Finite Difference Methods
for 1st Order in Time, 2nd Order in Space
Hyperbolic Systems
used in Numerical Relativity

Dissertation

von

Mihaela Chirvasa

eingereicht bei der

Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Potsdam

durchgeführt in Golm am

Max Planck Institut für Gravitationsphysik
Albert Einstein Institut

unter der Betreuung von

Prof. Dr. Bernard Schutz

Potsdam, July 2009

Published online at the
Institutional Repository of the University of Potsdam:
URL http://opus.kobv.de/ubp/volltexte/2010/4213/
URN urn:nbn:de:kobv:517-opus-42135
http://nbn-resolving.org/urn:nbn:de:kobv:517-opus-42135

Acknowledgments

I would like to thank my supervisor Prof. Bernard Schutz for his help and enthusiasm during this work. Special thanks go also to Prof. Edward Seidel for his guidance in the early stages of the thesis.

I am very much indebted to Sascha Husa, Manuel Tiglio, Bela Szilágyi, Denis Pollney and Steve White for their invaluable support and advice.

It was a privilege to work at the Max Planck Institute for Gravitational Physics and I am grateful for having this chance.

Finally, I would like to thank Thomas and Yanis for their moral support over the last years.

Contents

1 Introduction **1**

1.1 Initial (Boundary) Value Problem in Numerical Relativity . . 3

1.2 1st order time 2nd order space hyperbolic formulations 7

1.3 Numerical Discretization 10

1.4 Outline of the Thesis 13

2 Initial Value Problem for 2nd Order Systems in space and 1st order in time **15**

2.1 Space Discretization 15

2.1.1 Representation of periodic functions 16

2.1.2 High Order Finite Difference Operators 19

2.1.3 Artificial Dissipation Operator 23

2.1.4 Fourier Symbols: Properties I 24

2.1.5 Fourier Symbols: Properties II 26

2.2 Time Integration using Runge-Kutta Methods 31

2.2.1 Construction 31

2.2.2 Absolute Stability of Runge-Kutta methods 35

2.3 Well-Posedness and Numerical Stability 37

2.3.1 Well-Posedness 40

2.3.2 Numerical Stability 42

2.4 Dispersion and Dissipation 46

2.4.1 Mode Splitting 46

2.4.2 Amplification Factor and Speed Errors 47
2.4.3 Advection Equation 48

3 Initial Value Problem for the Wave Equation 50
3.1 Introduction 50
3.2 Continuum Problem 52
3.3 Discrete Problem 53
3.3.1 Semidiscrete Problem 54
3.3.2 Courant Limits and the Role of Dissipation 56
3.4 Dispersion and Dissipation 60
3.5 Phase and Group Speeds 64
3.5.1 Small Frequencies 65
3.5.2 Scaling of the Speeds Errors with the Order of Approximation when $\beta = 0$ 66
3.5.3 Scaling of the Speeds Errors with the Order of Approximation when $\beta \neq 0$ 67
3.5.4 Speeds Errors for Different Off-Centerings at the Same Order of Approximation 76
3.6 Numerical Experiments in 1-D 85
3.6.1 Centered Scheme vs One-Point Advected Scheme . . . 86
3.6.2 Accuracy and Convergence of Higher Orders 87

4 Initial Boundary Value Problem for the Wave Equation 94
4.1 Theoretical Background 95
4.1.1 Well-Posedness and Strong Well-Posedness 95
4.1.2 Discrete Schemes and Stability Concepts for IBVP . . 96
4.2 Continuum Problem 102
4.3 Ghost-Point Method 105
4.3.1 Boundary Prescriptions 105
4.3.2 Equivalent Systems and Known Results 107

4.3.3 Stability Analysis via Energy Method for Outflow Boundary . . . 111
4.3.4 Stability Analysis via Laplace Transform Method . . . 117
4.4 SBP-SAT Method for Inflow Boundary . . . 129
4.4.1 Another Continuum Energy Estimate . . . 129
4.4.2 Stability Analysis . . . 131
4.5 Numerical Tests . . . 136
4.5.1 General Setup . . . 136
4.5.2 Inflow-Inflow GP algorithm($|\beta| < 1$) . . . 138
4.5.3 Outflow-Completely Inflow GP algorithm($|\beta| > 1$) . . . 139
4.5.4 Outflow-Inflow GP algorithm ($\beta = 1$) . . . 141
4.5.5 Inflow-Inflow SBP-SAT algorithm($|\beta| < 1$) . . . 143
4.5.6 Disscussion . . . 147

5 BSSN System in Spherical Symmetry 149
5.1 Introduction . . . 149
5.2 Deducing the Equations . . . 152
5.2.1 ADM in Spherical Symmetry . . . 152
5.2.2 BSSN in Spherical Symmetry . . . 153
5.2.3 Minimal System with Densitized Lapse . . . 161
5.2.4 Analysis of the Principal Part . . . 162
5.3 Numerical Implementation . . . 164
5.3.1 Boundary Algorithms . . . 165
5.4 Numerical Results: Linear Case . . . 168
5.4.1 test 1: two timelike boundaries, zero shift . . . 169
5.4.2 test 2(a): two timelike boundaries, with shift (static) . 171
5.4.3 test 2(b): two timelike boundaries, with shift (dynamic) 173
5.4.4 test 3 : spacelike inner boundary and timelike outer boundary . . . 175
5.5 Numerical Results: Nonlinear Case . . . 175

6 Summary and Outlook **179**
6.1 Summary 179
6.2 Outlook 188

A Harmonic Formulation **189**

B 3+1 split and ADM equations **191**

C BSSN equations **193**

D SBP operators **195**

Notations

continuum

$\vec{x} = (x_1 \dots x_d)$	space vector
$\mathsf{u}, \mathsf{v}, \mathsf{K}, \Phi$	vector or scalar functions
$\hat{\mathsf{u}}, \hat{\mathsf{v}}, \hat{\mathsf{K}}, \hat{\Phi}$	Fourier coefficients
$\vec{\omega} = (\omega_1 \dots \omega_d)$	wave vector
$\partial_{i_1 \dots i_m}$	partial differential operator $\frac{d^m}{dx^{i_1} \dots dx^{i_m}}$
$\hat{\partial}_{i_1 \dots i_m}$	Fourier symbol associated to $\partial_{i_1 \dots i_m}$
v_p (v_g)	phase (group) speed
$\hat{\mathsf{P}}$, $\hat{\mathsf{H}}$, $\hat{\mathsf{T}}$, $\hat{\Delta}$, ...	expressions depending on Fourier symbols

discrete

N	number of grid points
h	space resolution
k	time resolution
λ	Courant factor
s	number of offcentered points
ϵ	sense of offcentering (left or right)
$\underline{x}$	grid coordinates
u, v, K, Φ	grid vectors
$\hat{u}, \hat{v}, \hat{K}, \hat{\Phi}$	discrete Fourier coefficients
$\underline{\omega} = (\omega_1, \dots \omega_d)$	grid wave vector
$\underline{\xi} = (\xi_1, \dots \xi_d)$	grid frequency
D_+, D_-	forward, backward FDO

$D^{(m,n,s,\epsilon)}$	$2n$-accurate FDO corresponding to ∂^m
$D^{(m,n)}$	$2n$-accurate CFDO corresponding to ∂^m
$D^{(m,n)}_{i_1\ldots i_m}$	$2n$-accurate CFDO corresponding to $\partial_{i_1\ldots i_m}$
$\hat{D}_\pm$, $\hat{D}^{(m,n,s,\epsilon)}$, $\hat{D}^{(m,n)}$, $\hat{D}^{(m,n)}_{i_1\ldots i_m}$	discrete Fourier symbols
$v_p, v_p^{(n,s)}, v_p^{(n)}$ $(v_g, v_g^{(n,s)}, v_g^{(n)})$	numerical phase (group) speeds
$\epsilon_p, \epsilon_p^{(n,s)}, \epsilon_p^{(n)}$ $(\epsilon_g, \epsilon_g^{(n,s)}, \epsilon_g^{(n)})$	numerical phase (group) speed errors
a_p	amplification factor
$\hat{P}$, $\hat{H}$, $\hat{T}$, $\hat{\Delta}$, ...	expressions depending on discrete Fourier symbols

Abstract

This thesis is concerned with the development of numerical methods using finite difference techniques for the discretization of initial value problems (IVPs) and initial boundary value problems (IBVPs) of certain hyperbolic systems which are first order in time and second order in space. This type of system appears in some formulations of Einstein equations, such as ADM, BSSN, NOR, and the generalized harmonic formulation.

For IVP, the stability method proposed in [14] is extended from second and fourth order centered schemes, to $2n$-order accuracy, including also the case when some first order derivatives are approximated with off-centered finite difference operators (FDO) and dissipation is added to the right-hand sides of the equations.

For the model problem of the wave equation, special attention is paid to the analysis of Courant limits and numerical speeds. Although off-centered FDOs have larger truncation errors than centered FDOs, it is shown that in certain situations, off-centering by just one point can be beneficial for the overall accuracy of the numerical scheme.

The wave equation is also analyzed in respect to its initial boundary value problem. All three types of boundaries - outflow, inflow and completely inflow- that can appear in this case, are investigated. Using the ghost-point method, $2n$-accurate ($n = \overline{1,4}$) numerical prescriptions are prescribed for each type of boundary. The inflow boundary is also approached using the SAT-SBP method.

In the end of the thesis, a 1-D variant of BSSN formulation is derived and some of its IBVPs are considered. The boundary procedures, based on the ghost-point method, are intended to preserve the interior $2n$-accuracy. Numerical tests show that this is the case if sufficient dissipation is added to the rhs of the equations.

Chapter 1

Introduction

The motivation of this thesis comes from the problem of solving numerically the Einstein equations,

$$G_{\mu\nu} = 8\pi T_{\mu\nu} \tag{1.1}$$

which describe the geometry of spacetime (Einstein tensor $G_{\mu\nu}$) as caused by matter and energy ($T_{\mu\nu}$). They are actually a set of ten quasilinear coupled partial differential equations for the ten components of the metric, $g_{\mu\nu}$.

The power of these equations resides in their ability to make predictions of physical or practical relevance ranging from cosmological models to the the orbits of communication satellites. But maybe the most exciting prediction which until now has not been *directly* verified is the existence of gravitational waves, as ripples in the spacetime fabric that propagate with the speed of light. If discovered, the gravitational waves will open a new window into the universe, allowing observations on its very distant or hidden regions such as black holes, neutron stars, or interior of supernovae. If the cosmic background electromagnetic radiation describes the universe as it was 10^5 years after the Big Bang, and studies of cosmological nucleosynthesis provide information about how it was after 3 minutes, the gravitational waves could picture the universe when it was only 10^{-24} seconds old, just at the end of inflation.

Although they are produced by any moving massive object, due to their

miniscule strength, they are expected to be detected in the near future, only if they originate in the most violent and distant cosmic laboratories, such as the process of coalescence of black holes or neutron stars.

Their detection is pushing not only the present-day technology to its limits by requiring extremely sensitive detectors (which should differentiate scales of 10^{-21}cm), but also at the theoretical level there is lots of provocation. In order to make possible the extraction of such a small signal from the noise, theoreticians have to tell the experimentalists what to look for, that is to provide them with accurate templates of the gravitational waves' signal.

For the binary system problem, the waveforms generated during the inspiral and after the merger are well understood. However, the gravitational radiation emitted at the time of coalescence is not, and the only way to gain insight into it is by simulating the full process on supercomputers.

This is where numerical relativity comes into play, using algorithms and supercomputers to analyze and solve the problem. Understanding the physics of coalescing compact objects and predicting the gravitational wave signal emitted in such processes is a major problem for numerical relativity nowadays, a field of research that has emerged in its own right from general relativity in the late '60s. Recent breakthroughs in the field [3, 11, 66] have lead to dramatic progress, providing a surge of astrophysical results. The accurate tracking of orbiting compact objects imposes high demands on computational accuracy. Correspondingly, in numerical relativity, discretizations higher than fourth order have lead to significant improvements in the quality of results from long evolutions, extending up to about 10 orbits before merger [21,38,42]. However, a systematic understanding of the numerical methods that underlie the field, and of "best practices" that should be employed have not yet been achieved.

The purpose of this thesis is to make another step in that direction, by highlighting several important aspects that come with the high order discretization of initial value problem and initial boundary value problem in numerical relativity.

But before going into finite differencing techniques, it is instructive to make

a stop at the continuum level to review the setup, the main requirements and the main approaches.

1.1 Initial (Boundary) Value Problem in Numerical Relativity

Setup

Modeling the Einstein equations for numerical purposes involves two steps:

1. **setup of the geometric arena**; The Einstein equations are covariant; this means one can choose any coordinate system in order to solve them. So naturally, the first step requires the introduction of a foliation of the manifold (physical or constructed from the physical one) in 3-hypersurfaces and the definition of some quantities adequate to the geometry. These quantities are to be measured and analyzed in order to extract physical information about the given phenomena.

2. In the second step, the Einstein equations are used in order to **set up an evolution algorithm** in the form of an initial value problem (IVP) or initial boundary value problem (IBVP) for these quantities; in all the cases, the evolution system is constructed using only some of the Einstein equations, while the rest form a *system of constraints* which should be compatible with the evolution system. [1]

Main Requirements

In order that the given setup to be useful in practical applications, the evolution system has to satisfy three requirements:

[1] in the absence of boundaries, they are actually compatible with the constraints, by virtue of the Bianchi identities.

1. it has to be **well-posed**; Well-posedness is a concept introduced by Hadamard [37] in order to describe mathematical models of physical phenomena. Depending of the problem (model), the definition for well-posedness might vary, but there are basically three properties that a model has to satisfy in order to be well-posed: a solution exists, is unique and depends continuously on the data.

2. it has to be **compatible with the constraints**; in other words, the solution should satisfy all the Einstein equations. For IBVP, this condition implies designing constraint-preserving boundary conditions (CPBC).

3. should allow **control over the gravitational radiation** emitted in the process. For IBVP, this requirement means that the boundary conditions have to be physically meaningful.

Main Approaches

Since the main goal of a numerical computation is the extraction of the gravitational wave signal, the approaches towards reaching this goal can be classified in three categories [24]: the standard Cauchy problem, the characteristic initial value problem and the Cauchy problem for the conformal field equations.

Before going on to briefly describe each approach, it is worth mentioning that one can sensibly talk of gravitational waves if they can be neatly separated from the background. In the present understanding, this is possible if the spacetime is flat to a good approximation, in other words, the detection makes sense if the spacetime is "asymptotically flat" [2] and the detector is placed in this asymptotic region. Asymptotically flatness can be regarded as a formalization of the concept of isolated systems in general relativity and is a natural assumption for many purposes, e.g. when we want to extract the signal "far away" from the source, but not so far that cosmological effects need to be taken into account. With this in mind, the following classification

[2] for a precise mathematical definition of this concept see e.g [80] or [24].

can be viewed in the context of the numerically treatment of asymptotically flat spacetimes.

1. the **standard Cauchy problem** (e.g. ADM-like and the harmonic formulations).
   In this approach the data is prescribed on a spacelike finite domain and on a timelike boundary. In the numerical codes to date the foliation is using spacelike hypersurfaces labeled by the time coordinate, and the evolution process is a IBVP for a set of PDEs which describe the propagation of some appropriate tensor fields from one slice to another. The truncation of the solution domain raises the problem of designing appropriate boundary conditions that satisfy all of the three main requirements mentioned above. Particularly important for this approach are the strongly and the symmetric hyperbolic formulations which lead to a well-posed IVP and respectively, a well-posed IBVP. Using spacelike foliations, the issue of constraints preservation has been also attained in some cases [13, 31, 49, 50, 52, 53, 68, 70]. However, boundary conditions which are able to control the gravitational degrees of freedom are still an open question. Up to date, the only formulation based on this approach which satisfies all three requirements, has been given in [29]. It is using a tetraed formalism —instead of metric components— and there are elliptic equations that the tetrads need to satisfy at each slice. Due to its complexity, this formulation did not find its way in numerical relativity so far.

2. the **characteristic initial value problem** for the Einstein equations in Bondi- or Newman-Penrose form
   Originated in the works of Bondi [6, 7] and Penrose [63] this approach uses for the foliation of spacetime null (characteristic) hypersurfaces, and the evolution algorithm is an IVP for a system of ODEs (null hypersurfaces are integrated along light-cones). With a proper coordinate rescaling, the spacetime infinity is transformed to a finite distance.

One of the advantages of this method is the elimination of the boundary issues (it manifestly only evolves the domain of dependence). Also it is very adequate for treating radiative systems. However, in strong gravitational fields, the characteristics develop caustics and spoil the evolution One work-around is to use a hybrid method which utilizes the Cauchy evolution within a prescribed world-tube matched onto a characteristic evolution in the exterior of this tube. The outer boundary for the Cauchy evolution is rather replaced with an interface between the two types of evolutions. Although complicated, this matching procedure can and has been done and successful computations have been performed using spacetimes with certain symmetries [81].

3. the **Cauchy problem for the conformal field equations**
   Pioneered by Penrose more then forty years ago, [63] this approach developed and embraced various formulations over time [25–28, 43]. For all of them, the common ingredient is the conformal compactification procedure which amounts to transforming the spacetime infinity to a finite distance not only by a mere change in coordinates, but also by rescaling the metric using a scalar field, named conformal factor (Ω). The arena is not the physical spacetime $(\tilde{\mathcal{M}}, \tilde{g}_{ab})$ anymore but rather the unphysical manifold $(\mathcal{M}, g_{ab})$, conformally related to it via $g_{ab} = \Omega^2 \tilde{g}_{ab}$. The unphysical manifold contains the physical one within a 3-dimensional smooth boundary, $\mathscr{I}$, named "conformal infinity". The foliation is done using hyperboloidal slices [3] which intersect $\mathscr{I}$ at a finite distance and go beyond it. When $\mathscr{I}$ is used as outer boundary, the evolution process takes the form of an IVP for a set of PDEs. The boundary issues disappear because $\mathscr{I}$ acts as an ingoing nullsurface where no boundary conditions are needed. This approach benefits of all the advantages of the characteristic approach regarding the treatment of radiative systems

[3] these are some special types of spacelike hypersurfaces whose induced physical metric behaves asymptotically like a surface of constant negative curvature

(by placing the detector at $\mathscr{I}$). The main difficulties are related to the equations at scri and keeping it fixed in the computational domain.

1.2 1st order time 2nd order space hyperbolic formulations

Due to its relative simplicity (apart from boundary issues!) the standard Cauchy problem is the most used approach in the numerical relativity codes nowadays. As mentioned already, harmonic and ADM-like formulations fall into this category. While both of them use spacelike foliations and advance the solution from one to the next one, they differ in the way they specify the coordinates.

The harmonic formulation imposes the coordinates to satisfy an inhomogeneous wave equation:

$$\Box x^{\alpha} = \frac{1}{\sqrt{-g}}\partial_{\mu}\left(\sqrt{-g}g^{\mu\beta}\partial_{\beta}x^{\alpha}\right) = F^{\alpha}\,, \tag{1.2}$$

with F^{α} freely specifiable in terms of a priori functions of spacetime.
The ADM-like formulations, use the 3+1 split which regards each slice as a differential manifold of its own described by the 3-metric $g_{ij} = {}^{(4)}g_{ij}$ with $i,\ j = \overline{1,3}$. The evolution of the coordinates is specified using a set of kinematical variables, α (lapse function) and $\{\beta^i\}_{i=1}^3$ (shift vector):

$$\begin{aligned} ds^2 &= {}^{(4)}g_{\mu\nu}dx^{\mu}dx^{\nu} \\ &= -\left(\alpha^2 - g_{ij}\beta^i\beta^j\right)dt^2 + 2g_{ij}\beta^j dt dx^i + g_{ij}dx^i dx^j\beta^i \end{aligned} \tag{1.3}$$

While the Einstein equations in harmonic coordinates form a coupled system of nonlinear wave equations (e.g. see appendix A), the ADM-like formulations (e.g. see the original ADM-system in appendix B, and the BSSN system in appendix C) have a more complicated structure.

However, in all these formulations, the general form of the evolution system is:

$$\frac{d}{dt}\mathsf{v}(t,\vec{x}) = \mathsf{P}\mathsf{v}(t,\vec{x}), \quad \mathsf{v} = (\mathsf{U},\mathsf{V})^T ,$$

with $\vec{x} \in \mathbb{R}^d$, $\mathsf{U} : \mathbb{R}\mathrm{x}\mathbb{R}^d \to \mathbb{R}^p$, $\mathsf{V} : \mathbb{R}\mathrm{x}\mathbb{R}^d \to \mathbb{R}^q$ and

$$\mathsf{P} \;=\; \begin{pmatrix} A^i\partial_i + B & C \\ D^{ij}\partial_{ij} + E^i\partial_i + F & G^i\partial_i + J \end{pmatrix} . \tag{1.4}$$

Although it is always possible to reduce such systems of equations to first order in space and time, it is not clear that this is generally a good idea from a numerical point of view. While the reduction to first order in time does not increase the solution space, this *is the case* for reductions to first order in space, leading e.g. to new constraints which have to be satisfied during the evolution.

The system (1.4) is the object of study for this thesis, especially at the discrete level. The whole Chapter 2 is dedicated to the analysis of $2n$-accurate stable discretizations of its IVP.

The shifted scalar wave equation written in first-order time form is the simplest nontrivial particular example for (1.4). This case will be thoroughly investigated: the associated IVP in Chapter 3 and various possible IBVP in Chapter 4.

Hyperbolicity and well-posedness

The first order hyperbolic formulations benefit from a rich mathematical background, in the sense that various concepts such as weak, strong or symmetric hyperbolicity are very well defined and related to the well-posedness of IVP or IBVP. However as mentioned already, introducing extra variables means introducing more constraints into the system that one has to take care of.

For second order systems, the notion of hyperbolicity was until recently

less-well established. A quasilinear system second order both in space and time is called hyperbolic if the principal part of the system corresponds to the wave equation. [4] Christodoulou has recently relaxed the positivity conditions on the principal part and introduced the concept of regular hyperbolicity [20]. Both concepts can be used as criteria for well-posedness of the Cauchy problem. The Einstein equations are second form but they meet the above conditions of hyperbolicity only when written in the harmonic gauge.

For 1st order time, 2nd order space systems, the hyperbolicity was defined by various authors in various ways either by requiring that some consistencies with the first order (pseudo)reduction are attained, or some reasonable energy (Soboloev norm) estimate holds [5,31,32,59,69]. In a recent paper, [33], these approaches have been unified through some equivalence theorems.

In [14], necessary and sufficient conditions for the well-posedness of the Cauchy problem (1.4) have been established, without reffering to hyperbolicity. The approach presented in [14] for testing well-posedness has the advantage that it can be mirrored at the discrete level by the stability analysis. This mirroring is the standard procedure for doing stability analysis in the case of first order systems. For second order systems, there are additional subtleties which will be discussed in more detail in Chapter 2.

The ADM equations with prescribed lapse and shift have been the workhorse in the field for many years despite being plagued with instabilities. Latter on it has been realized that this unstable behavior is due to the weakly hyperbolic nature of the equations [46]. Various second order hyperbolic formulations based on ADM system have been derived, which have better stability properties. The BSSN formulation is one that is actually very popular among numerical relativity groups around the world.

It is has been shown that in certain gauges, BSSN can take the form of a

[4] The second order system

$$P^{\mu\nu}(u, \partial u)u_{,\mu\nu} + S(\partial u, u) = 0 \tag{1.5}$$

is called hyperbolic if $P^{\mu\nu}$ is a Lorentzian metric.

strongly or even symmetric hyperbolic system ([31], [5]).

A 1-D version of this formulation using as gauge conditions a densitized lapse and an analytic shift is considered in Chapter 5. Discretization methods for some associated IBVPs are proposed and tested numerically.

1.3 Numerical Discretization

It is time now to turn to the discrete level and discuss some relevant issues. In principle, all the unresolved problems which appear at the continuum level will propagate to the discrete, and will be joined by others.

An ill-posed problem will be very likely to be plagued by numerical instabilities and no dissipation or other numerical trick would be able to cure that. If CPBC are not very well understood for the continuum problem, a discretization using some ad-hoc numerical boundary conditions will end up giving the wrong answer to the problem at hand.

In this thesis, the issue of constraints will be left to the side and the focus will be only on the discretization of a well-posed initial (boundary) value problem for a given hyperbolic system, first order in time, first or second order in space.

There are many ways one can feed this problem into a computer, and regardless of the implementation, the main issues which appear are stability, accuracy and convergence. This concepts are related and precise definitions will be given later on, in the thesis.

However, it is worth mentioning here, that when dealing with the discretization of I(B)VP for PDEs or ODEs, the concept of stability can embrace two nuances:

1. **stability** [5] : the numerical solution remains bounded while the stepsize is refined (in the continuum limit);

[5]Other names: "zero stability" or "D-stability" for ODEs and **"Lax stability"** or "LaxRichtmyer stability" for PDEs

2. **eigenvalue stability** [6].

These two concepts are the discrete analogus of well-posedness and respectively, Lyapunov stability, at the continuum level.

In this thesis both concepts will be encountered. In Chapter 2 the IVP will be analyzed for hyperbolic PDEs from the point of view of Lax stability. It will turn out that a necessary condition for this type of stability is related to the absolute stability of the time integrator. For IBVPs both types of stability are difficult to investigate. In Chapter 4, Lax stability will be investigated only for the semidiscrete problem (only space is discretized while time is continuous) of the wave equation with certain boundary conditions and certain types of discretization methods.

In numerical experiments, both stability and time stability will be tested.

The Method of Lines Approach

All the discretizations discussed in this thesis use the method of lines approach. This is a numerical technique for solving partial differential equations discretizing in all but one dimension (time), and then integrating the semidiscrete problem as a system of ordinary differential equations (ODE).

By decoupling the space discretization from the time integration, this method presents two significant advantages over other methods:

1. the code can be implemented in a modular way, allowing separation of the methods and the routines for spatial discretizations from the ones for time integration.

2. the stability analysis is simplified, and can be reduced to imposing separate conditions for the semidiscrete problem (time is kept continuous and only space is discretized) and for the ODE integrator.

[6] Other names: "weak stability" or **"absolute stability"** for ODEs and "time stability","practical stability" or "Pstability", "strict stability" for PDEs : the numerical solution remains bounded as time goes to infinity

Space Discretization

In this thesis the space is discretized using finite difference techniques. In general, for the equations describing the evolution of the inner points of the grid, $2n$-accurate centered finite difference operators will be used. However, the analysis of the discretization of IVP (Chapters 2-3) will include also the case when some of the first order derivatives corresponding to advection terms along the shift vector ($\beta^i \partial_i \mathsf{v}$) are approximated using off-centered stencils.

Definition If the FDO corresponding to a shift term is off-centered in the same/oposite direction with shift, it is called upwinded/downwinded.

The motivation for considering off-centered stencils in the analysis of IVP for general systems (1.4) (and in particular, for the wave equation) comes from numerical relativity experiments. In practical simulations it is found that upwinding the advection terms in the Einstein equations is essential to obtaining good accuracy in the orbital motion, while naively one might expect that centered stencils would yield better results.

Regarding the implementation of boundary conditions, this can be done in various ways. Two methods will be discussed:

1. **ghost-points method:** require populating fictive points outside the evolution domain using the boundary conditions and then applying the evolution equations until the last grid point.

2. **SBP-SAT method:** the derivatives are approximated using operators which satisfy summation by parts (SBP) rules, and the boundary conditions are implemented by adding *penalty terms* to the evolution equations of some points close to the boundary, such that the SBP-property is preserved.

Time Discretization

Discretizing in space, while keeping time continuous, yields a large set of ODEs for the time dependence of the field variables at the spacial grid points. This

system can be passed to a suitable ODE integrator to advance the solution in time. The time integrators considered in this thesis are some explicit Runge-Kutta methods (4th, 6th and 8th order accurate).

1.4 Outline of the Thesis

The scope of this thesis is to develop numerical methods based on finite difference techniques, for the implementation of initial value problem and initial boundary value problem of systems which are 1st order in time and 2nd order in space.

The following outlines the reminder of this thesis.

- Chapter 2 is dedicated to the discretization of IVP for second order space, first order time hyperbolic systems, using general $2n$-order accurate finite difference operators and one-step explicit time integrators such as Runge Kutta methods.
  It opens with a section meant to introduce the FDOs and present some of their general properties, followed by a section which introduces the RK-time integrators and discusses some associated stability issues. In 2.3.1-2.3.2 the stability analysis method presented in [14] is extended from 2nd and 4th order accuracy to arbitrary order of accuracy, including the case when some first order derivatives are discretized using non-centered FDOs or dissipation is added to the system. In 2.4 the numerical errors are related to the numerical speed errors (dispersive effects that modify only the phase of the signal) and amplification factors (dissipation effects that modify only the amplitude of the signal).

- Chapter 3 applies the stability method developed in the previous chapter to the scalar wave equation on a general background. For the 1-D case the Courant limits and numerical speeds are analyzed in relation to the order of approximation, off-centering of the first order derivative, dissipation and shift. Particular attention is paid to the comparison of

centered scheme with one-point upwinded scheme and it is shown, both analytically and numerically that, in contrast with the advection equation, there are cases when off-centering can be beneficial for accuracy. At the end of this chapter, the overall accuracy and convergence of the scheme are investigated numerically in respect to the order of spacial operators and the order of the time integrator.

- Chapter 4 discusses high order discretizations methods of some possible IBVPs for the shifted 1-D wave equation, using two different approaches: ghost-points method and SBP+SAT procedure.
  In the frame of ghost-points method, numerical prescriptions will be presented for outflow, inflow and completely inflow boundary. For the 2nd order accurate scheme with outflow boundary, stability will be shown using both the energy method and the Laplace transform method.

  In 4.4 it will be shown how the SBP-SAT procedure can be applied for implementating maximally dissipative conditions for an inflow boundary. Both approaches are validated numerically in 4.5.

- Chapter 5 provides a 1-D version of the BSSN system, using spherical symmetry.
  The equations are discretized using centered finite difference operators and ghost-point method is employed for implementing various types of boundaries. Again, the theoretical prescriptions are challenged with numerical tests.

- Chapter 6 summarizes the main results of the thesis.

Chapter 2

Initial Value Problem for 2nd Order Systems in space and 1st order in time

The purpose of this chapter is to provide methods for analyzing the stability and the accuracy for the numerical discretization of initial value problem for systems that match the form (1.4). The first two sections set up the tools for such an analysis: the finite difference operators and respectively, the time integrators. Stability is discussed in section 2.3 as an "almost" mirror at the discrete level of the concept of well-posedness. The last section analyses the error of the numerical scheme in terms of dispersive and dissipative effects.

2.1 Space Discretization

As mentioned already in the introduction, in this thesis the space derivatives are discretized using finite difference operators. In this section they will be thourougly analyzed in relation to their action on periodic grid functions. This is a natural restriction when considering the discretization of initial value problem of systems with constant coefficients. Before proceeding with the

discrete operators, it is instructive to give a brief overview over the Fourier representations of periodic functions at both continuum and discrete level.

2.1.1 Representation of periodic functions

continuum

Denote by $\langle .,. \rangle$ the usual Euclidean scalar product $\langle x, y \rangle = x^\dagger y = \sum_{i=1}^{d} \bar{x}_i y_i$, where $x = (x_1, \ldots, x_d)$, $y = (y_1, \ldots, y_d)$ and $x, y \in \mathbf{C}^d$.

If u and v are 2π-periodic functions, consider the following scalar product and the corresponding norm (l^2-norm)

$$(\mathsf{u}, \mathsf{v}) = \int_0^{2\pi} dx_1 \ldots \int_0^{2\pi} dx_d \, \mathsf{u}^\dagger \mathsf{v}\,, \quad \|\mathsf{u}\|^2 = (\mathsf{u}, \mathsf{u}).$$

Then the set of functions

$$\{\frac{1}{\sqrt{2\pi}} e^{i\langle \vec{\omega}, \vec{x} \rangle},\ \vec{\omega} = (\omega_1, \omega_2, \ldots, \omega_d), \omega_r \in \mathbf{Z}\}$$

forms a orthonormal basis in the space of square integrable functions endowed with the above scalar product.

Any 2π-periodic function, $\mathsf{v}(\vec{x}) \in C^1(\mathbb{R}^n)$, can be represented in Fourier space as

$$\mathsf{v}(\vec{x}) = (2\pi)^{-d/2} \sum_{\vec{\omega}} e^{i\langle \vec{\omega}, \vec{x} \rangle} \hat{\mathsf{v}}(\vec{\omega}), \tag{2.1}$$

where $\hat{\mathsf{v}}(\vec{\omega})$ are the Fourier coefficients. They satisfy

$$\hat{\mathsf{v}}(\vec{\omega}) = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_0^{2\pi} \ldots \int_0^{2\pi} e^{-i\langle \vec{\omega}, \vec{x} \rangle} v(\vec{x}) d\vec{x}. \tag{2.2}$$

The Parseval relation says that:

$$(\mathsf{u}, \mathsf{v}) = \sum_{\vec{\omega}} \hat{\mathsf{u}}^\dagger \hat{\mathsf{v}}. \tag{2.3}$$

The associated Fourier symbols are defined by considering the action of the partial derivative operators $\partial_{i_1 i_2 \ldots}$ on the basis vectors $\frac{1}{\sqrt{2\pi}} e^{i\langle \vec{\omega}, \vec{x} \rangle}$,

$$\hat{\partial}_{i_1 i_2 \ldots} = (i\omega_{i_1})\,(i\omega_{i_2}) \ldots$$

discrete

Consider a mesh of equidistant spatial points $\underline{x}_{(i)} = (x^1_{i_1}, \ldots, x^d_{i_d})$, with $i_r = 0, \ldots, N-1$. Denote the grid spacing by h, $h = 2\pi/N$ and a vector-valued grid function by $v_{(i)} = v(\underline{x}_{(i)}, h)$. Periodicity requires $v_{(i)} = v_{\mathrm{mod}((i),N)}$. All the grid functions in this chapter are considered to be periodic. For convenience, the multiple-index (i) will be dropped from now on.

The scalar product of two grid functions and the associated norm are defined as:

$$(v, u)_h = \sum_{\underline{x}} v^\dagger u h, \quad \|v\|_h^2 = (v, v)_h$$

Then the set of the exponential grid functions

$$\{\frac{1}{\sqrt{2\pi}} e^{i\langle \underline{\omega}, \underline{x} \rangle},\ \underline{\omega} = (\omega_1, \omega_2, \ldots, \omega_d), \omega_r = -N/2+1, \ldots, N/2\} \quad \text{for } N \text{ even}$$

or

$$\{\frac{1}{\sqrt{2\pi}} e^{i\langle \underline{\omega}, \underline{x} \rangle},\ \underline{\omega} = (\omega_1, \omega_2, \ldots, \omega_d), \omega_r = -(N-1)/2+1, \ldots, (N-1)/2\} \quad \text{for } N \text{ odd}$$

represents a orthonormal basis in the space of periodic grid functions endowed with the above scalar product. Denote with $\underline{\xi} = h\underline{\omega} = (\xi_1, \xi_2, \ldots, \xi_d)$ the vector of grid frequencies. Then $\xi_r = -\pi + 2\pi/N, -\pi + 4\pi/N, \ldots, \pi$ for N even and $\xi_r = -\pi + \pi/N, \ldots, \pi - \pi/N$ for N odd. Notice that when N is odd the frequencies $\pm\pi$ are never present in a finite grid, but only in the limit $N \to \infty$.

A grid function $v(\underline{x}, h)$ can be decomposed in the following way:

$$v(\underline{x}, h) = \frac{1}{(2\pi)^{\frac{d}{2}}} \sum_{\underline{\omega}} e^{i\langle \underline{\omega}, \underline{x} \rangle} \hat{v}(\underline{\omega}, \underline{\xi}) \,, \tag{2.4}$$

The quantities $\hat{v}(\underline{\omega}, \underline{\xi})$ represent the *discrete Fourier coefficients*. They satisfy

$$\hat{v}(\underline{\omega}, \underline{\xi}) = \frac{1}{(2\pi)^{\frac{d}{2}}} \sum_{\underline{x}} e^{-i\langle \underline{\omega}, \underline{x} \rangle} v(\underline{x}, h). \tag{2.5}$$

The discrete Parseval relation is

$$(v, u)_h = \sum_{\underline{\omega}} \hat{v}^\dagger \hat{u}. \tag{2.6}$$

Let S_j^k be the shift operator by k points in the j-direction

$$S_j^k v(t, \underline{x}) = v(t, \underline{x'}) \quad \text{with } \underline{x'} = (x_{i_1}, \ldots, x_j + kh, \ldots, x_{i_d}), \tag{2.7}$$

then $S_j \equiv S_j^1$ and $S_j^0 = I$ with I the identity operator. The shift operator S_j^k acting on the basis $e^{i\langle \underline{\omega}, \underline{x} \rangle}$ gives

$$S_j^k e^{\langle \underline{\omega}, \underline{x} \rangle} = \hat{S}_j^k(\xi_j) e^{\langle \underline{\omega}, \underline{x} \rangle}, \text{ with } \hat{S}_j^k(\xi_j) = e^{ik\xi_j} \,. \tag{2.8}$$

A finite difference operator D_j corresponding to the mth-order derivative in the j-direction, consists of a linear combination of shift operators of the type

$$D_j = h^{-m} \sum_k a_k S_j^k$$

Its *Fourier symbol* $\hat{D}_j$ is a function of the frequency ξ_j up to a factor h^{-m} and satisfies

$$D_j e^{\langle \underline{\omega}, \underline{x} \rangle} = \hat{D}_j(\xi_j, h) e^{\langle \underline{\omega}, \underline{x} \rangle} \text{ with } \hat{D}_j(\xi_j, h) = h^{-m} \sum_k a_k e^{ik\xi_j} \,. \tag{2.9}$$

In a similar way one can introduce the Fourier symbols of mixed derivatives (in $(i_1, i_2, \dots)$-direction), $\hat{D}_{i_1,i_2,\dots}(\xi_{i_1}, \xi_{i_2}, \dots, h)$.

A finite difference operator $D_{i_1,i_2,\dots}$ corresponding to the mth-order derivative in the $i_1, i_2, \dots$-directions, consists of a linear combination of shift operators of the type

$$D_{i_1,i_2,\dots} = h^{-m} \sum_{k_1,k_2,\dots} a_{k_1,k_2,\dots} S_{i_1}^{k_1} S_{i_2}^{k_2} \dots$$

and its *Fourier symbol* $\hat{D}_{i_1,i_2,\dots}$ is a function of the frequencies $\{\xi_{i_1}, \xi_{i_2}, \dots\}$ up to the factor h^{-m} and satisfies

$$\begin{aligned} D_{i_1,i_2,\dots} e^{\langle \underline{\omega}, \underline{x} \rangle} &= \hat{D}_{i_1,i_2,\dots}(\xi_{i_1}, \xi_{i_2}, \dots, h) e^{\langle \underline{\omega}, \underline{x} \rangle} \\ \hat{D}_{i_1,i_2,\dots}(\xi_{i_1}, \xi_{i_2}, \dots, h) &= h^{-m} \sum_{k_1,k_2,\dots} a_{k_1,k_2,\dots} e^{ik_1 i_1} e^{ik_2 i_2} \dots \end{aligned} \tag{2.10}$$

The weights a_k used in the construction of FDO will depend on the order of the derivative, on the order of the approximation and on the points used to compute the derivatives. The next section will derive explicitly these coefficients for the discrete $2n$-accurate first and second order derivative, constructed with $2n + 1$ points.

2.1.2 High Order Finite Difference Operators

We first restrict ourselves to one space-dimension. Accoding to [23], the finite difference operator using $2n + 1$ equidistant points which approximates the derivative of order m can be constructed starting from the Taylor expansion of the function

$$f^{m,n,s,\epsilon}(x) = x^{n-\epsilon s} (\log x)^m \tag{2.11}$$

around the point $x_0 = 1$ up to the term $(x - x_0)^{2n}$. Denote by $s \in \{0, 1, \dots, n\}$ the offset of these points from symmetry with respect to the center, ($s = 0$ for a centered operator) and by ϵ the direction of off-centering ($\epsilon = 1$ for

off-centering to the right, $\epsilon = -1$ for off-centering to the left).[1]

The coefficients of x in this expansion, $\tilde{f}_{m,n,s,\epsilon,k}$, will be the weights of the points which enter the construction of the FDO:

$$\begin{aligned} f^{m,n,s,\epsilon}(x) &= \sum_{k=0}^{2n} \frac{(x-1)^k}{k!} \left.\frac{d^k f^{m,n,s,\epsilon}(x)}{d^k x}\right|_{x=1} + O((x-1)^{2n+1}), \\ &= \sum_{k=0}^{2n} \tilde{f}_{m,n,s,\epsilon,k} x^k + O((x-1)^{2n+1}) \,. \end{aligned} \tag{2.12}$$

Then a general finite difference operator can be written as a linear combination of shift operators S^k:

$$D^{m,n,s,\epsilon} = \sum_{k=-n+\epsilon s}^{n+\epsilon s} \tilde{f}_{m,n,s,\epsilon,k} S^k \,. \tag{2.13}$$

In general, the accuracy of this operator will be $2n+1-m$. In this thesis, higher than second order derivatives in space will not be considered. The focus will be on the cases where a centered FDO is used for the second derivative, and a not necessarily centered FDO for the first derivative. The explicit expressions for these operators are:

$$D^{(1,n,s,\epsilon)} = \frac{1}{h} \sum_{j=-n+\epsilon s}^{n+\epsilon s} \alpha_{n,s,\epsilon,j} S^j, \tag{2.14}$$

$$D^{(1,n)} \equiv D^{(1,n,0,0)} = \frac{1}{h} \sum_{j=1}^{n} \frac{j\beta_{n,j}}{2} \left(S^j - S^{-j}\right), \tag{2.15}$$

$$D^{(2,n)} \equiv D^{(2,n,0,0)} = \frac{1}{h^2} \sum_{j=0}^{n} \beta_{n,j} \left(S^j + S^{-j}\right), \tag{2.16}$$

[1] Though one can simplify the notation by dropping ϵ and considering $s \in \{-n, \ldots, n\}$, it will later turn out useful to separate the sign of s and its absolute value.

where

$$\alpha_{n,s,\epsilon,j} = \begin{cases} \frac{(-1)^{j+1}(n+s)!(n-s)!}{j(n+\epsilon s-j)!(n-\epsilon s+j)!}, & j \neq 0 \\ \\ \epsilon\left(H_{n-s} - H_{n+s}\right), & j = 0 \end{cases} \tag{2.17}$$

and

$$\beta_{n,j} = \begin{cases} 2(-1)^{j+1}\frac{(n!)^2}{j^2(n+j)!(n-j)!} & j \geq 1 \\ \\ -\sum_{j=1}^{n}\beta_{n,j} & j = 0\,. \end{cases} \tag{2.18}$$

In the relation (2.17), $H_n = \sum_{i=1}^{n}\frac{1}{i}$ is the harmonic number. Note that $j\beta_{n,j} = 2\alpha_{n,0,0,j}$ for $j \geq 1$.

Truncation Errors

The (leading order) truncation errors for the discrete operators under study are defined by the differences:

$$\left.\frac{dv}{dx}\right|_{x_0} - D^{(1,n,s,\epsilon)}v_0 = T^{(1,n,s,\epsilon)}\left.\frac{d^{2n+1}v}{dx^{2n+1}}\right|_{x_0} h^{2n} + O(h^{2n+1})\,, \tag{2.19}$$

$$\left.\frac{d^2v}{d^2x}\right|_{x_0} - D^{(2,n)} = T^{(2,n)}\left.\frac{d^{2n+2}v}{dx^{2n+2}}\right|_{x_0} h^{2n} + O(h^{2n+2})\,. \tag{2.20}$$

One can show that $T^{(1,n,s,1)} = T^{(1,n,s,-1)} = T^{(1,n,s)}$, where,

$$\begin{aligned} T^{(1,n,s)} &= (-1)^{s+n}\frac{(n+s)!(n-s)!}{(2n+1)!}\,, \\ T^{(2,n)} &= (-1)^{n}\frac{2(n!)^2}{(2n+2)!}\,. \end{aligned} \tag{2.21}$$

Also for $s > 0$, the inequality

$$\left|T^{(1,n,0)}\right| < \left|T^{(1,n,s)}\right|, \tag{2.22}$$

$\left\lvert T^{(1,n,s)}\right\rvert$	$n=1$	$n=2$	$n=3$	$n=4$
$s=0$	$\frac{1}{6}$	$\frac{1}{30}$	$\frac{1}{140}$	$\frac{1}{630}$
$s=1$	$\frac{1}{3}$	$\frac{1}{20}$	$\frac{1}{105}$	$\frac{1}{504}$

Table 2.1: Leading order truncation errors for the first order discrete derivative $\left\lvert T^{(1,n,s)}\right\rvert$.

holds. This means that the centered FDO has the smallest leading order truncation error (see also table 2.1).

Note that the centered FDO constructed with $\frac{1}{2}\left(D^{(1,n,s,1)}+D^{(1,n,s,-1)}\right)$ (using $2(n+s)+1$ points), has the same truncation error as $D^{(1,n,s,\pm 1)}$, and thus a larger truncation error than $D^{(1,n)}$, though it is constructed using more points.

Define now the elementary dimensionless finite difference operators

$$\delta_0 = \frac{h}{2}\left(D_+ + D_-\right), \tag{2.23}$$

$$p = h(D_+ - D_-) = h^2 D_+ D_-, \tag{2.24}$$

where $D_+ v_i = (v_{i+1} - v_i)/h$ and $D_- v_i = (v_i - v_{i-1})/h$. Direct but lengthy calculations starting from the definitions (2.15) and (2.16) allows us to rewrite the finite difference operators in some more convenient forms. The results are stated in the following two lemmas:

Lemma 2.1.1 *In one dimension, centered FDOs of accuracy* $2n$ *satisfy:*

$$D^{(1,n)} = \frac{1}{h}\delta_0\left(1+\sum_{k=1}^{n-1} c_k p^k\right), \tag{2.25}$$

$$D^{(2,n)} = \frac{1}{h^2}p\left(1+\sum_{k=1}^{n-1} d_k p^k\right), \tag{2.26}$$

where the coefficients c_k *and* d_k *do not depend on* n*,*

$$c_k = (-1)^k\frac{(k!)^2}{(2k+1)!}, \qquad d_k = \frac{c_k}{k+1}. \tag{2.27}$$

For the rest $R^{(n)} = (D^{(1,n)})^2 - D^{(2,n)}$ *the identity*

$$R^{(n)} = \frac{1}{h^2}\frac{nc_{n-1}}{2}p^{n+1}\sum_{k=0}^{n-1}\frac{c_k}{n+1+k}p^k \tag{2.28}$$

holds.

Lemma 2.1.2 *The centered FDO* $2n$*-accurate* $D^{(1,n)}$ *can be written as:*

$$D^{(1,n)} = \frac{1}{h}\sum_{k=1}^{n}\frac{(-1)^{k+1}n!(2n-k)!}{k(2n)!(n-k)!}\left[(hD_+)^k + (-1)^{k+1}(hD_-)^k\right]. \tag{2.29}$$

The construction of FDOs in n dimensions is straightforward and is done by associating an index specifying the direction to all the FDOs defined above. The exception will be the second order operator defined as: $D_{ij}^{(2,n)} = D_i^{(1,n)}D_j^{(1,n)}$ for $i \neq j$ and $D_{ii}^{(2,n)} = D_i^{(2,n)}$.

2.1.3 Artificial Dissipation Operator

In order to achieve numerical stability for problems that go beyond the linear constant coefficient case, it is common practice to add artificial dissipation to

the right-hand sides of the time evolution equations as

$$\partial_t u \to \partial_t u + \mathcal{D}u. \tag{2.30}$$

This is usually done in a way that the dissipation term converges away fast enough so as not to change the convergence order of the scheme. In this thesis, the dissipation used for a $2m-2$ accurate scheme, will be given by the Kreiss-Oliger dissipation operator $\mathcal{D}^{(2m)}$ of order $2m$ [36],

$$\mathcal{D} \to \mathcal{D}^{(2m)} = -\frac{(-1)^m}{2^{2m}} h^{2m-1} \sum_{j=1}^{d} \sigma_j (D_{+j})^m (D_{-j})^m, \tag{2.31}$$

where the parameters $\sigma_j \geq 0$ regulate the strength of the dissipation.

The formula for the one-dimensional case is:

$$\mathcal{D}^{(2m)} = -\sigma \frac{(-1)^m}{2^{2m}} h^{2m-1} (D_+)^m (D_-)^m, \tag{2.32}$$

2.1.4 Fourier Symbols: Properties I

The FDOs are now analyzed in Fourier space, by considering their associated Fourier Symbols, formally introduced in 2.1.1.

Using the relation (2.9), the elementary discrete operators (2.24) have the following Fourier representations:

$$h\hat{D}_{+j}(\xi_j) = e^{i\xi_j} - 1, \qquad h\hat{D}_{-j}(\xi_j) = 1 - e^{-i\xi_j}, \tag{2.33}$$

$$\hat{\delta}_{0j}(\xi_j) = i\check{\delta}(\xi_j), \text{ where } \check{\delta}(\xi) \equiv \sin\xi, \tag{2.34}$$

$$\hat{p}_j(\xi_j) = -\check{\Omega}^2(\xi_j), \text{ where } \check{\Omega}(\xi) \equiv 2\sin\frac{\xi}{2}. \tag{2.35}$$

It is convenient to introduce the shorthand expression

$$\check{\Omega}_0 \equiv h\Omega_0 = \sqrt{\sum_{i=1}^{d} \left|\check{\Omega}(\xi_i)\right|^2}. \tag{2.36}$$

The symbols for the first and second order derivative operators are straightforwardly computed using (2.25)-(2.26),

$$h\hat{D}_i^{(1,n)}(\xi_i) = i\check{d}^{(1,n)}(\xi_i)\,, \tag{2.37}$$

$$h\hat{D}_{ij}^{(2,n)}(\xi_i,\xi_j) = \begin{cases} -\check{d}^{(1,n)}(\xi_i)\check{d}^{(1,n)}(\xi_j) & i \neq j \\ \\ -\check{d}^{(2,n)}(\xi_i) & i = j \end{cases}, \tag{2.38}$$

where

$$\check{d}^{(1,n)} \equiv \check{\delta}\sum_{k=0}^{n-1}|c_k|\,\check{\Omega}^{2k} \in \mathbb{R}\,, \tag{2.39}$$

$$\check{d}^{(2,n)} \equiv \check{\Omega}^2\sum_{k=0}^{n-1}|d_k|\,\check{\Omega}^{2k} > 0\,. \tag{2.40}$$

Starting from definition (2.14), and going to Fourier space, one can also compute the corresponding symbol for $\hat{D}_j^{(1,n,s,\epsilon)}$,

$$h\hat{D}_j^{(1,n,s,\epsilon)}(\xi_j) = \epsilon_j\check{\mathbf{d}}^{(1,n,s)}(\xi_j) + i\hat{d}^{(1,n,s)}(\xi_j)\,, \tag{2.41}$$

where $\epsilon_j \in \{-1,1\}$ gives the sense of off-centering for the derivative in the j-direction, $\check{\mathbf{d}}^{(1,n,s,\epsilon)} \equiv \epsilon\check{\mathbf{d}}^{(1,n,s)}$ and $\hat{d}^{(1,n,s)}$ represent the real and imaginary parts of the operator ($\check{\mathbf{d}}^{(1,n,s)}$, $\check{d}^{(1,n,s)} \in \mathbb{R}$). These quantities satisfy $\check{\mathbf{d}}^{(1,n,0)} = 0$, $\check{d}^{(1,n,0)} = \check{d}^{(1,n)}$ and, for $s \geq 1$:

$$\check{\mathbf{d}}^{(1,n,s)} = (-1)^s\frac{1}{2C_{2n}^{n+s}}\check{\Omega}^{2n+2}\sum_{k=0}^{s-1}\frac{(-1)^kC_{s+k}^{2k+1}}{(n+1+k)}\check{\Omega}^{2k}\,, \tag{2.42}$$

$$\check{d}^{(1,n,s)} = \check{d}^{(1,n)} + \check{\delta}\check{\Omega}^{2n}\sum_{k=0}^{s-1}(-1)^k\left[\sum_{j=k}^{s-1}(-1)^j\frac{C_{j+k}^{2k}}{(n-j)C_{2n}^{n+j}}\right]\check{\Omega}^{2k}\,. \tag{2.43}$$

Observations

- For s=1:

$$\begin{aligned} \check{\mathbf{d}}^{(1,n,1)} &= -\frac{n!(n-1)!}{2(2n)!}\check{\Omega}^{2n+2} \leq 0,\ \forall\check{\Omega} \in (-2,2]\,, \\ \check{d}^{(1,n,1)} &= \check{d}^{(1,n)} + \check{\delta}\frac{1}{2}\left|c_{n-1}\right|\check{\Omega}^{2n}\,. \end{aligned} \tag{2.44}$$

- $\check{\mathbf{d}}^{(1,n,s)}$ is an even function in $\check{\Omega}$, while $\check{d}^{(1,n,s)}$ is an odd function in $\check{\Omega}$.

The Fourier symbol of the dissipation operator defined in (2.31) is easy to write down in terms of $\check{\Omega}$:

$$h\hat{\mathcal{D}}^{(2m)}(\xi_1,\ldots,\xi_d) = -\frac{1}{2^{2m}}\sum_{j=1}^{d}\sigma_j\check{\Omega}^{2m}(\xi_j), \tag{2.45}$$

Convention For any function $\check{f} \in \{\check{\delta}, \check{\Omega},\ \check{d}^{(1,n,s)},\ \check{\mathbf{d}}^{(1,n,s)},\ \check{d}^{(2,n)}\}$ we will use the shorthand

$$\check{f}_i \equiv \hat{f}(\xi_i),\ i = \overline{1,d}\,.$$

2.1.5 Fourier Symbols: Properties II

In the following some further properties of the Fourier symbols are presented. These will be useful in the analysis of stability, Courant limits and numerical speeds.

1. From (2.40) it is straightforward to check that $\check{d}^{(2,n)}$ satisfies the inequalities

$$C_n^{-1}\check{\Omega}^2 \leq \check{\Omega}^2 \leq \check{d}^{(2,n)} \leq C_n\check{\Omega}^2, \quad \forall\check{\Omega} \in (-2,2]\,, \tag{2.46}$$

where

$$C_n \equiv 1 + \sum_{k=1}^{n-1}\left|d_k\right|4^k \geq 1\,. \tag{2.47}$$

2. **The $D^{(2,n)}$-norm**

The inequality (2.46) tells that the norm D_+ is equivalent [2] with the norm $D^{(2,n)}$ defined as

$$\|v\|^2_{h,D^{(2,n)}} = \frac{1}{h^2}\sum_{i=1}^{d}\sum_{k=1}^{n}|d_{k-1}|\left\|(hD_{+i})^k u\right\|^2_h + \|v\|^2_h \,. \tag{2.48}$$

This norm has been used to prove strong stability of the initial boundary value problem for the wave equation in [16] for the second and fourth order accuracy case.

3. **Rest between the second derivative and the square of the first derivative**.
   If $\check{r}^{(n)} \equiv h\hat{R}^n$ then

$$\check{r}^{(n)} \equiv \check{d}^{(2,n)} - (\check{d}^{(1,n)})^2 = \frac{n\,|c_{n-1}|}{2}\check{\Omega}^{2(n+1)}\sum_{k=0}^{n-1}\frac{|c_k|}{(n+k+1)}\check{\Omega}^{2k} > 0\,. \tag{2.49}$$

4. **derivatives of the $\check{}$–functions** with respect to ξ

$$\frac{d}{d\xi}\check{d}^{(2,n)} = 2\check{d}^{(1,n)}\,, \tag{2.50}$$

$$\frac{d}{d\xi}\check{r}^{(n)} = 2\frac{(n!)^2}{(2n)!}\check{\Omega}^{2n}\check{d}^{(1,n)} \tag{2.51}$$

$$\frac{d}{d\xi}\check{d}^{(1,n)} = 1 - \frac{(n!)^2}{(2n)!}\check{\Omega}^{2n}\,. \tag{2.52}$$

$$\frac{d}{d\xi}\check{\mathbf{d}}^{(1,n,s)} = \frac{(-1)^s}{C^{n-s}_{2n}}\sin(s\,\xi)\check{\Omega}^{2n} \tag{2.53}$$

$$\frac{d}{d\xi}\check{d}^{(1,n,s)} = 1 - \frac{(-1)^s}{C^{n-s}_{2n}}\cos(s\,\xi)\check{\Omega}^{2n} \tag{2.54}$$

[2] Two norms $\|v\|_{H_1}$ and $\|v\|_{H_2}$ are called equivalent if there exists a constant K such that $K^{-1}\|v\|_{H_2} \le \|v\|_{H_1} \le K\|v\|_{H_2}$, $\forall v$.

