x_F^{II}]$	$[x_F^{II}, x_F^{III}]$	$[x_F^{III}, 1]$
0.02326	3.8229e-10	1.0762e-08	1.0520e-05	7.8521e-10
0.1	2.3493e-10			3.3417e-11

Table 4.16: I_Err - conservation of the energy

As it could be seen the conservation of the energy is poor around the point where the gradient change its sign (i.e. the temperature gets its maximum), but again improves approaching the stationary state.

- *Cylindrical and Spherical geometry.*

Now, let us consider the cases of cylindrical and spherical geometry, that is $d = 2, 3$. In order to have not only transient anomalous transport we have chosen the value of the threshold to be $\bar{u} \leq \frac{1}{D_0} \int\limits_0^1 x^{d-1} S(x) dx$. In the next plots, Fig. 4.14 and Fig. 4.15, the solution and its gradient, respectively, are presented for $d = 1, 2$ and $d = 3$, for the following set of parameters

$$D_0 = 1,\ D_1 = 10,\ \bar{u} = 0.69,\ S_0 = 5,\ x_0 = 0.5,\ \delta^2 = 0.5, t \in [0, 0.1].$$

The geometry does not change, qualitatively, the behaviour of the heat

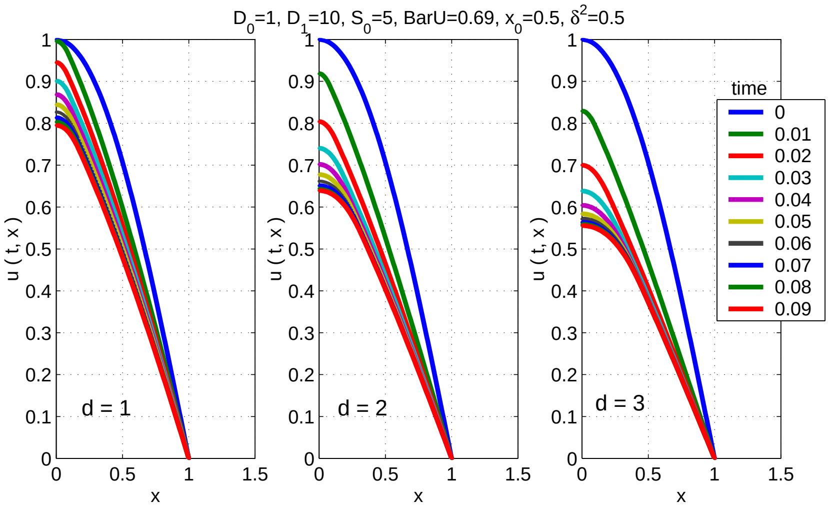

Figure 4.14: The evolution of the solution for one, two and three dimensions

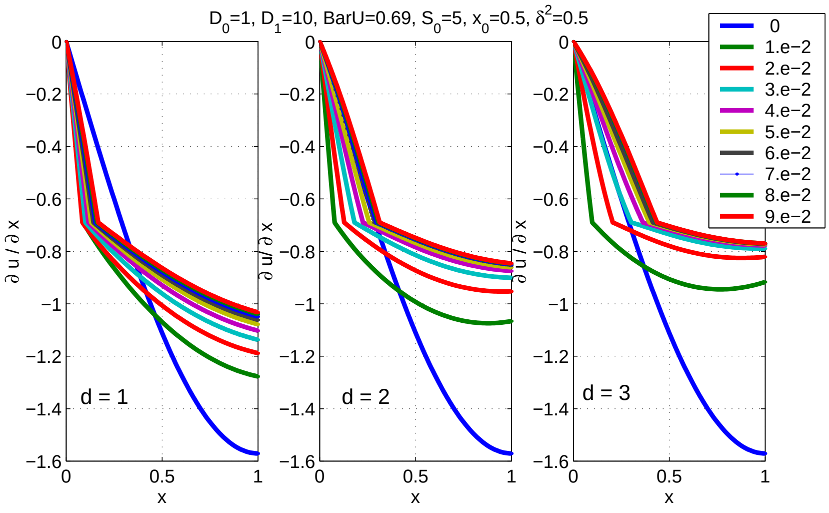

Figure 4.15: The evolution of the gradient for slab, cylindrical and spherical geometry

transport evolution. However, the larger the value of d the faster the temperature decreases and the faster the process goes to its stationary state.

In this chapter, we have, at first, demonstrated the efficiency of the Lobatto estimate in the time integrator of AIM. Further on we have shown the capabilities of the developed by us front tracking technique in solving the anomalous heat transport problem in the tokamak plasma. The latter does not encounter any problem in treating multiple front points or large jumps in the derivative of the heat conductivity coefficient.

Chapter 5

Summary & Conclusions

Adaptive Implicit Method (AIM)

In this thesis, we have analyzed and developed an adaptive algorithm (AIM) for solving nonlinear parabolic problems with a solution-dependent operator. It consists of discretization in space via Finite Element Method, which transforms the time-dependent PDE into a system of ODEs. This is solved by the implicit Trapezoidal rule coupled with a Newton-like iteration. We have proposed a novel strategy for controlling the number of the Newton iterations through monitoring the defect of the 4^{th} order Lobatto III A scheme. We have proven that the defect of the 4^{th} order Lobatto III A scheme could be used as a local estimate of the approximation error. Furthermore, this estimate of the local error is used for the automatic adaptation of the time step, as well as for estimating the global error. In Ch.4, we demonstrated the behaviour of the time integrator of AIM on a several of ODEs. For some of the considered problems the time integrator of AIM is better (in terms of computational effort) than for example the Lobatto IIIA 4^{th} order but not always. Particularly for the anomalous heat transport problem AIM is approximately 1.7 times better than the incomplete 4^{th} order Lobatto IIIA method (4.1). Moreover, we have demonstrated the high efficiency of Lobatto estimator on a system of ODEs with positive one-side Lipschitz constant and on a problem with a blow-up solution (called exploding by physicist).

