Applied Mathematics of space-time & space+time: Problems in General Relativity and Cosmology

Céline Cattoën

Supervisor: Prof. Matt Visser

A thesis submitted to the Victoria University of Wellington
in fulfilment of the requirements for the degree of
Doctor of Philosophy in Mathematics.

April 30, 2009

School of Mathematics, Statistics, and Computer Science
VICTORIA UNIVERSITY OF WELLINGTON
Te Whare Wānanga o te Ūpoko o te Ika a Māui

New Zealand

Abstract

Cosmography is the part of cosmology that proceeds by making minimal dynamic assumptions. That is, one does not assume the Friedmann equations (Einstein equations) unless and until absolutely necessary. On the other hand, cosmodynamics is the part of cosmology that relates the geometry to the density and pressure using the Friedmann equations. In both frameworks, we consider the amount of information and the nature of the constraints we can obtain from the Hubble flow in a FLRW universe. Indeed, the cosmological parameters contained in the Hubble relation between distance and redshift provide information on the behaviour of the universe (expansion, acceleration etc...). In the first framework, it is possible to concentrate more directly on the observational situation in a model-independent manner. We perform a number of inter-related cosmographic fits to supernova datasets, and pay particular attention to the extent to which the choice of distance scale and manner of representing the redshift scale affect the cosmological parameters. In the second framework, we use the class of w-parameter models which has become increasingly popular in the last decade. We explore the extent to which a constraint on the w-parameter leads to useful and non-trivial constraints on the Hubble flow in terms of cosmological parameters $H(z)$, density $\rho(z)$, density parameter $\Omega(z)$, distance scales $d(z)$, and lookback time $T(z)$.

On another front, Numerical Relativity has experienced many breakthroughs since 2005, with full inspiral-merger-ringdown simulations now possible. One of the main goals is to provide very accurate templates of gravitational waves for ground-based and space-based interferometers. We explore the potential of a very recent and accurate numerical method, the Spectral Element Method (SEM), for Numerical Relativity, by treating a singular Schwarszchild black hole evolution as a test case. Spectral elements combine the theory of spectral and pseudo-spectral methods for high order polynomials and the variational formulation of finite elements and the associated geometric flexibility. We use the BSSN formulation of the Einstein equations with the method of the moving punctures. After applying the variational formulation to the BSSN system, we present several possible weak forms of this system and its spectral element discretization in space. We use a Runge-Kutta fourth order time discretization. The accuracy of high order methods can deteriorate in the presence of discontinuities or sharp gradients. We show that we can treat the element that contains the puncture with a filtering method to avoid artificial and spurious oscillations. These might form and propagate into the domain coming from discontinuous initial data from the BSSN system.

LaTeX-ed Friday, August 7, 2009; 12:12pm

Acknowledgement

I cannot thank my supervisor Prof. Matt Visser enough for his assistance, guidance, friendship and support over the last couple of years. Conversations, discussions and explanations have been of a tremendous help to guide me through all my work and my life. I am most appreciative of the scientific guidance and the support and encouragement to develop my own scientific independence which I believe to be extremely important in research. I am forever in debt for that and so many other things.

I would like to thank Dr Mark Hannam. His knowledge, passion and enthusiasm for Numerical Relativity has been very contagious and a great source of inspiration and motivation for me. My three-week visit in Cork in april 2008 has been very fruitful in many ways. I have never felt like a graduate student, but more as a collaborator and friend. I will always remember all the late 3pm lunches resulting from a complete loss of track of time due to incredibly long and fascinating discussions.

I would like to give some special thanks to Petarpa Boonserm; most importantly for her friendship but also for her consistent and perpetual support during hard times and good times. She has helped me through so much so many times. It was a real pleasure to travel to conferences with her and most importantly share an office for so long.

I would like to thank Silke Weinfurtner for numerous interesting discussions relating to many scientific and casual topics. A very good friend and great inspiration.

During the course of this research I have benefited from systematic or occasional stimulating discussions with a number of people besides those mentioned above. I would like to express my gratitude to Gabriel Abreu and Jozef Skakala especially during the last few months.

I wish to acknowledge the staff of the School of Mathematics, Statistics, and Computer Science for providing me with office space and all the facilities. In particular, I am very grateful to Roger Cliffe for his help on numerous occasions during computer breakdowns. I would also like to thank Dr. Monique Cano-Damitio for sharing her knowledge of the subtleties of LaTeX.

I am also very grateful to the Victoria PhD Scholarship and the Marsden fund that have supported me financially during my work at Victoria University of Wellington and related travel and conferences.

Initial simulations were run on my first Macintosh laptop G4 which unfortunately did not long survive after the transition from the 3D wave equation to the full BSSN system simulations. Its replacement, a Macintosh laptop G5 has been barely switched off in the last 6 months. Finally but not least, "Jimmy" with its incredible 8GB of RAM and physically based on the other side of the world in Dr Mark Hannam's office, Ireland, has contributed many clock-cycles.

Finally, I would like to thank all my family for all their care and support over many years, and for just being there for me.

I am deeply grateful to Ian Gilbert for everything.

Contents

List of Figures

List of Tables

"I like mathematics because it is not human and has nothing particular to do with this planet or with the whole accidental universe – because, like Spinoza's God, it won't love us in return."

Bertrand Russell (1872–1970)

1 Introduction

This thesis mainly deals with problems in General Relativity and Cosmology, however, as indicated in the title, "Applied Mathematics of space-time and space+time: Problems in General Relativity and Cosmology", Applied Mathematics plays a huge part in this work. Mathematics offers wonderful tools that allows one to explore and compare reality with physics. Some of these tools (non-exhaustive) used here include: statistics, Taylor series, convergence, data fitting, integrals, inequalities, functional analysis, numerical analysis and numerical methods, polynomials, differential geometry and so on. Many different mathematical topics it may seem, but they are all related and linked together here to attack two main problems in Einstein's theory of gravitation. The "***space-time***", refers to Cosmology in a traditional $4D$ spacetime description of General Relativity, whereas, "***space+time***" emphasizes the time and space splitting of spacetime used in Numerical Relativity.

The first part of this thesis treats topics in General Relativity and Cosmology, most of these investigations have been conducted as a collaborative work with my superviser Matt Visser, whereas the second part considers the application of a very recently developed numerical method to Numerical Relativity, this part of the thesis is the result of a collaboration with Mark Hannam.

Regarding cosmology, what amount of information or constraints can one obtain from the Hubble flow in a FLRW (Friedmann-Lemaître-Robertson-Walker) universe? How general, precise, and useful, can results be under a minimum of theoretical assumptions? These are the key questions that motivate the first part of this thesis.

Chapter 2 introduces some of the main and basic notions of modern Cosmology in a FLRW universe. A ***F***riedmann-***L***emaître-***R***obertson-***W***alker universe relies on the Copernican principle of isotropy (direction independence) and homogeneity (position independence) of our universe. We will introduce the cosmological parameters whose values prescribe the behaviour of the universe, as well as standard definitions of cosmological distances. There are many notions of distance scale in Cosmology, which one should one use? We will clarify these concepts and introduce some new definitions leading to alternative Hubble laws. We will see how we can extract information on the cosmological parameters using the ***Supernovae type Ia*** data ("SNIa").

Chapter 3 discusses and presents results obtained in the context of Cosmography, that is, without assuming the Einstein field equations. In this framework, we minimize the

number of physics assumptions that go into the model. Is the expansion of the universe still accelerating in this context? What happens when considering realistic estimates of systematic uncertainties (based on the published data)? Moreover, can we obtain values for other cosmological parameters and therefore further characterize the behaviour of the universe? We will explore the aforementioned questions and try to answer them in this chapter.

Chapter 4 presents results developed within the framework of Cosmodynamics, where general relativity is now assumed. With these further assumptions, we can use the classical energy conditions to place very general and robust bounds on various cosmological parameters, and thereby get a qualitative and quantitative insight on how strange physics gets. Are the various energy conditions and their associated bounds on the cosmological parameters inter-related? Is a systematic and exhaustive analysis possible? Confronting some of these bounds with the supernova data, can we say anything concerning the universe? In the absence of any detailed understanding of the precise nature of the cosmological equation of state $\rho(p)$, just how much can be deduced with limited information? We will see that in fact, we can obtain even more general bounds by just assuming a general equation of state of the form $p = w\rho$.

In Numerical relativity, there have been many breakthroughs since 2005, with full inspiral-merger-ringdown simulations now possible. One of the main goals is to provide very accurate templates of gravitational waves for ground-based and space-based interferometers to detect. What is the potential of the Spectral Element Method for Numerical Relativity? Would this method allow for better accuracy and efficiency, and possibly contribute to gravitational wave detection?

Moving on to the second part of this thesis, we summarize basic notions of Numerical Relativity in Chapter 5. We outline the splitting of space and time necessary for numerical simulations, as well as stable reformulations of the Einstein equations in this context. We quickly describe the moving puncture method and BSSN formulation of the field equations.

Chapter 6 is devoted to the introduction of the Spectral Element Method (SEM) in a very general context. We discuss how to obtain the weak formulation from the variational formulation of a given problem. We show how the domain can be discretized into elements, and the numerical solution approximated with Lagrange-Legendre basis functions. We present the assembly process over each element to form a final global system of algeabric equations. We also mention some very general and powerful theorems of existence and uniqueness of a solution.

Chapter 7 illustrates the spectral element method in practice with a 1D and 3D wave equations. We show in much detail the SEM formulations and final discretized matrix systems obtained. We then present some numerical results one can achieve with this method.

In Chapter 8, we apply the SEM to the BSSN system with moving punctures. We start with the strong formulation of the system, and through the variational formulation obtain a weak formulation. After domain and elemental discretization, we describe the elemental matrix system. Consequent numerical results are presented in Chapter 9.

Finally, Chapter 10 contains the conclusions of this work, as well as suggestions and remarks about future work in respective fields.

Part I

General Relativity and Cosmology

κοσμολογια, *cosmology*
κοσμος, *kosmos, universe*
λογια, *logia, study*

2

Introduction to Cosmology in a FLRW universe

Cosmology is the study of the dynamical structure of the universe on the largest scales of space and time, considered as a whole.

Contemporary cosmological models are based on the idea that the universe is, on average, the same overall. That is, when describing the Universe as a whole one assumes that it is filled with a continuous medium (fluid, gas or radiation). This is based on a very simple principle, called ***the cosmological principle***, which is a generalization of the Copernican principle:

The cosmological principle: at each epoch, the universe presents the same aspect from every point, except for "***small***" local irregularities [7]. When averaged over sufficiently large volumes the universe and the matter in the universe should be *isotropic* and *homogeneous*.

- *Isotropy* states that space looks the same no matter what direction one looks at (direction independence).
- *Homogeneity* is the statement that the metric is the same throughout the space (position independence).

Astronomical observations suggest that the universe is homogeneous and isotropic when viewed on the largest scales. Figure 2.1 is a good illustration of homogeneity, the photo taken from the Hubble telescope covers an area 2.5 arcminutes across, two parts in a million of the whole sky, which is equivalent in angular size to a 65 mm tennis ball at a distance of 100 metres. The image was assembled from 342 separate exposures taken with the Space Telescope's Wide Field and Planetary Camera 2 over ten consecutive days between December 18 and December 28, 1995. Traditionally this homogeneity has been assumed up to ***small*** fluctuations that are ***large enough*** to include clusters of galaxies. Note that figure 2.1 is not an observational proof of homogeneity in any sense: The cosmological principle still has ***no direct*** observational verification at present. Indeed, homogeneity is a very crude assumption as the real Universe has more of a granular structure. The scale at which homogeneity sets in, is still not completely certain. Voids with diameters of order 10^8 light years are ubiquitous, forming at least 40% of the volume of the universe [8], [9], and are typically surrounded by bubble walls containing galaxy clusters. The largest feature observed (the Sloan Great Wall [10]) is 1.47×10^9 light years long. In the history of Cosmology there has been a constant evolution of the definition of the size of the elementary unit to save the assumption of homogeneity and isotropy of the Universe in the large. Hence we simply assume homogeneity for some ***suitable defined cell size***. Note that the universe is assumed

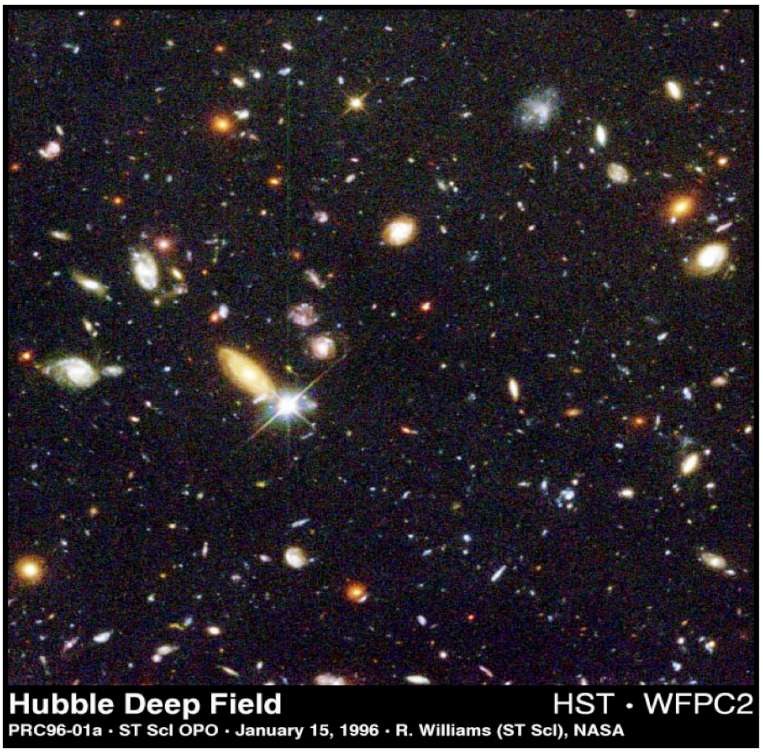


Figure 2.1: Small region in the constellation Ursa Major, montage constructed from a series of observations by the Hubble Space Telescope.

to be ***spatially*** homogeneous and isotropic at each instant of cosmic time, specifically, we are talking about homogeneity on each one of the 3-dimensional ***space-like hypersurfaces***.

When looking at distant galaxies, they seem to be receding from our galaxy. It appears that the universe is not static, but changing with time. Thus most cosmological models are built on the fact that the universe is homogeneous and isotropic in space, but not in time. Observationally, the universe today is significantly different from the universe of 10^{10} years ago, and radically different from the universe of 1.5×10^{10} years from now.

The realization that the Universe may have a ***history***, has been considered very gradually by astronomers and physicists over time. The expansion of the Universe was discovered by Hubble in 1929, and the expansion itself had already started to be described by the Friedman (1922) and Lemaître (1927) solutions to Einstein's equations. Einstein added a cosmological constant to his theory to try to force it to allow for a static universe with matter in it. However, the Einstein universe is unstable. Since then, cosmologists have been trying to reconstruct the possible sequence of events in the evolution of the Universe. The FLRW (Friedmann-Lemaître-Robertson-Walker) spacetime is the foundation stone of the standard Big Bang theory. Some of the main characteristics are that this metric satisfies the Einstein field equations for a perfect fluid, and is isotropic and homogeneous. The FLRW model has some clear successes, fitting the Hubble law (reasonably well), the cosmic microwave spectrum (to very high accuracy), and being compatible with the measured isotropies and the assumption of homogeneity. However, this standard model also has some weaknesses. For

example it lacks a full description of the formation of large-scale structure, and sometimes the inflation scenario introduces new problems, instead of solving all the uncertainties inherent in the naive FLRW universe.

2.1 A brief history of the Universe

Observations suggest that the universe is approximately 13.7 billion years old. The history of the universe is divided into different periods called epochs, according to the dominant forces and processes in each period. The main outline of the history of the universe has 3 main phases.

After the initial Big Bang explosion, the universe was very hot until at least 10^{-36} seconds. Between 10^{-36} seconds and 10^{-32} seconds after the Big Bang, there was a period of exponential growth called ***cosmic inflation***. Near the end of cosmological inflation, the universe was then cold and empty, and the immense heat and energy associated with the early stages of the big bang was re-created through the phase change associated with the end of inflation, through a reheating period. The very early universe was the split second in which the universe was so hot that particles had energies higher than those currently accessible in particle accelerators on Earth. Note that the details are largely based on educated guesses. The evolution of the universe then proceeded according to the known rules of general relativity and high energy physics.

The ***last scattering epoch*** occurred somewhere between 300 000 and 380 000 years after the Big Bang[1]. This is where hydrogen atoms appeared. Note that nuclei were formed earlier at the period of nucleosynthesis, whereas the first protons, electrons and neutrons were formed at reheating. With the formation of neutral hydrogen, the cosmic microwave background was emitted. Hydrogen and helium are at the beginning ionized, as the universe cools down, the electrons get captured by the ions making them neutral. This process is known as recombination, the photons can now travel freely, which means that the universe has become transparent. The photons emitted right after the recombination, that can therefore travel undisturbed, are those that we see in the cosmic microwave background (CMB) radiation. Therefore the CMB is a picture of the universe at the end of this epoch.

Finally, the ***epoch of structure formation*** began, when matter started to aggregate into the first stars and quasars, and ultimately galaxies, clusters of galaxies and superclusters formed. The oldest identified quasar (CFHQS 1641+3755) is at 12.7 billion light-years away. Our solar system was formed approximately 8 billion years ago.

The ultimate fate of the universe is not known, there are several scenarios but according to the standard ΛCDM model (see section 2.9 for more details), it will continue expanding forever.

[1] If last scattering is deemed to "begin" when the ionization fraction has dropped to 10% (a standard definition of recombination) then this occurs at $z \sim 1250$, whereas photon decoupling (which might be deemed to be the "end" of last scattering) occurs at $z = 1090$. Using cosmological parameters for the concordance cosmology as given by Komatsu ***et al*** [11], one is led to the expansion ages quoted above, using the exact solution for a spatially flat universe with matter and radiation (Ω_Λ being negligibly small at that epoch).

2.2 Cosmography

Simply by using the assumptions of isotropy and homogeneity, a cosmological model can be derived without yet using the Einstein equations. This homogeneous and isotropic cosmological model is called the Friedmann–Lemaître–Robertson–Walker (FLRW) geometry (2.1), and is given by:

$$ds^2 = -dt^2 + a(t)^2 \left\{ \frac{dr^2}{1-kr^2} + r^2 \left[d\theta^2 + \sin^2\theta \, d\phi^2 \right] \right\} \tag{2.1}$$

where $a(t)$ is the scale factor of the universe. There are only three values of interest for the parameter k:

- $k = -1$, this corresponds to a negative curvature (for the hyperboloid);
- $k = \ \ 0$, this corresponds to no curvature (flat space);
- $k = +1$, this corresponds to a positive curvature (for the 3-sphere).

That is, the assumptions of homogeneity and isotropy alone have determined the spacetime metric up to three discrete possibilities of spatial geometry k and the arbitrary positive function of the scale factor $a(t)$.

Observational evidence strongly suggests that our universe (or the part of our universe within our causal past), is well described by a Friedmann–Lemaître–Robertson–Walker model, and indeed a $k = 0$ model, at least as far back in time as the decoupling of matter and radiation.

For recent work on cosmographic analyses see [12, 13, 14, 15, 16].

2.3 Cosmodynamics

Now, by substituting the spacetime metric (2.1) into Einstein's equations (2.2):

$$G_{ab} = \frac{8\pi G_N}{c^2} T_{ab}, \tag{2.2}$$

some predictions for the dynamical evolution of the system can be obtained.

But first, we need to describe the matter content of the universe in terms of the stress-energy tensor. Using the assumptions of isotropy and homogeneity, the stress-energy tensor of matter in the present universe is approximated in an orthonormal frame by:

$$T^{\hat{a}\hat{b}} = \begin{bmatrix} \rho & 0 & 0 & 0 \\ 0 & p & 0 & 0 \\ 0 & 0 & p & 0 \\ 0 & 0 & 0 & p \end{bmatrix}. \tag{2.3}$$

Here ρ and p are the average density and pressure due to the galaxies, stars, clouds of dusts and so on.

Applying the Einstein equations imply the two following Friedmann equations (in the context of a FLRW universe):

$$8\pi G_N \rho = 3\left[\frac{\dot{a}^2}{a^2} + \frac{k}{a^2}\right] \tag{2.4}$$

$$8\pi G_N p = -\left[\frac{\dot{a}^2}{a^2} + \frac{k}{a^2} + 2\frac{\ddot{a}}{a}\right]. \tag{2.5}$$

And consequently, equation (2.4) and equation (2.5) imply:

$$8\pi G_N \left[\rho + 3p\right] = -6\frac{\ddot{a}}{a}. \tag{2.6}$$

The Friedmann equations completely specify the evolution of the universe as a function of time. By imposing homogeneity and isotropy as characteristics of the universe that remain unchanged with time (on suitably large scales), we have implicitly restricted any evolution to affect only one remaining characteristic: its size $a(t)$. The Friedmann equations are therefore equations for the scale factor $a(t)$, ultimately measuring the evolution of the size of ***any freely expanding length scale*** (gravitational, electromagnetic, etc...) in the universe.
Equation 2.4 tells us about the velocity of the expansion and or contraction of the universe, being an equation in $\dot{a}$. On the other hand, equation 2.5 involves $\ddot{a}$ which tells us about the acceleration/deceleration of the expansion or contraction.
In equation 2.4, if $k = 0$, the universe is spatially flat and equation 2.4 implies that it has to become infinite with the density ρ approaching zero, in order for the expansion to stop. If $k = 1$, the expansion can stop at a finite density at which the matter contribution is balanced by the space curvature. Thus, at a finite time, the universe will stop expanding and will recollapse. Finally, if $k = -1$, the universe will continue to expand forever even if matter is completely dispersed.
Notice that the spatial curvature is completely absent from 2.6, this means that the acceleration/deceleration of the expansion or contraction of the universe is ***independent*** of the spatial curvature k. This equation also reveals that gravity is always an attractive force.

The difficulty remains to detemermine a suitable matter model for ρ and p, that is to make even more progress, it is necessary to choose an ***equation of state*** between ρ and p.

2.4 Cosmological parameters

This section introduces some of the basic terminology associated with the cosmological parameters. The values of these parameters describe the behaviour of the universe. It is standard terminology in mechanics that the first four time derivatives of position are referred to as velocity, acceleration, jerk and snap. We will see that we have very similar definitions in cosmology for the cosmological parameters.

2.4.1 The Hubble parameter

The ***rate of expansion*** is characterized by the ***Hubble parameter***:

$$H(t) = +\frac{1}{a}\frac{\mathrm{d}a}{\mathrm{d}t} = \frac{\dot{a}}{a}. \tag{2.7}$$

The Hubble parameter quantifies the ***speed*** with which the size of the universe is increasing. The value of the Hubble parameter at the present epoch is the Hubble constant H_0. There has been a great deal of controversy about what its actual value is, but currently, the consensus measurements [17] (2006) give:

$$H_0 = 73\ {}^{+3}_{-4}\ (\text{km/sec})/\text{Mpc}, \tag{2.8}$$

where ***Mpc*** stands for ***megaparsec***, $1\ \text{Mpc} \cong 3 \times 10^{24}\text{cm}$. One parsec is $3.08 \times 10^{18}\text{cm} = 3.26$ light years.

The universe is expanding, therefore we know that $\dot{a} > 0$. From equation (2.6) we also know that $\ddot{a} < 0$ when assuming that the pressure p and the density ρ are both positive. The universe must have been expanding at a faster and faster rate when going back in time. If we consider that the universe has always been expanding at the present rate, then at the time $T = H^{-1} = a/\dot{a}$ ago, the scale factor a would be null, $a = 0$. However, the expansion rate was actually faster, therefore, the time at which $a = 0$ was even closer to the present. By assuming homogeneity and isotropy, general relativity makes the prediction that at a time less than H^{-1} ago, the universe was in a singular state. This singular point referred to as the Big Bang had an infinite density of matter and an infinite curvature of spacetime.

2.4.2 The deceleration parameter

The value of the Hubble parameter changes over time either increasing or decreasing depending on the sign of the ***deceleration parameter***:

$$q(t) = -\frac{1}{a}\frac{\mathrm{d}^2 a}{\mathrm{d}t^2}\left[\frac{1}{a}\frac{\mathrm{d}a}{\mathrm{d}t}\right]^{-2} = -\frac{\ddot{a}}{aH^2}. \tag{2.9}$$

The ***deceleration parameter*** is a dimensionless number which measures the rate of change of the rate of expansion H. Different values, or ranges of values, of q_0 correspond to different cosmological models. In principle, it should be possible to determine the value of q_0 observationally. For example, for a set of identical supernovae within remote galaxies, the relationship between apparent brightness and redshift is dependent on the value of the deceleration parameter. Although measurements of this kind are notoriously difficult to make and to interpret, recent observations tend to favor accelerating universe models. In [18], it was estimated that

$$q_0 = -0.55^{+0.26}_{-0.13}. \tag{2.10}$$

As we will see later on in this thesis this value of q_0 is to be taken ***very carefully***. It turns out to be highly model-dependent, and the argument in favor of an accelerating universe is not completely as tight as is commonly believed.

2.4.3 The jerk parameter

The ***Jerk parameter*** (the third time derivative) is also sometimes referred to as jolt. Less common alternative terminologies are pulse, impulse, bounce, surge, shock, and superacceleration. The dimensionless jerk parameter is defined by:

$$j(t) = +\frac{1}{a}\frac{\mathrm{d}^3 a}{\mathrm{d}t^3}\left[\frac{1}{a}\frac{\mathrm{d}a}{\mathrm{d}t}\right]^{-3} = \frac{\dddot{a}}{aH^3}. \tag{2.11}$$

2.4.4 The snap parameter

The ***Snap parameter*** (the fourth time derivative) is also sometimes called jounce. The fifth and sixth time derivatives are sometimes somewhat facetiously referred to as crackle and pop. The dimensionless snap parameter is defined by:

$$s(t) = +\frac{1}{a}\frac{\mathrm{d}^4 a}{\mathrm{d}t^4}\left[\frac{1}{a}\frac{\mathrm{d}a}{\mathrm{d}t}\right]^{-4} = \frac{\dddot{a}}{aH^4}. \tag{2.12}$$

2.4.5 The density parameter

Another useful quantity is the ***energy density parameter***,

$$\Omega = \frac{8\pi G}{3H^2}\rho = \frac{\rho}{\rho_{\mathrm{Hubble}}}, \tag{2.13}$$

where the Hubble density is

$$\rho_{\mathrm{Hubble}} = \frac{3H^2}{8\pi G}. \tag{2.14}$$

This quantity (which will generally change with time) is called the ***Hubble*** density (sometimes also referred to as the ***critical*** density) and current measurements[2] give:

$$\rho_{\mathrm{Hubble}} = 2.775 \times 10^{11}\ h^2\ M_\odot\ Mpc^{-3}, \tag{2.15}$$

where $M_\odot$ is the solar mass and h is the present day normalized Hubble expansion rate with $H_0 = h(100\mathrm{km/s/Mpc})$. Using the Friedmann equation (2.4), we can then write:

$$\Omega - 1 = \frac{k}{H^2 a^2}. \tag{2.16}$$

The sign of k is therefore determined by whether the energy density parameter Ω is greater than, equal to, or less than one. Indeed,

$$\begin{array}{lllll} \rho < \rho_{\mathrm{Hubble}} & \leftrightarrow & \Omega < 1 & \leftrightarrow & k = -1 \quad \leftrightarrow \quad \text{open} \\ \rho = \rho_{\mathrm{Hubble}} & \leftrightarrow & \Omega = 1 & \leftrightarrow & k = 0 \quad \leftrightarrow \quad \text{flat} \\ \rho > \rho_{\mathrm{Hubble}} & \leftrightarrow & \Omega > 1 & \leftrightarrow & k = +1 \quad \leftrightarrow \quad \text{closed} \end{array} \tag{2.17}$$

The density parameter, then, indicates which of the three Robertson-Walker geometries describes our universe. Determining it observationally is an area of intense investigation, however, presently, it is thought to be [11]:

$$\Omega = 1.02 \pm 0.02. \tag{2.18}$$

2.4.6 Analogy with mechanics

Cosmological parameters are used in a similar fashion as parameters used in mechanics, as illustrated in Table 2.1.