5. **Integral expressions**

$$\check{\mathbf{d}}^{(1,n,s)}(\xi) = \frac{(-1)^s}{C_{2n}^{n-s}} \int_0^{\xi} dx \sin(s\,x) \left(2\sin\frac{x}{2}\right)^{2n} \tag{2.55}$$

$$\check{d}^{(1,n,s)}(\xi) = \xi - \frac{(-1)^s}{C_{2n}^{n-s}} \int_0^{\xi} dx \cos(s\,x) \left(2\sin\frac{x}{2}\right)^{2n} \tag{2.56}$$

$$\check{d}^{(1,n)}(\xi) = \xi - \frac{(n!)^2}{(2n)!} \int_0^{\xi} dx \left(2\sin\frac{x}{2}\right)^{2n} \tag{2.57}$$

$$\check{d}^{(2,n)}(\xi) = \xi^2 - 2\frac{(n!)^2}{(2n)!} \int_0^{\xi} dy \int_0^{y} dx \left(2\sin\frac{x}{2}\right)^{2n} dx \tag{2.58}$$

The Fourier symbol $\hat{D}_j^{(1,n,s,\epsilon)}$ satisfies:

$$h\hat{D}_j^{(1,n,s,\epsilon)}(\xi_j) = i\xi_j - \frac{(-1)^s}{C_{2n}^{n-s}} \int_0^{\xi_j} dx \left(2\sin\frac{x}{2}\right)^{2n} e^{i\epsilon s x} \tag{2.59}$$

6. **Roots**:
The function $\check{\mathbf{d}}^{(1,n,s)}$ has $s-1$ roots in $(0,2]$ for $s \geq 1$, this means that only for $s=1$ the sign of $\check{\mathbf{d}}^{(1,n,s)}$ is constant (negative) for all frequencies. For $s>1$ there are ranges in frequency for which real part of $\check{\mathbf{d}}^{(1,n,s,1)}$ is positive.

7. **Recurrence relations**:

$$\check{d}^{(1,n+1)} = \check{d}^{(1,n)} + \check{\delta}\,|c_n|\,\check{\Omega}^{2n}\,. \tag{2.60}$$

$$\check{d}^{(2,n+1)} = \check{d}^{(2,n)} + |d_n|\,\check{\Omega}^{2n+2}\,. \tag{2.61}$$

$$\begin{aligned}\check{d}^{(1,n,s)} &= \check{d}^{(1,n,s-1)} - \frac{(-1)^s}{(n+s)C_{2n}^{n+s}}\check{\delta}\check{\Omega}^{2n}\sum_{k=0}^{s-1}(-1)^k C_{s+k-1}^{2k}\check{\Omega}^{2k} \\ &= \check{d}^{(1,n,s-1)} - \frac{(-1)^s}{(n+s)C_{2n}^{n+s}}\cos\left((2s-1)\arcsin\left(\frac{\check{\Omega}}{2}\right)\right)\check{\Omega}^{2n+1}\,.\end{aligned} \tag{2.62}$$

8. **Small frequency behavior**:

$$\begin{aligned}
\check{d}^{(1,n,s)} &\simeq \xi\left[1-(-1)^n T^{(1,n,s)}\xi^{2n}\right] = \xi\left[1-(-1)^s\frac{(n+s)!(n-s)!}{(2n+1)!}\xi^{2n}\right] \\
\check{\mathbf{d}}^{(1,n,s)} &\simeq (-1)^n s\frac{2n+1}{2n+2}T^{(1,n,s)}\xi^{2n+2} = (-1)^s\frac{s(n+s)!(n-s)!}{2(n+1)(2n)!}\xi^{2n+2} \\
\sqrt{\check{d}^{(2,n)}} &\simeq \xi\left[1-\frac{(-1)^n}{2}T^{(2,n)}\xi^{2n}\right] = \xi\left[1-\frac{(n!)^2}{(2n+2)!}\xi^{2n}\right] \qquad (2.63)
\end{aligned}$$

9. **inequalities**

$$0 \le \check{d}^{(2,n)} \le \check{d}^{(2,n+1)} \le \xi^2\,, \qquad \forall\xi\in(-\pi,\pi] \qquad (2.64)$$

$$1 \ge \frac{\check{r}^{(n)}}{\check{d}^{(2,n)}} \ge \frac{\check{r}^{(n+1)}}{\check{d}^{(2,n+1)}}, \qquad \forall\xi\in(-\pi,\pi]. \qquad (2.65)$$

10. **Limits $n\to\infty$:**
It is straightforward to show by Taylor expansion that the following limits exist for all $\check{\Omega}\in(-2,2]$,

$$\begin{aligned}
\lim_{n\to\infty}\check{d}^{(1,n,s)} &= 2\arcsin\frac{\check{\Omega}}{2} = \xi\,, \qquad (2.66)\\
\lim_{n\to\infty}\check{d}^{(2,n)} &= \left(2\arcsin\frac{\check{\Omega}}{2}\right)^2 = \xi^2\,, \qquad (2.67)\\
\lim_{n\to\infty}\check{r}^{(n)} &= 0, \quad \forall\,\check{\Omega}_j\in(-\pi,\pi)\,. \qquad (2.68)
\end{aligned}$$

11. **Scaling of the error with the order of the approximation**
Fig. 2.1 shows the Fourier symbols $\check{d}^{(1,n)}, \check{d}^{(2,n)}$ and $\check{r}^{(n)}$ as functions of the frequency ξ for different orders of accuracy. For increasing approximation order, the second derivative becomes more accurate for all frequencies, while the first derivative does not converge for $\xi=\pi$.

12. **Scaling of the error with the degree of off-centering**
In contrast with centered FDOs where the error scales with the approximation order for all frequencies $(-\pi,\pi)$, for non-centered FDOs this is

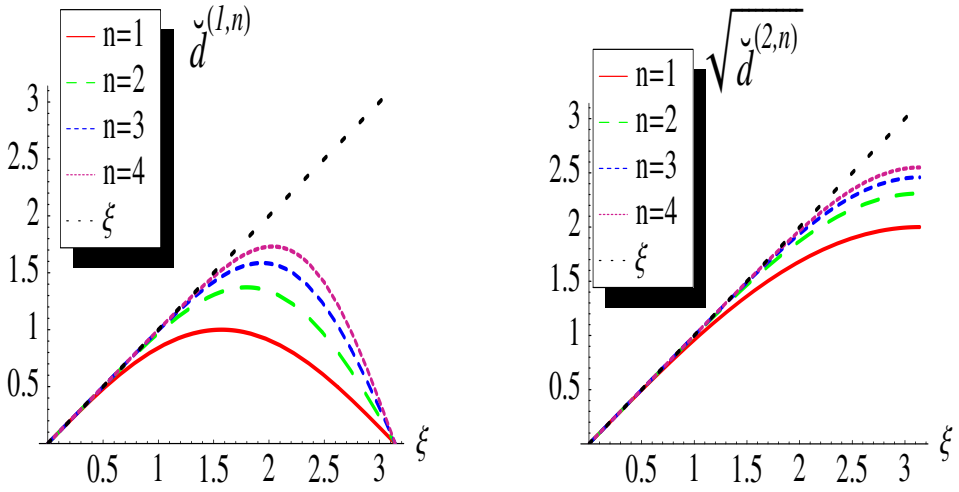


Figure 2.1: **First and second centered discrete derivatives** The figure shows the functions $\check{d}^{(1,n)}$ (left) and $\sqrt{\check{d}^{(2,n)}}$ (right) versus frequency ξ, for different orders of accuracy. Note that for increasing order of the approximation, the second derivative becomes more accurate for all frequencies, while the first derivative does not converge for $\xi = \pi$.

true only at small frequencies.

It is also interesting to see the scaling of the error for $\check{d}^{(1,n,s)}$ with off-centering, at fixed order of approximation. Fig. 2.3 shows this dependence when the order is $n = 1, 2, 3, 4$.

For $s = 1$ one can show that for each order, there is a frequency $\xi^{(n)}$ such that for all $\xi \geq \xi^{(n)}$ the error for $\check{d}^{(1,n,1)}$ is smaller then the error of $\check{d}^{(1,n)}$. This frequency can be computed numerically. For $n = \overline{1,4}$ it is $\xi^{(1)} = 1.3787$, $\xi^{(2)} = 1.0036$, $\xi^{(3)} = 0.8234$, $\xi^{(4)} = 0.7136$.

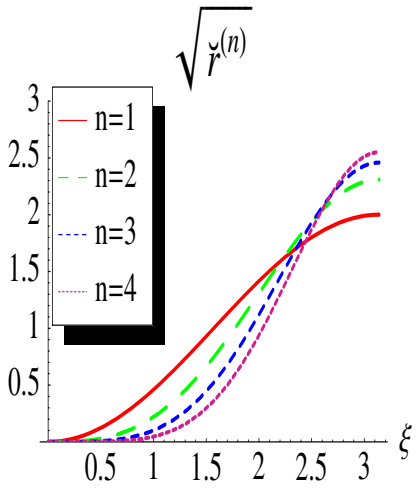


Figure 2.2: **The rest operator** $R^{(n)} = (D^{(1,n)})^2 - D^{(2,n)}$ The figure shows the corresponding function $\check{r}^{(n)}$ versus frequency ξ, for different orders of accuracy. The rest converges to zero for all frequencies apart from $\xi = \pi$.

2.2 Time Integration using Runge-Kutta Methods

Runge-Kutta algorithms are an important family of implicit and explicit iterative methods for the approximation of solutions of ordinary differential equations, developed around 1900 by the German mathematicians C.Runge and M.W. Kutta.
This section will sketch their construction and discuss some stability issues which will be needed when analyzing the discretizations of PDEs.

2.2.1 Construction

Let the initial value problem be specified as follows:

$$\begin{aligned} \frac{d\mathsf{y}}{dt} &= f(t,\mathsf{y}), \quad t \geq t_0 \\ \mathsf{y}(t_0) &= \mathsf{y}_0 \end{aligned} \tag{2.69}$$

The Runge-Kutta methods compute approximations y_n to $\mathsf{y}_n = \mathsf{y}(t_n)$ with initial values $y_0 = \mathsf{y}_0$, where $t_n = t_0 + nk$, $n \in \mathbb{N}$.

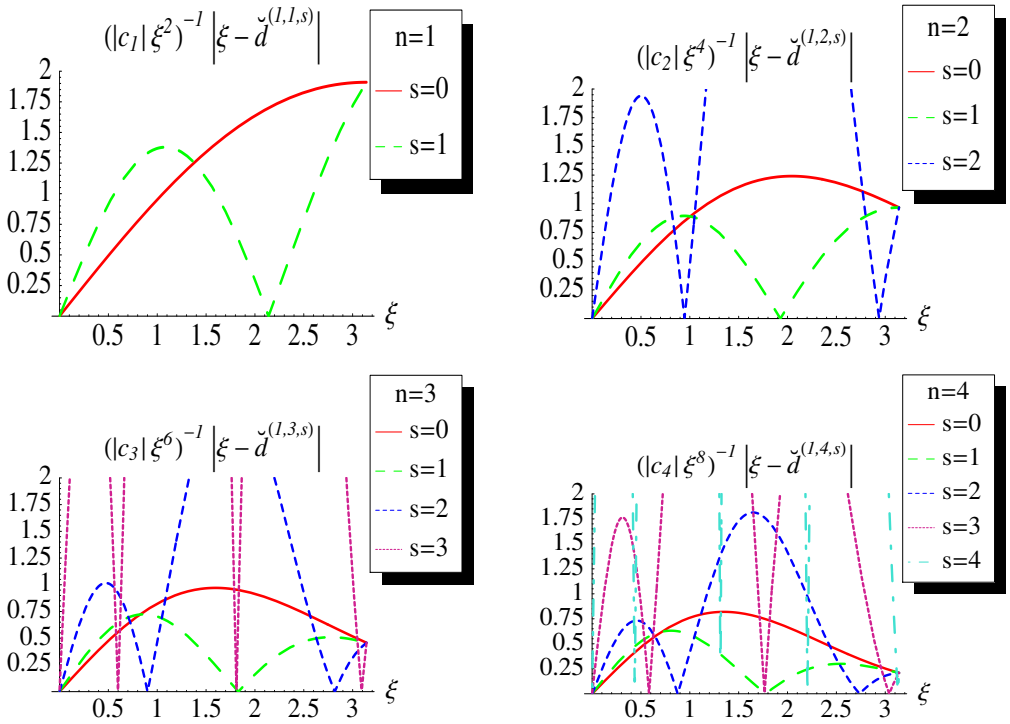


Figure 2.3: **First discrete derivative for different off-centerings** Each of these figures shows the error in absolute value for the symbol $\breve{d}^{(1,n,s)}$, downscaled with $|c_n|\,\xi^{2n}$, for various off-centerings, s, at fixed order of approximation, $2n$ ($n = 1, 2, 3, 4$). In the regime of small frequencies, the curves are actually straight lines with the slope given by $(n+s)!(n-s)!/(n!)^2$, according to the formula (2.63). In this region, the higher the off-centering, the larger the error. At larger frequencies this behavior changes. For each s, there are exactly s frequencies in $(0, \pi)$ where the error cancels. However, $s \geq 2$, there are large intervals where the error overcomes by far the error when $s = 0$. For $s = 1$, note that while at small frequencies, the error is slightly larger than for $s = 0$, for each order $2n$, there is a frequency, $\xi^{(n)}$, beyond which, the error is smaller than for the case $s = 0$.

A Runge-Kutta method is given by the following algorithm:

$$y_{n+1} = y_n + k\sum_{i=1}^{q} b_i k_i \tag{2.70}$$

$$k_i = f(t_n + c_i k, y_n + \sum_{j=1}^{N_i} a_{ij} k_j), \quad i = 1, \ldots, q \tag{2.71}$$

For explicit schemes, $N_i = i - 1$. For implicit schemes $N_i = q$. Note that for an explicit method the relation (2.71) can be solved for each k_i in turn, while for an implicit method, the evaluation of k_i involves the solution of a nonlinear (problem-dependent) system at each time step.

Despite this difficulty, implicit methods present certain advantages over the explicit schemes, such as high (possible unconditional) stability and higher order accuracy with fewer steps.

In this thesis I will consider only explicit Runge-Kutta methods.

An explicit Runge-Kutta method is specified by the number of stages, q, the nodes, $c = \{c_i\}_{i=1}^{s}$ the internal weights, $a = \{a_{ij}\}_{j=1,i=2}^{i-1,q}$, and the external weights, $b = \{b_i\}_{i=1}^{s}$. For convenience, these coefficients are usually displayed in an Butcher tableau [10] of the form (2.72).

General explicit Runge-Kutta

c_1							
c_2	a_{21}						
c_3	a_{31}	a_{32}					
.	.		.				
.	.			.			
.	.				.		
c_s	a_{s1}	a_{s2}	.	.	.	$a_{s,s-1}$	
	b_1	b_2	.	.	.	b_{s-1}	b_s

(2.72)

4th order Runge-Kutta

0				
1/2	1/2			
1/2	0	1/2		
1	0	0	1	
	1/6	1/3	1/3	1/6

(2.73)

The tableau (2.73) shows the parameters for classical 4th order Runge-Kutta method.
The number of stages and the parameters a, b, c are determined so that certain accuracy and consistency requirements are satisfied.

The method is said to be

1. *p-order accurate* if p is the largest integer for which

$$\mathsf{y}_{n+1} - \mathsf{y}_n = y_{n+1} - y_n + O(k^{p+1}) \tag{2.74}$$

2. *convergent* if

$$\lim_{\substack{k\to 0 \\ nk=t-t_0}} y_n = \mathsf{y}_n \tag{2.75}$$

3. *consistent* with the initial value problem

$$\lim_{\substack{k\to 0 \\ nk=t-t_0}} \frac{y_{n+1} - y_n}{k} = f(t_n, \mathsf{y}_n) \tag{2.76}$$

For Runge-Kutta methods it has been shown in [40] that consistency is a necessary and sufficient condition for convergence. [3]

In order to derive a Runge-Kutta algorithm of accuracy p, one starts from the Taylor expansion:

$$\mathsf{y}_{n+1} = \mathsf{y}_n + k\mathsf{y}'_n + \frac{1}{2}k^2\mathsf{y}''_n + \cdots + \frac{1}{p!}k^p\mathsf{y}^{(p)}_n + O(k^{p+1}) \tag{2.77}$$

Using 2.69 and the notations $\left(\frac{d^i f}{dt^i}\right)_n = \left(\frac{d^i f}{dt^i}\right)(t_n, y_n)$ this can be written as

$$\mathsf{y}_{n+1} = \mathsf{y}_n + kf_n + \frac{1}{2}k^2\left(\frac{df}{dt}\right)_n + \cdots + \frac{1}{p!}k^p\left(\frac{d^{p-1}f}{dt^{p-1}}\right)_n + O(k^{p+1}) \tag{2.78}$$

By matching the first $p+1$ terms in (2.78) with the equation (2.70) one

[3] note that this statement refers exclusively to ordinary differential equations

order p	**1**	**2**	**3**	**4**		**5**	**6**		**7**		**8**
conditions	1	2	4	8		17	37		85		200
stages	1	2	3	4	5	6	7	8	9	10	11
parameters	1	3	6	10	15	21	28	36	45	55	66

Table 2.2: Explicit Runge-Kutta Methods: number of stages to achieve a specified order

imposes conditions on the parameters a, b and c such that the scheme is p-order accurate (order conditions).

Using (2.70) in (2.76) one gets the consistency condition of the Runge-Kutta method:

$$\sum_{i=1}^{q} b_i = 1 \tag{2.79}$$

The order and consistency conditions form a nonlinear algebraic system to be solved for the parameters a, b and c. Whether the system has solutions or not, depends also on the number of stages, q. Due to computational cost, one is interested in choosing the smallest number of stages, q for which the system admits a solution. For general order, p, this problem of finding the algorithm with the minimum number of stages is not trivial to solve —in fact, it has not yet been solved beyond 8th order (see table 2.2).

2.2.2 Absolute Stability of Runge-Kutta methods

The stability analysis for explicit Runge-Kutta methods is carried out starting from the simplest model

$$\begin{aligned} \frac{d\mathsf{y}}{dt} &= v\mathsf{y}, \quad t \geq t_0 \\ \mathsf{y}(t_0) &= \mathsf{y}_0 \end{aligned} \tag{2.80}$$

with $v \in \mathbb{C}$. The analytical solution is $y(t) = y_0 e^{v(t-t_0)}$. The continuum system is called Lyapunvov stable if the solution is bounded as $t \to \infty$, that is if Re $v < 0$. If Re $v > 0$ the system is (Lyapunov) unstable.

One can show that integrating numerically (2.80) with a Runge-Kutta algorithm is equivalent with solving the following recurrence formula:

$$y_n = y_{n-1}\left[1 + (kv)b^T\left(I - (kv)A\right)^{-1}\mathbf{1}\right] = y_0\left[1 + (kv)b^T\left(I - (kv)a\right)^{-1}\mathbf{1}\right]^n \tag{2.81}$$

The numerical method is said to be (absolutely) stable if, for a fixed k, the solution of the recurrence relation is bounded as $n \to \infty$.

The stability function is defined as

$$\mathcal{P}(z) = 1 + zb^T\left(I - z\,a\right)^{-1}\mathbf{1} \tag{2.82}$$

The region of *absolute stability* is the set $z \in \mathbb{C}$ which satisfy

$$|\mathcal{P}(z)| \leq 1 \tag{2.83}$$

For explicit Runge-Kutta methods, the stability function is just a polynomial in z

$$\mathcal{P}(z) = \sum_{r=0}^{p}\frac{z^r}{r!} + \sum_{r=p+1}^{m}\alpha_j\frac{z^r}{r!} \tag{2.84}$$

Also if the order $p \leq 4$ the stability function does not depend on the parametrization of the method and takes the simpler form

$$\mathcal{P}(z) = \sum_{r=0}^{p}\frac{z^r}{r!}. \tag{2.85}$$

According to [48], the Runge-Kutta method is called *locally stable*, if there is an $R > 0$ such that the inequality (2.83) holds for all z with Re $(z) \leq 0$ and $|z| \leq R$.

The Runge-Kutta method is called *locally stable on the imaginary axis*, if there is an $R > 0$ such that the inequality (2.83) holds for all z with Re $(z) = 0$ and $|z| \leq R$.

Fig. 2.4 displays the stability diagrams for various orders Runge-Kutta

methods. The 1st, 2nd, 3rd and 4th orders are constructed according to (2.85), while the 6th and the 8th orders according to the general formula (2.84) using the coefficients given in [56] and respectively, [79]. The 3rd, 4th and the 8th order Runge-Kutta are locally stable while all the others are not. Time integrators which are not locally stable can still be used in solving numerically partial differential equations (within the method of lines approach) but require dissipation.

2.3 Well-Posedness and Numerical Stability

The previous two sections set up the basic tools for the discretization of an IVP using the MoL approach: FDOs and Runge-Kutta time integrators. It is time now to introduce and analyze several important issues which come with the discretization: stability, accuracy and convergence. As mentioned already in the introduction, there are two concepts of stability: one refers to the behaviour of the numerical solution in the continuum limit (Lax-stability) while the other one to the behaviour of the numerical solution in time.

This section is dedicated to the investigation of Lax-stability for the IVP of second order systems in space and first order in time. Lax-stability (from now on will be simply refered as stability) is the discrete analogous of well-posedness. In order to see how, consider the general IVP problem,

$$\begin{aligned} \mathsf{u}_t &= P(x,t,\frac{\partial}{\partial x})\mathsf{u} \\ \mathsf{u}(0,x) &= \mathsf{f}(x) \end{aligned} \tag{2.86}$$

The problem is called **well-posed** with respect to the norm $\|.\|_*$, if there are constant K and α such that

$$\|\mathsf{u}(t,.)\|_* \leq Ke^{\alpha t}\|\mathsf{f}(.)\|_* \tag{2.87}$$

holds for all initial data $\mathsf{f}(x)$.

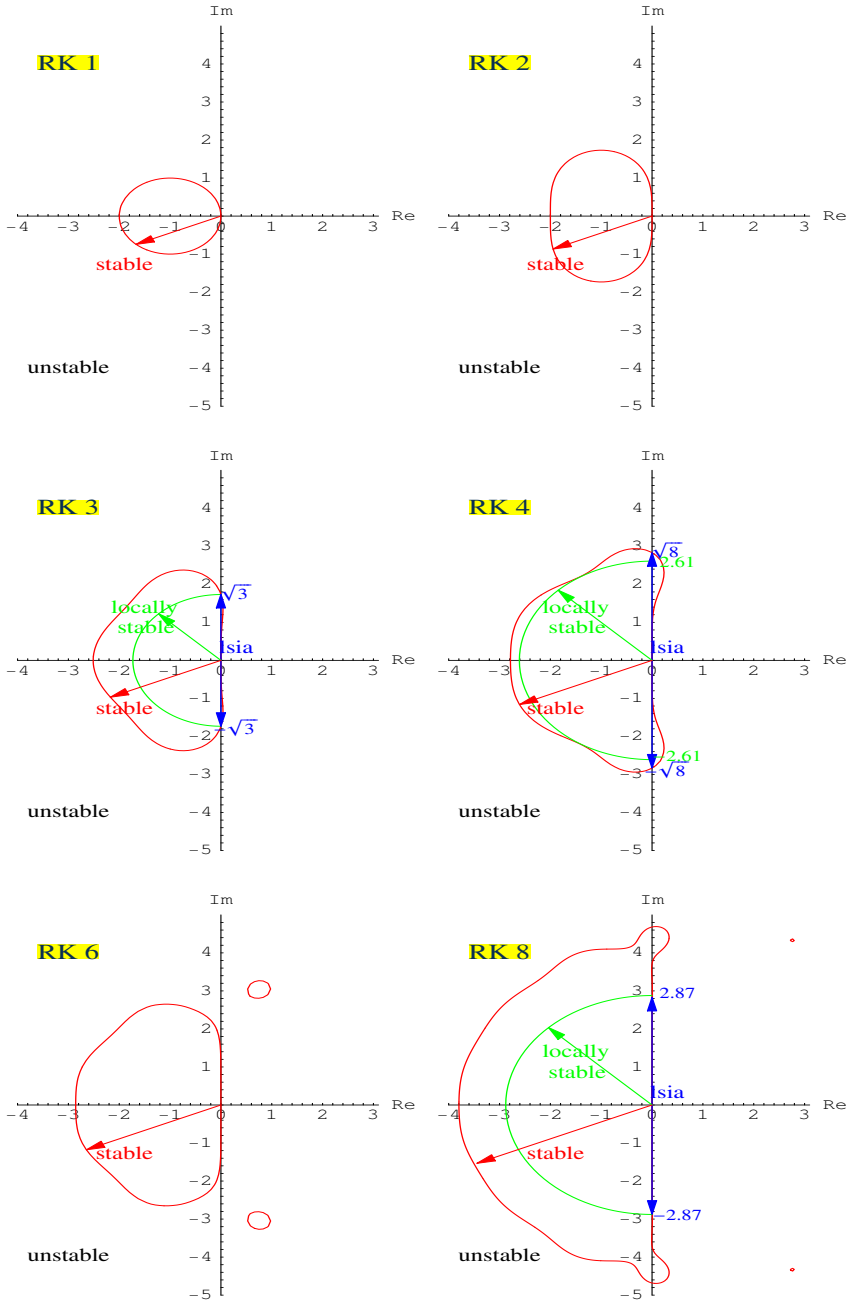


Figure 2.4: **Stability diagrams of some Runge-Kutta methods** The interiors of the red/green contours denote the regions of absolute/local stability. The blue segments correspond to the regions of local stability on the imaginary axis. The 1st, 2nd and 6th orders (RK1, RK2, RK6) are not locally stable.

Now discretize the problem (2.86) using the one-step scheme:

$$\begin{aligned} u^{n+1} &= Q_{h,k}(t_n)u^n \\ u^0 &= f \end{aligned} \tag{2.88}$$

where h is the grid-spacing and k is the time step.
Obviously,

$$u^n = S_{h,k}(t_n)u^0 \quad \text{where} \quad S_{h,k}(t_n) = \prod_{i=0}^{n-1} Q_{h,k}(t_i)$$

In respect to the discrete norm $\|\mathsf{u}(t,.)\|_{*,h}$, the discretization is called [36]

- **stable** if there are constant K, α such that

$$\|u^n\|_{*,h} \leq Ke^{\alpha t_n} \|f\|_{*,h} \tag{2.89}$$

  holds for all initial data f, $0 < h < h_0$ and $nk = t_n$.

- **accurate of order** p_1**,** p_2 if

$$\|\mathsf{u}(t_n + k,.) - Q_{h,k}\mathsf{u}(t_n,.)\|_{*,h} = \mathcal{O}(h^{p_1} + k^{p_2}). \tag{2.90}$$

  If $p_1 > 0$, $p_2 > 0$ then the scheme is called **consistent**.

- **convergent** if

$$\lim_{k,h\to 0} \|u^n - \mathsf{u}(t_n,.)\|_{*,h} = 0 \tag{2.91}$$

The most desirable property of the numerical scheme is convergence. However this is hard to investigate analytically. Fortunatly, there is a theorem, (Lax-Richtmyer equivalence theorem) that states: "a consistent finite difference scheme for a partial differential equation for which the initial-value problem is well-posed is convergent if and only if it is stable." (for a proof see [75])

Because consistency is relatively easy to check, showing stability becomes the main goal for such a discrete scheme.

Numerical implementation of first order linear hyperbolic systems with constant coefficients is greatly simplified by the following result [36]: If the Cauchy problem is well-posed, then the semidiscrete problem (only discretizing space and leaving time continuous) is stable when spatial derivatives are discretized with a centered finite difference operator. Furthermore, if the time integration is performed using an explicit one-step ODE integrator (e.g. the standard 4th order Runge-Kutta method), then, for sufficiently small Courant factor, the stability of the fully discrete problem is also achived.

Such a result does not hold, in general, for second order systems where first *and* second spatial derivatives appear. In order to obtain a stable semidiscrete scheme, the second order system needs to have additional properties. In [14] sufficient conditions for stability of the fully discrete problem for such systems were presented.

Although the method of [14] is in principle general (it applies to discretizations using a centered FDO of any order of accuracy), their focus was mainly on second and fourth order accurate centered discretizations. The following two sections recall the derivation of the stability criteria formulated in [14] and close the technical gap for arbitrary order discretizations in lemma 2.3.1. Even more, by this lemma it is shown that the method applies also in the case when some first order derivatives, which can be identified with advection terms, are approximated with noncentered finite difference operators. The path followed in [14] for deriving sufficient conditions for stability mirrors the path for showing well-posedness at the continuum level. The same approach applies also here.

2.3.1 Well-Posedness

The system under consideration matches the form (1.4). Note that the state vector v is split into two parts: U are those variables for which only first

spatial derivatives appear, while second spatial derivatives of the V-variables also enter the matrix P. For constant coefficient linear PDEs, it is convenient to treat the problem in Fourier space, assuming that the data are 2π-periodic in each spatial direction. The system is Fourier transformed ($\mathsf{P} \rightarrow \hat{\mathsf{P}}$) by using the representations for the functions and derivative operators described in 2.1.1. The evolution problem reduces in this way to solving a system of ODEs. By performing a pseudo first order reduction, it can be shown that the well-posedness is not influenced by the lower order terms of $\hat{\mathsf{P}}$. Thus, they can be dropped from the analysis and restrict attention to the so called *the second order principal symbol* (corresponding to the direction $\vec{n}$),

$$\hat{\mathsf{P}}' = \begin{pmatrix} i\omega_0 A^n & C \\ -\omega_0^2 D^{nn} & i\omega_0 G^n \end{pmatrix}, \tag{2.92}$$

where $\omega_0 = |\vec{\omega}|$, $\omega_i = \omega_0 n_i$ and $M^n = M_i n_i$. In [14] it is shown that if there exists a positive matrix $\hat{\mathsf{H}}(\vec{\omega}) = \hat{\mathsf{H}}^*(\vec{\omega})$ and a positive constant K, such that

$$\hat{\mathsf{H}}\hat{\mathsf{P}}' + \hat{\mathsf{P}}'^*\hat{\mathsf{H}} = 0 \tag{2.93}$$

$$K^{-1} I_{\omega_0} \leq \hat{\mathsf{H}} \leq K I_{\omega_0}, \quad \text{where} \quad I_{\omega_0} = \begin{pmatrix} \omega_0^2 I_p & 0 \\ 0 & I_q \end{pmatrix}, \tag{2.94}$$

then the problem is well-posed in the norm

$$\|\mathsf{v}\|_\partial^2 = \int \sum_{i=1}^{d} |\partial_i \mathsf{U}|^2 + |\mathsf{V}|^2 \ . \tag{2.95}$$

The problem is also well-posed in the norm $\|\mathsf{v}\|_{\mathsf{H}}$, defined as

$$\|\mathsf{v}\|_{\mathsf{H}}^2 = \sum_{\vec{\omega}} \hat{\mathsf{v}}^\dagger \hat{\mathsf{H}} \hat{\mathsf{v}}. \tag{2.96}$$

One can show that the norms (2.95) and (2.96) are equivalent.

Obs. A positive definite matrix $\hat{\mathsf{H}}$ which verifies (2.93) is called sym-

metrizer of the system.

2.3.2 Numerical Stability

We now turn to the discrete level. The stability analysis *almost* mirrors the well-posedness analysis and is greatly simplified by adopting the method-of-lines approach where initial time is kept continuous and only space is discretized.

The discrete system corresponding to (1.4) is:

$$\frac{d}{dt}v = Pv, \;\; v = (U, V)^T \,,$$

$$P \;\; = \;\; \left(\begin{array}{cc} A^i D_i^{(1,n)} + B & C \\ D^{ij} D_{ij}^{(2,n)} + E^i D_i^{(1,n)} + F & G^i D_i^{(1,n)} + J \end{array} \right) . \tag{2.97}$$

In the relations above, $D_i^{(1,n)}$ and $D_{ij}^{(2,n)}$ are taken as the $2n$-accurate centered discretizations of the first and second derivative in the i-direction and in the i and j directions, respectively.

The problem is now analyzed in Fourier space, representing grid functions in terms of discrete Fourier coefficients and FDOs in terms of discrete Fourier symbols. After Fourier transforming the system using the relations (2.4) and (2.9), a first order reduction is performed by introducing the variable $\hat{w}$,

$$\hat{w} \equiv i\Omega_0 \hat{u}, \;\; \Omega_0^2 = \sum_{i=1}^{d} \left| \hat{D}_{+i} \right|^2 \,, \tag{2.98}$$

where $\hat{D}_{+i}$ is the Fourier representation of the usual forward finite difference operator in the i-direction, D_{+i}.

This yields to the following system of ODEs:

$$\begin{aligned}\frac{d}{dt}\hat{v}_R &= \hat{P}_R\hat{v}_R \text{ with } \hat{v}_R = (\hat{u},\hat{w},\hat{v})^T\,,\\ \hat{P}_R &= \begin{pmatrix} B & (i\Omega_0)^{-1}A^i\hat{D}_i^{(1,n)} & C \\ 0 & A^i\hat{D}_i^{(1,n)}+B & i\Omega_0 C \\ F & (i\Omega_0)^{-1}\left(D^{ij}\hat{D}_{ij}^{(2,n)}+E^i\hat{D}_i^{(1,n)}\right) & G^i\hat{D}_i^{(1,n)}+J \end{pmatrix}\end{aligned} \tag{2.99}$$

Here $\hat{D}_i^{(1,n)}$ and $\hat{D}_{ij}^{(2,n)}$ are the Fourier symbols of the discrete derivatives. By theorem 5.1.2 of [36] the terms which correspond to the continuum lower order terms can be dropped from $\hat{P}_R$ without affecting the stability analysis if

$$(i\Omega_0)^{-1}\hat{D}_i^{(1,n)}, k\hat{D}_i^{(1,n)}, k\Omega_0^{-1}\hat{D}_{ij}^{(2,n)} \tag{2.100}$$

are bounded. The lemma 2.3.1 will show that this is indeed the case for any centered FDO. Having proved this, the rest of the discussion in [14] applies. The problem now reduces to the analysis of a first order system with the principal part:

$$\hat{P'}_R = \begin{pmatrix} A^i\hat{D}_i^{(1,n)} & i\Omega_0 C \\ (i\Omega_0)^{-1}D^{ij}\hat{D}_{ij}^{(2,n)} & G^i\hat{D}_i^{(1,n)} \end{pmatrix}. \tag{2.101}$$

Following [14], define the principal part of the second order system as:

$$\hat{P'} = \begin{pmatrix} A^i\hat{D}_i^{(1,n)} & C \\ D^{ij}\hat{D}_{ij}^{(2,n)} & G^i\hat{D}_i^{(1,n)} \end{pmatrix}. \tag{2.102}$$

For first order systems semidiscrete stability can be discussed in terms of a symmetrizer $\hat{H}_R$, that is a positive matrix $\hat{H}_R = \hat{H}_R(\underline{\xi}, h)$ such that $\hat{H}_R\hat{P'}_R + \hat{P'}^*_R\hat{H}_R = 0$. If such a symmetrizer exists and additionally satisfies $K^{-1}I \leq \hat{H}_R \leq KI$ for some positive constant K, then the semidiscrete problem is stable [36]. [14] show that $\hat{P'} = J^{-1}\hat{P'}_R J$, where $J = diagonal[i\Omega_0, 1]$. This implies that if H_R is a symmetrizer for $\hat{P}'_R$, then $H = J^*\hat{H}_R J$ is a symmetrizer

for $\hat{P}'$. It is also true that $\hat{P}'_R$ and $\hat{P}'$ have the same eigenvalues. In other words, the second order system is stable if:

1. There exists $\hat{H}(\underline{\xi}, h) = \hat{H}^*(\underline{\xi}, h) > 0$ such that

$$K^{-1} I_{\Omega_0} \leq \hat{H} \leq K I_{\Omega_0},\ I_{\Omega_0} = diagonal\left\{\Omega_0^2, 1\right\}, \tag{2.103}$$

$$\hat{H}\hat{P}' + \hat{P}'^* \hat{H} = 0, \tag{2.104}$$

for some positive constant K. This condition implies that the *semidiscrete problem* is stable with respect to the norms $D_\pm$ defined as:

$$\|v\|^2_{h,D_\pm} = \sum_{i=1}^{d} \|D_{\pm i} U\|^2_h + \|V\|^2_h \ . \tag{2.105}$$

In one space dimension the derivative term in the norm is

$$\|D_\pm u\|^2_h = \sum_{j=0}^{N-2} |D_+ u_j|^2 + |D_- u_{N-1}|^2 = |D_+ u_0|^2 + \sum_{j=1}^{N-1} |D_- u_j|^2 \ . \tag{2.106}$$

The generalization to d dimensions is straightforward.

Remark Consider the scalar product $(v, w)_{h,H} = \sum_{\underline{\omega}} \hat{v}^T \hat{H} \hat{w}$ and the corresponding norm

$$\|v\|_{h,H} = \sum_{\underline{\omega}} \hat{v}^T \hat{H} \hat{v}\,, \tag{2.107}$$

Then the problem is well-posed also in this norm because $\|v\|_{h,H}$ and $\|v\|^2_{h,D_\pm}$ are equivalent:

$$K^{-1/2} \|v(t,.)\|_{h,D_+} \leq \|v(t,.)\|_{h,H} = \|v(0,.)\|_{h,H} \leq K^{1/2} \|v(0,.)\|_{h,D_+} \ . \tag{2.108}$$

2. The time is discretized by using a *locally stable* (implicit or explicit) Runge-Kutta method. Then, according to [48] the resulting fully discretized scheme the eigenvalues of $\hat{P}'$ have non-positive real parts for all

frequencies and the Courant factor λ is chosen such that

$$\lambda \leq \frac{\alpha_0}{\sigma(h\hat{P}')}, \tag{2.109}$$

where $\sigma(h\hat{P}')$ is the maximum spectral radius of $h\hat{P}'$ and α_0 is a constant specific to the time integrator. In the general case α_0 will denote the radius of local stability (e.g. for the classical fourth order Runge-Kutta method $\alpha_0 = 2.61$) and in the particular case when the eigenvalues are purely imaginary, α_0 can be taken as the radius of local stability on the imaginary axis, leading to a relaxation of the Courant limit (e.g for the classical fourth order Runge-Kutta method $\alpha_0 = \sqrt{8} = 2.83$).

Remark If the right hand sides of the equations are modified by adding artificial dissipation (via the operator $\mathcal{D}$ defined in (2.31)) and/or by adding advection terms of the form $I\beta^i D_i^{(1,n,s,\epsilon)}$ (where $D_i^{(1,n,s,\epsilon)}$ is the non-centered FDO defined in (2.14)), these modifications only have effect on the diagonal entries of the principal part. The new system will have different eigenvalues than $\hat{P}'$ but the same set of eigenvectors. The symmetrizer will not depend on the way the advection terms are discretized, nor on the dissipation operator. This implies that the stability conditions (2.103)-(2.104) and (2.109) remain valid if

$$(i\Omega_0)^{-1}\hat{D}_i^{(1,n,s,\epsilon)}, (i\Omega_0)^{-1}\hat{\mathcal{D}} \tag{2.110}$$

are bounded, which will be shown in lemma 2.3.1.

Lemma 2.3.1 *The terms*

$$(i\Omega_0)^{-1}\hat{D}_j^{(1,n)},\ k\hat{D}_j^{(1,n)},\ k\Omega_0^{-1}\hat{D}_{ij}^{(2,n)},\ (i\Omega_0)^{-1}\hat{D}_j^{(1,n,s,\epsilon)}, (i\Omega_0)^{-1}\hat{\mathcal{D}} \tag{2.111}$$

are bounded.

Using the relations (2.37), (2.38), (2.41) and (2.45), the proof is reduced to

showing the boundedness of

$$\check{\Omega}_0^{-1}\check{d}_j^{(1,n)},\ i\lambda\check{d}_j^{(1,n)},\ \lambda\check{\Omega}_0^{-1}\check{d}_j^{(2,n)},\ \lambda\check{\Omega}_0^{-1}(\check{d}_j^{(1,n)})^2, \check{\Omega}_0^{-1}\check{d}_j^{(1,n,s)}, i\check{\Omega}_0^{-1}\check{\mathbf{d}}_j^{(1,n,s)}, (i\check{\Omega}_0)^{-1}\check{\Omega}_j^{2m}\,.$$

From (2.39), (2.40), (2.42) and (2.43) observe that each of these quantities can be written formally as a product $\check{\Omega}_j\check{\Omega}_0^{-1}F(\check{\Omega}_j)$, with $F(\check{\Omega}_j)$ a continuous and bounded function in $(-2,2]$. Because $\check{\Omega}_j\check{\Omega}_0^{-1}$ is bounded in $(-2,2]\times(-2,2]\times(-2,2]$ ($\left|\check{\Omega}_j/\check{\Omega}_0\right|\le 1$) the desired result is obtained.

2.4 Dispersion and Dissipation

2.4.1 Mode Splitting

Let $\hat{\mathsf{T}} = \hat{\mathsf{T}}(\vec{\omega})$ be the matrix of the eigenvectors of the second order principal symbol, $\hat{\mathsf{P}}'$, and $\hat{\mathsf{D}} = \hat{\mathsf{D}}(\vec{\omega})$ the matrix of its eigenvalues, ($\hat{\mathsf{P}}' = \hat{\mathsf{T}}\hat{\mathsf{D}}\hat{\mathsf{T}}^{-1}$). Then the characteristics of the second order system in Fourier space are defined by $\hat{\mathsf{c}} = \hat{\mathsf{T}}^{-1}\hat{\mathsf{v}}$. The discrete characteristics are constructed in a similar way, $\hat{c} = \hat{T}^{-1}\hat{v}$ from the matrix of eigenvectors of $\hat{P}'$, $\hat{T} = \hat{T}(\underline{\omega},\underline{\xi})$

Let $\hat{\mathsf{C}}$ be a characteristic of the continuum system, $\hat{\Lambda}$ the corresponding eigenvalue, and the pair ($\hat{C}$, $\hat{\Lambda}$) their discrete analogs. Consistent initial data, $\hat{C}_0(\underline{\omega},\underline{\xi})$ is provided for both the continuum and discrete system. Then the evolution equation for the characteristics, $\hat{\mathsf{C}}$ and $\hat{C}$ are:

$$\hat{\mathsf{C}}(t,\underline{\omega},\underline{\xi}) = \hat{C}_0(\underline{\omega},\underline{\xi})e^{\hat{\Lambda}t} \quad \text{and } \hat{C}(t,\underline{\omega},\underline{\xi}) = \hat{C}_0(\underline{\omega},\underline{\xi})e^{\hat{\Lambda}t} \tag{2.112}$$

and the numerical solution can be written in terms of the continuum solution:

$$\hat{C}(t,\underline{\omega},\underline{\xi}) = \hat{\mathsf{C}}(t,\underline{\omega},\underline{\xi})e^{\hat{\Lambda}^{Re}t}e^{i\left(\hat{\Lambda}^{Im}-\hat{\Lambda}^{Im}\right)t} \tag{2.113}$$

where the superscripts Re and Im denote the real and imaginary parts of the eigenvalues. ($\hat{\Lambda}^{Re} = 0$ because the system if hyperbolic.) We now want to discriminate between mechanisms which modify only the phase of the mode

(leading to phase errors) and mechanisms which modify only the amplitude (damping or amplification of the signal).

2.4.2 Amplification Factor and Speed Errors

The following definitions are made:

- **amplification factor** of one mode, $a_p = a_p(\underline{\xi})$, in terms of the real part of the eigenvalues:
$$a_p = \frac{1}{\xi_0}(h\hat{\Lambda}^{Re})(\underline{\xi}) \tag{2.114}$$

- **phase and group speeds** in terms of the imaginary part of the eigenvalues:

$$\begin{aligned} & & \underline{\text{phase}} & & \underline{\text{group}} & \\ \text{continuum:} & & \mathsf{v}_p \equiv \tfrac{\hat{\Lambda}^{Im}(\vec{\omega})}{\omega_0}, & \mathsf{v}_g \equiv n^i \frac{d}{d\omega_i}(\hat{\Lambda}^{Im})(\vec{\omega}) & (2.115) \\ \text{discrete:} & & v_p \equiv \tfrac{\hat{\Lambda}^{Im}(\underline{\omega},\underline{\xi})}{\omega_0}, & v_g \equiv n^i \frac{d}{d\xi_i}(h\hat{\Lambda}^{Im})(\underline{\xi}) & (2.116) \end{aligned}$$

- **phase/group speed errors**:
$$\epsilon_p \equiv v_p - \mathsf{v}_p, \quad \epsilon_g \equiv v_g - \mathsf{v}_g \tag{2.117}$$

- **phase error**:
$$\hat{\mathfrak{E}} \equiv \left(\hat{\Lambda}^{Im} - \hat{\mathsf{\Lambda}}^{Im}\right) t = e_p \omega_0 t \tag{2.118}$$

- **relative error** of one mode is $\hat{E} = \hat{E}(t, \underline{\omega}, \underline{\xi})$,
$$\hat{E} \equiv \frac{\hat{C} - \hat{\mathsf{C}}}{\hat{\mathsf{C}}} = e^{(\hat{\Lambda} - \hat{\mathsf{\Lambda}})t} - 1 = e^{a_p \omega_0 t} e^{i\mathfrak{E}} - 1 = e^{a_p \omega_0 t} e^{i\epsilon_p \omega_0 t} - 1 \tag{2.119}$$

The absolute value of the relative error will evolve according to:

$$\left|\hat{E}\right|^2 = 4e^{a_p\omega_0 t}\sin^2\left(\frac{\epsilon_p\omega_0 t}{2}\right) + \left(e^{a_p\omega_0 t} - 1\right)^2 \tag{2.120}$$

The relation (2.120) tells that: a) if $a_p > 0$ the error of the semidiscrete problem will grow exponentially in time; b) if $a_p = 0$ the error will have an oscillatory behavior with the period $\frac{2\pi}{\epsilon_p\omega_0}$ and constant amplitude 2; c) if $a_p < 0$ the relative error is the superposition of two effects: damped oscillatory effect and a growing effect (asymptotically to 1)— the first effect will be dominant at early times while the second will dominate at later times.

In the **linear regime**, defined by $t \ll \frac{1}{\omega_0}\min\{\frac{1}{\epsilon_p}, \frac{1}{a_p}\}$, the errors of the semidiscrete problem scale linearly with time,

$$\begin{aligned} \hat{E} &\simeq \left(\hat{\Lambda} - \hat{\mathsf{\Lambda}}\right)t = i\epsilon_p\omega_0 t + a_p\omega_0 t\,, \\ \mathfrak{E} &\simeq \epsilon_p\omega_0 t\,, \\ \left|\hat{E}\right| &\simeq \omega_0 t\sqrt{\epsilon_p^2 + a_p^2}\,. \end{aligned} \tag{2.121}$$

- **total error**

If $\hat{C}_1, \hat{C}_2, ...$ are the characteristics of the system and $\hat{E}_1, \hat{E}_2 \ldots$ their relative errors given by (2.119) then the norm $\|.\|_{h,H}$ of numerical solution v is given by

$$\|v\|_{h,H}^2 = \sum_j \sum_{\underline{\omega}} \left|\hat{C}_j(t, \underline{\omega}, h\underline{\omega})\right|^2 \tag{2.122}$$

and the total error satisfies:

$$\|v - \mathsf{v}\|_{h,H}^2 = \sum_j \sum_{\underline{\omega}} \left|\hat{\mathsf{C}}_j(t, \underline{\omega})\hat{E}_j(t, \underline{\omega}, h\underline{\omega})\right|^2 \tag{2.123}$$

2.4.3 Advection Equation

For illustration, the definitions introduced in 2.4.2 are now applied on the particular case of the advection equation.

The continuum and the discrete advection equations are:

$$\text{continuum:} \quad \dot{\mathsf{C}}(t,\vec{x}) = \beta\partial\mathsf{C}(t,\vec{x}) \tag{2.124}$$

$$\text{discrete:} \quad \dot{C}(t,\underline{x}) = \beta D^{(1,n,s,\epsilon)}C(t,\underline{x}) \tag{2.125}$$

One can show that the amplification factor and the phase speed error are:

$$a_p = \beta\epsilon\frac{\check{\mathbf{d}}^{(1,n,s)}}{|\xi|} \quad \text{and} \quad \epsilon_p = \beta\left(\frac{\check{d}^{(1,n,s)}}{\xi} - 1\right) \tag{2.126}$$

In case the first derivative is approximated with a centered FDO, ($s = 0$), the amplification factor is zero ($a_p = 0$) for all frequencies. For an one-point upwinded scheme ($s = 1$ and sign β = sign ϵ) the scheme is dissipative ($a_p < 0$ for all frequencies). For all the other cases there are frequencies where $\check{\mathbf{d}}^{(1,n,s)}$ changes sign, so there are modes with $a_p > 0$, that is, modes that exhibit an exponential growth. In order to cure that, dissipation has to be added to the scheme.

According to the relations (2.63), for small frequencies, the amplification factor and the phase speed error are given by:

$$a_p \simeq \beta\epsilon\frac{2n+1}{2n+2}s\,T^{(1,n,s)}\xi^{2n+2} \quad \text{and} \quad \epsilon_p = \beta(-1)^{n+1}T^{(1,n,s)}\xi^{2n} \tag{2.127}$$

where $T^{(1,n,s)}$ is defined in (2.21) and satisfies the inequality (2.22). This means that in case of advection equation, for any β, off-centering by s points will lead to larger phase speed errors in the small frequency regime and consequently to larger total errors, in comparison to the centered scheme.
At higher frequencies this situation can change and one can show that there are intervals in the spectrum where off-centerings improve the phase speed errors (see for example the Fig. 2.3).

Chapter 3

Initial Value Problem for the Wave Equation

This chapter applies the methods presented in Chapter 2 for studying the discretization of initial value problem for general second order systems, to the particular example of the wave equation on a general curved background. It analyses the well-posedness of the continuum problem and the stability and the accuracy of the numerical scheme using $2n$-accurate finite differencing operators. In the case of 1-D wave equation with shift, special attention is paid to the investigation of Courant limits and numerical speeds in connection to the order of approximation and off-centering of some first order discrete derivatives.

3.1 Introduction

The standard wave equation in d space dimensions is

$$-\frac{\partial^2 \Phi}{\partial \tilde{t}^2} + \Delta \Phi = -\frac{\partial^2 \Phi}{\partial \tilde{t}^2} + \sum_{i=1}^{d} \frac{\partial^2 \Phi}{\partial \tilde{x}^{i\,2}} = 0\,. \tag{3.1}$$

It is a special case of the curved spacetime scalar wave equation

$$g^{\alpha\beta}\partial_\alpha\partial_\beta\Phi = 0, \tag{3.2}$$

where $g_{\alpha\beta}$ is the spacetime metric and the summation is done over repeated indices $\alpha, \beta = \overline{0, d}$.

The equation (3.1) is obtained from (3.2) by considering flat spacetime and choosing standard Cartesian coordinates corresponding to the line element

$$ds^2 = -d\tilde{t}^2 + \sum_i d\tilde{x_i}^2.$$

Although much simpler, the curved spacetime wave equation can serve as a very useful model for the numerical solution of the Einstein equations — in particular since the generalized harmonic formulation of the Einstein equations takes the form of a system of wave equations (however with very complicated source terms, see [30] for a detailed discussion).

For simplicity and without reducing the generality, assume a uniform time slicing, $g^{00} = -1$. Then perform a $d+1$ split introducing a positive definite d-metric $\gamma^{ij} = g^{ij} + \beta^i\beta^j$ with $i, j = \overline{1, d}$ and a shift vector $\beta^i = g^{0i}$ (see e.g. [80]). The wave equation (3.2) becomes

$$\partial_{tt}\Phi = 2\beta^i\partial_i\partial_t\Phi + \left(\gamma^{ij} - \beta^i\beta^j\right)\partial_i\partial_j\Phi\,. \tag{3.3}$$

Now, in analogy with the York-ADM-system ([82], Appendix B), the variable K is introduced by

$$\mathsf{K} = \partial_t\Phi - \beta^i\partial_i\Phi \tag{3.4}$$

which transforms the wave equation into the first order in time, second order in space system:

$$\begin{aligned} \partial_t\Phi &= \beta^i\partial_i\Phi + \mathsf{K}\,, \\ \partial_t\mathsf{K} &= \gamma^{ij}\partial_{ij}\Phi + \beta^i\partial_i\mathsf{K}\,. \end{aligned} \tag{3.5}$$

Obs. In the particular case of a flat metric and one space dimension ($d = 1$), the system (3.5) has only one parameter ($\beta \equiv \beta_1$). This case will be extensively analyzed in this thesis so it merits being written down explicitly:

$$\begin{aligned} \partial_t \Phi &= \beta \partial_x \Phi + \mathsf{K}\,, \\ \partial_t \mathsf{K} &= \partial_{xx} \Phi + \beta \partial_x \mathsf{K}\,. \end{aligned} \tag{3.6}$$

Well-posedness for the Cauchy problem for the system (3.2) is a standard textbook result both in the original second order form and for the reduction to first order symmetric hyperbolic form. Here the well-posedness and numerical stability are proved for the first order in time, second order in space equivalent system (3.5), using the methods presented in the previous chapter.

3.2 Continuum Problem

It is easy to see that the initial boundary value problem for the wave equation (3.5), is indeed well-posed. Following the procedure outlined in 2.3.1, the system is investigated in Fourier space. Showing well-posedness amounts to proving the existence of a symmetrizer (a positive matrix that satisfies (2.93)) which obeys the boundeness condition (2.94).

Define $\hat{\Delta} \equiv \sqrt{\gamma^{ij}\omega_i\omega_j}$. Then the second order principal symbol, the diagonalizing matrix and the eigenvalues are:

$$\hat{\mathsf{P}}' = \begin{pmatrix} i\beta^j\omega_j & 1 \\ -\hat{\Delta}^2 & i\beta^j\omega_j \end{pmatrix}, \quad \hat{\mathsf{T}}^{-1} = \begin{pmatrix} i\hat{\Delta} & 1 \\ -i\hat{\Delta} & 1 \end{pmatrix}, \quad \hat{\Lambda}_\pm = i\left(\beta^j\omega_j \pm \hat{\Delta}\right) \tag{3.7}$$

Because γ^{ij} is positive definite, $\hat{\Delta} \geq 0$. This means that the eigenvalues are purely imaginary and also that

$$\hat{\mathsf{H}} \equiv \frac{1}{2}\hat{\mathsf{T}}^{-1*}\hat{\mathsf{T}}^{-1} = \begin{pmatrix} \hat{\Delta}^2 & 0 \\ 0 & 1 \end{pmatrix} \tag{3.8}$$

is a symmetrizer for the system ($\hat{\mathsf{H}} > 0$ and $\hat{\mathsf{H}}\hat{\mathsf{P}}' + \hat{\mathsf{P}}'^{*}\hat{\mathsf{H}} = 0$).

The positivity of the matrix γ^{ij} implies, also, that there exists a constant $c_1 > 0$ such that

$$\gamma^{ij}\omega_i\omega_j \geq c_1\omega_0^2, \quad \forall\omega_i \in \mathbb{R}\,, \tag{3.9}$$

$$\min\gamma^{ii} \geq c_1\,. \tag{3.10}$$

Because $|\gamma^{ij}| < \infty$ there also exists a constant $c_2 > 0$ such that

$$\gamma^{ij}\omega_i\omega_j \leq c_2\omega_0^2 \quad \forall\omega_i \in \mathbb{R}\,, \tag{3.11}$$

$$\max\gamma^{ii} \leq c_2\,. \tag{3.12}$$

Take $K = \max\{c_1^{-1}, c_2, 1\}$ and consequently, $K^{-1} = \min\{c_1, c_2^{-1}, 1\}$. The definition for $\hat{\Delta}$ and the above inequalities lead to

$$K^{-1}\omega_0^2 \leq \hat{\Delta}^2 \leq K\omega_0^2 \tag{3.13}$$

The boundedness condition (2.94) for the symmetrizer follows immediatelly from (3.13). With this the proof of well-posedness ends.

The conserved quantity in physical space, corresponding to the symmetrizer $\hat{\mathsf{H}}$ via the Parseval relation (2.3), is:

$$\|\mathsf{v}\|_{\mathsf{H}} \;=\; \int_{\mathbb{R}^d} dx\,\left(\gamma^{ij}\partial_i\Phi\partial_j\Phi + \mathsf{K}^2\right)\,, \tag{3.14}$$

with $\mathsf{v} = (\Phi, \mathsf{K})^T$.

3.3 Discrete Problem

The system (3.5) is discretized using the MoL approach, first, by leaving continuous in time and discretizing only in space and then by integrating the system of ODEs using a locally stable Runge-Kutta method.

According to the general analysis presented in 2.3.2, the stability of the fully discrete problem is achieved if: 1) the semidiscrete problem is stable and 2) the conditions for local stability are satisfied. The first issue is going to be addressed in the 3.3.1 while the second in 3.3.2.

3.3.1 Semidiscrete Problem

The semidiscrete system corresponding to (3.5) is:

$$\frac{d}{dt}\Phi = \beta^i D_i^{(1,n,s,\epsilon)}\Phi + K\,, \tag{3.15}$$

$$\frac{d}{dt}K = \gamma^{ij} D_{ij}^{(2,n)}\Phi + \beta^i D_i^{(1,n,s,\epsilon)} K\,. \tag{3.16}$$

This way of discretizing the first order derivative terms, which correspond to advection along the shift vector β^i, with off-centered derivatives has become customary in numerical relativity (see e.g. [1, 42, 83]).

The stability of the semidiscrete problem is analyzed in Fourier space in a way similar to the well-posedness analysis for the continuum problem.