Application of AIM to the Anomalous heat transport

The understanding of the mechanism governing the anomalous heat transport in the tokamak plasma is one of the substantial issues in the fusion research. It requires an adequate model of the problem and an accurate solution of it. The anomalous heat transport appears when the gradient of the temperature gets above a certain value. In this respect, the underlying model is a non-standard problem where the derivative of the conductivity coefficient depends discontinuously on the temperature gradient. Our objective was to solve the problem accurately in order to get insight into the nature of the phenomenon " anomalous transport". This led us while using the explicit front tracking technique. The evolution differential equation describing the heat transport has been split into two subproblems at the point where the gradient of the solution reaches the threshold. A theoretical investigation of the equivalence of the entire problem to the obtained two subproblems has been carried out. We have proven that each of the problems can be treated separately and that their solutions match continuously at the inner boundary (the front point). The proof is based mainly on the work of Ladyzhenskaia [25].

The AIM algorithm has been applied separately on each of the subproblems on a grid that partially varies from one time step to another. The front point is a grid point, around which we have formed an irregular sub-mesh that moves together with the interface (Ch. 3.5.4). Furthermore, employing the continuity of the flux over the interface, we have derived an equation for the speed of the front. This information has been used for choice of the sub-irregular grid at the next time step.

Using the idea of the front tracking technique, specifically the separation of the original problem into two subproblems, we have found two families of analytic solutions for a simpler version of the problem (Ch.3.4). They have been used in the numerical verification of the front tracking technique.

The algorithm, applied to the radially symmetric anomalous heat transport model, has been implemented in a C routine for linear finite elements method, and in a C-Fortran routine for Hermitian finite elements. The analytic solution has been used as a test example to examine the code, and to investigate the order of convergence of the algorithm - Ch.4.3.1. On this test example we have not detected, numerically, the theoretically proven order reduction for linear FEM. It turns out that the Hermitian elements do not increase the order of convergence (this is due to the fact that the second derivative of the solution is discontinuous) but conserve the energy to higher accuracy. Justification of the application of the front tracking technique instead of a non front tracking has been given. Shortly said, the FTT is twice faster and needs 1.5 times less grid points than a non front tracking method. Additional comparison with a problem with smoothed out conductivity coefficient is carried out. Once again the FTT performs better, not only for big values but also for small values of the jump in the derivative of the conductivity coefficient.

Investigation of the behaviour of the anomalous transport has been performed for values of the parameters related to real physical situations -

Ch.4.3.2. The code was tested for a large range of values of the jump in the slope of the flux without encountering any problems. Moreover, the size of the jump in the slope of the flux does not affect the steady state, but it does have influence on the speed of reaching this steady state. Analytical support of this statement has been given.

We have derived an analytic expression for the temperature gradient at the steady state. This expression has been used in a comparison with the numerically obtained values for the gradient. The high accuracy of the FTT has shown up. The error in the computation of the temperature gradient at the stationary state is of order $h^{3.5}$ (for h the spatial step).

One important case is considered - pulse source: an additional time dependent source is introduced once the solution reaches its stationary state. That is to simulate the usual technique used in the plasma experiments for getting information about the size of some of the constants taking part in the model. We have observed that a source depending monotonic on time drives a monotonic front motion.

For a specific set of parameters multiple front points occur, which correspond to a turbulence regime in the heat transport. These additional front points are only temporal and they dissappear on a very short time scale. At the stationary state at most one front point is possible. The position of front point depends mostly on the power of the source introduced in the medium.

Appendix A

Logarithmic matrix norm and Positive definite matrix

A.1 Logarithmic matrix norm

Let A be a matrix, i.e. $A \in \mathbb{R}^{s \times s}$ and $\lambda_i,\ i = 1, \cdots, s$ are the eigenvalues of A. For real eigenvalues we define $\lambda_{max} = \max\limits_{i} \lambda_i$. By I we denote the identity matrix and by A^t the transpose of A.

Definition 13 (Logarithmic matrix norm)**.** *(Verwer,Dekker[5])* **:**
For a given matrix $A \in \mathbb{R}^{s \times s}$ the **logarithmic norm** $\mu[A]$ *is defined by*

$$\mu[A] = \lim_{\delta \to 0+} \frac{||I + \delta A|| - 1}{\delta}$$

Properties:

$$\left.\begin{array}{l} 1)\quad \mu[A] = \max_{\xi \neq 0} \frac{<A\xi,\xi>}{||\xi||^2}, \\ \\ 2)\quad \mu[A] = \lambda_{max}[\frac{A+A^t}{2}], \end{array}\right\} \text{valid for } L_2 \text{ norm}$$

3) $\mu[A] \geq \alpha[A], \quad \alpha[A] = \max_i Re(\lambda_i[A])$,

4) $\mu[cA] = c\mu[A], \quad \forall c \geq 0, c \in \mathbb{R}$,

5) $\mu[A + cI] = \mu[A] + c, \quad \forall c \in \mathbb{R}$,

6) $\max(\mu[A] - \mu[-B], -\mu[-A] + \mu[B]) \leq \mu[A + B] \leq \mu[A] + \mu[B]$,

7) $|\mu[A] - \mu[B]| \leq \max(\mu[A - B], \mu[B - A]) \leq ||A - B||$,

8) $||A\xi|| \geq \max(-\mu[-A], -\mu[A])||\xi||,\ \forall \xi$.