[2] See S. Eidelman ***et al.*** from the Particle Data Group [19] for recent measurement values.

Table 2.1: Analogy between Mechanics and Cosmology

Mechanics	**Cosmology**
position $x(t)$	scale factor $a(t)$
velocity $v(t)$	Hubble parameter $H(t)$
acceleration $a(t)$	deceleration $q(t)$
jerk $j(t)$	jerk parameter $j(t)$
snap $s(t)$	snap parameter $s(t)$
crackle	...
pop	...

The deceleration, jerk, and snap parameters are dimensionless, and we can write a relation between the scale factor and these cosmological parameters:

$$a(t) = a_0 \left\{ 1 + H_0\,(t-t_0) - \frac{1}{2}\,q_0\,H_0^2\,(t-t_0)^2 + \frac{1}{3!}\,j_0\,H_0^3\,(t-t_0)^3 + \frac{1}{4!}\,s_0\,H_0^4\,(t-t_0)^4 + O([t-t_0]^5) \right\}. \tag{2.19}$$

Equation (2.19) is a key formula that links the cosmological parameters to the behaviour of the scale factor, and by extension the behaviour of the universe.

2.5 Cosmological Distance Scales

In cosmology there are many different and equally natural definitions of the notion of ***distance*** between two objects or events, whether directly observable or not. Before defining these distance scales, we first need to introduce the cosmological redshift.

2.5.1 The Cosmological Redshift

The energy of a particle will change as it moves in a spacetime geometry similarly to the way it would move in a time-dependent potential. The energy of a photon is proportional to frequency, that change in energy is called the ***cosmological redshift***. Figure 2.2 illustrates this change of energy for a light ray emitted at t_e and observed at t_o.

General definition of the redshift (model independent)

Let λ_e be the wavelength of light ray emitted from some galaxy and λ_o the wavelength of the same light ray observed on Earth. The redshift is defined in its familiar form by:

$$z \;=\; \frac{\lambda_o - \lambda_e}{\lambda_e} = \frac{\lambda_o}{\lambda_e} - 1. \tag{2.20}$$

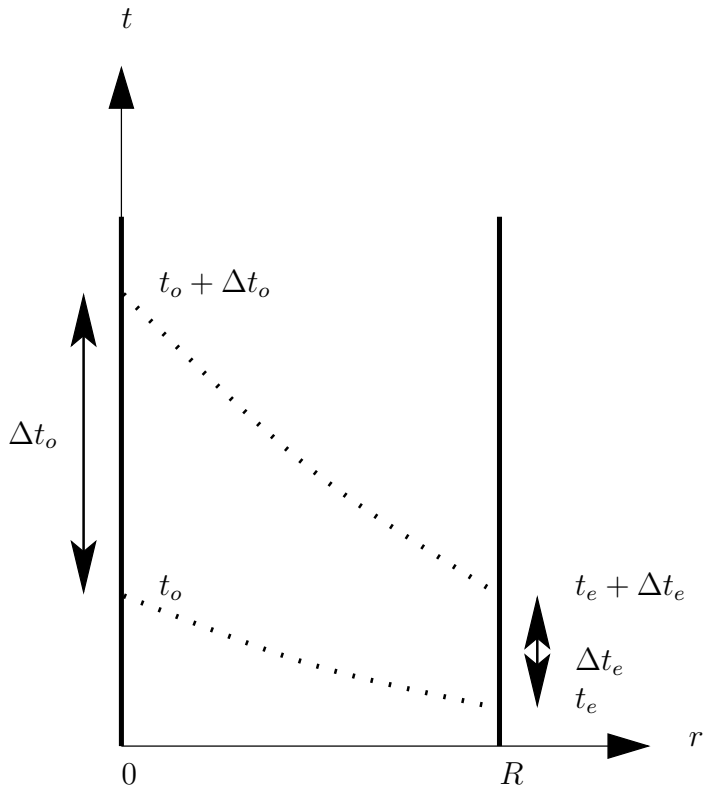


Figure 2.2: The cosmological redshift is the change of energy between a light ray emitted at t_e and observed at t_o.

The rate of change of phase of the light wave v_p can be measured by a first observer moving with the 4-velocity u_1^α by

$$v_p \;=\; k_{1\alpha}\, u_1^\alpha, \tag{2.21}$$

where k_α is the wave vector of the ray. For a short time interval Δt_1, the phase will change by $\Delta P = k_\alpha u^\alpha \Delta t_1$. A second observer moving with the velocity u_2^α and measuring the change of phase at another spacetime point, the same change of phase ΔP will take a different time interval Δt_2. $k_{1\alpha}$ and $k_{2\alpha}$ are respectively affinely parametrized tangent to the null curve, the light ray has to be geodesic. Therefore we have the following ratio:

$$\frac{\Delta t_1}{\Delta t_2} \;=\; \frac{(k_\alpha\, u^\alpha)_2}{(k_\alpha\, u^\alpha)_1}. \tag{2.22}$$

There is a relationship between the change of phase and the frequency ν which results in the following:

$$\frac{\nu_2}{\nu_1} \;=\; \frac{\Delta t_1}{\Delta t_2} = \frac{(k_\alpha\, u^\alpha)_2}{(k_\alpha\, u^\alpha)_1}. \tag{2.23}$$

Since $\lambda_o/\lambda_e = \nu_e/\nu_o$, we have the general model independent cosmological redshift:

$$1+z \quad = \quad \frac{\nu_e}{\nu_o} = \frac{(k_\alpha\, u^\alpha)_2}{(k_\alpha\, u^\alpha)_1}. \tag{2.24}$$

Finally, in terms of light emitted and observed notations, the general cosmological redshift formula becomes:

$$1+z \quad = \quad \frac{(k_\alpha\, u^\alpha)_e}{(k_\alpha\, u^\alpha)_o}. \tag{2.25}$$

In order to apply the above equation (2.25) to observational results, one has to integrate the equations of a null geodesic, which can be very difficult in general.

The redshift in a FLRW universe

To determine the redshift formula in a FLRW universe, one has to know the field of vectors tangent to light rays k^α. Using the spatial homogeneity property of the metric, all points within the same space $t = constant$ are equivalent, therefore a calculation will be independent of the spatial position of the observer. We can then assume an observer to be at the origin $r = 0$. A null geodesic sent off radially lies in the surface $\theta = \phi = constant$ and obeys the following equation:

$$0 \quad = \quad dt^2 - \frac{a^2(t)}{(1-kr^2)}dr^2. \tag{2.26}$$

For an incoming light ray (proceeding towards the observer) we have the following relation:

$$\int_{t_e}^{t_o} \frac{dt}{a(t)} \quad = \quad -\int_{r_e}^{r_o} \frac{dr}{\sqrt{1-kr^2}}. \tag{2.27}$$

We can define the following affine parameter v on the geodesic

$$\frac{dt}{dv} \quad = \quad \frac{1}{a(t)}, \tag{2.28}$$

the tangent vector in this parametrisation can be written as

$$k^\alpha \quad = \quad \left(\frac{-1}{a(t)}, \frac{1}{a^2(t)}\sqrt{1-kr^2}, 0, 0\right). \tag{2.29}$$

Since the velocity field is $u^\alpha = \delta^\alpha_0$, we have

$$k^\alpha\, u_\alpha \quad = \quad \frac{1}{a(t)}. \tag{2.30}$$

Consequently, the cosmological redshift in a FLRW universe can be written as:

$$1+z \quad = \quad \frac{a(t_0)}{a(t_e)}. \tag{2.31}$$

We now have an implicit relation between redshift and time, and we can define distance scales as a function of the redshift rather than time.

2.5.2 Original Hubble law

The original Hubble law gives a simple linear relation between the velocity of recession of an object V and its observed distance d:

$$V = H_0\, d. \tag{2.32}$$

For sufficiently close galaxies this relation is a very good approximation. The recession of galaxies away from us does not imply that we are at the centre of the universe: Hubble's law implies that there is no centre that can be deduced from the expansion itself. Figure 2.5 illustrates this relation with observational data.

2.5.3 Standard (Popular) distance scales

The *luminosity distance* is:

$$d_L(z) \;=\; a_0\,(1+z)\,\sin_k\left\{\frac{c}{H_0\,a_0}\int_0^z \frac{H_0}{H(z)}\,\mathrm{d}z\right\}, \tag{2.33}$$

where

$$\sin_k(x) \;=\; \begin{cases} \sin(x), & k=+1;\\ x, & k=0;\\ \sinh(x), & k=-1. \end{cases} \tag{2.34}$$

By changing variables and adopting definitions as in equations (2.33) and (2.34), we can rewrite the luminosity diameter distance in an alternative exact general form, $\forall\ z \in [-1, +\infty)$ and $\forall$ fixed Ω_0:[3]

$$d_L(z) \;=\; \frac{c}{H_0}\,(1+z)\,\frac{\sinh\left[\sqrt{1-\Omega_0}\int_0^z \frac{H_0}{H(z)}\,\mathrm{d}z\right]}{\sqrt{1-\Omega_0}}, \tag{2.35}$$

where we note

$$\Omega_0 \quad \begin{cases} >1, & k=+1;\\ =1, & k=0;\\ <1, & k=-1. \end{cases} \tag{2.36}$$

Observe that by continuity of the functions $\sin x/x$ and $\sinh x/x$ as $x \to 0$, the function $d_L(z)$ is also continuous as $\Omega_0 \to 1^{\pm}$. For convenience, from equation (2.35), the luminosity distance is given by

$$d_L(z) \;=\; (1+z)\,\frac{c}{H_0}\,\frac{\sinh\left[\sqrt{1-\Omega_0}\,J\right]}{\sqrt{1-\Omega_0}}, \tag{2.37}$$

where J is the integral defined by

$$J \;=\; \int_0^z \frac{H_0}{H(z)}\,\mathrm{d}z \;=\; H_0\,a_0\int_a^{a_0}\frac{\mathrm{d}a}{a\,\dot{a}}. \tag{2.38}$$

[3] Another notation that is sometimes used is $\Omega_k = 1-\Omega_0$, so that $k = -\mathrm{sign}(\Omega_k)$.

It is quite standard to write the luminosity distance versus redshift relation [20, 21] as a Taylor expansion series in z:

$$d_L(z) = \frac{c\,z}{H_0}\left\{1 + \frac{1}{2}\left[1 - q_0\right] z + O(z^2)\right\}, \tag{2.39}$$

and its higher-order extension [22, 23, 24, 25]

$$\begin{aligned} d_L(z) \;=\; & \frac{c\,z}{H_0}\left\{1 + \frac{1}{2}\left[1 - q_0\right] z + \frac{1}{6}\left[q_0 + 3q_0^2 - (j_0 + \Omega_0)\right] z^2 \right. \\ & \left. + \frac{1}{24}\left[2 - 2q_0 - 15q_0^2 - 15q_0^3 + 10q_0 j_0 + 5j_0 + s_0 + 2(1+3q_0)\Omega_0\right] z^3 + O(z^4)\right\}. \end{aligned} \tag{2.40}$$

The *distance modulus* is:

$$\mu_D = 5\ \log_{10}[d_L/(10\ \text{pc})] = 5\ \log_{10}[d_L/(1\ \text{Mpc})] + 25. \tag{2.41}$$

Note that the distance modulus can be rewritten in terms of traditional stellar magnitudes as

$$\mu_D = \mu_{\text{apparent}} - \mu_{\text{absolute}}. \tag{2.42}$$

The continued use of stellar magnitudes and the distance modulus in the context of cosmology is largely a matter of historical tradition, though we shall soon see that the logarithmic nature of the distance modulus has interesting and useful side effects. Note that we prefer as much as possible to deal with natural logarithms: $\ln x = \ln(10)\ \log_{10} x$. Indeed

$$\mu_D = \frac{5}{\ln 10}\ \ln[d_L/(1\ \text{Mpc})] + 25, \tag{2.43}$$

so that

$$\ln[d_L/(1\ \text{Mpc})] = \frac{\ln 10}{5}[\mu_D - 25]. \tag{2.44}$$

2.5.4 More distance scales

Instead of using the standard default choice of luminosity distance d_L, let us now consider using one or more of:

The *photon flux distance*:

$$d_F = \frac{d_L}{(1+z)^{1/2}}. \tag{2.45}$$

The ***photon flux distance*** d_F is based on the fact that it is often technologically easier to count the photon flux (photons/sec) than it is to bolometrically measure total energy flux (power) deposited in the detector. If we are counting photon number flux, rather than energy flux, then the photon number flux contains one fewer factor of $(1+z)^{-1}$. Converted to a distance estimator, the "photon flux distance" contains one extra factor of $(1+z)^{-1/2}$ as compared to the (power-based) luminosity distance.

The *photon count distance*:

$$d_P = \frac{d_L}{(1+z)}. \tag{2.46}$$

The ***photon count distance*** d_P is related to the total number of photons absorbed without regard to the rate at which they arrive. Thus the "photon count distance" contains one extra factor of $(1+z)^{-1}$ as compared to the (power-based) luminosity distance. Indeed D'Inverno [7] uses what is effectively this photon count distance as his nonstandard definition for luminosity distance. Furthermore, though motivated very differently, this quantity is equal to Weinberg's definition of proper motion distance [20], and is also equal to Peebles' version of angular diameter distance [21]. That is:

$$d_P = d_{L,\text{D'Inverno}} = d_{\text{proper,Weinberg}} = d_{A,\text{Peebles}}. \tag{2.47}$$

The *deceleration distance*:

$$d_Q = \frac{d_L}{(1+z)^{3/2}}. \tag{2.48}$$

The quantity d_Q is (as far as we can tell) a previously un-named quantity that seems to have no simple direct physical interpretation — but we shall soon see why it is potentially useful, and why it is useful to refer to it as the ***deceleration distance***.

The *angular diameter distance*:

$$d_A = \frac{d_L}{(1+z)^2}. \tag{2.49}$$

The quantity d_A is Weinberg's definition of angular diameter distance [20], corresponding to the physical size of the object ***when the light was emitted***, divided by its current angular diameter on the sky. This differs from Peebles' definition of angular diameter distance [21], which corresponds to what the size of the object would be at the current cosmological epoch if it had continued to co-move with the cosmological expansion (that is, the "comoving size"), divided by its current angular diameter on the sky. Weinberg's d_A exhibits the (at first sight perplexing, but physically correct) feature that beyond a certain point d_A can actually ***decrease*** as one moves to older objects that are clearly "further" away. In contrast Peebles' version of angular diameter distance is always increasing as one moves "further" away. Note that

$$d_{A,\text{Peebles}} = (1+z)\, d_A. \tag{2.50}$$

See reference [26] for more details on distance measures in cosmology. Obviously

$$d_L \geq d_F \geq d_P \geq d_Q \geq d_A. \tag{2.51}$$

Furthermore these particular distance scales satisfy the property that they converge on each other, and converge on the naive Euclidean notion of distance, as $z \to 0$.

To simplify subsequent formulae, it is now useful to define the ***Hubble distance*** [4]

$$d_H = \frac{c}{H_0}, \tag{2.52}$$

[4] The ***Hubble distance*** $d_H = c/H_0$ is sometimes called the ***Hubble radius***, or the ***Hubble sphere***, or even the "speed of light sphere" [SLS] [27]. Sometimes ***Hubble distance*** is used to refer to the naive estimate $d = d_H\ z$ coming from the linear part of the Hubble relation and ignoring all higher-order terms — this is definitely ***not*** our intended meaning.

so that for $H_0 = 73\ ^{+3}_{-4}$ (km/sec)/Mpc [17] we have

$$d_H = 4100\ ^{+240}_{-160}\ \text{Mpc}. \tag{2.53}$$

Furthermore we choose to set

$$\Omega_0 = 1 + \frac{kc^2}{H_0^2 a_0^2} = 1 + \frac{k\ d_H^2}{a_0^2}. \tag{2.54}$$

For our current purposes Ω_0 is a purely cosmographic definition without dynamical content. (Only if one additionally invokes the Einstein equations in the form of the Friedmann equations does Ω_0 have the standard interpretation as the ratio of total density to the Hubble density, but we would be prejudging things by making such an identification in the current cosmographic framework.) In the cosmographic framework k/a_0^2 is simply the present day curvature of space (not spacetime), while $d_H^{-2} = H_0^2/c^2$ is a measure of the contribution of expansion to the spacetime curvature of the FLRW geometry. More precisely, in a FRLW universe the Riemann tensor has (up to symmetry) only two non-trivial components. In an orthonormal basis:

$$R_{\hat\theta\hat\phi\hat\theta\hat\phi} = \frac{k}{a^2} + \frac{\dot a^2}{c^2\ a^2} = \frac{k}{a^2} + \frac{H^2}{c^2}; \tag{2.55}$$

$$R_{\hat t\hat r\hat t\hat r} = -\frac{\ddot a}{c^2\ a} = \frac{q\ H^2}{c^2}. \tag{2.56}$$

Then at arbitrary times Ω can be defined purely in terms of the Riemann tensor of the FLRW spacetime as

$$\Omega = \frac{R_{\hat\theta\hat\phi\hat\theta\hat\phi}(\dot a \to 0)}{R_{\hat\theta\hat\phi\hat\theta\hat\phi}(k \to 0)}. \tag{2.57}$$

2.6 Lookback time

The lookback time-redshift relation, is defined as the difference between the present age of the Universe t_0 and its age $t(z)$ when a particular light ray at redshift z was emitted. In the context of a FLRW universe, it is given by:

$$T(z) = t_o - t(z) = \int_a^{a_0} \mathrm{d}t \tag{2.58}$$

$$= \int \frac{\mathrm{d}t}{\mathrm{d}a}\mathrm{d}a = \int \frac{a}{\dot a}\frac{\mathrm{d}a}{a} \tag{2.59}$$

$$= \int \frac{1}{H}\frac{\mathrm{d}(a_0/(1+z)}{a_0/(1+z)} = -\int \frac{1}{H}\frac{\mathrm{d}z/(1+z)^2}{1/(1+z)}. \tag{2.60}$$

That is, in a FLRW universe, the lookback time $T(z)$ is:

$$T(z) = \int_0^z \frac{1}{(1+z)\ H(z)}\ \mathrm{d}z$$

2.7 Supernovae

Supernovae are catastrophic explosions of stars whose peak brightness can rival that of the whole host galaxy. They cause a burst of radiation and are detectable at great distances before fading from view over several weeks or months. During this short interval, a supernova can radiate as much energy as the Sun could emit over its life span. Most of the star's material is expelled during the explosion and the consequent shock waves sweep up an expanding shell of gas and dust called a supernova remnant. Figure 2.3 shows an X-ray of the remnant (leftover) of a supernova explosion (Tycho's nova).

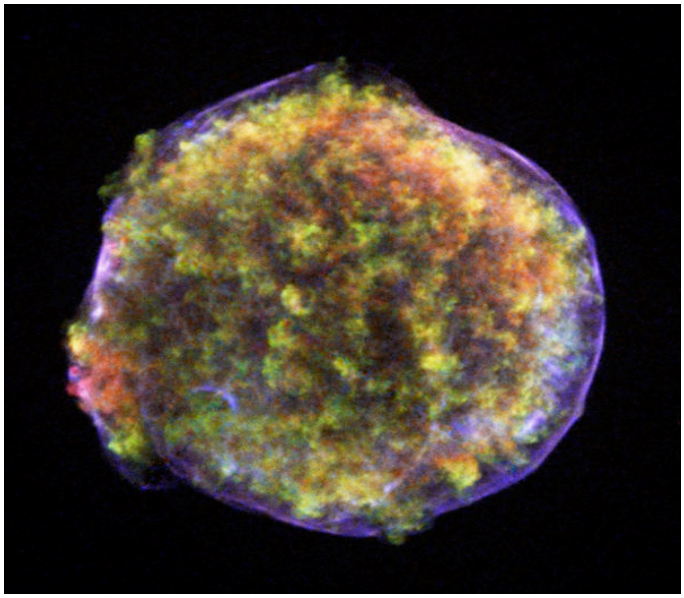


Figure 2.3: X-ray of SN 1572 (Tycho's Nova) remnant as seen by Chandra X-Ray Observatory, Spitzer Space Telescope, and Calar Alto Observatory

There are several kinds of supernovae, they may be triggered in one of two ways, either turning off or suddenly turning on the production of energy through nuclear fusion. Table 2.2 describes the classification of several types of supernovae. On average, supernovae occur about once every 50 years in a galaxy the size of the Milky Way.

A predominant interest in supernova is as "***standard candles***" for measuring distances (or more precisely "***standardizable candles***"). This requires an observation of their peak luminosity. It is therefore important to discover them well before they reach their maximum.

2.7.1 Standard candles

Objects of known brightness are termed ***standard candles***, they are classified into various brightness classes. By comparing the known luminosity of the latter to its observed brightness, the distance to the object can be inferred. Specifically, the luminosity L of a supernova can be determined from its apparent brightness f (energy flux measured on Earth) and the

Table 2.2: Supernovae classifications

Type	Characteristics
Type Ia	***Lacks hydrogen and presents a singly-ionized silicon line at*** 615 nm, ***near peak light***
Type Ib	***Non-ionized helium line at*** 587.6 nm ***and no strong silicon absorption feature near*** 615 nm
Type Ic	***Weak or no helium lines and no strong silicon absorption feature nea***r 615 nm
Type IIP	***Reaches a "plateau" in its light curve***
Type IIL	***Displays a "linear" decrease in its light curve (linear in magnitude versus time)***

luminosity distance d_L can be determined by the inverse square law:

$$f = \frac{L}{4\pi d_L^2}. \tag{2.61}$$

Practically, the luminosity L can be inferred (from the shape and spectral properties of the light curve), the flux f can be measured and therefore the luminosity distance d_L can be measured.

In astronomy, the brightness of an object is given in terms of its absolute magnitude. This quantity is derived from the logarithm of its luminosity as seen from a distance of 10 parsecs. The apparent magnitude, or the magnitude as seen by the observer, can be used to determine the distance D to the object in kiloparsecs (where 1 kpc equals 10^3 parsecs) as follows:

$$5 \cdot \log_{10} \frac{D}{\text{kpc}} \; = \; m \; - \; M \; - \; 5, \tag{2.62}$$

where m is the apparent magnitude and M is the absolute magnitude. For this to be accurate, both magnitudes must be in the same frequency band and there must be no relative motion in the radial direction.

Some means of accounting for interstellar extinction, which also makes objects appear fainter and more red, is also needed.

2.7.2 Problems

Two problems exist for any class of standard candle.

- Calibration: determining exactly what the absolute magnitude of the candle is.
  Classes need to be defined well enough so that members can be recognized. It also means finding enough members with well-known distances that their true absolute magnitude can be determined with enough accuracy.
- Recognition: recognizing members of the class.
  At extreme distances, which is where one most wishes to use a distance indicator, this recognition problem can be quite serious.

The most important issue with standard candles is the recurring question of how standard they are. For example, all observations seem to indicate that type Ia supernovae that are of known distance have all the same brightness. However the possibility that the distant type Ia supernovae have different properties than nearby type Ia supernovae exists.

That this is not merely a philosophical issue can be seen from the history of distance measurements using Cepheid variables. In the 1950s, Walter Baade discovered that the nearby Cepheid variables used to calibrate the standard candle were of a different type than the ones used to measure distances to nearby galaxies. The nearby cepheid variables were population I stars with much higher metal content than the distant population II stars. As a result, the population II stars were actually much brighter than believed, and this had the effect of doubling the distances to the globular clusters, the nearby galaxies, and the diameter of the Milky Way.

2.7.3 Type Ia light curves

Type Ia supernovae are some of the best ways to determine distances. SNIa occur when a binary white dwarf star begins to accrete matter from its companion. As the white dwarf gains matter, eventually it reaches its Chandrasekhar Limit of $1.4M_{\odot}$, once reached, the star becomes unstable and undergoes a runaway nuclear fusion reaction. Because all Type Ia supernovae explode at about the same mass, their absolute magnitudes are all the same. Moreover there is some similarity in basic mechanism between one SNIa and the next and hence some similarity in their peak luminosity. There is an even tighter correlation between peak brightness and time it takes for the brightness to decay. This makes them very useful as standard candles. All type Ia SN have a standard blue and visual magnitude of

$$M_B \approx M_V \approx -19.3 \pm 0.03.$$

Figure 2.4 shows the light curve of a supernova.

When observing a type Ia supernova, if it is possible to determine what its peak magnitude was, then its distance can be calculated. It is not intrinsically necessary to capture the supernova directly at its peak magnitude; using the multicolor light curve method (MCLS), the shape of the light curve (taken at any reasonable time after the initial explosion) is compared to a family of parameterized curves that will determine the absolute magnitude at the maximum brightness.

Using Type Ia supernovae is one of the most accurate methods. Much time has been devoted to the refining of this method.