Define the shorthand quantity $\hat{\Delta}$ as

$$\hat{\Delta} \equiv \sqrt{-\gamma^{il}\hat{D}_{il}^{(2,n)}} = \frac{1}{h}\sqrt{\gamma^{ij}\check{d}_i^{(1,n)}\check{d}_j^{(1,n)} + \sum_i \gamma^{ii}(\check{d}_i^{(2,n)} - \check{d}_i^{(1,n)}\check{d}_i^{(1,n)})}\,. \tag{3.17}$$

Then the discrete symbol, the diagonalizing matrix and the eigenvalues can be written (respectively) as

$$\hat{P}' = \begin{pmatrix} \beta^j \hat{D}_j^{(1,n,s,\epsilon)} & 1 \\ -\hat{\Delta}^2 & \beta^j \hat{D}_j^{(1,n,s,\epsilon)} \end{pmatrix}, \tag{3.18}$$

$$\hat{T}^{-1} = \begin{pmatrix} i\hat{\Delta} & 1 \\ -i\hat{\Delta} & 1 \end{pmatrix}, \qquad \hat{\Lambda}_\pm = \beta^j \hat{D}_j^{(1,n,s,\epsilon)} \pm i\hat{\Delta}\,. \tag{3.19}$$

Because of the relation (2.49) and the positive definiteness of the matrix γ^{ij},

the quantity $\hat{\Delta}$ is real and $\hat{\Delta} \geq 0$ with equality only when all $\check{\Omega}_j$ are zero. This means that

$$\hat{H} \equiv \frac{1}{2}\hat{T}^{-1*}\hat{T}^{-1} = \begin{pmatrix} \hat{\Delta}^2 & 0 \\ 0 & 1 \end{pmatrix} \tag{3.20}$$

is a symmetrizer for the system (3.15, 3.16). Note that the symmetrizer does not depend on the diagonal entries of the symbol $\hat{P}'$, e.g. does not depend on the way the shift terms are advected.

One still has to prove that the symmetrizer obeys the boundeness condition, (2.103), that is there exists a constant $K \geq 1$ such that

$$K^{-1}\Omega_0^2 \leq \hat{\Delta}^2 \leq K\Omega_0^2 \,. \tag{3.21}$$

Using the positive-matrix condition (3.10), the definition for $\check{r}_i^{(n)}$ and the property (2.46) the following chain of inequalities is obtained:

$$h^2\hat{\Delta}^2 \geq (\min \gamma^{ii}) \sum_{i=1}^{d} \check{r}_i^{(n)} + c_1 \sum_{i=1}^{d} (\check{d}_i^{(1,n)})^2 \geq c_1 \sum_{i=1}^{d} \check{d}_i^{(2,n)} \geq c_1 \check{\Omega}_0^2 \,. \tag{3.22}$$

On the other hand, applying the finite-matrix condition (3.12) together with the definition for $\check{r}_i^{(n)}$ and the property (2.46) leads to

$$h^2\hat{\Delta}^2 \leq (\max \gamma^{ii}) \sum_{i=1}^{d} \check{r}_i^{(n)} + c_2 \sum_{i=1}^{d} (\check{d}_i^{(1,n)})^2 \leq c_2 \sum_{i=1}^{d} \check{d}_i^{(2,n)} \leq c_2 C_n \check{\Omega}_0^2 \,. \tag{3.23}$$

Like in the continuum case, chose $K = \max\{c_1^{-1}, (c_2 C_n), 1\}$ and the relation (3.21) is obtained.

The conserved discrete quantity in physical space associated with $\hat{H}$, that is the norm $\|v\|_{h,H}$ defined in (2.107), is:

$$\|v\|_{h,H}^2 = \frac{1}{h^2} \left[\sum_{i=1}^{d} \gamma^{ii} \sum_{k=1}^{n} |d_{k-1}| \left\| (hD_{+i})^k \Phi \right\|_h^2 + \sum_{i \neq j} \gamma^{ij} \left\| hD_i^{(1,n)} \Phi \right\|_h^2 \right] + \|K\|_h^2 \,,$$

where $v = (\Phi^T, K^T)^T$.

By proving the existence of a symmetrizer subjected to the boundeness condition (2.103), the stability for the semidscrete problem has actually been proven with respect to the norms D_+ and H. Note again that the stability property (of the semidiscrete problem!) is in particular independent of how the shift terms are discretized. However these terms become important in the next step of the analysis of the fully discrete problem, as shown below.

3.3.2 Courant Limits and the Role of Dissipation

In order to ensure that the fully discrete problem is stable, the conditions for local stability are now imposed. That means the eigenvalues $\hat{\Lambda}_\pm = \hat{\Lambda}_\pm(\underline{\xi})$ have nonpositive real parts for all frequencies, and the Courant factor is restricted according to (2.109):

$$Re(\hat{\Lambda}_\pm) \leq 0\,, \quad \forall \xi_j \in (-\pi, \pi] \tag{3.24}$$

$$\lambda \leq \frac{\alpha_0}{\max\limits_{\Omega_i \in (-2,2]} \left| h\hat{\Lambda}_\pm \right|}\,. \tag{3.25}$$

Lemma 3.3.1 *The first condition for local stability, (3.24), is satisfied if the advection terms are approximated either by centered FDOs, or by one-point upwinded FDOs.*

Proof According to the formula (3.19) for the eigenvalues,

$$Re(\hat{\Lambda}_\pm) = \frac{1}{h} \sum_j \beta^j \epsilon_j \check{\mathbf{d}}_j^{(1,n,s)}.$$

The relation (3.24) holds for all $\xi_j \in (-\pi, \pi]$ if and only if $\epsilon_j = \text{sign}\, \beta^j$ (upwind) and $\check{\mathbf{d}}^{(1,n,s)}(\xi) \leq 0$ for the whole spectrum. From 2.1.4-2.1.5 we know that $\check{\mathbf{d}}^{(1,n,0)}(\xi) = 0$, $\check{\mathbf{d}}^{(1,n,1)}(\xi) \leq 0$ for all frequencies, while for $s \geq 2$, $\check{\mathbf{d}}^{(1,n,s)}(\xi)$ changes sign in $(-\pi, \pi]$. So, (3.24) holds only for centered or one-point upwinded schemes, and the lemma is proved.

Given Lemma 3.3.1, the following result is straightforward:

Lemma 3.3.2 *For sufficiently small Courant factor, the centered and the one-point upwinded schemes are stable in the following way:*

- *for centered schemes, the problem is in the regime of local stability on the imaginary axis, if (3.25) is satisfied with* $\alpha_0 = \alpha_{lsia}$.
- *for one-point upwinded schemes, the problem is in the regime of local stability, if (3.25) is satisfied with* $\alpha_0 = \alpha_{ls}$.

All the other cases (one-point downwind and off-centerings by $s \geq 2$ points) are unstable. However,

Lemma 3.3.3 *No matter whether the shift terms are upwinded or downwinded, stability can always be achieved by adding dissipation.*

Proof If Kreiss-Oliger dissipation (2.31) is added to the system, the eigenvalues become

$$h\hat{\Lambda}_{\pm} = \beta^j \check{\mathbf{d}}_j^{(1,n,s,\epsilon)} - \frac{1}{2^{2(n+1)}}\sigma^j \check{\Omega}_j^{2(n+1)} + i\left(\beta^j \check{d}_j^{(1,n,s)} \pm h\hat{\Delta}\right). \tag{3.26}$$

According to the relation, (2.42)

$$\frac{\check{\mathbf{d}}^{(1,n,s,\epsilon)}}{\check{\Omega}^{2(n+1)}} = \epsilon\frac{\check{\mathbf{d}}^{(1,n,s)}}{\check{\Omega}^{2(n+1)}} = \epsilon(-1)^s \frac{1}{2C_{2n}^{n+s}} \sum_{k=0}^{s-1} \frac{(-1)^k C_{s+k}^{2k+1}}{(n+1+k)} \check{\Omega}^{2k}.$$

Now, imposing $Re(\hat{\Lambda}_{\pm}) \leq 0$ for all $\Omega_j \in (-2, 2]$, $j = \overline{1,d}$, gives that the minimum dissipation required to obtain a stable scheme is:

$$\sigma^j \geq \begin{cases} 2^{2(n+1)} \left|\beta^j\right| \bar{\sigma}_+^{(n,s)}, & \epsilon_j = \text{sign } \beta_j \text{ (upwind)} \\ \\ 2^{2(n+1)} \left|\beta^j\right| \bar{\sigma}_-^{(n,s)}, & \epsilon_j = -\text{sign } \beta_j \text{ (downwind)} \end{cases} \tag{3.27}$$

where

$$\bar{\sigma}_+^{(n,s)} \equiv \max_{\check{\Omega}\in(0,2]} \frac{\check{\mathbf{d}}^{(1,n,s)}}{\check{\Omega}^{2(n+1)}}, \quad \bar{\sigma}_-^{(n,s)} \equiv -\min_{\check{\Omega}\in(0,2]} \frac{\check{\mathbf{d}}^{(1,n,s)}}{\check{\Omega}^{2(n+1)}}. \tag{3.28}$$

	$\bar{\sigma}_+^{(n,s)}$	$\bar{\sigma}_-^{(n,s)}$
s=1	$-\frac{1}{2C_{2n}^{n+1}}\frac{1}{n+1}$	$\frac{1}{2C_{2n}^{n+1}}\frac{1}{n+1}$
s=2	$\frac{1}{2C_{2n}^{n+2}}\frac{2}{n+1}$	$\frac{1}{2C_{2n}^{n+2}}\frac{2n}{n^2+3n+2}$
s=3	$\frac{1}{2C_{2n}^{n+3}}\frac{n(n+4)}{(n+1)(n+2)^2}$	$\frac{1}{2C_{2n}^{n+3}}\frac{3}{n+1}$

Table 3.1: Formulas for the dissipation parameters $\bar{\sigma}_\pm^{(n,s)}$ when $s = 1, 2, 3$. The quantity $2^{2(n+1)}\,|\beta^j|\,\bar{\sigma}_\pm^{(n,s)}$ ($\pm$ stands for upwind/downwind) represents the minimum dissipation required to make the numerical scheme stable.

The table 3.1 gives the formulas of $\bar{\sigma}_\pm^{(n,s)}$ for $s = 1, 2, 3$.

Note that for all the orders of approximation $2n$, the minimum dissipation required in case of one-point upwind scheme is negative ($\bar{\sigma}_+^{(n,1)} < 0$, so no dissipation is actually needed for stability!), while for all the other cases is positive ($\bar{\sigma}_-^{(n,s=1)}$, $\bar{\sigma}_\pm^{(n,s\geq 2)} > 0$, dissipation is needed!).

By allowing only "positive" dissipation, the minimum amount required for stability is $\sigma^j = 2^{2(n+1)}\,|\beta^j|\,\sigma_\pm^{(n,s)}$ with $\sigma_\pm^{(n,s)} = \max\{0, \bar{\sigma}_\pm^{(n,s)}\}$. That is $\sigma_\pm^{(n,s)} = 0$ for $s = \overline{0,1}$ and $\sigma_\pm^{(n,s)} = \bar{\sigma}_\pm^{(n,s)}$ for $s \geq 2$.

Now, for each choice of dissipation parameters σ^j satisfying (3.27) the Courant factor will be limited according to (3.25). One can easily show that the smaller the dissipation parameters, the higher the Courant factor limit. This means that by choosing σ^j corresponding to equality in (3.27) the Courant limit is maximized.

one-point off-centered scheme

In the case of upwinding by one point, the problem is turned from locally stable to locally stable on the imaginary axis by adding "negative" dissipation. This means that when using one-point upwinded stencils, "negative" dissipation can be used and still obtain a stable scheme. In fact, the following situations are equivalent:

- Upwind one point and add dissipation with $\sigma^j = 2^{2(n+1)} \left|\beta^j\right| \bar{\sigma}_+^{(n,1)} < 0$.
- Downwind one point and add dissipation with $\sigma^j = 2^{2(n+1)} \left|\beta^j\right| \bar{\sigma}_-^{(n,1)} > 0$.
- Use the centered FDO operator constructed with $2(n+s)+1$ points,

$$\frac{1}{2}\left(D^{(1,n,s,1)} + D^{(1,n,s,-1)}\right)$$

  with $s = 1$, and do not add dissipation, $\sigma^j = 0$.

In any of the above three situations, the real part of the eigenvalues is zero, so the local stability condition (3.25) can be relaxed to a condition for local stability on the imaginary axis (same formula, with a larger constant, $\alpha_0 \to \alpha_{lsia}$, $\alpha_{lsia} \geq \alpha_{ls}$)

Computing Courant limits

To explicitly compute the limit of the Courant factor as a function of β^j, order of approximation, $2n$, advection stencil s, direction of advection, $\check{\epsilon}_j$, dissipation parameters, σ^j is not easy in the general case.

- In the particular case of a **centered scheme evolving a flat d-metric with zero shift**, however, dissipation is not needed and the Courant limit is easy to write down:

$$\lambda \leq \frac{\alpha_0}{2\sqrt{dC_n}}\,, \tag{3.29}$$

where C_n is given in (2.47) and α_0 stands for the constant of local stability on the imaginary axis.

In the general case one usually has to evaluate the Courant limit numerically by maximizing eq. (3.25) over $\check{\Omega}$.

- **For the 1-D wave equation with shift $\beta > 0$ with upwind discretization** of the advection term and adding the minimal amount of dissipation if necessary, the limit of the Courant factor is given by

$$\lambda^{(n,s)}(\beta) \equiv \frac{\alpha_0}{\max\limits_{\check{\Omega}\in(0,2]} \left| \beta\check{\Omega}^{2(n+1)} \left(\frac{\check{\mathbf{d}}^{(1,n,s)}}{\check{\Omega}^{2(n+1)}} - \sigma_+^{(n,s)} \right) + i \left(\beta\check{d}^{(1,n,s)} + \sqrt{\check{d}^{(2,n)}} \right) \right|} . \tag{3.30}$$

Fig. 3.1 shows the Courant limits for different orders of approximation at fixed advection stencil. Note that if $s = 0$, the higher the order of approximation, the lower the Courant limit. For $s \geq 1$, this is not true anymore beyond a certain value of the shift. For large shifts, increasing the order of approximation, actually *decreases* the Courant limit.

Fig. 3.2 compares Courant limits at fixed order of approximation for different advection stencils. By advecting points the Courant limit is decreased, and there is a significant drop in the Courant factor between $s = 1$ and $s = 2$, for all orders of approximation.

3.4 Dispersion and Dissipation

This section applies the general methodology introduced in 2.4 regarding the mode splitting, speeds and amplification factors, in the case of the wave equation.

The characteristics of the wave equation at the continuum level are $\hat{\mathsf{C}}_\pm \equiv \hat{\mathsf{K}} \pm i\hat{\Delta}\hat{\Phi}$. The Fourier coefficients of the main variables, $\hat{\mathsf{K}}$ and $\hat{\Phi}$ are a

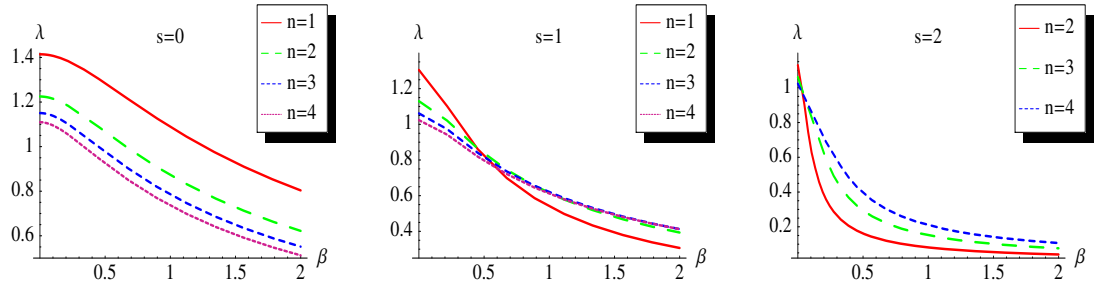


Figure 3.1: **Courant Factor Limit as a function of β, for different orders, at fixed off-centering** s For $s = 0$ (left plot) no dissipation is needed, ($\sigma = 0$), and the problem is in the regime of local stability on the imaginary axis ($\alpha_0 = 2.83$). For $s = 1$ (middle plot), again no dissipation is needed ($\sigma = 0$), but now the problem is in the regime of local stability ($\alpha_0 = 2.61$). For $s = 2$ (right plot) dissipation is required and the minimum amount is added in order to attain stability.

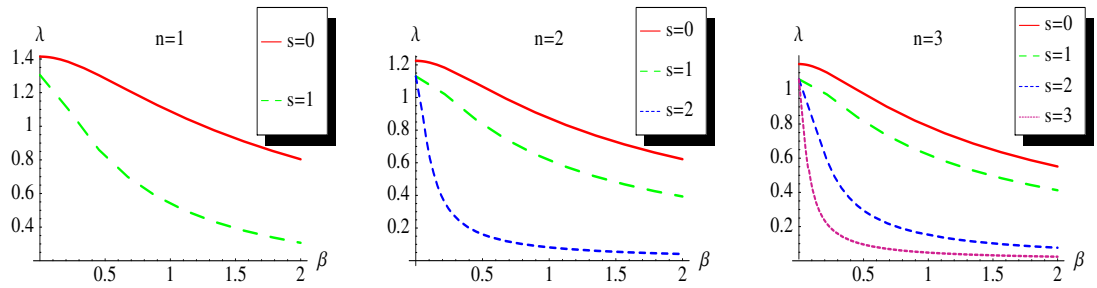


Figure 3.2: **Courant Factor Limit as a function of β, for different advection stencils, at fixed order of spatial accuracy**. From left to right: Courant limits at approximation orders $2, 4, 6$. As in Fig. 3.1 the Courant limit calculation takes into account whether the problem is in the regime of local stability on the imaginary axis (the case $s = 0$), or only local stability (for $s \geq 1$), and the minimal amount of Kreiss-Oliger dissipation is added for $s \geq 2$.

superposition of the $\pm$ modes:

$$\hat{\Phi}(\vec{\omega},t) = \frac{1}{2i\hat{\Delta}}\left(\hat{\mathsf{C}}_{0+}(\vec{\omega})e^{\hat{\Lambda}_+ t} - \hat{\mathsf{C}}_{0-}(\vec{\omega})e^{\hat{\Lambda}_- t}\right) \tag{3.31}$$

$$\hat{\mathsf{K}}(\vec{\omega},t) = \frac{1}{2}\left(\hat{\mathsf{C}}_{0+}(\vec{\omega})e^{\hat{\Lambda}_+ t} + \hat{\mathsf{C}}_{0-}(\vec{\omega})e^{\hat{\Lambda}_- t}\right) \tag{3.32}$$

Similarly, the discrete characteristics are $\hat{C}_\pm \equiv \hat{K} \pm i\Delta\hat{\Phi}$ and the discrete Fourier coefficients, $\hat{K}$ and $\hat{\Phi}$ are given by:

$$\hat{\Phi}(\underline{\omega},\underline{\xi},t) = \frac{1}{2i\hat{\Delta}}\left(\hat{C}_{0+}(\underline{\omega},\underline{\xi})e^{\hat{\Lambda}_+ t} - \hat{C}_{0-}(\underline{\omega},\underline{\xi})e^{\hat{\Lambda}_- t}\right) \tag{3.33}$$

$$\hat{K}(\underline{\omega},\underline{\xi},t) = \frac{1}{2}\left(\hat{C}_{0+}(\underline{\omega},\underline{\xi})e^{\hat{\Lambda}_+ t} + \hat{C}_{0-}(\underline{\omega},\underline{\xi})e^{\hat{\Lambda}_- t}\right) \tag{3.34}$$

Each mode has an associated pair of phase/group speeds defined in (2.115) for the continuum level and in (2.116) for the discrete.

Let $\beta^n = \beta^i n_i$ and $\gamma^{nn} = \gamma^{ij} n_i n_j$ where $n_i = \omega_i/\omega_0 = \xi_i/\xi_0$. The eigenvalues of the continuum (3.7) and of the discrete (3.26) problems determine the speeds and the amplification factor.

The continuum speeds are:

$$\mathsf{v}_{p\pm} = \mathsf{v}_{g\pm} = \beta^n \pm \sqrt{\gamma^{nn}} \tag{3.35}$$

while the discrete speeds are:

$$v_{p\pm} = \left(\sum_j \beta^j n_j \frac{\check{d}_j^{(1,n,s)}}{\xi_j}\right) \pm \frac{(h\hat{\Delta})(\underline{\xi})}{\xi_0} \tag{3.36}$$

$$v_{g\pm} = \sum_j \left[\beta^j n_j \frac{\partial \check{d}_j^{(1,n,s)}}{\partial \xi_j} \pm \frac{\partial (h\hat{\Delta})(\underline{\xi})}{\partial \xi_j} n_j\right] \tag{3.37}$$

The phase speed errors are obtained with $\epsilon_{p\pm} = v_{p\pm} - \mathsf{v}_{p\pm}$.

The amplification factor defined in (2.114) is the same for both $\pm$ modes:

$$a_p = \sum_j \left(\beta^j n_j \epsilon_j \frac{\check{\mathbf{d}}_j^{(1,n,s)}}{\xi_j} - \frac{1}{2^{2(n+1)}} \sigma^j n_j \frac{\check{\Omega}_j^{2(n+1)}}{\xi_j} \right) \tag{3.38}$$

In the regime of **small frequencies**, $\check{\Omega} \simeq \xi$ while the functions $\check{d}^{(1,n,s)}$, $\check{\mathbf{d}}^{(1,n,s)}$ and $\check{d}^{(2,n)}$ behave according to the relations (2.63). Then the amplification factor is easily computed:

$$\begin{aligned} a_p &\simeq \sum_j \mathbf{T}^j \xi_j^{2n+1} \quad \text{where} \\ \mathbf{T}^j &= \beta^j n_j \epsilon_j (-1)^s \frac{s(n+s)!(n-s)!}{2(n+1)(2n)!} - \frac{1}{2^{2(n+1)}} \sigma_j n_j \end{aligned} \tag{3.39}$$

In order to find out the behaviour of phase speed errors, $\epsilon_{p\pm}$, first the relation

$$\hat{\Delta}^2 \simeq \gamma^{nn} \omega_0^2 - \frac{\omega_0^2}{2\sqrt{\hat{\Delta}}} \frac{(n!)^2}{(2n)!} \left(\frac{2}{2n+1} \gamma^{nn} - \frac{1}{n+1} \sum_i \gamma^{ii} n_i^2 \right) \sum_i \xi_i^{2n}$$

is proven using the definition of $\hat{\Delta}$, (3.17), and the properties (2.63). Then the phase speed errors are:

$$\begin{aligned} \epsilon_{p\pm} &\simeq \sum_j T_\pm^j \xi_j^{2n} \quad \text{where} \\ T_\pm^j &= \beta^j n_j (-1)^s \frac{(n+s)!(n-s)!}{(2n+1)!} \pm \frac{(n!)^2}{(2n+1)!} \left(-\sqrt{\gamma^{nn}} + \frac{2n+1}{2(n+1)} \frac{\sum_i \gamma^{ii} n_i^2}{\sqrt{\gamma^{nn}}} \right) \end{aligned} \tag{3.40}$$

The eigenvalues of the continuum and of the discrete problems are related via:

$$\hat{\Lambda}_\pm \simeq \hat{\Lambda}_\pm + \omega_0 \left(\mathbf{T}^k \xi_k^{2n+1} + i T_\pm^k \xi_k^{2n} \right) \tag{3.41}$$

As shown in 2.4.2, for short enough time, that is, in the linear regime ($t \ll \frac{1}{\omega_0} \min\{\frac{1}{\epsilon_p}, \frac{1}{a_p}\}$) the relative errors of the semidiscrete problem, $\hat{E}_\pm$, grow linearly with time and are proportional with a_p and $\epsilon_\pm$. From (3.39)-(3.40),

one can see that, for small frequencies, the phase speed errors, $\epsilon_{\pm}$, scale with ξ^{2n} while the amplification factor, a_p with ξ^{2n+1}. This means that the accuracy of the phase speeds will be dominant in determining the accuracy of the semidiscrete scheme.

3.5 Phase and Group Speeds

In the following, the phase and the group speed errors are analyzed by restricting to the one dimensional case.

Because the speeds corresponding to positive and negative modes interchange when ξ changes sign, it is enough to consider only the "+" speed over the whole spectrum $\xi \in (-\pi, \pi]$. Also because the speeds are compared at different orders of approximation or at different stencils, the superscript (n, s) (or only (n) in case $s = 0$) will be attached to the symbols representing the discrete speeds and the corresponding errors:

$$v_p^{(n,s)}(\xi) = \frac{1}{\xi}\left(\beta \check{d}^{(1,n,s)} + \sqrt{\check{d}^{(2,n)}}\right), \tag{3.42}$$

$$v_g^{(n,s)}(\xi) = \frac{d}{d\xi}\left(\beta \check{d}^{(1,n,s)} + \sqrt{\check{d}^{(2,n)}}\right). \tag{3.43}$$

The continuum limits for both phase and group speeds are: $\beta + 1$ for $\xi > 0$ and $\beta - 1$ for $\xi < 0$. In the reminder of this section, the behavior of the speed errors defined as

$$\epsilon_p^{(n,s)} \equiv \beta\left(\frac{\check{d}^{(1,n,s)}}{\xi} - 1\right) + \left(\frac{\sqrt{\check{d}^{(2,n)}}}{\xi} - \text{sign } \xi\right), \tag{3.44}$$

$$\epsilon_g^{(n,s)} \equiv \beta\left(\frac{d}{d\xi}\check{d}^{(1,n,s)} - 1\right) + \left(\frac{d}{d\xi}\sqrt{\check{d}^{(2,n)}} - \text{sign } \xi\right) \tag{3.45}$$

will be analyzed in detail. Without restricting generality, the shift is assumed positive, $\beta \geq 0$. (If $\beta \to -\beta$, then $\epsilon_{p,g}^{(n,s)}(\xi) \to -\epsilon_{p,g}^{(n,s)}(-\xi)$.)

3.5.1 Small Frequencies

When $\xi \simeq 0$ one can show that the phase and group speed errors satisfy

$$\begin{aligned}
\epsilon_p^{(n,s)} &= -|c_n| \left[(-1)^s \frac{(n+s)!(n-s)!}{(n!)^2} \beta + \frac{\text{sign } \xi}{2(n+1)} \right] \xi^{2n} + O(\xi^{2n+2}) \,, \\
\epsilon_g^{(n,s)} &= -(2n+1)\,|c_n| \left[(-1)^s \frac{(n+s)!(n-s)!}{(n!)^2} \beta + \frac{\text{sign } \xi}{2(n+1)} \right] \xi^{2n} + O(\xi^{2n+2}) \,.
\end{aligned} \tag{3.46}$$

Because the errors scale with ξ^{2n}, for small enough frequencies, higher order approximations will improve the phase and group errors for all the values of the shift and for all advection stencils.
If the order, $2n$, is kept fixed and the speeds corresponding to an off-centering by $s \geq 1$-points are compared with the ones corresponding to the centered scheme, $s = 0$, then one can easily show (by comparing the coefficients of ξ^{2n} in the relations above) that the off-centered scheme improves over the centered one the accuracy of

- the "+" numerical speeds ($\xi > 0$) if s is odd and β is small enough
- the "−" numerical speeds ($\xi < 0$) if s is even and β is small enough

where small enough means

$$\beta < \frac{1}{(n+1)} \frac{1}{\frac{(n+s)!(n-s)!}{(n!)^2} - 1}. \tag{3.47}$$

Obs. For $s = 1$, the inequality (3.47) becomes $\beta < \frac{n}{(n+1)}$. Also notice that with increasing s the above limit on β decreases.
The subsections (3.5.2–3.5.4) analyse in some more detail the behavior of the numerical speeds over the whole spectrum.

Comparison with the First Order Form Wave Equation

By introducing the extra variable $\mathsf{X} = \partial_x \Phi$, the first order reduction of the wave equation (3.6) is obtained:

$$\begin{aligned} \partial_t \Phi &= \beta \mathsf{X} + \mathsf{K}\,, \\ \partial_t \mathsf{X} &= \beta \partial_x \mathsf{X} + \partial_x \mathsf{K}\,, \\ \partial_t \mathsf{K} &= \partial_x \mathsf{X} + \beta \partial_x \mathsf{K}\,. \end{aligned} \tag{3.48}$$

If the first derivatives from (3.48) are approximated using the centered FDO $D^{(1,n)}$, then the nonzero eigenvalues of the corresponding semidiscrete system are $(h\hat{\Lambda}_{\pm})(\xi) = i\,(\beta \pm 1)\,\check{d}^{(1,n)}$. For $\xi \simeq 0$ the speed errors behave according to:

$$\begin{aligned} \epsilon_p^{(n)} &= -\,(\beta + \text{sign } \xi)\,|c_n|\,\xi^{2n} + O(\xi^{2n+2}) \\ \epsilon_g^{(n)} &= -\,(\beta + \text{sign } \xi)\,(2n+1)\,|c_n|\,\xi^{2n} + O(\xi^{2n+2}). \end{aligned} \tag{3.49}$$

Now the relations (3.46) with $s = 0$ are compared with (3.49) by matching the corresponding coefficients of ξ^{2n}. This gives that at a given order of approximation, $2n$, if $|\beta| \leq \frac{2n+3}{4(n+1)}$, then the second order system discretized with centered FDO, has smaller phase and group errors than the first order system for both eigenvalues. If $|\beta| > \frac{2n+3}{4(n+1)}$ then one pair of speeds (phase and group) is better approximated by the second order system, while the other one is better approximated by the first order system.

3.5.2 Scaling of the Speeds Errors with the Order of Approximation when $\beta = 0$

Lemma 3.5.1 *If $\beta = 0$ then higher order approximations bring an improvement in the phase and group errors for all frequencies.*

Proof Using the relation (2.50) in the definitions of the speeds leads to:

$$v_p^{(n)} = \frac{\sqrt{\check{d}^{(2,n)}}}{\xi}, \qquad \epsilon_p^{(n)} = \frac{\sqrt{\check{d}^{(2,n)}}}{\xi} - \text{sign } \xi \tag{3.50}$$

$$v_g^{(n)} = \frac{\check{d}^{(1,n)}}{\sqrt{\check{d}^{(2,n)}}}, \qquad \epsilon_g^{(n)} = \frac{\check{d}^{(1,n)}}{\sqrt{\check{d}^{(2,n)}}} - \text{sign } \xi \tag{3.51}$$

Using the inequalities (2.64) and (2.65) one can easily show that $\left|\epsilon_p^{(n+1)}\right| < \left|\epsilon_p^{(n)}\right|$ and $\left|\epsilon_g^{(n+1)}\right| < \left|\epsilon_g^{(n)}\right|$ for all the frequencies. The situation is illustrated in Fig. 3.3 where the speeds $v_p^{(n)}$ and $v_g^{(n)}$ are ploted versus ξ.

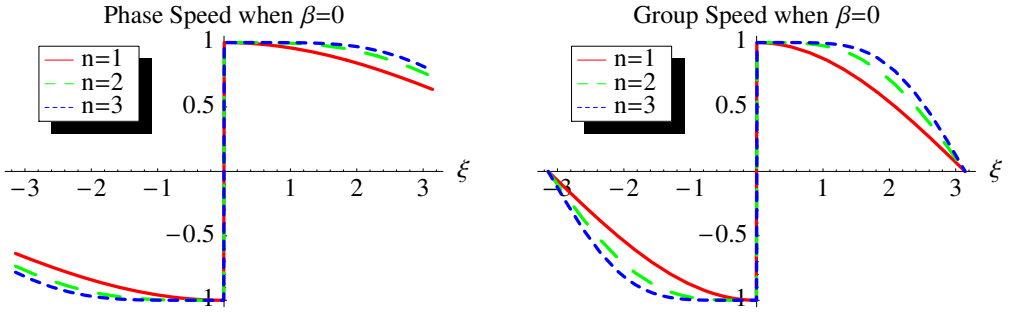


Figure 3.3: **Phase and Group Speeds for $\beta = 0$.** The higher the order of the approximation, the more accurate the phase and group speeds for all frequencies.

3.5.3 Scaling of the Speeds Errors with the Order of Approximation when $\beta \neq 0$

For $\beta \neq 0$ the situation is complicated by the presence of the shift terms that can be advected in different ways.

It will be shown that, at a fixed advection stencil (off-centering), it is not true anymore that higher order approximations improve the numerical speeds

for all frequencies (not even in the case of approximating the shift terms with centered FDOs!).

Phase Speeds

Lemma 3.5.2 *At a given order, n, at fixed advection stencil s, if $|\beta|$ is sufficiently small, then $\left|\epsilon_p^{(n+1,s)}\right| < \left|\epsilon_p^{(n,s)}\right|$, for all $\xi \in (\pi, \pi]$. Otherwise there are frequency intervals where $\left|\epsilon_p^{(n+1,s)}\right| > \left|\epsilon_p^{(n,s)}\right|$.*

In the following an argument of this lemma is presented for some particular cases of n and s. I call it "argument" and not "proof" because although for $s = 0$ and $n \geq 1$ the proof is analytic and complete, for other cases the problem can be only graphically investigated due to the transcendental nature of some functions that come in the analysis.

The graphical investigation is done for $s = \overline{1,3}$ and $n = \overline{1,5}$. It is conjectured that the lemma holds for all cases, $s \geq 1$ and $n \geq s$.

Argument Imposing $\left|\epsilon_p^{(n+1,s)}\right| < \left|\epsilon_p^{(n,s)}\right|$ and using the definition (3.44), gives the following inequality:

$$a_2^{(n,s)}(\xi)\beta^2 + 2a_1^{(n,s)}(\xi)\beta + a_0^{(n,s)}(\xi) > 0 \tag{3.52}$$

where $a_0^{(n,s)}$, $a_0^{(n,s)}$ and $a_2^{(n,s)}$ are functions of the frequency:

$$\begin{aligned}
a_2^{(n,s)}(\xi) &= \left(\check{d}^{(1,n+1,s)} - \check{d}^{(1,n,s)}\right)\left(2\xi - \check{d}^{(1,n,s)} - \check{d}^{(1,n+1,s)}\right) \\
a_1^{(n,s)}(\xi) &= \left(\check{d}^{(1,n,s)}\sqrt{\check{d}^{(2,n)}} - \check{d}^{(1,n+1,s)}\sqrt{\check{d}^{(2,n+1)}}\right) \\
&+ \xi\left(\sqrt{\check{d}^{(2,n+1)}} - \sqrt{\check{d}^{(2,n)}}\right) + |\xi|\left(\check{d}^{(1,n+1,s)} - \check{d}^{(1,n,s)}\right) \\
a_0^{(n,s)}(\xi) &= \check{d}^{(2,n)} - \check{d}^{(2,n+1)} + 2|\xi|\left(\sqrt{\check{d}^{(2,n+1)}} - \sqrt{\check{d}^{(2,n)}}\right)
\end{aligned} \tag{3.53}$$

The associated equation has real solutions because the determinant is positive

for all ξ:

$$\begin{aligned}\Delta(\xi) &= (a_1^{(n,s)}(\xi))^2 - a_0^{(n,s)}(\xi)a_2^{(n,s)}(\xi) \\ &= \left[\check{d}^{(1,n+1,s)}\sqrt{\check{d}^{(2,n)}} - \check{d}^{(1,n,s)}\sqrt{\check{d}^{(2,n+1)}}\right. \\ &+ \left.|\xi|\left(\check{d}^{(1,n,s)} - \check{d}^{(1,n+1,s)}\right) + \xi\left(\sqrt{\check{d}^{(2,n+1)}} - \sqrt{\check{d}^{(2,n)}}\right)\right]^2 \geq 0\end{aligned} \tag{3.54}$$

Define:

$$\begin{aligned}f_1^{(n,s)}(\xi) &\equiv \check{d}^{(1,n+1,s)} - \check{d}^{(1,n,s)} \\ f_2^{(n,s)}(\xi) &\equiv 2\xi - \check{d}^{(1,n,s)} - \check{d}^{(1,n+1,s)} \\ g_1^{(n)}(\xi) &\equiv -\left[\sqrt{\check{d}^{(2,n+1)}} - \sqrt{\check{d}^{(2,n)}}\right] \\ g_2^{(n)}(\xi) &\equiv -\left[2|\xi| - \sqrt{\check{d}^{(2,n+1)}} - \sqrt{\check{d}^{(2,n)}}\right].\end{aligned} \tag{3.55}$$

Then it's easy to show that $a_2^{(n,s)}$ and the roots of the equations (3.52), $\beta_{1,2}^{(n,s)}$ satisfy:

$$\begin{aligned}a_2^{(n,s)}(\xi) &= f_1^{(n,s)}f_2^{(n,s)} \\ \beta_1^{(n,s)}(\xi) &= \frac{g_1^{(n)}}{f_1^{(n,s)}} \\ \beta_2^{(n,s)}(\xi) &= \frac{g_2^{(n)}}{f_2^{(n,s)}}\end{aligned} \tag{3.56}$$

By inequality (2.64), $g_{1,2}^{(n)}(\xi) < 0$ for all $\xi \in (-\pi, 0) \cup (0, \pi]$. This gives sign $(a_2^{(n,s)}) =$ sign $(\beta^1\beta^2)$ and the following proposition holds.

Proposition 3.5.3 *The phase speed at a certain frequency is better approximated by the next order of approfimation, ($\left|\epsilon_p^{(n+1,s)}\right| < \left|\epsilon_p^{(n,s)}\right|$), if and only if one of the following two cases holds:*

1. *the roots $\beta_{1,2}^{(n,s)}(\xi)$ have the same sign and β is outside the interval be-*

tween the roots

2. *the roots $\beta_{1,2}^{(n,s)}(\xi)$ have opposite sign and β lies inside the interval between the roots*

The solution to this problem requires determining the zeros and the signs of the functions $f_{1,2}^{(n,s)}$, establishing the monotony of $\beta_{1,2}^{(n,s)}(\xi)$, and solving the equations $\beta = \beta_{1,2}^{(n,s)}(\xi)$. For general n and s, this is not an trivial task. Some properties of these functions can be easily inferred from the properties of Fourier symbols (2.1.4 -–2.1.5) and are listed bellow:

- parity: $f_{1,2}^{(n,s)}(\xi)$ are odd functions, while $g_{1,2}^{(n)}(\xi)$, $a_2^{(n,s)}(\xi)$ and $\beta_{1,2}^{(n,s)}(\xi)$ are even functions.

- values and limits in $\xi = 0$:

$$f_{1,2}^{(n,s)}(0) = g_{1,2}^{(n)}(0) = 0$$

$$\lim_{\xi \searrow 0} \beta_{1,2}^{(n,s)} = -\lim_{\xi \nearrow 0} \beta_{1,2}^{(n,s)} = (-1)^{s+1} \frac{(n!)^2}{2(n+1)(n-s)!(n+s)!} \tag{3.57}$$

- values and limits in $\xi = \pi$:

$$\begin{aligned} f_1^{(n,s)}(\pi) &= 0 \qquad & g_1^{(n)}(\pi) &= -2\left(\sqrt{C_{n+1}} - \sqrt{C_n}\right) \\ f_2^{(n,s)}(\pi) &= 2\pi \qquad & g_2^{(n)}(\pi) &= -2\left(\pi - \sqrt{C_{n+1}} - \sqrt{C_n}\right) \\ \left|\lim_{\xi \nearrow \pi} \beta_1^{(n,s)}(\xi)\right| &= \infty \qquad & \beta_2^{(n,s)}(\pi) &< 0 \end{aligned} \tag{3.58}$$

In the particular **case** $s = 0$,

$$\begin{aligned} f_1^{(n,0)}(\xi) &= \hat{\delta}\,|c_n|\,\tilde{\Omega}^{2n} \\ f_2^{(n,0)}(\xi) &= (\xi - \check{d}^{(1,n,0)}) + (\xi - \check{d}^{(1,n+1,0)}) \end{aligned} \tag{3.59}$$

Because of the symmetry properties of these functions in respect to the y-axis, it is enough to restrict the analysis to $\xi > 0$. One can analytically

determine the roots of $f_{1,2}^{(n,0)}(\xi)$ (0 and π for $f_1^{(n,0)}$, and 0 for $f_2^{(n,0)}(\xi)$), the sign (both positive), and the monotony of $\beta_{1,2}^{(n,0)}(\xi)$ (descending), and show that the equations $\beta = \beta_{1,2}^{(n,0)}(\xi)$ have no solution for $\xi \in (0,\pi)$ and at most one solution each for $\xi \in (-\pi, 0)$ (denoted with $\xi_{1,2}^{-}$).

It follows that the inequality $\left|\epsilon_p^{(n+1,0)}\right| < \left|\epsilon_p^{(n,0)}\right|$ is satisfied if and only if one of the following cases holds:

- $0 < \beta < \frac{1}{2(n+1)}$ and $\xi \in (-\pi, 0) \cup (0, \pi]$
- $\frac{1}{2(n+1)} < \beta < \left|\beta_2^{(n,0)}(\pi)\right|$ and $\xi \in (-\pi, \xi_2^-) \cup (\xi_1^-, 0) \cup (0, \pi)$
- $\left|\beta_2^{(n,0)}(\pi)\right| < \beta$ and $\xi \in (\xi_1^-, 0) \cup (0, \pi)$

In the **case** $s = 1$,

$$\begin{aligned} f_1^{(n,1)}(\xi) &= -\hat{\delta}\,|c_n|\,\tilde{\Omega}^{2n}\left(\frac{1}{n} + \cos(\xi)\right) \\ f_2^{(n,1)}(\xi) &= 2(\xi - \check{d}^{(1,n+1,0)}) + \hat{\delta}\,|c_n|\,\tilde{\Omega}^{2n}\left(\frac{1}{n} + 2 - \cos(\xi)\right) \end{aligned} \tag{3.60}$$

While the roots of $f_1^{(n,1)}$ are easy to compute analytically, the roots of $f_2^{(n,1)}$ require numerical evaluation due to the transcendental nature of the function. For $s \geq 2$ the functions are even more difficult to analyze, involving also evaluation of transcendental equations. From now on, the proof is limited to plotting the quantities $\beta_{1,2}^{(n,s)}(\xi)$ (see Fig. 3.4) and interpreting the figures according to the proposition 3.5.3.

Interpreting the plots: The plots from Fig.3.4 show that, at a given order $2n$, if β is sufficiently small, that is, if $|\beta| < \min_{\xi\in(0,\pi)} \left|\beta_{1,2}^{(n,s)}(\xi)\right|$, then $\left|\epsilon_p^{(n+1,s)}\right| < \left|\epsilon_p^{(n,s)}\right|$ for the whole spectrum. Otherwise this relation will hold everywhere apart from some intervals. The number of intervals, their location and their length depend on s, n and β, and they are difficult to determine analytically. For sufficiently large β, these intervals will be located close to the branches of discontinuity of $\beta_{1,2}^{(n,s)}(\xi)$. The graphs show that the number

of intervals increases with s and also that their length decreases with n. So, by increasing the off-centering, there will be a better scaling with the order of approximation.

As an example for the previous considerations, Fig. 3.5 shows the phase speeds at different orders of approximation with fixed advection stencil at a particular value of the shift, $\beta = 0.5$.

Group Speeds

Lemma 3.5.4 *At a given order $2n$, and fixed advection stencil s, the inequality $\left|\epsilon_g^{(n+1,s)}\right| < \left|\epsilon_g^{(n,s)}\right|$, does not hold for all $\xi \in (-\pi, \pi]$.*

As in the case of phase speeds, it is hard to give a complete proof due to the analysis of some transcendental equations.

Argument Imposing $\left|\epsilon_g^{(n+1,s)}\right| < \left|\epsilon_g^{(n,s)}\right|$ gives the following inequality:

$$a_2^{(n,s)}(\xi)\beta^2 + 2a_1^{(n,s)}(\xi)\beta + a_0^{(n,s)}(\xi) > 0 \tag{3.61}$$

where

$$\begin{aligned}
a_2^{(n,s)}(\xi) &= \left(\partial_\xi \check{d}^{(1,n+1,s)} - \partial_\xi \check{d}^{(1,n,s)}\right)\left(2 - \partial_\xi \check{d}^{(1,n,s)} - \partial_\xi \check{d}^{(1,n+1,s)}\right) \\
a_1^{(n,s)}(\xi) &= \left(\partial_\xi \check{d}^{(1,n,s)} \partial_\xi \sqrt{\check{d}^{(2,n)}} - \partial_\xi \check{d}^{(1,n+1,s)} \partial_\xi \sqrt{\check{d}^{(2,n+1)}}\right) \\
&+ \left(\partial_\xi \sqrt{\check{d}^{(2,n+1)}} - \partial_\xi \sqrt{\check{d}^{(2,n)}}\right) + \text{sign } \xi \left(\partial_\xi \check{d}^{(1,n+1,s)} - \partial_\xi \check{d}^{(1,n,s)}\right) \\
a_0^{(n,s)}(\xi) &= \partial_\xi \check{d}^{(2,n)} - \partial_\xi \check{d}^{(2,n+1)} + 2\,\text{sign } \xi \left(\partial_\xi \sqrt{\check{d}^{(2,n+1)}} - \partial_\xi \sqrt{\check{d}^{(2,n)}}\right)
\end{aligned} \tag{3.62}$$

One can show that the associated equation has two real solutions (the deter-

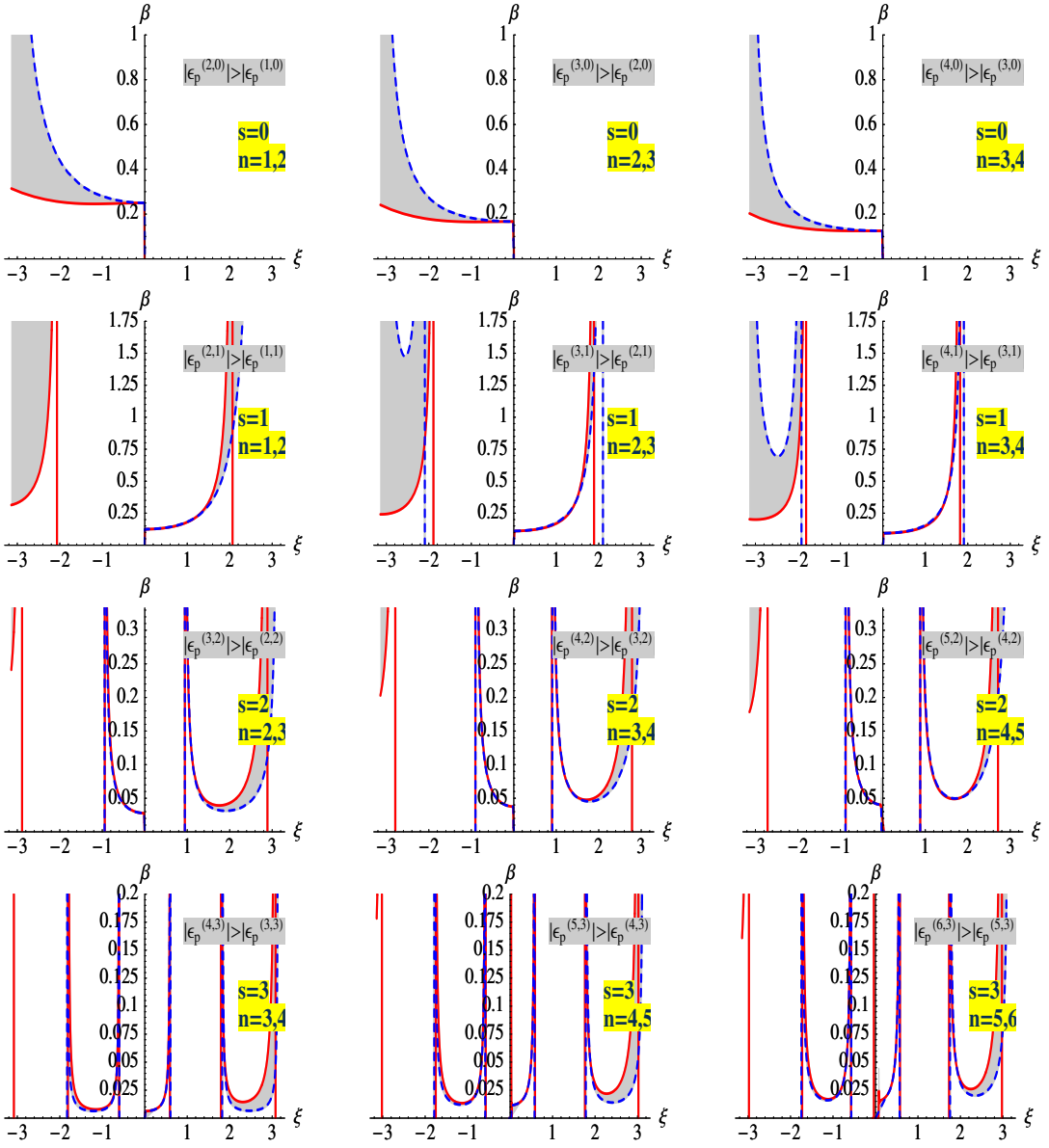


Figure 3.4: **Scaling of the phase speed errors with the order of approximation at different advection stencils** Shown are the regions where at a fixed advection stencil ($s = 0, 1, 2, 3$) the phase speed error does not scale with the order of approximation ($\left|\epsilon_p^{(n+1,s)}\right| > \left|\epsilon_p^{(n,s)}\right|$). The regions are delimited by the quantities $\beta_{1,2}^{(n,s)}$ defined in (3.56).

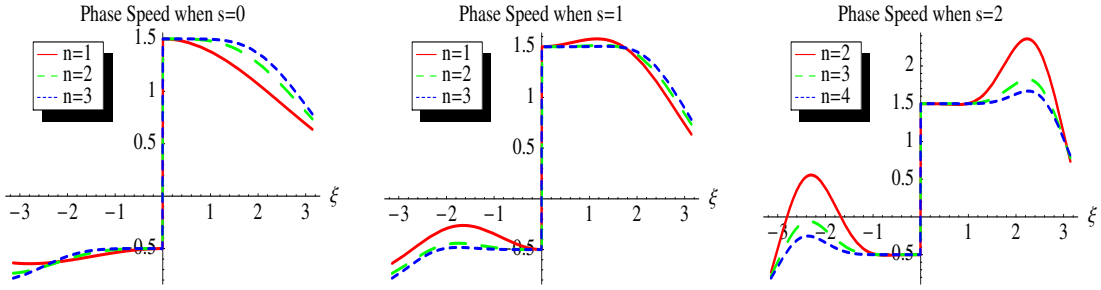


Figure 3.5: **The phase speeds at different orders of approximation with the same advection stencil when** $\beta = 0.5$

minant is positive). Now introduce:

$$F_1^{(n,s)} \equiv \partial_\xi f_1^{(n,s)}, \qquad G_1^{(n)} \equiv \partial_\xi g_1^{(n)} = -\left(\frac{\check{d}^{(1,n+1)}}{\sqrt{\check{d}^{(2,n+1)}}} - \frac{\check{d}^{(1,n)}}{\sqrt{\check{d}^{(2,n)}}}\right),$$
$$F_2^{(n,s)} \equiv \partial_\xi f_2^{(n,s)}, \qquad G_2^{(n)} \equiv \partial_\xi g_2^{(n)} = -\left(2\text{sign }\xi - \frac{\check{d}^{(1,n+1)}}{\sqrt{\check{d}^{(2,n+1)}}} - \frac{\check{d}^{(1,n)}}{\sqrt{\check{d}^{(2,n)}}}\right) \tag{3.63}$$

where $f_{1,2}^{(n,s)}$ and $g_{1,2}^{(n)}$ are defined in (3.55). Then $a_0^{(n,s)}(\xi)$ and the roots $\beta_{1,2}^{(n,s)}(\xi)$ satisfy:

$$\begin{aligned} a_2^{(n,s)}(\xi) &= F_1^{(n,s)} F_2^{(n,s)} \\ \beta_1^{(n,s)}(\xi) &= \frac{G_1^{(n)}}{F_1^{(n,s)}} \\ \beta_2^{(n,s)}(\xi) &= \frac{G_2^{(n)}}{F_2^{(n,s)}} \end{aligned} \tag{3.64}$$

The properties of the Fourier symbols (2.1.4–2.1.5) give $G_{1,2}^{(n)}(\xi) < 0$ for $\xi \in (0, \pi]$ and $G_{1,2}^{(n)}(\xi) > 0$ for $\xi \in (-\pi, 0)$. This means sign $(a_2^{(n,s)})$ = sign $(\beta^1\beta^2)$, so a similar proposition as in the case of the phase speeds analysis, follows:

Proposition 3.5.5 *The inequality* $\left|\epsilon_g^{(n+1,s)}\right| < \left|\epsilon_g^{(n,s)}\right|$ *holds if and only if one of the following statements is true:*

1. *the roots $\beta_{1,2}^{(n,s)}(\xi)$ have the same sign and β is outside the interval between the roots*

2. *the roots $\beta_{1,2}^{(n,s)}(\xi)$ have opposite sign and β lies inside the interval between the roots*

The functions $\beta_{1,2}^{(n,s)}(\xi)$ will not be analyzed in detail (apart from the case of $s = 0$). However, some properties of $F_{1,2}^{(n,s)}(\xi)$ and $G_{1,2}^{(n)}(\xi)$ are listed below.
some properties

- parity: $F_{1,2}^{(n,s)}(\xi)$ are even functions, while $G_{1,2}^{(n)}(\xi)$, $a_2^{(n,s)}(\xi)$ and $\beta_{1,2}^{(n,s)}(\xi)$ are odd functions.

- values and limits in $\xi = 0$:

$$F_{1,2}^{(n,s)}(0) = \lim_{\xi \searrow 0} G_{1,2}^{(n)}(\xi) = \lim_{\xi \nearrow 0} G_{1,2}^{(n)}(\xi) = 0$$

$$\lim_{\xi \searrow 0} \beta_{1,2}^{(n,s)} = -\lim_{\xi \nearrow 0} \beta_{1,2}^{(n,s)} = (-1)^{s+1} \frac{(n!)^2}{2(n+1)(n-s)!(n+s)!} \tag{3.65}$$

In the **case** $s = 0$,

$$\begin{aligned} F_1^{(n,0)}(\xi) &= (n+1)\,|c_n|\,\tilde{\Omega}^{2n}\left(\frac{n}{n+1} + \cos\xi\right) \\ F_2^{(n,0)}(\xi) &= (2n+1)\,|c_n|\,\tilde{\Omega}^{2n}\left(1 + \frac{n+1}{2(2n+1)}\tilde{\Omega}^2\right) \end{aligned} \tag{3.66}$$

One can analytically show that the equation $\beta = \beta_1^{(n,0)}(\xi)$ has one solution for in $\xi \in (0, \pi)$ (denoted with ξ_1^+) and at most one solution for $\xi \in (-\pi, 0)$ (denoted by ξ_1^-). Also the equation $\beta = \beta_2^{(n,0)}(\xi)$ has no solution for $\xi \in (0, \pi)$ and at most one solution for $\xi \in (-\pi, 0)$ (denoted by ξ_2^-).

The inequality $\left|\epsilon_g^{(n+1,0)}\right| < \left|\epsilon_g^{(n,0)}\right|$ holds

- if $0 < \beta < \frac{1}{2(n+1)}$ and $\xi \in (-\pi, 0) \cup (0, \xi_1^+)$

- if $\frac{1}{2(n+1)} < \beta < \left|\beta_2^{(n,0)}(\pi)\right|$ and $\xi \in (-\pi, \xi_2^-) \cup (\xi_1^-, 0) \cup (0, \xi_1^+)$,

- if $\left|\beta_2^{(n,0)}(\pi)\right| < \beta$ and $\xi \in (-\xi_1^-, 0) \cup (0, \xi_1^+)$

In the case of centered FDO, notice that, in contrast with the phase speed analysis, no matter how small is the shift β, there are regions in the spectrum where the error does not scale with the order of approximation. However, these regions are located at high frequencies. And this actually holds also for non-centered schemes. In figure 3.6 these roots are plotted against frequency for some particular advection stencils.

Interpreting the plots If $s = 1$ and $n = 1$ then for sufficiently small β, $\left|\epsilon_g^{(2,1)}\right| < \left|\epsilon_g^{(1,1)}\right|$ for the whole spectrum. In all the other cases there will be regions in the intervals where the error does not scale with the order of approximation. Also notice that with increasing the off-centering there is an overall improvement in scaling with the order of approximation.

As an illustration on a particular example, Fig. 3.7 shows the group speeds at different orders of approximation with the same advection stencil when $\beta = 0.5$.