Definition 14 (Singular values of matrix)**. :**
The singular values of an $m \times s$ matrix A are the square roots of the eigenvalues of the $s \times s$ matrix AA^t.

A.2 Positive definite matrix

Definition 15. *([17], p.7)* **:**
A square matrix $A \in \mathbb{R}^{s \times s}$ is

- *positive definite if $x^t Ax > 0$ for $0 \neq x \in \mathbb{R}^s$.*
- *positive semi-definite if $x^t Ax >= 0$ for $0 \neq x \in \mathbb{R}^s$.*
- *indefinite if $(x^t Ax)(y^t Ay) < 0$ for some $x,\ y \in \mathbb{R}^s$.*
- *diagonally dominant if $|a(i,i)| > \sum\limits_{j \neq i} |a(i,j)|$, for all i.*

Corollary 2. *(from Gershgorin Circle Theorem [17], p.200)* **:**
If the diagonal elements of a square matrix $A \in \mathbb{R}^{s \times s}$ are positive and both A and A^t are diagonally dominant then A is positive definite.

Proof:
Let $\sigma[A] = \{\lambda_i\}$ be the spectrum of the matrix $A = \{a_{ij}\},\ i,j = 1,\cdots,s$. Then the Gershgorin Circle Theorem says that $\sigma[A] \subset \bigcup\limits_{i=1}^{s} D_i$, where

$$D_i = \left\{ z \in \mathbb{C} : \quad |z - a_{ii}| \leq \sum_{k=1}^{s} |a_{ki}|,\ i = 1,\cdots,s \right\}.$$

Since A and A^t are diagonally dominant and $a_{ii} > 0, \quad i = 1,\cdots,s$ it follows that $\lambda_i[A] > 0$.

Proposition 3. :
If $A,\ B \in \mathbb{R}^{s \times s}$ are definite matrices, both positive or both negative definite, then the matrix AB is positive definite.

Proof:
Let us suppose that AB is indefinite, that is, there exist vectors x and y, $x^t x = 1$, $y^t y = 1$ such that $x^t ABx < 0$ and $y^t ABy > 0$, in the same time for the same x, $x^t Ax > 0$ and $x^t Bx > 0$ (if A and B are negative definite than we can use instead of A $(-A)$ and instead of B, $(-B)$) therefore $x^t Axx^t Bx > 0$. Now let us consider

$$0 > x^t ABx - x^t Axx^t Bx = x^t(AB - Axx^t B)x = x^t A(I - xx^t)Bx, \tag{A.1}$$

and let us denote $C_0 := I - xx^t$. The matrix C_0 is symmetric, with positive diagonal elements $(C_0)_{ii} = 1 - x_i^2 > 0$, and the absolute values of all the others are less than 1. If the matrix C_0 is not diagonally dominant then we add to $(A.1)$ $x^t ABx$ and we denote with $C_2 := 2I - xx^t$. If the matrix C_2 is not diagonally dominant we proceed till we obtain a matrix C_N which is. From the property of a diagonally dominant symmetric matrix with positive diagonal elements follows that our matrix C_N is positive definite. Now let us consider the product AC_N. Because A and B are positive definite their diagonal elements are positive and $(AC_N)_{ii} = a_{ii}(N - x_i^2) - \sum\limits_{j \neq i} a_{ij}x_i x_j$, the term $a_{ii}(N - x_i^2) > 0$ and if the whole $(AC_N)_{ii}$ is not positive then we can choose bigger N to assure this and also to provide that the matrices AC_N and $(AC_N)^t$ are diagonally dominant. So the product of A and C_N will be also positive matrix, and in the same way we can prove that $AC_N B$ is also positive matrix for big enough N. Finally we have that $AC_N B$ is positive definite and there exists a vector x such that $x^t AC_N Bx < 0$ which is contradiction.

Appendix B

Linear Parabolic Equations. Comparison theorem of Nagumo type

B.1 Linear Parabolic Equations

Let $D \subset \mathbb{R}^n$ be a bounded domain with boundary $\partial D =: S$, let $T > 0$ and

$$\Omega_T \equiv (0,T) \times D \equiv \{(t,x) \in \mathbb{R}^{n+1} : \ 0 < t < T, x \in D\}, \tag{B.1}$$

$S_T := (0,T) \times S$ be the lateral surface of the domain Ω_T, and $\bar{\Omega}_T$ is the closure of Ω_T.

Define a differential operator L by

$$Lu = \sum_{i,j=1}^{n} a_{ij}(t,x) \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{n} a_i(t,x) \frac{\partial u}{\partial x_i} + a(t,x)u - \frac{\partial u}{\partial t}. \tag{B.2}$$

with coefficients $a_{ij}, \ a_j, \ a$ defined in the cylinder $\bar{\Omega}_T$. Without loss of generality we assume that the matrix $\{a_{ij}\}$ is symmetric.

Definition 16. *We say that the operator L is* **parabolic** *at point $(t,x) \in \Omega_T$ if the symmetric matrix $\{a_{ij}(t,x)\}$ is positive definite. If L is parabolic at all points of Ω_T then we say that L is parabolic in Ω_T. If in addition there exist $\bar{\lambda}_0, \bar{\lambda}_1 > 0$ such that, for all $\xi \in \mathbb{R}^n$*

$$\bar{\lambda}_0 |\xi|^2 \le \sum_{i,j=1}^{n} a_{ij}(t,x) \xi_i \xi_j \le \bar{\lambda}_1 |\xi|^2, \quad \textit{for all } (t,x) \in \Omega_T,$$

then L is **uniformly parabolic** *in Ω_T.*

Now, first we give a theorem for existence and uniqueness of solution for the linear parabolic problem with Neuman boundary conditions, and then we consider the case of mixed boundary conditions.