2.7.4 The legacy05 dataset

The supernova data is available in published form [28], and in a slightly different format, via internet [2]. (The differences amount to minor matters of choice in the presentation.) The final processed result reported for each 115 of the supernovae is a redshift z, a luminosity modulus μ_B, and an uncertainty in the luminosity modulus. The luminosity modulus can be converted into a luminosity distance via the formula

$$d_L = (1 \text{ Megaparsec}) \times 10^{(\mu_B + \mu_{\text{offset}} - 25)/5}. \qquad (2.63)$$

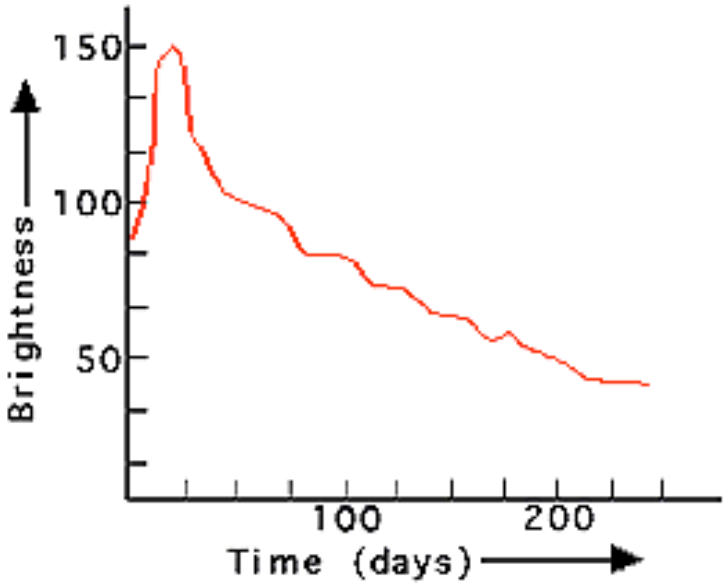


Figure 2.4: Supernova light curve "standard candles" (NASA)

The reason for the ***offset*** is that supernovae by themselves only determine the ***shape*** of the Hubble relation (*i.e.*, q_0, j_0, *etc.*), but not its absolute ***slope*** (*i.e.*, H_0) — this is ultimately due to the fact that we do not have good control of the absolute luminosity of the supernovae in question. The offset μ_{offset} can be chosen to match the known value of H_0 coming from other sources. (In fact the data reported in the published article [28] has already been normalized in this way to the ***standard value*** $H_{70} = 70$ (km/sec)/Mpc, corresponding to Hubble distance $d_{70} = c/H_{70} = 4283$ Mpc, whereas the data available on the website [2] has ***not*** been normalized in this way — which is why μ_B as reported on the website is systematically 19.308 stellar magnitudes smaller than that in the published article.)

The other item one should be aware of concerns the error bars: The error bars reported in the published article [28] are photometric uncertainties only — there is an additional source of error to do with the intrinsic variability of the supernovae. In fact, if you take the ***photometric*** error bars seriously as estimates of the ***total*** uncertainty, you would have to reject the hypothesis that we live in a standard FLRW universe. Instead, intrinsic variability in the supernovae is by far the most widely accepted interpetation. Basically one uses the ***nearby*** dataset to estimate an intrinsic variability that makes chi-squared look reasonable. This intrinsic variability of 0.13104 stellar magnitudes [2, 12]) has been estimated by looking at low redshift supernovae (where we have good measures of absolute distance from other techniques), and has been included in the error bars reported on the website [2]. Indeed

$$(\text{uncertainty})_{\text{website}} = \sqrt{(\text{intrinsic variability})^2 + (\text{uncertainty})^2_{\text{article}}}. \qquad \boxed{2.64}$$

With these key features of the supernovae data kept in mind, conversion to luminosity distance and estimation of scientifically reasonable error bars (suitable for chi-square analysis) is straightforward.

2.7.5 The gold06 dataset

Our second collection of supernova data is the gold06 dataset [3]. This dataset contains 206 supernovae (including ***most but not all*** of the legacy05 supernovae) and reaches out considerably further in redshift, with one outlier at $z = 1.755$, corresponding to $y = 0.6370$. Though the dataset is considerably more extensive it is unfortunately heterogeneous — combining observations from five different observing platforms over almost a decade. In some cases full data on the operating characteristics of the telescopes used does not appear to be publicly available. The issue of data inhomogeneity has been specifically addressed by Nesseris and Perivolaropoulos in [29]. (For related discussion, see also [30].) In the gold06 dataset one is presented with distance moduli and total uncertainty estimates, in particular, including the intrinsic dispersion.

A particular point of interest is that the HST-based high-z supernovae previously published in the gold04 dataset [2] have their estimated distances reduced by approximately 5% (corresponding to $\Delta\mu_D = 0.10$), due to a better understanding of nonlinearities in the photodetectors. [5] Furthermore, the authors of [3] incorporate (most of) the supernovae in the legacy dataset [28, 2], but do so in a modified manner by reducing their estimated distance moduli by $\Delta\mu_D = 0.19$ (corresponding naively to a 9.1% reduction in luminosity distance) — however this is only a change in the normalization used in reporting the data, not a physical change in distance. Based on revised modelling of the light curves, and ignoring the question of overall normalization, the overlap between the gold06 and legacy05 datasets is argued to be consistent to within 0.5% [3].

The critical point is this: Since one is still seeing $\approx 5\%$ variations in estimated supernova distances on a two-year timescale, this strongly suggests that the unmodelled systematic uncertainties (the so-called ***unknown unknowns***) are not yet fully under control in even the most recent data. It would be prudent to retain a systematic uncertainty budget of at least 5% (more specifically, $\Delta\mu_D = 0.10$), and not to place too much credence in any result that is not robust under possible systematic recalibrations of this magnitude. Indeed the authors of [3] state:

- "... we adopt a limit on redshift-dependent systematics to be 5% per $\Delta z = 1$";
- "At present, none of the ***known***, well-studied sources of systematic error rivals the statistical errors presented here."

We shall have more to say about possible systematic uncertainties, both "known unknowns" and ***unknown unknowns*** later in chapter 3.

2.8 Some history

The need for a certain amount of caution in interpreting the observational data can clearly be inferred from a dispassionate reading of history. We reproduce below Hubble's original 1929 version of what is now called the Hubble plot (Figure 2.5(a)) [31], a modern update

[5] Changes in stellar magnitude are related to changes in luminosity distance via equations 2.43 and 2.44. Explicitly $\Delta(\ln d_L) = \ln 10\ \Delta\mu_D/5$, so that for a given uncertainty in magnitude the corresponding luminosity distances are multiplied by a factor $10^{\Delta\mu_D/5}$. Then 0.10 magnitudes $\to$ 4.7% $\approx$ 5%, and similarly 0.19 magnitudes $\to$ 9.1%.

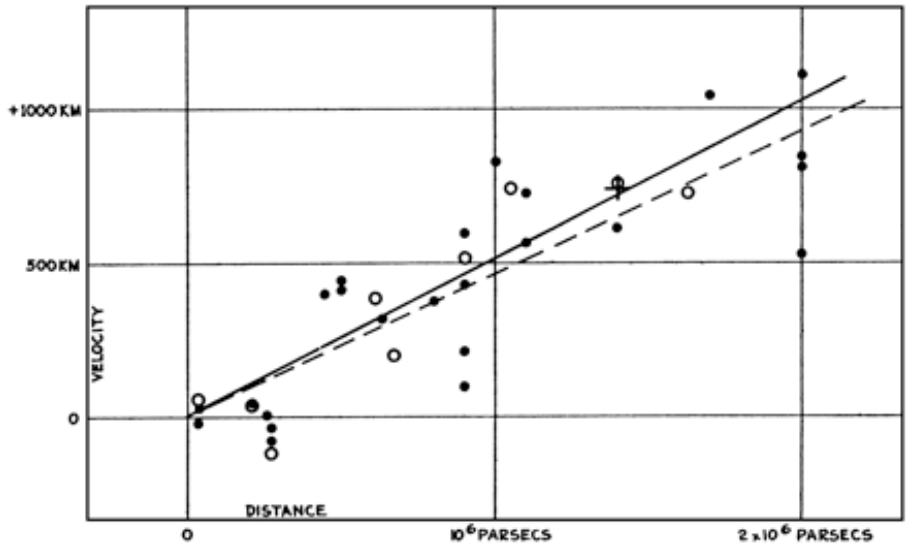
from 2004 (Figure 2.5(b)) [32], and a very telling plot of the estimated value of the Hubble parameter ***as a function of publication date*** (Figure 2.6) [32]. Regarding this last plot, Kirshner is moved to comment [32]:

> "At each epoch, the estimated error in the Hubble constant is small compared with the subsequent changes in its value. This result is a symptom of underestimated systematic errors."

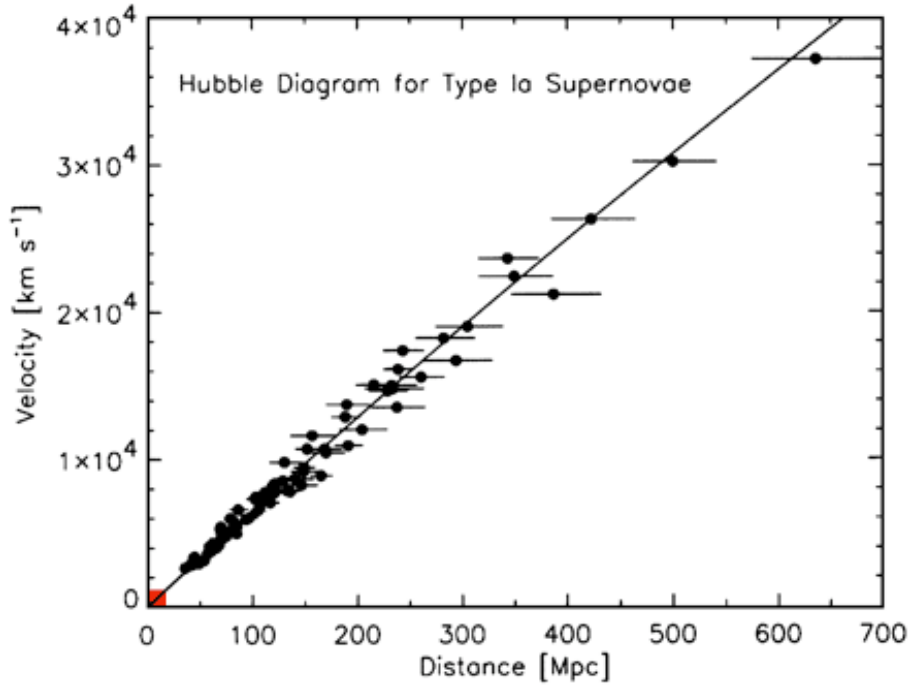
It is important to realise that the systematic under-estimating of systematic uncertainties is a generic phenomenon that cuts across disciplines and sub-fields, it is not a phenomenon that is limited to cosmology. For instance, the ***Particle Data Group*** [http://pdg.lbl.gov/] in their bi-annual ***Review of Particle Properties*** publishes fascinating plots of estimated values of various particle physics parameters ***as a function of publication date*** (Figure 2.7) [17]. These plots illustrate an aspect of the experimental and observational sciences that is often overlooked:

> ***It is simply part of human nature to always think the situation regarding systematic uncertainties is better than it actually is — systematic uncertainties are systematically under-reported.***

This historical perspective should be kept in focus — ultimately the treatment of systematic uncertainties will prove to be an important component in estimating the reliability and robustness of any conclusions we can draw from the data.



(a) Hubble's original 1929 plot [31]. Note the rather large scatter in the data.



(b) Modern 2004 version of the Hubble plot. From Kirshner [32]. The original 1929 Hubble plot is confined to the small red rectangle at the bottom left.

Figure 2.5: The original Hubble law with observational data.

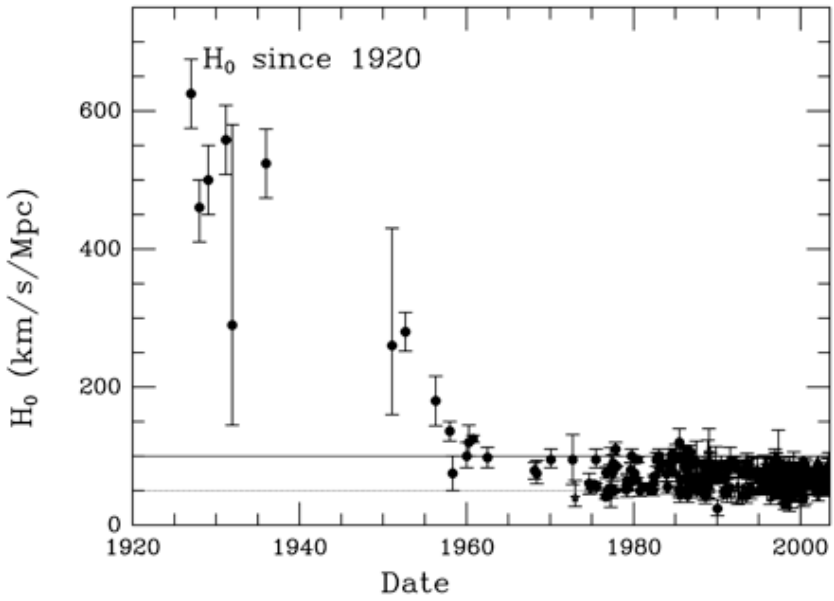


Figure 2.6: Estimates of the Hubble parameter as a function of publication date. From Kirshner [32]. Quote: "At each epoch, the estimated error in the Hubble constant is small compared with the subsequent changes in its value. This result is a symptom of underestimated systematic errors."

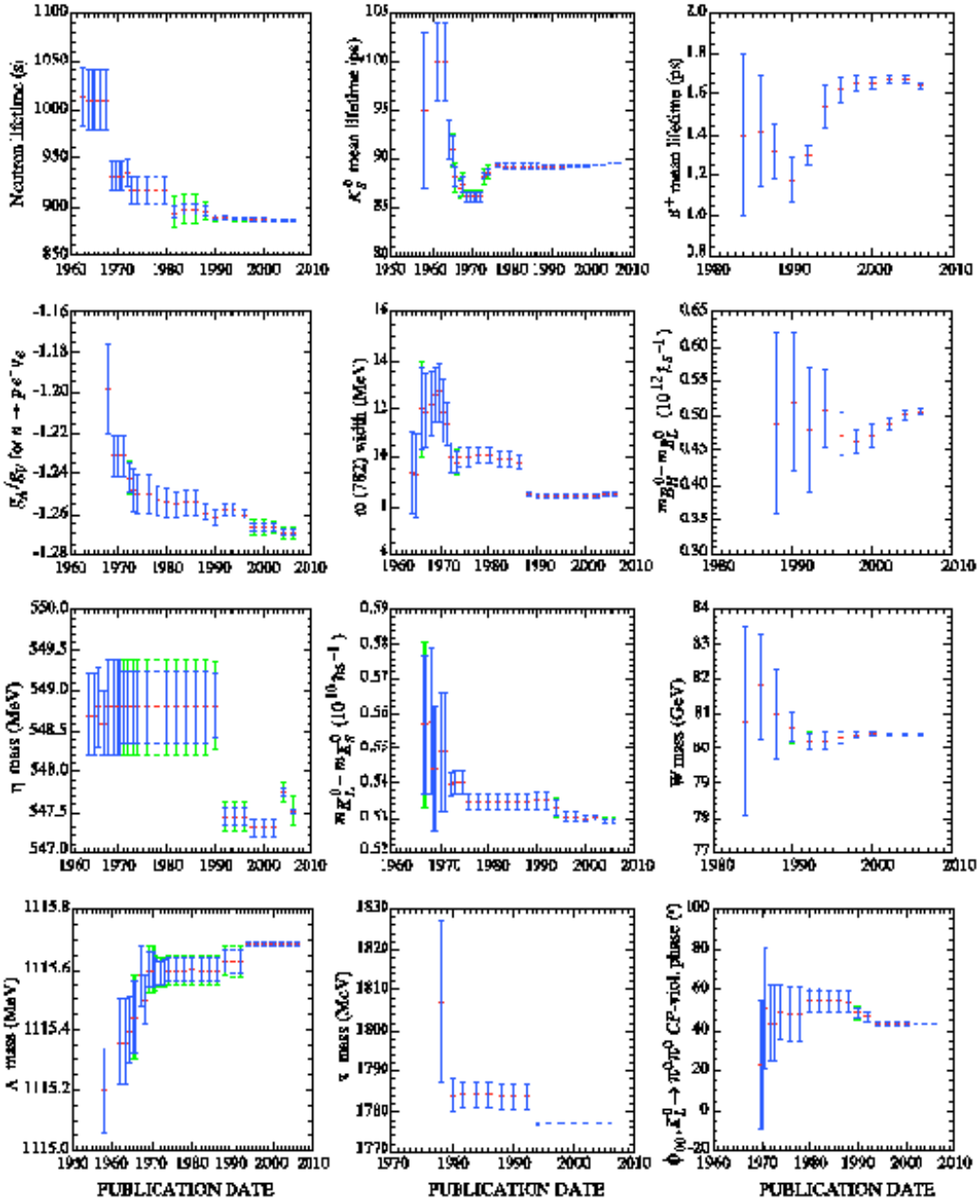


Figure 2.7: Some historical plots of particle physics parameters as a function of publication date. From the Particle Data Group's 2006 Review of Particle Properties [17]. These plots strongly suggest that the systematic under-estimating of systematic uncertainties is a generic phenomenon that cuts across disciplines and sub-fields, it is not a phenomenon that is limited to cosmology.

2.9 The *standard* Cosmological Model (ΛCDM)

ΛCDM or Lambda-CDM is an abbreviation for Lambda-Cold Dark Matter. This model is referred to as the concordance model of big bang cosmology, it attempts to explain cosmic microwave background observations, large scale structure observations and supernovae observations of the ***accelerating*** expansion of the universe. In this model Λ is the cosmological constant that stands for dark energy.

This model has very strong assumptions, the simplest are:

- Nearly scale-invariant spectrum of primordial perturbations.
- A universe without spatial curvature ($k = 0$).
- No observable topology, so that the universe is much larger than the observable particle horizon.
- Cosmic inflation.
- FLRW metric, the Friedmann equations (Einstein field equations) and the cosmological equations of state to describe the universe from right after the inflationary epoch to present and future.

This model has 6 basic parameters: 3 parameters relevant to the Friedmann equations, the Hubble parameter H_0, the baryon density Ω_b, the total matter density (baryon + dark matter) Ω_m, and 3 other parameters related to the CMB and perturbative structure, the optical depth to reionization τ, the scalar fluctuation amplitude A_s and the scalar spectral index n_s. The model also has some derived parameters including the critical density ρ_0, the dark energy density Ω_Λ and the age of the universe t_0.

There are some concerns on some of these assumptions. In particular, cosmic inflation predicts spatial curvature at the level of 10^{-4} to 10^{-5}. Moreover, the ΛCDM says nothing about the fundamental physical origin of dark matter, dark energy and the nearly scale-invariant spectrum of primordial curvature perturbations.

2.10 Energy conditions

In classical general relativity, there are several types of energy conditions [33]:

- the null energy condition (NEC);
- the weak energy condition (WEC);
- the strong energy condition (SEC);
- the dominant energy condition (DEC).

The energy conditions of general relativity permit one to deduce very powerful and general theorems about the behaviour of strong gravitational fields and cosmological geometries. There are also ***Averaged Energy Conditions*** (AEC), but they are of less relevance in FLRW cosmology. These conditions can most easily be stated in terms of the components of the stress energy tensor $T^{\hat{\mu}\hat{\nu}}$ in an orthonormal frame. Ultimately, however, constraints

on the stress-energy are converted, via the Einstein equations, to constraints on the spacetime geometry — in particular in a FLRW spacetime one is ultimately imposing conditions on the scale factor and its time derivatives (and implicitly cosmological parameters). In FLRW cosmology, it is sufficiently general to assume that the energy momentum tensor is of Hawking–Ellis type one (type I) [34, p 89]. In an orthonormal frame, the components of the stress energy tensor are given by:

$$T^{\hat{a}\hat{b}} = \begin{bmatrix} \rho & 0 & 0 & 0 \\ 0 & p_1 & 0 & 0 \\ 0 & 0 & p_2 & 0 \\ 0 & 0 & 0 & p_3 \end{bmatrix}. \tag{2.65}$$

The components of $T^{\hat{a}\hat{b}}$ are the energy density and the three principal pressures.

2.10.1 Null Energy condition (NEC)

For all future pointing null vectors k^a, we ask that:

$$T_{ab}\, k^a k^b \geqslant 0 \tag{2.66}$$

In terms of pressures and density, we have:

$$\forall\, i \qquad \rho + p_i \geqslant 0.$$

Hawking's area theorem for black hole horizon relies on the NEC, and hence evaporation of a black hole must violate the NEC.

2.10.2 Weak Energy condition (WEC)

Sometimes it is useful to think about Einstein's equations without specifying the theory of matter from which $T^{\hat{a}\hat{b}}$ is derived. This leaves us with a great deal of arbitrariness, in the absence of some constraints on $T^{\hat{a}\hat{b}}$, any metric can satisfy the Einstein equations. The real concern is the existence of solutions to Einstein's equations with ***realistic*** sources of energy and momentum. The most common property that is demanded of $T^{\hat{a}\hat{b}}$ is that it represent positive energy densities — no negative masses are allowed. In a locally inertial frame this requirement can be stated as $\rho = T_{00} > 0$. To turn this into a coordinate-independent statement, we ask that:

$$T_{ab}V^aV^b \geqslant 0 \qquad \forall \text{ timelike vector } V$$

In terms of pressures and density, we have:

$$\rho \geqslant 0 \qquad \text{and} \qquad \forall\, i \qquad \rho + p_i \geqslant 0.$$

Any timelike vector can be a tangent to an observer's world line. The WEC condition states that the energy density measured by any timelike observer is non-negative. It seems like a fairly reasonable requirement, and many of the important theorems about solutions to general relativity (such as the singularity theorems of Hawking and Penrose ([35, p 240])

rely on this condition or something very close to it. Unfortunately it is not set in stone; indeed, it is straightforward to invent otherwise respectable classical field theories which violate the WEC, and almost impossible to invent a quantum field theory which obeys it. Nevertheless, it is legitimate to assume that the WEC holds in all but the most extreme conditions.

2.10.3 Strong Energy Condition (SEC)

For any timelike vectors V^a, we ask that:

$$\left(T_{ab} - \frac{T}{2} g_{ab}\right) V^a V^b \geqslant 0$$

where T is the trace of the stress-energy tensor: $T = T_{ab} g^{ab}$.

In terms of pressures and density, we have:

$$T = -\rho + \sum_i p_i$$

$$\forall i \qquad \rho + p_i \geqslant 0 \quad \text{and} \quad \rho + \sum_i p_i \geqslant 0.$$

Note that the SEC implies the NEC, it does not imply the WEC. For example, matter with a negative energy density but sufficiently high pressures could satisfy the SEC but would violate the WEC.

The Penrose–Hawking singularity theorem relevant to the cosmological singularity uses the SEC. See [36, 37] for strong energy condition violations.

2.10.4 Dominant Energy Condition (DEC)

For any timelike vectors V^a, we ask that:

$$T_{ab} V^a V^b \geqslant 0 \qquad \text{and that} \qquad T_{ab} V^b \text{ is a future directed non-spacelike vector.}$$

The DEC assumes that the WEC holds, and that for all future directed timelike vectors V^a that $T_{ab} V^b$ is a future directed non-spacelike vector. This ensures that the net energy flow does not exceed the speed of light. The dominant energy condition implies the weak energy condition and also the null energy condition, but does not necessarily imply the strong energy condition.

In terms of pressures and density, we have:

$$\rho \geqslant 0 \qquad \text{and} \qquad \forall\, i \quad -\rho \leqslant p_i \leqslant \rho.$$

The dominant energy condition can be interpreted as saying that the speed of energy flow of matter is always less than the speed of light.

2.10.5 Comments on the Energy Conditions

Note that the null energy condition implies the weak energy condition, but otherwise the NEC, the WEC and the SEC are mathematically independent assumptions. In particular, the SEC does not imply the WEC. It is stronger only in the sense that it appears to be a stronger physical requirement to assume equation (2.10.3) rather than equation (2.10.2). Violating the NEC implies violating the DEC, SEC and WEC as well.

The energy conditions are looking a lot less secure than they once seemed:

- There are quantum effects that violate all of the energy conditions.
- There are even relatively benign looking classical systems that violate all the energy conditions [33].

Note that ideal relativistic fluids satisfy the DEC, and certainly all the known forms of normal matter encountered in our solar system satisfy the DEC. With sufficiently strong self-intereactions relativistic fluids can be made to violate the SEC (and DEC); but classical relativistic fluids always seem to satisfy the NEC. Most classical fields (apart from non-minimally coupled scalars) satisfy the NEC. Violating the NEC seems to require either quantum physics (which is unlikely to be a major contributor to the overall cosmological evolution of the universe) or non-minimally coupled scalars (implying that one is effectively adopting some form of scalar-tensor gravity).

"Statistics:
The only science that enables different experts using the same figures to draw different conclusions."

Evan Esar (1899–1995)

3

Cosmography in a FLRW universe

From various observations of the Hubble relation, most recently including the supernova data [28, 1, 2, 3, 38, 39], one is by now very accustomed to seeing many plots of luminosity distance d_L versus redshift z. But are there better ways of representing the data?

For instance, consider cosmography (cosmokinetics) which is the part of cosmology that proceeds by making minimal dynamic assumptions. One keeps the geometry and symmetries of FLRW spacetime,

$$\mathrm{d}s^2 = -c^2\,\mathrm{d}t^2 + a(t)^2\left\{\frac{\mathrm{d}r^2}{1-k\,r^2} + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\right\}, \tag{3.1}$$

at least as a working hypothesis, but does not assume the Friedmann equations (Einstein equations), unless and until absolutely necessary. By doing so it is possible to defer questions about the equation of state of the cosmological fluid, minimize the number of theoretical assumptions one is bringing to the table, and so concentrate more directly on the observational situation.

In particular, the "***big picture***" is best brought into focus by performing a global fit of all available supernova data to the Hubble relation, from the current epoch at least back to redshift $z \approx 1.75$. Indeed, all the discussion over acceleration versus deceleration, and the presence (or absence) of jerk (and snap) ultimately boils down, in a cosmographic setting, to doing a finite-polynomial truncated–Taylor series fit of the distance measurements (determined by supernovae and other means) to some suitable form of distance–redshift or distance–velocity relationship. Phrasing the question to be investigated in this way keeps it as close as possible to Hubble's original statement of the problem, while minimizing the number of extraneous theoretical assumptions one is forced to adopt. For instance, it is quite standard to phrase the investigation in terms of the luminosity distance versus redshift relation [20, 21]:

$$d_L(z) = \frac{c\,z}{H_0}\left\{1 + \frac{1}{2}\,[1-q_0]\,z + O(z^2)\right\}, \tag{3.2}$$

and its higher-order extension [22, 23, 24, 25]

$$d_L(z) = \frac{c\,z}{H_0}\left\{1+\frac{1}{2}\left[1-q_0\right]z+\frac{1}{6}\left[q_0+3q_0^2-(j_0+\Omega_0)\right]z^2 + \frac{1}{24}\left[2-2q_0-15q_0^2-15q_0^3+10q_0j_0+5j_0+s_0+2(1+3q_0)\Omega_0\right]z^3+O(z^4)\right\}. \tag{3.3}$$

A central question thus has to do with the choice of the luminosity distance as the primary quantity of interest — there are several other notions of cosmological distance that can be used, some of which (we shall see) lead to simpler and more tractable versions of the Hubble relation. Furthermore, as will quickly be verified by looking at the derivation (see, for example, [20, 21, 22, 23, 24, 25], the standard Hubble law is actually a Taylor series expansion derived for small z, whereas much of the most interesting recent supernova data occurs at $z > 1$. Should we even trust the usual formalism for large $z > 1$? Two distinct things could go wrong:

- The underlying Taylor series could fail to converge.
- Finite truncations of the Taylor series might be a bad approximation to the exact result.