3.5.4 Speeds Errors for Different Off-Centerings at the Same Order of Approximation

Phase Speeds

Imposing $\left|\epsilon_{p\pm}^{(n,s)}\right| < \left|\epsilon_{p\pm}^{(n,0)}\right|$ and using the definition (3.44) yields the inequality

$$f_1^{(n,s)}(\xi) f_2^{(n,s)}(\xi) \left(\beta - \beta^{(n,s)}(\xi)\right) < 0\,, \tag{3.67}$$

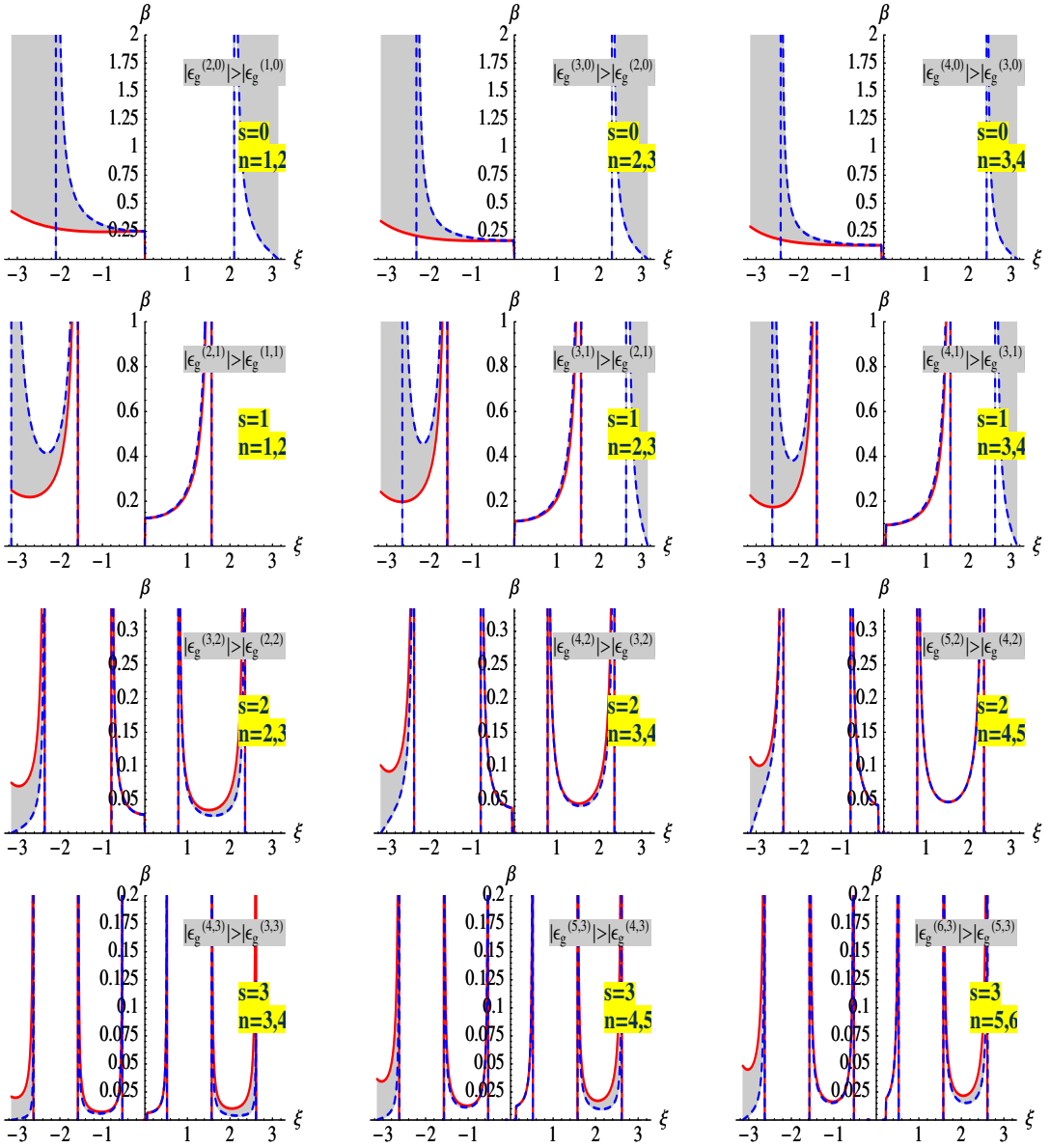


Figure 3.6: **Scaling of the group speed errors with the order of approximation at different advection stencils** Shown are the regions where at a fixed advection stencil ($s = 0, 1, 2, 3$) the group speed error does not scale with the order of approximation ($\left|\epsilon_g^{(n+1,s)}\right| > \left|\epsilon_g^{(n,s)}\right|$). The regions are delimited by the quantities $\beta_{1,2}^{(n,s)}$ defined in (3.64).

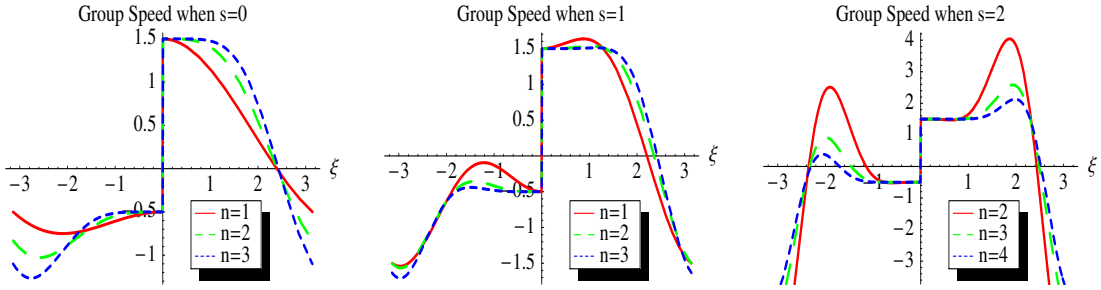


Figure 3.7: **The group speeds at different orders of approximation with the same advection stencil when** $\beta = 0.5$

where

$$\begin{aligned} f_1^{(n,s)}(\xi) &\equiv \check{d}^{(1,n,s)} - \check{d}^{(1,n,0)}\,, \\ f_2^{(n,s)}(\xi) &\equiv \check{d}^{(1,n,s)} + \check{d}^{(1,n,0)} - 2\xi\,, \\ g^{(n)}(\xi) &\equiv 2\left(|\xi| - \sqrt{\check{d}^{(2,n)}}\right)\,, \\ \beta^{(n,s)}(\xi) &\equiv \frac{g^{(n)}(\xi)}{f_2^{(n,s)}(\xi)}\,. \end{aligned} \tag{3.68}$$

The function $g^{(n)}$ satisfies $g^{(n)}(\xi) > 0$ but $f_{1,2}^{(n,s)}$ can change sign over the spectrum. The inequality (3.67) holds at a given frequency ξ, if $\beta > \beta^{(n,s)}(\xi)$ and sign $f_{1,2}^{(n,s)}(\xi) < 0$ or $\beta < \beta^{(n,s)}(\xi)$ and sign $f_{1,2}^{(n,s)}(\xi) > 0$. In general, the regions in (ξ, β) plane where at fixed order of approximation, off-centering by s points improves the accuracy of the phase speed, are difficult to determine analytically and the proof is restricted to a numerical evaluation (see Fig. 3.73). These plots tell that if s is odd (even), then for sufficiently small β, the "+" ("−") speed has smaller error compared with the case of centered FDO in some intervals of the spectrum that include the small frequency range. However these regions become narrower with increasing the off-centering, so that the strongest effect appears for $s = 1$. This case is analyzed in more detail below.

If $s = 1$, the functions $f_{1,2}^{(n,s)}$ and $\beta^{(n,s)}$, defined in (3.68) become

$$\begin{aligned} f_1^{(n,1)}(\xi) &= \frac{|c_{n-1}|}{2}(\sin\xi)\check{\Omega}_j^{2n}\,, \\ f_2^{(n,1)}(\xi) &= \hat{\delta}\frac{|c_{n-1}|}{2}\check{\Omega}_j^{2n} + 2\left(\check{d}^{(1,n)} - \xi\right)\,, \\ \beta^{(n,1)}(\xi) &= \frac{g^{(n)}(\xi)}{f_2^{(n,1)}(\xi)}\,. \end{aligned} \tag{3.69}$$

We have sign $f_1^{(n,1)}(\xi) = \text{sign}\ \xi$ and $f_1^{(n,1)}(\pm\pi) = 0$.
Then the inequality $\left|\epsilon_p^{(n,1)}\right| < \left|\epsilon_p^{(n,0)}\right|$ holds

- for $\xi > 0$ if $\beta^{(n,1)}(\xi) < 0$ or $0 < \beta < \beta^{(n,1)}(\xi)$,
- for $\xi < 0$ if $\beta > \beta^{(n,1)}(\xi) > 0$.

The values and limits of $\beta^{(n,1)}(\xi)$ in 0 and π are

$$\beta^{(n,1)}(\pi) = -1 + 2\frac{\sqrt{C_n}}{\pi} < 0\,, \tag{3.70}$$

$$\lim_{\xi\searrow 0}\beta^{(n,1)}(\xi) = -\lim_{\xi\nearrow 0}\beta^{(n,1)}(\xi) = \frac{n}{n+1}\,. \tag{3.71}$$

One can show that the equation $\beta = \beta^{(n,1)}(\xi)$ has at most one solution in each of the branches $\xi > 0$ and $\xi < 0$, which will be denoted by ξ^+ and ξ^-, respectively. Also if ξ_2 is the zero of the function $f_2^{(n)}$ in $(0,\pi)$, then $\xi^+ \in (0,\xi_2)$ and $\xi^- \in (-\pi,-\xi_2)$. These zeros can be evaluated numerically: for $n = 1,2,3,4$ they are $\frac{\pi}{2.27862}, \frac{\pi}{3.1304}, \frac{\pi}{3.81538}, \frac{\pi}{4.40246}$. The inequality $\left|\epsilon_p^{(n,1)}\right| < \left|\epsilon_p^{(n,0)}\right|$ holds if

- $\beta < 1 - 2\frac{\sqrt{C_n}}{\pi}$ and $\xi \in (0,\pi)$,
- $1 - 2\frac{\sqrt{C_n}}{\pi} < \beta < \frac{n}{n+1}$ and $\xi \in (-\pi,\xi^-)\cup(0,\pi)$,
- $\beta > \frac{n}{n+1}$ and $\xi \in (-\pi,\xi^-)\cup(\xi^+,\pi)$.

The previous results can be formulated now in the following:

Lemma 3.5.6 *At a given order of approximation, $2n$,*

1. *if $\beta < \frac{n}{n+1}$, then the "+" speed has smaller error in the case of one-point advected scheme than in the case of centered scheme, for all frequencies $0 < \xi \leq \pi$, but the "−" speed will have larger error, at least for small and medium frequencies.*

2. *if $\beta > \frac{n}{n+1}$ then for both $\pm$ speeds, in the regime of small frequencies, the centered scheme has smaller error than the one-point advected scheme, while for medium and high frequencies the situation reverses. The interval of small frequencies where the centered algorithm is more accurate than the advected one shrinks with increasing order of approximation.*

As an illustration on a particular example, Fig 3.9 shows the phase speeds at different advection stencils with the same order of approximation when $\beta = 0.5$.

Group Speeds

Imposing $\left|\epsilon_g^{(n,s)}\right| < \left|\epsilon_g^{(n,0)}\right|$ and using the definition (3.45) yields the inequality

$$F_1^{(n,s)}(\xi) F_2^{(n,s)}(\xi) \left(\beta - \beta^{(n,s)}(\xi)\right) < 0\,, \tag{3.72}$$

where

$$\begin{aligned} F_1^{(n,s)}(\xi) &\equiv \partial_\xi f_1^{(n,s)}(\xi)\,, \\ F_2^{(n,s)}(\xi) &\equiv \partial_\xi f_2^{(n,s)}(\xi)\,, \\ G^{(n)}(\xi) &\equiv \partial_\xi g^{(n)}(\xi)\,, \\ \beta^{(n,s)}(\xi) &\equiv \frac{G^{(n)}(\xi)}{F_2^{(n,s)}(\xi)}\,, \end{aligned} \tag{3.73}$$

and $f_{1,2}^{(n,s)}$ and $g^{(n)}$ are given by (3.68). It is easy to see that $G^{(n)}(\xi) = -G^{(n)}(-\xi)$. However the signs of $F_{1,2}^{(n,s)}(\xi)$ are more difficult to determine.

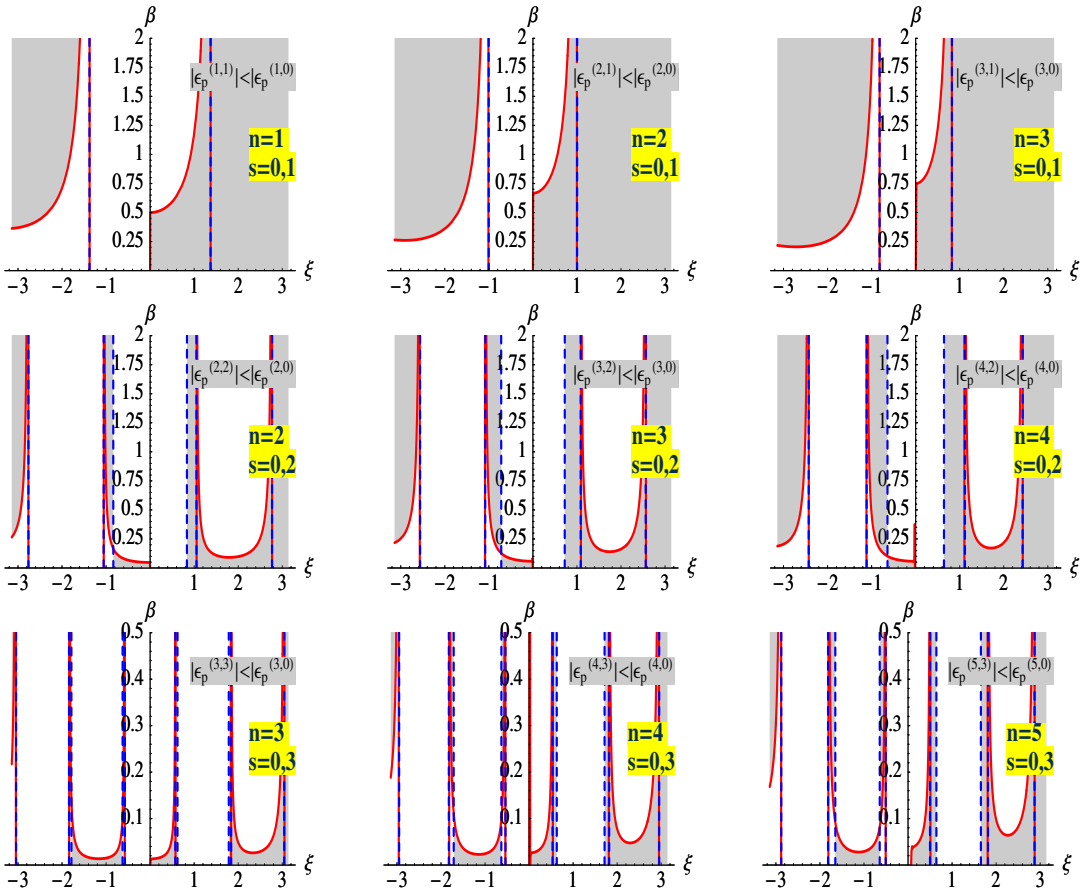


Figure 3.8: Shown are the regions where advected stencils improve the phase speed error over the centered scheme. The regions are delimited by the quantity $\beta^{(n,s)}$ and the zeros of the function $f_1^{(n,s)}$ as defined in (3.68).

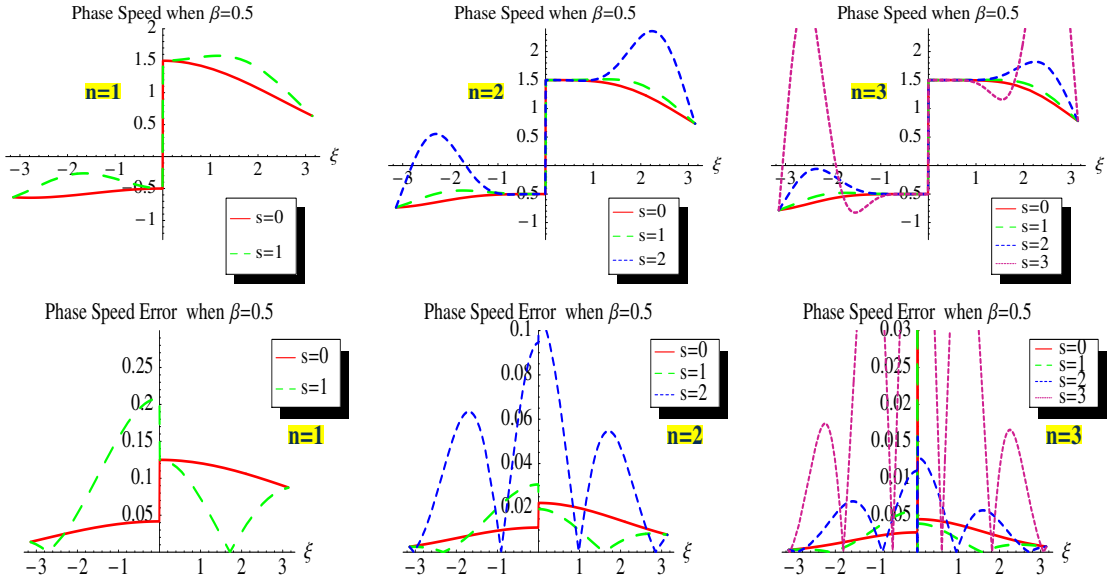


Figure 3.9: **The phase speeds at different advection stencils with the same order of approximation when** $\beta = 0.5$ top: phase speeds; bottom: phase speed error in absolute value scaled with ξ^{2n}.

As in the case of phase speeds analysis, the regions in (ξ, β) plane where at fixed order of approximation, off-centering by s points improves the accuracy of the group speed, are determined graphically (see Fig. 3.73). The same qualitative behavior appears as for the phase speeds, in the sense that for sufficiently small β, the "+" ("−") speed has smaller error compared with the centered scheme at least at small frequencies, and off-centering decreases of the extent of these regions in (ξ, β) parameter space.
The particular **case** $s = 1$ is analyzed in more detail below. The relations (3.73) become

$$\begin{aligned} F_1^{(n,1)}(\xi) &= \frac{(n+1)\,|c_{n-1}|}{2}\left(\frac{n}{n+1}+\cos\xi\right)\check{\Omega}_j^{2n} \\ F_2^{(n,1)}(\xi) &= \frac{(n+1)\,|c_{n-1}|}{2}\left(-\frac{n}{n+1}+\cos\xi\right),\check{\Omega}_j^{2n}\,, \\ \beta^{(n,1)}(\xi) &= \frac{G^{(n)}(\xi)}{F_2^{(n,1)}(\xi)}\,. \end{aligned} \tag{3.74}$$

Notice that $F_{1,2}^{(n,1)}(\xi) = F_{1,2}^{(n,1)}(-\xi)$, for $\xi \in (-\pi, \pi]$, $F_1^{(n,1)}(\xi) > 0$ for $\xi \in (0, \pi - \arccos(\frac{n}{n+1}))$, $F_2^{(n,1)}(\xi) > 0$ for $\xi \in (0, \arccos(\frac{n}{n+1}))$.

Then the inequality $\left|\epsilon_g^{(n,1)}\right| < \left|\epsilon_g^{(n,0)}\right|$ holds

- for $\xi > 0$ if $\xi \in (\xi_2, \pi - \arccos\frac{n}{n+1}) \subset (\arccos\frac{n}{n+1}, \pi - \arccos\frac{n}{n+1})$,
- for $\xi < 0$ if $\xi \in (-\pi + \arccos\frac{n}{n+1}, -\xi_2) \subset (-\pi + \arccos\frac{n}{n+1}, -\arccos\frac{n}{n+1})$.

The previous results are put now in the following:

Lemma 3.5.7 *At a given order of approximation* $2n$*,*

1. *if* $\beta < \frac{n}{n+1}$*, the "+" group speed has smaller error in the case of one-point advected scheme than in the case of centered scheme for all frequencies* $0 < \xi < \pi - \arccos\frac{n}{n+1}$*, but the "−" speed will have larger error, at least for small and mid frequencies.*

2. *if $\beta > \frac{n}{n+1}$ then for both $\pm$ speeds, in the regime of small frequencies, the centered scheme has smaller error than the one-point advected scheme, while for mid and high frequencies, the situation reverses. The interval of small frequencies where the centered algorithm is more accurate than advected one narrows with increasing the order of approximation.*

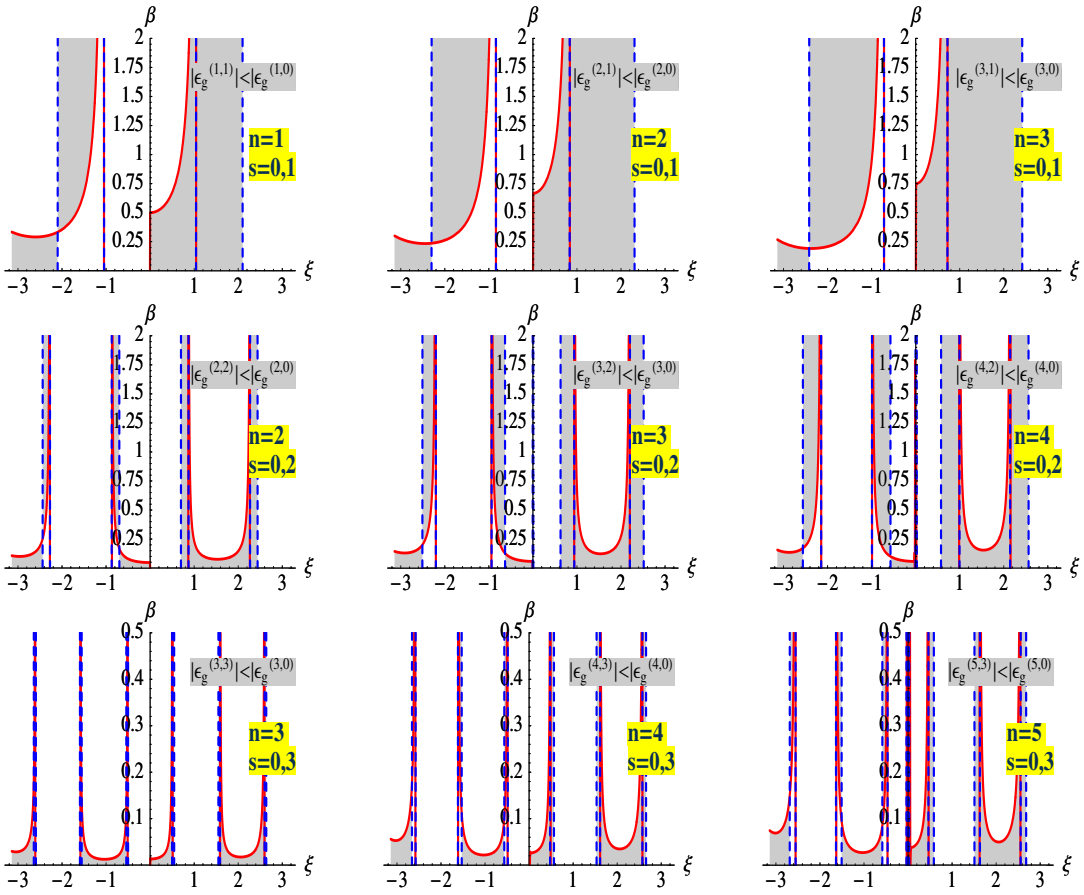


Figure 3.10: Shown are the regions where, at fixed order of approximation, advected stencils bring an improvement in the group speed error in comparison with the centered scheme. The regions are delimited by the quantity $\beta_2^{(n,s)}$ and the zeros of the function $F_1^{(n,s)}$, both defined in (3.73)

As an illustration on a particular example, Fig. 3.11 shows the group speeds

at different advection stencils with the same order of approximation when $\beta = 0.5$.

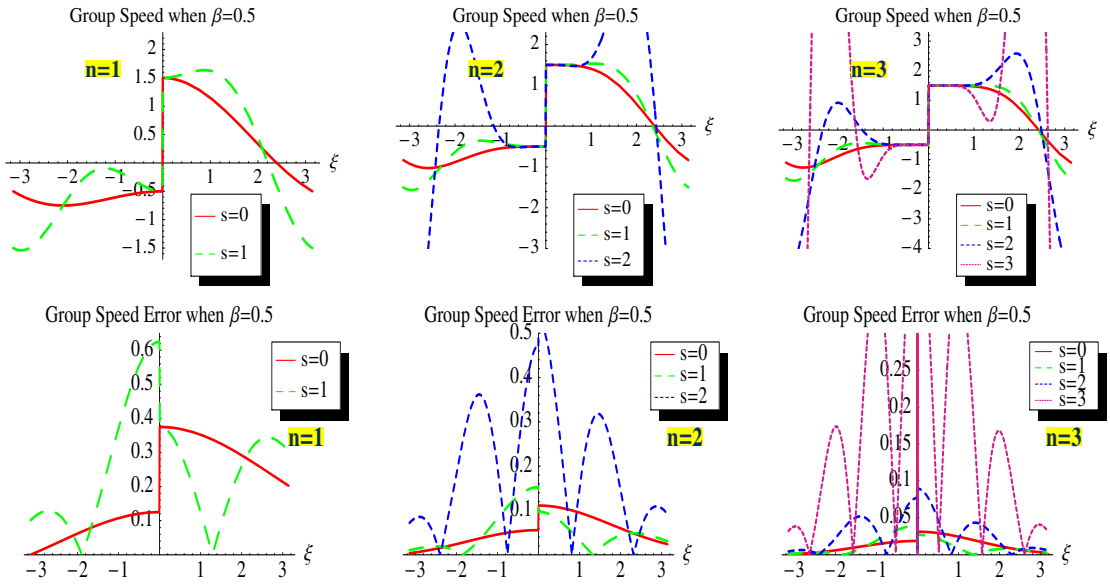


Figure 3.11: **The group speeds at different advection stencils with the same order of approximation when** $\beta = 0.5$ top: group speeds; bottom: group speed error in absolute value scaled with ξ^{2n}.

3.6 Numerical Experiments in 1-D

This section shows the results of some numerical tests performed with the 1-D wave equation and periodic boundary conditions. The tests are grouped in two categories:

1. the first set of tests compares the accuracy of the centered scheme with the one-point upwinded scheme.

2. the second set investigates the accuracy and convergence of high order centered schemes

(a) when different orders Runge-Kutta-integrators are used

(b) when the resolution and/or the Courant factor are varied.

In all the tests, the accuracy and convergence is measured in respect to the analytical solution of the continuum problem.

The analytical solution of the system (3.6) with initial data $\Phi(0,x) = f^\Phi(x)$ and $\mathsf{K}(0,x) = f^K(x)$ is

$$\begin{aligned} \Phi(t,x) &= \frac{f^\Phi(x+\lambda_+ t) + f^\Phi(x+\lambda_- t)}{2} + \int_{r+\lambda_- t}^{r+\lambda_+ t} f^K(\tau) d\tau \\ \mathsf{K}(t,x) &= \frac{f^K(x+\lambda_+ t) + f^K(x+\lambda_- t)}{2} + \partial_x f^\Phi(x+\lambda_+ t) - \partial_x f^\Phi(x+\lambda_- t) \end{aligned} \tag{3.75}$$

where $\lambda_\pm = \beta \pm 1$.

The D_+ norm of the state vector $v = (\Phi, K)$ and the D_+ norm of the error in respect to the analytical solution $\mathsf{v} = (\Phi, \mathsf{K})$ are:

$$D_+ \text{ norm} \equiv \|v\|_{h,D_\pm} = \sqrt{\|D_\pm \Phi\|_h^2 + \|K\|_h^2} \tag{3.76}$$

$$D_+ \text{ norm error} \equiv \|v - \mathsf{v}\|_{h,D_\pm} = \sqrt{\|D_\pm(\Phi - \Phi)\|_h^2 + \|K - \mathsf{K}\|_h^2} \tag{3.77}$$

with $\|D_\pm \Phi\|_h^2$ and $\|D_\pm(\Phi - \Phi)\|_h^2$ computed according to (2.106).

The convergence factor is defined as:

$$\frac{1}{\log 2} \log \frac{\|v - \mathsf{v}\|_{h,D_\pm}}{\|v - \mathsf{v}\|_{h/2,D_\pm}} \tag{3.78}$$

3.6.1 Centered Scheme vs One-Point Advected Scheme

In 3.5.4 (lemmas 3.5.6 and 3.5.7) it was proven that for $0 < \beta \leq \frac{n}{n+1}$ the numerical "+" speeds are better approximated with one-point off-centered schemes than with centered schemes (at least up to very high frequencies in

the grid). This subsection shows some simple numerical tests to illustrate this fact.

Consider l-periodic initial data:

$$\begin{aligned} \Phi(0,x) &= A_1 e^{-A_2 \sin^2\left(\frac{\pi}{l}x-\frac{\pi}{2}\right)}, \\ \mathsf{K}(0,x) &= a\partial_x \Phi(0,x), \quad x \in [0,l)\,. \end{aligned} \tag{3.79}$$

with $A_1 = 1$ and $A_2 = 50/\pi$.

The parameter $a \in [-1,1]$ sets the amplitude of the "$\pm$" components of the signal,

$$C_\pm = (a \pm 1)\partial_x \Phi\,. \tag{3.80}$$

When $a = 1(-1)$ the signal is purely "left" ("right") going and when $a = 0$, the signal is equally distributed between both modes.

The grid has $N = 101$ points and the resolution is $h = 0.01$, so the grid-length is $l = Nh = 1.01$. The shift is chosen $\beta = 0.5$. The wave equation is integrated using fourth order FDOs for space derivatives and the fourth order Runge-Kutta as time integrator.

Let $a \in \{1, 0, -1\}$. For each value of a, two runs are made: once using centered FDOs ($s = 0$) and once using one-point upwinded shift terms ($s = 1$). For each pair of runs, the errors of the main variables are computed and compared (Fig. 3.12). The numerical results show that, indeed, when the signal is "left" going, the upwinded scheme has less error than the centered scheme, while when the signal is going "right", the centered scheme is to be preferred.

3.6.2 Accuracy and Convergence of Higher Orders

In this subsection only centered schemes are considered. The numerical experiments are meant to study how the overall accuracy and convergence are influenced by the choice of the (a) time integrator, (b) space resolution and Courant factor.

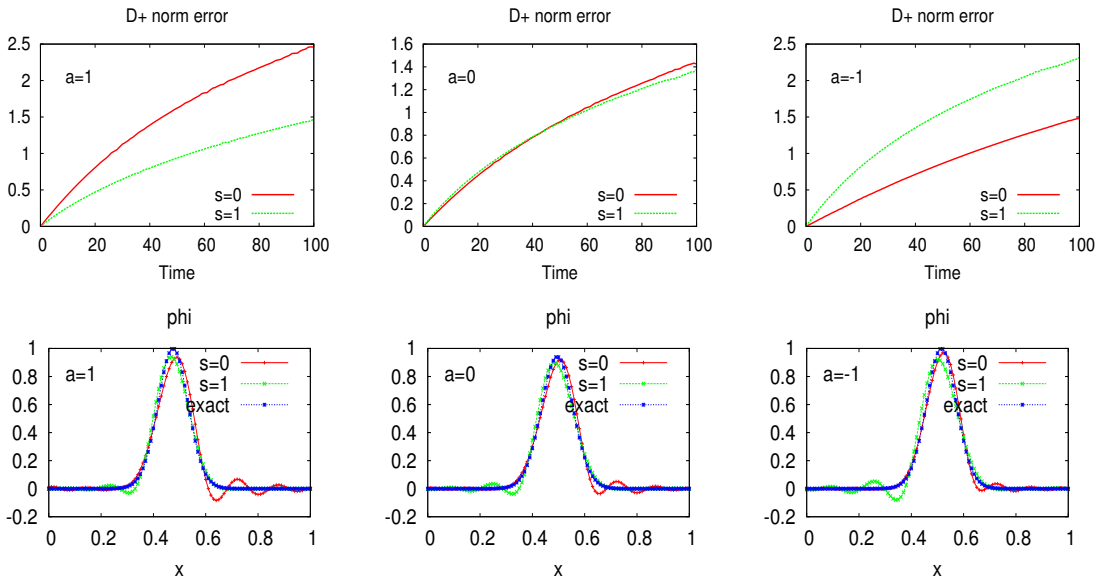


Figure 3.12: **Centered scheme (red lines) versus one-point advected stencil (green lines)** The plots show the total errors (D_+ norm) (top) and the snapshots at the end of the evolution for the variable ϕ (bottom), when the signal has different distribution on "+/−" modes ("left/right" going). If the signal is left-going ($a = 1$) upwind is more accurate (see left column); The situation reverses if the signal is right-going ($a = -1$) (see right column). No big difference appears between upwind and centered scheme if the left and right are equal ($a = 0$) (see middle column). The blue lines correspond to the exact solution.

(a) Scaling with the order of the Runge-Kutta Time Integrator

In this test, three different types of explicit Runge-Kutta methods are considered —the classical 4th order method (RK4), the explicit 6th order method from [56] (RK6) and the embedded 8th order algorithm from [79] (RK8). For each of them, the space derivatives are approximated with centered FDOs of increasing order from 2 to 10.

The initial data is chosen as in (3.79) with $A_1 = l/(2\pi)$, $A_2 = 1$ and $l = 1.01$. The shift is fixed to $\beta = 0.5$. For the lowest resolution the grid has $N = 51$ points and the space resolution is $h = 0.02$. In all the runs the Courant factor is $\rho = 0.5$. When using RK6, dissipation was needed (because this time integrator is not locally stable on the imaginary axis) and this has been added with the coefficients $\sigma = 0.01$, 0.03, 0.05, 0.2, 0.4 for $2n = 2$, 4, 6, 8, 10.

The results are presented in figures (3.13) which show the D_+ norm of the error and the convergence factor. The plots tell us that

- a $2p$ order Runge-Kutta time integrator discriminates between $2n$-order spacial finite difference schemes as long as $2n \leq 2p + 2$. If the order of the centered FDOs is higher than $2p + 2$ then the error of the time integrator will be dominant and there is practically no improvement in the accuracy over the previous order.

- the higher the order of the spatial approximation, the better the time behavior of the convergence factor. (e.g. for $2n = 2$ the convergence factor drops to 1 in less than 50CT, and the convergence is lost completely at the end of the evolution; for $2n > 2$ the convergence factors decrease also in time but at much lower rates.)

(b) Scaling with the Grid Spacing and with the Courant Factor

The purpose of this test is to measure the influence of the grid spacing (h) and of the Courant factor (λ) on the accuracy and convergence of the numerical scheme.

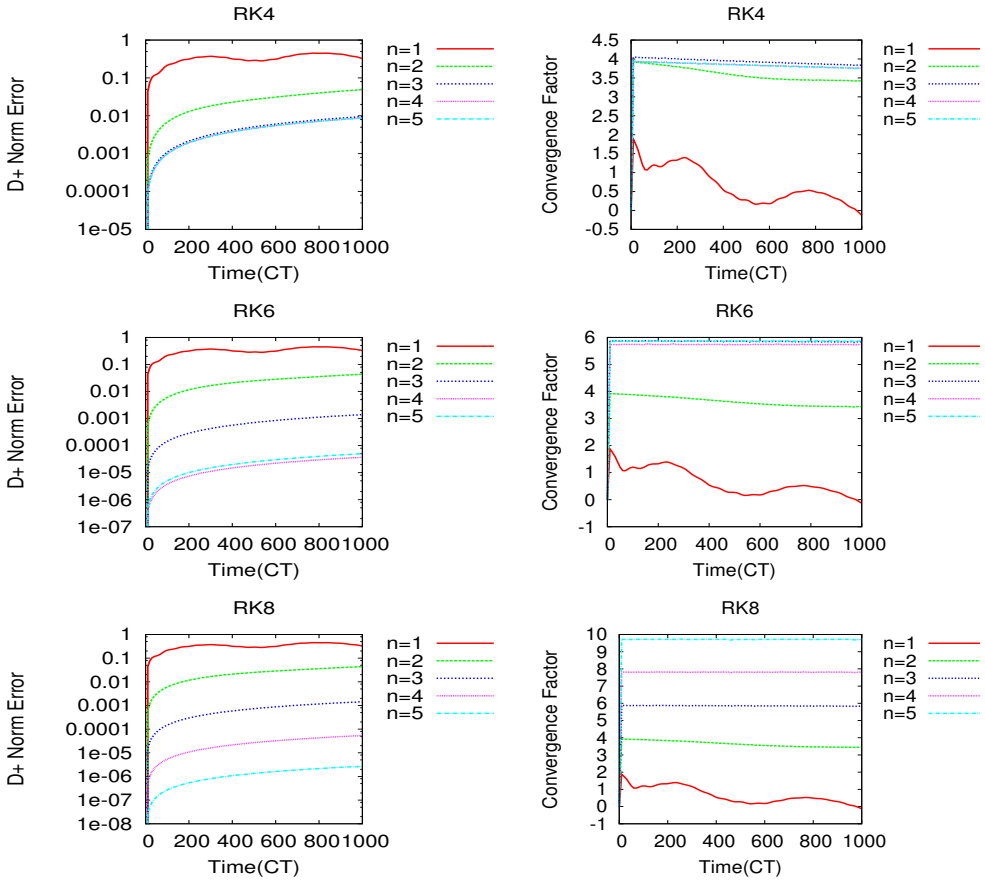


Figure 3.13: **Error and Convergence when using Different Time Integrators** Shown are the D_+ norm of error (left column) and the convergence factor (right column) when centered FDOs of different orders are used in interior in combinations with various orders for the time integrator (from top to bottom: Runge-Kutta 4, 6, 8). Increasing the order of the spatial discrete operators beyond $2(p+1)$ does not improve the accuracy anymore, the total error being dominated by the error of the time integrator. Higher orders exhibit better convergence over time.

The time integrator is the 4th order Runge-Kutta method. The shift is fixed to $\beta = 0.5$. The initial data is chosen as in (3.79) with $A_1 = l/(2\pi)$, $A_2 = 1$ and $l = 1.01$).

For each order of approximation $2n = 2, 4, 6, 8$, the runs are performed at various resolutions (21 values h in $[0.000625, 0.25]$) and various Courant limits (5 values λ in $[0.0625, 0.5]$). For each test (n, h, λ), the total error and the convergence factor are measured at the same time, $t = 3$. The results are displayed in Fig. 3.14-3.15.

Interpretation of the plots

For 2nd and 4th order, the accuracy and the convergence factor vary very little (they increase) with the Courant factor and the space resolution plays the determinant role. The accuracy increases with the space resolution, and the higher the resolution, the sharper the increase. The best one gets is an error of $\sim 10^{-5}$ with $2n = 2$ and $\sim 10^{-9}$ with $2n = 4$ at the lowest resolution $h = 0.000625$.

For the 6th and 8th order, with increasing space resolution, the Courant factor becomes important. Although the time integrator is only 4th order, the overall convergence can increase to 6 and respectively, 8, at high resolutions and high Courant factors. In these regions, the accuracy reaches its maximum, $\sim 10^{-12}$, (determined by the precision of the numerical evaluation of the analytical solution). Increasing even more the Courant factor or the resolution, leads to a decrease in accuracy and convergence.

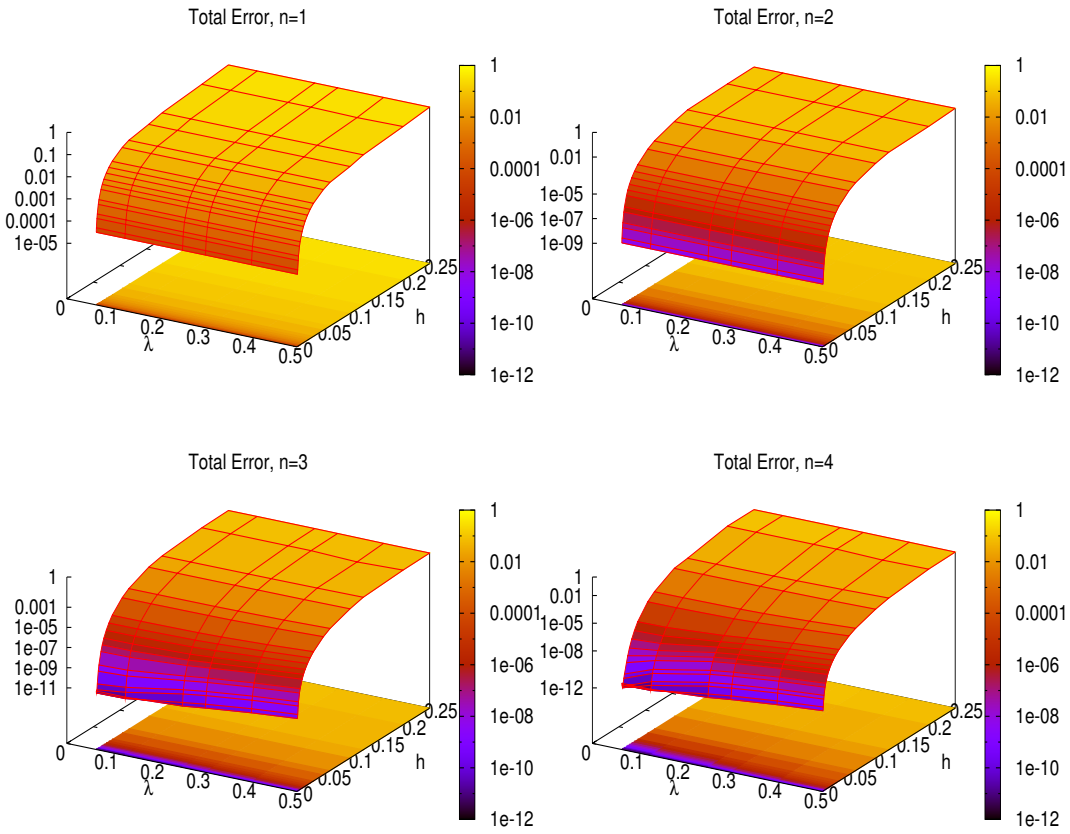


Figure 3.14: **Total error as function of Courant factor (λ) and grid spacing (h);** The error is computed at $t = 3$; the shift is $\beta = 0.5$ and the time integrator is RK4.

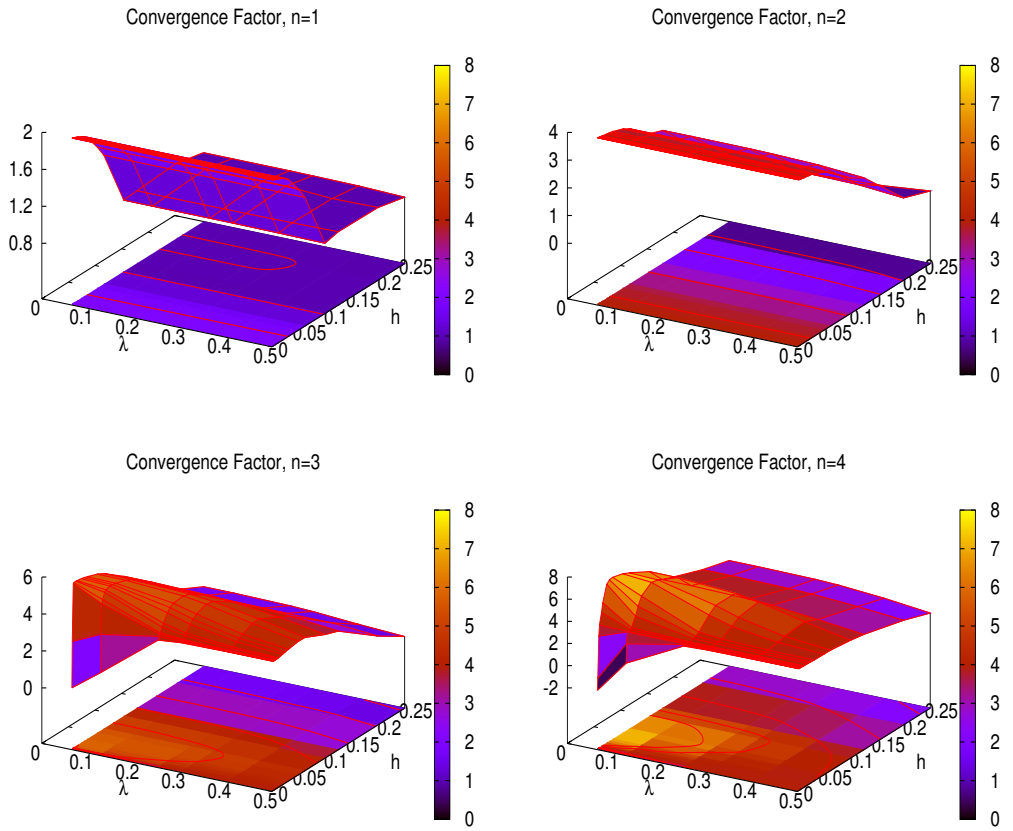


Figure 3.15: **Convergence factor as function of Courant factor (λ) and grid spacing (h);** The convergence factor is computed at $t = 3$; the shift is $\beta = 0.5$ and the time integrator is RK4.

Chapter 4

Initial Boundary Value Problem for the Wave Equation

This chapter investigates two high order discretization methods of the IBVP for the wave equation with shift, written as a first order in time and second order in space hyperbolic system. The first method discussed is called the ghost-point method and uses the boundary conditions to populate sufficient points outside the grid domain in such a way that the discrete evolution equations can be applied up to the last points of the grid. The second method is the SBP-SAT procedure, which uses summation by parts (SBP) operators in conjunction with a modification of the rhs of the evolution equations by addition of simultaneous approximation terms (SAT) dictated by the boundary conditions. The chapter opens with a section that frames the general theoretical background and discusses the advantages and drawbacks of each method. It is followed by a section that investigates the strong well-posedness of the continuum problem with maximally dissipative boundary conditions. Then the two discretization methods are discussed in detail. The last section presents numerical results.

4.1 Theoretical Background

4.1.1 Well-Posedness and Strong Well-Posedness

Consider the initial boundary value problem for the following system of PDEs in one dimension:

$$\begin{aligned} \dot{\mathsf{v}} &= \mathsf{P}\mathsf{v} + \mathsf{F}, \quad x_0 \le x \le x_1, t \ge t_0 \\ \mathsf{v}(t_0, x) &= \mathsf{f}(x) \\ \mathsf{L}_0\mathsf{v}(t, x_0) &= \mathsf{g}_0(t) \\ \mathsf{L}_1\mathsf{v}(t, x_1) &= \mathsf{g}_1(t) \end{aligned} \tag{4.1}$$

where v is a vector function, P and $\mathsf{L}_{0,1}$ are matrix differential operators of certain orders. [1]

According to [36], the problem (4.1) is called

1. **well-posed** in the norm $\|.\|_*$ if, considering $\mathsf{F} = \mathsf{g}_0 = \mathsf{g}_1 = 0$, for every function $\mathsf{f} \in C^\infty$ that vanishes in a neighborhood of the boundary points $x_{0,1}$, it has a smooth solution that satisfies the estimate:

$$\|\mathsf{v}(t, .)\|_* \le K_c e^{\alpha_c (t - t_0)} \|\mathsf{f}\|_* \tag{4.2}$$

2. **strongly well-posed** in the norm $\|.\|_*$ if it is well-posed and in addition satisfies:

$$\|\mathsf{v}(t, .)\|_*^2 \le W_c(t, t_0) \left(\|\mathsf{f}\|_*^2 + \int_{t_0}^{t} d\tau \left(\|\mathsf{F}\|_*^2 + |\mathsf{g}_0(\tau)|^2 + |\mathsf{g}_1(\tau)|^2 \right) \right) \tag{4.3}$$

where $W_c(t, t_0)$ is a function that is bounded in every time interval and does not depend on the data.

In order to show (strong) well-posedness for a given IBVP, two methods are available: the energy method and the Laplace transform method.

[1] usually the boundary operators are one order lower than the evolution operator, P.

It is known [36] that the IBVP of first order symmetric hyperbolic systems[2] with maximally dissipative boundary conditions is strongly well-posed in the l^2-norm, $\|.\|$. Second order systems which admit a reduction to first order symmetric hyperbolic systems are called also symmetric hyperbolic and they are strongly well-posed in a Soboloev norm, $\|.\|_\partial$, containing derivatives of some of the variables.

4.1.2 Discrete Schemes and Stability Concepts for IBVP

In order to discretize the problem (4.1) there are several procedures available. This thesis considers two of them: the ghost-point method and the SBP-SAT method.

Ghost-Point Method

Let us suppose that the grid has a number of $N+1$ points, x_j, with $j = \overline{0,N}$, and denote the grid representation of $\mathsf{v}(x)$ with $v(t)$, where $v(t) \equiv (v_0(t)^T, \ldots, v_N^T(t))^T$, and $v_i = \mathsf{v}(t, x_i)$.
In this method the evolution is carried out up to the the last points of the grid (x_0, x_N). The extra points (ghost points) required by the evaluation of the discrete derivatives close to the boundaries are computed by solving an algebraic system formed by discretizing the physical boundary conditions and introducing an appropriate set of numerical boundary conditions[3]. The semi-discrete scheme is of the form:

$$\begin{aligned} \dot{v} &= Pv + F \\ v(t_0) &= f \\ L_0 v_0^b(t) &= g_0(t) \\ L_1 v_N^b(t) &= g_1(t) \end{aligned} \tag{4.4}$$

[2] in 1D strongly hyperbolic systems are also symmetric hyperbolic
[3] numerical boundary conditions do not have an equivalent at the continuum level

where P and $L_{0,1}$ are discrete matrix differential operators, F and f are grid vectors such that $F_i(t) = \mathsf{F}(t, x_i)$ and $f_i(t) = \mathsf{f}(t, x_i)$. The quantities $v^b_{0,N}$ represent vectors defined at the boundaries, containing the ghost points and a certain number of interior points. Because the operators $L_{0,1}$ include also the numerical boundary conditions (which can be inhomogeneous), the terms $g_{0,1}$ are vectors containing the continuum $\mathsf{g}_{0,1}$ but also the contributions from the numerical boundary conditions.

Using the algebraic system defined at the boundary, one can eliminate the ghost points and rewrite (4.4) as an equivalent system of ODEs:

$$\begin{aligned} \dot{v} &= \tilde{P}v + F \\ v(t_0) &= f \end{aligned} \tag{4.5}$$

where $\tilde{P}$ is the operator P modified at the points close to the boundaries and includes also the inhomogeneous terms $g_{0,1}$.

The semi-discrete approximation is called:

1. **stable** in the norm $\|.\|_{*,h}$ if, considering $F = g_0 = g_1 = 0$, for all $h \leq h_0$ there are constants K_d and α_d such that for all t_0 and all initial data f,

$$\|v(t)\|_{*,h} \leq K_d e^{\alpha_d (t-t_0)} \|f\|_{*,h} \tag{4.6}$$

2. **strongly stable** in the norm $\|.\|_{*,h}$, if it is stable in this norm and in addition satisfies:

$$\|v(t)\|^2_{*,h} \leq W_d(t, t_0) \left(\|f\|^2_{*,h} + \max_{t_0 \leq \tau \leq t} \|F\|^2_{*,h} + \max_{t_0 \leq \tau \leq t} \left(|g_0(\tau)|^2 + |g_1(\tau)|^2 \right) \right) \tag{4.7}$$

where $W_d(t, t_0)$ is a function that is bounded in every time interval and does not depend on the data.

If the semi-discrete scheme has the same asymptotic time growth as the continuum problem (e.g. $\alpha_c \leq \alpha_d + \mathcal{O}(h)$) then it is called **strictly stable**.

Similar to the continuum problem, in order to prove (strong) stability of the semi-discrete problem, one can employ the energy or the Laplace transform method. While the energy method can be tricky in the sense that one has to "guess" an appropriate energy depending on the system, the Laplace method is very well formalized and can be applied to a large variety of systems (including our systems of interest, hyperbolic and second order in space). However, there is a price to be payed in the form of analysis of the roots of some high order polynomials with complex coefficients.

For the fully discrete problem, similar concepts for stability can be defined: the GKS-stability theory, which gives necessary and sufficient conditions for stability, has been developed in [35]. GKS-stability is based on the Laplace transform method and insures that the solution remains bounded by a function of time, as the space and time stepping goes to zero. However it does not capture the behavior of the solution as $t \to \infty$. The fully discrete solution can manifest a nonphysical exponential growth in time even if the semi-discrete scheme is strictly stable.

For applications which require long time computations, one has to devise schemes which do not allow a growth in time of the discrete solution if it is not inherited from the continuum problem. The SBP-SAT procedure is designed for this purpose and allows one to construct schemes which are not only GKS-stable but also time-stable [17].

SBP-SAT method

The basic idea of this approach is to approximate the derivatives by finite difference operators using only grid points (no ghost points) so that certain summation by parts rules hold, and to implement the boundary conditions so that the SBP property is preserved. The procedure amounts to constructing a discrete energy that mimics the behavior of the corresponding energy at the continuum level.

In the space of real functions $u, v \in L^2[x_0, x_1]$, define the following scalar

product and norm:

$$(u,v)=\int_{x_0}^{x_1}u^T(x)v(x)dx,\quad |u|^2=\int_{x_0}^{x_1}u^T(x)u(x)dx \tag{4.8}$$

With respect to the above scalar product, the differential operators ∂_x and ∂_{xx} obey the following:

integration by parts rules

$$\begin{aligned}(u_x,v)+(u,v_x) &= \langle u,v\rangle|_{x_0}^{x_1}\\ (u_{xx},v)+(u,v_{xx}) &= \langle u_x,v\rangle|_{x_0}^{x_1}+\langle u,v_x\rangle|_{x_0}^{x_1}-2(u_x,v_x)\\ (u_{xx},v)-(u,v_{xx}) &= \langle u_x,v\rangle|_{x_0}^{x_1}-\langle u,v_x\rangle|_{x_0}^{x_1}\\ (u_{xx},v) &= \langle u_x,v\rangle|_{x_0}^{x_1}-(u_x,v_x)\end{aligned} \tag{4.9}$$

where $\langle u,v\rangle|_{x_0}^{x_1}=u^T(x_1)v(x_1)-u^T(x_0)v(x_0)$.

At the discrete level, in the space of grid functions, a positive $(N+1)\times(N+1)$ real matrix $\Sigma=\Sigma^T$ is used to define the scalar product and the norm:

$$(u,v)_\Sigma=u^T\Sigma v,\quad \|u\|_\Sigma=(u,u)_\Sigma \tag{4.10}$$

An SBP operator corresponding to a differential operator ∂_* [4] is a matrix $(N+1)\times(N+1)$ which mimics the integration by parts rules of ∂_* in respect to (4.8), by summation by parts rules in respect to (4.10). In [74] and [57] SBP operators associated with ∂_x, and respectively, with ∂_{xx}, have been designed, based on the same norms, Σ. They are of the following form:

$$\begin{aligned}D_1 &= \Sigma^{-1}Q\\ D_2 &= \Sigma^{-1}\left(-A+BS\right)\end{aligned} \tag{4.11}$$

[4] ∂_* is ∂_x or ∂_{xx}.

where Q,A, B and S are matrices such that:

$$\begin{aligned} &B \equiv E_0 + E_1 \quad \text{where} \quad E_0 \equiv \text{diag}[-1, 0, ..., 0], \quad E_1 \equiv \text{diag}[0, 0, ..., 1] \\ &Q + Q^T = B, \quad A + A^T > 0, \end{aligned} \tag{4.12}$$

and S includes a first order derivative at the boundary. The operators constructed in this way satisfy the following

summation by parts rules

$$\begin{aligned} (D_1 u, v)_\Sigma + (u, D_1 v)_\Sigma &= u^T B v \\ (D_2 u, v)_\Sigma + (u, D_2 v)_\Sigma &= (Su)^T B v + u^T B (Sv) - u^T (A^T + A) v \\ (D_2 u, v)_\Sigma - (u, D_2 v)_\Sigma &= (Su)^T B v - u^T B (Sv) - u^T (A^T - A) v \\ (D_2 u, v)_\Sigma &= (Su)^T B v - u^T A^T v \end{aligned} \tag{4.13}$$

Note that for the first derivative there is only one integration/summation by parts rule, while for the second derivative one can construct three rules. The SBP operator D_1 mimics completely the behavior of ∂_x, while D_2 does that only if the matrix A can be written as $A = A^T = \tilde{D}_1^T \Sigma \tilde{D}_1$ where $\tilde{D}_1$ is a consistent approximation of ∂_x. In this case the second discrete derivative for the inner points is no longer the standard minimal bandwidth centered FDO (that uses $2n + 1$ points to attain $2n$-accuracy) but rather the first discrete derivative applied twice, which is disadvantageous from the numerical point of view (see [57]).

Now the SBP property alone does not guarantee that an energy estimate exists and the scheme is strictly stable. In order to attain this, a special boundary treatment is necessary. To this date, there are two methods available which implement boundary conditions without destroying the SBP-property: the projection method [60, 61] and the SAT method [17]. The latter uses the

boundary conditions as penalty terms to the evolution equations:

$$\begin{aligned} \dot{v} &= Pv + F + \tau_0 \left(L_0 v_0^b(t) - \mathrm{g}_0(t) \right) + \tau_1 \left(L_1 v_1^b(t) - \mathrm{g}_1(t) \right) \\ v(t_0) &= f \end{aligned} \tag{4.14}$$

choosing the parameters $\tau_{0,1}$ in such a way that the system with the modified operator $\tilde{P} = P + \tau_0 \left(L_0 v_0^b(t) - \mathrm{g}_0(t) \right) + \tau_1 \left(L_1 v_1^b(t) - \mathrm{g}_1(t) \right)$ and $F = 0$ admits a non-increasing energy, given that the continuum problem behaves as such.

The SBP-SAT procedure has been applied with success in the case of first order symmetric hyperbolic systems [22, 62] and parabolic systems [57], both well-posed in l^2-norm. For these types of systems the procedure is straightforward once the continuum energy estimate has been established. Also, when the continuum problem does not require any boundary condition, the SBP operators give the numerical boundary condition "for free", in the sense that nothing more is to be done at the boundary —the one-sided stencils provided by the SBP operators guarantee stability.

In the case of parabolic systems, it is not necessary that D_2 mimics completely ∂_{xx}, in order to get an energy estimate; the condition $A + A^T > 0$ is sufficient.