We consider the second boundary value problem

$$\begin{aligned} Lu(t,x) &= f(t,x), \qquad \text{in } \Omega_T \\ u(0,x) &= u_0(x), \qquad \text{in } \bar{D} \\ \mathcal{B}u &\equiv \sum_{i=1}^{n} b_i(t,x)\frac{\partial u}{\partial x_i} + b_0(t,x)u = \Phi(t,x) \qquad \text{on } S_T. \end{aligned} \tag{B.3}$$

We assume that the functions $b_i(t,x)$ satisfy everywhere on S_T the condition

$$(b,\nu) \geq \delta > 0, \quad b = (b_1, \cdots, b_n)$$

where ν is the outer normal to S_T (i.e. the vector b does not at any point lie in the tangent plane to $S = \partial D$).

In order to define compatibility conditions we introduce the notation

$$u^{(k)}(x) = \left.\frac{\partial^k u(t,x)}{\partial t^k}\right|_{t=0},$$

$$\mathcal{A}u = \sum_{i,j=1}^{n} a_{ij}(t,x)\frac{\partial^2 u}{\partial x_i \partial x_j} - \sum_{i=1}^{n} a_i(t,x)\frac{\partial u}{\partial x_i} - au.$$

The functions $u^{(k)}$, $(k = 0, 1)$, are determined in the following manner:

$$u^{(0)}(x) = u(0,x), \quad u^{(1)}(x) = \mathcal{A}u(0,x) + f(0,x),$$

while the remaining functions are found from the recursion relation

$$u^{(k+1)}(x) = \frac{\partial^k}{\partial t^k}\mathcal{A}u(t,x) + \left.\frac{\partial^k f(t,x)}{\partial t^k}\right|_{t=0}.$$

Definition 17. *We say that the compatibility conditions of order $m \geq 0$ are fulfilled for problem (B.3) if*

$$\left.\frac{\partial^k}{\partial t^k}\mathcal{B}u\right|_{t=0} = \Phi^{(k)}(0,x), \ (k = 0, \cdots, m).$$

Theorem 9. [25, Th.5.3, Ch.IV]:
Suppose that $l = q + \alpha$ for $q \in \mathbb{N}$, $\alpha \in (0,1)$. Assume further that $S \in C^{l+2}$, the coefficients of the operator L belong to the class $C^{l/2,l}(\bar{\Omega}_T)$, and finally, b_i, $b_0 \in C^{l/2+1/2,l+1}(\bar{S}_T)$. Then problem (B.3) has a unique solution from the class $C^{1+l/2,2+l}(\bar{\Omega}_T)$ for any $f \in C^{l/2,l}(\bar{\Omega}_T)$, $u_0 \in C^{l+2}(\bar{D})$, $\Phi \in C^{(l+1)/2,l+1}(\bar{S}_T)$ satisfying the compatibility condition of order $[(l+1)/2]$.

Remark 11. [25, Ch.IV, pp.319]:
Theorem 9 is also applicable in the case of unbounded D, since the solution u is considered in the function space $C^{1+l/2,2+l}(\bar{\Omega}_T)$, the elements of which are bounded.

For the case of mixed boundary value problems, we define the operator L by

$$Lu \equiv u_t - \frac{\partial}{\partial x}(a_1(t,x)u_x). \tag{B.4}$$

The problem under consideration is

$$Lu = \frac{\partial f}{\partial x} - f(t,x) \tag{B.5}$$

$$\left.\frac{\partial u}{\partial \nu}\right|_{x=0} = \psi := const, \quad u\,|_{x=1} = 0, \tag{B.6}$$

$$u(0,x) = u_0(x). \tag{B.7}$$

In the next theorem we make use of the spaces $W_2^{0,1}(\Omega_T)$, $V_2^{0,1}(\Omega_T)$ and $V_2^{1/2,1}(\Omega_T)$ defined as:

- $W_2^{0,1}(\Omega_T)$ is the Hilbert space with scalar product

$$(u,v)_{W_2^{0,1}(\Omega_T)} = \int\limits_{\Omega_T}(uv + u_x v_x)dxdt.$$

- The space $V_2^{0,1}(\Omega_T)$ is the Banach space consisting of all elements of $W_2^{0,1}(\Omega_T)$ that are continuous in t in the norm of $L_2(D)$, with norm

$$|u|_{\Omega_T} = \max_{0\le t\le T}||u(t,x)||_{2,D} + ||u_x||_{2,\Omega_T},$$

where

$$||u||_{2,D} = \sqrt{\int\limits_D u^2 dx}, \quad ||u_x||_{2,\Omega_T} = \sqrt{\int\limits_{\Omega_T} u_x^2 dxdt}.$$

- The space $V_2^{1/2,1}(\Omega_T)$ is the subset of those elements $u(t,x)$ of $V_2^{0,1}(\Omega_T)$ for which

$$\int\limits_0^{T-\tau}\int\limits_D \tau^{-1}[u(t+\tau,x) - u(t,x)]^2 dxdt \to 0 \text{ for } \tau \to 0.$$

Theorem 10. [25, Th.5.1,pp.170]:
Suppose that the operator L defined by (B.4) is uniformly parabolic in Ω_T, S is piecewise-smooth boundary and

$$||f|| = (\int_{\Omega_T} f^2 dxdt)^{1/2} \leq \mu, \quad \mu = const > 0$$

$$||f||_{q,r,\Omega_T} = \left(\int_0^T (\int_D |f^q| dx)^{r/q} dt \right)^{1/r} \leq \mu, \quad q \in [1,2], \ r \in [1,4/3].$$

Then for any initial function $u_0(x) \in L_2(D)$ there exists a unique solution of the problem (B.5)-(B.7) in the class $V_2^{1/2,1}(\Omega_T)$.