In fact, ***both*** things happen. There are good mathematical and physical reasons for this undesirable behaviour, as we shall discuss below. We shall carefully explain just what goes wrong — and suggest various ways of improving the situation. Our ultimate goal will be to find suitable forms of the Hubble relation that are well adapted to performing fits to all the available distance ***versus*** redshift data.

Moreover — once one stops to consider it carefully — why should the cosmology community be so fixated on using the luminosity distance d_L (or its logarithm, proportional to the distance modulus) and the redshift z as the relevant parameters? In principle, in place of luminosity distance $d_L(z)$ versus redshift z one could just as easily plot $f(d_L, z)$ versus $g(z)$, choosing $f(d_L, z)$ and $g(z)$ to be arbitrary locally invertible functions, and ***exactly the same physics*** would be encoded. Suitably choosing the quantities to be plotted and fit will not change the physics, ***but it might improve statistical properties and insight***. (And we shall soon see that it will definitely improve the behaviour of the Taylor series.)

By comparing cosmological parameters obtained using multiple different fits of the Hubble relation to different distance scales and different parameterizations of the redshift we can then assess the robustness and reliability of the data fitting procedure. In performing this analysis we had hoped to verify the robustness of the Hubble relation, and to possibly obtain improved estimates of cosmological parameters such as the deceleration parameter and jerk parameter, thereby complementing other recent cosmographic and cosmokinetic analyses such as [12, 13, 14, 15, 16], as well as other analyses that take a sometimes skeptical view of the totality of the observational data [40, 41, 30, 42, 43]. The actual results of our current cosmographic fits to the data are considerably more ambiguous than we had initially expected, and there are many subtle issues hiding in the simple phrase "fitting the data".

In this chapter we first discuss the various cosmological distance scales, and the related versions of the Hubble relation. We then discuss technical problems with the usual redshift

variable for $z > 1$, and how to ameliorate them, leading to yet more versions of the Hubble relation. After discussing key features of the supernova data, we perform, analyze, and contrast multiple fits to the Hubble relation — providing discussions of model-building uncertainties (some technical details being relegated to the appendices) and sensitivity to systematic uncertainties. Finally we present our results and conclusions: There is a disturbingly strong model-dependence in the resulting estimates for the deceleration parameter. Furthermore, once realistic estimates of systematic uncertainties (based on the published data) are budgeted for it becomes clear that purely statistical estimates of goodness of fit are dangerously misleading. While the "***preponderance of evidence***" certainly suggests an accelerating universe, we would argue that this conclusion is not currently supported "***beyond reasonable doubt***" — the supernova data (considered by itself) certainly ***suggests*** an accelerating universe, it is not sufficient to allow us to reliably conclude that the universe is accelerating.[1]

3.1 New versions of the Hubble law

As illustrated in section 2.5 on ***Cosmological Distance Scales***, there are many different and equally natural definitions of the notion of ***distance*** between two objects or events, whether directly observable or not.

For the vertical axis of the Hubble plot, instead of using the standard default choice of luminosity distance d_L, let us now consider using one or more of:

- The ***photon flux distance***:
$$d_F = \frac{d_L}{(1+z)^{1/2}}. \tag{3.4}$$
- The ***photon count distance***:
$$d_P = \frac{d_L}{(1+z)}. \tag{3.5}$$
- The ***deceleration distance***:
$$d_Q = \frac{d_L}{(1+z)^{3/2}}. \tag{3.6}$$
- The ***angular diameter distance***:
$$d_A = \frac{d_L}{(1+z)^{2}}. \tag{3.7}$$
- The ***distance modulus***:
$$\mu_D = 5\ \log_{10}[d_L/(10\,\text{pc})] = 5\ \log_{10}[d_L/(1\,\text{Mpc})] + 25. \tag{3.8}$$
- Or possibly some other surrogate for distance.

[1] If one adds additional theoretical assumptions, such as by specifically fitting to a Λ-CDM model, the situation at first glance looks somewhat better — but this is then telling you as much about one's choice of theoretical model as it is about the observational situation.

Remember the relation between the distances

$$d_L \geq d_F \geq d_P \geq d_Q \geq d_A. \tag{3.9}$$

Furthermore these particular distance scales satisfy the property that they converge on each other, and converge on the naive Euclidean notion of distance, as $z \to 0$.

New versions of the Hubble law are easily calculated for each of these cosmological distance scales. Explicitly:

$$\begin{aligned} d_L(z) &= d_H\, z\Bigg\{1+\frac{1}{2}\left[1-q_0\right]z+\frac{1}{6}\left[q_0+3q_0^2-(j_0+\Omega_0)\right]z^2 \\ &\quad +\frac{1}{24}\left[2-2q_0-15q_0^2-15q_0^3+10q_0j_0+5j_0+s_0+2(1+3q_0)\Omega_0\right]z^3+O(z^4)\Bigg\}. \end{aligned} \tag{3.10}$$

$$\begin{aligned} d_F(z) &= d_H\, z\Bigg\{1-\frac{1}{2}q_0z+\frac{1}{24}\left[3+10q_0+12q_0^2-4(j_0+\Omega_0)\right]z^2 \\ &\quad +\frac{1}{48}\left[2-17q_0-42q_0^2-30q_0^3+20q_0j_0+14j_0+2s_0+4(2+3q_0)\Omega_0\right]z^3+O(z^4)\Bigg\}. \end{aligned} \tag{3.11}$$

$$\begin{aligned} d_P(z) &= d_H\, z\Bigg\{1-\frac{1}{2}\left[1+q_0\right]z+\frac{1}{6}\left[3+4q_0+3q_0^2-(j_0+\Omega_0)\right]z^2 \\ &\quad +\frac{1}{24}\left[-3-9q_0-16q_0^2-10q_0^3+8q_0j_0+7j_0+s_0+4(1+q_0)\Omega_0\right]z^3+O(z^4)\Bigg\}. \end{aligned} \tag{3.12}$$

$$\begin{aligned} d_Q(z) &= d_H\, z\Bigg\{1-\frac{1}{2}\left[2+q_0\right]z+\frac{1}{24}\left[27+22q_0+12q_0^2-4(j_0+\Omega_0)\right]z^2 \\ &\quad +\frac{1}{48}\left[-44-61q_0-66q_0^2-30q_0^3+20q_0j_0+22j_0+2s_0+4(4+3q_0)\Omega_0\right]z^3+O(z^4)\Bigg\}. \end{aligned} \tag{3.13}$$

$$\begin{aligned} d_A(z) &= d_H\, z\Bigg\{1-\frac{1}{2}\left[3+q_0\right]z+\frac{1}{6}\left[12+7q_0+3q_0^2-(j_0+\Omega_0)\right]z^2 \\ &\quad +\frac{1}{24}\left[-50-46q_0-39q_0^2-15q_0^3+10q_0j_0+13j_0+s_0+2(5+3q_0)\Omega_0\right]z^3+O(z^4)\Bigg\}. \end{aligned} \tag{3.14}$$

If one simply wants to deduce (for instance) the ***sign*** of q_0, then it seems that plotting the ***photon flux distance*** d_F versus z would be a particularly good test — simply check if the first nonlinear term in the Hubble relation curves up or down.

In contrast, the Hubble law for the distance modulus itself is given by the more complicated expression

$$\begin{aligned}\mu_D(z) &= 25 + \frac{5}{\ln(10)}\Bigg\{ \ln(d_H/\text{Mpc}) + \ln z \\ &+\frac{1}{2}\left[1-q_0\right] z - \frac{1}{24}\left[3 - 10q_0 - 9q_0^2 + 4(j_0+\Omega_0)\right] z^2 \\ &+\frac{1}{24}\left[5 - 9q_0 - 16q_0^2 - 10q_0^3 + 8q_0 j_0 + 7j_0 + s_0 + 4(1+q_0)\Omega_0\right] z^3 + O(z^4)\Bigg\}.\end{aligned} \tag{3.15}$$

However, when plotting μ_D versus z, most of the observed curvature in the plot comes from the universal ($\ln z$) term, and so carries no real information and is relatively uninteresting. It is much better to rearrange the above as:

$$\begin{aligned}\ln[d_L/(z\ \text{Mpc})] &= \frac{\ln 10}{5}[\mu_D - 25] - \ln z \\ &= \ln(d_H/\text{Mpc}) \\ &\quad -\frac{1}{2}\left[-1+q_0\right] z + \frac{1}{24}\left[-3 + 10q_0 + 9q_0^2 - 4(j_0+\Omega_0)\right] z^2 \\ &\quad +\frac{1}{24}\left[5 - 9q_0 - 16q_0^2 - 10q_0^3 + 8q_0 j_0 + 7j_0 + s_0 + 4(1+q_0)\Omega_0\right] z^3 + O(z^4),\end{aligned} \tag{3.16}$$

In a similar manner one has

$$\begin{aligned}\ln[d_F/(z\ \text{Mpc})] &= \frac{\ln 10}{5}[\mu_D - 25] - \ln z - \frac{1}{2}\ln(1+z) \\ &= \ln(d_H/\text{Mpc}) \\ &\quad -\frac{1}{2}q_0 z + \frac{1}{24}\left[3 + 10q_0 + 9q_0^2 - 4(j_0+\Omega_0)\right] z^2 \\ &\quad +\frac{1}{24}\left[1 - 9q_0 - 16q_0^2 - 10q_0^3 + 8q_0 j_0 + 7j_0 + s_0 + 4(1+q_0)\Omega_0\right] z^3 + O(z^4),\end{aligned} \tag{3.17}$$

$$\begin{aligned}\ln[d_P/(z\ \text{Mpc})] &= \frac{\ln 10}{5}[\mu_D - 25] - \ln z - \ln(1+z) \\ &= \ln(d_H/\text{Mpc}) \\ &\quad -\frac{1}{2}\left[1+q_0\right] z + \frac{1}{24}\left[9 + 10q_0 + 9q_0^2 - 4(j_0+\Omega_0)\right] z^2 \\ &\quad +\frac{1}{24}\left[3 - 9q_0 - 16q_0^2 - 10q_0^3 + 8q_0 j_0 + 7j_0 + s_0 + 4(1+q_0)\Omega_0\right] z^3 + O(z^4),\end{aligned} \tag{3.18}$$

$$\begin{aligned}\ln[d_Q/(z\ \text{Mpc})] &= \frac{\ln 10}{5}[\mu_D - 25] - \ln z - \frac{3}{2}\ln(1+z) \\ &= \ln(d_H/\text{Mpc}) \\ &\quad -\frac{1}{2}\left[2+q_0\right] z + \frac{1}{24}\left[15 + 10q_0 + 9q_0^2 - 4(j_0+\Omega_0)\right] z^2 \\ &\quad +\frac{1}{24}\left[-7 - 9q_0 - 16q_0^2 - 10q_0^3 + 8q_0 j_0 + 7j_0 + s_0 + 4(1+q_0)\Omega_0\right] z^3 + O(z^4),\end{aligned} \tag{3.19}$$

$$\begin{aligned}\ln[d_A/(z\ \text{Mpc})] &= \frac{\ln 10}{5}[\mu_D - 25] - \ln z - 2\ln(1+z) \\ &= \ln(d_H/\text{Mpc}) \\ &\quad -\frac{1}{2}\left[3+q_0\right] z + \frac{1}{24}\left[21 + 10q_0 + 9q_0^2 - 4(j_0+\Omega_0)\right] z^2 \\ &\quad +\frac{1}{24}\left[-3 - 9q_0 - 16q_0^2 - 10q_0^3 + 8q_0 j_0 + 7j_0 + s_0 + 4(1+q_0)\Omega_0\right] z^3 \\ &\quad +O(z^4).\end{aligned} \tag{3.20}$$

These logarithmic versions of the Hubble law have several advantages — fits to these relations are easily calculated in terms of the observationally reported distance moduli μ_D and their estimated statistical uncertainties [28, 1, 2, 3, 38]. (Specifically there is no need to transform the statistical uncertainties on the distance moduli beyond a universal multiplication by the factor $[\ln 10]/5$.) Furthermore the deceleration parameter q_0 is easy to extract as it has been ***untangled*** from both Hubble parameter and the combination $(j_0 + \Omega_0)$.

Note that it is always the combination $(j_0 + \Omega_0)$ that arises in these third-order terms of the Hubble relations, and that it is even in principle impossible to separately determine j_0 and Ω_0 in a cosmographic framework. When looking at the fourth-order terms, it becomes impossible to separately determine j_0, s_0 and Ω_0 in this framework. The reason for this degeneracy is (or should be) well-known [20, p. 451]: Consider the ***exact*** expression for the luminosity distance in any FLRW universe, which is usually presented in the form [20, 21]

$$d_L(z) = a_0\ (1+z)\ \text{sin}_k\left\{\frac{c}{H_0\, a_0}\int_0^z \frac{H_0}{H(z)}\,\mathrm{d}z\right\}, \tag{3.21}$$

where

$$\text{sin}_k(x) = \begin{cases} \sin(x), & k=+1; \\ x, & k=0; \\ \sinh(x), & k=-1. \end{cases} \tag{3.22}$$

By inspection, even if one knows $H(z)$ exactly for all z one cannot determine $d_L(z)$ without independent knowledge of k and a_0. Conversely even if one knows $d_L(z)$ exactly for all z one cannot determine $H(z)$ without independent knowledge of k and a_0. Indeed let us rewrite this exact result in a slightly different fashion as

$$d_L(z) = a_0\ (1+z)\ \frac{\sin\left\{\dfrac{\sqrt{k}\,d_H}{a_0}\displaystyle\int_0^z \frac{H_0}{H(z)}\,\mathrm{d}z\right\}}{\sqrt{k}}, \tag{3.23}$$

where this result now holds for all k provided we interpret the $k=0$ case in the obvious limiting fashion. Equivalently, using the cosmographic Ω_0 as defined above we have the ***exact*** cosmographic result that for all Ω_0:

$$d_L(z) = d_H\ (1+z)\ \frac{\sin\left\{\sqrt{\Omega_0 - 1}\displaystyle\int_0^z \frac{H_0}{H(z)}\,\mathrm{d}z\right\}}{\sqrt{\Omega_0-1}}. \tag{3.24}$$

This form of the exact Hubble relation makes it clear that an independent determination of Ω_0 (equivalently, k/a_0^2), is needed to complete the link between $a(t)$ and $d_L(z)$. When Taylor

expanded in terms of z, this expression leads to a degeneracy at third-order, which is where Ω_0 [equivalently k/a_0^2] first enters into the Hubble series [24, 25].

What message should we take from this discussion? There are many physically equivalent versions of the Hubble law, corresponding to many slightly different physically reasonable definitions of distance, and whether we choose to present the Hubble law linearly or logarithmically. If one were to have arbitrarily small scatter/error bars on the observational data, then the choice of which Hubble law one chooses to fit to would not matter. In the presence of significant scatter/uncertainty there is a risk that the fit might depend strongly on the choice of Hubble law one chooses to work with. (And if the resulting values of the deceleration parameter one obtains do depend significantly on which distance scale one uses, this is evidence that one should be very cautious in interpreting the results.) Note that the two versions of the Hubble law based on "photon flux distance" d_F stand out in terms of making the deceleration parameter easy to visualize and extract.

3.2 Why is the redshift expansion badly behaved for $z > 1$?

In addition to the question of which distance measure one chooses to use, there is a basic and fundamental physical and mathematical reason why the traditional redshift expansion breaks down for $z > 1$.

3.2.1 Convergence

Consider the exact Hubble relation (3.21). This is certainly nicely behaved, and possesses no obvious poles or singularities, (except possibly at a turnaround event where $H(z) \to 0$, more on this below). However if we attempt to develop a Taylor series expansion in redshift z, using what amounts to the ***definition*** of the Hubble H_0, deceleration q_0, and jerk j_0 parameters, then:

$$\frac{1}{1+z} = \frac{a(t)}{a_0} = 1 + H_0\,(t-t_0) - \frac{q_0\,H_0^2}{2!}\,(t-t_0)^2 + \frac{j_0\,H_0^3}{3!}\,(t-t_0)^3 + \frac{1}{4!}\,s_0\,H_0^4\,(t-t_0)^4 + O([t-t_0]^5). \tag{3.25}$$

Now this particular Taylor expansion manifestly has a pole at $z = -1$, corresponding to the instant (either at finite or infinite time) when the universe has expanded to infinite volume, $a = \infty$. Note that a ***negative*** value for z corresponds to $a(t) > a_0$, that is: In an expanding universe $z < 0$ corresponds to the ***future***. Since there is an explicit pole at $z = -1$, by standard complex variable theory the radius of convergence is ***at most*** $|z| = 1$, so that this series ***also*** fails to converge for $z > 1$, when the universe was less than half its current size.

Consequently when reverting this power series to obtain lookback time $T = t_0 - t$ as a function $T(z)$ of z, we should not expect that series to converge for $z > 1$. Ultimately, when written in terms of a_0, H_0, q_0, j_0, and a power series expansion in redshift z you should not expect $d_L(z)$ to converge for $z > 1$.

Note that the ***mathematics*** that goes into this result is that the radius of convergence of any power series is the distance to the closest singularity in the complex plane, while the relevant ***physics*** lies in the fact that on physical grounds we should not expect to be able to extrapolate forwards beyond $a = \infty$, corresponding to $z = -1$. Physically we should

expect this argument to hold for any observable quantity when expressed as a function of redshift and Taylor expanded around $z = 0$ — the radius of convergence of the Taylor series must be less than or equal to unity. (Note that the radius of convergence might actually be ***less*** than unity, this occurs if some other singularity in the complex z plane is closer than the breakdown in predictability associated with attempting to drive $a(t)$ "***past***" infinite expansion, $a = \infty$.) Figure 3.1 illustrates the radius of convergence in the complex plane of the Taylor series expansion in terms of z.

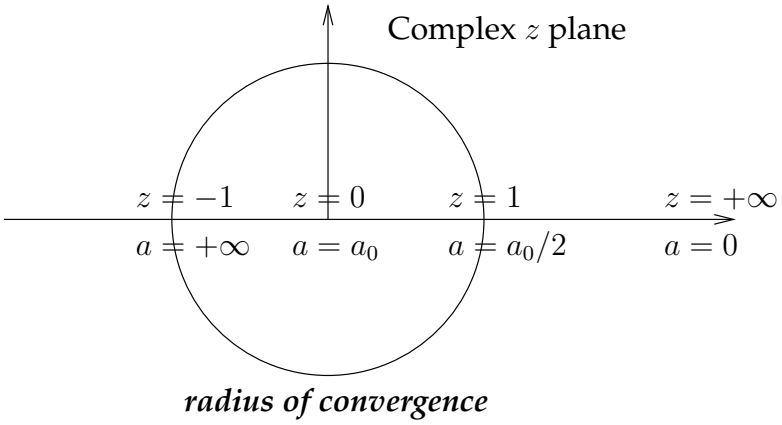


Figure 3.1: Qualitative sketch of the behaviour of the scale factor a and the radius of convergence of the Taylor series in z-redshift.

Consequently, we must conclude that observational data regarding $d_L(z)$ for $z > 1$ is not going to be particularly useful in fitting a_0, H_0, q_0, and j_0, to the usual ***traditional*** version of the Hubble relation.

3.2.2 Pivoting

A trick that is sometimes used to improve the behaviour of the Hubble law is to Taylor expand around some nonzero value of z, which might be called the "pivot". That is, we take

$$z = z_{pivot} + \Delta z, \tag{3.26}$$

and expand in powers of Δz. If we choose to do so, then observe

$$\begin{aligned} \frac{1}{1 + z_{pivot} + \Delta z} &= 1 + H_0\,(t - t_0) - \frac{1}{2}\,q_0\,H_0^2\,(t - t_0)^2 + \frac{1}{3!}\,j_0\,H_0^3\,(t - t_0)^3 \\ &\quad - \frac{1}{4!}\,s_0\,H_0^4\,(t - t_0)^4 + O([t - t_0]^5). \end{aligned} \tag{3.27}$$

The pole is now located at:

$$\Delta z = -(1 + z_{pivot}), \tag{3.28}$$

which again physically corresponds to a universe that has undergone infinite expansion, $a = \infty$. The radius of convergence is now

$$|\Delta z| \leq (1 + z_{pivot}), \tag{3.29}$$

and we expect the pivoted version of the Hubble law to fail for

$$z > 1 + 2\, z_{pivot}. \tag{3.30}$$

So pivoting is certainly helpful, and can in principle extend the convergent region of the Taylor expanded Hubble relation to somewhat higher values of z, but maybe we can do even better?

3.2.3 Other singularities

Other singularities that might further restrict the radius of convergence of the Taylor expanded Hubble law (or any other Taylor expanded physical observable) are also important. Chief among them are the singularities (in the Taylor expansion) induced by turnaround events. If the universe has a minimum scale factor $a_{\rm min}$ (corresponding to a ***bounce***) then clearly it is meaningless to expand beyond

$$1 + z_{\rm max} = a_0/a_{\rm min}; \qquad z_{\rm max} = a_0/a_{\rm min} - 1; \tag{3.31}$$

implying that we should restrict our attention to the region

$$|z| < z_{\rm max} = a_0/a_{\rm min} - 1. \tag{3.32}$$

Since for other reasons we had already decided we should restrict attention to $|z| < 1$, and since on observational grounds we certainly expect any ***bounce***, if it occurs at all, to occur for $z_{\rm max} \gg 1$, this condition provides no new information.

On the other hand, if the universe has a moment of maximum expansion, and then begins to recollapse, then it is meaningless to extrapolate beyond

$$1 + z_{\rm min} = a_0/a_{\rm max}; \qquad z_{\rm min} = -[1 - a_0/a_{\rm max}]; \tag{3.33}$$

implying that we should restrict our attention to the region

$$|z| < 1 - a_0/a_{\rm max}. \tag{3.34}$$

This relation now does provide us with additional constraint, though (compared to the $|z| < 1$ condition) the bound is not appreciably tighter unless we are "***close***" to a point of maximum expansion. Other singularities could lead to additional constraints.

3.3 Improved redshift variable for the Hubble relation

Now it must be admitted that the traditional redshift has a particularly simple physical interpretation:

$$1 + z = \frac{\lambda_0}{\lambda_e} = \frac{a(t_0)}{a(t_e)}, \tag{3.35}$$

so that

$$z = \frac{\lambda_0 - \lambda_e}{\lambda_e} = \frac{\Delta\lambda}{\lambda_e}. \tag{3.36}$$

That is, z is the change in wavelength divided by the ***emitted*** wavelength. This is certainly simple, but there's at least one other ***equally simple*** choice. Why not use:

$$y = \frac{\lambda_0 - \lambda_e}{\lambda_0} = \frac{\Delta\lambda}{\lambda_0} \; ? \tag{3.37}$$

That is, define y to be the change in wavelength divided by the ***observed*** wavelength. This implies

$$1 - y = \frac{\lambda_e}{\lambda_0} = \frac{a(t_e)}{a(t_0)} = \frac{1}{1+z}. \tag{3.38}$$

Now similar expansion variables have certainly been considered before. (See, for example, Chevalier and Polarski [44], who effectively worked with the dimensionless quantity $b = a(t)/a_0$, so that $y = 1 - b$. Similar ideas have also appeared in several related works [45, 46, 47, 48]. Note that these authors have typically been interested in parameterizing the so-called w-parameter, rather than specifically addressing the Hubble relation.)

Indeed, the variable y introduced above has some very nice properties:

$$y = \frac{z}{1+z}; \qquad z = \frac{y}{1-y}. \tag{3.39}$$

In the past (of an expanding universe)

$$z \in (0, \infty); \qquad y \in (0, 1); \tag{3.40}$$

while in the future

$$z \in (-1, 0); \qquad y \in (-\infty, 0). \tag{3.41}$$

So the variable y is both easy to compute, and when extrapolating back to the Big Bang has a nice finite range $(0, 1)$. We will refer to this variable as the ***y-redshift***. (Originally when developing these ideas we had intended to use the variable y to develop orthogonal polynomial expansions on the finite interval $y \in [0, 1]$. This is certainly possible, but we shall soon see that given the current data, this is somewhat overkill, and simple polynomial fits in y are adequate for our purposes.)

In terms of the variable y it is easy to extract a new version of the Hubble law by simple substitution:

$$\begin{aligned} d_L(y) &= d_H\, y \Bigg\{ 1 - \frac{1}{2}\left[-3 + q_0\right] y + \frac{1}{6}\left[12 - 5q_0 + 3q_0^2 - (j_0 + \Omega_0)\right] y^2 \\ &\quad + \frac{1}{24}\left[50 - 26q_0 - 21q_0^2 - 15q_0^3 + 10q_0 j_0 - 7j_0 + s_0 + 2(-5 + 3q_0)\Omega_0\right] y^3 + O(y^4) \Bigg\}. \end{aligned} \tag{3.42}$$

This still looks rather messy, in fact as messy as before — one might justifiably ask in what sense is this new variable any real improvement?

First, when expanded in terms of y, the formal radius of convergence covers much more of the physically interesting region. Consider:

$$\begin{aligned} 1 - y &= 1 + H_0\,(t - t_0) - \frac{1}{2}\, q_0\, H_0^2\,(t - t_0)^2 + \frac{1}{3!} j_0\, H_0^3\,(t - t_0)^3 \\ &\quad - \frac{1}{4!}\, s_0\, H_0^4\,(t - t_0)^4 + O([t - t_0]^5). \end{aligned} \tag{3.43}$$

This expression now has no poles, so upon reversion of the series lookback time $T = t_0 - t$ should be well behaved as a function $T(y)$ of y — at least all the way back to the Big Bang. (We now expect, on physical grounds, that the power series is likely to break down if one tries to extrapolate backwards ***through*** the Big Bang.) Based on this, we now expect $d_L(y)$, as long as it is expressed as a Taylor series in the variable y, to be a well-behaved power series all the way to the Big Bang. In fact, since

$$y = +1 \qquad \Leftrightarrow \qquad \text{Big Bang}, \tag{3.44}$$

we expect the radius of convergence to be given by $|y| = 1$, so that the series converges for

$$|y| < 1. \tag{3.45}$$

Consequently, when looking into the future, in terms of the variable y we expect to encounter problems at $y = -1$, when the universe has expanded to twice its current size. Figure 3.2 illustrates the radius of convergence in the complex plane of the Taylor series expansion in terms of y.

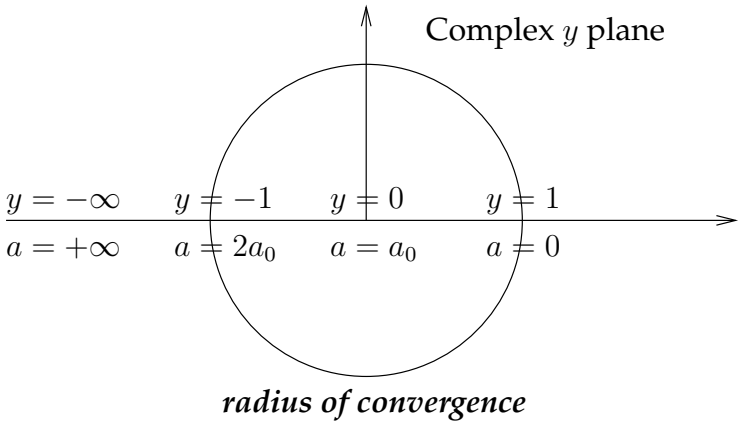


Figure 3.2: Qualitative sketch of the behaviour of the scale factor a and the radius of convergence of the Taylor series in y-redshift.