For general second order hyperbolic systems that are well-posed in Soboloev norms, there is no standard way to construct the discrete energy and the method can be applied only on a case-by-case basis, with lots if imagination, usually by requiring the SBP operators to satisfy additional properties which do not appear in their standard construction. Section 4.4 shows how the SBP-SAT procedure can be used to implement inflow boundaries for the shifted wave equation. It will turn out that the extra condition that has to be enforced is $A - Q^T H Q \geq 0$.

4.2 Continuum Problem

Consider the shifted one-dimensional wave equation in the domain $x_0 \leq x$, $t \geq t_0$:

$$\begin{aligned} \dot{\Phi}(t,x) &= \beta\partial_x\Phi + \mathsf{K} + \mathsf{F}^{\Phi} \\ \dot{\mathsf{K}}(t,x) &= \partial_{xx}\Phi + \beta\partial_x\mathsf{K} + \mathsf{F}^{\mathsf{K}} \\ \Phi(t_0,x) &= \mathsf{f}^{\Phi}(x) \\ \mathsf{K}(t_0,x) &= \mathsf{f}^{\mathsf{K}}(x) \\ \mathsf{L}_0(\partial_x\Phi,\mathsf{K}) &= \mathsf{g}_0(t) \end{aligned} \tag{4.15}$$

where L is the boundary operator. The system is symmetric hyperbolic so maximally dissipative boundary conditions lead to strong stability of the IBVP. This can be shown this in various ways, e.g. by introducing the variable $\mathsf{X} = \partial_x\Phi$, and constructing an equivalent first order symmetric hyperbolic system which is strongly well-posed with respect to the l^2-norm. Then by restoring the variables of the second order system, and denoting with $v \equiv (\Phi, \mathsf{K})$, one obtains that the IBVP, (4.15) is strongly well-posed in the Soboloev norm:

$$\|\mathsf{v}\|_{\partial}^2 = \int_{x_0}^{\infty} dx \left(|\partial_x\Phi|^2 + |\mathsf{K}|^2\right) . \tag{4.16}$$

Obs. If $\mathsf{C}_{\pm} = \mathsf{K} \pm \partial_x\Phi$ are the characteristics of the system, and $\lambda = \beta \pm 1$ are the speeds, then the norm (4.16) can be written as:

$$\|\mathsf{v}\|_{\partial}^2 = \frac{1}{2}\int_{x_0}^{\infty} dx \left(\mathsf{C}_+^2 + \mathsf{C}_-^2\right) \tag{4.17}$$

The energy estimates for each type of boundary condition are presented below. Denote $\mathsf{F} = (\mathsf{F}^{\Phi}, \mathsf{F}^{\mathsf{K}})$. Then :

$$\frac{d}{dt}\|\mathsf{v}\|_{\partial}^2 \leq \frac{1}{2}\left(\lambda_+\mathsf{C}_+^2 + \lambda_-\mathsf{C}_-^2\right)\Big|_{x_0}^{\infty} + \|\mathsf{v}\|_{\partial}^2 + \|\mathsf{F}\|_{\partial}^2 \tag{4.18}$$

According to the sign of the speeds, one of the following situations appears:

1. **Outflow boundary:** $\beta \geq 1$
   No boundary conditions are needed because $-\frac{1}{2}\left(\lambda_+ \mathsf{C}_+^2 + \lambda_- \mathsf{C}_-^2\right) \leq 0$.

   The energy estimate is:

$$\frac{d}{dt}\|\mathsf{v}\|_\partial^2 \;\leq\; \|\mathsf{v}\|_\partial^2 + \|\mathsf{F}\|_\partial^2 \tag{4.19}$$

2. **Inflow boundary:** $-1 \leq \beta < 1$
   One boundary condition is needed: $\mathsf{C}_- = R_0 \mathsf{C}_+ + \mathsf{g}_0(t)$ which leads to:

$$\lambda_+ \mathsf{C}_+^2 + \lambda_- \mathsf{C}_-^2 = \lambda_+ \mathsf{C}_+^2 + \lambda_- \left(R_0 \mathsf{C}_+ + \mathsf{g}_0\right)^2 \geq \mathsf{C}_+^2 \left(\lambda_+ + 2\lambda_- R_0^2\right) + 2\lambda_- \mathsf{g}_0^2 \tag{4.20}$$

   If $R_0^2 \leq \frac{\lambda_+}{2|\lambda_-|}$ then $-\frac{1}{2}\left(\lambda_+ \mathsf{C}_+^2 + \lambda_- \mathsf{C}_-^2\right) \leq |\lambda_-|\, g_0^2$. The following energy estimate is obtained:

$$\frac{d}{dt}\|\mathsf{v}\|_\partial^2 \;\leq\; |\lambda_-|\, g_0^2 + \|\mathsf{v}\|_\partial^2 + \|\mathsf{F}\|_\partial^2 \tag{4.21}$$

3. **Completely Inflow Boundary:** $\beta < -1$
   Two boundary condition are needed: $\mathsf{C}_\pm = \mathsf{g}_{0\pm}(t)$ which give $-\frac{1}{2}\left(\lambda_+ \mathsf{C}_+^2 + \lambda_- \mathsf{C}_-^2\right) \leq \frac{1}{2}\left(|\lambda_-|\, \mathsf{g}_{0-}^2 + |\lambda_+|\, \mathsf{g}_{0+}^2\right) \leq |\lambda_-|\, \mathsf{g}_0^2$, where $g_0^2(t) \equiv g_{0+}^2(t) + g_{0-}^2(t)$.
   The energy estimate is:

$$\begin{aligned} \frac{d}{dt}\|\mathsf{v}\|_\partial^2 \;&\leq\; \frac{1}{2}\left(|\lambda_-|\, \mathsf{g}_{0-}^2 + |\lambda_+|\, \mathsf{g}_{0+}^2\right) + \|\mathsf{v}\|_\partial^2 + \|\mathsf{F}\|_\partial^2 \\ &\leq\; |\lambda_-|\, \mathsf{g}_0^2 + \|\mathsf{v}\|_\partial^2 + \|\mathsf{F}\|_\partial^2 \end{aligned} \tag{4.22}$$

Integrating in time any of the relations above [5] leads to the following inequality:

$$\|\mathsf{v}(t,.)\|_{\partial}^2 \leq K_d e^{t-t_0}\left(\|\mathsf{f}\|_{\partial}^2 + \int_{t_0}^{t} d\tau \left(\|\mathsf{F}\|_{\partial}^2 + |\mathsf{g}_0(\tau)|^2\right)\right) \tag{4.23}$$

So all the above IBVP problems are strongly well-posed.
Obs. In case $\mathsf{F} = 0$, the energy estimate is just (4.23) without the exponential factor e^{t-t_0}.

Each of the three cases above requires a different numerical treatment in order to insure some type stability mentioned in 4.1.2.

When two boundaries are present (one at x_0 and one at x_1) according to the value of the shift, β, 5 possible initial boundary value problems can be formulated (see table 4.1), out of which only 3 are distinct: inflow-inflow for $-1 < \beta < 1$, outflow-inflow for $\beta = 1$ and outflow-completely inflow for $\beta > 1$.

The next section shows how the ghost-point method can be used to implement all the above IBVPs and obtain $2n$-accurate stable schemes.

For outflow boundary, 2nd order accuracy, stability is proved through both Laplace method and energy method. For higher orders, I was not able to reach my goal of a stablility proof. However, I will show the path and the mathematical obstacles encountered on the way.

[5] If $y(t), w_1(t), w_2(t) \geq 0$ and

$$\dot{y}(t) \leq y(t) + w_1(t) + w_2(t), \quad t \geq t_0$$

denote $\dot{Y}(t) = \dot{y}(t) - y(t)$ and $W(t) = w_1(t) + w_2(t) \geq 0$. Then $y(t) = e^{t-t_0}\left(y(t_0) + \int_{t_0}^{t} \dot{Y}(\tau)e^{-\tau}d\tau\right)$ and $\dot{Y}(t) \leq W(t)$, for all $t \geq t_0$. This leads to:

$$y(t) \leq e^{t-t_0}\left(y(t_0) + \int_{t_0}^{t} \left(w_1(\tau) + w_2(\tau)\right) e^{-\tau} d\tau\right)$$

β	x_0	x_1
$\beta < -1$	completely inflow	outflow
$\beta = -1$	inflow	outflow
$-1 < \beta < 1$	inflow	inflow
$\beta = 1$	outflow	inflow
$\beta > 1$	outflow	completely inflow

Table 4.1: Types of IBVPs for the wave equation with boundaries at x_0 and x_1

4.3 Ghost-Point Method

The system (4.15) is discretized in the following way:

$$\begin{aligned} \dot{\Phi}(t) &= \beta D^{(1,n)}\Phi + K + F^{\Phi}(t) \\ \dot{K}(t) &= D^{(2,n)}\Phi + \beta D^{(1,n)}K + F^{K} \\ \Phi(0) &= f^{\Phi} \\ K(0) &= f^{K} \\ L_0(\Phi, K) &= g_0(t) \end{aligned} \tag{4.24}$$

where $D^{(1,n)}$ and $D^{(2,n)}$ are the $2n$-order accurate centered FDOs corresponding to the first and second derivative, and L_0 is the discrete boundary operator which can be read out from the following boundary prescriptions.

4.3.1 Boundary Prescriptions

Consider the boundary $x = x_0$. According to the value of β the following cases are possible:

1. **Outflow boundary:** $\beta \geq 1$
   No boundary condition needs to be applied. The ghost points $i = \overline{-n, -1}$ can be computed with one of the following prescriptions:

$$\begin{aligned} D_+^m K_i &= 0 \\ D_+^{m+1}\Phi_i &= 0 \quad (4.25) \end{aligned} \quad \text{or} \quad \begin{aligned} D_+^{m+1} K_i &= 0 \\ D_+^{m+1}\Phi_i &= 0 \quad (4.26) \end{aligned}$$

where $m \geq 2n$.

2. **Inflow boundary:** $-1 \leq \beta < 1$
   One boundary condition needs to be applied: $C_- = R_0 C_+ + \mathrm{g}_0(t)$. If $R_0 \neq -1$ the ghost points are computed with following prescription:

$$\begin{aligned} K_0 - D^{(1)}\Phi_0 &= R_0(K_0 + D^{(1)}\Phi_0) + \mathrm{g}_0(t) \\ D_+^{2n} K_i &= 0 \ , i = \overline{-n, -1} \\ D_+^{2n+1}\Phi_i &= 0 \ , i = \overline{-n, -2} \end{aligned} \tag{4.27}$$

   That is the ghost points for K are computed from extrapolation conditions and the ghost points for Φ are computed by solving the linear system given by the boundary condition and extrapolation conditions for the ghost points $i = -n, -2$. The case $R_0 = -1$ (Dirichlet) needs a special treatment and it is not considered in this thesis. The focus will be only on the Sommerfeld case, $R_0 = 0$.

3. **Completely Inflow Boundary:** $\beta < -1$
   Two boundary condition needs to be applied: $C_\pm = \mathrm{g}_{0\pm}(t)$. The ghost points are computed using the following algorithm: first, the value K_0 computed with the evolution equations is saved in a temporary variable:

$$K_0^{\text{temp}} = K_0$$

   then the following linear system is solved for the unknowns K_{-i}, $i = \overline{0, n}$

and Φ_{-i}, $i = \overline{1,n}$:

$$\begin{aligned} K_0 - D^{(1)}\Phi_0 &= \mathrm{g}_{0-}(t) \\ K_0 + D^{(1)}\Phi_0 &= \mathrm{g}_{0+}(t) \\ D_+^{2n} K_i &= 0 \ , i = \overline{-n,-1} \\ D_+^{2n+1}\Phi_i &= 0 \ , i = \overline{-n,-2} \end{aligned} \tag{4.28}$$

In the end, the value of K_0 is restored:

$$K_0 = K_0^{\text{temp}}.$$

Notice that proceeding in this way we make use of the boundary conditions to determine all the ghost points without overwriting the already computed K_0 by the evolution equations. For 2nd and 4th order accuracy, stable schemes are obtained also when the point K_0 is given by the boundary conditions (by overwriting the value computed with evolution equations). However the schemes proved (experimentally) to be less accurate and could not be generalized to higher orders.

4.3.2 Equivalent Systems and Known Results

Second Order Time System

By deriving in time the evolution equation for Φ in (4.24) we get the following second-order time, second-order space discrete system:

$$\begin{aligned} \ddot{\Phi}(t) &= 2\beta D^{(1,n)}\dot{\Phi} + \left(D^{(2,n)} - \beta^2 D^{(1,n)} D^{(1,n)}\right)\Phi + F \\ \Phi(0) &= f^{\Phi} \\ \dot{\Phi}(0) &= f^{\dot{\Phi}} \\ L_0(\Phi, K) &= g_0(t) \end{aligned} \tag{4.29}$$

where $F = -\beta(D^{(1,n)}\Phi)F^{\Phi}(t)+F^{K}+\dot{F}^{\Phi}(t)$ and $f^{\dot{\Phi}} = \beta D^{(1,n)}f^{\Phi}+f^{K}+F^{\Phi}(t_0)$. The system (4.29) with $n = 1$, has been considered in [77] under the name "horizon algorithm", to discretize the outflow boundary. Using the Laplace technique, it was shown that the numerical boundary conditions

$$D_{+}^{3}\Phi_0 = D_{+}^{3}\Phi_{-1} = 0$$

lead to stability. These extrapolation conditions are the equivalent of (4.25) with $q_1 = q_2 = 3$.

In [77], the inflow boundary is also considered, however, the discretization scheme which is proposed, ("outer algorithm") uses the following approximation for the evolution equation of Φ:

$$\ddot{\Phi}(t) = 2\beta D^{(1,n)}\dot{\Phi} + \left(1-\beta^2\right) D^{(2,n)}\Phi + F \tag{4.30}$$

This no longer corresponds to our system, (4.24), but rather to (4.24) with the evolution equation for K replaced by:

$$\dot{K}(t) = \beta^2 D^{(1,n)}D^{(1,n)} + (1-\beta^2)D^{(2,n)} + \beta D^{(1,n)}K + F^{K} \tag{4.31}$$

First Order System

Introducing the extra variable $X = D_{+}\Phi$ transforms (4.24) into a first order system:

$$\begin{aligned}
\dot{X}(t) &= \beta D^{(1,n)}X + D_{+}K + F^{X} \\
\dot{K}(t) &= D_{\sim}^{(n)}X + \beta D^{(1,n)}K + F^{K} \\
X(0) &= f^{X} \\
K(0) &= f^{K} \\
L_R(X,K) &= g(t)
\end{aligned} \tag{4.32}$$

plus the evolution equation for Φ:

$$\begin{aligned} \dot{\Phi}(t) &= \beta M^{(n)} X + K + F^{\Phi} \\ \Phi(0) &= f^{\Phi} \end{aligned}$$

where $F^X = D_+ F^{\Phi}$, $f^X = D_+ f^{\Phi}$ and

$$\begin{aligned} M^{(n)} &\equiv \frac{1}{2}\left(I + S^{-1}\right) \sum_{k=0}^{n-1} c_k p^k \\ D_{\sim}^{(n)} &\equiv D_- \sum_{k=0}^{n-1} d_k p^k \end{aligned} \tag{4.33}$$

Because the evolution equation for Φ is just an ODE, for stability analysis it is enough to consider only the reduced system for the variables X and K, (4.32).

The boundary conditions (4.25)–(4.26), (4.27) and (4.28) are now expressed in terms of K and X.

1. outflow: The relations (4.25)–(4.26) become

$$\begin{aligned} D_+^m K_i &= 0 \\ D_+^m X_i &= 0 \end{aligned} \quad (4.34) \quad \text{or} \quad \begin{aligned} D_+^{m+1} K_i &= 0 \\ D_+^m X_i &= 0 \end{aligned} \quad (4.35)$$

   where $m \geq 2n$ and $i = \overline{-n, -1}$.

2. inflow: The lemma 2.29 together with the relation $D_+^p v_i = D_-^p v_{i-p}$ for $p \geq 0$, lead to

$$D^{(1)}\Phi_0 = \sum_{k=1}^{n} \frac{(-1)^{k+1} n!(2n-k)!}{k(2n)!(n-k)!} \left[(hD_+)^{k-1} X_0 + (-1)^{k+1} (hD_+)^{k-1} X_{-k}\right] \tag{4.36}$$

Then the boundary conditions are easy to write down:

$$\begin{aligned} K_0 - D^{(1)}\Phi_0 &= R_0(K_0 + D^{(1)}\Phi_0) + \mathrm{g}_0(t) \\ D_+^{2n} K_i &= 0 \ , i = \overline{-n, -1} \\ D_+^{2n} X_i &= 0 \ , i = \overline{-n, -2} \end{aligned} \tag{4.37}$$

with $D^{(1)}\Phi_0$ replaced by (4.36).

3. completely inflow: The boundary conditions (4.28) become

$$\begin{aligned} K_0 - D^{(1)}\Phi_0 &= \mathrm{g}_{0-}(t) \\ K_0 + D^{(1)}\Phi_0 &= \mathrm{g}_{0+}(t) \\ D_+^{2n} K_i &= 0 \ , i = \overline{-n, -1} \\ D_+^{2n} X_i &= 0 \ , i = \overline{-n, -2} \end{aligned} \tag{4.38}$$

with $D^{(1)}\Phi_0$ replaced by (4.36).

Using this first order reduction, in [16] both second and fourth order accuracy have been considered for the inflow and outflow boundaries and strong stability has been shown by checking numerically the Kreiss condition at some particular value of the shift.

No other results in the literature, apart from [77] and [16] concerning the discretization of IBVP (4.15) using the ghost-points method are known to me.

The results of this chapter are:

- using the energy method, in 4.3.3 stability analysis is performed for the 2nd order accurate discretization of (4.15) with outflow boundary
- using the Laplace transform method, 4.3.4 derives necessary conditions for the stability of general $2n$-accurate discretizations of (4.15) and shows strong stability for the case of 2nd order accurate schemes with outflow boundary.

- section 4.5 presents the results of some numerical experiments meant to test the stability of the discrete scheme corresponding to the boundary prescriptions (4.25), (4.27) and (4.28).

4.3.3 Stability Analysis via Energy Method for Outflow Boundary

In case of second order in space hyperbolic systems with the outflow boundary, the standard method employed for stability analysis is the Laplace transform. This will be analyzed in more detail in 4.3.4.

In the following the outflow boundary is investigated using the energy method. The discussion restricts to 2nd order accuracy.

Denote with $u = (X, K)$ the state vector and consider the system (4.32) with the extrapolation conditions (4.34) or (4.35).

Using a set of positive numbers a_i for $i = 1, m$, the discrete energy is constructed in the following way:

$$E = \|u\|^2_{a_i,h} = \|X\|^2_h + \|K\|^2_h + \sum_{i=1}^{m} a_i \left(\left\|(hD_+)^i X\right\|^2_h + \left\|(hD_+)^i K\right\|^2_h \right) \quad (4.39)$$

Obs. In the limit $h \to 0$ and $N \to \infty$ the terms $\left\|hD^i_+ X\right\|_h$ and $\left\|hD^i_+ K\right\|_h$ will cancel out and the discrete energy converges to the continuum energy (4.16). In order to show stability it is enough to find $a_i > 0$ such that the energy does not increase in time,

$$\dot{E} \leq 0.$$

The reminder of this subsection proves the existence of such numbers.

Without loss of generality, take $h = 1$.

- The energy estimate is:

$$\begin{aligned}\dot{E} &= 2\sum_{i=0}^{m} a_i \left[(D_+^i K, D_+^i \dot{K}) + (D_+^i X, D_+^i \dot{X})\right] \\ &= 2\sum_{i=0}^{m} a_i \left[\beta(D_+^i K, D_+^i D^{(1,n)} K) + (D_+^i K, D_+^i D_- X)\right. \\ &+ \left.\beta(D_+^i X, D_+^i D^{(1,n)} X) + (D_+^i X, D_+^{i+1} K)\right] \qquad (4.40)\end{aligned}$$

Using the fact that all the FD operators commute and

$$(v, D^{(1,n)} v) = -\frac{1}{2} v_0 v_{-1} \qquad (4.41)$$

$$(D_+^i X, D_+^{i+1} K) = -(D_+^i K, D_- D_+^i X) - (D_+^i K_0)(D_+^i X_{-1}) \qquad (4.42)$$

the energy estimate is written in terms of operators D_+^i applied to the ghost and boundary points:

$$\dot{E} = -\sum_{i=0}^{m} a_i \left[\beta\left(D_+^i K_0 D_+^i K_{-1} + D_+^i X_0 D_+^i X_{-1}\right) + 2(D_+^i K_0)(D_+^i X_{-1})\right] \qquad (4.43)$$

Using the extrapolation conditions (4.34), the quantities containing ghost points in (4.43) ($D_+^i K_{-1}$ and $D_+^i K_{-1}$) are written using only boundary and inner points:

$$D_+^i K_{-1} = (-1)^i \sum_{k=i}^{m} (-1)^k D_+^k K_0, \quad i = \overline{0, m} \qquad (4.44)$$

$$D_+^i X_{-1} = (-1)^i \sum_{k=i}^{m} (-1)^k D_+^k X_0, \quad i = \overline{0, m} \qquad (4.45)$$

It is useful to introduce

$$v_k \equiv (-1)^k D_+^k K_0, \quad z_k \equiv (-1)^k D_+^k X_0 \qquad (4.46)$$

and write the energy estimate in terms of v_k and z_k:

$$\dot{E} = -\beta \sum_{i=0}^{m} a_i v_i \sum_{k=i}^{m} v_k - \beta \sum_{i=0}^{m} a_i z_i \sum_{k=i}^{m} z_k - 2 \sum_{i=0}^{m} a_i v_i \sum_{k=i}^{m} z_k \tag{4.47}$$

The form (4.47) is further transformed by making the following notations:

$$\begin{aligned} \tilde{v}_i &\equiv \sum_{k=i}^{m} v_k \Rightarrow v_i = \tilde{v}_i - \tilde{v}_{i+1} \\ \tilde{z}_i &\equiv \sum_{k=i}^{m} z_k \Rightarrow z_i = \tilde{z}_i - \tilde{z}_{i+1} \end{aligned} \tag{4.48}$$

with $\tilde{v}_{m+1} = \tilde{z}_{m+1} = 0$.
Then

$$\begin{aligned} \dot{E} &= -\beta \sum_{i=0}^{m} a_i (\tilde{v}_i - \tilde{v}_{i+1}) \tilde{v}_i - \beta \sum_{i=0}^{m} a_i (\tilde{z}_i - \tilde{z}_{i+1}) \tilde{z}_i - 2 \sum_{i=0}^{m} a_i (\tilde{v}_i - \tilde{v}_{i+1}) \tilde{z}_i \\ &= -\left[\beta \left(\tilde{v}_0^2 - \tilde{v}_1 \tilde{v}_0 + \tilde{z}_0^2 - \tilde{z}_1 \tilde{z}_0\right) + 2 \left(\tilde{v}_0 \tilde{z}_0 - \tilde{v}_1 \tilde{z}_0\right)\right] \\ &\quad - \sum_{i=1}^{m-1} a_i \left[\beta \left(\tilde{v}_i^2 - \tilde{v}_{i+1} \tilde{v}_i + \tilde{z}_i^2 - \tilde{z}_{i+1} \tilde{z}_i\right) + 2 \left(\tilde{v}_i \tilde{z}_i - \tilde{v}_{i+1} \tilde{z}_i\right)\right] \\ &\quad - a_m \left[\beta \left(v_m^2 + z_m^2\right) + 2 v_m z_m\right] \end{aligned} \tag{4.49}$$

The existence of $a_i > 0$ such that $\dot{E} < 0$ will be proved by requiring the energy estimate to be of the following form:

$$\dot{E} = -\frac{1}{2} (\beta + 1) \sum_{i=0}^{m-1} (C_i^+)^2 - \frac{1}{2} (\beta - 1) \sum_{i=0}^{m-1} (C_i^-)^2 \tag{4.50}$$

where

$$\begin{aligned} C_i^+ &\equiv \tilde{v}_i + \gamma_i^+ \tilde{v}_{i+1} + \left(\tilde{z}_i + \rho_i^+ \tilde{z}_{i+1}\right) \\ C_i^- &\equiv \tilde{v}_i + \gamma_i^- \tilde{v}_{i+1} - \left(\tilde{z}_i + \rho_i^- \tilde{z}_{i+1}\right) \end{aligned} \tag{4.51}$$

and $\gamma_i^{\pm}, \rho_i^{\pm} \in \mathbb{R}$ are to be determined.

Using (4.51) the following sums are computed:

$$\begin{aligned} \sum_{i=0}^{m-1} (C_i^+)^2 &= \left[\tilde{v}_0^2 + \tilde{z}_0^2 + 2\gamma_0^+ \left(\tilde{v}_0\tilde{v}_1 + \tilde{v}_1\tilde{z}_0\right) + 2\tilde{v}_0\tilde{z}_0 + 2\rho_0^+ \left(+\tilde{v}_0\tilde{z}_1 + \tilde{z}_0\tilde{z}_1\right)\right] \\ &\quad + \sum_{i=1}^{m-1} \left[\left(1 + (\gamma_{i-1}^+)^2\right)\tilde{v}_i^2 + \left(1 + (\rho_{i-1}^+)^2\right)\tilde{z}_i^2 \right. \\ &\quad \left. + 2\gamma_i^+ \left(\tilde{v}_i\tilde{v}_{i+1} + \tilde{v}_{i+1}\tilde{z}_i\right) + 2\tilde{v}_i\tilde{z}_i\left(1 + \gamma_{i-1}^+\rho_{i-1}^+\right) + \rho_i^+ \left(+\tilde{v}_i\tilde{z}_{i+1} + \tilde{z}_i\tilde{z}_{i+1}\right)\right] \\ &\quad + (\gamma_{m-1}^+)^2 v_m^2 + (\rho_{m-1}^+)^2 z_m^2 + 2\gamma_{m-1}^+\rho_{m-1}^+ v_m z_m \end{aligned} \tag{4.52}$$

and

$$\begin{aligned} \sum_{i=0}^{m-1} (C_i^-)^2 &= \left[\tilde{v}_0^2 + \tilde{z}_0^2 + 2\gamma_0^- \left(\tilde{v}_0\tilde{v}_1 - \tilde{v}_1\tilde{z}_0\right) - 2\tilde{v}_0\tilde{z}_0 + 2\rho_0^- \left(-\tilde{v}_0\tilde{z}_1 + \tilde{z}_0\tilde{z}_1\right)\right] \\ &\quad + \sum_{i=1}^{m-1} \left[\left(1 + (\gamma_{i-1}^-)^2\right)\tilde{v}_i^2 + \left(1 + (\rho_{i-1}^-)^2\right)\tilde{z}_i^2 \right. \\ &\quad \left. + 2\gamma_i^- \left(\tilde{v}_i\tilde{v}_{i+1} - \tilde{v}_{i+1}\tilde{z}_i\right) - 2\tilde{v}_i\tilde{z}_i\left(1 + \gamma_{i-1}^-\rho_{i-1}^-\right) + \rho_i^- \left(-\tilde{v}_i\tilde{z}_{i+1} + \tilde{z}_i\tilde{z}_{i+1}\right)\right] \\ &\quad + (\gamma_{m-1}^-)^2 v_m^2 + (\rho_{m-1}^-)^2 z_m^2 + 2\gamma_{m-1}^-\rho_{m-1}^- v_m z_m \end{aligned} \tag{4.53}$$

with $\gamma_{-1}^{\pm} = \rho_{-1}^{\pm} = 0$.

Now the sums (4.52)–(4.53) are replaced in (4.50) and the new relation

is matched with (4.49). This leads to determining $\gamma_i^{\pm}$, $\rho_i^{\pm}$ and a_i:

$$\begin{aligned} \gamma_i^{\pm} &= -\frac{\beta \pm 2}{2(\beta \pm 1)} a_i, \quad i = 0, m-1 \\ \rho_i^{\pm} &= -\frac{\beta}{2(\beta \pm 1)} a_i, \quad i = 0, m-1 \end{aligned} \tag{4.54}$$

and

$$a_i = 1 + \frac{\beta^2}{4(\beta^2 - 1)} a_{i-1}^2, \quad i = 1, m-1 \tag{4.55}$$

$$a_m = \frac{\beta^2}{4(\beta^2 - 1)} a_{m-1}^2 \tag{4.56}$$

Because $\beta > 1$ all the numbers $a_i > 0$ are strictly positive.

After restoring the original variables (K and Φ), and h, the energy estimate is

$$\dot{E} = -\frac{1}{2}\lambda_+ C_+^2 - \frac{1}{2}\lambda_- C_-^2 \leq 0 \tag{4.57}$$

with

$$\begin{aligned} C_+^2 &\equiv \sum_{i=0}^{m-1} \left[\alpha_i^+ \sum_{k=i}^{m} (-1)^k (hD_+)^k K_0 + \frac{1}{h} \beta_i^+ \sum_{k=i}^{m} (-1)^k (hD_+)^{k+1} \Phi_0 \right]^2 \\ C_-^2 &\equiv \sum_{i=0}^{m-1} \left[\alpha_i^- \sum_{k=i}^{m} (-1)^k (hD_+)^k K_0 - \frac{1}{h} \beta_i^- \sum_{k=i}^{m} (-1)^k (hD_+)^{k+1} \Phi_0 \right]^2 \end{aligned} \tag{4.58}$$

where

$$\begin{aligned}
\alpha_0^{\pm} &= \beta_0^{\pm} = 1, \\
\alpha_i^{\pm} &= (-1)^i \left(1 - \frac{\beta \pm 2}{2(\beta \pm 1)} a_i\right), \quad i = 1, m \\
\beta_i^{\pm} &= (-1)^{i+1} \left(1 - \frac{\beta}{2(\beta \pm 1)} a_i\right), \quad i = 1, m
\end{aligned} \tag{4.59}$$

Summarizing, a set of strictly positive numbers, $\{a_i\}_{i=0,m}$ has been found ($a_0 = 1$ and a_i are given by (4.56) for $i \geq 1$) such that the norm (4.39) is not increasing in time. And this ends the proof of stability in respect to this norm.

Obs. In the limit $h \to 0$ and $N \to \infty$ the discrete energy estimate converges to the continuum energy estimate (4.18).

One can show, following the same steps as before, that using the extrapolation conditions (4.35), the same energy estimate as in (4.57) is obtained. But now,

$$\begin{aligned}
C_+^2 &\equiv \sum_{i=0}^{m-1} \left[\alpha_i^+ \sum_{k=i}^{m} (-1)^k (hD_+)^k K_0 + \frac{1}{h}\beta_i^+ \sum_{k=i}^{m-1} (-1)^k (hD_+)^{k+1} \Phi_0\right]^2 \\
C_-^2 &\equiv \sum_{i=0}^{m-1} \left[\alpha_i^- \sum_{k=i}^{m} (-1)^k (hD_+)^k K_0 - \frac{1}{h}\beta_i^- \sum_{k=i}^{m-1} (-1)^k (hD_+)^{k+1} \Phi_0\right]^2
\end{aligned} \tag{4.60}$$

In the particular case $m = 1$ the energy is:

$$E = \|u\|_{a_i,h}^2 = \|D_+\Phi\|_h^2 + \|K\|_h^2 + a_1 \left(\left\|(hD_+)^{i+1}\Phi\right\|_h^2 + \left\|(hD_+)^i K\right\|_h^2\right) \tag{4.61}$$

With the extrapolation conditions $D_+ K_{-1} = 0$ and $D_+^2 \Phi_{-1} = 0$, the

energy estimate is (4.57) with:

$$C_+ \equiv K_0 + D_+\Phi_0 - hD_+(K_0 + D_+\Phi_0) \tag{4.62}$$
$$C_- \equiv K_0 - D_+\Phi_0 - hD_+(K_0 - D_+\Phi_0) \tag{4.63}$$

while with the extrapolation conditions $D_+^2 K_{-1} = 0$ and $D_+^2\Phi_{-1} = 0$, the energy estimate is (4.57) with:

$$C_+ \equiv K_0 + D_+\Phi_0 - hD_+K_0 \tag{4.64}$$
$$C_- \equiv K_0 - D_+\Phi_0 - hD_+K_0 \tag{4.65}$$

This means that the problem is stable also when $m = 1$ in (4.26) or (4.25). However, in these cases the scheme is only first order convergent.

4.3.4 Stability Analysis via Laplace Transform Method

General Order Discussion

With $u_j = (X_j, K_j)^T$ the system (4.32) to be analyzed is put in the form:

$$\begin{aligned} \dot{v}_j(t) &= Qu_j + F, \quad Q = \begin{pmatrix} \beta D^{(1,n)} & D_+ \\ D_{\sim}^{(n)} & \beta D^{(1,n)} \end{pmatrix} \\ u_j(0) &= f_j \\ L_R(v) &= g(t); \end{aligned} \tag{4.66}$$

Note that Q can be written as

$$Q = \frac{1}{h}\sum_{\nu=-p}^{p} B_\nu S^\nu \tag{4.67}$$

where S^ν is the shift operator by ν points and B_ν are 2×2 matrices.

For the general systems described by the matrix operator Q, (4.67) necessary and sufficient conditions for stability and strong stability have been

derived in [36].

The first condition to be met is the semi-boundness of Q for the Cauchy problem. This will be shown below. The others are conditions on the solutions of the Laplace transformed system and are presented in the paragraph b).

a) Semi-Boundness of Q for the Cauchy Problem

Q is semi-bounded for the Cauchy problem — that is,

$$(w, Qw)_h + (Qw, w)_h \leq 2\alpha(w, w)_h \tag{4.68}$$

Proof The proof is done in Fourier space. If $\hat{Q}$ is the Fourier symbol of Q then the relation (4.68) becomes:

$$\hat{Q} + \hat{Q}^* \leq 2\alpha I \tag{4.69}$$

It is easy to see that the Fourier symbol of the operator $D_{\sim}^{(n)}$ (defined in (4.33)) is $\hat{D}_{\sim}^{(n)} = -\frac{1}{h}\check{d}_+^* \frac{\check{d}^{(2)}}{\Omega^2}$. Then

$$\hat{Q} = \frac{1}{h}\begin{pmatrix} i\beta\check{d}^{(1,n)} & \check{d}_+ \\ -\check{d}_+^* \frac{\check{d}^{(2)}}{\Omega^2} & i\beta\check{d}^{(1,n)} \end{pmatrix} \tag{4.70}$$

For x and y arbitrary complex numbers,

$$\begin{aligned} & (x*, y*)(\hat{Q} + \hat{Q}^*)(x, y)^T = 2(1 - \frac{d^{(2)}}{\Omega^2})\mathrm{Re}\ (\check{d}_+ x^* y) \\ \leq\ & 2\left|1 - \frac{d^{(2)}}{\Omega^2}\right| \left|\check{d}_+ x^* y)\right| = 2\left(|\Omega| \sum_{k=1}^{n-1} |d_k|\, \Omega^{2k}\right) |x|\, |y| \\ \leq\ & 2\sum_{k=1}^{n-1} |d_k|\, 2^{2k} \left(|x|^2 + |y|^2\right) \end{aligned} \tag{4.71}$$

Taking $\alpha = \sum_{k=1}^{n-1} |d_k|\, 2^{2k}$, the inequality (4.69) follows. Using Parseval relation, (4.68) is immediately obtained.

Notice that the semi-boundess condition is equivalent with the well-posedness of the Cauchy problem (and also with the strong hyperbolicity of the system).

b) Laplace Transformed System

The system (4.66) with zero initial data and zero forcing terms is Laplace transformed using $\hat{u}(s) = \int_0^\infty e^{-st} u(t) dt$:

$$\begin{aligned} \tilde{s}\hat{\mathbf{v}}_j &= Q\hat{\mathbf{v}}_j \\ L\hat{v}_0 &= \hat{g}(t); \end{aligned} \tag{4.72}$$

According to [36], if Q has the form (4.67) and is semi-bounded for the Cauchy problem then

1. a necessary condition for stability is: the problem (4.72) does not admit bounded solutions of the type $\hat{v}_j = k^j \hat{v}_0$ for Re $(\tilde{s}) > 0$ (Ryabenkii-Godonov condition).

2. a sufficient condition for strong stability is: the problem (4.72) does not admit bounded solutions of the type $\hat{v}_j = k^j \hat{v}_0$ for Re $(\tilde{s}) \geq 0$ (Kreiss condition). The Kreiss condition is equivalent with requiring

$$\sum_{i=-n}^{n-1} |\hat{v}_i|^2 \leq \text{const} \, |\hat{g}|^2 \tag{4.73}$$

that is, that the values of the points close to the boundary ($i = \overline{-n, n-1}$) are bounded in terms of boundary data.

c) Looking for Bounded Solutions

Consider now bounded solutions of the type $\hat{v}_j = k^j \hat{v}_0$. Inserting this in the system gives:

$$\begin{aligned} \tilde{s}\hat{\mathbf{v}}_0 &= \bar{Q}\hat{\mathbf{v}}_0, \quad \bar{Q} = \frac{1}{h}\begin{pmatrix} \beta\bar{d}^{(1,n)} & \bar{d}_+ \\ \bar{d}_\sim & \beta\bar{d}^{(1,n)} \end{pmatrix} \\ \hat{L}\hat{v}_0 &= \hat{g}(t); \end{aligned} \tag{4.74}$$

and $\bar{d} \equiv k^{-j} D k^j$ for any FDO D.
The barred operators corresponding to the elementary operators are:

$$\begin{array}{lllllll} \bar{d}_+ &=& k-1 & \bar{\delta} &=& \frac{1}{2}\left(k-k^{-1}\right), & \bar{\delta}' &=& \frac{1}{2}(1+\frac{1}{k^2}) \\ \bar{d}_- &=& 1-k^{-1} & \bar{p} &=& k^{-1}(k-1)^2 & \bar{p}' &=& 1-\frac{1}{k^2} = 2\frac{\delta}{k} \end{array} \tag{4.75}$$

Notice that $\bar{d}_- = k^{-1}\bar{d}_+$ and $\bar{\delta}^2 = \bar{p} + \frac{\bar{p}^2}{4}$.
The barred operators corresponding to the discrete derivatives $D^{(1,n)}$, $D^{(2)}$ and the rest $R = D^{(1,n)}D^{(1,n)} - D^{(2)}$ satisfy:

$$\begin{aligned} \bar{d}^{(1,n)} &= \frac{1}{2}\left(k-k^{-1}\right)\sum_{l=0}^{n-1} c_l k^{-l}(k-1)^{2l} \\ \bar{d}^{(2)} &= k^{-1}(k-1)^2 \sum_{l=0}^{n-1} d_l k^{-l}(k-1)^{2l} \\ \bar{r} &= \kappa_n \bar{p}^{n+1} \sum_{l=0}^{n-1} (-1)^l \frac{(l!)^2}{(2l+1)!(s+l+1)} \bar{p}^l \\ \bar{d}^{(2)'} &= \frac{p'}{\delta}\bar{d}^{(1,n)} = \frac{2}{k}\bar{d}^{(1,n)} \\ \bar{d}^{(1,n)'} &= \frac{1}{k}\left(1+\kappa_n\bar{p}^n\right) \\ \bar{r}' &= \kappa_n \bar{p}^n \frac{2}{k}\bar{d}^{(1,n)} = \kappa_n \bar{p}^n \bar{d}^{(2)'} \end{aligned} \tag{4.76}$$

d) Characteristic Equation

The characteristic equation of the system (4.74) is a polynomial of order $4n$ in k:

$$\left(\tilde{s} - \beta \bar{d}^{(1,n)}\right)^2 - \bar{d}^{(2)} = 0 \tag{4.77}$$

properties:

Lemma 4.3.1 *For $Re(s) > 0$ there are no solutions with $|k| = 1$.*

Proof Suppose that for $Re(s) > 0$ there is a solution $k = e^{i\xi}$. In this case, $\bar{d}^{(1,n)}$ and $\bar{d}^{(2)}$ are actually the Fourier symbols of the operators $D^{(1,n)}$ and $D^{(2)}$:

$$\bar{d}^{(1,n)} = i\hat{d}^{(1,n)} \in i\Re \tag{4.78}$$

$$\bar{d}^{(2)} = -\hat{d}^{(2)} \in \Re_- \tag{4.79}$$

where $\hat{d}^{(1,n)} = \hat{d}^{(1,n)}$ and $\hat{d}^{(2,n)} = \hat{d}^{(2,n)}$ with $\hat{d}^{(1,n)}$ and $\hat{d}^{(2,n)}$ computed in 2.1.4. Inserting these relations together with $s = a + ib$ (with $a > 0$ and $b \in \Re$) in (4.77) gives:

$$\left[a^2 - b^2 + \hat{d}^{(2)} + \beta(\hat{d}^{(1,n)}) \left(2b - \beta(\hat{d}^{(1,n)})\right)\right] + 2ia \left(b - \beta \hat{d}^{(1,n)}\right) = 0 \tag{4.80}$$

So it is necessary to have $b = \beta \hat{d}^{(1,n)}$ which leads to

$$a^2 + \hat{d}^{(2)} = 0 \tag{4.81}$$

which gives a contradiction with $a = Re(s) > 0$ and $\hat{d}^{(2)} \geq 0$.

Lemma 4.3.2 *For $Re(s) > 0$ there are 2n and only 2n roots inside the unit circle.*

Proof The roots are continuous functions of s and for large s,

$$s \simeq \beta\bar{\mathcal{D}}^{(1,n)} \pm \sqrt{\bar{\mathcal{D}}^{(2)}}, \quad \text{where} \tag{4.82}$$

$$\bar{\mathcal{D}}^{(1,n)} = \frac{1}{2}k\sum_{l=0}^{n} c_l k^l \tag{4.83}$$

$$\bar{\mathcal{D}}^{(2)} = k\sum_{l=0}^{n} d_l k^l\,. \tag{4.84}$$

The relation (4.82) is a polynomial of order $2n$ in k so it has $2n$ solutions (outside the unit circle). So the other $2n$ solutions of the characteristics equation must be inside the unit circle. Note also that $\lim_{s\to\infty} k_i(s) = 0$ for $|k_i(s)| < 1$.

e) General Solution

Denote with ν the number of distinct roots inside the unit circle. The roots and their multiplicities are collected in pairs (k_i, m_i), $i = \overline{1,\nu}$ with $\sum_{i=1}^{\nu} m_i = 2n$. It is convenient to introduce the following notations:

$$z_i \equiv \bar{d}_+(k_i), \quad t_i \equiv \frac{1}{z_i}\left(s - \beta\bar{d}_1(k_i)\right) \tag{4.85}$$

From the characteristic equation (4.77), t_i obeys also:

$$t_i^2 = \frac{1}{z_i^2}\bar{d}^{(2)}(k_i) = \frac{1}{k_i}\left(1 + \sum_{l=1}^{n-1} d_l \frac{z_i^{2l}}{k_i^l}\right) \tag{4.86}$$

Then the general solution is:

$$\begin{aligned} X_j &= \sum_{i=1}^{\nu} p_i(j)k_i^j \\ K_j &= \sum_{i=1}^{\nu} p_i(j)t_i k_i^j \end{aligned} \tag{4.87}$$

where $p_i(j)$ is a polynomial in j of order $m_i - 1$, with coefficients $\{\sigma_{\sum_{l=1}^{m_{i-1}}}, \ldots, \sigma_{\sum_{l=1}^{m_i}}\}$.

By defining the matrices:

$$A = \begin{pmatrix} 1 & 1 & ... & 1 \\ t_1 & t_2 & ... & t_\nu \end{pmatrix}, \quad P_i = \begin{pmatrix} 1 & j & \dots & j^{m_i-1} \end{pmatrix} \tag{4.88}$$

$$I_k^j = \begin{pmatrix} k_1^j P_1 & 0 & & \\ 0 & k_2^j P_2 & 0 & \\ & 0 & \dots & 0 \\ & & 0 & k_\nu^j P_\nu \end{pmatrix}, \quad \sigma = \begin{pmatrix} \sigma_1 \\ \sigma_2 \\ . \\ . \\ \sigma_{2n} \end{pmatrix} \tag{4.89}$$

the general solution can be written in matrix form as:

$$\hat{v}_j = A I_k^j \sigma \tag{4.90}$$

Obs. If $\nu = 2n$ (all the roots are distinct) then $p_i(j) = \sigma_i$, $P_i = 1$ for all $i = \overline{1, 2n}$ and I_k^j becomes a diagonal matrix with entries $\{k_i^j\}_{i=1}^{2n}$.

f) Applying Boundary Conditions

Inserting the general solution (4.90) in the Laplace transformed boundary conditions, leads to a linear system in unknowns σ:

$$B\sigma = \hat{g} \tag{4.91}$$

For stability it is necessary that the matrix B is invertible. If this is the case, the solution $\sigma = B^{-1}\hat{g}$ of the linear system is inserted in the general solution (4.90), leading to

$$\hat{v}_j = A I_k^j B^{-1} \hat{g} \tag{4.92}$$

The solutions (4.92) with $j = \overline{-n, n-1}$ are now plugged into the Kreiss condition (4.73). If the condition holds then the problem is strongly stable.

The paragraphs a)–f) presented the path for analyzing the stability of the semi-discrete boundary value problem (4.66). Also several properties of the

system have been discussed. However, a full analysis requires more knowledge about the roots of the characteristic equation (4.77). These are not easy to determine for general order $2n$ and remain an open issue for further investigation. The case of 2nd order accuracy with outflow boundary condition is relatively easy to analyze and it is presented below.

Particular Case $n = 1$

When $n = 1$ the characteristic equation

$$\left(\tilde{s} - \beta\bar{\delta}\right)^2 - \bar{p} = 0 \tag{4.93}$$

admits 4 roots: $k_{1,2,3,4}(s)$. According to the lemmas 4.3.1–4.3.2, for Re $(s) > 0$, two roots are inside the unit circle ($k_{1,2}$) and two outside ($k_{3,4}$). For analysis it is needed to know the behavior of the roots when $s = 0$ and in a neighborhood of $s = 0$.

The Roots when $s = 0$

The characteristic equation with $s = 0$ is

$$\beta^2 \frac{(k^2-1)^2}{4k^2} = \frac{(k-1)^2}{k}\,. \tag{4.94}$$

If $\beta = 0$ there are only two solutions: $k_1^* = k_2^* = 1$.
If $\beta \neq 0$ there are four solutions:

$$k_{1,2}^* = \frac{2 - \beta^2 \pm 2\sqrt{1-\beta^2}}{\beta^2}, \quad k_{3,4}^* = 1\,. \tag{4.95}$$

- If $\beta > 1$ then $\left|k_{1,2}^*\right| = 1$ and $k_1^* = \bar{k}_2^*$. So all the four roots are on the unit circle
- If $|\beta| < 1$ and $\beta \neq 0$ then $k_1^* > 1$, $0 < k_2^* < 1$ ($k_1^* k_2^* = 1$). So there is one root inside, one outside and two on the unit circle.

The Roots in a Neighborhood of $s = 0$

By perturbation arguments one can show that in a neighborhood of $s = 0$ the roots have the form

$$k_1^*(s) = \frac{2 - \beta^2 + 2\sqrt{1-\beta^2}}{\beta^2} + \mathcal{O}(|s|) \tag{4.96}$$

$$k_2^*(s) = \frac{2 - \beta^2 - 2\sqrt{1-\beta^2}}{\beta^2} + \mathcal{O}(|s|) \tag{4.97}$$

$$k_3^*(s) = 1 + \frac{1}{\beta+1}s + \mathcal{O}(|s|^2) \tag{4.98}$$

$$k_4^*(s) = 1 + \frac{1}{\beta-1}s + \mathcal{O}(|s|^2)\,. \tag{4.99}$$

Lemma 4.3.3 *In the limit $s \to 0$, the roots inside the unit circle for Re $(s) > 0$, $k_{1,2}$ become:*

- *k_1^* and k_2^*, if $\beta > 1$*
- *k_2^* and k_4^*, if $-1 < \beta < 1$*
- *k_2^* and k_4^*, if $\beta < -1$.*

Multiplicity of the Roots

Lemma 4.3.4 *For Re $(s) > 0$ the roots $k_{1,2}$ are*

- *distinct if $\beta > 1$ or $-1 < \beta < 0$*
- *not necessarily distinct if $0 < \beta < 1$.*

Proof Assume that k is a double root of the characteristic equation. Then

$$-2\beta(s - \beta\bar{\delta})\bar{\delta}' - \bar{p}' = 0 \tag{4.100}$$

Together with the characteristic equation (4.93), this forms a system to be solved for β and s:

$$\beta = \pm\frac{\sqrt{k}(1+k)}{1+k^2} \tag{4.101}$$

$$s = \pm\frac{(1-k)^3}{2\sqrt{k}\,(k^2+1)}\,. \tag{4.102}$$

With $k = re^{i\xi}$, $\xi \in (-\pi, \pi]$, $r \in (0,1)$ the previous relations give

$$\text{Im}\,(\beta) = \pm\frac{-(r-1)\sqrt{r}\,[(r+1)^2 + 2r\cos\xi]\sin\frac{\xi}{2}}{r^4 + 2r^2\cos 2\xi + 1}. \tag{4.103}$$

Because β is real, it follows that $\xi = 0$. This means $k = r \in (0,1)$ and $s \in \Re$.

- If $|\beta| \geq 1$, one can show that the equation (4.101) has no solutions in $(0,1)$. So there are no double roots in this case, $\forall \tilde{s} \in \mathbb{C}$
- If $-1 < \beta < 0$ then $s < 0$ which is in contradiction with $Re(s) > 0$. So there are no double roots.
- If $0 < \beta < 1$ then the equation (4.101) has one and only one solution in $(0,1)$.

Boundness Relations for the Roots

Lemma 4.3.5 *For $\beta > 1$, there is a constant $\rho > 0$ such that, for any compact set $|s| \leq C$, $Re(s) \geq 0$, the roots k_1, k_2 satisfy:*

$$|k_i - 1| \geq \rho, \quad i = 1, 2 \tag{4.104}$$

Proof The relation holds for large enough s because when $s \to \infty$, $k_i(s) \to 0$, $\forall\, |k_i(s)| < 1$. This means that it can be violated, if for some s_0 there is a root k_i such that $k_i(s_0) = 1$. Then it follows that $s_0 = 0$. However, by lemma 4.3.3, for $\beta > 1$, $k_i(0) = k_i^* \neq 1$, $i = 1, 2$, which is a contradiction.

Applying Boundary Conditions: Outflow Boundary

In case $\beta > 1$ the roots k_1 and k_2 are distinct. For second order accuracy, according to the notations (4.85)–(4.86), $k_i = \frac{1}{t_i^2}$ and $z_i = \frac{1}{t_i^2} - 1$. The general solution is given by (4.90),

$$\hat{v}_j = AI_k^j \sigma \quad \text{with} \quad A = \begin{pmatrix} 1 & 1 \\ t_1 & t_2 \end{pmatrix} \quad \text{and} \quad I_k^j = \begin{pmatrix} k_1^j & 0 \\ 0 & k_2^j \end{pmatrix}. \tag{4.105}$$

Now the numerical boundary conditions are applied, using the prescriptions (4.34). For the Laplace transformed system they are:

$$(hD_+)^m \hat{v}_{-1} = \hat{g} \tag{4.106}$$

Inserting the general solution (4.105) in (4.106) the following system is obtained:

$$B\sigma = \hat{g} \quad \text{with} \quad B = AI_z^m I_k^{-1} \quad \text{and} \quad I_z^m = \begin{pmatrix} z_1^m & 0 \\ 0 & z_2^m \end{pmatrix}. \tag{4.107}$$

Because the roots $k_{1,2}$ are distinct it follows that $\det A \neq 0$ and the matrix $B = AI_z^m I_k^{-1}$ is invertible. The system is solved now for σ and the boundary vectors $\hat{v}_{-1,0}$ are evaluated: $\hat{v}_{-1} = AI_z^{-m} A^{-1} \hat{g}$ and $\hat{v}_0 = AI_k I_z^{-m} A^{-1} \hat{g}$. Proving the Kreiss condition reduces to checking the following inequality:

$$\left| AI_z^{-m} A^{-1} \hat{g} \right|^2 + \left| AI_k I_z^{-m} A^{-1} \hat{g} \right|^2 \leq \text{const} \, |\hat{g}|^2 \tag{4.108}$$

The direct evaluation of the product $AI_k^p I_z^{-m} A^{-1}$ (where $p \in \{0, 1\}$) gives:

$$AI_k^p I_z^{-m} A^{-1} = z_1^{-m} z_2^{-m} \begin{pmatrix} -\frac{z_1^m k_2^p t_1 - z_2^m k_1^p t_2}{t_2 - t_1} & \frac{\left(z_1^m k_2^p - z_2^m k_1^p\right)}{t_2 - t_1} \\ -\frac{\left(z_1^m k_2^p - z_2^m k_1^p\right) t_1 t_2}{t_2 - t_1} & \frac{z_1^m k_2^p t_2 - z_2^m k_1^p t_1}{t_2 - t_1} \end{pmatrix}$$

By lemma 4.3.5, the quantities $z_i^{-m} = (k_i - 1)^{-m}$ are bounded for any $m \in \mathbb{N}$. This means that it is enough to prove that each element of the matrix $M \equiv z_1^m z_2^m A I_k^p I_z^{-m} A^{-1}$ is bounded in order to satisfy the Kreiss condition.

If $x \equiv \frac{1}{t_1}$ and $y \equiv \frac{1}{t_2}$ then $|x| < 1$, $|y| < 1$ and M becomes:

$$M = \begin{pmatrix} \frac{x^{2n+1}\left(y^2-1\right)^m - \left(x^2-1\right)^m y^{2n+1}}{x-y} & \frac{xy\left[\left(x^2-1\right)^m y^{2n} - x^{2n}\left(y^2-1\right)^m\right]}{x-y} \\ \frac{x^{2n}\left(y^2-1\right)^m - \left(x^2-1\right)^m y^{2n}}{x-y} & \frac{x\left(x^2-1\right)^m y^{2n} - x^{2n} y\left(y^2-1\right)^m}{x-y} \end{pmatrix}$$

It is easy to see that for any $m, n \in \mathbb{N}$ each element of M is bounded as long as x, y are bounded. So the Kreiss condition holds and the outflow discretization (4.34) is strongly stable for $2n = 2$. In a similar way one can prove that also the extrapolation conditions (4.35) lead to strong stability.

4.4 SBP-SAT Method for Inflow Boundary

This section shows how to apply the SBP-SAT procedure presented in 4.1.2 to the implementation of the inflow boundary ($-1 < \beta < 1$) for the shifted wave equation. The discrete analysis is based on a different energy estimate at the continuum level than presented in 4.2, and this energy is well defined only for this case (of inflow boundary).