Theorem 11. [25, Th. 12.1,pp 223]:
Suppose u is a weak solution from $V_2^{0,1}(\Omega_T)$ of equation (B.5), the coefficients and the free terms satisfy the conditions from Theorem 10. If the coefficients and the free terms, and also their derivatives are elements of $C^{m+\alpha/2,2m+\alpha}(\Omega_T)$, $m \geq 0$ then u belongs to $C^{m+2+\alpha/2,2m+2+\alpha}(\Omega_T)$.

B.2 Comparison theorem of Nagumo type

The Comparison theorem of Nagumo type is helping one to get super - and sub - solutions for parabolic boundary value problems. The results in this section follow the Walter paper [50].

For $n = 1$ and $D = [a,b] \subset \mathbb{R}$ let Ω_T be defined by (B.1). We consider a nonlinear elliptic operator $\mathcal{L}$ which acts on a function $u = u(t,x)$ and is defined by

$$(\mathcal{L}u)(t,x) = \frac{1}{\psi(x)}(\varphi(x, u_x(t,x)))_x.$$

- The function $\varphi(x,p)$ is assumed to be
  - i.) $\varphi \in C([a,b] \times \mathbb{R})$,
  - ii.) strictly increasing in p for $x \in [a,b]$ if $a > 0$, and for $x \in (0,b]$ if $a = 0$,
  - iii.) $\varphi(x,0) = 0$.
- The function $\psi(x)$ is continuous and positive in $[a,b]$, but $\psi(0) = 0$ is allowed in case $a = 0$.

The following type of problems are under consideration

$$\begin{aligned} u_t &= \mathcal{L}u + h(t,x) \qquad\qquad (B.8) \\ u(0,x) &= u_0(x) \text{ in } [a,b] \\ \mathcal{B}_a u &= u_a(t) \text{ at } x = a, \ \mathcal{B}_b u = u_b(t) \text{ at } x = b \end{aligned}$$

where $u_a,\ u_b$ are given functions, and $\mathcal{B}_a u = \alpha_1 u - \alpha_2 u_x$ and $\mathcal{B}_b u = \beta_1 u + \beta_2 u_x$. The coefficients $\alpha_i,\ \beta_i$ are nonnegative and $\alpha_1 + \alpha_2 > 0,\ \beta_1 + \beta_2 > 0$.

Let us consider the case $\alpha_2 > 0$ and $\beta_2 = 0$ and define the following class of functions $Z = \{y \in C(\bar{\Omega}_T) :\ y_t, y_x, \varphi(x, y_x), \mathcal{L}y \in C\}$.

Theorem 12. *Let us consider the case $\alpha_2 > 0$ and $\beta_2 = 0$ and let v and $w \in Z$ If*

$$v_t - \mathcal{L}v - h(t,x) \leq w_t - \mathcal{L}w - h(t,x)$$

and $\mathcal{B}_a v \leq \mathcal{B}_a w,\ \mathcal{B}_b v \leq \mathcal{B}_b w$, and $v_0 \leq w_0$ then $v \leq w$ in $\bar{\Omega}_T$.

Definition 18. *Sub-solutions v and super-solutions w for problem (B.8) are defined by $v, w \in Z$*

$$\begin{aligned} v_t &\leq \mathcal{L}v + h(t,x), \quad \mathcal{B}_a v \leq u_a(t),\ \mathcal{B}_b v \leq u_b(t) \\ w_t &\geq \mathcal{L}w + h(t,x), \quad \mathcal{B}_a w \geq u_a(t),\ \mathcal{B}_b w \geq u_b(t). \end{aligned}$$

If $u \in Z$ and u satisfies problem (B.8), moreover v and w are sub- and super-solutions for (B.8), respectively, and $v \leq w$ then $v \leq u \leq w$.

Appendix C

Miscellaneous

C.1 Lemma 1

Here, we give the proof of Lemma 1, Ch.2, using Taylor expansion.
Lemma 1:
Let U_{j+1} be an approximate solution, at $t = t_{j+1}$, of the initial value problem (2.8) computed through a difference scheme of type (2.9) of order $1 \leq p \leq 4$. Furthermore, let $U_j \equiv u_j$. Then, there exists the following connection between the truncation error, T_{j+1}, of the difference scheme and the defect of the 4th order Lobatto III A, $L_j(\tau, U_j, U_{j+1})$,

$$L_j(\tau, U_j, U_{j+1}) = T_{j+1}\Big(-1 + \frac{\tau}{2} f'(U_j)\Big) + \mathcal{O}(\tau^4).$$

Proof:
Let us consider the defect $L_j(\tau, U_j, U_{j+1})$

$$\begin{aligned} L_j(\tau, U_j, U_{j+1}) &= U_{j+1} - U_j - \frac{\tau}{6}(f(U_{j+1}) + f(U_j)) \\ &\quad -\frac{2\tau}{3} f\left(\frac{U_{j+1}+U_j}{2} + \frac{\tau}{8}(f(U_j) - f(U_{j+1}))\right). \end{aligned}$$