Note the tradeoff here — z is a useful expansion parameter for arbitrarily large universes, but breaks down for a universe half its current size or less; in contrast y is a useful expansion parameter all the way back to the Big Bang, but breaks down for a universe double its current size or more. Whether or not y is more suitable than z depends very much on what you are interested in doing. This is illustrated in Figures 3.1 and 3.2. For the purposes of this chapter we are interested in high-redshift supernovae — and we want to probe rather early times — so it is definitely y that is more appropriate here. Indeed the furthest supernova for which we presently have both spectroscopic data and an estimate of the distance occurs at $z = 1.755$ [3], corresponding to $y = 0.6370$. Furthermore, using the variable y it is easier to plot very large redshift datapoints. For example, (though we shall not pursue this point in this chapter), the Cosmological Microwave Background is located at $z_{\text{CMB}} = 1088$, which corresponds to $y_{\text{CMB}} = 0.999$. This point is not ***out of range*** as it would be if one uses the variable z.

3.4 More versions of the Hubble law

In terms of this new redshift variable, the ***linear in distance*** Hubble relations are:

$$d_L(y) = d_H\, y\Bigg\{1-\frac{1}{2}\left[-3+q_0\right]y+\frac{1}{6}\left[12-5q_0+3q_0^2-(j_0+\Omega_0)\right]y^2 \qquad (3.46)$$
$$+\frac{1}{24}\left[50-26q_0-21q_0^2-15q_0^3+10q_0j_0-7j_0+s_0+2(-5+3q_0)\Omega_0\right]y^3+O(y^4)\Bigg\}.$$

$$d_F(y) = d_H\, y\Bigg\{1-\frac{1}{2}\left[-2+q_0\right]y+\frac{1}{24}\left[27-14q_0+12q_0^2-4(j_0+\Omega_0)\right]y^2 \qquad (3.47)$$
$$+\frac{1}{24}\left[44-29q_0+30q_0^2-30q_0^3+20q_0j_0-10j_0+2s_0+4(-4+3q_0)\Omega_0\right]y^3+O(y^4)\Bigg\}.$$

$$d_P(y) = d_H\, y\Bigg\{1-\frac{1}{2}\left[-1+q_0\right]y+\frac{1}{6}\left[3-2q_0+3q_0^2-(j_0+\Omega_0)\right]y^2 \qquad (3.48)$$
$$+\frac{1}{24}\left[6-6q_0+9q_0^2-15q_0^3+10q_0j_0-3j_0+s_0+6(1+q_0)\Omega_0\right]y^3+O(y^4)\Bigg\}.$$

$$d_Q(y) = d_H\, y\Bigg\{1-\frac{q_0}{2}y+\frac{1}{12}\left[3-2q_0+12q_0^2-4(j_0+\Omega_0)\right]y^2 \qquad (3.49)$$
$$+\frac{1}{48}\left[-2-q_0+6q_0^2-30q_0^3+20q_0j_0-2j_0+s_0+4(-2+3q_0)\Omega_0\right]y^3+O(y^4)\Bigg\}.$$

$$d_A(y) = d_H\, y\Bigg\{1-\frac{1}{2}\left[1+q_0\right]y+\frac{1}{6}\left[q_0+3q_0^2-(j_0+\Omega_0)\right]y^2 \qquad (3.50)$$
$$+\frac{1}{24}\left[-2+2q_0-3q_0^2-15q_0^3+10q_0j_0+j_0+s_0+2(-1+3q_0)\Omega_0\right]y^3+O(y^4)\Bigg\}.$$

Note that in terms of the y variable it is the "deceleration distance" d_Q that has the deceleration parameter q_0 appearing in the simplest manner. Similarly, the "logarithmic in distance" Hubble relations are:

$$\ln[d_L/(y\,\text{Mpc})] = \frac{\ln 10}{5}[\mu_D-25]-\ln y \qquad (3.51)$$
$$= \ln(d_H/\text{Mpc})$$
$$-\frac{1}{2}\left[-3+q_0\right]y+\frac{1}{24}\left[21-2q_0+9q_0^2-4(j_0+\Omega_0)\right]y^2$$
$$+\frac{1}{24}\left[11-q_0+2q_0^2-10q_0^3+8q_0j_0-j_0+s_0+4(-1+q_0)\Omega_0\right]y^3+O(y^4),$$

$$\begin{aligned}\ln[d_F/(y\ \text{Mpc})] &= \frac{\ln 10}{5}[\mu_D - 25] - \ln y + \frac{1}{2}\ln(1-y) \\ &= \ln(d_H/\text{Mpc}) \\ &\quad -\frac{1}{2}\left[-2+q_0\right]y + \frac{1}{24}\left[15 - 2q_0 + 9q_0^2 - 4(j_0+\Omega_0)\right]y^2 \\ &\quad +\frac{1}{24}\left[7 - q_0 + 2q_0^2 - 10q_0^3 + 8q_0 j_0 - j_0 + s_0 + 4(-1+q_0)\Omega_0\right]y^3 + O(y^4),\end{aligned} \tag{3.52}$$

$$\begin{aligned}\ln[d_P/(y\ \text{Mpc})] &= \frac{\ln 10}{5}[\mu_D - 25] - \ln y + \ln(1-y) \\ &= \ln(d_H/\text{Mpc}) \\ &\quad -\frac{1}{2}\left[-1+q_0\right]y + \frac{1}{24}\left[9 - 2q_0 + 9q_0^2 - 4(j_0+\Omega_0)\right]y^2 \\ &\quad +\frac{1}{24}\left[3 - q_0 + 2q_0^2 - 10q_0^3 + 8q_0 j_0 - j_0 + s_0 + 4(-1+q_0)\Omega_0\right]y^3 + O(y^4),\end{aligned} \tag{3.53}$$

$$\begin{aligned}\ln[d_Q/(y\ \text{Mpc})] &= \frac{\ln 10}{5}[\mu_D - 25] - \ln y + \frac{3}{2}\ln(1-y) \\ &= \ln(d_H/\text{Mpc}) \\ &\quad -\frac{1}{2}q_0\, y + \frac{1}{24}\left[3 - 2q_0 + 9q_0^2 - 4(j_0+\Omega_0)\right]y^2 \\ &\quad +\frac{1}{24}\left[-1 - q_0 + 2q_0^2 - 10q_0^3 + 8q_0 j_0 - j_0 + s_0 + 4(-1+q_0)\Omega_0\right]y^3 + O(y^4),\end{aligned} \tag{3.54}$$

$$\begin{aligned}\ln[d_A/(y\ \text{Mpc})] &= \frac{\ln 10}{5}[\mu_D - 25] - \ln y + 2\ln(1-y) \\ &= \ln(d_H/\text{Mpc}) \\ &\quad -\frac{1}{2}\left[1+q_0\right]y + \frac{1}{24}\left[-3 - 2q_0 + 9q_0^2 - 4(j_0+\Omega_0)\right]y^2 \\ &\quad +\frac{1}{24}\left[3 - q_0 + 2q_0^2 - 10q_0^3 + 8q_0 j_0 - j_0 + s_0 + 4(-1+q_0)\Omega_0\right]y^3 + O(y^4).\end{aligned} \tag{3.55}$$

Again note that the ***logarithmic in distance*** versions of the Hubble law are attractive in terms of maximizing the disentangling between Hubble distance, deceleration parameter, and jerk. Now having a selection of Hubble laws on hand, we can start to confront the observational data to see what it is capable of telling us.

3.5 Cosmic microwave background

A particularly interesting feature of the plots we generate is that we can meaningfully obtain a "***global***" view of the situation by plotting data coming from the CMB on the same plot as the supernovae. Recall that decoupling of the CMB occurs at $z_{\text{CMB}} = 1088$, which corresponds to $y_{\text{CMB}} = 0.999$. The distance to the CMB fireball is estimated as [13]

$$d_{P,\text{CMB}} = 13.8 \pm 1.1\ \text{Gpc}. \tag{3.56}$$

This is a (Peebles-style) angular diameter distance: It depends on the observed angular size of CMB fluctuations,

$$\theta_A = 0.6 \pm 0.01 \text{ degrees} \tag{3.57}$$

and the estimated size of the acoustic horizon at decoupling

$$r_{\text{sound horizon}} = 146 \pm 10 \text{ Mpc}. \tag{3.58}$$

In contast to the supernova data, which is completely cosmographic, the CMB data point does depend on the Friedmann equations, but only rather "***weakly***" [13]. The fact that we can put plot both CMB data and the supernovae on the same plot ultimately works because

$$H_0 = 73 \text{ (km/s)/Mpc} \qquad \Rightarrow \qquad \frac{c}{H_0} = 4110 \text{ Mpc} \approx \frac{d_{P,\text{CMB}}}{3}. \tag{3.59}$$

There is a subtle coincidence behind this. From [12] we have

$$d_{P,\text{CMB}} = \frac{c}{H_0} \; \frac{\displaystyle\int_{y_{\text{CMB}}}^{1} \frac{1}{\sqrt{\Omega_m(1-y)+\Omega_r}} \; \frac{1}{1+(3\Omega_b[1-y])/(4\Omega_\gamma)} \; \mathrm{d}y}{\theta_A}, \tag{3.60}$$

which depends on the interval $y \in (y_{\text{CMB}}, 1)$ — from decoupling to the big bang — or more precisely from decoupling to the end of cosmological inflation. Observationally the integral in the numerator is about a factor of 3 larger than the angular diameter in the denominator. This appears to be an instance of cosmological fine tuning.

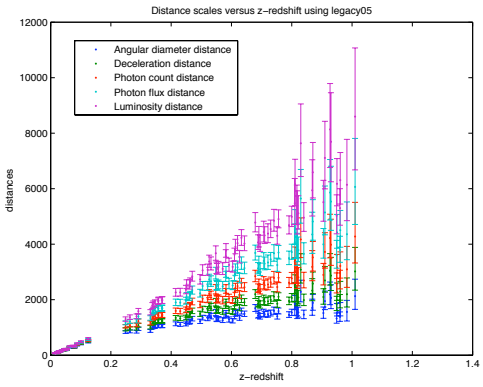
3.6 Supernova data

For the plots below we have used data from the ***supernova legacy survey*** (legacy05) [28, 2] and the Riess *et. al.* ***gold*** dataset of 2006 (gold06) [3]. Refer to sections 2.7.4 and 2.7.5 for details on the supernova data.

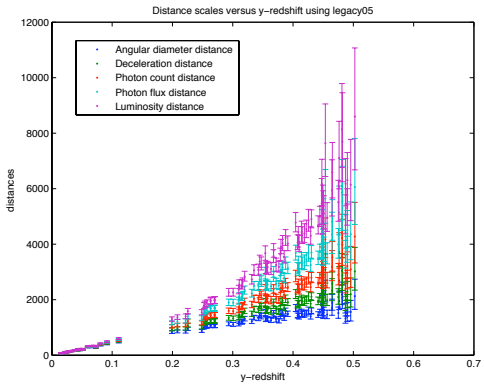
3.6.1 The legacy05 dataset

Figures 3.3(a) and 3.3(c) illustrate the distance scales defined previously plotted as a function of the z-redshift, whereas figures 3.3(b) and 3.3(d) show these same distances plotted as a function of the y-redshift . Mainly, Note that all versions of the Hubble law are linear at low redshift, and that differences first arise in the non-linear part of the relation beyond $y \approx 0.2$.

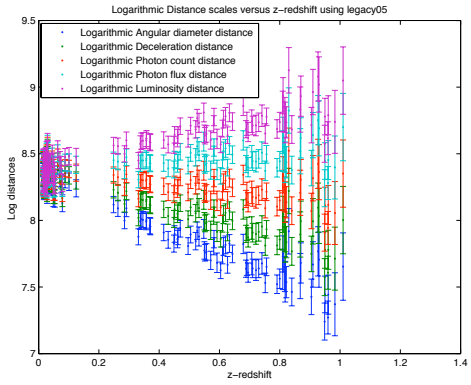
To orient oneself, figure 3.4(a) focuses on the deceleration distance $d_Q(y)$, and plots $\ln(d_Q/[y \text{ Mpc}])$ versus y. Visually, the curve appears close to flat, at least out to $y \approx 0.4$, which is an unexpected oddity that merits further investigation — since it seems to imply an "***eyeball estimate***" that $q_0 \approx 0$. Note that this is ***not*** a plot of ***statistical residuals*** obtained after curve fitting — rather this can be interpreted as a plot of "***theoretical residuals***", obtained by first splitting off the linear part of the Hubble law (which is now encoded in the intercept with the vertical axis), and secondly choosing the quantity to be plotted so as to make the slope of the curve at zero particularly easy to interpret in terms of the deceleration parameter. The fact that there is considerable "***scatter***" in the plot should not be



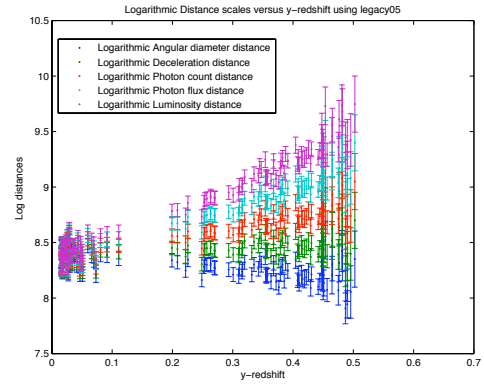
(a) Distance scales versus the z-redshift



(b) Distance scales versus the y-redshift

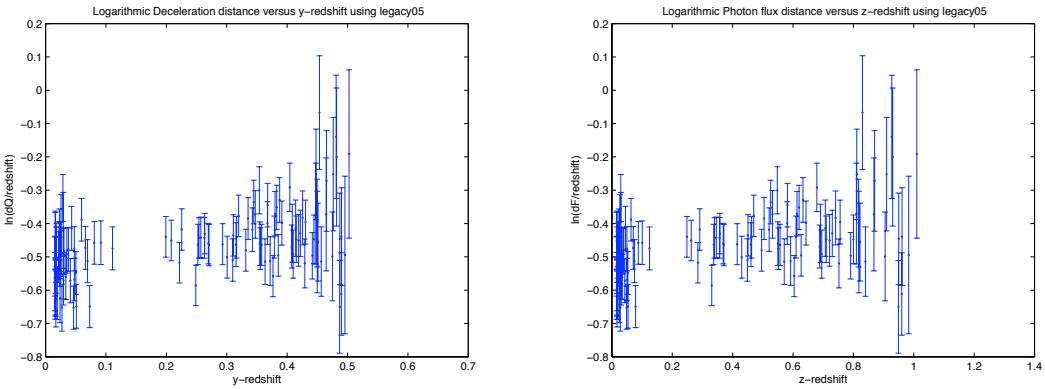


(c) Logarithmic distance scales versus the z-redshift



(d) Logarithmic distance scales versus the y-redshift

Figure 3.3: Various distance scales as a function of the z and y-redshift using the nearby and legacy dataset [1].



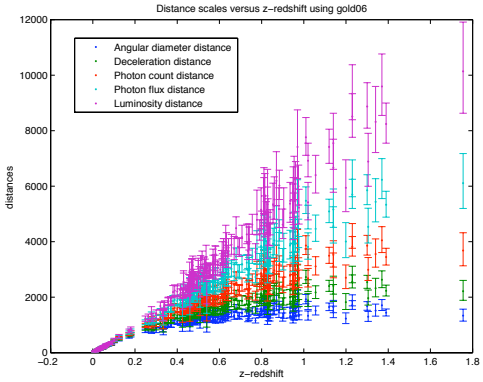
(a) $\ln(d_Q/(y\,\text{Mpc}))$ versus the y-redshift

(b) $\ln(d_F/(z\,\text{Mpc}))$ versus the z-redshift

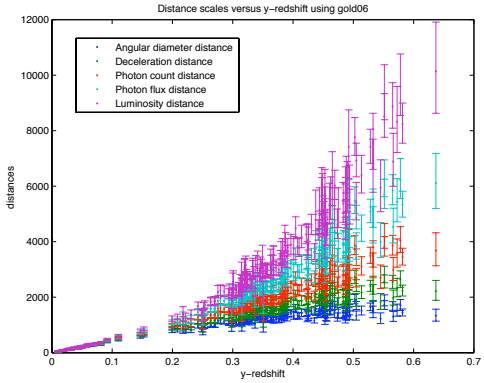
Figure 3.4: The normalized logarithms of the deceleration distance $\ln(d_Q/(y\,\text{Mpc}))$ as a function of the y-redshift (a) and of the photon flux distance $\ln(d_F/(z\,\text{Mpc}))$ as a function of the z-redshift (b), using the legacy05 dataset [1].

thought of as an artifact due to a "***bad***" choice of variables — instead this choice of variables should be thought of as "***good***" in the sense that they provide an honest basis for dispassionately assessing the quality of the data that currently goes into determining the deceleration parameter. Similarly, figure 3.4(b) focuses on the photon flux distance $d_F(z)$, and plots $\ln(d_F/[z\,\text{Mpc}])$ versus z. Visually, this curve is again very close to flat, at least out to $z \approx 0.4$. This again gives one a feel for just how tricky it is to reliably estimate the deceleration parameter q_0 from the data.

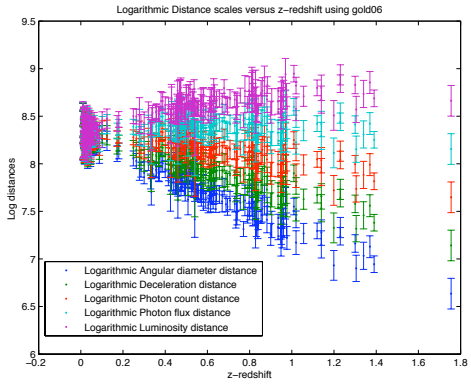
3.6.2 The gold06 dataset



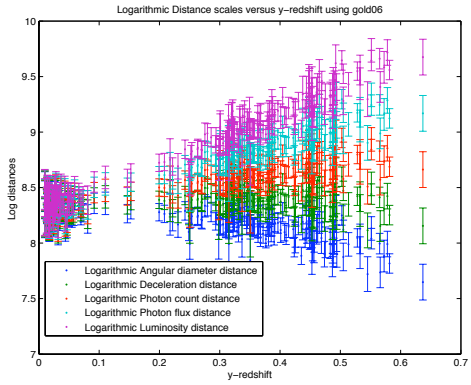
(a) Distance scales versus the z-redshift



(b) Distance scales versus the y-redshift



(c) Logarithmic distance scales versus the z-redshift

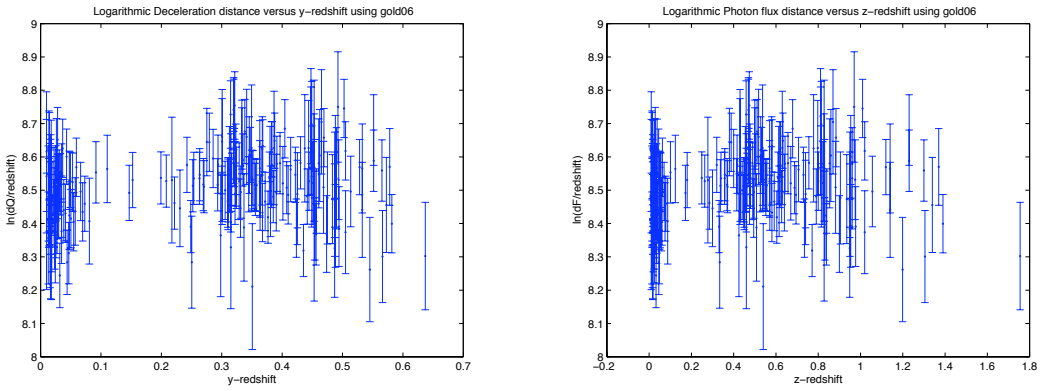


(d) Logarithmic distance scales versus the y-redshift

Figure 3.5: Various distance scales as a function of the z and y-redshift using the gold06 dataset [2, 3].

Figures 3.5(a) and 3.5(c) illustrate the distance scales defined previously plotted as a function of the z-redshift, whereas figures 3.5(b) and 3.5(d) show these same distances plotted as a function of the y-redshift . To orient oneself, figure 3.6(a) again focusses on the normalized logarithm of the deceleration distance $d_Q(y)$ as a function of y-redshift. Similarly, figure 3.6(b) focusses on the normalized logarithm of the photon flux distance $d_F(z)$ as a function of z-redshift. Visually, these curves are again very close to flat out to $y \approx 0.4$ and $z \approx 0.4$ respectively, which seems to imply an "eyeball estimate" that $q_0 \approx 0$. Again, this gives one a feel for just how tricky it is to reliably estimate the deceleration parameter q_0 from the data.

Note the outlier at $y = 0.6370$, that is, $z = 1.755$. In particular, observe that adopting the y-redshift in place of the z-redshift has the effect of pulling this outlier "closer" to the main body of data, thus reducing its "leverage" effect on any data fitting one undertakes — apart from the theoretical reasons we have given for preferring the y-redshift, (improved conver-



(a) $\ln(d_Q/(y\,\mathrm{Mpc}))$ versus the y-redshift

(b) $\ln(d_F/(z\,\mathrm{Mpc}))$ versus the z-redshift

Figure 3.6: The normalized logarithms of the deceleration distance $\ln(d_Q/(y\,\mathrm{Mpc}))$ as a function of the y-redshift (a) and of the photon flux distance $\ln(d_F/(z\,\mathrm{Mpc}))$ as a function of the z-redshift (b), using the gold06 dataset [2, 3].

gence behaviour for the Taylor series), the fact that it automatically reduces the leverage of high redshift outliers is a feature that is considered highly desirable purely for statistical reasons. In particular, the method of least-squares is known to be non-robust with respect to outliers. One could implement more robust regression algorithms, but they are not as easy and fast as the classical least-squares method. We have also implemented least-squares regression against a reduced dataset where we have trimmed out the most egregious high-z outlier, and also eliminated the so-called "***Hubble bubble***" for $z < 0.0233$ [49, 50]. While the precise numerical values of our estimates for the cosmological parameters then change, there is no great qualitative change to the points we wish to make in this chapter, nor to the conclusions we will draw.

3.6.3 Peculiar velocities

One point that should be noted for both the legacy05 and gold06 datasets is the way that peculiar velocities have been treated. While peculiar velocities would physically seem to be best represented by assigning an uncertainty to the measured redshift, in both these datasets the peculiar velocities have instead been modelled as some particular function of z-redshift and then lumped into the reported uncertainties in the distance modulus. Working with the y-redshift ***ab initio*** might lead one to re-assess the model for the uncertainty due to peculiar velocities. We expect such effects to be small and have not considered them in detail.

3.7 Data fitting: Statistical uncertanties

We shall now compare and contrast the results of multiple least-squares fits to the different notions of cosmological distance, using the two distinct redshift parameterizations discussed above. Specifically, we use a finite-polynomial truncated Taylor series as our model, and perform classical least-squares fits. This is effectively a test of the robustness of the

data-fitting procedure, testing it for model dependence. For general background information see [51, 52, 53, 54, 55, 56, 57].

3.7.1 Finite-polynomial truncated-Taylor-series fit

Working (for purposes of the presentation) in terms of y-redshift, the various distance scales can be fitted to finite-length power-series polynomials $d(y)$ of the form

$$P(y): \quad d(y) = \sum_{j=0}^{n} a_j \, y^j, \tag{3.61}$$

where the coefficients a_j all have the dimensions of distance. In contrast, logarithmic fits are of the form

$$P(y): \quad \ln[d(y)/(y \text{ Mpc})] = \sum_{j=0}^{n} b_j \, y^j, \tag{3.62}$$

where the coefficients b_j are now all dimensionless. By fitting to finite polynomials we are implicitly making the assumption that the higher-order coefficients are all exactly zero — this does then implicitly enforce assumptions regarding the higher-order time derivatives $\mathrm{d}^m a/\mathrm{d}t^m$ for $m > n$, but there is no way to avoid making at least some assumptions of this type [51, 52, 53, 54, 55, 56, 57].

The method of least squares requires that we minimize

$$\chi^2 = \sum_{I=1}^{N} \left(\frac{P_I - P(y_I)}{\sigma_I} \right)^2, \tag{3.63}$$

where the N data points (y_I, P_I) represent the relevant function $P_I = f(\mu_{D,I}, y_I)$ of the distance modulus $\mu_{D,I}$ at corresponding y-redshift y_I, as inferred from some specific supernovae dataset. Furthermore $P(y_I)$ is the finite polynomial model evaluated at y_I. The σ_I are the total statistical uncertainty in P_I (including, in particular, intrinsic dispersion). The location of the minimum value of χ^2 can be determined by setting the derivatives of χ^2 with respect to each of the coefficients a_j or b_j equal to zero.

Note that the theoretical justification for using least squares assumes that the statistical uncertainties are normally distributed Gaussian uncertainties — and there is no real justification for this assumption in the actual data. Furthermore if the data is processed by using some nonlinear transformation, then in general Gaussian uncertainties will not remain Gaussian — and so even if the untransformed uncertainties are Gaussian the theoretical justification for using least squares is again undermined unless the scatter/uncertainties are small, [in the sense that $\sigma \ll f''(x)/f'(x)$], in which case one can appeal to a local linearization of the nonlinear data transformation $f(x)$ to deduce approximately Gaussian uncertainties [51, 52, 53, 54, 55, 56, 57]. As we have already seen, in figures 3.4(a)–3.6(b), there is again no real justification for this "small scatter" assumption in the actual data — nevertheless, in the absence of any clearly better data-fitting prescription, least squares is the standard way of proceeding. More statistically sophisticated techniques, such as "robust regression", have their own distinct draw-backs and, even with weak theoretical underpinning, χ^2 data-fitting is still typically the technique of choice [51, 52, 53, 54, 55, 56, 57].

We have performed least squares analyses, both linear in distance and logarithmic in distance, for all of the distance scales discussed above, d_L, d_F, d_P, d_Q, and d_A, both in terms of z-redshift and y-redshift, for finite polynomials from $n = 1$ (linear) to $n = 7$ (septic). We stopped at $n = 7$ since beyond that point the least squares algorithm was found to become numerically unstable due to the need to invert a numerically ill-conditioned matrix — this ill-conditioning is actually a well-known feature of high-order least-squares polynomial fitting. We carried out the analysis to such high order purely as a diagnostic — we shall soon see that the "most reasonable" fits are actually rather low order $n = 2$ quadratic fits.

3.7.2 χ^2 goodness of fit

A convenient measure of the goodness of fit is given by the reduced chi-square:

$$\chi^2_\nu = \frac{\chi^2}{\nu}, \tag{3.64}$$

where the factor $\nu = N - n - 1$ is the number of degrees of freedom left after fitting N data points to the $n + 1$ parameters.