4.4.1 Another Continuum Energy Estimate

Suppose $\mathsf{v}_{(i)} = (\mathsf{K}_{(i)}, \Phi_{(i)})^T$ for $i = 1, 2$, and $\mathsf{v} = (\mathsf{K}, \Phi)^T$ where $\mathsf{K}_{(1)}, \mathsf{K}_{(2)}, \mathsf{K} \in C^0(\mathbb{R})$ and $\Phi_{(1)}, \Phi_{(2)}, \Phi \in C^1(\mathbb{R})$. Consider the following scalar product and the associated norm:

$$(\mathsf{v}_{(1)}, \mathsf{v}_{(2)})_{\text{inflow}} \equiv \int_{x_0}^{x_1} \left[\mathsf{K}_{(1)}\mathsf{K}_{(2)} + (\partial_x \Phi_{(1)})(\partial_x \Phi_{(2)}) + \beta \left(\mathsf{K}_{(1)}\partial_x \Phi_{(2)} + \mathsf{K}_{(2)}\partial_x \Phi_{(1)}\right)\right] dx \tag{4.109}$$

$$\|\mathsf{v}\|^2_{\text{inflow}} = \int_{x_0}^{x_1} \left[\mathsf{K}^2 + (\partial_x \Phi)^2 + 2\beta \mathsf{K}\partial_x \Phi\right] dx \tag{4.110}$$

where $\|\mathsf{v}\|^2_{\text{inflow}} \equiv (\mathsf{v}, \mathsf{v})_{\text{inflow}}$. These two quantities are well-defined only for $|\beta| < 1$. In terms of characteristic variables $\mathsf{C}_{(1)\pm} = \mathsf{K}_{(1)} \pm \partial_x \Phi_{(1)}$, $\mathsf{C}_{(2)\pm} = \mathsf{K}_{(2)} \pm \partial_x \Phi_{(2)}$, $\mathsf{C}_{\pm} = \mathsf{K} \pm \partial_x \Phi$, and speeds $\lambda_{\pm} = \beta \pm 1$, the relations (4.109–4.110) are written as:

$$(\mathsf{v}_{(1)}, \mathsf{v}_{(2)})_{\text{inflow}} = \frac{1}{2}\left(\lambda_+ \mathsf{C}_{(1)+}\mathsf{C}_{(2)+} - \lambda_+ \mathsf{C}_{(1)-}\mathsf{C}_{(2)-}\right) \tag{4.111}$$

$$\|\mathsf{v}\|^2_{\text{inflow}} = \frac{1}{2}\int_{x_0}^{x_1} \left(\lambda_+ \mathsf{C}_+^2 - \lambda_- \mathsf{C}_-^2\right) dx. \tag{4.112}$$

Lemma 4.4.1 *The inflow boundary problem (4.15) with maximally dissipative boundary conditions is strongly well-posed in respect to the energy norm*

defined in (4.110) if the reflection coefficients are chosen so that

$$\lambda_+^2 - \lambda_-^2 R_0^2 > 0 \quad \textit{and} \quad \lambda_-^2 - \lambda_+^2 R_1^2 > 0. \tag{4.113}$$

Proof Applying the evolution equations (4.15) leads to

$$\frac{d}{dt}\mathsf{C}_\pm = \lambda_\pm \partial_x \mathsf{C}_\pm + \mathsf{F}^{\mathsf{K}} \pm \partial_x \mathsf{F}^{\Phi}. \tag{4.114}$$

The energy estimate is

$$\frac{d}{dt}\left\|\mathsf{v}\right\|_{\text{inflow}}^2 = \frac{1}{2}\lambda_+^2 \mathsf{C}_+^2 - \frac{1}{2}\lambda_-^2 \mathsf{C}_-^2 \Big|_{x_0}^{x_1} + 2(\mathsf{v},\mathsf{F})_{\text{inflow}} \tag{4.115}$$

where $\mathsf{F} = (\mathsf{F}^K, \mathsf{F}^\Phi)^T$. Imposing the maximally dissipative boundary conditions, $\mathsf{C}_- = R_0 \mathsf{C}_+ + \mathsf{g}_0(t)$ and $\mathsf{C}_+ = R_1 \mathsf{C}_- + \mathsf{g}_1(t)$, the energy estimate becomes:

$$\begin{aligned}
\frac{d}{dt}\left\|\mathsf{v}\right\|_{\text{inflow}}^2 &= \frac{1}{2}\left[\left(\lambda_-^2 R_0^2 - \lambda_+^2\right)\mathsf{C}_+^2(t,x_0) + 2\lambda_-^2 R_0 \mathsf{C}_+(t,x_0)\mathsf{g}_0(t) + \lambda_-^2 \mathsf{g}_0^2(t)\right. \\
&+ \left.\left(\lambda_+^2 R_1^2 - \lambda_-^2\right)\mathsf{C}_-^2(t,x_1) + 2\lambda_+^2 R_1 \mathsf{C}_-(t,x_1)\mathsf{g}_1(t) + \lambda_+^2 \mathsf{g}_1^2(t)\right] \\
&+ 2(\mathsf{v},\mathsf{F})_{\text{inflow}}
\end{aligned} \tag{4.116}$$

With $\mathsf{V} \equiv \lambda_+ \mathsf{C}_+ - \lambda_- \mathsf{C}_-$ the energy estimate is written:

$$\begin{aligned}
\frac{d}{dt}\left\|\mathsf{v}\right\|_{\text{inflow}}^2 &= -\frac{(\lambda_+ + \lambda_- R_0)^2}{2\left(\lambda_+^2 - \lambda_-^2 R_0^2\right)}\left[\mathsf{V}(t,x_0) + \frac{\lambda_-\lambda_+}{\lambda_+ + \lambda_- R_0}\mathsf{g}_0(t)\right]^2 + \frac{\lambda_-^2\lambda_+^2}{2\left(\lambda_+^2 - \lambda_-^2 R_0^2\right)}(\mathsf{g}_0(t))^2 \\
&\quad -\frac{(\lambda_- + \lambda_+ R_1)^2}{2\left(\lambda_-^2 - \lambda_+^2 R_1^2\right)}\left[\mathsf{V}(t,x_1) + \frac{\lambda_+\lambda_-}{\lambda_- + \lambda_+ R_1}\mathsf{g}_1(t)\right]^2 + \frac{\lambda_+^2\lambda_-^2}{2\left(\lambda_-^2 - \lambda_+^2 R_1^2\right)}(\mathsf{g}_1(t))^2 \\
&+ 2(\mathsf{v},\mathsf{F})_{\text{inflow}}
\end{aligned} \tag{4.117}$$

Choosing $R_{0,1}$ according to (4.113) and applying also the inequality $2(\mathsf{v},\mathsf{F})_{\text{inflow}} \leq$

$\|\mathsf{v}\|^2_{\text{inflow}} + \|\mathsf{F}\|^2_{\text{inflow}}$, leads to

$$\frac{d}{dt}\|\mathsf{v}\|^2_{\text{inflow}} \leq \frac{\lambda_-^2\lambda_+^2}{2\left(\lambda_+^2 - \lambda_-^2 R_0^2\right)}(\mathsf{g}_0(t))^2 + \frac{\lambda_+^2\lambda_-^2}{2\left(\lambda_-^2 - \lambda_+^2 R_1^2\right)}(\mathsf{g}_1(t))^2 + \|\mathsf{v}\|^2_{\text{inflow}} + \|\mathsf{F}\|^2_{\text{inflow}} \tag{4.118}$$

By integrating in time the relation (4.118), the strongly well-posed estimate (4.3) is obtained.

4.4.2 Stability Analysis

The discretization of the wave equation in SBP-SAT fashion is:

$$\begin{aligned} \dot{\Phi} &= \beta D_1 \Phi + K + F^{\Phi} \\ \dot{K} &= \beta D_1 K + D_2 \Phi + \tau^0 P_0 + \tau^1 P_1 + F^K \end{aligned} \tag{4.119}$$

In the relations above, D_1 and D_2 are SBP operators corresponding to the first and second derivative — they satisfy the properties (4.11), (4.12) and (4.13). The terms $P_{0,1}$ are the penalty terms corresponding to the boundary conditions. These penalty terms are added to the equations with some factors $\tau^{0,1}$ in such a way that an energy estimate exists and mimics the continuum one. Now, before defining a discrete energy corresponding to the continuum (4.110), it is useful to introduce the following notations:

$$\begin{aligned} \Delta &\equiv A - Q^T \Sigma^{-1} Q \\ \delta &\equiv D_1 - S \end{aligned} \tag{4.120}$$

The following assumptions are made:

$$A = A^T > 0 \tag{4.121}$$

$$\Delta > 0 \tag{4.122}$$

Obs. All the SBP operators based on diagonal norms from [57] obey (4.121). However, the second condition, (4.122), is only satisfied by the SBP operators with 2nd and 4th order accurate interior stencils and 1st and respectively 2nd accuracy boundary closures (Appendix C.1–C.2 from [57]). In the appendix D of this thesis is presented another set of SBP operators corresponding to 6th order accuracy in the interior and 3rd accuracy at the boundary, which satisfies in addition (4.122).

Suppose $v_{(i)} = (K_{(i)}^T, \Phi_{(i)}^T)^T$ for $i = 1, 2$ and $v = (K^T, \Phi^T)^T$ are grid vector functions and consider the following scalar product and associated norm:

$$(v_{(1)}, v_{(2)})_{\text{inflow},\Sigma} \equiv K_{(1)}^T \Sigma K_{(2)} + \Phi_{(1)}^T A \Phi_{(2)} + \beta \left(K_{(1)}^T Q \Phi_{(2)} + K_{(2)}^T Q \Phi_{(1)} \right) \tag{4.123}$$

$$\|v\|^2_{\text{inflow},\Sigma} \equiv (v, v)_{\text{inflow},\Sigma} = K^T \Sigma K + \Phi^T A \Phi + 2\beta K^T Q \Phi \tag{4.124}$$

These two quantities are well-defined if $|\beta| < 1$ and the assumptions (4.121–4.122) hold. In terms of the discrete characteristics, $C_{(i)\pm} = K_{(i)} \pm D_1 \Phi_{(i)}$, $i = 1, 2$, $C_\pm = K \pm D_1 \Phi$, the energy (4.124) is:

$$\begin{aligned} (v_{(1)}, v_{(2)})_{\Sigma\ ,\text{inflow}} &= \frac{1}{2} \left(\lambda_+ C_{(1)}^T \Sigma C_{(2)} - \lambda_- C_{(2)}^T \Sigma C_{(1)} \right) + \Phi_{(1)} \Delta \Phi_{(2)} \\ \|v\|^2_{\text{inflow},\Sigma} &= \frac{1}{2} \left(\lambda_+ C_+^T \Sigma C_+ - \lambda_- C_-^T \Sigma C_- \right) + \Phi \Delta \Phi \end{aligned} \tag{4.125}$$

Lemma 4.4.2 *The discretization (4.119) of the inflow boundary problem is strongly and strictly stable in respect to the norm (4.124) if the following conditions hold:*

1. *the SBP operators satisfy the assumptions (4.121–4.122)*

2. the penalty terms are given by [6]

$$\begin{aligned} P_0 &= \lambda_- \Sigma^{-1} E_0 \left(C_- - R_0 C_+ - \mathrm{g}_0 \right) + \left(-1 + \frac{\lambda_-}{\lambda_+} R_0 \right) \Sigma^{-1} E_0 \delta \\ P_1 &= \lambda_+ \Sigma^{-1} E_1 \left(C_+ - R_1 C_- - \mathrm{g}_1 \right) + \left(-1 + \frac{\lambda_+}{\lambda_-} R_1 \right) \Sigma^{-1} E_1 \delta \end{aligned} \tag{4.127}$$

3. the coefficients $\tau^{0,1}$ of the penalty terms are:

$$\tau^0 = \frac{1}{1 - R_0 \frac{\lambda_-}{\lambda_+}}, \quad \tau^1 = -\frac{1}{1 - R_1 \frac{\lambda_+}{\lambda_-}} \tag{4.128}$$

4. the reflection coefficients $R_{0,1}$ satisfy (4.113)

Proof Take a time derivative of the energy (4.125):

$$\frac{d}{dt} \left\| v \right\|^2_{\mathrm{inflow},\Sigma} = \lambda_+ C_+^T \Sigma \frac{d}{dt} C_+ - \lambda_- C_-^T \Sigma \frac{d}{dt} C_- + 2 \Phi \Delta \frac{d}{dt} \Phi \tag{4.129}$$

Computing $\frac{d}{dt} C_\pm$ using the discrete evolution equations gives:

$$\frac{d}{dt} C_\pm = \dot{K} \pm D_1 \dot{\Phi} = \lambda_\pm D_1 C_\pm - \left(D_1^2 - D_2 \right) \Phi + \tau^i P_i + F^K \pm D_1 F^\Phi \tag{4.130}$$

Evaluate now $C_\pm^T \Sigma \frac{d}{dt} C_\pm$ and apply the relations $D_1 = \Sigma^{-1} Q$ and $D_1^2 - D_2 = \Sigma^{-1} \left(\Delta + B \delta \right)$:

$$C_\pm^T \Sigma \frac{d}{dt} C_\pm = \frac{1}{2} \lambda_\pm C_\pm^T B C_\pm - C_\pm^T \left(\Delta + B \delta \right) \Phi + C_\pm^T \Sigma (\tau^i P_i + F^K \pm D_1 F^\Phi) \tag{4.131}$$

[6] With the notation $\tilde{C}_\pm \equiv C_\pm - \frac{1}{\lambda_\pm} (\delta \Phi)$ the penalty terms can be written in a more compact form as:

$$\begin{aligned} P_0 &= \lambda_- \Sigma^{-1} E_0 \left(\tilde{C}_- - R_0 \tilde{C}_+ - \mathrm{g}_0 \right) \\ P_1 &= \lambda_+ \Sigma^{-1} E_1 \left(\tilde{C}_+ - R_1 \tilde{C}_- - \mathrm{g}_1 \right) \end{aligned} \tag{4.126}$$

Replace the above relation in (4.129):

$$\begin{aligned}\frac{d}{dt}\|v\|^2_{\text{inflow},\Sigma} &= \frac{1}{2}\left(\lambda_+^2 C_+^T B C_+ + \lambda_-^2 C_-^T B C_-\right)\\ &- \left(\lambda_+ C_+ - \lambda_- C_-\right)\left[\Sigma\tau^i P_i - (\Delta + B\delta)\,\Phi\right]\\ &+ \left[\lambda_+ C_+^T \Sigma(F^K + D_1 F^\Phi) - \lambda_- C_-^T \Sigma(F^K - D_1 F^\Phi)\right]\\ &+ 2\Phi^T \Delta \frac{d}{dt}\Phi \end{aligned} \tag{4.132}$$

The evolution equation for Φ can also be written as:

$$2\frac{d}{dt}\Phi = \lambda_+ C_+ - \lambda_- C_- + 2F^\Phi.$$

By using this relation in (4.132), all the terms containing Δ disappear and the sum of all terms containing source functions gives the scalar product $2(v,F)_{\text{inflow},\Sigma}$.

$$\begin{aligned}\frac{d}{dt}\|v\|^2_{\text{inflow},\Sigma} &= \frac{1}{2}\left(\lambda_+^2 C_+^T B C_+ - \lambda_-^2 C_-^T B C_-\right)\\ &+ \left(\lambda_+ C_+ - \lambda_- C_-\right)\left[\Sigma\tau^i P_i - B\delta\Phi\right] + 2(v,F)_{\text{inflow},\Sigma}\end{aligned} \tag{4.133}$$

It is time now to replace the penalty terms by their formulas, (4.127). The quantity $\Sigma\tau^i P_i - B\delta\Phi$ is the sum of the following terms:

$$\begin{aligned}\tau^0\Sigma P_0 + E_0\delta\Phi &= \tau^0\lambda_- E_0\left(C_- - R_0 C_+ - \mathsf{g}_0\right) + \left[\tau^0\left(-1 + \frac{\lambda_-}{\lambda_+}R_0\right) + 1\right]E_0\delta\Phi\\ \tau^1\Sigma P_1 - E_1\delta\Phi &= \tau^1\lambda_- E_1\left(C_+ - R_1 C_- - \mathsf{g}_1\right) + \left[\tau^1\left(-1 + \frac{\lambda_+}{\lambda_-}R_1\right) - 1\right]E_1\delta\Phi\end{aligned} \tag{4.134}$$

Now, notice that choosing $\tau^{0,1}$ as in (4.128) annihilates the coefficients of $E_0\delta\Phi$

and $E_1\delta\Phi$ in the previous relations, and the energy estimate is:

$$\begin{aligned}\frac{d}{dt}\left\|v\right\|^2_{\text{inflow},\Sigma} &= -\frac{1}{2}\left(\lambda_+^2 C_+^T E_0 C_+ - \lambda_-^2 C_-^T E_0 C_-\right)\\ &+\frac{\lambda_-\lambda_+}{\lambda_+ - R_0\lambda_-}\left(\lambda_+ C_+ - \lambda_- C_-\right)^T E_0\left(C_- - R_0 C_+ - \mathsf{g}_0\right)\\ &-\frac{1}{2}\left(\lambda_-^2 C_-^T E_1 C_- - \lambda_+^2 C_+^T E_1 C_+\right)\\ &+\frac{\lambda_+\lambda_-}{\lambda_- - R_1\lambda_+}\left(\lambda_- C_- - \lambda_+ C_+\right)^T E_1\left(C_+ - R_1 C_- - \mathsf{g}_1\right)\\ &+2(v,F)_{\text{inflow},\Sigma}\end{aligned} \tag{4.135}$$

Denote $V \equiv \lambda_+ C_+ - \lambda_- C_-$. Then (4.135) becomes:

$$\begin{aligned}\frac{d}{dt}\left\|v\right\|^2_{\text{inflow},\Sigma} &= -\frac{(\lambda_+ + \lambda_- R_0)^2}{2\left(\lambda_+^2 - \lambda_-^2 R_0^2\right)}\left[V_0 + \frac{\lambda_-\lambda_+}{\lambda_+ + \lambda_- R_0}\mathsf{g}_0(t)\right]^2 + \frac{\lambda_-^2\lambda_+^2}{2\left(\lambda_+^2 - \lambda_-^2 R_0^2\right)}(\mathsf{g}_0(t))^2\\ &-\frac{(\lambda_- + \lambda_+ R_1)^2}{2\left(\lambda_-^2 - \lambda_+^2 R_1^2\right)}\left[V_1 + \frac{\lambda_+\lambda_-}{\lambda_- + \lambda_+ R_1}\mathsf{g}_1(t)\right]^2 + \frac{\lambda_+^2\lambda_-^2}{\lambda_-^2 - \lambda_+^2 R_1^2}(\mathsf{g}_1(t))^2\\ &+2(v,F)_{\text{inflow},\Sigma}\end{aligned} \tag{4.136}$$

By comparing the discrete estimate (4.136) with the continuum estimate (4.117) note that there is a one to one correspondence. Like in the continuum case, if (4.113) holds then

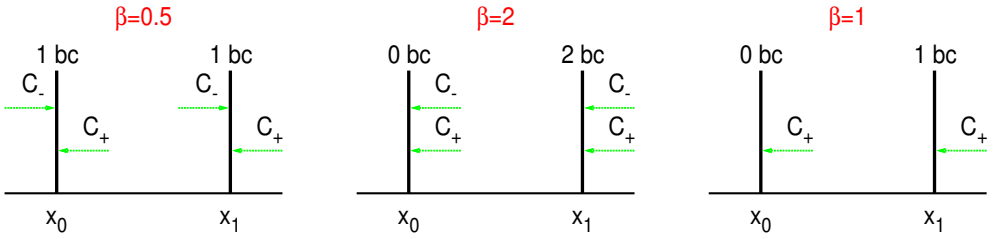
$$\frac{d}{dt}\left\|v\right\|^2_{\text{inflow},\Sigma} \leq \frac{\lambda_-^2\lambda_+^2}{\lambda_+^2 - \lambda_-^2 R_0^2}(\mathsf{g}_0(t))^2 + \frac{\lambda_+^2\lambda_-^2}{\lambda_-^2 - \lambda_+^2 R_1^2}(\mathsf{g}_1(t))^2 + \left\|v\right\|^2_{\text{inflow},\Sigma} + \left\|F\right\|^2_{\text{inflow},\Sigma} \tag{4.137}$$

The above relation tells that the scheme is strongly well-posed and strictly stable in respect to the energy (4.124).

Remark The estimate (4.137) is not optimal and as consequence, it is not sufficient for studying the convergence of the numerical scheme. Optimal energy estimates can be obtained by considering the Laplace transformed problem. However the analysis ends here and the convergence will be investigated numerically in the following section.

4.5 Numerical Tests

This section presents the results of some numerical experiments regarding the implementation of various IBVPs for the wave equation using the two discretization methods discussed in the previous sections: the ghost-point and SBP-SAT procedures. Figure 4.1 illustrates the IBVPs under consideration: (a) two inflow boundaries if $|\beta| < 1$, (b) one outflow and one completely inflow if $|\beta| > 1$, (c) one outflow and one inflow if $|\beta| = 1$. All three cases will be implemented using the ghost-point method. The case (a) will be also analyzed using the SBP-SAT method.



(a) Inflow-Inflow (b) Outflow-Completely Inflow (c) Outflow-Inflow

Figure 4.1: The initial boundary value problems under numerical investigation

4.5.1 General Setup

In order to test the validity of a specific numerical algorithm, three types of tests are performed:

- **stability test:** Random initial and boundary data (noise), with amplitude $a = 10^{-10}$ are evolved at two different resolutions. [7] The $D_{\pm}$ norm of the state vector $v = (\Phi, K)$ defined in (2.95) is computed and checked for scaling with the resolution.

[7] The amplitude of the noise is not important because the system is linear.

- run setup

  The two resolutions are $h = 0.02/r$, where $r \in \{1, 2\}$. The grid has $N + 1$ points where $N = 50r$. The length of the grid is $l = Nh = 1$. The Courant factor is fixed to $\lambda = 0.5$, apart from the case of the outflow-completely inflow algorithm, where it is lowered to $\lambda = 0.25$. For a $2n$-order spatial discretization, $2n + 2$-order Kreiss-Oliger dissipation, (2.32), is added only if necessary (the scheme is unstable otherwise). In this case, the corresponding dissipation coefficient σ will be specified. The evolution time is 1000CT (1CT=l=1). The time integration is done using either the standard 4th order Runge-Kutta method (RK4) or the 8th order Runge-Kutta method from [79] (RK8).

- **accuracy test:** Analytic initial and boundary data are evolved. For the variable K, the initial data is the zero function, while for Φ, it is a modified Gaussian as defined in (3.79) with $A_1 = l/(2\pi)$ and $A_2 = 1$.

  The boundary data is constructed from the analytical solution of the Cauchy problem (3.75). So the solutions of the IVP and IBVP are the same. The test evaluates the $D_\pm$ norm of the error and the convergence factor, both in respect to the analytic solution. The accuracy test uses the same run setup as the stability test.

- **Courant-stability/accuracy test:** The same analytic initial and boundary data as for the accuracy test are evolved. The resolution is kept fixed at $h = 0.02$ and varied is only the Courant factor. The goal of this test is to check if by decreasing the Courant factor, the scheme remains stable (local stability) but also how the accuracy and the convergence factor are affected by this decrease. For this purpose, the $D_\pm$ norm of the error and the convergence factor are computed at a certain time, ($t = 10$CT) in respect to the analytic solution.

In all the cases the maximally dissipative boundary conditions are of Sommerfeld type (reflection coefficients $R_0 = R_1 = 0$).

4.5.2 Inflow-Inflow GP algorithm($|\beta| < 1$)

In this test the shift is fixed to $\beta = 0.5$.
Algorithm: inner points are computed using (4.24), both boundaries are treated according to the prescription (4.27)
Orders tested: $2n = 2, 4, 6, 8$
Courant limit: $\lambda = 0.5$
Dissipation: added in the case of $2n = 8 \rightarrow \sigma = 0.05$;
The results of the stability and accuracy tests are presented in the figures 4.2.

The stability test indicates that the discrete system loses energy (damps the high frequencies of the noise) until it reaches a certain saturation value beyond which it just fluctuates with small amplitudes around this value. At fixed order of approximation, the saturation value for RK8 is higher than for RK4. Notice that in general, high orders spatial approximations lose energy faster than lower orders, but they settle down to a higher saturation value.

Regarding the accuracy test, the results depend also on the time integrator used. With RK4 the 4th order improves significantly over the 2nd order (the error at low resolution ($r = 1$) is $\sim 10^2$ times smaller) and the 6th order improves over the 4th order (the error at low resolution ($r = 1$) is ~ 10 times smaller), while the improvement of the 8th order over the 6th is very small. (the error is only ~ 1.1 times smaller).
With RK8, the 2nd and 4th order accurate schemes have practically the same error as with RK4, however the 6th and the 8th order become clearly differentiated (by a factor of ~ 13).

Table 4.2 shows the results of the Courant stability/accuracy test. The same general pattern is encountered as in the case with periodic boundary conditions analyzed in 3.6.2. The accuracy and the convergence factor of the 2nd and 4th orders vary little with the Courant factor while the the 6th and

λ	n=1	n=2	n=3	n=4 ($\sigma = 0.05$)
0.5	2.394144e-3	3.86393e-5	4.71159e-6	4.25473e-6
0.25	2.392045e-3	3.66228e-5	1.16704e-6	2.61326e-7
0.125	2.391910e-3	3.65077e-5	1.03374e-6	6.39926e-8
0.0625	2.391902e-3	3.65007e-5	1.02668e-6	6.23034e-8

(a) Total Error

λ	n=1	n=2	n=3	n=4 ($\sigma = 0.05$)
0.5	1.9882	3.9571	4.0666	3.8931
0.25	1.9873	3.9568	5.3655	3.9925
0.125	1.9872	3.9568	5.8616	5.9512
0.0625	1.9872	3.9568	5.8870	7.9096

(b) Convergence Factor

Table 4.2: **Inflow-inflow GP-algorithm:** The $D_{\pm}$ error and the convergence factor after 10CT, for various Courant factors and orders of approximations

the 8th, show drastic improvements at small λ.

4.5.3 Outflow-Completely Inflow GP algorithm($|\beta| > 1$)

In this test the shift is $\beta = 2$.

Algorithm: the inner points are computed using (4.24), the x_0 boundary is treated according to the outflow extrapolation conditions (4.25), with $m = 2n$, and the x_1 boundary according to the completely-inflow prescription (4.28)

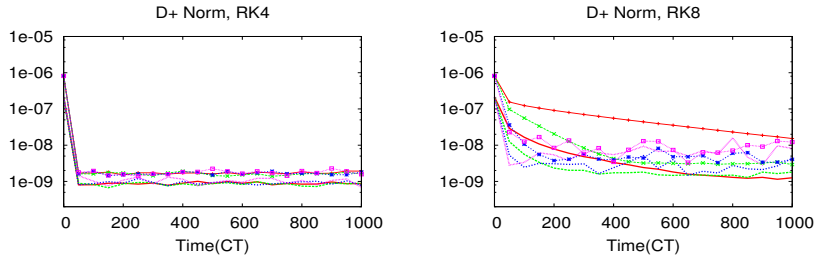
Orders tested: $2n = 2$, 4, 6, 8

Courant limit: $\lambda = 0.25$

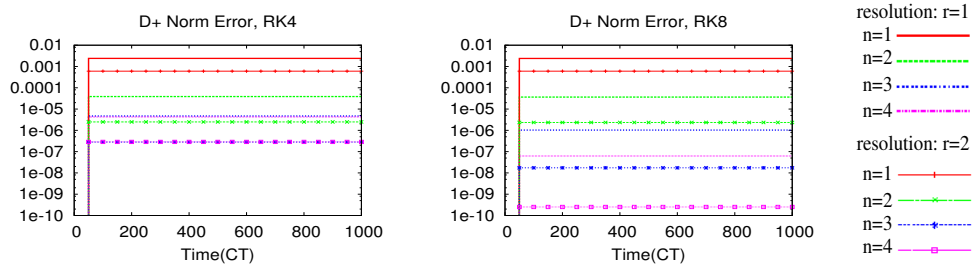
Dissipation: added in the case of $2n = 6 \rightarrow \sigma = 2$ and $2n = 8 \rightarrow \sigma = 8$;

The results of the stability and accuracy tests are presented in the figures 4.3.

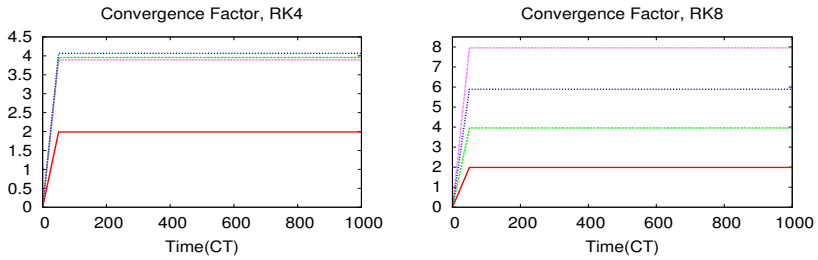
The stability test exhibits the same qualitative behavior as in the case of the inflow-inflow GP-algorithm, described in 4.5.2, with higher orders losing energy faster than lower orders, but they enter the saturation regime at a higher energy value.



(a) Stability Test: $D_{\pm}$ norm of the state vector



(b) Accuracy Test: $D_{\pm}$ norm of error



(c) Accuracy Test: Convergence Factor

Figure 4.2: **Inflow-inflow GP-algorithm,** $\beta = 0.5$**.** The figures show the results of the stability (a) and accuracy (b)–(c) tests, using two types of time-integrators: RK4 (left) and RK8 (right). Details of the runs are given in 4.5.2

λ	n=1	n=2	n=3 ($\sigma = 2$)	n=4 ($\sigma = 8$)
0.5	6.6468755e-3	1.40524e-4	unstable	unstable
0.4	6.6375947e-3	1.16444e-4	2.715e-5	unstable
0.25	6.6316708e-3	1.06148e-4	5.858e-6	3.938e-6
0.125	6.6305799e-3	1.04848e-4	3.058e-6	2.516e-7
0.0625	6.6305078e-3	1.04776e-4	2.948e-6	1.476e-7

(a) Total Error

λ	n=1	n=2	n=3 ($\sigma = 2$)	n=4 ($\sigma = 8$)
0.5	2.0429	4.2088	unstable	unstable
0.4	2.0415	4.1949	4.0441	unstable
0.25	2.0405	4.1597	4.4324	3.9701
0.125	2.0404	4.1519	5.7700	4.0270
0.0625	2.0403	4.1514	5.9342	7.0486

(b) Convergence Factor

Table 4.3: **Outflow-completely inflow GP algorithm ($\beta = 2$):** The $D_{\pm}$ error and the convergence factor after 10CT, for various Courant factor and orders of approximations

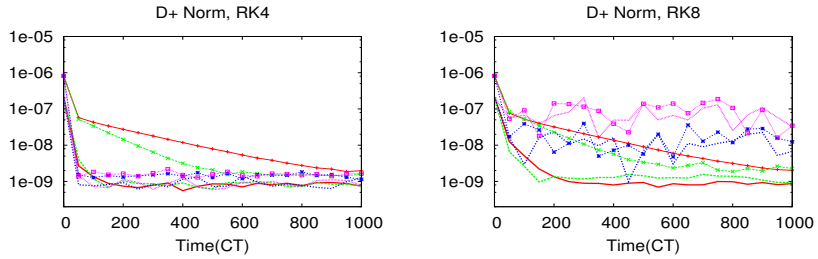
Regarding the accuracy, with RK4, the error of the 4th order scheme is ~ 60 times smaller than the one of the 2nd order scheme and ~ 20 times larger than the error of the 6th order. As in 4.5.2, the 8th order becomes differentiated from the 6th only when using RK8.

Table 4.3 shows the results of the Courant stability/accuracy test. The same general pattern is encountered as in the case with periodic boundary conditions analyzed in 3.6.2. The accuracy and the convergence factor of the 2nd and 4th orders vary little with the Courant factor while the the 6th and the 8th, show drastic improvements at small λ.

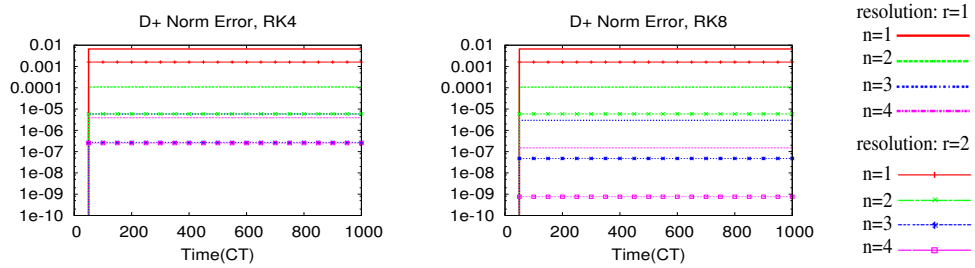
4.5.4 Outflow-Inflow GP algorithm ($\beta = 1$)

The shift is fixed to $\beta = 1$.

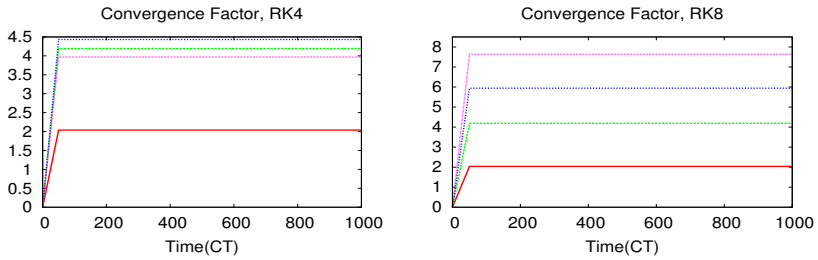
Algorithm: inner points are computed using (4.24), the x_0 boundary is treated according to the outflow extrapolation conditions (4.25), with $m = 2n$,



(a) Stability Test: $D_{\pm}$ norm of the state vector



(b) Accuracy Test: $D_{\pm}$ norm of error



(c) Accuracy Test: Convergence Factor

Figure 4.3: **Outflow-Completely Inflow GP algorithm,** $\beta = 2$**.** The figures show the results of the stability (a) and accuracy (b)–(c) tests, using two types of time-integrators: RK4 (left) and RK8 (right). Details of the runs are given in 4.5.3

and the x_1 boundary according to the inflow prescription (4.27)
Orders tested: $2n = 2$, 4, 6, 8
Courant limit: $\lambda = 0.5$
Dissipation: added in the case of $2n = 8 \rightarrow \sigma = 0.1$;

The results of the stability and accuracy tests are presented in the figures 4.4.

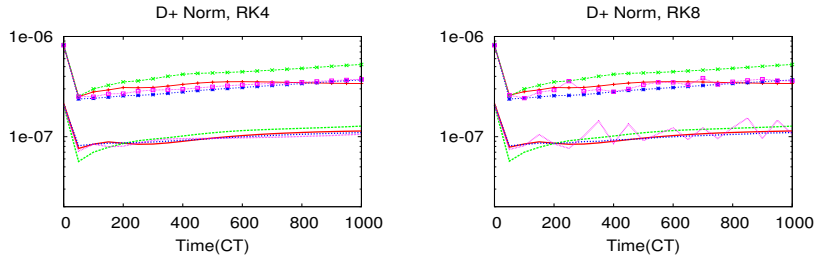
The stability tests show that, for all the orders, and independent of the time integrator, the energy of the noise decreases in the first few crossing times and afterward exhibits an almost linear growth which is practically independent of the grid spacing. This is to be contrasted with the other two algorithms presented in 4.5.2 and 4.5.3 where the energy was settling down at a saturation value. Notice also that for both RK4 and RK8, the energy manifests practically the same quantitative time behavior at a fixed order of spatial approximation (except the case of $2n = 8$).

Regarding the accuracy, for all the orders, the error grows linearly with time. For both RK4 and RK8, the convergence is lost after few crossing times when $2n = 2$ and decreases slowly in time for $2n = 4$ and $2n = 6$. For $2n = 8$ with RK4, the convergence increases in time from ~ 4 initially, to ~ 6.5 at the end of the evolution. With RK8, the case $2n = 8$ has practically constant convergence factor over time (8).

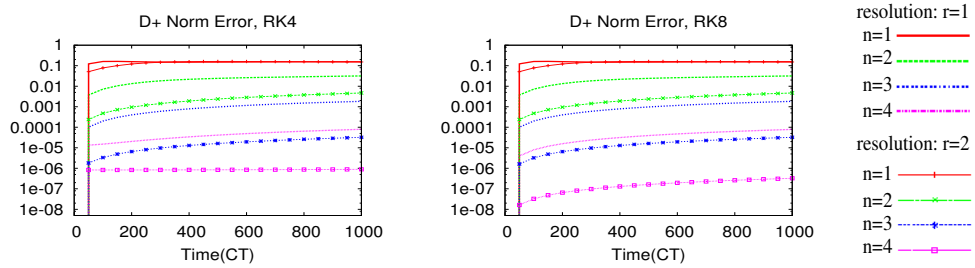
Table 4.4 shows the results of the Courant stability/accuracy test. Again, the accuracy and the convergence factor of the 2nd and 4th orders vary little with the Courant factor, while the the 6th and the 8th show improvement at small λ.

4.5.5 Inflow-Inflow SBP-SAT algorithm($|\beta| < 1$)

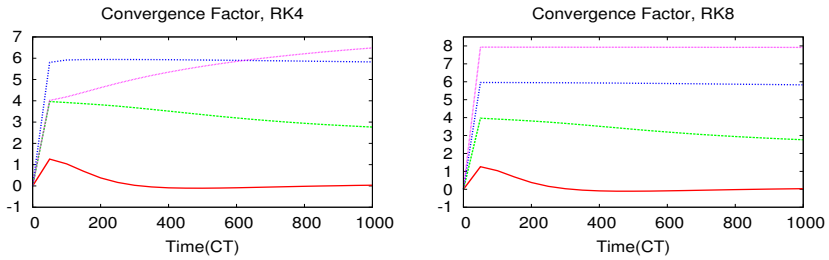
The shift is fixed to $\beta = 0.5$.
Algorithm: SBP-SAT method, (4.119)-(4.127)
Orders tested: the 2nd, 4th and 6th order SBP operators given in [57] (labeled as $n = 1, 2, 3$) and the 6th order SBP operator given in the appendix



(a) Stability Test: $D_{\pm}$ norm of the state vector



(b) Accuracy Test: $D_{\pm}$ norm of error



(c) Accuracy Test: Convergence Factor

Figure 4.4: **Outflow-Inflow GP algorithm,** $\beta = 1$**.** The figures show the results of the stability (a) and accuracy (b)–(c) tests, using two types of time-integrators: RK4 (left) and RK8 (right). Details of the runs are given in 4.5.4

λ	n=1	n=2	n=3	n=4 ($\sigma = 0.1$)
0.5	4.29536182e-2	7.586432e-4	2.402065e-5	1.2817e-5
0.25	4.29532059e-2	7.582713e-4	2.019353e-5	1.1296e-6
0.125	4.29531797e-2	7.582529e-4	2.016324e-5	7.9683e-7
0.0625	4.29531781e-2	7.582516e-4	2.016220e-5	7.9477e-7

(a) Total Error

λ	n=1	n=2	n=3	n=4 ($\sigma = 0.1$)
0.5	1.8899	3.9812	4.7633	3.9545
0.25	1.8899	3.9813	5.9392	4.4834
0.125	1.8899	3.9813	5.9580	7.457
0.0625	1.8899	3.9813	5.9582	7.9303

(b) Convergence Factor

Table 4.4: **Outflow-inflow GP-algorithm ($\beta = 1$):** The $D_{\pm}$ error and the convergence factor after 10CT, for various Courant limits and orders of approximations

D (labeled as $n = 3^*$)

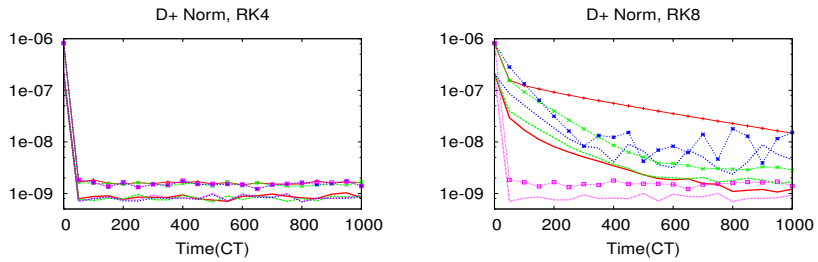
Courant limit: $\lambda = 0.5$

Dissipation: no dissipation.

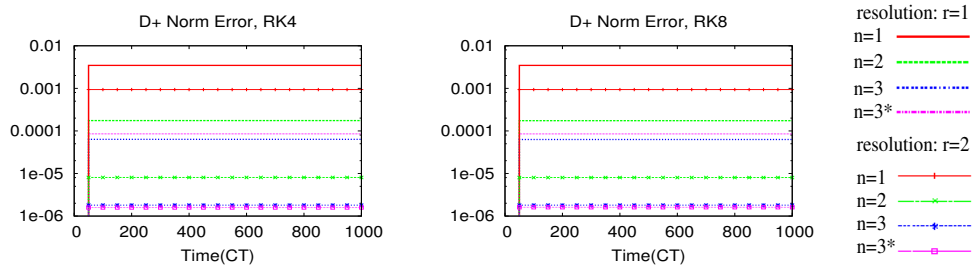
The results of the stability and accuracy tests are presented in fig. 4.5.

The stability test shows that the scheme using the new 6th order SBP operator has a much cleaner energy behavior (the noise is damped much faster and to lower saturation value) in comparison with all the other tested SBP-schemes.

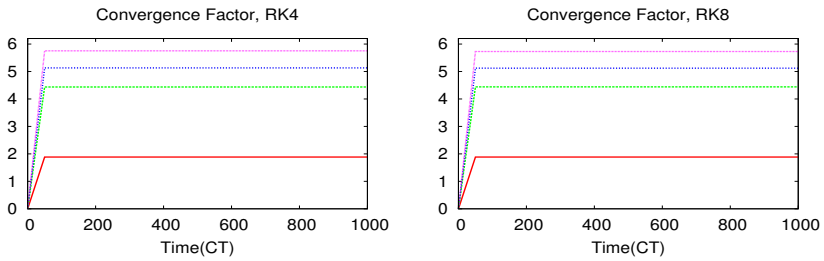
As in the case of inflow-inflow GP algorithm, the RK8 does not show a significant accuracy benefit over RK4 when the order of spacial discretization is at most 6. Notice that $n = 3$ gives less error than $n = 3^*$. However the convergence rate for $n = 3$ is only 5 while for $n = 3^*$ is 6, such that at the next resolution (with $r = 2$), the case $n = 3^*$ is more accurate than $n = 3$.



(a) Stability Test: $D_{\pm}$ norm of the state vector



(b) Accuracy Test: $D_{\pm}$ norm of error



(c) Accuracy Test: Convergence Factor

Figure 4.5: **Stability Test for the inflow-inflow SBP-SAT algorithm** The figures show the results of the stability (a) and accuracy (b)–(c) tests, using two types of time-integrators: RK4 (left) and RK8 (right). $n = 1, 2, 3$ stand for 2nd, 4th and 6th order SBP operators given in [57], $n = 3^*$ is a new 6th SBP operator that verifies the condition (4.122); details of the runs in 4.5.5

λ	n=1	n=2	n=3	n=3*
0.5	3.462716e-3	1.748814e-4	6.33678e-5	8.542455e-5
0.25	3.461758e-3	1.743466e-4	6.28864e-5	8.526654e-5
0.125	3.461920e-3	1.743138e-4	6.28823e-5	8.526288e-5
0.0625	3.461923e-3	1.743118e-4	6.28816e-5	8.526271e-5

(a) Total Error

λ	n=1	n=2	n=3	n=3*
0.5	1.8858	4.4352	5.1312	5.7523
0.25	1.8855	4.4390	5.1199	5.7298
0.125	1.8858	4.4392	5.1304	5.7253
0.0625	1.8858	4.4392	5.1293	5.7250

(b) Convergence Factor

Table 4.5: **Inflow-inflow SBP-SAT-algorithm:** The $D_{\pm}$ error and the convergence factor after 10CT, for various Courant factors and orders of approximations

4.5.6 Disscussion

In 4.5.2-4.5.5 all three possible strip-IBVPs of the wave equation have been investigated numerically. Two of them (outflow-completely inflow and outflow-inflow) have been implemented using only the ghost-point method while the inflow-inflow case, also using the SBP-SAT algorithm. The numerical results suggest that, regardless of the implementation method, the inflow-inflow and outflow-completely inflow cases have similar qualitative behaviour of the error and convergence rate. The energy in the grid defined by the D_+ norm of the state vector, remains practically constant during the time evolution. Constant remain also the total error and the convergence factor. Notice that each of them required in total two boundary conditions (see also fig.4.1).

In contrast, the outflow-inflow case, which only required (in total) one boundary condition, exhibits a linear growth in the total energy and in the total error, while ideally it should also stay constant. Also the time behaviour of the convergence factor is not flat anymore but it depends on the order of approximation used, higher orders giving better results. This case is similar

to the numerical codes using the standard-Cauchy approach (e.g. BSSN) that use an excision boundary (inside the black hole horizon, so outflow) and an inflow (but not completely-inflow!) outer boundary.

I also want to point out that designing stable algorithms for all the possible IBVPs for the wave equation is a necessary step towards modelling the interface boundaries in a grid with multiples domains (i.e. multipatch) for more complicated second order systems.

Chapter 5

BSSN System in Spherical Symmetry

5.1 Introduction

The scalar wave equation is a simple but powerful model for numerical relativity. It has been analyzed in detail in the previous chapters , however, there is still a big gap to be filled between this simple model and the second order formulations of Einstein's equations such as ADM, BSSN or NOR. Difficulties arise at both continuum and the discrete level. They come not only from the nonlinearity, the complicated source and lower order terms, and the existence of constraints or nontrivial boundary conditions, but from the very structure of the principal part of the Einstein equations in these particular formulations.

For the BSSN system, the main achievements at the continuum level were to construct gauges (Bona-Masso type-gauge for lapse in [5] and densitized lapse in [31], and shift prescribed analytically in both cases) such that the resulting PDE system is symmetric hyperbolic. That means, maximally dissipative boundary conditions can be imposed in order to achieve well-posedness. The issue of constraints preservation has been raised in [31] and constraint preserving boundary conditions have been designed but it is not clear if they lead

to a well-posed problem.

In [5] a live-gauge often used in NR ("hyperbolic K-driver" for lapse "hyperbolic Gamma-driver") has been also analyzed but only strong hyperbolicity has been proved, a necessary but not sufficient condition for the system to be well-posed with MDBC.

Concerning the discrete level very little is known, already leaving to the side the issue of constraints. Because the BSSN system matches the form (1.4), the discretization of its IVP can be investigated using the method from [14], generalized in Chapter 2 to accommodate $2n$-order accuracy and off-centered first order derivatives. Regarding the IBVP of BSSN, to this date it is not known how to discretize it in a stable way, because in general, there is no prescription of how to treat the boundaries for the systems (1.4).

At the moment an ad-hoc "radiation boundary condition" is used to update the boundary points for each variable f of the BSSN system. This condition assumes that at large distances from the source, each variable f is a superposition of the type:

$$f = f_0 + \frac{u(r - vt)}{r} + \frac{h(r + vt)}{r} \tag{5.1}$$

where f_0 is the correct asymptotic value and the last two terms stand for outgoing and respectively for ingoing spherical waves propagating with the speed v. Taking the time derivative in (5.1) and projecting onto a boundary with the normal x^i, gives:

$$\frac{\partial f}{\partial t} = -\frac{vr}{x_i}\frac{\partial f}{\partial x^i} - \frac{v}{r}(f - f_0) + H\frac{v}{r} \tag{5.2}$$

where $H = 2dh(s)/ds$. The function H is assumed to be of the form $H = \frac{\text{const}}{r^m}$. The exponent m is called the radiation power and it is adjusted ad-hoc (usually it has the value 2).

The boundary condition (5.2) is easy to apply, by discretizing the first order derivative with one-sided stencils (2nd order accurate) and then passing to the

time integrator as an evolution equation for the variable f at the boundary.

Although it is very likely to give an ill-posed problem (because the boundary condition is applied to *each* variable) and consequently to lead to an unstable scheme, this condition has been extensively used in long term simulations [64] where the boundaries have been pushed to large distances.

In this chapter, a simple one-dimensional model of the BSSN system is constructed assuming spherical symmetry and several possible associated IBVPs are considered. The system is discretized using $2n$-accurate centered FDOs ($n = \overline{1,4}$) and the boundaries are implemented using the ghost-point method. The numerical treatment of the timelike boundary is different in the case when the shift is zero at the boundary from the case when there is shift.

The outline of the chapter is:

- deduce equations in spherical symmetry (ADM then BSSN) and decide which variables are relevant in spherical symmetry
- setup the characteristics and the corresponding speeds (show strong hyperbolicity)
- provide boundary prescriptions for different types of boundary conditions (outflow, timelike with and without shift).
- present numerical tests for the case when the principal part of evolution equations is frozen to analytical quantities (functions of spacetime) and the lower order terms are given analytically (in other words, the tested system is linear with non-constant coefficients and has source terms).
- show some numerical tests for the fully nonlinear case

5.2 Deducing the Equations

The starting point for deducing the BSSN equations in spherical symmetry, is the spacetime element[1]:

$$ds^2 = \left(-\alpha^2 + \mathsf{a}\beta^2\right) dt^2 + 2\beta \mathsf{a} dt dr + \mathsf{a} dr^2 + \mathsf{b}\left(d\theta^2 + \sin^2 d\varphi^2\right) \tag{5.3}$$

where α, β, a and b are functions of the coordinates t and r.

5.2.1 ADM in Spherical Symmetry

The fully 3-D ADM system is presented in Appendix B. Assuming spherical symmetry, the ADM quantities — physical metric (γ_{ij}) and extrinsic curvature (K_{ij}) — are of the following form:

$$\gamma_{ij} = \begin{pmatrix} \mathsf{a} & 0 & 0 \\ 0 & \mathsf{b} & 0 \\ 0 & 0 & \mathsf{b}\sin^2\theta \end{pmatrix}, \quad K_{ij} = \begin{pmatrix} \mathsf{X} & 0 & 0 \\ 0 & \mathsf{Y} & 0 \\ 0 & 0 & \mathsf{Y}\sin^2\theta \end{pmatrix} \tag{5.4}$$

where X and Y are functions of t and r. With the notation $\partial_0 \equiv \partial_t - \beta\partial_r$, the ADM equations in spherical symmetry are:

$$\begin{aligned}
\partial_0 \mathsf{a} &= -2\alpha\mathsf{X} + 2\mathsf{a}\beta_{,r} \\
\partial_0 \mathsf{b} &= -2\alpha\mathsf{Y} \\
\partial_0 \mathsf{X} &= -\alpha_{,rr} - \frac{\alpha \mathsf{b}_{,rr}}{\mathsf{b}} - \frac{\alpha\mathsf{X}^2}{\mathsf{a}} + \frac{\alpha \mathsf{b}_{,r}{}^2}{2\mathsf{b}^2} \\
&+ \mathsf{a}_{,r}\left(\frac{\alpha_{,r}}{2\mathsf{a}} + \frac{\alpha\mathsf{b}_{,r}}{2\mathsf{ab}}\right) + \mathsf{X}\left(\frac{2\alpha\mathsf{Y}}{\mathsf{b}} + 2\beta_{,r}\right) \\
\partial_0 \mathsf{Y} &= -\alpha\frac{\mathsf{b}_{,rr}}{2\mathsf{a}} + \frac{\alpha\mathsf{XY}}{\mathsf{a}} + \frac{\alpha\mathsf{a}_{,r}\mathsf{b}_{,r}}{4\mathsf{a}^2} - \frac{\alpha_{,r}\mathsf{b}_{,r}}{2\mathsf{a}} + \alpha
\end{aligned} \tag{5.5}$$

[1] the most general spherically symmetric spacetime

The evolution equations for the nondiagonal terms of the metric and extrinsic curvature are automatically zero. The energy-momentum constraints are:

$$\begin{aligned} \mathcal{H} &\equiv \frac{4\,\mathsf{X}\,\mathsf{Y}}{\mathsf{a}\,\mathsf{b}} + \frac{2\,\mathsf{Y}^2}{\mathsf{b}^2} + \frac{\mathsf{a}_{,r}\,\mathsf{b}_{,r}}{\mathsf{a}^2\,\mathsf{b}} + \frac{{\mathsf{b}_{,r}}^2}{2\,\mathsf{a}\,\mathsf{b}^2} + \frac{2\,(\mathsf{a}-\mathsf{b}_{,rr})}{\mathsf{a}\,\mathsf{b}} = 0 \\ \mathcal{M} &\equiv \frac{\mathsf{X}\,\mathsf{b}_{,r}}{\mathsf{a}^2\,\mathsf{b}} + \frac{\mathsf{Y}\,\mathsf{b}_{,r}}{\mathsf{a}\,\mathsf{b}^2} - \frac{2\,\mathsf{Y}_{,r}}{\mathsf{a}\,\mathsf{b}} = 0 \end{aligned} \tag{5.6}$$

5.2.2 BSSN in Spherical Symmetry

For the BSSN equations there is one technical obstacle when trying to reduce them using spherical symmetry. This comes from the fact that they assume that the determinant of the conformal metric is one. This condition is not generally covariant, e.g, the conformal metric corresponding to a flat physical metric is not necessarily flat. In practice this might be a problem if one wants to evolve flat spacetimes. In [9] this difficulty has been overcome by modifying the BSSN system to allow for a conformal metric with non-unit determinant.

This is not the approach followed here. In this thesis the 1-D model is constructed from the standard equations (so the determinant of the conformal metric remains one) assuming that the BSSN variables correspond to a spherically symmetric spacetime. The advantage of this approach is that it leads to a system with the same structure as the original BSSN system — similar principal part and no additional evolution equations [2]. The disadvantage is, as mentioned above, the difficulty to evolve flat spacetimes.

The 3-D BSSN system is presented in Appendix C. The procedure for deducing the form of the BSSN variables —conformal metric, $\tilde{\gamma}_{ij}$, conformal traceless extrinsic curvature, $\tilde{A}_{ij}$, conformal factor ϕ, trace of extrinsic curvature, K and connection functions, $\tilde{\Gamma}^i$— and their evolution equations in spherical symmetry is similar with the one followed for the ADM system in 5.2.1. Start by assuming a spherical symmetric spacetime, (5.3).

[2] the general BSSN system from [9] has an additional evolution equation for the determinant of the conformal metric

Then the trace K of K_{ij} is

$$\mathsf{K} = \frac{\mathsf{X}}{\mathsf{a}} + 2\frac{\mathsf{Y}}{\mathsf{b}} \tag{5.7}$$

which means that K is only a function of t and r.

$$\mathsf{K} = \mathsf{K}(t, r) \tag{5.8}$$

In spherical symmetry the determinant of the physical metric is $\gamma = \mathsf{a}\mathsf{b}^2 \sin^2\theta$. This means that the conformal factor $\phi = \frac{1}{12}\ln\gamma$ can be written as a sum of two terms, one depending only on time and radial coordinate and the other depending only on the angle θ:

$$\phi = \Phi(t, r) + \Theta(\theta) \tag{5.9}$$

with

$$\Phi = \frac{1}{12}\log(\mathsf{a}\,\mathsf{b}^2) \text{ and } \Theta(\theta) = \frac{1}{6}\log(\sin\theta).$$

It follows that the conformal metric $\tilde{\gamma}_{ij}$ and the conformal extrinsic curvature $\tilde{A}_{ij}$ have the form:

$$\tilde{\gamma}_{ij} = e^{-4\Theta(\theta)} \begin{pmatrix} \tilde{\mathsf{a}} & 0 & 0 \\ 0 & \tilde{\mathsf{b}} & 0 \\ 0 & 0 & \tilde{\mathsf{b}}\sin^2\theta \end{pmatrix} \quad \text{and} \quad \tilde{A}_{ij} = e^{-4\Theta(\theta)} \begin{pmatrix} \tilde{\mathsf{X}} & 0 & 0 \\ 0 & \tilde{\mathsf{Y}} & 0 \\ 0 & 0 & \tilde{\mathsf{Y}}\sin^2\theta \end{pmatrix} \tag{5.10}$$

with $\tilde{\mathsf{a}}$, $\tilde{\mathsf{b}}$, $\tilde{\mathsf{X}}$ and $\tilde{\mathsf{Y}}$ functions of (t, r):

$$\begin{aligned} \tilde{\mathsf{a}} &= e^{-4\Phi}\mathsf{a} \\ \tilde{\mathsf{b}} &= e^{-4\Phi}\mathsf{b} \\ \tilde{\mathsf{X}} &= e^{-4\Phi}\left(\mathsf{X} - \frac{1}{3}\frac{\mathsf{K}}{\mathsf{a}}\right) \\ \tilde{\mathsf{Y}} &= e^{-4\Phi}\left(\mathsf{Y} - \frac{1}{3}\frac{\mathsf{K}}{\mathsf{b}}\right) \end{aligned}$$

The algebraic constraints are:

$$\det\tilde{\gamma}_{ij} = 1 \quad \Leftrightarrow \quad \tilde{\mathsf{a}}\tilde{\mathsf{b}}^2 = 1 \tag{5.11}$$

$$\mathrm{Tr}\tilde{A}_{ij} = 0 \quad \Leftrightarrow \quad \frac{\tilde{\mathsf{X}}}{\tilde{\mathsf{a}}} + \frac{2\tilde{\mathsf{Y}}}{\tilde{\mathsf{b}}} = 0 \tag{5.12}$$

The conformal connection functions are

$$\{\tilde{\Gamma}^i\} = \begin{pmatrix} \tilde{\Gamma}^r & \tilde{\Gamma}^\theta & \tilde{\Gamma}^\varphi \end{pmatrix} \tag{5.13}$$

where $\tilde{\Gamma}^r$, $\tilde{\Gamma}^\theta$ and $\tilde{\Gamma}^\phi$ are functions of (t, r, θ).
Using the definition $\tilde{\Gamma}^i = \tilde{\gamma}^{jk}\tilde{\Gamma}^i_{jk}$, one can show that the $\tilde{\mathsf{G}}$ constraints are:

$$\begin{aligned} \tilde{\Gamma}^r - \sin^{2/3}\theta \frac{\tilde{\mathsf{b}}\tilde{\mathsf{a}}_{,r} - 2\tilde{\mathsf{a}}\tilde{\mathsf{b}}_{,r}}{2\tilde{\mathsf{a}}^2\tilde{\mathsf{b}}} &= 0 \\ \tilde{\Gamma}^\theta + \frac{2}{3\tilde{\mathsf{b}}}\frac{\cos\theta}{\sin^{1/3}\theta} &= 0 \\ \tilde{\Gamma}^\varphi &= 0 \end{aligned} \tag{5.14}$$

Now the relations (5.8), (5.9), (5.10), and (5.13) are inserted in the BSSN equations (C.3)–(C.7).

The evolution equations for the **conformal metric** give:

$$\begin{aligned} \partial_0\tilde{\mathsf{a}} &= -2\,\alpha\,\tilde{\mathsf{X}} + \frac{4\,\tilde{\mathsf{a}}\,\beta'}{3} \\ \partial_0\tilde{\mathsf{b}} &= -2\,\alpha\,\tilde{\mathsf{Y}} - \frac{2\,\tilde{\mathsf{b}}\,\beta'}{3} \\ \partial_0\tilde{\gamma}_{33} &= (\partial_0\tilde{\gamma}_{22})\sin^2\theta \\ \partial_0\tilde{\gamma}_{ij} &= 0, \quad \forall i \neq j \end{aligned} \tag{5.15}$$

The rhs for all the nondiagonal terms of the metric are identically zero.