By adding and subtracting the exact solution $u(t)$ at t_{j+1} and expanding it at t_j we get

$$\begin{aligned} L_j(\tau, U_j, U_{j+1}) = \; & -(u_{j+1} - U_{j+1}) + \overbrace{u_j + \tau \dot{u}_j + \frac{\tau^2}{2}\ddot{u}_j + \frac{\tau^3}{6}\dddot{u}_j + \mathcal{O}(\tau^4)}^{u_{j+1}} - U_j \\ & -\tfrac{\tau}{6}(f(U_{j+1}) + f(U_j)) - \tfrac{2\tau}{3} f\left(\tfrac{U_{j+1}+U_j}{2} + \tfrac{\tau}{8}(f(U_j) - f(U_{j+1}))\right). \end{aligned}$$

We use the fact that $T_{j+1} = u_{j+1} - U_{j+1}$ for $U_j = u_j$ and expand the function $f(.)$ at u_j

$$\begin{aligned} L_j(\tau, U_j, U_{j+1}) = \quad & -T_{j+1} - \tfrac{\tau}{6}(U_{j+1} - u_j)(3f'(u_j) + f''(u_j)(U_{j+1} - u_j)) \\ & + \tfrac{\tau^2}{2}(\ddot{u}_j + \tfrac{1}{6}f'^2(u_j)(U_{j+1} - u_j)) + \tfrac{\tau^3}{6}\dddot{u}_j + \mathcal{O}(\tau^k(U_{j+1} - u_j)^l). \end{aligned} \tag{C.1}$$

The indexes k and l in the term $\mathcal{O}(\tau^k(U_{j+1} - u_j)^l)$ are such that $k + l \geq 4$, for $k,\ l \geq 0$. Let us have a look at $(U_{j+1} - u_j)$

$$(U_{j+1} - u_j) = -T_{j+1} + (u_{j+1} - u_j) = -T_{j+1} + \tau f(u_j) + \frac{\tau^2}{2} f(u_j)f'(u_j) + \mathcal{O}(\tau^3)$$

and substitute it in (C.1)

$$\begin{aligned} L_j(\tau, U_j, U_{j+1}) = \quad & -T_{j+1} + \tfrac{\tau^2}{2}\ddot{u}_j + \tfrac{\tau}{2}f'(u_j)T_{j+1} - \tfrac{\tau^2}{2}f'(u_j)f(u_j)(1 + \tfrac{\tau}{2}f'(u_j)) \\ & -\tfrac{\tau}{6}f''(u_j)T_{j+1}^2 - \tfrac{\tau^3}{6}f''(u_j)f^2(u_j) + \tfrac{\tau^3}{6}\dddot{u}_j - \tfrac{\tau^2}{12}f'(u_j)T_{j+1} \\ & -\tfrac{\tau^2}{12}f'^2(u_j)T_{j+1} + \tfrac{\tau^3}{12}f'^2(u_j)f(u_j) + \mathcal{O}(\tau^4). \end{aligned}$$

Due to the fact that T_{j+1} is at least of second order and $\ddot{u}_j = f'(u_j)f(u_j)$, $\dddot{u}_j = f'^2(u_j)f(u_j) + f''(u_j)f^2(u_j)$ we write that

$$L_j(\tau, U_j, U_{j+1}) = T_{j+1}(-1 + \frac{\tau}{2}f'(U_j)) + \mathcal{O}(\tau^4). \tag{C.2}$$

♢

C.2 Butcher tableaux

The Butcher tableaux are shorter (tableaux) way of writing the coefficients for the Runge-Kutta methods. For example the Butcher tableaux

$$\begin{array}{c|ccc} c_1 & a_{11} & a_{12} & a_{22} \\ c_2 & a_{21} & a_{22} & a_{23} \\ c_3 & a_{31} & a_{32} & a_{33} \\ \hline & b_1 & b_2 & b_3 \end{array}$$

defines the following three stage Runge-Kutta method

$$\begin{aligned} k_1 &= f(t + c_1h, U_j + a_{11}hk_1 + a_{12}hk_2 + a_{13}hk_3) \\ k_2 &= f(t + c_2h, U_j + a_{21}hk_1 + a_{22}hk_2 + a_{23}hk_3) \\ k_3 &= f(t + c_3h, U_j + a_{31}hk_1 + a_{32}hk_2 + a_{33}hk_3) \end{aligned}$$

$$U_{j+1} = U_j + b_1hk_1 + b_2hk_2 + b_3hk_3.$$

C.3 Implicit Function Theorem

Here, we give the Implicit function theorem [45].

Theorem 13. [45, Th.C.7,pp. 658]:
Suppose that X, Y and Z are Banach spaces, $U \subset X$, $V \subset Y$ are open sets, $F \in C^k(U \times V, Z)$, $k \geq 1$ and that $(x_0, y_0) \in U \times V$, $F(x_0, y_0) = 0$ and $\mathcal{D}_x F(x_0, y_0)$ has a bounded inverse. Then there is a neighbourhood $U_1 \times V_1 \subseteq U \times V$ of (x_0, y_0) and a function $f \in C^k(V_1, U_1)$ such that $f(y_0) = x_0$ and $F(x, y) = 0$ for $(x, y) \in U_1 \times V_1$ if and only if $x = f(y)$. Furthermore, if $k > 1$, then the size of $U_1 \times V_1$ may be bounded from below in terms of the norm of $[\mathcal{D}_x F(x_0, y_0)]^{-1}$ and the norms of the second derivatives of $F(x, y)$ in $U \times V$.