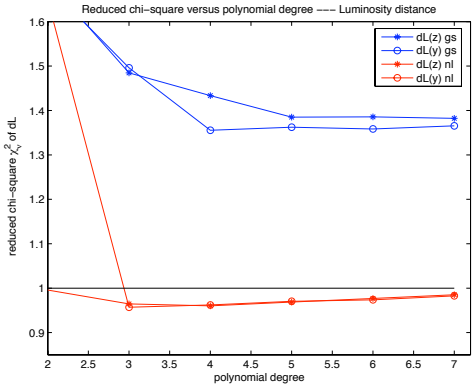
If the fitting function is a good approximation to the parent function, then the value of the reduced chi-square should be approximately unity $\chi^2_\nu \approx 1$. If the fitting function is not appropriate for describing the data, the value of χ^2_ν will be greater than 1. Also, "***too good***" a chi-square fit ($\chi^2_\nu < 1$) can come from over-estimating the statistical measurement uncertainties. Again, the theoretical justification for this test relies on the fact that one is assuming, without a strong empirical basis, Gaussian uncertainties [51, 52, 53, 54, 55, 56, 57]. Note that when the data is normalized to $H_0 = 70$ (km/sec)/Mpc, we are dealing with only $n - 1$ free parameters.

In all the cases we considered, for polynomials of order $n = 2$ and above, we found that $\chi^2_\nu \approx 1$ for the legacy05 dataset, and $\chi^2_\nu \approx 0.8 < 1$ for the gold06 dataset. Note that in Figure 3.7, the goodness of fit is $\chi^2_\nu \approx 1.4$ for the ***gs*** cases. The difference is that the larger χ^2_ν shows the effect of including the silver data points as well as the gold ones. Linear $n = 1$ fits often gave high values for χ^2_ν. We deduce that:

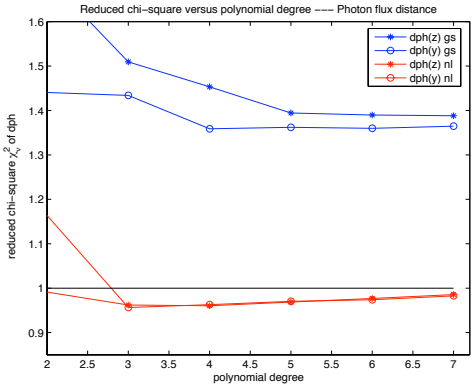
- It is desirable to keep at least quadratic $n = 2$ terms in all data fits.
- Caution is required when interpreting the reported statistical uncertainties in the gold06 dataset.

(In particular, note that some of the estimates of the statistical uncertainties reported in gold06 have themselves been determined through statistical reasoning — essentially by adjusting χ^2_ν to be "***reasonable***". The effects of such pre-processing become particularly difficult to untangle when one is dealing with a heterogeneous dataset.)

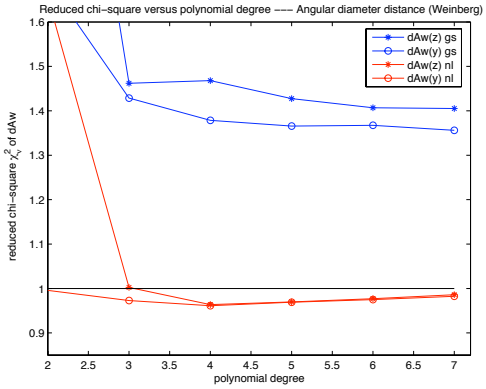
Figures 3.7(a), 3.7(b), 3.7(c), 3.7(d), and 3.7(e), show the goodness of fit of polynomials of degree 2 to 7 for each distance scale both in z-redshift and y-redshift for the two datasets ***gold and silver*** and ***nearby and legacy***. Note that for all the distance scales and the gold and silver dataset, the y-redshift seems to give better fits than the z-redshift, whereas, nothing can be said for the nearby and legacy dataset. This could result from the fact that there are more data with $z > 1$ in the gold and silver dataset than in the other set.



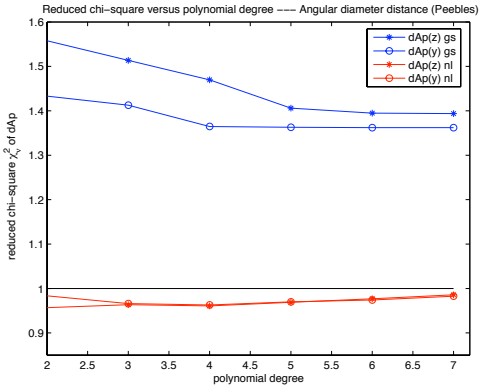
(a) Logarithm Luminosity distance scale



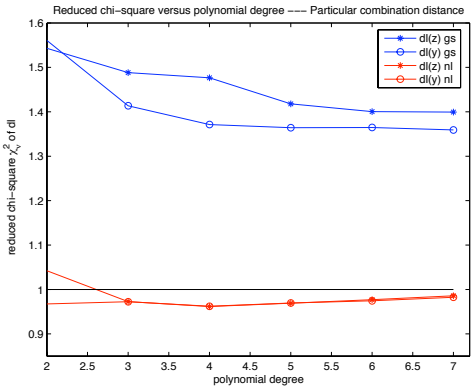
(b) Logarithm photon flux distance scale



(c) Logarithm Angular diameter distance scale (Weinberg)



(d) Logarithm Angular diameter distance scale (Peebles)



(e) Logarithm deceleration distance scale

Figure 3.7: Goodness of fit of polynomial data fitting to various distance scales as a function of the z-redshift and y-redshift, using the gold and silver dataset, and the nearby and legacy dataset.

3.7.3 F-test of additional terms

How many polynomial terms do we need to include to obtain a good approximation to the parent function?

The difference between two χ^2 statistics is also distributed as χ^2. In particular, if we fit a set of data with a fitting polynomial of $n-1$ parameters, the resulting value of chi-square associated with the deviations about the regression $\chi^2(n-1)$ has $N-n$ degrees of freedom. If we add another term to the fitting polynomial, the corresponding value of chi-square $\chi^2(n)$ has $N-n-1$ degrees of freedom. The difference between these two follows the χ^2 distribution with one degree of freedom.

The F_χ statistic follows a F distribution with $\nu_1 = 1$ and $\nu_2 = N-n-1$,

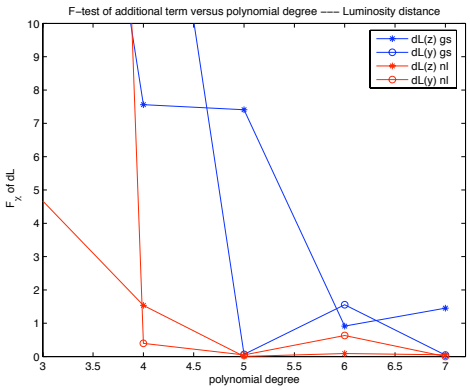
$$F_\chi = \frac{\chi^2(n-1) - \chi^2(n)}{\chi^2(n)/(N-n-1)}. \tag{3.65}$$

This ratio is a measure of how much the additional term has improved the value of the reduced chi-square. F_χ should be small when the function with n coefficients does not significantly improve the fit over the polynomial fit with $n-1$ terms.

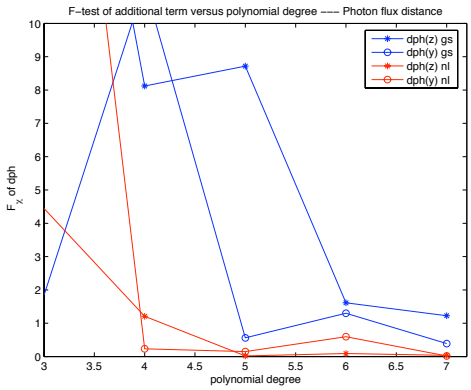
In all the cases we considered, the F_χ statistic was not significant when one proceeded beyond $n = 2$. We deduce that:

- It is statistically meaningless to go beyond $n = 2$ terms in the data fits.
- This means that one can ***at best*** hope to estimate the deceleration parameter and the jerk (or more precisely the combination $j_0 + \Omega_0$). There is no meaningful hope of estimating the snap parameter from the current data.

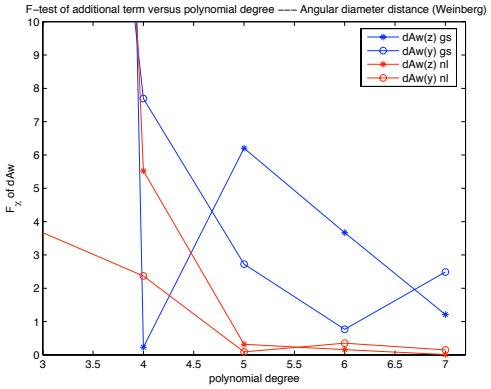
Figures 3.8(a), 3.8(b), 3.8(c), 3.8(d) and 3.8(e), illustrate the F-test of additional terms of polynomial data fitting for various distance scales as a function of the z-redshift and y-redshift for both the ***gold and silver*** and the ***nearby and legacy*** datasets. At the polynomial degree n, we calculate the terms in equation (3.65), if F_χ is small the term of order n is not needed in the polynomial fit. Overall, it seems that fitting the distance scales by a polynomial of order 3 or 4 is a rather good approximation in the case of ***nearby and legacy*** data points. With the ***gold and silver*** data, the degree of fitting polynomials needs to be higher especially when using the z-redshift. In this case F_χ seems to be oscillating significantly, however, it seems to be more consistently decreasing when using the variable y.



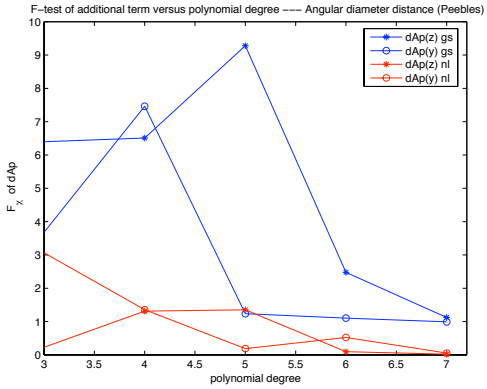
(a) Logarithm Luminosity distance scale



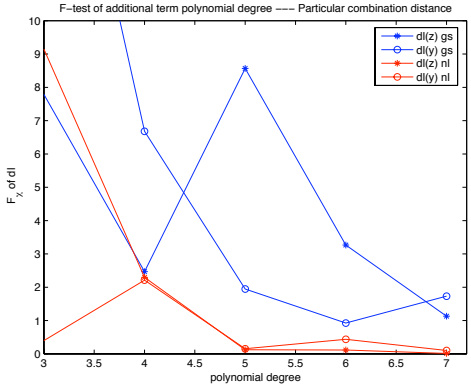
(b) Logarithm photon flux distance scale



(c) Logarithm Angular diameter distance scale (Weinberg)



(d) Logarithm Angular diameter distance scale (Peebles)



(e) Logarithm deceleration distance scale

Figure 3.8: F-test of additional terms for various distance scales as a function of the z-redshift and y-redshift, using the gold and silver dataset, and the nearby and legacy dataset.

3.7.4 Uncertainties in the coefficients a_j and b_j

From the fit one can determine the standard deviations σ_{a_j} and σ_{b_j} for the uncertainty of the polynomial coefficients a_j or b_j. It is the root sum square of the products of the standard deviation of each data point σ_i, multiplied by the effect that the data point has on the determination of the coefficient a_j [51]:

$$\sigma_{a_j}^2 = \sum_I \left[\sigma_I^2 \left(\frac{\partial a_j}{\partial P_I} \right)^2 \right]. \tag{3.66}$$

Similarly the covariance matrix between the estimates of the coefficients in the polynomial fit is

$$\sigma_{a_j a_k}^2 = \sum_I \left[\sigma_I^2 \left(\frac{\partial a_j}{\partial P_I} \right) \left(\frac{\partial a_k}{\partial P_I} \right) \right]. \tag{3.67}$$

Practically, the σ_{a_j} and covariance matrix $\sigma_{a_j a_k}^2$ are determined as follows [51]:

- Determine the so-called ***curvature matrix*** α for our specific polynomial model, where the coefficients are given by

$$\alpha_{jk} = \sum_I \left[\frac{1}{\sigma_I^2} \, (y_I)^j \, (y_I)^k \right]. \tag{3.68}$$

- Invert the symmetric matrix α to obtain the so-called ***error matrix*** ϵ:

$$\epsilon = \alpha^{-1}. \tag{3.69}$$

- The uncertainty and covariance in the coefficients a_j is characterized by:

$$\sigma_{a_j}^2 = \epsilon_{jj}; \qquad \sigma_{a_j a_k}^2 = \epsilon_{jk}. \tag{3.70}$$

- Finally, for any function $f(a_i)$ of the coefficients a_i:

$$\sigma_f = \sqrt{\sum_{j,k} \sigma_{a_j a_k}^2 \, \frac{\partial f}{\partial a_j} \, \frac{\partial f}{\partial a_k}}. \tag{3.71}$$

Note that these rules for the propagation of uncertainties implicitly assume that the uncertainties are in some suitable sense "small" so that a local linearization of the functions $a_j(P_I)$ and $f(a_i)$ is adequate.

Now for each individual element of the curvature matrix

$$0 < \frac{\alpha_{jk}(z)}{(1+z_{\max})^{2n}} < \frac{\alpha_{jk}(z)}{(1+z_{\max})^{j+k}} < \alpha_{jk}(y) < \alpha_{jk}(z). \tag{3.72}$$

Furthermore the matrices $\alpha_{jk}(z)$ and $\alpha_{jk}(y)$ are both positive definite, and the spectral radius of $\alpha(y)$ is definitely less than the spectral radius of $\alpha(z)$. After matrix inversion this means that the minimum eigenvalue of the error matrix $\epsilon(y)$ is definitely greater than the minimum eigenvalue of $\epsilon(z)$ — more generally this tends to make the statistical uncertainties when one works with y greater than the statistical uncertainties when one works with z. (However this naive interpretation is perhaps somewhat misleading: It might be more appropriate to say that the statistical uncertainties when one works with z are anomalously low due to the fact that one has artificially stretched out the domain of the data.)

3.7.5 Estimates of the deceleration and jerk

For all five of the cosmological distance scales discussed in this chapter, we have calculated the coefficients b_j for the logarithmic distance fits, and their statistical uncertainties, for a polynomial of order $n = 2$ in both the y-redshift and z-redshift, for both the legacy05 and gold06 datasets. The constant term b_0 is (as usual in this context) a "nuisance term" that depends on an overall luminosity calibration that is not relevant to the questions at hand. These coefficents are then converted to estimates of the deceleration parameter q_0 and the combination $(j_0 + \Omega_0)$ involving the jerk. A particularly nice feature of the logarithmic distance fits is that logarithmic distances are linearly related to the reported distance modulus. So assumed Gaussian errors in the distance modulus remain Gaussian when reported in terms of logarithmic distance — which then evades one potential problem source — whatever is going on in our analysis it is ***not*** due to the nonlinear transformation of Gaussian errors. We should also mention that for both the lagacy05 and gold06 datasets the uncertainties in z have been folded into the reported values of the distance modulus: The reported values of redshift (formally) have no uncertainties associated with them, and so the nonlinear transformation $y \leftrightarrow z$ does not (formally) affect the assumed Gaussian distribution of the errors.

The results are presented in tables 3.1–3.4. Note that even after we have extracted these numerical results there is still a considerable amount of interpretation that has to go into understanding their physical implications. In particular note that the differences between the various models, (Which distance do we use? Which version of redshift do we use? Which dataset do we use?), often dwarf the statistical uncertainties within any particular model.

Table 3.1: Deceleration and jerk parameters (legacy05 dataset, y-redshift).

distance	q_0	$j_0 + \Omega_0$
d_L	-0.47 ± 0.38	-0.48 ± 3.53
d_F	-0.57 ± 0.38	$+1.04 \pm 3.71$
d_P	-0.66 ± 0.38	$+2.61 \pm 3.88$
d_Q	-0.76 ± 0.38	$+4.22 \pm 4.04$
d_A	-0.85 ± 0.38	$+5.88 \pm 4.20$

With 1-σ statistical uncertainties.

The statistical uncertainties in q_0 are independent of the distance scale used because they are linearly related to the statistical uncertainties in the parameter b_1, which themselves depend only on the curvature matrix, which is independent of the distance scale used. In contrast, the statistical uncertainties in $(j_0 + \Omega_0)$, while they depend linearly the statistical uncertainties in the parameter b_2, depend nonlinearly on q_0 and its statistical uncertainty.

Table 3.2: Deceleration and jerk parameters (legacy05 dataset, z-redshift).

distance	q_0	$j_0 + \Omega_0$
d_L	-0.48 ± 0.17	$+0.43 \pm 0.60$
d_F	-0.56 ± 0.17	$+1.16 \pm 0.65$
d_P	-0.62 ± 0.17	$+1.92 \pm 0.69$
d_Q	-0.69 ± 0.17	$+2.69 \pm 0.74$
d_A	-0.75 ± 0.17	$+3.49 \pm 0.79$

With 1-σ statistical uncertainties.

Table 3.3: Deceleration and jerk parameters (gold06 dataset, y-redshift).

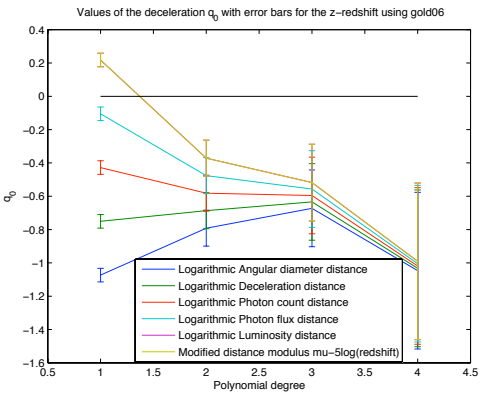
distance	q_0	$j_0 + \Omega_0$
d_L	-0.62 ± 0.29	$+1.66 \pm 2.60$
d_F	-0.78 ± 0.29	$+3.95 \pm 2.80$
d_P	-0.94 ± 0.29	$+6.35 \pm 3.00$
d_Q	-1.09 ± 0.29	$+8.87 \pm 3.20$
d_A	-1.25 ± 0.29	$+11.5 \pm 3.41$

With 1-σ statistical uncertainties.

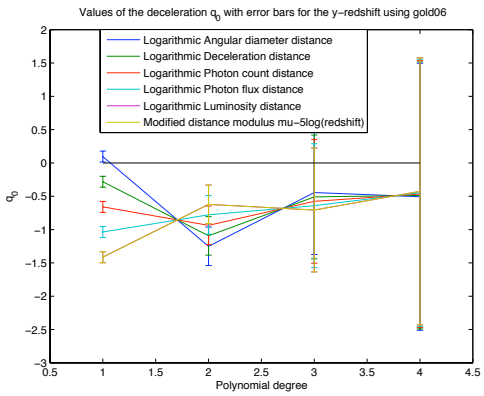
Table 3.4: Deceleration and jerk parameters (gold06 dataset, z-redshift).

distance	q_0	$j_0 + \Omega_0$
d_L	-0.37 ± 0.11	$+0.26 \pm 0.20$
d_F	-0.48 ± 0.11	$+1.10 \pm 0.24$
d_P	-0.58 ± 0.11	$+1.98 \pm 0.29$
d_Q	-0.68 ± 0.11	$+2.92 \pm 0.37$
d_A	-0.79 ± 0.11	$+3.90 \pm 0.39$

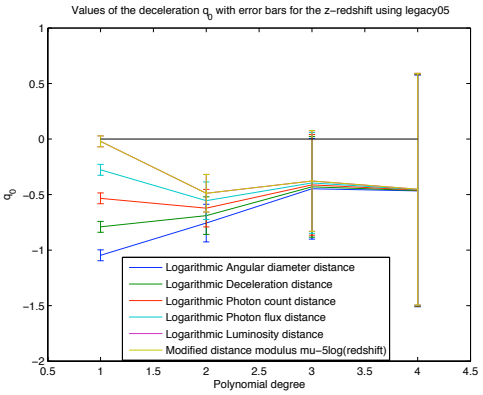
With 1-σ statistical uncertainties.



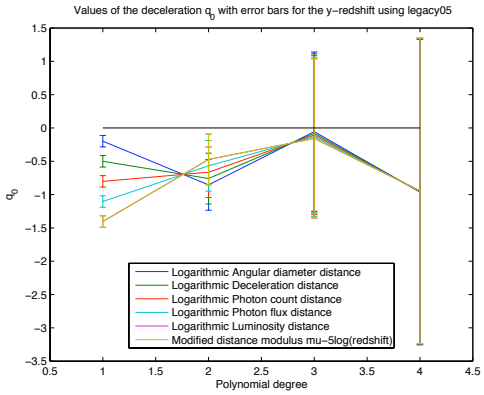
(a) Value of q_0 with the z-redshift and gold06



(b) Value of q_0 with the y-redshift and gold06

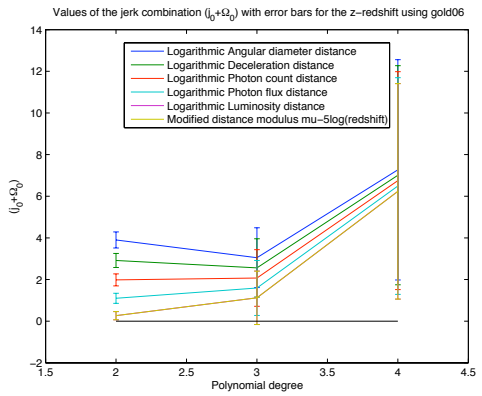


(c) Value of q_0 with the z-redshift and legacy05

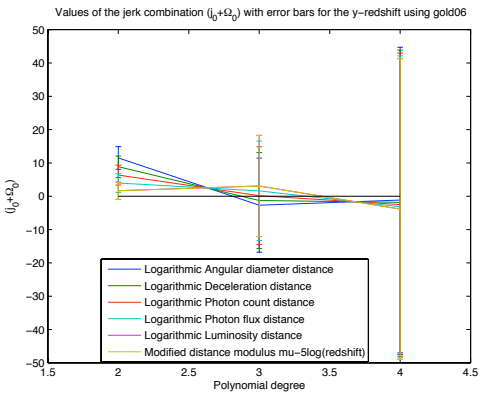


(d) Value of q_0 with the y-redshift and legacy05

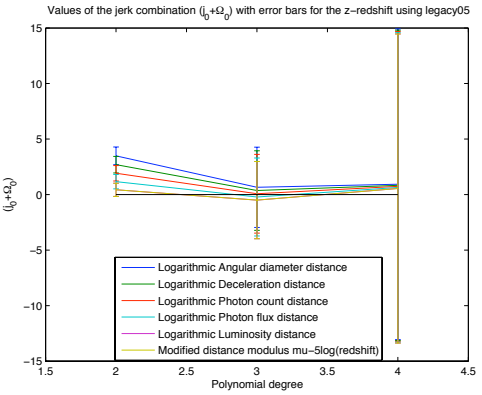
Figure 3.9: Values of the deceleration parameter q_0 for varying polynomial order fits as a function of various distance scales, with the z and y-redshift, and with the gold06 dataset [3] (a), (b), and legacy05 dataset [1] (c), (d).



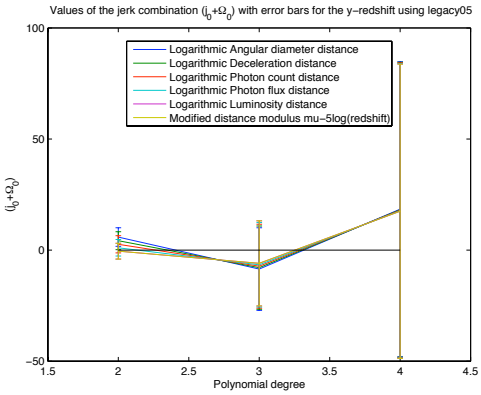
(a) Value of $(j_0 + \Omega_0)$ with the z-redshift and gold06



(b) Value of $(j_0 + \Omega_0)$ with the y-redshift and gold06



(c) Value of $(j_0 + \Omega_0)$ with the z-redshift and legacy05



(d) Value of $(j_0 + \Omega_0)$ with the y-redshift and legacy05

Figure 3.10: Values of the deceleration parameter q_0 for varying polynomial order fits as a function of various distance scales, with the z and y-redshift, and with the gold06 dataset [3] (a), (b), and legacy05 dataset [1] (c), (d).

3.8 Model-building uncertainties

The fact that there are such large differences between the cosmological parameters deduced from the different models should give one pause for concern. These differences do not arise from any statistical flaw in the analysis, nor do they in any sense represent any "***systematic***" error, rather they are an intrinsic side-effect of what it means to do a least-squares fit — to a finite-polynomial approximate Taylor series — in a situation where it is physically unclear as to which if any particular measure of ***distance*** is physically preferable, and which particular notion of ***distance*** should be fed into the least-squares algorithm. In Appendix A we present a brief discussion of the most salient mathematical issues.

The key numerical observations are that the different notions of cosmological distance lead to equally spaced least-squares estimates of the deceleration parameter, with equal statistical uncertainties; the reason for the equal-spacing of these estimates being analytically explainable by the analysis presented in A. Furthermore, from the results in Appendix A we can explicitly calculate the magnitude of this modelling ambiguity as

$$[\Delta q_0]_{\rm modelling} = -1 + \left[\sum_I z_I^{i+j}\right]^{-1}_{1j} \left[\sum_I z_I^j \, \ln(1+z_I)\right], \qquad (3.73)$$

while the corresponding formula for y-redshift is

$$[\Delta q_0]_{\rm modelling} = -1 - \left[\sum_I y_I^{i+j}\right]^{-1}_{1j} \left[\sum_I y_I^j \, \ln(1-y_I)\right]. \qquad (3.74)$$

Note that for the quadratic fits we have adopted this requires calculating a $(n+1)\times(n+1)$ matrix, with $\{i,j\} \in \{0,1,2\}$, inverting it, and then taking the inner product between the first row of this inverse matrix and the relevant column vector. The Einstein summation convention is implied on the j index. For the z-redshift (if we were to restrict our z-redshift dataset to $z < 1$, *e.g.*, using legacy05 or a truncation of gold06) it makes sense to Taylor series expand the logarithm to alternatively yield

$$[\Delta q_0]_{\rm modelling} = -\sum_{k=n+1}^{\infty} \frac{(-1)^k}{k} \left[\sum_I z_I^{i+j}\right]^{-1}_{1j} \left[\sum_I z_I^{j+k}\right]. \qquad (3.75)$$

For the y-redshift we do not need this restriction and can simply write

$$[\Delta q_0]_{\rm modelling} = \sum_{k=n+1}^{\infty} \frac{1}{k} \left[\sum_I y_I^{i+j}\right]^{-1}_{1j} \left[\sum_I y_I^{j+k}\right]. \qquad (3.76)$$

As an extra consistency check we have independently calculated these quantities (which depend only on the redshifts of the supernovae) and compared them with the spacing we find by comparing the various least-squares analyses. For the $n = 2$ quadratic fits these formulae reproduce the spacing reported in tables 3.1–3.4. As the order n of the polynomial increases, it was seen that the differences between deceleration parameter estimates based

on the different distance measures decreases — unfortunately the size of the purely statistical uncertainties was simultaneously seen to increase — this being a side effect of adding terms that are not statistically significant according to the F test.