For the **conformal factor** the evolution equation is:

$$\partial_0\Phi = -\frac{\alpha}{6}\mathsf{K} + \frac{\beta_{,r}}{6} \tag{5.16}$$

while for the **trace K**:

$$\partial_0 \mathsf{K} = -e^{-4\,\Phi}\frac{\alpha_{,rr}}{\tilde{\mathsf{a}}} + \alpha\left(\frac{\mathsf{K}^2}{3} + \frac{\tilde{\mathsf{X}}^2}{\tilde{\mathsf{a}}^2} + \frac{2\,\tilde{\mathsf{Y}}^2}{\tilde{\mathsf{b}}^2}\right) + e^{-4\,\Phi}\alpha_{,r}\left(\csc^{\frac{2}{3}}\theta\,\tilde{\Gamma}^r - \frac{2}{\tilde{\mathsf{a}}}\Phi_{,r}\right) \tag{5.17}$$

The **conformal trace-free extrinsic curvature** components, $\tilde{\mathsf{X}}$ and $\tilde{\mathsf{Y}}$, evolve according to:

$$\begin{aligned}
\partial_0\tilde{\mathsf{X}} &= -\frac{\alpha}{3}e^{-4\,\Phi}\left(\frac{\tilde{\mathsf{a}}_{,rr}}{\tilde{\mathsf{a}}} - \frac{\tilde{\mathsf{b}}_{,rr}}{\tilde{\mathsf{b}}} + 2\,\frac{\alpha_{,rr}}{\alpha} + 4\Phi_{,rr}\right) \\
&+ \alpha\left(-\frac{2}{\tilde{\mathsf{a}}}\tilde{\mathsf{X}}^2 + \mathsf{K}\,\tilde{\mathsf{X}}\right) + \frac{4}{3}\tilde{\mathsf{X}}\beta_{,r} \\
&+ \frac{e^{-4\,\Phi}}{3}\left[3\alpha\left(\frac{\tilde{\mathsf{a}}_{,r}^2}{2\tilde{\mathsf{a}}^2} - \frac{2\tilde{\mathsf{b}}_{,r}^2}{3\tilde{\mathsf{b}}^2}\right) + \left(\frac{\tilde{\mathsf{a}}_{,r}}{\tilde{\mathsf{a}}} + \frac{\tilde{\mathsf{b}}_{,r}}{\tilde{\mathsf{b}}}\right)(\alpha_{,r} + 2\alpha\Phi_{,r}) + 8\left(\alpha_{,r}\Phi_{,r} + \alpha{\Phi_{,r}}^2\right)\right] \\
&+ \alpha\frac{e^{-4\,\Phi}}{3}\csc^{\frac{2}{3}}\theta\,\tilde{\Gamma}^r\left(\tilde{\mathsf{a}}_{,r} - \frac{\tilde{\mathsf{a}}\tilde{\mathsf{b}}_{,r}}{\tilde{\mathsf{b}}}\right) \\
&- \alpha\frac{e^{-4\,\Phi}}{3}\frac{\tilde{\mathsf{a}}}{\tilde{\mathsf{b}}}\left[\frac{4}{9}\left(2 + \csc^2\theta\right) + \tilde{\mathsf{b}}\csc^{\frac{2}{3}}\theta\left(\cot\theta\,\tilde{\Gamma}^\theta + \tilde{\Gamma}^\theta_{,\theta} - 2\,\tilde{\Gamma}^r_{,r}\right)\right]
\end{aligned} \tag{5.18}$$

$$
\begin{aligned}
\partial_0 \tilde{\mathsf{Y}} &= \frac{\alpha}{3}\frac{\tilde{\mathsf{b}}}{\tilde{\mathsf{a}}} e^{-4\,\Phi}\left(\frac{\tilde{\mathsf{a}}_{,rr}}{2\tilde{\mathsf{a}}} - \frac{\tilde{\mathsf{b}}_{,rr}}{2\tilde{\mathsf{b}}} + \frac{\alpha_{,rr}}{\alpha} + 2\Phi_{,rr}\right) \\
&+ \alpha\tilde{\mathsf{Y}}\left(\mathsf{K} - 2\,\frac{\tilde{\mathsf{Y}}}{\tilde{\mathsf{b}}}\right) - \frac{2}{3}\tilde{\mathsf{Y}}\beta_{,r} - \frac{1}{3}e^{-4\,\Phi}\alpha \\
&+ \frac{\tilde{\mathsf{b}}}{\tilde{\mathsf{a}}} e^{-4\,\Phi}\left\{-\alpha\left(\frac{\tilde{\mathsf{a}}_{,r}^2}{4\tilde{\mathsf{a}}^2} - \frac{\tilde{\mathsf{b}}_{,r}^2}{3\tilde{\mathsf{b}}^2}\right) - \frac{1}{6}\left(\frac{\tilde{\mathsf{a}}_{,r}}{\tilde{\mathsf{a}}} + \frac{\tilde{\mathsf{b}}_{,r}}{\tilde{\mathsf{b}}}\right)(\alpha_{,r} + 2\alpha\Phi_{,r}) - \frac{4}{3}\left(\alpha_{,r}\Phi_{,r} + \alpha(\Phi_{,r})^2\right)\right\} \\
&- \frac{1}{6}e^{-4\,\Phi}\alpha\,\csc^{\frac{2}{3}}\theta\,\tilde{\Gamma}^r\left(\frac{\tilde{\mathsf{b}}}{\tilde{\mathsf{a}}}\tilde{\mathsf{a}}_{,r} - \tilde{\mathsf{b}}_{,r}\right) \\
&- \frac{1}{3}e^{-4\,\Phi}\alpha\left[\frac{2}{9}\left(-8 + 5\csc^2\theta\right) + \tilde{\mathsf{b}}\,\csc^{\frac{2}{3}}\theta\,\left(\cot(\theta)\,\tilde{\Gamma}^\theta - 2\,\tilde{\Gamma}^\theta_{,\theta} + \tilde{\Gamma}^r_{,r}\right)\right]
\end{aligned}
\tag{5.19}
$$

In contrast with the evolution equations for the nondiagnonal terms of the conformal metric which are identically zero, for the curvature, the nondiagonal terms have a nontrivial evolution:

$$
\begin{aligned}
\partial_0 \tilde{A}_{12} &= \frac{\alpha}{6}e^{-4\,\Phi}\sin^{-\frac{4}{3}}\theta\left[-\frac{\tilde{\mathsf{a}}_{,r}}{\tilde{\mathsf{a}}}\cot\theta\,\sin^{\frac{2}{3}}\theta + 3\left(\tilde{\mathsf{a}}\tilde{\Gamma}^r_{,\theta} + \tilde{\mathsf{b}}\tilde{\Gamma}^\theta_{,r}\right)\right] \\
\partial_0 \tilde{A}_{13} &= \frac{\alpha}{2}e^{-4\,\Phi}\,\tilde{\mathsf{b}}\,\tilde{\Gamma}^\varphi_{,r}\sin^{\frac{2}{3}}\theta \\
\partial_0 \tilde{A}_{23} &= \frac{\alpha}{2}e^{-4\,\Phi}\tilde{\mathsf{b}}\,\tilde{\Gamma}^\varphi_{,\theta}\sin^{\frac{2}{3}}\theta
\end{aligned}
\tag{5.20}
$$

Also the spherical symmetry condition $\partial_0\left(\tilde{A}_{33} - \tilde{A}_{22}\sin^2\theta\right) = 0$ is not automatically accomplished because:

$$
\partial_0\left(\tilde{A}_{33} - \tilde{A}_{22}\sin^2\theta\right) = e^{-4\,\Phi}\frac{\alpha}{9}\sin^{-\frac{2}{3}}\theta\left[8 - 2\sin^2\theta + 9\,\tilde{\mathsf{b}}\sin^{\frac{4}{3}}\theta\left(\cot\theta\,\tilde{\Gamma}^\theta - \tilde{\Gamma}^\theta_{,\theta}\right)\right]
\tag{5.21}
$$

For the **connection functions** the evolution equations are:

$$\begin{aligned}
\partial_0\tilde{\Gamma}^r &= \alpha \sin^{\frac{2}{3}}\theta \left(-\frac{4\,m}{3\,\tilde{\mathsf{a}}}\mathsf{K}_{,r} + \frac{2\,(m-1)}{\tilde{\mathsf{a}}^2}\tilde{\mathsf{X}}_{,r}\right) \\
&+ \alpha\frac{\sin^{\frac{2}{3}}\theta}{\tilde{\mathsf{a}}}\left(-2\,m\frac{\tilde{\mathsf{Y}}}{\tilde{\mathsf{b}}}\frac{\tilde{\mathsf{b}}_{,r}}{\tilde{\mathsf{b}}} + \frac{\tilde{\mathsf{X}}}{\tilde{\mathsf{a}}}\left[(4-3\,m)\,\frac{\tilde{\mathsf{a}}_{,r}}{\tilde{\mathsf{a}}} + 12\,m\Phi_{,r}\right]\right) \\
&+ \frac{\sin^{\frac{2}{3}}\theta}{\tilde{\mathsf{a}}}\left(-\frac{2\tilde{\mathsf{X}}}{\tilde{\mathsf{a}}}\alpha_{,r} + \frac{4}{3}\beta_{,rr}\right) - \frac{1}{3}\tilde{\Gamma}^r\beta_{,r} \\
\partial_0\tilde{\Gamma}^\theta &= \frac{2\alpha}{3}\frac{\cos}{\tilde{\mathsf{b}}\sin^{\frac{1}{3}}}\left(m\frac{\tilde{\mathsf{X}}}{\tilde{\mathsf{a}}} + (m-1)\,\frac{2\tilde{\mathsf{Y}}}{\tilde{\mathsf{b}}}\right) + \frac{2}{3}\tilde{\Gamma}^\theta\,\beta_{,r} \\
\partial_0\tilde{\Gamma}^\varphi &= \frac{2}{3}\tilde{\Gamma}^\varphi\,\beta_{,r}
\end{aligned} \tag{5.22}$$

Now the spherical symmetry conditions are imposed by requiring that:

1. the rhs of $\partial_0\mathsf{K}$, $\partial_0\tilde{\mathsf{X}}$, $\partial_0\tilde{\mathsf{Y}}$ do not depend on θ, that is, the following quantities do not have dependence on θ:

$$\partial_0\mathsf{K} \;\rightarrow\; \csc^{\frac{2}{3}}\theta\,\tilde{\Gamma}^r \tag{5.23}$$

$$\partial_0\tilde{\mathsf{X}} \;\rightarrow\; \frac{4}{9}\left(2+\csc^2\theta\right) + \tilde{\mathsf{b}}\csc^{\frac{2}{3}}\theta\left(\cot\theta\,\tilde{\Gamma}^\theta + \tilde{\Gamma}^\theta_{,\theta} - 2\,\tilde{\Gamma}^r_{,r}\right) \tag{5.24}$$

$$\partial_0\tilde{\mathsf{Y}} \;\rightarrow\; \frac{2}{9}\left(-8+5\csc^2\theta\right) + \tilde{\mathsf{b}}\,\csc^{\frac{2}{3}}\theta\,\left(\cot\theta\,\tilde{\Gamma}^\theta - 2\tilde{\Gamma}^\theta_{,\theta} + \tilde{\Gamma}^r_{,r}\right) \tag{5.25}$$

2. the rhs of the nondiagonal terms of the curvature, $\partial_0\tilde{A}_{12}$, $\partial_0\tilde{A}_{13}$, $\partial_0\tilde{A}_{23}$ cancel out, that is:

$$\partial_0\tilde{A}_{13} \;\rightarrow\; \tilde{\Gamma}^\varphi_{,r}\sin^{\frac{2}{3}}\theta = 0 \tag{5.26}$$

$$\partial_0\tilde{A}_{23} \;\rightarrow\; \tilde{\Gamma}^\varphi_{,\theta}\sin^{\frac{2}{3}}\theta = 0 \tag{5.27}$$

$$\partial_0\tilde{A}_{12} \;\rightarrow\; -\frac{\tilde{\mathsf{a}}_{,r}}{\tilde{\mathsf{a}}}\cot\theta\,\sin^{\frac{2}{3}}\theta + 3\left(\tilde{\mathsf{a}}\tilde{\Gamma}^r_{,\theta} + \tilde{\mathsf{b}}\tilde{\Gamma}^\theta_{,r}\right) = 0 \tag{5.28}$$

3. the evolution equations for the variables $\tilde{A}_{22}$ and $\tilde{A}_{33}$ are related via the relation $\partial_0 \left(\tilde{A}_{33} - \tilde{A}_{22} \sin^2 \theta \right) = 0$

$$\partial_0 \left(\tilde{A}_{33} - \tilde{A}_{22} \sin^2 \theta \right) \quad \rightarrow \quad 9\, \tilde{\mathsf{b}}\, \sin^{\frac{4}{3}} \theta \left(\cot \theta\, \tilde{\Gamma}^{\theta} - \tilde{\Gamma}^{\theta}_{,\theta} \right) = -8 + 2 \sin^2 \theta \tag{5.29}$$

It is easy to show that if the $\tilde{\Gamma}$-constraints, (5.14), are maintained during the evolution, then the spherical symmetry conditions 1., 2., 3. are satisfied.
The reciprocal is also true. From the relations (5.26), (5.27), (5.22) one can see that spherical symmetry implies $\tilde{\Gamma}^{\varphi} = 0$. Assume that $\tilde{\Gamma}^{r}$ and $\tilde{\Gamma}^{\theta}$ are of the form

$$\tilde{\Gamma}^{r} = \sin^{\frac{2}{3}} \theta\, \tilde{\mathsf{G}}(t, r) \tag{5.30}$$

$$\tilde{\Gamma}^{\theta} = -\frac{2 \cos \theta}{3 \sin^{\frac{1}{3}} \theta} \tilde{\mathsf{L}}(t, r) \tag{5.31}$$

Then, by (5.29) it is obtained that $\tilde{\mathsf{L}} = \frac{1}{\tilde{\mathsf{b}}}$ and by (5.28), $\tilde{\mathsf{G}} = \frac{\tilde{\mathsf{b}}\tilde{\mathsf{a}}_{,r} - 2\tilde{\mathsf{a}}\tilde{\mathsf{b}}_{,r}}{2\tilde{\mathsf{a}}^2 \tilde{\mathsf{b}}}$.

In the following, the $\tilde{\Gamma}^{\theta}$ and $\tilde{\Gamma}^{\varphi}$ constraints will be imposed while $\tilde{\mathsf{G}}^{r} = \sin^{-\frac{2}{3}} \theta\, \tilde{\Gamma}^{r}$ will stand as an independent variable.

The variables of the BSSN system in spherical symmetry are:

$$\{\tilde{\mathsf{a}}, \tilde{\mathsf{b}}, \Phi, \mathsf{K}, \tilde{\mathsf{X}}, \tilde{\mathsf{Y}}, \tilde{\mathsf{G}}\}$$

and their evolution equations:

$$\partial_0 \tilde{\mathsf{a}} = -2\,\alpha\,\tilde{\mathsf{X}} + \frac{4\,\tilde{\mathsf{a}}\,\beta_{,r}}{3} \tag{5.32}$$

$$\partial_0 \tilde{\mathsf{b}} = -2\,\alpha\,\tilde{\mathsf{Y}} - \frac{2\,\tilde{\mathsf{b}}\,\beta_{,r}}{3} \tag{5.33}$$

$$\partial_0 \Phi = -\frac{\alpha}{6}\mathsf{K} + \frac{\beta_{,r}}{6} \tag{5.34}$$

$$\partial_0 \mathsf{K} = -e^{-4\,\Phi}\frac{\alpha_{,rr}}{\tilde{\mathsf{a}}} + \alpha\left(\frac{\mathsf{K}^2}{3} + \frac{\tilde{\mathsf{X}}^2}{\tilde{\mathsf{a}}^2} + \frac{2\,\tilde{\mathsf{Y}}^2}{\tilde{\mathsf{b}}^2}\right) + e^{-4\,\Phi}\alpha_{,r}\left(\tilde{\mathsf{G}} - \frac{2}{\tilde{\mathsf{a}}}\Phi_{,r}\right) \tag{5.35}$$

$$\begin{aligned}
\partial_0 \tilde{\mathsf{X}} &= -\frac{\alpha}{3}e^{-4\,\Phi}\left(\frac{\tilde{\mathsf{a}}_{,rr}}{\tilde{\mathsf{a}}} - \frac{\tilde{\mathsf{b}}_{,rr}}{\tilde{\mathsf{b}}} + 2\,\frac{\alpha_{,rr}}{\alpha} + 4\Phi_{,rr} - 2\tilde{\mathsf{a}}\tilde{\mathsf{G}}_{,r}\right) \\
&+ \alpha\left(-\frac{2}{\tilde{\mathsf{a}}}\tilde{\mathsf{X}}^2 + \mathsf{K}\,\tilde{\mathsf{X}}\right) + \frac{4}{3}\tilde{\mathsf{X}}\beta_{,r} \\
&+ \frac{e^{-4\,\Phi}}{3}\left[3\alpha\left(\frac{\tilde{\mathsf{a}}_{,r}^2}{2\tilde{\mathsf{a}}^2} - \frac{2\tilde{\mathsf{b}}_{,r}^2}{3\tilde{\mathsf{b}}^2}\right) + \left(\frac{\tilde{\mathsf{a}}_{,r}}{\tilde{\mathsf{a}}} + \frac{\tilde{\mathsf{b}}_{,r}}{\tilde{\mathsf{b}}}\right)(\alpha_{,r} + 2\alpha\Phi_{,r}) + 8\left(\alpha_{,r}\Phi_{,r} + \alpha{\Phi_{,r}}^2\right)\right] \\
&+ \alpha\frac{e^{-4\,\Phi}}{3}\left[\tilde{\mathsf{G}}\left(\tilde{\mathsf{a}}_{,r} - \frac{\tilde{\mathsf{a}}\tilde{\mathsf{b}}_{,r}}{\tilde{\mathsf{b}}}\right) - 2\frac{\tilde{\mathsf{a}}}{\tilde{\mathsf{b}}}\right]
\end{aligned} \tag{5.36}$$

$$\begin{aligned}
\partial_0 \tilde{\mathsf{Y}} &= \frac{\alpha}{3}e^{-4\,\Phi}\left[\frac{\tilde{\mathsf{b}}}{\tilde{\mathsf{a}}}\left(\frac{\tilde{\mathsf{a}}_{,rr}}{2\tilde{\mathsf{a}}} - \frac{\tilde{\mathsf{b}}_{,rr}}{2\tilde{\mathsf{b}}} + \frac{\alpha_{,rr}}{\alpha} + 2\Phi_{,rr}\right) - \tilde{\mathsf{b}}\tilde{\mathsf{G}}_{,r}\right] \\
&+ \alpha\tilde{\mathsf{Y}}\left(\mathsf{K} - 2\,\frac{\tilde{\mathsf{Y}}}{\tilde{\mathsf{b}}}\right) - \frac{2}{3}\tilde{\mathsf{Y}}\beta_{,r} - \frac{1}{3}e^{-4\,\Phi}\alpha \\
&+ \frac{\tilde{\mathsf{b}}}{\tilde{\mathsf{a}}}e^{-4\,\Phi}\left\{-\alpha\left(\frac{\tilde{\mathsf{a}}_{,r}^2}{4\tilde{\mathsf{a}}^2} - \frac{\tilde{\mathsf{b}}_{,r}^2}{3\tilde{\mathsf{b}}^2}\right) - \frac{1}{6}\left(\frac{\tilde{\mathsf{a}}_{,r}}{\tilde{\mathsf{a}}} + \frac{\tilde{\mathsf{b}}_{,r}}{\tilde{\mathsf{b}}}\right)(\alpha_{,r} + 2\alpha\Phi_{,r}) - \frac{4}{3}\left(\alpha_{,r}\Phi_{,r} + \alpha(\Phi_{,r})^2\right)\right\} \\
&- \frac{\alpha}{6}e^{-4\,\Phi}\left[\tilde{\mathsf{G}}\left(\frac{\tilde{\mathsf{b}}}{\tilde{\mathsf{a}}}\tilde{\mathsf{a}}_{,r} - \tilde{\mathsf{b}}_{,r}\right) - 4\right]
\end{aligned} \tag{5.37}$$

$$\begin{aligned}\partial_0\tilde{\mathsf{G}} &= \alpha\left(-\frac{4\,m}{3\,\tilde{\mathsf{a}}}\mathsf{K}_{,r}+\frac{2\,(m-1)}{\tilde{\mathsf{a}}^2}\tilde{\mathsf{X}}_{,r}\right)+\frac{4}{3\tilde{\mathsf{a}}}\beta_{,rr}\\ &+ \alpha\left[-2\,m\frac{\tilde{\mathsf{Y}}}{\tilde{\mathsf{a}}\tilde{\mathsf{b}}}\frac{\tilde{\mathsf{b}}_{,r}}{\tilde{\mathsf{b}}}+\frac{\tilde{\mathsf{X}}}{\tilde{\mathsf{a}}^2}\left((4-3\,m)\,\frac{\tilde{\mathsf{a}}_{,r}}{\tilde{\mathsf{a}}}+12\,m\Phi_{,r}\right)\right]-\frac{2\tilde{\mathsf{X}}}{\tilde{\mathsf{a}}^2}\alpha_{,r}-\frac{1}{3}\tilde{\mathsf{G}}\beta_{,r}\end{aligned} \tag{5.38}$$

5.2.3 Minimal System with Densitized Lapse

Imposing the algebraic constraints (5.12)–(5.12), the variables $\tilde{\mathsf{b}}$ and $\tilde{\mathsf{Y}}$ are eliminated from the system (5.32–5.38). Furthermore the lapse is densitized using $\alpha = e^{6\tau\phi}Q$ where τ is a constant and Q is a given function of the coordinates, t and r.

Evolution Equations

With the notation: $\rho := \frac{1}{\sqrt{a}} = \frac{e^{-2\phi}}{\sqrt{\tilde{\mathsf{a}}}}$ the system of equations is now:

$$\partial_t\tilde{\mathsf{a}} = -2\,\alpha\,\tilde{\mathsf{X}}+\beta\tilde{\mathsf{a}}_{,r}+loA \tag{5.39}$$

$$\partial_t\Phi = -\frac{1}{6}\alpha\,\mathsf{K}+\beta\Phi_{,r}+loP \tag{5.40}$$

$$\partial_t\mathsf{K} = -6\tau\rho^2\alpha\Phi_{,rr}+\beta\mathsf{K}_{,r}+loK \tag{5.41}$$

$$\partial_t\tilde{\mathsf{X}} = \rho^2\alpha\left[\frac{2\,\tilde{\mathsf{a}}^2}{3}\tilde{\mathsf{G}}_{,r}-\frac{1}{2}\tilde{\mathsf{a}}_{,rr}-\frac{4\,(1+3\,\tau)}{3}\tilde{\mathsf{a}}\Phi_{,rr}\right]+\beta\tilde{\mathsf{X}}_{,r}+loX \tag{5.42}$$

$$\partial_t\tilde{\mathsf{G}} = \frac{-4\,\alpha\,m}{3\,\tilde{\mathsf{a}}}\mathsf{K}_{,r}+\frac{2\,\alpha\,(-1+m)}{\tilde{\mathsf{a}}^2}\tilde{\mathsf{X}}_{,r}+\beta\tilde{\mathsf{G}}_{,r}+loG \tag{5.43}$$

where,

$$
\begin{aligned}
loA &= \frac{4\,\tilde{\mathsf{a}}\,\beta_{,r}}{3} \\
loP &= \frac{1}{6}\beta_{,r} \\
loK &= \alpha\left(\frac{1}{3}\,\mathsf{K}^2 + \frac{3}{2\,\tilde{\mathsf{a}}^2}\,\tilde{\mathsf{X}}^2\right) \\
&+ \rho^2\,\alpha\left[-6\,\tau\,\Phi_{,r}\left(-\tilde{\mathsf{a}}\,\tilde{\mathsf{G}} + 2\,(1+3\,\tau)\,\Phi_{,r}\right) + \left(\tilde{\mathsf{a}}\,\tilde{\mathsf{G}} - 2\,(1+6\,\tau)\,\Phi_{,r}\right)\frac{Q_{,r}}{Q} - \frac{Q_{,rr}}{Q}\right] \\
loX &= \alpha\,\rho^2\frac{\tilde{\mathsf{a}}}{6}\left[-4\,\tilde{\mathsf{a}}^{\frac{3}{2}} + \frac{7}{2}\,\frac{(\tilde{\mathsf{a}}_{,r})^2}{\tilde{\mathsf{a}}^2} + \frac{\tilde{\mathsf{a}}_{,r}}{\tilde{\mathsf{a}}}\left(2\,(1+3\,\tau)\,\Phi_{,r} + \frac{Q_{,r}}{Q}\right) + 3\,\tilde{\mathsf{G}}\,\tilde{\mathsf{a}}_{,r}\right. \\
&+ \left. 16\,\left(1+6\,\tau-9\,\tau^2\right)\,{\Phi_{,r}}^2 - 16\,(-1+3\,\tau)\,\Phi_{,r}\,\frac{Q_{,r}}{Q} - 4\,\frac{Q_{,rr}}{Q}\right] + \mathsf{K}\,\tilde{\mathsf{X}} + \frac{4}{3}\,\tilde{\mathsf{X}}\,\beta_{,r} - 2\,\alpha\,\tilde{\mathsf{X}}^2 \\
loG &= \frac{1}{6\,\tilde{\mathsf{a}}^3}\left[-2\,\tilde{\mathsf{a}}^3\,\tilde{\mathsf{G}}\,\beta_{,r} - 3\,\alpha\,\tilde{\mathsf{X}}\left((-8+7\,m)\,\tilde{\mathsf{a}}_{,r} + 24\,(-m+\tau)\,\tilde{\mathsf{a}}\,\Phi_{,r} + 4\,\tilde{\mathsf{a}}\,\frac{Q_{,r}}{Q}\right) + 8\,\tilde{\mathsf{a}}^2\,\beta_{,rr}\right]
\end{aligned}
\tag{5.44}
$$

This will be the system to be implemented numerically and further analyzed in this thesis.

The energy-momentum constraints for the BSSN system, (C.8)–(C.9), become:

$$
\begin{aligned}
\mathcal{H} &\equiv \rho^2\left(\frac{\tilde{\mathsf{a}}_{,rr}}{\tilde{\mathsf{a}}} - 8\,\Phi_{,rr}\right) + 2\rho^2\left[\tilde{\mathsf{a}}^{\frac{3}{2}} + \frac{1}{16}\frac{(\tilde{\mathsf{a}}_{,r})^2}{\tilde{\mathsf{a}}^2} - \left(2\Phi_{,r} - \frac{\tilde{\mathsf{a}}_{,r}}{\tilde{\mathsf{a}}}\right)^2\right] + \frac{1}{6}\left(4\mathsf{K}^2 - 9\frac{\tilde{\mathsf{X}}^2}{\tilde{\mathsf{a}}^2}\right) = 0 \\
\mathcal{M} &\equiv -\frac{2\,\mathsf{K}_{,r}}{3} + \frac{\tilde{\mathsf{X}}_{,r}}{\tilde{\mathsf{a}}} + \frac{\tilde{\mathsf{X}}}{4\tilde{\mathsf{a}}}\left(24\,\Phi_{,r} - 7\,\frac{\tilde{\mathsf{a}}_{,r}}{\tilde{\mathsf{a}}}\right) = 0
\end{aligned}
\tag{5.45}
$$

5.2.4 Analysis of the Principal Part

With $\mathsf{v}_R = (\tilde{\mathsf{a}}_{,r},\ \Phi_{,r},\ \mathsf{K},\ \tilde{\mathsf{X}},\ \tilde{\mathsf{G}})$ the system (5.39)–(5.43) can be written up to the lower order terms as:

$$
\partial_t \mathsf{v}_R \simeq (\alpha A + \beta I)\,\mathsf{v}_{R,r} \tag{5.46}
$$

where

$$A = \begin{pmatrix} 0 & 0 & 0 & -2 & 0 \\ 0 & 0 & -\frac{1}{6} & 0 & 0 \\ 0 & -6\tau\rho^2 & 0 & 0 & 0 \\ -\frac{1}{2}\rho^2 & -\frac{4(1+3\tau)}{3}\tilde{\mathsf{a}}\rho^2 & 0 & 0 & \frac{2\tilde{\mathsf{a}}^2}{3}\rho^2 \\ 0 & 0 & -\frac{4m}{3\tilde{\mathsf{a}}} & \frac{2(m-1)}{\tilde{\mathsf{a}}^2} & 0 \end{pmatrix} \tag{5.47}$$

With $\eta = \frac{4m-1}{3}$, the eigenvalues of A are:

$$\lambda_{u0} = \beta, \quad \lambda_{u\pm} = \beta \pm \alpha\rho\sqrt{\eta}, \quad \lambda_{v\pm} = \beta \pm \alpha\rho\sqrt{\tau} \tag{5.48}$$

The matrix of eigenvectors and its inverse are:

$$\mathsf{T}_R = \begin{pmatrix} \frac{4\tilde{\mathsf{a}}^2}{3\eta} & \frac{1}{\rho} & -\frac{1}{\rho} & \frac{1}{\rho\eta} & -\frac{1}{\rho\eta} \\ 0 & \frac{1}{8\tilde{\mathsf{a}}\rho} & -\frac{1}{8\tilde{\mathsf{a}}\rho} & 0 & 0 \\ 0 & \frac{3\sqrt{\tau}}{4\tilde{\mathsf{a}}} & \frac{3\sqrt{\tau}}{4\tilde{\mathsf{a}}} & 0 & 0 \\ 0 & \frac{\sqrt{\tau}}{2} & \frac{\sqrt{\tau}}{2} & \frac{1}{2\sqrt{\eta}} & \frac{1}{2\sqrt{\eta}} \\ \frac{1}{\eta} & \frac{1}{\tilde{\mathsf{a}}^2\rho} & -\frac{1}{\tilde{\mathsf{a}}^2\rho} & -\frac{m-1}{\tilde{\mathsf{a}}^2\rho\eta} & \frac{m-1}{\tilde{\mathsf{a}}^2\rho\eta} \end{pmatrix}, \quad \mathsf{T}_R^{-1} = \begin{pmatrix} \frac{m-1}{\tilde{\mathsf{a}}^2} & -\frac{8m}{\tilde{\mathsf{a}}} & 0 & 0 & 1 \\ 0 & 4\tilde{\mathsf{a}}\rho & \frac{2\tilde{\mathsf{a}}}{3\sqrt{\tau}} & 0 & 0 \\ 0 & -4\tilde{\mathsf{a}}\rho & \frac{2\tilde{\mathsf{a}}}{3\sqrt{\tau}} & 0 & 0 \\ \frac{\rho}{2} & \frac{4\tilde{\mathsf{a}}\rho}{3} & -\frac{2\tilde{\mathsf{a}}\sqrt{\eta}}{3} & \sqrt{\eta} & -\frac{2\tilde{\mathsf{a}}^2\rho}{3} \\ -\frac{\rho}{2} & -\frac{4\tilde{\mathsf{a}}\rho}{3} & -\frac{2\tilde{\mathsf{a}}\sqrt{\eta}}{3} & \sqrt{\eta} & \frac{2\tilde{\mathsf{a}}^2\rho}{3} \end{pmatrix} \tag{5.49}$$

One can easily see that if $\tau, \eta > 0$, the eigenvalues are real and there exist a complete set of eigenvectors, which means that the system is strongly hyperbolic.

The characteristics of the system are constructed with $\mathsf{C} = \mathsf{T}_R^{-1} v$,

$$\mathsf{C} = (\mathsf{U}_0, \quad \mathsf{V}_-, \quad \mathsf{V}_+, \quad \mathsf{U}_-, \quad \mathsf{U}_+)^T$$

where

$$\begin{aligned}
\lambda_{u0}: \quad \mathsf{U}_0 &= \tilde{\mathsf{G}} + \frac{m-1}{\tilde{\mathsf{a}}^2}\tilde{\mathsf{a}}_{,r} - \frac{8m}{\tilde{\mathsf{a}}}\Phi_{,r} \\
\lambda_{u\pm}: \quad \mathsf{U}_\pm &= \sqrt{\eta}\left(\tilde{\mathsf{X}} - \frac{2\tilde{\mathsf{a}}}{3}\mathsf{K}\right) \pm \rho\left(-\frac{1}{2}\tilde{\mathsf{a}}_{,r} - \frac{4\tilde{\mathsf{a}}}{3}\Phi_{,r} + \frac{2\tilde{\mathsf{a}}^2}{3}\tilde{\mathsf{G}}\right) \\
\lambda_{v\pm}: \quad \mathsf{V}_\pm &= \frac{2\tilde{\mathsf{a}}}{3\sqrt{\tau}}\mathsf{K} \mp 4\tilde{\mathsf{a}}\rho\Phi_{,r}
\end{aligned} \tag{5.50}$$

Obs. If $\eta = \tau = 1$ then the characteristics $\mathsf{U}_\pm$ and $\mathsf{V}_\pm$ propagate along the light cones. ($\lambda_\pm \equiv \lambda_{u\pm} = \lambda_{v\pm} = \beta \pm \frac{\alpha}{\sqrt{a}} = \beta \pm \alpha\rho$).

The system (5.39)–(5.43) is equivalent with the following second order in time system:

$$\ddot{\Phi} \simeq \left[\tau(\alpha\rho)^2 - \beta^2\right]\Phi_{,rr} + 2\beta\dot{\Phi}_{,r} \tag{5.51}$$

$$\dot{\mathsf{U}}_0 \simeq \beta\,\mathsf{U}_{0,r} \tag{5.52}$$

$$\ddot{\tilde{\mathsf{a}}} \simeq \left[\eta(\alpha\rho)^2 - \beta^2\right]\tilde{\mathsf{a}}_{,rr} + 2\beta\dot{\tilde{\mathsf{a}}}_{,r} - \frac{4(\alpha\rho\tilde{\mathsf{a}})^2}{3}\mathsf{U}_{0,r} - 8(\alpha\rho)^2(\eta-\tau)\tilde{\mathsf{a}}\Phi_{,rr} \tag{5.53}$$

Notice that Φ obeys a standalone wave equation, while $\tilde{\mathsf{a}}$ obeys an wave equation coupled with an advection equation and with the wave equation for Φ. The equation for $\tilde{\mathsf{a}}$, (5.53) decouples from the one for Φ, (5.51) if $\eta = \tau$. It decouples from the advection equation (5.52) only if the shift is zero.

5.3 Numerical Implementation

Now the system (5.39)–(5.43) is integrated numerically in a finite spatial domain $[r_{\min}, r_{\max}]$. The free parameters m and τ are fixed to unity ($m = \tau = \eta = 1$). The grid has $N+1$ equidistant points $r_i = r_{min} + i\,h$, $i = \overline{0, N}$, with $h = \frac{1}{N}(r_{\max} - r_{\min})$ the grid spacing. The discretization uses the method of lines approach. The space derivatives are approximated by centered FDOs and the time integration is performed using the classical 4th order Runge-

Kutta method.
It is customary in numerical relativity to use one-point upwinded stencils for the Lie derivatives. In this chapter *all* the derivatives will be approximated using centered stencils. At the edges of the grid, finite differencing requires additional points. These "ghost points" are populated according to the boundary prescriptions specific for each type of boundary as will be described later in 5.3.1.

5.3.1 Boundary Algorithms

Denote by $\tilde{a}$, Φ, K, $\tilde{X}$ and $\tilde{G}$ the grid functions corresponding to the continuum functions, $\mathsf{\tilde{a}}$, $\mathsf{\Phi}$, K, $\mathsf{\tilde{X}}$ and, respectively, $\mathsf{\tilde{G}}$. Also U_0, $U_\pm$, $V_\pm$ will stand for the discrete characteristics corresponding to the continuum U_0, $\mathsf{U}_\pm$, $\mathsf{V}_\pm$, defined in (5.50), with the continuum derivative operator replaced by the associated $2n$-accurate centered FDO.

timelike boundary zero shift (Alg:2bc)

Consider the boundary at $x = x_1$. If $\beta = 0$ then $\lambda_+ > 0$ and $\lambda_- < 0$, so two boundary conditions (maximally disipative) are necessary. The Sommerfeld type conditions are imposed:

$$\begin{aligned} \mathsf{V}_+ &= v_+(t) \\ \mathsf{U}_+ &= u_+(t). \end{aligned} \tag{5.54}$$

The proposed numerical boundary prescription is the following algebraic system to be solved for the ghost points of the main variables.

$$\tilde{X}_N - \frac{2\tilde{a}}{3}K_N + \rho\left(\frac{2}{3}\tilde{a}^2\tilde{G}_N - \frac{1}{2}D^{(1,n)}\tilde{a}_N - \frac{4}{3}\tilde{a}D^{(1,n)}\Phi_N\right) = u_+(t)$$
$$\frac{2\tilde{a}}{3}K_N - 4\rho\,\tilde{a}D^{(1,n)}\Phi_N = v_+(t)$$

$$\begin{aligned}
&D_-^{2n+1}\Phi_{N+i} = 0, \qquad i = \overline{1,n}\\
&D_-^{2n+1}\tilde{a}_{N+i} = 0, \qquad i = \overline{1,n}\\
&D_-^{2n}K_{N+i} = 0, \qquad i = \overline{1,n}\\
&D_-^{2n}\tilde{X}_{N+i} = 0, \qquad i = \overline{1,n}\\
&D_-^{2n}\tilde{G}_{N+i} = 0, \qquad i = \overline{1,n}
\end{aligned} \tag{5.55}$$

timelike boundary with $\beta > 0$ (Alg:3bc)

Consider again the boundary at $x = x_1$. If $\beta > 0$, $\lambda_+ > 0$ and $\lambda_- < 0$, then the outer boundary requires three boundary conditions. These are:

$$V_+ = v_+(t) \tag{5.56}$$
$$U_+ = u_+(t) \tag{5.57}$$
$$U_0 = u_0(t). \tag{5.58}$$

The numerical prescription is the following:

(1.) save the value G_N computed with the evolution equations:
$G_N^{\text{temp}} = G_N$

(2.) solve the following linear system for ghost points and G_N:
$$\tilde{X}_N - \frac{2\tilde{a}}{3}K_N + \rho\left(\frac{2}{3}\tilde{a}^2\tilde{G}_N - \frac{1}{2}D^{(1,n)}\tilde{a}_N - \frac{4}{3}\tilde{a}D^{(1,n)}\Phi_N\right) = u_+(t)$$
$$\frac{2\tilde{a}}{3}K_N - 4\rho\,\tilde{a}D^{(1,n)}\Phi_N = v_+(t)$$
$$\tilde{G}_N - \frac{8}{\tilde{a}}D^{(1,n)}\Phi_N = u_0(t)$$

$$D_-^{2n+1}\Phi_{N+i} = 0, \qquad i = \overline{1,n}$$
$$D_-^{2n+1}\tilde{a}_{N+i} = 0, \qquad i = \overline{1,n}$$
$$D_-^{2n}K_{N+i} = 0, \qquad i = \overline{1,n}$$
$$D_-^{2n}\tilde{X}_{N+i} = 0, \qquad i = \overline{1,n}$$
$$D_-^{2n}\tilde{G}_{N+i} = 0. \qquad i = \overline{1,n}$$

(3.) restore G_N:
$$G_N = G_N^{\text{temp}} \tag{5.59}$$

The procedure (5.59) enables us to populate the ghost points without overwriting any of the values already computed by the evolution equations.

outflow boundary (Alg:0bc)

Consider the boundary at $x = x_0$. If $\beta > 0$ and $\lambda_\pm > 0$ then the inner boundary does not require any boundary conditions. However, numerically it is necessary to provide a way to populate the ghost points. This is achieved

with the following extrapolations conditions

$$\begin{aligned} D_+^{2n+1}\Phi_{-i} &= 0, \qquad i = \overline{1,n} \\ D_+^{2n+1}\tilde{a}_{-i} &= 0, \qquad i = \overline{1,n} \\ D_+^{2n+1}K_{-i} &= 0, \qquad i = \overline{1,n} \\ D_+^{2n+1}\tilde{X}_{-i} &= 0, \qquad i = \overline{1,n} \\ D_+^{2n+1}\tilde{G}_{-i} &= 0. \qquad i = \overline{1,n} \end{aligned} \tag{5.60}$$

5.4 Numerical Results: Linear Case

This section presents the results of some numerical tests performed for the system (5.39)–(5.43) discretized as described in 5.3. In order to check the validity of a certain boundary algorithm, different types of initial data are evolved, with boundaries placed at various locations. The boundary data is given analytically.

In all the simulations presented in this section, the lower order terms and the coefficients of the main variables that appear in the principal part are given analytically (linear case).

In some cases, Kreiss-Oliger dissipation, (2.32), is added to the rhs of the equations, including the boundary points.

If v is the analytical value of one of the main variables and v its numerical value, then its error is given by:

$$\operatorname{err} v = h\sqrt{\sum_{i=0}^{N} |v_i(t) - \mathsf{v}(t, r_i)|^2} \tag{5.61}$$

The total error is defined in the following way:

$$\begin{aligned} \text{Total Error}^2 &= |\operatorname{err}\tilde{a}|^2 + |\operatorname{err}\Phi|^2 + |\operatorname{err} D_+\tilde{a}|^2 + |\operatorname{err} D_+\Phi|^2 \\ &+ |\operatorname{err} K|^2 + \left|\operatorname{err}\tilde{X}\right|^2 + \left|\operatorname{err}\tilde{G}\right|^2 \end{aligned} \tag{5.62}$$

All the tests assume $N = 20$ and $h = 0.5$ for the lowest resolution. They evaluate the total error, the overall convergence factor and the individual errors versus time, up to 1000 crossing times, for different orders of approximations ($2n = 2, 4, 6, 8$).

5.4.1 test 1: two timelike boundaries, zero shift

In this test, a Schwarzschild black hole is evolved in isotropic coordinates:

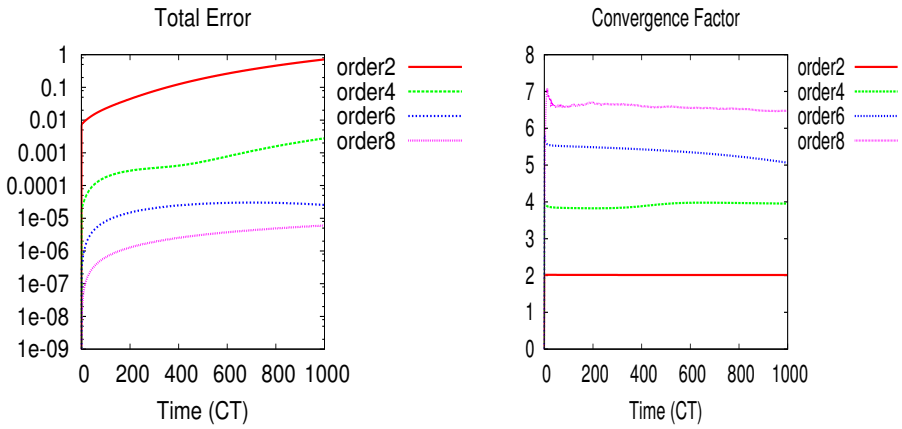
$$ds^2 = -\left(\frac{1-2r}{1+2r}\right)^2 dt^2 + \left(1+\frac{1}{2r}\right)^4 dr^2 + r^2\left(1+\frac{1}{2r}\right)^4 \left(d\theta + \sin^2\theta d\phi\right) \tag{5.63}$$

in the domain $r \in [r_{\min}, r_{\max}] = [10, 20]$.
The ADM variables, the gauge and the BSSN variables are:

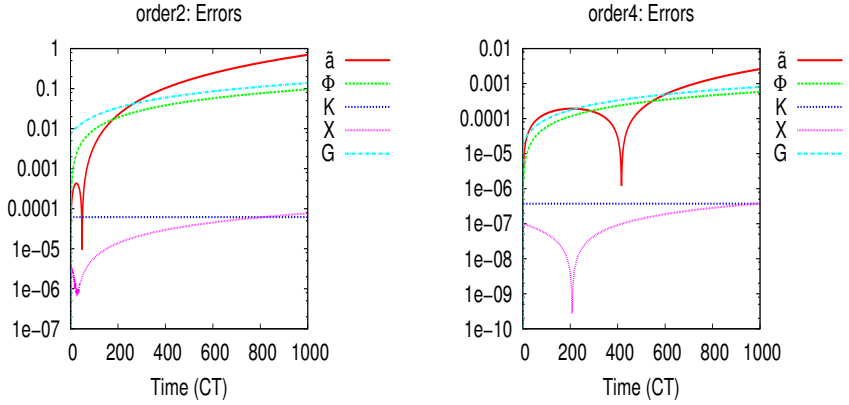
$$\begin{aligned}
\textbf{ADM var:} \quad & \mathsf{a} = \left(1+\frac{1}{2r}\right)^4, \quad \mathsf{b} = \frac{(2r+1)^4}{16r^2}, \quad \mathsf{X} = 0 \quad \mathsf{Y} = 0 \\
\textbf{Gauge:} \quad & \alpha = 1 - \frac{2}{2r+1}, \quad \beta = 0 \\
\textbf{BSSN var:} \quad & \tilde{\mathsf{a}} = \frac{1}{r^{4/3}}, \quad \Phi = \log\left(r+\frac{1}{2}\right) - \frac{2\log(r)}{3}, \quad \tilde{\mathsf{X}} = \mathsf{K} = 0, \quad \tilde{\mathsf{G}} = -\frac{4\sqrt[3]{r}}{3}
\end{aligned} \tag{5.64}$$

The distinct speeds are $\{0, \pm\frac{4r^2(2r-1)}{(2r+1)^3}\}$. This means that both boundaries are timelike and are treated with the algorithm **Alg:2bc** given in (5.55).
The Courant factor is $\lambda = 0.25$ and no artificial dissipation is used. The results are presented in fig. 5.1.

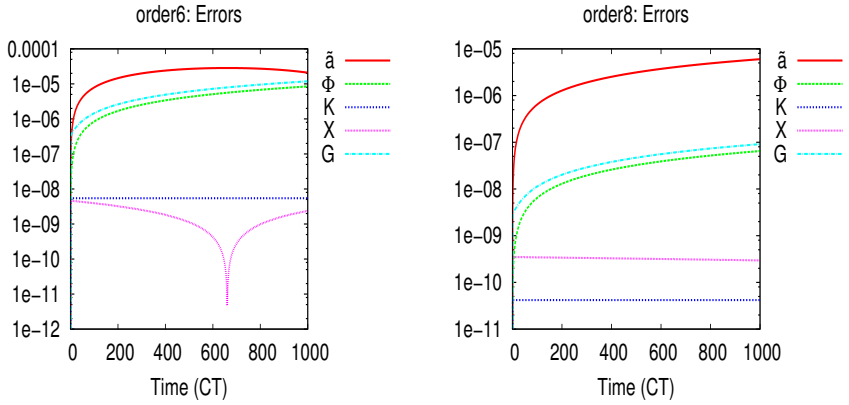
For all orders of approximation tested, the scheme is stable and the convergence factor of the scheme practically coincides with the order of the spatial discretization. The error of K remains constant while the errors of all the other variables have at a bounded growth.



(a) Total Error and Convergence Factor



(b) Individual Errors (orders 2 and 4)



(c) Individual Errors (orders 6 and 8)

Figure 5.1: **test 1: Schwarzschild Black Hole in isotropic coordinates (Alg:2bc–Alg:2bc)** Both boundaries ($r_{\min} = 10$ and $r_{\max} = 20$) are timelike and are treated using the same algorithm, (**Alg:2bc**). No artificial dissipation has been used.

5.4.2 test 2(a): two timelike boundaries, with shift (static)

In this test a Schwarzschild black hole is evolved in Eddington-Finkelstein coordinates:

$$ds^2 = -\left(1-\frac{2}{r}\right)dt^2 + \left(1+\frac{2}{r}\right)dr^2 + \frac{4}{r}dtdr + r^2\left(d\theta + \sin^2\theta d\phi\right) \qquad (5.65)$$

in the domain $r \in [r_{\min}, r_{\max}] = [10, 20]$. The ADM variables, the gauge and the BSSN variables are:

$$\textbf{ADM var:} \qquad \mathsf{a} = \tfrac{r+2}{r}, \quad \mathsf{b} = r^2 \qquad \mathsf{X} = -\frac{2(r+1)}{\sqrt{r^5(r+2)}}, \quad \mathsf{Y} = 2\sqrt{\frac{r}{r+2}}$$

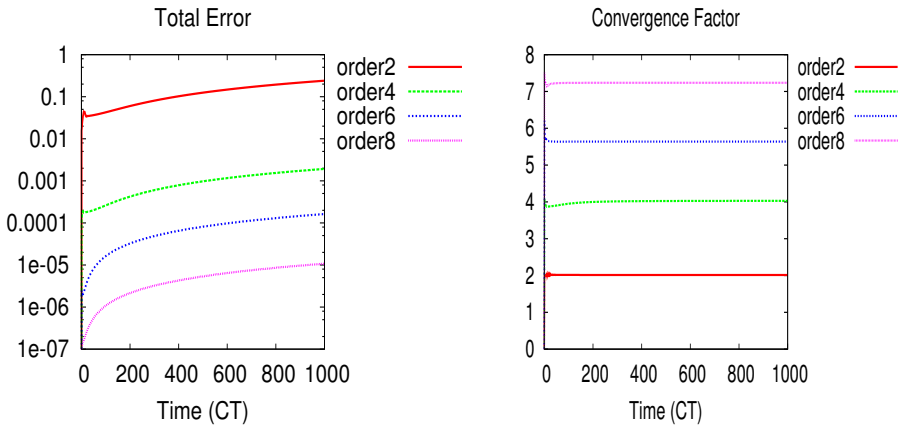
$$\textbf{Gauge:} \quad \alpha = \sqrt{\tfrac{r}{r+2}}, \quad \beta = \tfrac{2}{r+2} \qquad (5.66)$$

$$\textbf{BSSN var:} \qquad \tilde{\mathsf{a}} = \tfrac{(r+2)^{2/3}}{r^2}, \qquad \Phi = \frac{1}{12}\log\left(r^3(r+2)\right),$$
$$\tilde{\mathsf{X}} = -\frac{4(2r+3)}{3r^{7/2}(r+2)^{5/6}}, \qquad \mathsf{K} = \tfrac{2(r+3)}{(r(r+2))^{3/2}}, \qquad \tilde{\mathsf{G}} = -\frac{4r(r+3)}{3(r+2)^{5/3}} \qquad (5.67)$$

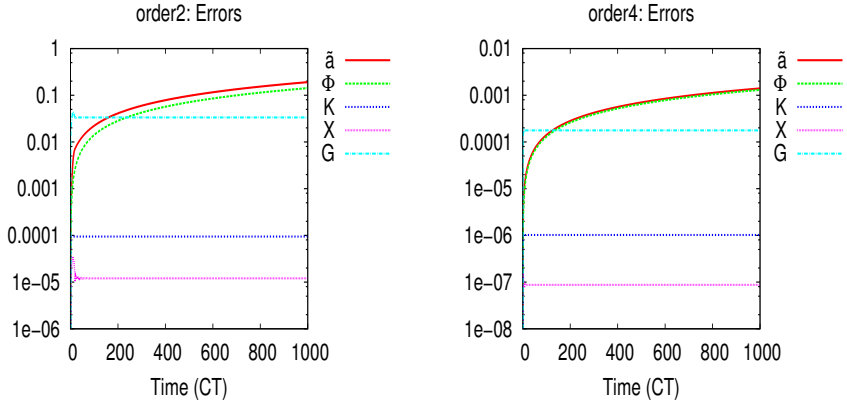
The distinct speeds are $\{\frac{2}{r+2}, 1, \frac{2-r}{r+2}\}$. This means that both boundaries are timelike but because of the shift, the inner boundary asks for two boundary conditions (implemented according to the boundary algorithm **Alg:2bc**), while the outer boundary asks for three (algorithm **Alg:3bc**).

The Courant factor is $\lambda = 0.25$ and dissipation is added for 4th, 6th and 8th order approximations with the following coefficients $\sigma = 0.1$, 0.3 and respectively, 0.5. The results are presented in fig. 5.2.

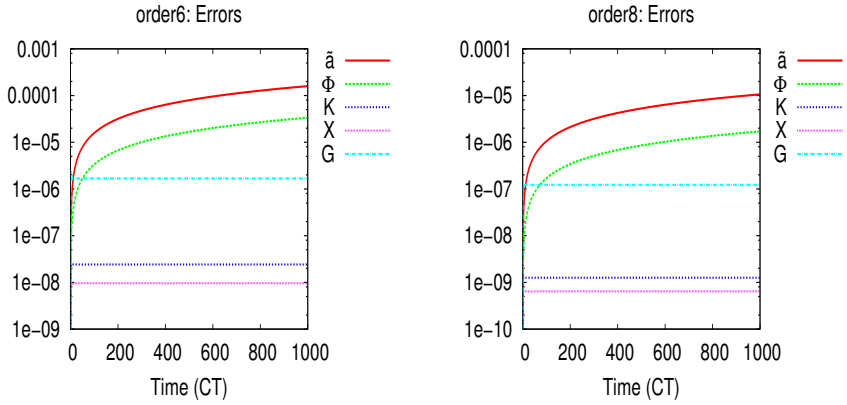
For all the orders tested the scheme is stable and the convergence factor of the scheme practically coincides with the order of the spatial discretization. The errors of K, $\tilde{\mathsf{X}}$ and $\tilde{\mathsf{G}}$ remain constant while the errors of $\tilde{\mathsf{a}}$ and Φ grow linearly in time.



(a) Total Error and Convergence Factor



(b) Individual Errors (orders 2 and 4)



(c) Individual Errors (orders 6 and 8)

Figure 5.2: **test 2(a): Schwarzschild Black Hole in Eddington-Finkelstein coordinates (Alg:2bc–Alg:3bc)** Both boundaries ($r_{\min} = 10$ and $r_{\max} = 20$) are timelike, but the inner boundary asks for two boundary conditions (**Alg:2bc**) while the the outer boundary asks for three boundary conditions (**Alg:3bc**). For 4th, 6th and 8th order approximations, artificial dissipation has been used. For all the orders tested, the errors of the variables $\tilde{a}$ and Φ grow linearly with time, while the errors of K, $\tilde{X}$ and $\tilde{G}$ remain constant.

5.4.3 test 2(b): two timelike boundaries, with shift (dynamic)

This test evolves the flat spacetime with shift given by:

$$ds^2 = -\left(1-\beta^2\right)dt^2 + dr^2 + 2\beta dt dr + (r+\beta t)^2\left(d\theta + \sin^2\theta d\phi\right) \tag{5.68}$$

$$\textbf{ADM var:} \quad \mathsf{a}=1, \quad \mathsf{b}=(r+\beta t)^2 \quad \mathsf{X}=0, \quad \mathsf{Y}=0$$

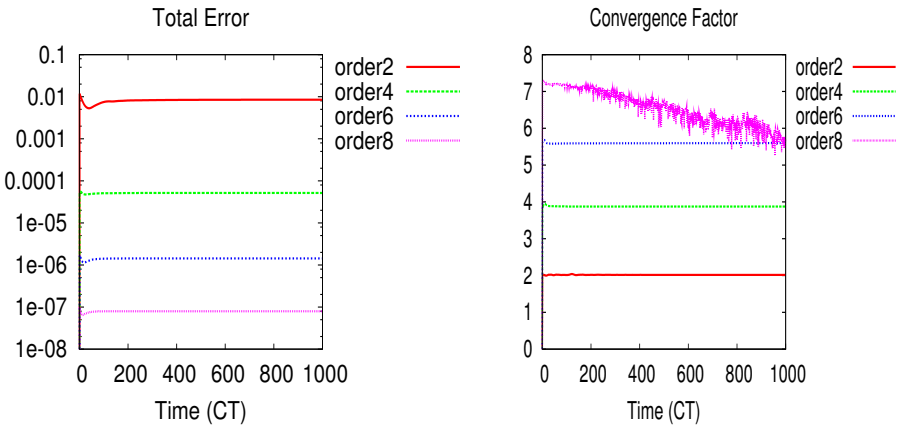
$$\textbf{Gauge:} \qquad \alpha = 1, \qquad \beta = \text{const} > 0 \tag{5.69}$$

$$\textbf{BSSN var:} \qquad \tilde{\mathsf{a}} = \tfrac{1}{(r+\beta t)^{4/3}}, \qquad \Phi = \frac{1}{3}\log(r+\beta t),$$
$$\tilde{\mathsf{X}}=0, \qquad \mathsf{K}=0, \qquad \tilde{\mathsf{G}} = -\frac{4}{3}\sqrt[3]{r+\beta t} \tag{5.70}$$

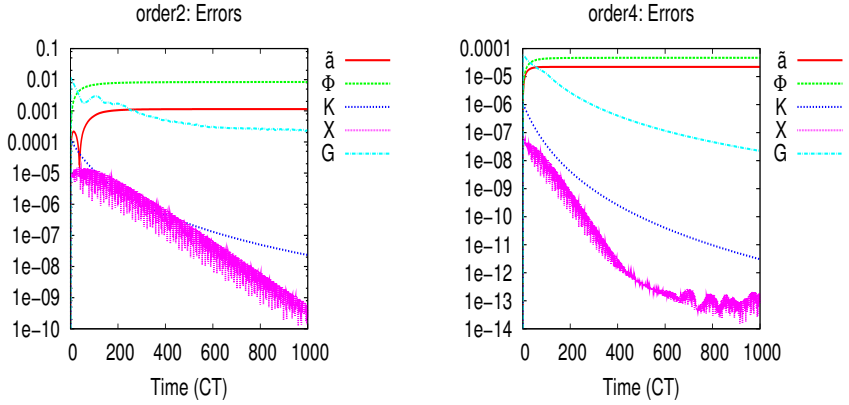
The distinct speeds are $\{\beta,\, \beta+1,\, \beta-1\}$. The shift is chosen to be $\beta = 0.001$. Both boundaries are timelike but because of the shift, the inner boundary asks for two boundary conditions (**Alg:2bc**) while the outer boundary asks for three (**Alg:3bc**). The case is similar to the one presented in 5.4.2 with the difference that now the field variables and the boundary data are also time dependent.