One often has to deal with a problem where a $x \in \mathbb{R}^n$ is an implicit function of $y \in \mathbb{R}^m$ defined by

$$H(y, x) = 0, \tag{C.3}$$

where H is only locally Lipschitz function mapping $\mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n$. Before giving the implicit function theorem due to Clarke we define a generalized Jacobian or Clarke derivative of a Lipschitz function.

If $f : \mathbb{R}^m \to \mathbb{R}^n$ is locally Lipschitz continuous, then its **generalized Jacobian** at a point $x \in \mathbb{R}^m$, denoted $\partial f(x)$, is given by

$$\partial f(x) = conv \left\{ A : \; A = \lim_{x' \to x, x' \in D_f} Df(x') \right\} \tag{C.4}$$

where $conv(B)$ denotes the closed convex hull of the set B, and D_f denotes the set of points at which f is differentiable. In particular, note that the generalized Jacobian may be multi-valued at some points.

Let $\pi_x \partial H(y, x) = \{M \in \mathbb{R}^{n \times n}$: for some $N \in \mathbb{R}^{n \times m}$the matrix$[N, M] \in \partial H(y, x) \subset \mathbb{R}^{n \times (m+n)}\}$. Denote $\pi_y \partial H(y, x)$ be such that $[\pi_y \partial H(y, x), \pi_x \partial H(y, x)] = \partial H(y, x)$.

Then the implicit function theorem due to Clarke is:

Theorem 14. [46, Th.1.1]:
Suppose that $H : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n$ is a locally Lipschitz function in a neighbourhood of $(\bar{y}, \bar{x})$, which is a solution of (C.3), i.e., $H(\bar{y}, \bar{x}) = 0$. If $\pi_x \partial H(\bar{y}, \bar{x})$ is of maximal rank, then there exist an open neighbourhood Y of $\bar{y}$ and a function $G(.) : Y \to \mathbb{R}^n$ such that G is locally Lipschitz in Y , $G(\bar{y}) = \bar{x}$ and for every $y \in Y$, $H(y, G(y)) = 0$.

Bibliography

[1] D. J. Benson, *Volume of fluid interface reconstruction methods for multi-material problems,* Appl. Mech. Rev. 55 (2002), pp. 151-165

[2] F. A. Bornemann, *An adaptive multilevel approach to parabolic equations. General theory and 1D implementation,* Impact of Computing in Science and Engineering 2 (1990), pp. 279-317

[3] C. J. Budd, R. Carretero-Gonzalez, R. D. Russell, *Precise computations of chemotactic collapse using moving mesh methods,* Journal of Computational Physics, Vol.202, Issue 2 (2005), pp. 463-487

[4] W. Cao, W. Huang, R. D. Russell, *Approaches for generating moving adaptive meshes: location versus velocity,* Applied Num. Math. 47 (2003), pp. 121-138

[5] K. Dekker, J. G. Verwer, *Stability of Runge-Kutta methods for stiff nonlinear differential equations,* North-Holland Publ, Amsterdam (1984)

[6] P. Deuflhard, *Uniqueness theorems for stiff ODE initial value problems,* D. F. Griffiths and G. A. Watson(eds), Numerical analysis 1989, Proceedings of the 13th Dundee Conference, pp.74-87,Pitman Research Notes in Math. Series 228, Longman Scientific and Technical (1990)

[7] P. Deuflhard, F. Bornemann, *Scientific computing with ordinary differential equations,* Springer-Verlag New York, Inc (2002)

[8] St. Dimova, D. P. Vasileva, *Numerical heat transfer, part B:Lumped-Mass finite element method with interpolation of the nonlinear coefficients for a quasilinear heat transfer equation,* Taylor& Francis (1995), pp. 199-215

[9] G. G. Elenin, S. P. Kurdymov, A. Samarski, *Nonstationary dissipative structures in a nonlinear heat-conducting medium,* Zh. Vychisl Mat. i Mat. Fiz, vol 23 (1983), pp. 380-390

[10] K. Eriksson, C. Johnson, V. Thome, *Time discretization of parabolic problems by the discontinuous Galerkin method,* RAIRO Model. Math. Anal. Numéer. 19(4) (1985), pp.611-643

[11] Erwan Faou, Ernst Hairer, Truong-Linh Pham, *Energy conservation with non-symplectic methods: examples and counter-examples* (2004)

[12] K. Finckenstein, *Difference methods for quasilinear parabolic systems from plasma physics,* Numerical Methods for PDEs,3 (1987), pp. 289-311

[13] C. A. J. Fletcher, *Computational Galerkin methods,* Springer-Verlag (1984)

[14] A. Friedman, *Partial differential equation of parabolic type,* Prentice-Hall, INC (1964)

[15] M.J. Gander, R. Meyer-Spasche,*An introduction to numerical integrators preserving physical properties*, World Scientific (2002)

[16] J. Glimm, Xiao Lin Li, Yingjie Liu, Ning Zhao, *Conservative front tracking and level set algorithms,* PNAS, Vol 98, no 25 (2001)

[17] G. H. Golub, C. F. Van Loan, *Matrix computations,* Baltimore, MD, Johns Hopkins Univ. Press (1983)

[18] E. Hairer, S. P. Norsett, G. Wanner, *Solving ordinary differential equations I- nonstiff problems,* Springer-Verlag, Berlin/Heidelberg (1987)

[19] E. Hairer, G. Wanner, *Solving ordinary differential equations II- stiff and differential-algebraic problems,* Springer-Verlag, Berlin/Heidelberg (1991)