Thus to minimize "model building ambiguities" one wishes the parameter "n" to be as large as possible, while to minimize statistical uncertainties, one does not want to add statistically meaningless terms to the polynomial.

Note that if one were to have a clearly preferred physically motivated ***best*** distance this whole model building ambiguity goes away. In the absence of a clear physically justifiable preference, the best one can do is to combine the data as per the discussion in B, which is based on NIST recommended guidelines [58], and report an additional model building uncertainty (beyond the traditional purely statistical uncertainty).

Note that we do limit the modelling uncertainty to that due to considering the five reasonably standard definitions of distance d_A, d_Q, d_P, d_F, and d_L. The reasons for this limitation are partially practical (we have to stop somewhere), and partly physics-related (these five definitions of distance have reasonably clear physical interpretations, and there seems to be no good physics reason for constructing yet more notions of cosmological distance).

Turning to the quantity $(j_0+\Omega_0)$, the different notions of distance no longer yield equally spaced estimates, nor are the statistical uncertainties equal. This is due to the fact that there is a nonlinear quadratic term involving q_0 present in the relation used to convert the polynomial coefficient b_2 into the more physical parameter $(j_0 + \Omega_0)$. Note that while for each specific model (choice of distance scale and redshift variable) the F-test indicates that keeping the quadratic term is statistically significant, the variation among the models is so great as to make measurements of $(j_0 + \Omega_0)$ almost meaningless. The combined results are reported in tables 3.5–3.6. Note that these tables do not yet include ***any*** budget for "systematic" uncertainties.

Table 3.5: Deceleration parameter summary: Statistical plus modelling.

dataset	redshift	$q_0 \pm \sigma_{\text{statistical}} \pm \sigma_{\text{modelling}}$
legacy05	y	$-0.66 \pm 0.38 \pm 0.13$
legacy05	z	$-0.62 \pm 0.17 \pm 0.10$
gold06	y	$-0.94 \pm 0.29 \pm 0.22$
gold06	z	$-0.58 \pm 0.11 \pm 0.15$

With 1-σ statistical uncertainties and 1-σ model building uncertainties, no budget for "systematic" uncertainties.

Again, we reiterate the fact that there are distressingly large differences between the cosmological parameters deduced from the different models — this should give one pause for concern above and beyond the purely formal statistical uncertainties reported herein.

3.9 Systematic uncertainties

Beyond the statistical uncertainties and model-building uncertainties we have so far considered lies the issue of systematic uncertainties. Systematic uncertainties are extremely

Table 3.6: Jerk parameter summary: Statistical plus modelling.

dataset	redshift	$(j_0 + \Omega_0) \pm \sigma_{\text{statistical}} \pm \sigma_{\text{modelling}}$
legacy05	y	$+2.65 \pm 3.88 \pm 2.25$
legacy05	z	$+1.94 \pm 0.70 \pm 1.08$
gold06	y	$+6.47 \pm 3.02 \pm 3.48$
gold06	z	$+2.03 \pm 0.31 \pm 1.29$

With 1-σ statistical uncertainties and 1-σ model building uncertanties, no budget for "systematic" uncertainties.

difficult to quantify in cosmology, at least when it comes to distance measurements — see for instance the relevant discussion in [3, 38], or in [39]. The method by which light curves are empirically reduced to make "standard candles" may itself introduce systematic uncertainties, as the recent analysis of Hicken ***et al.*** [59] demonstrates. In particular, while the legacy05 dataset is a homogeneously reduced sample, it has been reduced by the SALT method in which empirical light curve parameters are simultaneously fit with cosmological parameters assuming a ΛCDM cosmology. Hicken ***et al.*** find that the SALT reduction gives greater scatter at larger redshifts than data reduced by the MLCS method. Type Ia supernovae are only "standardizable candles", and the empirical method of standardizing them can introduce systematic errors.

What is less difficult to quantify, but still somewhat tricky, is the extent to which systematics propagate through the calculation.

3.9.1 Major philosophies underlying the analysis of statistical uncertainty

When it comes to dealing with systematic uncertainties there are two major philosophies on how to report and analyze them:

- Treat all systematic uncertainties ***as though*** they were purely statistical and report 1-sigma "effective standard uncertainties". In propagating systematic uncertainties treat them ***as though*** they were purely statistical and uncorrelated with the usual statistical uncertainties. In particular, this implies that one is to add estimated systematic and statistical uncertainties in quadrature

$$\sigma^2_{\text{combined}} = \sqrt{\sigma^2_{\text{statistical}} + \sigma^2_{\text{systematic}}}. \tag{3.77}$$

  This manner of treating the systematic uncertainties is that currently recommended by NIST [58], this recommendation itself being based on ISO, CPIM, and BIPM recommendations. This is also the language most widely used within the supernova community, and in particular in discussing the gold05 and legacy05 datasets [28, 2, 2, 3, 38], so we shall standardize our language to follow these norms.

- An alternative manner of dealing with systematics (now deprecated) is to carefully segregate systematic and statistical effects, somehow estimate "***credible bounds***" on

the systematic uncertainties, and then propagate the systematics through the calculation — if necessary using ***interval arithmetic*** to place "***credible bounds***" on the final reported systematic uncertainty. The measurements results would then be reported as a number with two independent sources of uncertainty — the statistical and systematic uncertainties, and within this philosophy there is no justification for adding statistical and systematic effects in quadrature.

It is important to realise that the systematic uncertainties reported in gold05 and legacy05 are of the first type: effective equivalent 1-sigma error bars [28, 1, 2, 3, 38]. These reported uncertainties are based on what in the supernova community are referred to as "***known unknowns***".

(The NIST guidelines [58] also recommend that all uncertainties estimated by statistical methods should be denoted by the symbol s, not σ, and that uncertainties estimated by non-statistical methods, and combined overall uncertainties, should be denoted by the symbol u — but this is rarely done in practice, and we shall follow the traditional abuse of notation and continue to use σ throughout.)

3.9.2 Deceleration

For instance, assume we can measure distance moduli to within a systematic uncertainty $\Delta\mu_{\text{systematic}}$ over a redshift range $\Delta(\text{redshift})$. If all the measurements are biased high, or all are biased low, then the systematic uncertainty would affect the Hubble parameter H_0, but would not in any way disturb the deceleration parameter q_0. However there may be a systematic drift in the bias as one scans across the range of observed redshifts. The worst that could plausibly happen is that all measurements are systematically biased high at one end of the range, and biased low at the other end of the range. For data collected over a finite width $\Delta(\text{redshift})$, this "***worst plausible***" situation leads to a systematic uncertainty in the slope of

$$\Delta\left[\frac{d\mu}{dz}\right]_{\text{systematic}} = \frac{2\,\Delta\mu_{\text{systematic}}}{\Delta(\text{redshift})}, \tag{3.78}$$

which then propagates to an uncertainty in the deceleration parameter of

$$\sigma_{\text{systematic}} = \frac{2\ln 10}{5}\,\Delta\left[\frac{d\mu}{dz}\right]_{\text{systematic}} = \frac{4\ln 10}{5}\,\frac{\Delta\mu_{\text{systematic}}}{\Delta(\text{redshift})} \approx 1.8\,\frac{\Delta\mu_{\text{systematic}}}{\Delta(\text{redshift})}. \tag{3.79}$$

For the situation we are interested in, if we take at face value the reliability of the assertion "...we adopt a limit on redshift-dependent systematics to be 5% per $\Delta z = 1$" [3], meaning up to 2.5% high at one end of the range and up to 2.5% low at the other end of the range. A 2.5% variation in distance then corresponds, via $\Delta\mu_D = 5\Delta(\ln d_L)/\ln 10$, to an uncertainty $\Delta\mu_{\text{systematic}} = 0.05$ in stellar magnitude. So, (taking $\Delta z = 1$), one has to face the somewhat sobering estimate that the "***equivalent 1-σ uncertainty***" for the deceleration parameter q_0 is

$$\sigma_{\text{systematic}} = 0.09. \tag{3.80}$$

When working with y-redshift, one really should reanalyze the entire corpus of data from first principles — failing that, (not enough of the raw data is publicly available), we shall simply observe that

$$\frac{\mathrm{d}z}{\mathrm{d}y} \to 1 \qquad \text{as} \qquad y \to 0, \tag{3.81}$$

and use this as a justification for assuming that the systematic uncertainty in q_0 when using y-redshift is the same as when using z-redshift.

3.9.3 Jerk

Turning to systematic uncertainties in the jerk, the worst that could plausibly happen is that all measurements are systematically biased high at both ends of the range, and biased low at the middle, (or low at both ends and high in the middle), leading to a systematic uncertainty in the second derivative of

$$\frac{1}{2}\,\Delta\left[\frac{d^2\mu}{dz^2}\right]_{\text{systematic}}\left[\frac{\Delta(\text{redshift})}{2}\right]^2 = 2\Delta\mu_{\text{systematic}}, \tag{3.82}$$

where we have taken the second-order term in the Taylor expansion around the midpoint of the redshift range, and asked that it saturate the estimated systematic error $2\Delta\mu_{\text{systematic}}$. This implies

$$\Delta\left[\frac{d^2\mu}{dz^2}\right]_{\text{systematic}} = \frac{16\;\Delta\mu_{\text{systematic}}}{\Delta(\text{redshift})^2}, \tag{3.83}$$

which then propagates to an uncertainty in the jerk parameter $(j_0 + \Omega_0)$ of ***at least***

$$\sigma_{\text{systematic}} \geq \frac{3\ln 10}{5}\,\Delta\left[\frac{d^2\mu}{dz^2}\right]_{\text{systematic}} = \frac{48\ln 10}{5}\,\frac{\Delta\mu_{\text{systematic}}}{\Delta(\text{redshift})^2} \approx 22\,\frac{\Delta\mu_{\text{systematic}}}{\Delta(\text{redshift})^2}. \tag{3.84}$$

There are additional contributions to the systematic uncertainty arising from terms linear and quadratic in q_0. They do not seem to be important in the situations we are interested in so we content ourselves with the single term estimated above. Using $\Delta\mu_{\text{systematic}} = 0.05$ and $\Delta z = 1$ we see that the "equivalent 1-σ uncertainty" for the combination $(j_0 + \Omega_0)$ is:

$$\sigma_{\text{systematic}} = 1.11. \tag{3.85}$$

Thus direct cosmographic measurements of the jerk parameter are plagued by ***very*** high systematic uncertainties. Note that the systematic uncertainties calculated in this section are completely equivalent to those reported in [3].

3.10 Historical estimates of systematic uncertainty

We now turn to the question of possible additional contributions to the uncertainty, based on what the NIST recommendations call "***type B evaluations of uncertainty***" — namely "any method of evaluation of uncertainty by means other than the statistical analysis of a series of observations" [58]. (This includes effects that in the supernova community are referred to as "***unknown unknowns***", which are ***not*** reported in any of their estimates of systematic uncertainty.)

The key point here is this: "A type B evaluation of standard uncertainty is usually based on scientific judgment using all of the relevant information available, which may include: previous measurement data, ***etc...***" [58]. It is this recommendation that underlies what we might wish to call the "***historical***" estimates of systematic uncertainty — roughly speaking, we suggest that in the systematic uncertainty budget it is prudent to keep an extra

"historical uncertainty" at least as large as the most recent major re-calibration of whatever measurement method you are currently using.

Now this "***historical uncertainty***" contribution to the systematic uncertainty budget that we are advocating is based on 100 years of unanticipated systematic errors ("***unknown unknowns***") in astrophysical distance scales — from Hubble's reliance on mis-calibrated Cepheid variables (leading to distance estimates that were about 666% too large), to last decade's debates on the size of our own galaxy (with up to 15% disagreements being common), to last year's 5% shift in the high-z supernova distances [3, 38] — and various other re-calibration events in between. That is, 5% variations in estimates of cosmological distances on a 2 year time scale seem common, 10% on a 10 year time scale, and 500% or more on an 80 year timescale? A disinterested outside observer does detect a certain pattern here. (These re-calibrations are of course not all related to supernova measurements, but they are historical evidence of how difficult it is to make reliable distance measurements in cosmology.)

3.10.1 Deceleration

Based on the historical evidence we feel that it is currently prudent to budget an additional "***historical uncertainty***" of approximately 5% in the distances to the furthest supernovae, (corresponding to 0.10 stellar magnitudes), while for the nearby supernovae we generously budget a "***historical uncertainty***" of 0%, based on the fact that these distances have not changed in the last 2 years [3, 38].[2]

This implies

$$\Delta \left[\frac{d\mu}{dz}\right]_{\text{historical}} = \frac{\Delta\mu_{\text{historical}}}{\Delta(\text{redshift})}. \qquad (3.86)$$

Note the ***absence*** of a factor 2 compared to equation (3.78), this is because in this "historical" discussion we have taken the nearby supernovae to be accurately calibrated, whereas in the discussion of systematic uncertainties in equation (3.78) both nearby and distant supernovae are subject to "***known unknown***" systematics. This then propagates to an uncertainty in the deceleration parameter of

$$\sigma_{\text{historical}} = \frac{2\ln 10}{5}\,\Delta\left[\frac{d\mu}{dz}\right]_{\text{historical}} = \frac{2\ln 10}{5}\,\frac{\Delta\mu_{\text{historical}}}{\Delta(\text{redshift})} \approx 0.9\;\frac{\Delta\mu_{\text{historical}}}{\Delta(\text{redshift})}. \qquad (3.87)$$

Noting that a 5% shift in luminosity distance is equivalent to an uncertainty of $\Delta\mu_{\text{historical}} = 0.10$ in stellar magnitude, this implies an "equivalent 1-σ uncertainty" for the deceleration parameter q_0 is

$$\sigma_{\text{historical}} = 0.09. \qquad (3.88)$$

This (coincidentally) is equal to the systematic uncertainties based on "known unknowns".

[2] Some researchers have argued that the present "***historical***" estimates of uncertainty confuse the notion of "error" with that of "***uncertainty***". We disagree. What we are doing here is to use the most recently detected (significant) error to estimate one component of the uncertainty — this is simply a "scientific judgment using all of the relevant information available". We should add that other researchers have argued that our historical uncertainties should be even larger. By using the most recent major re-calibration as our basis for historical uncertainty we feel we are steering a middle course between placing too much ***versus*** to little credence in the observational data.

3.10.2 Jerk

Turning to the second derivative a similar analysis implies

$$\frac{1}{2}\,\Delta\left[\frac{d^2\mu}{dz^2}\right]_{\text{historical}}\Delta(\text{redshift})^2 = \Delta\mu_{\text{historical}}. \tag{3.89}$$

Note the absence of various factors of 2 as compared to equation 3.82. This is because we are now assuming that for "***historical***" purposes the nearby supernovae are accurately calibrated and it is only the distant supernovae that are potentially uncertain — thus in estimating the historical uncertainty the second-order term in the Taylor series is now to be saturated using the entire redshift range. Thus

$$\Delta\left[\frac{d^2\mu}{dz^2}\right]_{\text{historical}} = \frac{2\,\Delta\mu_{\text{historical}}}{\Delta(\text{redshift})^2}, \tag{3.90}$$

which then propagates to an uncertainty in the jerk parameter of ***at least***

$$\sigma_{\text{historical}} \geq \frac{3\ln 10}{5}\,\Delta\left[\frac{d^2\mu}{dz^2}\right]_{\text{historical}} = \frac{6\ln 10}{5}\,\frac{\Delta\mu_{\text{historical}}}{\Delta(\text{redshift})^2} \approx 2.75\,\frac{\Delta\mu_{\text{historical}}}{\Delta(\text{redshift})^2}. \tag{3.91}$$

Again taking $\Delta\mu_{\text{historical}} = 0.10$ this implies an "***equivalent 1-σ uncertainty***" for the combination $j_0 + \Omega_0$ is

$$\sigma_{\text{historical}} = 0.28. \tag{3.92}$$

Note that this is (coincidentally) one quarter the size of the systematic uncertainties based on "***known unknowns***", and is still quite sizable.

Table 3.7: Deceleration parameter summary: Statistical, modelling, systematic, and historical.

dataset	redshift	$q_0 \pm \sigma_{\text{statistical}} \pm \sigma_{\text{modelling}} \pm \sigma_{\text{systematic}} \pm \sigma_{\text{historical}}$
legacy05	y	$-0.66 \pm 0.38 \pm 0.13 \pm 0.09 \pm 0.09$
legacy05	z	$-0.62 \pm 0.17 \pm 0.10 \pm 0.09 \pm 0.09$
gold06	y	$-0.94 \pm 0.29 \pm 0.22 \pm 0.09 \pm 0.09$
gold06	z	$-0.58 \pm 0.11 \pm 0.15 \pm 0.09 \pm 0.09$

With 1-σ effective statistical uncertainties for all components.

The systematic and historical uncertainties are now reported in tables 3.7–3.8. The estimate for systematic uncertainties are equivalent to those presented in [3], which is largely in accord with related sources [28, 1, 2]. Our estimate for "***historical***" uncertainties is likely to be more controversial — with, we suspect, many cosmologists arguing that our estimates are too generous — and that $\sigma_{\text{historical}}$ should perhaps be ***even larger*** than we have estimated. What is not (or should not) be controversial is the need for ***some*** estimate of $\sigma_{\text{historical}}$. Previous history should not be ignored, and as the NIST guidelines emphasize, previous history is an essential and integral part of making the scientific judgment as to what the overall uncertainties are.

Table 3.8: Jerk parameter summary: Statistical, modelling, systematic, and historical.

dataset	redshift	$(j_0 + \Omega_0) \pm \sigma_{\text{statistical}} \pm \sigma_{\text{modelling}} \pm \sigma_{\text{systematic}} \pm \sigma_{\text{historical}}$
legacy05	y	$+2.65 \pm 3.88 \pm 2.25 \pm 1.11 \pm 0.28$
legacy05	z	$+1.94 \pm 0.70 \pm 1.08 \pm 1.11 \pm 0.28$
gold06	y	$+6.47 \pm 3.02 \pm 3.48 \pm 1.11 \pm 0.28$
gold06	z	$+2.03 \pm 0.31 \pm 1.29 \pm 1.11 \pm 0.28$

With 1-σ effective statistical uncertainties for all components.

3.11 Combined uncertainties

We now combine these various uncertainties, purely statistical, modelling, "***known unknown***" systematics, and "***historical***" ("***unknown unknowns***"). Adopting the NIST philosophy of dealing with systematics, these uncertainties are to be added in quadrature [58]. Including all 4 sources of uncertainty we have discussed:

$$\sigma_{\text{combined}} = \sqrt{\sigma^2_{\text{statistical}} + \sigma^2_{\text{modelling}} + \sigma^2_{\text{systematic}} + \sigma^2_{\text{historical}}}. \tag{3.93}$$

That the statistical and modelling uncertainties should be added in quadrature is clear from their definition. Whether or not systematic and historical uncertainties should be treated this way is very far from clear, and implicitly presupposes that there are no correlations between the systematics and the statistical uncertainties — within the "credible bounds" philosophy for estimating systematic uncertainties there is no justification for such a step. Within the "***all errors are effectively statistical***" philosophy adding in quadrature is standard and in fact recommended — this is what is done in current supernova analyses, and we shall continue to do so here. The combined uncertainties σ_{combined} are reported in tables 3.9–3.10.

3.12 Expanded uncertainty

An important concept under the NIST guidelines is that of "***expanded uncertainty***"

$$U_k = k\, \sigma_{\text{combined}}. \tag{3.94}$$

Expanded uncertainty is used when for either scientific or legal/regulatory reasons one wishes to be ***certain*** that the actual physical value of the quantity being measured lies within the stated range. We shall take $k = 3$, this being equivalent to the well-known particle physics aphorism "if it's not three-sigma, it's not physics". Note that this is not an invitation to randomly multiply uncertainties by 3, rather it is a scientific judgment that if one wishes to be 99.5% certain that something is or is not happening one should look for a 3-sigma effect. Bitter experience within the particle physics community has led to the consensus that 3-sigma is the absolute minimum standard one should look for when claiming "***new physics***". There is now a growing consensus in the particle physics community that ***5-sigma*** should be the new standard for claiming "***new physics***" [60]. In other words, 2-sigma

is generally used in "***sociology***", 3-sigma is "***evidence for***" whereas 5-sigma is "***discovery of new physics***". We take

$$U_3 = 3\,\sigma_{\rm combined}, \tag{3.95}$$

and also present the results one would obtain with U_5. The best estimates, combined uncertainties $\sigma_{\rm combined}$, and expanded uncertainties U, are reported in tables 3.9–3.10.

Table 3.9: Deceleration parameter summary: Combined and expanded uncertainties.

dataset	redshift	$q_0 \pm \sigma_{\rm combined}$	$q_0 \pm U_3$	$q_0 \pm U_5$
legacy05	y	-0.66 ± 0.42	-0.66 ± 1.26	-0.66 ± 2.1
legacy05	z	-0.62 ± 0.23	-0.62 ± 0.70	-0.62 ± 1.26
gold06	y	-0.94 ± 0.39	-0.94 ± 1.16	-0.94 ± 1.95
gold06	z	-0.58 ± 0.23	-0.58 ± 0.68	-0.58 ± 1.15

Table 3.10: Jerk parameter summary: Combined and expanded uncertainties.

dataset	redshift	$(j_0 + \Omega_0) \pm \sigma_{\rm combined}$	$(j_0 + \Omega_0) \pm U_3$	$(j_0 + \Omega_0) \pm U_5$
legacy05	y	$+2.65 \pm 4.63$	$+2.65 \pm 13.9$	$+2.65 \pm 23.1$
legacy05	z	$+1.94 \pm 1.72$	$+1.94 \pm 5.17$	$+1.94 \pm 8.6$
gold06	y	$+6.47 \pm 4.75$	$+6.47 \pm 14.2$	$+6.47 \pm 23.7$
gold06	z	$+2.03 \pm 1.75$	$+2.03 \pm 5.26$	$+2.03 \pm 8.75$

3.13 Results

What can we conclude from this? While the "***preponderance of evidence***" is certainly that the universe is currently accelerating, $q_0 < 0$, this is not yet a "gold plated" result. We emphasise the fact that (as is or should be well known) there is an enormous difference between the two statements:

- "***the most likely value for the deceleration parameter is negative***", and
- "***there is significant evidence that the deceleration parameter is negative***".

When it comes to assessing whether or not the evidence for an accelerating universe is physically significant, the first rule of thumb for combined uncertainties is the well known aphorism "***if it's not three-sigma, it's not physics***". The second rule is to be conservative in your systematic uncertainty budget. We cannot in good faith conclude that the expansion of the universe is accelerating. It is more likely that the expansion of the universe is accelerating, than that the expansion of the universe is decelerating — but this is a very long way from having definite evidence in favour of acceleration. The summary regarding the jerk parameter, or more precisely $(j_0 + \Omega_0)$, is rather grim reading, and indicates the need for considerable caution in interpreting the supernova data. Note that while use of

the y-redshift may improve the theoretical convergence properties of the Taylor series, and will not affect the uncertainties in the distance modulus or the various distance measures, it does seem to have an unfortunate side-effect of magnifying statistical uncertainties for the cosmological parameters.

As previously mentioned, we have further checked the robustness of our analysis by first excluding the outlier at $z = 1.755$, then excluding the so-called "***Hubble bubble***" at $z < 0.0233$ [49, 50], and then excluding both — the precise numerical estimates for the cosmological parameters certainly change, but the qualitative picture remains as we have painted it here.

3.14 Conclusions on Cosmography

Why do our conclusions seem to be so much at variance with currently perceived wisdom concerning the acceleration of the universe? The main reasons are twofold:

- Instead of simply picking a single model and fitting the data to it, we have tested the overall robustness of the scenario by encoding the same physics (H_0, q_0, j_0) in multiple different ways (d_L, d_F, d_P, d_Q, d_A; using both z and y) to test the robustness of the data fitting procedures.
- We have been much more explicit, and conservative, about the role of systematic uncertainties, and their effects on estimates of the cosmological parameters.

If we ***only*** use the statistical uncertainties and the "***known unknowns***" added in quadrature, then the case for cosmological acceleration is much improved, and is (in some cases we study) "statistically significant at three-sigma", but this does not mean that such a conclusion is either robust or reliable. (By "cherry picking" the data, and the particular way one analyzes the data, one can find statistical support for almost any conclusion one wants.)

The modelling uncertainties we have encountered depend on the distance variable one chooses to do the least squares fit (d_L, d_F, d_P, d_Q, d_A). There is no good physics reason for preferring any one of these distance variables over the others. One can always minimize the modelling uncertainties by going to a higher-order polynomial — unfortunately at the price of unacceptably increasing the statistical uncertainties — and we have checked that this makes the overall situation worse. This does however suggest that things might improve if the data had smaller scatter and smaller statistical uncertainties: We could then hope that the F-test would allow us to go to a cubic polynomial, in which case the dependence on which notion of distance we use for least-squares fitting should decrease.

> ***We wish to emphasize the point that, regardless of one's views on how to combine formal estimates of uncertainty, the very fact that different distance scales yield data-fits with such widely discrepant values strongly suggests the need for extreme caution in interpreting the supernova data.***

Though we have chosen to work on a cosmographic framework, and so minimize the number of physics assumptions that go into the model, we expect that similar modelling uncertainties will also plague other more traditional approaches. (For instance, in the present-day consensus scenario there is considerable debate as to just when the universe switches

from deceleration to acceleration, with different models making different statistical predictions [61].) One lesson to take from the current analysis is that purely statistical estimates of error, while they can be used to make statistical deductions within the context of a specific model, are often a bad guide as to the extent to which two different models for the same physics will yield differing estimates for the same physical quantity.

There are a number of other more sophisticated statistical methods that might be applied to the data to possibly improve the statistical situation. For instance, ridge regression, robust regression, and the use of orthogonal polynomials and loess curves. However one should always keep in mind the difference between ***accuracy*** and ***precision*** [51]. More sophisticated statistical analyses may permit one to improve the precision of the analysis, but unless one can further constrain the systematic uncertainties such precise results will be no more accurate than the current situation. Excessive refinement in the statistical analysis, in the absence of improved bounds on the systematic uncertainties, is counterproductive and grossly misleading.

However, we are certainly not claiming that all is grim on the cosmological front — and do not wish our views to be misinterpreted in this regard — there are clearly parts of cosmology where there is plenty of high-quality data, and more coming in, constraining and helping refine our models. But regarding some specific cosmological questions the catch cry should still be "***Precision cosmology? Not just yet***" [62].