The Courant factor is $\lambda = 0.1$ and dissipation is added for 4th, 6th and 8th order approximations with the following coefficients $\sigma = 0.1, 0.3$ and respectively, 0.5. The results are presented in fig. 5.3.

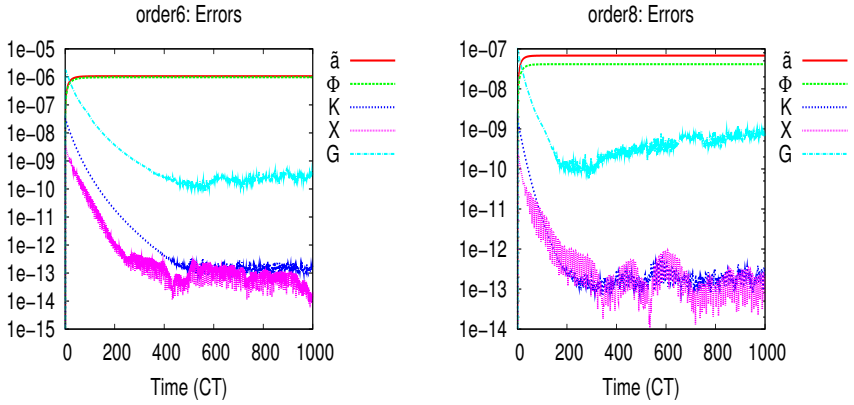
For all the orders tested the scheme is stable and the convergence factor of the scheme practically coincides with the order of the spatial discretization. The total error remains constant in time. The errors of the variables $\tilde{\mathsf{a}}$ and Φ grow linearly with time, while K, $\tilde{\mathsf{X}}$ and $\tilde{\mathsf{G}}$ decrease in time to the machine precision.



(a) Total Error and Convergence Factor



(b) Individual Errors (orders 2 and 4)



(c) Individual Errors (orders 6 and 8)

Figure 5.3: **test 2(b): Flat Space Time with Shift (Alg:2bc–Alg:3bc)**($\beta = 0.001$) This case has the same boundary treatment as the one presented in fig.5.2. Also the dissipation is added in the same way. The drop in convergence of the 8th order scheme is due to the fact that the error of some variables is in the order of machine precision. The total error remains constant in time. The errors of the variables $\tilde{a}$ and Φ grow linearly with time, while K, $\tilde{X}$ and $\tilde{G}$ decrease in time to the machine precision.

5.4.4 test 3 : spacelike inner boundary and timelike outer boundary

The spacetime evolved is a Schwarzschild black hole in Eddington-Finkelstein coordinates (presented in 5.4.2) but now the evolution domain is $r \in [r_{\min}, r_{\max}] = [1.85, 11.85]$. The inner boundary is spacelike and is treated numerically according to **Alg:0bc**, while the outer boundary is timelike and requires three boundary conditions, **Alg:3bc**.

The Courant factor is $\lambda = 0.25$. For 4th, 6th and 8th order approximations, artificial dissipation has been used with the following coefficients $\sigma = 0.3$, 0.5 and respectively, 0.7. The results are presented in fig. 5.4.

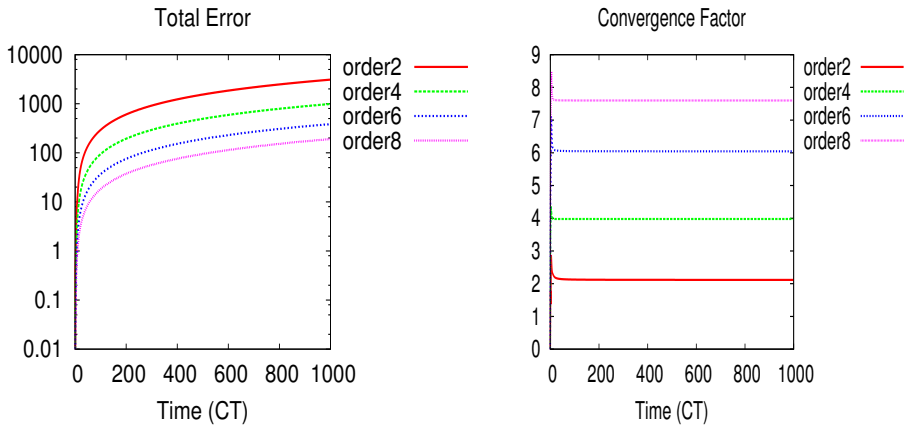
The error grows faster than in the previous cases, but the growth remains bounded and the scheme is convergent. For 2nd, 4th and 6th order of approximation, the errors of the variables $\tilde{\mathsf{a}}$ and Φ grow linearly with time, while K, $\tilde{\mathsf{X}}$ and $\tilde{\mathsf{G}}$ remain constant. The 8th order manifests the same behavior up to $\sim 600CT$ when the errors for $\tilde{\mathsf{X}}$ and $\tilde{\mathsf{G}}$ start to grow exponentially. This growth is convergent as one can see in fig. 5.5

5.5 Numerical Results: Nonlinear Case

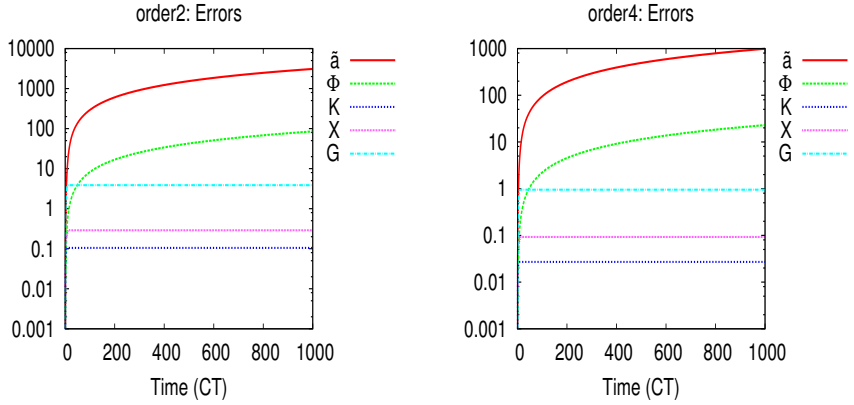
In the previous section it has been shown numerically that the implementation of the linear version of the BSSN system with the numerical boundary conditions (5.60), (5.55) or (5.59) leads to stable evolutions.

This section presents what happens in the nonlinear case for one of the spacetimes investigated before. That is, the initial data is constructed using the metric of a Schwarzschild black hole in isotropic coordinates and the evolution domain is $r \in [r_{\min}, r_{\max}] = [10, 20]$.

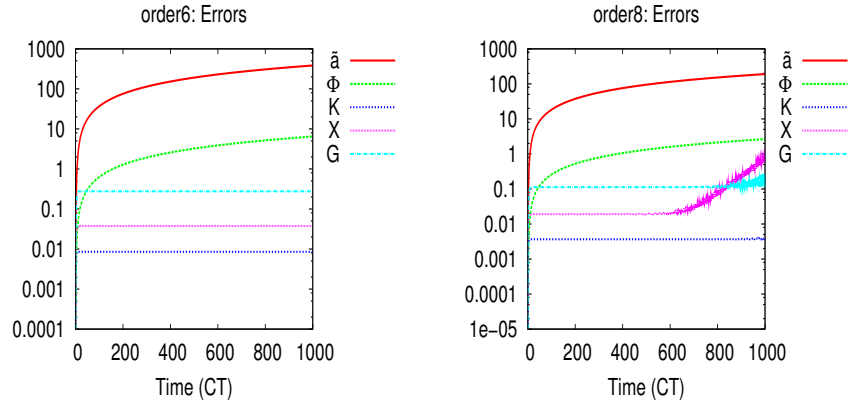
Both boundaries are timelike, requiring each two boundary conditions, which are implemented using the algorithm (**Alg:2bc**). If in the linear case no dissipation was needed to achieve stability, in the nonlinear case this is required ($\sigma = 10$ for all the orders tested). The results are presented in fig.



(a) Total Error and Convergence Factor



(b) Individual Errors (orders 2 and 4)



(c) Individual Errors (orders 6 and 8)

Figure 5.4: **test 3: Schwarzschild BH in Eddington-Finkelstein coordinates (Alg:0bc–Alg:3bc)** The inner boundary ($r_{\min} = 1.85$) is spacelike and the algorithm **Alg:0bc** is applied while the outer boundary ($r_{\max} = 11.85$) is timelike asking for three boundary conditions (**Alg:3bc**). For 4th, 6th and 8th order, artificial dissipation has been used. The errors of the variables $\tilde{a}$ and Φ grow linearly with time, while K, $\tilde{X}$ and $\tilde{G}$ remain constant. With 8th order, the errors of $\tilde{X}$ and $\tilde{G}$ exhibit an exponential growth which vanishes at higher resolution (see also fig. 5.5)

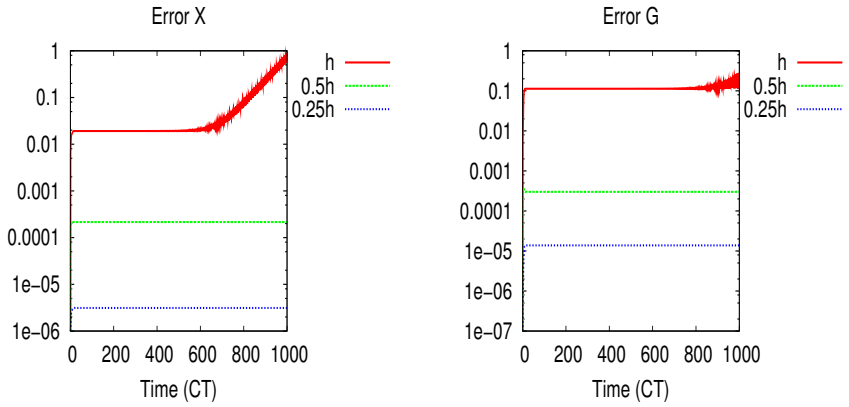
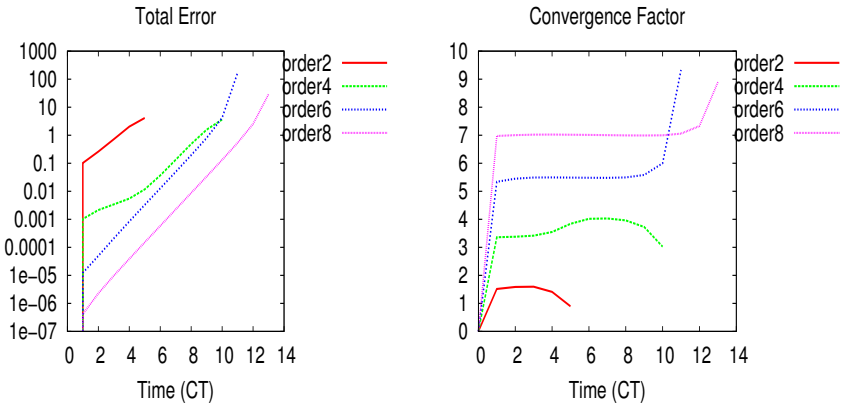
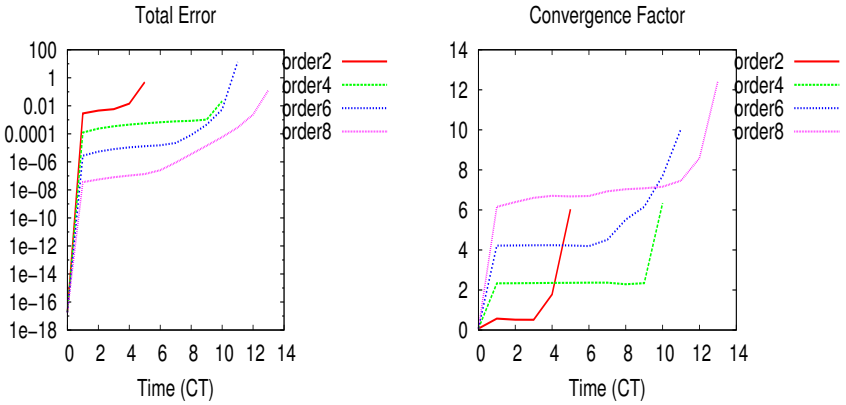


Figure 5.5: **test 3: Schwarzschild Black Hole in Eddington-Finkelstein coordinates (Alg:0bc–Alg:3bc)** This plot shows the l^2 norm of the errors for the main variables $\tilde{\mathsf{X}}$ and $\tilde{\mathsf{G}}$ at three different resolutions. It illustrates that the exponential growth seen in fig. 5.4 for the 8th order, goes away with increasing the resolution.

5.6 which shows the total error of the main variables and the norm of the constraints. As one can see, the runs do not last very long (at most 13 CT at low resolution) and the error manifests an exponential growth. However, the growth is convergent (the lower the resolution, the later it crashes). The reason for crash is not because the numerical scheme would be unstable, but is due to the fact that the error of $\tilde{\mathsf{a}}$ increases and makes $\tilde{\mathsf{a}}$ zero or negative. One work-around would be to redefine the variables by some proper rescaling. Also in the nonlinear case dissipation plays a crucial role. Further investigation is needed.



(a) Main Variables



(b) Constraints

Figure 5.6: **Nonlinear Effects** This plot shows the results of evolving of Schwarzschild Black Hole in isotropic coordinates (same setup as in fig. 5.1) using the fully nonlinear system. The total error of the main variables (a) and the norm of the constraints (b) have an exponentially growth but this growth is convergent. Also the higher the order of approximation the longer the life of the numerical evolution. The code crashes when $\tilde{a}$ becomes zero or negative.

Chapter 6

Summary and Outlook

6.1 Summary

In this thesis I have explored different high order finite differencing techniques for implementing the initial value problem (Chapter 2 and 3) and initial boundary value problem (Chapter 4 and 5) for hyperbolic system which are first order in time and second order in space. The main results of the thesis are summarized below, chapter by chapter.

Chapter 2: Initial Value Problem for 2nd Order Systems

- **Analysis of first and second discrete derivative operators**

  A series of properties have been derived an collected in 2.1.4–2.1.5. Some relevant ones are enumerated below.

  While first derivatives (centered or off-centered) do not converge in the limit $n \to \infty$ at the maximum grid frequency ($\xi = \pi$), second derivatives do converge at all frequencies (that is the highest frequency in the grid will not be captured by the first order derivative, regardless of increasing the order of approximation or the off-centering, while the second centered derivative can "see" it and approximates it better with increasing

order).

For first order derivatives, increasing the off-centering (s) at a fixed order of approximation increases the error of the derivative at small frequencies. At larger frequencies this behavior changes. For operators that are off-centered by s grid points, there are exactly s frequencies in $(0, \pi)$ where the error cancels. However, for $s \geq 2$ there are large intervals where the error is far larger than for $s = 0$. For $s = 1$ it is shown that while at small frequencies the error is slightly larger than for $s = 0$, for each order n, there is a frequency $\xi^{(n)}$ beyond which the error is smaller than for the case $s = 0$.

As a consequence, off-centering of the first order derivative in the case of the advection equation, increases the error at small frequencies, while at high frequencies, this situation can change. Off-centering by more than one point requires dissipation for stability.

- **Generalization of the stability analysis method from [14]**

  The method proposed in [14] for analyzing the stability of second order in space and first order in time systems in terms of discrete symmetrizers is extended from second and fourth order accurate operators to the general $2n$-order case, including also the case when some derivatives are approximated with non-centered FDOs, as is customary for treating black hole spacetimes in numerical relativity.

  It is pointed out that neither adding artificial dissipation (as defined in (2.31)) nor shift advection terms affect the eigenvectors of the discrete symmetrizer, and thus the conditions for semidiscrete numerical stability (the Courant limit will of course be affected in general).

Chapter 3: Initial Value Problem for the Wave Equation

The general methods discussed in Chapter 2 are applied to the case of the shifted wave equation in first order in time, second order in space form.

- The **well-posedenss** of the continuum IVP and the **stability of the semi-discrete** scheme (constructed with $2n$-accurate centered FDOs for the second derivative and not necessarily centered $2n$-accurate FDOs for the first derivative) are proven.

- **Courant limits**

  Off-centerings by more than one point require dissipation for stability. In these cases, the minimal Kreiss-Oliger dissipation needed for stability has been computed and found to be proportional to the shift β.

  For centered schemes, higher order approximations have lower Courant limits. Interestingly, this does not hold for off-centered schemes (when adding just dissipation to be in the local stability regime) — for large enough shift, the Courant limit is actually larger for higher order schemes. In particular, for the one-point off-centered scheme, which does not require artificial dissipation for stability, the turn-around point is at $\beta \approx 0.5$.

  Off-centering generally reduces the Courant limit drastically, except for at least fourth order accurate schemes, when only one-point off-centering is used: for higher than fourth order schemes one-point off-centering only leads to a minor reduction of the CFL factor.

- **Numerical speeds**

  Without shift, higher order approximations always result in more accurate numerical speeds, with nonzero shift this is not generally true at higher frequencies.

  Although the truncation error for the first order derivative increases with the off-centering, the mixing with the second order discrete derivative in the scheme, causes upwinded stencils to give a higher overall accuracy in some situations.

  More precisely, in the case of the 1-D wave equation, it is shown that

advecting shift terms by an odd (even) number of points reduces the errors of the "+" ("−") numerical speeds in some intervals of the spectrum that include the small frequency range, if the shift is not too large. The extent of the regions in the (frequency, shift)-parameter space where this improvement appears decreases with off-centering, such that for $s = 1$ one gets the strongest effect.

Thus, at a given order $2n$, if the shift satisfies $0 < \beta < \frac{n}{n+1}$, then off-centering by one point has in comparison with the centered scheme, better "+" phase speed error for all frequencies, and better "+" group speed error for all frequencies up to a very high frequency in the grid, $\pi - \arccos \frac{n}{n+1}$.

If the semidiscrete problem is not dissipative (e.g. the real part of the eigenvalues is cancelled by adding appropriate dissipation) or the dissipative effects can be neglected (as is the case at small frequency), smaller speed errors will result in better overall accuracy.

If the wave equation is written in first order form (approximating the first derivatives with the corresponding centered FDO), then for a given order of approximation, the second order system discretized with centered FDO has smaller phase and group errors than the first order one (for both eigenvalues), if and only if $|\beta| \leq \frac{2n+3}{4(n+1)}$. If $|\beta|$ is not in this interval then one pair of speeds (phase and group) is better approximated by the second order system, while the other one is better approximated by the first order system.

- **Overall accuracy with different orders Runge-Kutta methods** This issue has been investigated numerically and some conclusions are presented below.

  A $2p$ order Runge-Kutta time integrator discriminates between $2n$-order spacial finite difference schemes as long as $2n \leq 2p + 2$. If the order of the centered FDOs is higher than $2p + 2$ then the error of the time

integrator will be dominant and there is practically no improvement in the accuracy over the previous order.

The higher the order of the spatial approximations, the better the time behavior of the convergence factor.

- **Overall accuracy with the classical 4th order Rungge-Kutta**

  Numerically it is found that: if $2n \geq 6$, with a 4th order Rungge-Kutta time integrator, the overall convergence factor is usually 4, however, at high resolutions and small Courant limits, a $2n$ overall convergence factor can be attained.

Chapter 4: Initial Boundary Value Problem for the Wave Equation

This chapter discusses two discretization methods (the ghost-point method and the SBP-SAT method) for the some types of IBVP that can be formulated for the one-dimensional shifted scalar wave equation.

For the quarter-space problem, $x \in [x_0, \infty)$ there are three types of boundaries at the continuum level, according to the number of boundary conditions that need to be specified in order to obtain a well-posed problem: outflow ($\beta \geq 1$, requires no boundary condition), inflow ($-1 \leq \beta < 1$, requires one boundary condition) and completely-inflow ($\beta < -1$, requires two boundary conditions). Each of them requires at the numerical level a different boundary treatment.

- **ghost-point method**

  For each type of boundary a $2n$-accurate algorithm is proposed.

  Stability analysis is performed for second order accuracy, outflow boundary using both energy method and Laplace transform method.

  The boundary procedures are validated numerically on a grid with two boundaries, via three tests: 1) a stability test with random initial and boundary data (to check the growth of the noise), 2) an accuracy test

(to evaluate the time behaviour of the error and convergence factor in respect to the analytic solution at a fixed Courant factor) and 3) a Courant test (evaluate the error and the convergence factor in respect to the analytic solution at a given time for various Courant factors).

All the algorithms are found to be stable without aid of artificial dissipation if the order is 2 or 4.

For higher orders dissipation is usually needed as detailed below.

The inflow-inflow ($\beta = 0.5$) and outflow-inflow ($\beta = 1$) algorithms require dissipation for $2n = 8$; the outflow-completely inflow algorithm ($\beta = 2$) require dissipation for both $2n = 6$ and $2n = 8$.

For the outflow-inflow case the noise is damped for a few crossing times and afterward the noise grows in time at a rate independent of the grid spacing. Also the error grows linearly in time. In the other two cases, the noise is damped to a constant value which depends on the order of approximation and time integrator and the error stays constant in time.

For 2nd and 4th order the error and the convergence factor depend only slightly on the Courant limit, while for 6th and 8th order, there is a drastic improvement in the accuracy and overall convergence (till 6th and respectively 8th, even if the time integrator is only 4th order) at small Courant factors.

- **SBP-SAT method**
  It is used to implement numerically the inflow boundaries.

  By imposing certain conditions on the SBP operators — a) they should be based on diagonal norms and b) satisfy the positivity condition (4.122)—, an energy can be defined in respect to which the scheme is strongly stable.

  The SBP operators from [57] with 2nd and 4th order interior stencils satisfy these properties while the 6th order SBP-operator not (condi-

tion b) is violated). A new 6th order operator is constructed (given in appendix D) which obeys also this condition.

The SBP-SAT algorithm is validated numerically using the same tests as for the ghost point method.

When evolving random data, the noise is damped to a constant value which depends on the order of approximation and time integrator, and, when evolving analytic data, the error stays constant in time.

Using the 6th order SBP from [57] leads to a stable scheme although it does not satisfy the relation (4.122). However the new SBP operator damps better the noise and exhibits better properties in terms of accuracy and overall convergence.

Chapter 5: BSSN System in Spherical Symmetry

This chapter investigates a 1-D model (with spherical symmetry) of the BSSN formulation of Einstein equations. Several IBVPs are considered and implemented numerically using the ghost-point method.

- **construction:**
  First the ADM equations in spherical symmetry are deduced starting from the most general spacetime element in spherical symmetry; The ADM equations preserve such a symmetry regardless of preservation of (energy-momentum) constraints.

  Then the BSSN equations are derived assuming that the determinant of the conformal metric is unity also in spherical symmetry. In contrast with the ADM equations, the BSSN equations do not automatically preserve spherical symmetry. This is achieved only if the Γ-constraints are obeyed during the evolution.

  Using as gauge conditions a densitized lapse and analytic shift, and imposing the algebraic (trace and determinant) constraints, a PDE system with five equations for five unknown functions is constructed.

The system is first order in time, second order in space and has the structure (1.4).

- **hyperbolicity:**

  With a proper choice of parameters, the system is strongly hyperbolic, with four characteristics ($U_\pm$, $V_\pm$) propagating along the light-cones $\lambda_\pm$ and one (U_0) along the shift vector, β.

- **numerical implementation**

  The system is implemented numerically using centered FDOs for the interior points. Three types of boundaries are considered and for each of them $2n$-accurate prescriptions are given based on the ghost point method.

  1. timelike outer boundary when $\beta = 0$, $\lambda_+ > 0$ and $\lambda_- < 0$ at r_{max}. This case requires two boundary conditions (algorithm **Alg:2bc**).
  2. timelike outer boundary when $\beta > 0$, $\lambda_+ > 0$ and $\lambda_- < 0$ at r_{max}. This case requires three boundary conditions (algorithm **Alg:3bc**).
  3. spacelike inner boundary when $\beta > 0$, $\lambda_+ > 0$ and $\lambda_- > 0$ at r_{min}. This case does not require boundary conditions at the continuum level (**Alg:0bc**).

- **testing the numerical scheme**

  The algorithms are tested using various types of spacetimes:

  – a Schwarzschild black hole in isotropic coordinates, in a domain outside the horizon. Both boundaries are timelike and each of them asks for two boundary conditions (**Alg:2bc**-**Alg:2bc**).

  – a Schwarzschild black hole in Eddington-Finkelstein coordinates in a domain outside the horizon. Both boundaries are timelike, however the inner boundary requires two boundary conditions while

the outer boundary requires three boundary conditions (**Alg:2bc**-**Alg:3bc**).

- a flat metric with constant shift in a domain $[r_{min}, r_{max}] \in (0, \infty)$. The boundary treatment is the same as in the previous case (**Alg:2bc**-**Alg:3bc**). This case provides a test with a dynamically nontrivial spacetime because although the physical metric is flat and has no time dependence, the conformal metric is not flat and also evolves in time.
- a Schwarzschild black hole in Eddington-Finkelstein coordinates in a domain with the inner boundary inside the horizon (spacelike, no boundary conditions needed) and outer boundary outside the horizon, timelike, requires 3 boundary conditions (**Alg:0bc**-**Alg:3bc**).

The accuracy and stability of the numerical schemes are asessed by measuring the total error and the convergence factor in respect to the analytical solution.

In the linear case, that is, when the coefficients of the principal part and the lower order terms are given analytically (linear system with nonconstant coefficients and forcing terms), all the algorithms with $2n = 2, 4, 6, 8$, are stable until the end of the runs (1000 crossing times).

In the fully nonlinear case, the runs last only a few (~ 10) crossing times. However, the cause of the crash does not point to an incorrect boundary treatment, because the convergence is not lost, but rather to a very fast growth of the error which makes the variable $\tilde{a}$ to become zero or negative (and this is not allowed since the equations contain terms proportional with $\sqrt{\tilde{a}}$ or $1/\tilde{a}$). A higher resolution prolonges the lifetime of the runs, but a very important ingredient for stability seems to be the choice of dissipation.

6.2 Outlook

- The general stability method presented in Chapter 2 could be applied to the BSSN system, in a way similar to that used for the wave equation in Chapter 3. Stability analysis of $2n$-accurate finite discretizations for the Cauchy problem of this system is an important step before approaching the discretization of its IBVP.

- For the shifted wave equation, it would be interesting to study the numerical speeds in the multidimensional case, e.g. when the wave propagates in a direction that is not aligned with the grid. A similar analysis for the full Einstein equations can be done but it will require a substantial use of computer algebra methods.

- In Chapter 4 it has been shown using the energy method, that in case of the shifted wave equation, extrapolation conditions imposed at an outflow boundary lead to stability, for a 2nd order accurate scheme. A generalization to higher orders would be desirable because in general, the energy method is simpler than the Laplace transform method, and can open the way to discretizing the IBVP for more complicated systems.

- Using the spherically symmetric model for BSSN, proposed in Chapter 5 one can try to find boundary treatments also for the case when the interior equations use upwinded Lie terms. Also one can try other types of gauges. However, a stability analysis it will be possible only after the simplest model of the wave equation is perfectly understood for all its possible initial boundary value problems.

Appendix A

Harmonic Formulation

In this formulation each coordinate x^α satisfies a wave equation:

$$\Box x^\alpha = \frac{1}{\sqrt{-g}}\partial_\mu\left(\sqrt{-g}g^{\mu\beta}\partial_\beta x^\alpha\right) = F^\alpha\,, \tag{A.1}$$

The gauge degrees of freedom are fixed by choosing the four functions F^α. The condition (A.1) gives rise to the following constraint:

$$C^\alpha := \Box x^\alpha - F^\alpha = 0\,, \tag{A.2}$$

which used in combination with the Einstein tensor $G^{\mu\nu}$ leads to the generalized harmonic evolution system

$$E^{\mu\nu} := G^{\mu\nu} - \nabla^{(\mu}C^{\nu)} + \frac{1}{2}g^{\mu\nu}\nabla_\alpha C^\alpha = 0\,. \tag{A.3}$$

A densitized inverse metric is introduced by $\tilde{g}^{\mu\nu} := \sqrt{-g}g^{\mu\nu}$. Assume that the gauge source functions F^α do not depend on the derivatives of the metric. Then after some algebraic manipulation involving also constraints adjustments, the system (A.3) can be written as a set of ten wave equations:

$$\partial_\rho\left(g^{\rho\sigma}\partial_\sigma\tilde{g}^{\mu\nu}\right) = S^{\mu\nu}\,, \tag{A.4}$$

where $S^{\mu\nu}$ are non-principle source terms consisting of at most first derivatives of the evolution variables.

By introducing the auxiliary variables,

$$Q^{\mu\nu} \equiv n^{\rho}\partial_{\rho}\tilde{g}^{\alpha\beta}\,, \tag{A.5}$$

where n^{ρ} is timelike and tangential to the outer boundary, the system (A.4) can be cast in a form that is first differential-order in time, and second-differential order in space:

$$\begin{aligned}
\partial_t \tilde{g}^{\mu\nu} &= -\frac{g^{it}}{g^{tt}}\partial_i \tilde{g}^{\mu\nu} + \frac{1}{g^{tt}}Q^{\mu\nu}\,, \\
\partial_t Q^{\mu\nu} &= -\partial_i\left(\left(g^{ij} - \frac{g^{it}g^{jt}}{g^{tt}}\right)\partial_j \tilde{g}^{\mu\nu}\right) - \partial_i\left(\frac{g^{it}}{g^{tt}}Q^{\mu\nu}\right) + \tilde{S}^{\mu\nu}(\tilde{g}, \partial\tilde{g}, F, \partial F)\,,
\end{aligned} \tag{A.6}$$

where $\tilde{S}^{\mu\nu}(\tilde{g}, \partial\tilde{g}, F, \partial F)$ are again, non-principle source terms consisting of at most first derivatives of the evolution variables and are determined by the choice of gauge.

Appendix B

3+1 split and ADM equations

The four-dimensional spacetime (described by a metric ${}^{(4)}g_{\mu\nu}$ in some coordinates x^μ with $\mu,\, \nu = \overline{0,3}$, $x^0 = t$) is a foliation of three-dimensional spacelike surfaces (slices), each of them being labeled by the so-called "time" coordinate. Each slice is a differential manifold in its own described by the induced metric $g_{ij} = {}^{(4)}g_{ij}$ with $i,\, j = \overline{1,3}$. The connection between slices is realized using a set of kinematical variables, α (lapse function) and $\{\beta^i\}_{i=1}^3$ (shift vector) which describe the time evolution of the coordinates.

3+1 split

$$\begin{aligned} ds^2 &= {}^{(4)}g_{\mu\nu}dx^\mu dx^\nu \\ &= -\left(\alpha^2 - g_{ij}\beta^i\beta^j\right)dt^2 + 2g_{ij}\beta^j dt dx^i + g_{ij}dx^i dx^j\beta^i \end{aligned} \tag{B.1}$$

with

$$g_{ij} = {}^{(4)}g_{ij}, \quad \alpha^2 = -\frac{1}{{}^{(4)}g^{00}}, \quad \beta_i = {}^{(4)}g_{0i}, \quad \beta^i = g^{ij}\beta_j = -\frac{{}^{(4)}g^{0i}}{{}^{(4)}g^{00}}$$

The (timelike) future-pointing unit vector normal to the slices is given by $\{n^\mu\}_{\mu=0}^3$ where $n^0 = 1/\alpha$ and $n^i = -\beta^i/\alpha$.

A 3-tensor quantity, (the extrinsic curvature), K_{ij} is introduced by

$$K_{ij} = -\frac{1}{2}\mathcal{L}_n g_{ij} = -\nabla_i n_j \tag{B.2}$$

It describes how the slices are embedded in the 4-dimensional spacetime, in other words, it measures the curvature of the 3-manifold, relative to that of the 4-geometry.

Then the Einstein equations can be cast in 2 subsystems of equations in unknowns the two sets of 3-tensors fields (g_{ij} and K_{ij}). The first subsystem (hyperbolic, 12 equations) is the *evolution system* (B.3), and it describes the time development of these tensors from one slice to another, while the second one (elliptic, 4 equations) is the *constraints system* (B.4) that must be obeyed by these quantities on each time slice. The Bianchi identities guarentee that the evolution system is compatible with the constraints system.

ADM evolution system

$$\begin{aligned} (\partial_t - \mathcal{L}_\beta)\, \gamma_{ij} &= -2\alpha K_{ij} \\ (\partial_t - \mathcal{L}_\beta)\, K_{ij} &= -D_i D_j \alpha + \alpha(R_{ij} + K K_{ij} - 2K_{ik}K^k{}_j), \end{aligned} \tag{B.3}$$

ADM constraints system

$$\begin{aligned} \mathcal{H} &\equiv R + K^2 - K_{ij}K^{ij} = 0, \\ \mathcal{D}^i &\equiv D_j(K^{ij} - \gamma^{ij}K) = 0. \end{aligned} \tag{B.4}$$

Here $\mathcal{L}_\beta$ is the Lie derivative with respect to the shift vector β^i, D_i is the covariant derivative associated with the 3-metric γ_{ij}, R_{ij} is the three-dimensional Ricci tensor, R the Ricci scalar, and K is the trace of K_{ij}.

Appendix C

BSSN equations

In order to deduce the BSSN formulation of Einstein equations, start from the ADM-equations, (B.3), and consider the following decomposition of the metric and extrinsic curvature:

$$\gamma_{ij} = e^{4\phi}\tilde{\gamma}_{ij} \,, \tag{C.1}$$

$$K_{ij} = e^{4\phi}\left(\tilde{A}_{ij} + \frac{1}{3}\tilde{\gamma}_{ij}K\right), \tag{C.2}$$

In these relations, ϕ is the conformal factor — chosen such that the conformal metric $\tilde{\gamma}_{ij}$ has unit determinant—, $K = \gamma^{ij}K_{ij}$ is the mean curvature and $\tilde{A}_{ij}$ is the conformal traceless extrinsic curvature. The conformal connection functions are introduced by:

$$\tilde{\Gamma}^i = \tilde{\gamma}^{jk}\tilde{\Gamma}^i_{jk}$$

where $\tilde{\Gamma}^i_{jk}$ is the Christoffel symbol of the conformal metric. If the determinant of the conformal 3-metric is one then $\tilde{\Gamma}^i = -\partial_j\tilde{\gamma}^{ij}$.

The BSSN evolution equations are:

$$\partial_0\phi = -\frac{\alpha}{6}K + \frac{1}{6}\partial_k\beta^k, \tag{C.3}$$

$$\partial_0\tilde{\gamma}_{ij} = -2\alpha\tilde{A}_{ij} + 2\tilde{\gamma}_{k(i}\partial_{j)}\beta^k - \frac{2}{3}\tilde{\gamma}_{ij}\partial_k\beta^k, \tag{C.4}$$

$$\partial_0 K = -e^{-4\phi}\left[\tilde{D}^i\tilde{D}_i\alpha - 2\partial_i\phi\cdot\tilde{D}^i\alpha\right] + \alpha\left(\tilde{A}^{ij}\tilde{A}_{ij} + \frac{1}{3}K^2\right) \tag{C.5}$$

$$\begin{aligned}\partial_0\tilde{A}_{ij} = {} & e^{-4\phi}\left[\alpha\tilde{R}_{ij} + \alpha R^{\phi}_{ij} - \tilde{D}_i\tilde{D}_j\alpha - 4\partial_{(i}\phi\cdot\tilde{D}_{j)}\alpha\right]^{TF} \\ & +\alpha K\tilde{A}_{ij} - 2\alpha\tilde{A}_{ik}\tilde{A}^k_j + 2\tilde{A}_{k(i}\partial_{j)}\beta^k - \frac{2}{3}\tilde{A}_{ij}\partial_k\beta^k\end{aligned} \tag{C.6}$$

$$\begin{aligned}\partial_0\tilde{\Gamma}^i = {} & \tilde{\gamma}^{kl}\partial_k\partial_l\beta^i + \frac{1}{3}\tilde{\gamma}^{ij}\partial_j\partial_k\beta^k + \partial_k\tilde{\gamma}^{kj}\cdot\partial_j\beta^i - \frac{2}{3}\partial_k\tilde{\gamma}^{ki}\cdot\partial_j\beta^j - 2\tilde{A}^{ij}\partial_j\alpha \\ & +2\alpha\left[(m-1)\partial_k\tilde{A}^{ki} - \frac{2m}{3}\tilde{D}^iK + m(\tilde{\Gamma}^i{}_{kl}\tilde{A}^{kl} + 6\tilde{A}^{ij}\partial_j\phi)\right],\end{aligned} \tag{C.7}$$

where $\partial_0 \equiv \partial_t - \beta^j\partial_j$. Here, all quantities with a tilde refer to the conformal three metric $\tilde{\gamma}_{ij}$, and the latter is used in order to raise and lower their indices. The expression $[...]^{TF}$ denotes the traceless part (with respect to the metric $\tilde{\gamma}_{ij}$) of the expression inside the parentheses, and

$$\begin{aligned}\tilde{R}_{ij} &= -\frac{1}{2}\tilde{\gamma}^{kl}\partial_k\partial_l\tilde{\gamma}_{ij} + \tilde{\gamma}_{k(i}\partial_{j)}\tilde{\Gamma}^k - \tilde{\Gamma}_{(ij)k}\partial_j\tilde{\gamma}^{jk} + \tilde{\gamma}^{ls}\left(2\tilde{\Gamma}^k_{l(i}\tilde{\Gamma}_{j)ks} + \tilde{\Gamma}^k_{is}\tilde{\Gamma}_{klj}\right), \\ R^{\phi}_{ij} &= -2\tilde{D}_i\tilde{D}_j\phi - 2\tilde{\gamma}_{ij}\tilde{D}^k\tilde{D}_k\phi + 4\tilde{D}_i\phi\,\tilde{D}_j\phi - 4\tilde{\gamma}_{ij}\tilde{D}^k\phi\,\tilde{D}_k\phi.\end{aligned}$$

The parameter m controls how the momentum constraint is added to the evolution equations for the variable $\tilde{\Gamma}^i$. The constraint system is :

$$H \equiv \frac{1}{2}\left(\gamma^{ij}R^{(3)}_{ij} + K^2 - K^{ij}K_{ij}\right) = 0, \tag{C.8}$$

$$M_i \equiv \tilde{D}^j\tilde{A}_{ij} - \frac{2}{3}\tilde{D}_iK + 6\tilde{A}_{ij}\tilde{D}^j\phi = 0, \tag{C.9}$$

$$C^i_{\Gamma} \equiv \tilde{\Gamma}^i + \partial_j\tilde{\gamma}^{ij} = 0\,, \tag{C.10}$$

Appendix D

SBP operators

The following SBP operators satisfy $\Delta \equiv A - Q^T \Sigma^{-1} Q > 0$:

- 2nd order in interior and 1st order at the boundary (Appendix C.1 [57])
- 4th order in interior and 2nd order at the boundary (Appendix C.2 [57])
- 6th order in interior and 3rd order boundary

$$\Sigma = \begin{pmatrix} \frac{13649}{43200} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{12013}{8640} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{2711}{4320} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{5359}{4320} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{7877}{8640} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{43801}{43200} \end{pmatrix}$$

$$BS = \begin{pmatrix} -\frac{25}{12} & 4 & -3 & \frac{4}{3} & -\frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$D_1 = \begin{pmatrix} -\frac{21600}{13649} & \frac{83096}{40947} & -\frac{10271}{81894} & -\frac{6477}{13649} & \frac{9875}{81894} & \frac{1333}{40947} & 0 & 0 & 0 & 0 \\ -\frac{83096}{180195} & 0 & \frac{3341}{12013} & \frac{19973}{72078} & -\frac{995}{12013} & -\frac{1351}{120130} & 0 & 0 & 0 & 0 \\ \frac{10271}{162660} & -\frac{3341}{5422} & 0 & \frac{4601}{8133} & \frac{191}{10844} & -\frac{821}{27110} & 0 & 0 & 0 & 0 \\ \frac{6477}{53590} & -\frac{19973}{64308} & -\frac{4601}{16077} & 0 & \frac{713}{1398} & -\frac{15287}{321540} & \frac{72}{5359} & 0 & 0 & 0 \\ -\frac{1975}{47262} & \frac{995}{7877} & -\frac{191}{15754} & -\frac{16399}{23631} & 0 & \frac{6048}{7877} & -\frac{1296}{7877} & \frac{144}{7877} & 0 & 0 \\ -\frac{1333}{131403} & \frac{1351}{87602} & \frac{821}{43801} & \frac{15287}{262806} & -\frac{30240}{43801} & 0 & \frac{32400}{43801} & -\frac{6480}{43801} & \frac{720}{43801} & 0 \\ 0 & 0 & 0 & -\frac{1}{60} & \frac{3}{20} & -\frac{3}{4} & 0 & \frac{3}{4} & -\frac{3}{20} & \frac{1}{60} \end{pmatrix}$$

$$D_2 = \begin{pmatrix} \frac{35}{12} & -\frac{26}{3} & \frac{19}{2} & -\frac{14}{3} & \frac{11}{12} & 0 & 0 & 0 & 0 & 0 \\ \frac{163526}{180195} & -\frac{77883}{48052} & \frac{14714}{36039} & \frac{30637}{72078} & -\frac{1552}{12013} & \frac{6611}{720780} & 0 & 0 & 0 & 0 \\ \frac{131}{54220} & \frac{7357}{8133} & -\frac{26717}{16266} & \frac{1290}{2711} & \frac{11237}{32532} & -\frac{3487}{40665} & 0 & 0 & 0 & 0 \\ -\frac{9143}{80385} & \frac{30637}{64308} & \frac{1290}{5359} & -\frac{46693}{32154} & \frac{13733}{16077} & -\frac{67}{4660} & \frac{48}{5359} & 0 & 0 & 0 \\ \frac{20539}{472620} & -\frac{1552}{7877} & \frac{11237}{47262} & \frac{27466}{23631} & -\frac{82147}{31508} & \frac{178774}{118155} & -\frac{1296}{7877} & \frac{96}{7877} & 0 & 0 \\ 0 & \frac{6611}{525612} & -\frac{6974}{131403} & -\frac{1541}{87602} & \frac{178774}{131403} & -\frac{1390165}{525612} & \frac{64800}{43801} & -\frac{6480}{43801} & \frac{480}{43801} & 0 \\ 0 & 0 & 0 & \frac{1}{90} & -\frac{3}{20} & \frac{3}{2} & -\frac{49}{18} & \frac{3}{2} & -\frac{3}{20} & \frac{1}{90} \end{pmatrix}$$

Bibliography

[1] Alcubierre, M., Brügmann, B.: "Simple excision of a black hole in 3+1 numerical relativity", Phys. Rev. D63, 104006 (2001).

[2] Arnowitt R., Deser S., and Misner, C.W.: "The Dynamics of General Relativity", in L. Witten, ed., Gravitation: An Introduction to Current Research, 227-265, Wiley, New York, U.S.A., (1962)

[3] Baker, J.G., Centrella, J., Choi, D.I., Koppitz, M. and van Meter, J.: "Gravitational wave extraction from an inspiraling configuration of merging black holes", Phys. Rev. Lett. 96 (2006) 111102.

[4] Baumgarte, T.W. and Shapiro, S.L.: "On the numerical integration of Einstein's field equations", Phys. Rev. D59 024007 (1999).

[5] Beyer, H., Sarbach, O.: "On the well posedness of the Baumgarte-Shapiro-Shibata-Nakamura formulation of Einstein's field equations" Phys.Rev D70 104004 (2004)

[6] Bondi, H.: "Gravitational waves in general relativity", Nature, 186, 535-535, (1960).

[7] Bondi, H., van der Burg, M.J.G., and Metzner, A.W.K.: "Gravitational waves in general relativity VII. Waves from axi-symmetric isolated systems", Proc. R. Soc. London, Ser. A

[8] Boyle, M. *et al.*: "High-accuracy comparison of numerical relativity simulations with post-Newtonian expansions," Phys. Rev. D76 124038 (2007).

[9] Brown, J.D.: "BSSN in spherical symmetry", Class. Quantum Grav. 25 205004 (2008)

[10] Butcher, J.C.: "The Numerical Analysis of Ordinary Differential Equations: Runge–Kutta and General Linear Methods", Wiley (1987).

[11] Campanelli, M., Lousto, C.O., Marronetti, P. and Zlochower, Y.: "Accurate evolutions of orbiting black-hole binaries without excision," Phys. Rev. Lett. 96 111101 (2006).

[12] Campanelli, M., Lousto, C.O., Nakano, H. and Zlochower, Y.: "Comparison of Numerical and Post-Newtonian Waveforms for Generic Precessing Black-Hole Binaries", arXiv:0808.0713 [gr-qc].

[13] Calabrese, G., Lehner, L., Tiglio, M.: "Constraint-preserving boundary conditions in numerical relativity" Phys.Rev. D65-104031 (2002).

[14] Calabrese, G., Hinder, I., Husa, S.: "Numerical stability for finite difference approximations of Einstein's equations", J. Comput. Phys 218 607 (2006).

[15] Calabrese, G., "Finite differencing second order systems describing black hole spacetimes", Phys. Rev. D71 027501 (2005).

[16] Calabrese, G., Gundlach, C.: "Discrete boundary treatment for the shifted wave equation in second-order form and related problems" Class. Quantum Grav. 23, S343-S367 (2006).

[17] Carpenter, M.H., Gottlieb, D., Abarbanel, S.: "The stability of numerical boundary treatments for compact high-order finite-difference schemes", J. Comput. Phys. 108 (2) (1994).

[18] Cartwright, J.H.E. and Piro, O.: "The Dynamics of Runge-Kutta Methods.", Int. J. Bifurcations Chaos 2, 427-449 (1992).

[19] Chirvasa, M., Husa, S.: "Discretization of the Cauchy problem for second order in space, first order in time systems using high order finite difference operators", arXiv:0812.3752 [gr-qc] (2009).

[20] Christodoulou, D.: "The Action Principle and Partial Differential Equations", Princeton University Press, Princeton (2000).

[21] Damour, T., Nagar, A., Hannam, M., Husa, S. and Brügmann, B.: "Accurate Effective-One-Body waveforms of inspiralling and coalescing black-hole binaries", Phys. Rev. D78: 044049 (2008).

[22] Diener,P., Dorband N., Schnetter E., Tiglio M.: "New, efficient, and accurate high order derivative and dissipation operators satisfying summation by parts, and applications in three-dimensional multi-block evolutions", J. Sci. Comput.32:109-145 (2007).

[23] Fornberg, B., "Calculation of weights in finite difference formulas": SIAM Review 40 No. 3, 685-691 (1998).

[24] Frauendiener, J.: "Conformal Infinity", Living Rev. Relativity 7, (January 2004), [Online Journal Article]:cited on February 2004, http://www.livingreviews.org/lrr-2004-1.

[25] Frauendiener, J.: "Numerical treatment of the hyperboloidal initial value problem for the vacuum Einstein equations. I. The conformal field equations", Phys. Rev. D58, 064002 (1998).

[26] Friedrich, H.: "Cauchy problems for the conformal vacuum field equations in general relativity", Commun. Math. Phys. 91 No 4 (1983).

[27] Friedrich, H.: "Gravitational fields near space-like and null infinity", Journal of Geometry and Physics 24, 83 (1998).

[28] Friedrich, H.: "Einstein's Equation and Geometric Asymptotics". In: Proceedings of the GR-15 conference. ed. by N. Dadhich, J. Narlikar (IUCAA, Pune, 1998).

[29] Friedrich, H. and Nagy, G.: "The initial boundary value problem for Einsteins vacuum field equations", Commun. Math. Phys. 201 619 - 655 (1999).

[30] Friedrich, H. and Rendall, A.D.: "The Cauchy Problem for the Einstein Equations", Lect. Notes Phys. 540 127 (2000).

[31] Gundlach, C. and Martin-Garcia, J.M.: "Symmetric hyperbolicity and consistent boundary conditions for second-order Einstein equations", Phys.Rev D70 044032 (2004).

[32] Gundlach, C. and Martin-Garcia, J.M.: "Symmetric hyperbolic form of systems of second-order evolution equations subject to constraints", Phys. Rev. D70, 044031 (2004).

[33] Gundlach, C. and Martin-Garcia, J.M.: "Hyperbolicity of second-order in space systems of evolution equations", Class.Quant.Grav. 23 S387-S404, (2006).

[34] Gundlach, C. and Martin-Garcia, J.M.: "Well-posedness of formulations of the Einstein equations with dynamical lapse and shift conditions," Phys. Rev. D74 024016 (2006).

[35] Gustafsson, G., Kreiss, H.O. and Sundström, A.: "Stability theory of difference approximations for mixed initial boundary value problems", Math. Comp. 26, 649-68 (1972).

[36] Gustafsson, G., Kreiss, H.O. and Oliger, J.: "Time dependent problems and difference methods", J.Wiley&Sons, New-York (1995).

[37] Hadamard, J.: "Sur les problèmes aux dérivées partielles et leur signification physique", Princeton University Bulletin, 49–52 (1902).

[38] Hannam, M., Husa, S., Brügmann, B. and Gopakumar, A.: "Comparison between numerical-relativity and post-Newtonian waveforms from spinning binaries: the orbital hang-up case", arXiv:0712.3787 [gr-qc].

[39] Hannam, M., Husa, S., Pollney, D., Brügmann, B. and O'Murchadha, N.: "Geometry and Regularity of Moving Punctures", Phys. Rev. Lett. 99 241102 (2007).

[40] Henrici, P.: "Discrete Variable Methods in Ordinary Differential Equations", Wiley (1962).

[41] Hinder, I., Herrmann, F., Laguna P. and Shoemaker, D.: "Comparisons of eccentric binary black hole simulations with post-Newtonian models", arXiv:0806.1037 [gr-qc].

[42] Husa, S., Gonzalez, J.A., Hannam, M., Brügmann, B. and Sperhake, U.: "Reducing phase error in long numerical binary black hole evolutions with sixth order finite differencing", arXiv:0706.0740 [gr-qc].

[43] Hübner, P.: Class. Quantum Grav. 16, 2145 (1999).

[44] Kreiss, H.O. and Scherer, G.: "Finite element and finite difference methods for hyperbolic partial differential equations", Mathematical Aspects of Finite Elements in Partial Differential Equations, Academic Press, New York (1974).

[45] Kreiss, H.O. and Scherer, G.: "On the existence of energy estimates for difference approximations for hyperbolic systems", Technical Report, Department of Scientific Computing, Uppsala University (1977).

[46] Kreiss, H.O. and Ortiz, O.E.: "Some Mathematical And Numerical Questions Connected With First And Second Order Time Dependent

Systems Of Partial Differential Equations", Lect.Notes Phys. 604 359 (2002).

[47] Kreiss, H.O. and Wu, L.: "On the stability definition of difference approximations for the initial boundary value problem", Applied Numer. Math., Vol.12, 213-227 (1993).

[48] Kreiss, H.O. and Scherer, G.: "Method of lines for hyperbolic differential equations", SIAM Journal on Numerical Analysis Volume 29, Issue 3 (1992).

[49] Kreiss, H.O. and Winicour, J.: "Problems which are well-posed in a generalized sense with applications to the Einstein equations", Class. Quantum Grav. 23 S405-20 (2006).

[50] Kreiss, H.O., Reula, O., Sarbach, O. and Winicour, J.: "Boundary conditions for coupled quasilinear wave equations with application to isolated systems ", Commun. Math. Phys. 289:1099-1129 (2009).

[51] Lambert, J. D., "Numerical Methods for Ordinary Differential Systems", Wiley (1991).

[52] Lindblom, L., Scheel, M., Kidder, L., Pfeiffer, H., Shoemaker, D., Teukolsky, S.: "Controlling the Growth of Constraints in Hyperbolic Evolution Systems", Phys.Rev. D69 124025 (2004).

[53] Kidder, L., Lindblom, L., Scheel, M., Buchman, L., Pfeiffer, H.: "Boundary Conditions for the Einstein Evolution System", Phys.Rev. D71 064020 (2005).

[54] Lindelöf, M.E.: "Sur l'application de la méthode des approximations successives aux équations différentielles ordinaires du premier ordre"; Comptes rendus hebdomadaires des séances de l'Académie des sciences, Vol. 114, pp. 454 (1894).

[55] Lubich, C. and Nevanlinna, O.: "On resolvent conditions and stability estimates", BIT 31, 2, 293-313 (1991).

[56] Luther, H.A.: "An Explicit Sixth-Order Runge-Kutta Formula", Mathematics of Computation, Vol.22, No.102, pp434-436 (1968).

[57] Mattsson, K., Nordström, J.: "Summation by parts operators for finite difference approximations of second derivatives", J. Comput. Phys. 199, 503-540 (2004).

[58] Motamed, M., Babiuc, M., Szilágyi, B., Kreiss, H.O., Winicour, J.: "Finite difference schemes for second order systems describing black holes", Phys.Rev. D73 124008 (2006).

[59] Nagy, G., Ortiz, O. and Reula, O.: "Strongly hyperbolic second order Einstein's evolution equations", Phys. Rev. D70, 044012 (2004).

[60] Olsson P.: "Summation by parts, projections, and stability. I", Math. Comput. 64 1035 (1995).

[61] Olsson, P.: "Summation by parts, projections, and stability. II", Math. Comput. 64 1473 (1995).

[62] Pazos, E., Tiglio M., Duez M., Kidder L., Teukolsky, S.: "Orbiting binary black hole evolutions with a multipatch high order finite-difference approach", gr-qc/09040493 (2009).

[63] Penrose, R.: "Asymptotic properties of fields and space-times", Phys. Rev. Lett. 10, 66 - 68 (1963).

[64] Pollney, D. et al: "Recoil velocities from equal-mass binary black-hole mergers: A systematic investigation of spin-orbit aligned configurations", Phys. Rev. D 76 124002 (2007).

[65] Polyanin A.D. and Manzhirov, A.V.: "Handbook of Mathematics for Engineers and Scientists", Chapman & Hall/CRC Press, Boca Raton-London (2007).

[66] Pretorius, F.: "Evolution of binary black hole spacetimes", Phys. Rev. Lett. 95 121101 (2005).

[67] Reddy, S.C. and Trefethen, L.N.: "Stability of the method of lines", Numerische Mathematik, Volume 62, Number 1 (1992).

[68] Ruiz, M., Rinne, O. and Sarbach, O.: "Outer boundary conditions for Einstein's field equations in harmonic coordinates ", Class.Quant.Grav.24:6349-6378 (2007).

[69] Sarbach, O., Calabrese, G., Pullin, J. and Tiglio, M.: "Hyperbolicity of the Baumgarte-Shapiro-Shibata-Nakamura system of Einstein evolution equations", Phys. Rev. D 66, 064002 (2002).

[70] Sarbach, O. and Tiglio, M.: "Boundary conditions for Einstein's field equations: Analytical and numerical analysis ", J.Hyperbol.Diff.Equat. 2-839 (2005).

[71] Scheel, M.A., Boyle, M., Chu, T., Kidder, L.E., Matthews, K.D. and Pfeiffer,H.P.: "High-accuracy waveforms for binary black hole inspiral, merger, and ringdown", arXiv:0810.1767 [gr-qc].

[72] Schmitt, K.: "Nonlinear Analysis and Differential Equations: An Introduction", http://www.freescience.info.

[73] Shibata, M. and Nakamura, T.: "Evolution of three-dimensional gravitational waves: Harmonic slicing case", Phys. Rev. D52 5428 (1995).

[74] Strand, B.: "Summation by parts for finite difference approximations for d/dx", J. Comput. Phys. Volume 110, Issue 1 (1994).

[75] Strikwerda, J.C.: "Finite Difference Schemes and Partial Differential Equations", Pacific Grove, CA: Wadsworth and Brooks (1989).

[76] Szilágyi, B., Pollney, D., Rezzolla, L., Thornburg, J. and Winicour, J.: "An explicit harmonic code for black-hole evolution using excision", Class. Quant. Grav. 24 S275 (2007).

[77] Szilágyi, B., Kreiss, H.O., Winicour, J.: " Modeling the black hole excision problem", Phys. Rev. D 71, 104035 (2005).

[78] Trefethen, L.N.: "Finite Difference and Spectral Methods for Ordinary and Partial Differential Equations" "Stability of the method of lines" unpublished text available at http://www.comlab.ox.ac.uk/nick.trefethen/pdetext.html (1996).

[79] Verner, H.: "Explicit Runge–Kutta methods with estimates of the Local Truncation Error", SIAM NA 772-790 (1978).

[80] Wald, R.M.: *General Relativity*, Univ. Chicago Press (1984).

[81] Winicour, J.: "Characteristic evolution and matching", Living Rev. Relativity 4, (March, 2001), [Online Journal Article]: cited on 23 July 2003, http://www.livingreviews.org/lrr-2001-3.

[82] York, J.W. in *Sources of Gravitational Radiation*, Smarr, L. (ed.), Cambridge University Press (1979).

[83] Zlochower, Y., Baker, J.G., Campanelli, M. and Lousto, C.O.: "Accurate black hole evolutions by fourth-order numerical relativity", Phys. Rev. D72, 024021 (2005).