[20] E. Hairer, C. Lubich, G. Wanner, *Geometric numerical investigation - structure preserving algorithms for ordinary differential equations,* Springer-Verlag, Berlin/Heidelberg (2002)

[21] T. S. Hahm, K. H. Burrell, *Flow shear induced fluctuation suppression in finite aspect ratio shaped tokamak plasma,* Physics of Plasma 2, Vol 2, Issue 5 (1995), pp. 1648-1651

[22] Ph. Hartman, *Ordinary Differential Equations,* John Wiley & Sons Inc (1964)

[23] P. Helander, D. J. Sigmar, *Collisional transport in magnetized plasmas,* Cambridge University Press (2002)

[24] W. Hundsdorfer, J.G. Verwer, *Numerical solution of time-dependent advection-diffusion-reaction equations,* Springer-Verlag (2003)

[25] O. A. Ladyzhenskaia, V. A. Solonnikov, N. N. Ural'ceva, *Linear and quasilinear equations of parabolic type,* Translations of Mathematical Monographs, Vol. 23, American Mathematical Society (1968)

[26] J. Lang, *Adaptive multilevel solution of nonlinear parabolic PDE systems,* Springer (2001)

[27] J. Lang, *Adaptive FEM for reaction-diffusion equations,* Applied Numerical Mathematics 26 (1998), pp.105-116

[28] M.-N. Le Roux, H. Wilhelmsson, *External boundary effects of simultaneous diffusion and reaction processes,* Physica Scripta, Vol 40 (1989), pp. 674-681

[29] C. Lederman, J. L. Vazquez, N. Wolanski, *A mixed semilinear parabolic problem from combustion theory,* Electron.J.Diff.Eques.,Conf.06 (2001), pp. 203-214

[30] M.-N. Le Roux, J. Weiland, H. Wilhelmsson, *Simulation of a coupled dynamic system of temperature and density in a fusion plasma,* Physica Script, Vol. 46 (1992), pp. 457-462

[31] G. M. Lieberman, *Second order parabolic differential equations,* World Scientific (1996)

[32] J. López, J. Hernández, P. Gómez, F. Fauraa, *A volume of fluid method based on multidimensional advection and spline interface reconstruction,* Journal of Computational Physics Vol. 195, Issue 2 (2004), pp. 718-742

[33] N. J. Lopes Cardozo, *Perturbative transport studies in fusion plasmas,* Plasma Phys. and Control. Fusion 37, Issue 8 (1995), pp. 799-852

[34] A. Lunardi, *Analytic semigroups and optimal regularity in parabolic problems,* Birkhäuser (1995)

[35] Rita Meyer-Spasche, *Difference schemes of optimum degree of implicitness for a family of simple ODEs with blow-up solutions,* Journal of Comput. and Appl. Math. 97 (1998), pp. 137-152

[36] J.D. Murray *Mathematical Biology*, Springer Verlag, Heidelberg (1989)

[37] K. Nishikawa, M. Wakatani, *Plasma physics - basic theory with fusion applications,* Springer Series in Atoms and Plasmas Berlin/Heidelberg, Vol. 8 (1994)

[38] Stanely Osher, Ronald P. Fedkiw, *Level set methods: an overview and some recent results,* Journal of Comp. Physics 169 (2000), pp. 463-502

[39] Linda Petzold, Laurent Jay and Jeng Yen, *Numerical solution of highly oscillatory ordinary differential equations*, Acta Numerica 6 (1997), pp. 437-484

[40] H.H. Rosenbrock, *Some general implicit processes for numerical solution of differential equations,* Computar J.,5 (1963), pp.329-331

[41] A. Samarski, V. Galaktionov, S. P. Kurdymov, A. P. Mikhailov *Blow-up in quasilinear parabolic equations,* Walter de Gruyter (1995)

[42] W. E. Schiesser, *The numerical method of lines. Integration of partial differential equations,* Boston, MA Academic Press (1991)

[43] J. A. Sethian, *Level Set methods and fast marching methods. Evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science,* Cambridge University Press (1999)

[44] G. Strang, G. Fix, *An analysis of the finite element method,* Wellesley-Cambridge Press (1997)

[45] A. M. Stuart, A. R. Hymphries, *Dynamical systems and numerical analysis,* Cambridge Univ. Press (1996)

[46] D. Sun, *A further result on an implicit function theorem for locally Lipschitz functions,* Operations Research Letters **28** (2001) 193 - 198

[47] V. Thomee, *Galerkin finite element methods for parabolic problems,* Springer (1997)

[48] G. Tryggvason, A. Esmaeeli, N. Al-Rawahi, *Direct numerical simulations of flows with phase change,* Computers& Structures, Vol 83, Iss 6-7, (2005), pp.445-453

[49] S. O. Unverdi, G. Tryggvason, *A front tracking method for viscous, incompressible, multi-fluid flows,* Journal of Computational Physics 100 (1992), pp.25-37

[50] W. Walter, *Nonlinear parabolic differential equations and inequalities,* Discrete and Continuous Dynamical Systems, Vol. 8, Num. 2 (2002), pp.451-468

[51] J. Wesson, *Tokamaks,* Clarendon Press, Oxford (1997)

[52] H. Wilhelmsson, *Diffusion, creation and decay processes in plasma dynamics: evolution towards equilibria and the role of bifurcated states,* Nuclear Phys. A, Vol 518 (1990), pp. 84-98

[53] H. Wilhelmsson, B. Etlicher, R. A. Carins, M.-N. La Roux, *Evolution of temperature profiles in fusion reactor plasma,* Physica Scripta, Vol 45 (1992), pp. 184-187