In particular, in order for the current technique to become a tool for precision cosmology, we would need more data, smaller scatter in the data, and smaller uncertainties. For instance, by performing the F-test we found that it was almost always statistically meaningless to go beyond quadratic fits to the data. If one can obtain an improved dataset of sufficient quality for cubic fits to be meaningful, then ambiguities in the deceleration parameter are greatly suppressed.

In closing, we strongly encourage readers to carefully contemplate figures 3.4(a)–3.6(b) as an inoculation against over-interpretation of the supernova data. In those figures we have split off the linear part of the Hubble law (which is encoded in the intercept) and chosen distance variables so that the slope (at redshift zero) of whatever curve one fits to those plots is directly proportional to the acceleration of the universe (in fact the slope is equal to $-q_0/2$). Remember that these plots only exhibit the statistical uncertainties. Remembering that we prefer to work with natural logarithms, not stellar magnitudes, one should add systematic uncertainties of $\pm[\ln(10)/5] \times (0.05) \approx 0.023$ to these statistical error bars, presumably in quadrature. Furthermore a good case can be made for adding an additional "***historical***" uncertainty, using the past history of the field to estimate the "***unknown unknowns***".

> ***Ultimately however, it is the fact that figures 3.4(a)–3.6(b) do not exhibit any overwhelmingly obvious trend that makes it so difficult to make a robust and reliable estimate of the sign of the deceleration parameter.***

"Every generation of humans believed it had all the answers it needed, except for a few mysteries they assumed would be solved at any moment. And they all believed their ancestors were simplistic and deluded. What are the odds that you are the first generation of humans who will understand reality?"

Scott Adams (born 1957)

4 Cosmodynamics in a FLRW universe

Some 10 years ago Matt Visser initiated a programme of using the classical energy conditions of general relativity to place very general and robust bounds on various cosmological parameters [63, 64, 65]. In that early work, attention was mainly focussed on the energy density $\rho(z)$ and lookback time $T(z)$. Since then, the classical energy conditions have (on the one hand) seen continued use in studying issues such as the minimal requirements for cosmological bounces [66, 67] and other ***cosmological milestones*** [68, 69, 70], and (on the other hand) have seen further applications to bounding cosmological distances $d(z)$ [71, 72], and lookback time $T(z)$ [73]. In the first part of this chapter we shall try to draw these various threads together and establish several simple and rugged energy-condition-induced bounds on cosmological parameters. Several of these bounds are completely new, several are significant extensions of known results, and all are now generalized to arbitrary spatial curvature. For some generic cosmological parameter, say represented by $X(z)$, we shall seek bounds of the form

$$X(z) \gtrless X_{\rm bound} \equiv X_0\, f(\Omega_0, z), \tag{4.1}$$

where X_0 is the value of $X(z)$ at the present epoch, the direction of the inequality may depend both on the bound being considered and the redshift region of interest, and $f(\Omega_0, z)$ is some dimensionless function to be determined. Typically $f(\Omega_0, z)$ will be a polynomial, rational, algebraic, or elementary function, though for the particular case of the dominant energy condition applied to the lookback time we shall encounter a specific hypergeometric function.

There is (at least) one important caveat: It should be kept clearly in mind that the classical energy conditions are ***not*** fundamental physics — in fact the classical energy conditions are known to be violated by quantum effects [74, 75, 76, 76, 77, 78, 79, 80], at least to some extent, and so the energy conditions should always be viewed provisionally — as a way of characterizing whether or not a certain situation is describable by "normal" physics [79, 80].

In all of these analyses with the classical energy conditions there is a trade-off between the ***precision*** and ***generality*** of the constraints one obtains — the art lies in choosing a form of the input assumptions that is as general as possible, but not too general, for the precision of the output constraints one wishes to derive.

In the second part of this chapter we shall derive some very general bounds in terms of assumptions about the w-parameter, where as usual $w = p/\rho$. Specifically, we shall ask the question: If we know for theoretical reasons, or can observationally determine, that w lies

in some restricted range

$$w(z) \in [w_-, w_+], \tag{4.2}$$

between redshift zero and redshift z, what constraint does that place on the cosmological expansion? We shall see that considerable useful information can be extracted regarding the density $\rho(z)$, Hubble parameter $H(z)$, density parameter $\Omega(z)$, various cosmological distances $d_X(z)$, and lookback time $T(z)$. Specifically, for some generic cosmological parameter $X(z)$, we shall be looking for bounds of the form

$$X_{w_\pm}(z) \leq X(z) \leq X_{w_\mp}(z), \tag{4.3}$$

Conversely, observational constraints on these cosmological parameters can be used to infer features of the cosmological fluid in a largely model-independent manner. In contrast to other partial results scattered throughout the literature, we carry out the computations for arbitrary values of the space curvature $k \in [-1, 0, +1]$, equivalently for arbitrary $\Omega_0 \lesseqgtr 1$.

4.1 Basic formulae

In standard cosmology, one assumes the cosmological principle, that is, our universe is isotropic and homogeneous on large scales. This assumption leads one to consider cosmological spacetimes of the idealized FLRW form [81, 20, 82, 21, 83, 84]:

$$\mathrm{d}s^2 \;=\; -\mathrm{d}t^2 + a(t)^2 \left\{ \frac{\mathrm{d}r^2}{1-kr^2} + r^2 \left[\mathrm{d}\theta^2 + \sin^2\theta \, \mathrm{d}\phi^2\right] \right\}. \tag{4.4}$$

If we further assume that gravitational interactions at large scales are described by general relativity, we can use the Friedmann equations that relate the total density ρ and the total pressure p to a function of the scale factor a and its time derivatives. Indeed, in units where $8\pi G_N = 1$, but explicitly retaining the speed of light c, we have [1]

$$\rho(t) \;=\; 3\left(\frac{\dot{a}^2}{a^2} + \frac{k\,c^2}{a^2}\right), \tag{4.5}$$

$$p(t) \;=\; -2\,\frac{\ddot{a}}{a} - \frac{\dot{a}^2}{a^2} - \frac{k\,c^2}{a^2}, \tag{4.6}$$

$$\rho(t) + 3\,p(t) \;=\; -6\,\frac{\ddot{a}}{a}. \tag{4.7}$$

The classical energy conditions of general relativity, to the extent that one believes that they are a useful guide [79, 80], allow one to deduce physical constraints on the behaviour of matter fields in strong gravitational fields or cosmological geometries. For a perfect fluid cosmology, and in terms of pressure and density, the so-called ***Null***, ***Weak***, ***Strong*** and ***Dominant*** energy conditions reduce to [88]:

[1] Note that we are now specifically assuming Friedmann dynamics for the universe, one is thus explicitly stepping outside the "cosmographic" or "cosmokinetic" framework of [22, 23, 24, 25, 12, 13, 85, 86, 87].

NEC: $\rho + p \geq 0$.
In view of the Friedmann equations this then reduces to

$$-\frac{\ddot{a}}{a} + \frac{\dot{a}^2}{a^2} + \frac{k\,c^2}{a^2} \geq 0; \qquad \text{that is} \qquad \frac{\ddot{a}}{a} \leq \frac{\dot{a}^2}{a^2} + \frac{k\,c^2}{a^2}. \tag{4.8}$$

WEC: This specializes to the NEC plus $\rho \geq 0$.
This then reduces to the NEC plus the condition

$$\dot{a}^2 + k\,c^2 \geq 0. \tag{4.9}$$

This condition is ***vacuous*** for $k \in \{0, +1\}$ and only for $k = -1$ does it convey even a little information.

SEC: This specializes to the NEC plus $\rho + 3p \geq 0$.
This then reduces to the NEC plus the deceleration condition

$$\frac{\ddot{a}}{a} \leq 0. \tag{4.10}$$

DEC: $\rho \pm p \geq 0$.
This reduces to the NEC plus the condition

$$\frac{\ddot{a}}{a} \geq -2\left(\frac{\dot{a}^2}{a^2} + \frac{k\,c^2}{a^2}\right). \tag{4.11}$$

Note particularly that the condition (4.10) is independent of the space curvature k. Now, DEC implies WEC implies NEC, and SEC implies NEC, but otherwise the NEC, WEC, SEC, and DEC are mathematically independent assumptions. In particular, the SEC does ***not*** imply the WEC. Violating the NEC implies violating the DEC, SEC, and WEC as well [88].

Using this dynamical formulation of the energy conditions, and adopting the outlook of [63, 64, 65], Santos ***et al.*** [71] have recently derived some bounds, for the special case $k = 0$, on the luminosity distance d_L of supernovae, and have then compared these bounds with the legacy [28, 1] and gold [3] datasets. In reference [72] bounds on the distance modulus are presented for general values of $k \in \{-1, 0, +1\}$, while in reference [73] they concentrate on the lookback time. Herein, we shall use a similar but distinct approach to obtain rugged and more general bounds on the Hubble parameter, the Omega parameter, the density, the lookback time, and on the various distance scales defined previously in [85, 86] ***for all values of k space curvature*** $\in \{-1, 0, +1\}$.

4.2 Energy conditions and the Hubble parameter $\mathrm{H}(\mathrm{z})$

The energy conditions translate, in a FLRW setting, into the inequalities (4.8), (4.9), (4.10), and (4.11), from which we deduce bounds on the Hubble function $H(z)$ in terms of the Hubble parameter H_0, the Omega parameter Ω_0, and the z-redshift.

4.2.1 NEC:

Using inequality (4.8) we obtain:

$$\frac{\dot{a}}{a}\,\frac{\mathrm{d}}{\mathrm{d}a}\left(\frac{\dot{a}}{a}\right) \;\leq\; \frac{k\,c^2}{a^3}, \tag{4.12}$$

which can be integrated to yield

$$\int_a^{a_0} \frac{\mathrm{d}}{\mathrm{d}a}\left(\frac{1}{2}\,\left(\frac{\dot{a}}{a}\right)^2\right)\,\mathrm{d}a \;\leq\; \int_a^{a_0} \frac{k\,c^2}{a^3}\,\mathrm{d}a. \tag{4.13}$$

That is

$$H_0^2 - H(z)^2 \leq -k\,c^2\,\{a_0^{-2} - a^{-2}\} \tag{4.14}$$

Now using

$$\frac{a_0}{a} \;=\; 1+z, \tag{4.15}$$

and the relation

$$\Omega_0 \;=\; 1 + \frac{kc^2}{a_0^2\,H_0^2}, \tag{4.16}$$

after a few rearrangements we obtain a bound in terms of a simple algebraic function: [2]

$$H(z) \;\geq\; H_{\mathrm{NEC}} \;\equiv\; H_0\sqrt{\Omega_0 + [1-\Omega_0]\,(1+z)^2}. \tag{4.17}$$

In order to obtain this inequality we have assumed that $z > 0$, so that one is looking into the past. When looking into the future, $z < 0$, the inequality is reversed.[3] Physically, we see that for $\Omega_0 \leq 1$ equation (4.17) implies $H(z) \geq H_0$, so the Hubble parameter is nondecreasing (nonincreasing) as we look into the past (future) — this implies that the expansion is always less than exponential.

Technical point: To be useful, the bound on the Hubble parameter must be a real number. For $\Omega_0 \leqslant 1$ this is automatic. In contrast, note that when $\Omega_0 > 1$, there exists a z value for which the expression in the square root becomes negative or zero. This specific value of the z-redshift is given by:

$$z_{\mathrm{NEC}} \;=\; \sqrt{\frac{\Omega_0}{\Omega_0 - 1}} - 1. \tag{4.18}$$

Note that at $z = z_{\mathrm{NEC}}$, we get $H_{\mathrm{NEC}}(z_{\mathrm{NEC}}) = 0$. Also note that z_{NEC} is positive as long as Ω_0 is positive. Nothing unusual need happen to the universe itself at z_{NEC}, it is only the ***bound***

[2] In fact, all the Hubble bounds derived below will be algebraic.

[3] Note that looking back into the past $z > 0$, with $z = \infty$ corresponding to the big bang. In contrast, looking forward into the future $z < 0$, with $z = -1$ corresponding to infinite expansion [85, 86].

that loses its predictive usefulness. In practice, given that current observational estimates are

$$\Omega_0 = 1.02 \pm 0.02 \qquad \text{(PDG 2004 [89])}, \tag{4.19}$$

$$\Omega_0 = 1.003^{+0.013}_{-0.017} \qquad \text{(PDG 2006 [17])}, \tag{4.20}$$

we see that (for instance) $z_{\rm NEC}(\Omega_0 = 1.04) = 4.1$ and $z_{\rm NEC}(\Omega_0 = 1.01) = 9.0$. So given current observational estimates of Ω_0, the fact that the highest-z supernovae seen to date have $z \lesssim 2$, and the fact that we are expected to run out of galaxies by the time we reach $z \lesssim 7$, the limitations associated with $z_{\rm NEC}$ are unlikely to be significant in any realistic setting.

Overall, the bound on the Hubble function (4.17) is valid for $\Omega_0 \leq 1, \ \forall\ z \in \ [0, +\infty)$, and for $\Omega_0 > 1$ under the condition that $z \ \in [0, z_{\rm NEC}]$.

4.2.2 WEC:

From inequality (4.9), we can deduce that

$$\text{for } k = -1, \qquad \dot{a} \ \leq \ \sqrt{-k}\, c. \tag{4.21}$$

To obtain a constraint on the Hubble function, we divide equation (4.21) by a, and obtain the rather weak bound:

$$H(z) \ \geq \ H_{\rm WEC} \equiv H_0\, (1+z)\, \sqrt{1-\Omega_0} \qquad \forall\, \Omega_0 \in (0,1). \tag{4.22}$$

We have assumed that $z > 0$ in inequality (4.22) so that one is looking into the past. When looking into the future $z < 0$, the inequality is reversed. Note that this bound is only valid $\forall\, \Omega_0 \in (0,1)$ and $\forall\, z > 0$.

Important remark: Note that as long as $\Omega_0 > 0$ we have:

$$H_{\rm NEC} \geq H_{\rm WEC}. \tag{4.23}$$

That is, the WEC really does not give us anything extra beyond the statement that Ω_0 is positive.

4.2.3 SEC:

From inequality (4.10), we deduce that

$$\forall\, a \ < \ a_0 \qquad \frac{1}{\dot{a}} \ \leq \ \frac{1}{H_0\, a_0}. \tag{4.24}$$

Further, to obtain a relation on the Hubble function and Hubble parameter, we multiply equation (4.24) by a, and we obtain the bound: [4]

$$H(z) \ \geq \ H_{\rm SEC} \equiv H_0\, (1+z). \tag{4.25}$$

[4] We have assumed that $z > 0$ in inequality (4.25) so that one is looking into the past. When looking into the future $z < 0$, the inequality is reversed. Note that this specific bound can also be found in [71, 72].

Physically we see that this bound can be rewritten as $\dot{a} \geq \dot{a}_0$, implying that H must decrease at least as rapidly as free expansion — it is this particular energy condition that is violated by the now-usual interpretation of the observational supernovae data in terms of cosmic acceleration. (Equivalently, the "dark energy" is specifically designed to violate this energy condition.)

4.2.4 DEC:

To satisfy this energy condition, the NEC must hold as well as inequality (4.11). We use the same approach as for the NEC, rewriting (4.11) as:

$$\frac{\mathrm{d}(a^2\dot{a})}{\mathrm{d}t} + 2kc^2a \quad \geq \quad 0, \tag{4.26}$$

that is,

$$\frac{\mathrm{d}a}{\mathrm{d}t}\frac{\mathrm{d}}{\mathrm{d}a}(a^2\dot{a}) + 2kc^2a \quad \geq \quad 0. \tag{4.27}$$

Multiplying by a^2, this inequality leads to

$$\frac{\mathrm{d}}{\mathrm{d}a}\left[\frac{1}{2}\left(a^2\,\dot{a}\right)^2 + \frac{k\,c^2}{2}\,a^4\right] \quad \geq \quad 0 \qquad \forall\, a. \tag{4.28}$$

Integrating, we can deduce the new inequality,

$$\forall\, a < a_0 \qquad \left(a^2\,\dot{a}\right)^2 + k\,c^2\,a^4 \quad \leq \left(a_0^2\dot{a}_0\right)^2 + k\,c^2\,a_0^4. \tag{4.29}$$

Now, we multiply or divide appropriately by some combination of a and a_0 to force the appearance of the Hubble function $H(z)$ and the Hubble parameter H_0. We also use equations (4.15) and (4.16) and substitute, leading to:

$$H(z) \quad \leq \quad H_{\text{DEC}} \;\equiv\; H_0\,(1+z)\sqrt{1+\Omega_0\;[(1+z)^4-1]}, \tag{4.30}$$

Again, we have assumed that $z > 0$ in inequality (4.30) so that one is looking into the past. When looking into the future $z < 0$, the inequality is reversed.

Thus the DEC is satisfied if and only if:

$$H_{\text{NEC}} \;\leq\; H(z) \;\leq\; H_{\text{DEC}}, \tag{4.31}$$

where H_{NEC} and H_{DEC} are defined respectively in equations (4.17), and (4.30).

Technical point: Similarly to the situation for the NEC, the bound is guaranteed to be real for all values of $z > -1$ (and hence $z \geq 0$) if $\Omega_0 \in (0,1)$. However, note that when $\Omega_0 \geqslant 1$, there exists a z value for which the expression in the square root becomes negative or zero. This specific value of the z-redshift is given by:

$$z_{\text{DEC}} \quad = \quad \left(\frac{\Omega_0 - 1}{\Omega_0}\right)^{1/4} - 1. \tag{4.32}$$

Note that z_{DEC} is always negative so it is never a problem when looking back into the past. In fact $z_{\text{DEC}}(\Omega_0 = 1.04) = -0.56$ and $z_{\text{DEC}}(\Omega_0 = 1.01) = -0.68$ are well into the future.

Thus the bound on the Hubble function (4.30) is valid for $\Omega_0 \in (0,1),\ \forall\, z \geq -1$, and for $\Omega_0 \geqslant 1$ under the condition that $z \in [z_{\text{DEC}}, +\infty]$. If we are only interested in looking into the past, then the DEC bound holds for $\Omega_0 > 0$ and $z > 0$.

4.3 Energy conditions and the distance scales

In order to obtain bounds on the various distance scales, it is enough to obtain a bound on Peebles' angular diameter distance [21] and then use the different relations between the various distance scales presented in [85, 86]. [5] We choose to work primarily with Peebles' angular diameter distance because it minimizes the number of factors of $(1+z)$ occurring in the various formulae. Recall from section 2.5 that Peebles' angular diameter distance can be defined in its exact form as [21]:

$$d_P(z) \;=\; a_0\,\sin_k\left\{\frac{c}{H_0\,a_0}\int_0^z\frac{H_0}{H(z)}\,\mathrm{d}z\right\}, \tag{4.33}$$

where

$$\sin_k(x) \;=\; \begin{cases}\sin(x), & k=+1;\\ x, & k=0;\\ \sinh(x), & k=-1.\end{cases} \tag{4.34}$$

By changing variables and adopting definitions as in equations (4.15) and (4.16), we can rewrite Peebles' angular diameter distance in an alternative exact general form, $\forall\ z\in[-1,+\infty)$ and $\forall$ fixed Ω_0:[6]

$$d_P(z) \;=\; \frac{c}{H_0}\,\frac{\sinh\left[\sqrt{1-\Omega_0}\int_0^z\frac{H_0}{H(z)}\,\mathrm{d}z\right]}{\sqrt{1-\Omega_0}}, \tag{4.35}$$

where we note

$$\Omega_0 \quad \begin{cases}>1, & k=+1;\\ =1, & k=0;\\ <1, & k=-1.\end{cases} \tag{4.36}$$

Observe that by continuity of the functions $\sin x/x$ and $\sinh x/x$ as $x\to 0$, the function $d_P(z)$ is also continuous as $\Omega_0\to 1^{\pm}$. For convenience, from equation (4.35), the angular diameter distance is given by

$$d_P(z) \;=\; \frac{c}{H_0}\,\frac{\sinh\left[\sqrt{1-\Omega_0}\;J\right]}{\sqrt{1-\Omega_0}}, \tag{4.37}$$

where J is the integral defined by

$$J \;=\; \int_0^z\frac{H_0}{H(z)}\,\mathrm{d}z \;=\; H_0\,a_0\int_a^{a_0}\frac{\mathrm{d}a}{a\,\dot{a}}. \tag{4.38}$$

The procedure now is as follows: The energy conditions provide bounds on $H(z)$, which allow us to obtain a bound on the integral J. Then provided the function $\sin_k$ is monotonic on the interval $z\in[0,+\infty]$, (or at least some sub-interval $z\in[0,z_{\max}]$), we can derive a bound on the angular diameter distance on this same domain.

[5] Peebles' angular diameter distance is equal to Weinberg's proper motion distance [20], and is also equal to D'Inverno's version of luminosity distance [7]. Details on how the various distance scales are inter-related can be found in section 2.5.

[6] Another notation that is sometimes used is $\Omega_k = 1-\Omega_0$, so that $k=-\mathrm{sign}(\Omega_k)$.

4.3.1 NEC:

The null energy condition gives a bound on H in equation (4.17) leading to the inequality,

$$J \;=\; \int_0^z \frac{H_0}{H(z)}\,\mathrm{d}z \;\leq\; J_{\mathrm{NEC}} \;\equiv\; \int_0^z \frac{\mathrm{d}z}{\sqrt{\Omega_0 + [1-\Omega_0]\,(1+z)^2}}. \tag{4.39}$$

We integrate, and substitute the resulting bound back into the angular diameter distance. In the general case, we obtain the algebraic bound

$$d_P(z) \;\leq\; d_{P_{\mathrm{NEC}}} = \tfrac{c}{H_0\,\Omega_0}\left[1+z-\sqrt{\Omega_0+(1-\Omega_0)\,(1+z)^2}\right];$$

$$\begin{cases} \forall\,\Omega_0 \;\leq\; 1,\ \forall\,z \;\in [0,+\infty]; \\ \forall\,\Omega_0 \;>\; 1,\ \forall\,z \;\in [0,z_{\mathrm{NEC}}). \end{cases} \tag{4.40}$$

Special case: Note that as $\Omega_0 \to 1$ ($k=0$), we have

$$d_P(z) \;\leq\; d_{P_{\mathrm{NEC}}} = \frac{c\,z}{H_0} \qquad \Omega_0 = 1; \quad \forall\, z \in [0,+\infty]. \tag{4.41}$$

so we find the same particular result as in [71], that is,

$$d_L(z) \;\leq\; d_{L_{\mathrm{NEC}}} = \frac{c\,z\,(1+z)}{H_0} \qquad \Omega_0 = 1; \quad \forall\, z \in [0,+\infty]. \tag{4.42}$$

Comment: In contrast, note that equation (4.40) is the general case, now valid for all values of $k \in \{-1,0,+1\}$. The apparently rather different equation (15) of reference [72], which involves (hyperbolic) trigonometric functions and their inverses, can be viewed as an intermediate step in deriving the comparatively simple and general algebraic result in equation (4.40) above.

Note that as $\Omega_0 \to 1$, equation (4.40) can be developed in a Taylor series as

$$d_{P_{\mathrm{NEC}}}(z) \;=\; \frac{c\,z}{H_0} + \frac{c\,z^2}{2\,H_0}\,(\Omega_0-1) + \mathcal{O}\left([\Omega_0-1]^2\right). \tag{4.43}$$

If instead one performs a low-redshift expansion, then for general Ω_0

$$d_{P_{\mathrm{NEC}}}(z) = \frac{cz}{H_0}\left\{1+\frac{(\Omega_0-1)\,z}{2}+\mathcal{O}(z^2)\right\}. \tag{4.44}$$

4.3.2 WEC:

The weak energy condition gives a new (but weak) bound on $H(z)$ as in equation (4.22), but only for $\Omega_0 \in (0,1)$, leading to the inequality:

$$J \;=\; \int_0^z \frac{H_0}{H(z)}\mathrm{d}z \;\leq\; J_{\mathrm{WEC}} \;\equiv\; \int_0^z \frac{\mathrm{d}z}{\sqrt{1-\Omega_0}\,(1+z)} = \frac{\ln(1+z)}{\sqrt{1-\Omega_0}}. \tag{4.45}$$

We integrate, and substitute the resulting bound back into the angular diameter distance to obtain:

$$d_P(z) \;\leq\; d_{P_{\rm WEC}} = \tfrac{c}{2\,H_0\,\sqrt{1-\Omega_0}}\;\tfrac{z(2+z)}{(1+z)}; \qquad \boxed{4.46}$$

$$\forall\,\Omega_0 \;\in\; (0,1),\; \forall\; z \;\in [0,+\infty].$$

Comment: Note that $d_{P_{\rm NEC}} \leq d_{P_{\rm WEC}}$. Thus the bound $d_{P_{\rm WEC}}$ is not very useful.

4.3.3 SEC:

This energy condition gives a bound on $H(z)$ in (4.25), and therefore

$$J \;\;\leq\;\; J_{\rm SEC} \equiv \int_0^z \frac{{\rm d}z}{1+z} = \ln(1+z). \qquad \boxed{4.47}$$

In the general case, we obtain the bound on the angular diameter distance (*cf*. the related equation (17) of [72]):

$$d_P(z) \;\leq\; d_{P_{\rm SEC}}(z) = \tfrac{c}{H_0}\;\tfrac{\sinh\left[\sqrt{1-\Omega_0}\;\ln(1+z)\right]}{\sqrt{1-\Omega_0}}; \qquad \boxed{4.48}$$

$$\begin{cases} \forall\,\Omega_0 \;\leq\; 1,\; \forall\; z \;\in [0,+\infty]; \\ \forall\,\Omega_0 \;>\; 1,\; \forall\; z \;\in [0,z_{\rm max}]. \end{cases}$$

In particular, we can make this much more explicit than the analysis in [72] by evaluating:

- For $k=-1$, that is $\forall\;\Omega_0<1$, and $\forall\;z\in[0,+\infty]$,

$$d_{P_{\rm SEC}}(z) = \frac{c}{H_0}\;\frac{(1+z)^{\sqrt{1-\Omega_0}} - (1+z)^{-\sqrt{1-\Omega_0}}}{2\sqrt{1-\Omega_0}}; \qquad \boxed{4.49}$$

- For $k=0$, that is when $\Omega_0=1$, and $\forall\;z\in[0,+\infty]$ [7]

$$d_{P_{\rm SEC}}(z) = \frac{c}{H_0}\;\ln\left(1+z\right); \qquad \boxed{4.50}$$

- For $k=+1$, that is $\forall\;\Omega_0>1$, and $\forall\;z\in[0,z_{\rm max}]$,

$$d_{P_{\rm SEC}}(z) = \frac{c}{H_0}\;\frac{\sin\left[\sqrt{\Omega_0-1}\;\ln\left(1+z\right)\right]}{\sqrt{\Omega_0-1}}, \qquad \boxed{4.51}$$

  where [8]

$$z_{\rm max} = \exp\left(\frac{\pi}{2\sqrt{\Omega_0-1}}\right) - 1. \qquad \boxed{4.52}$$

  In fact, since $z_{\rm max}(\Omega_0=1.04)=2575$ and $z_{\rm max}(\Omega_0=1.01)=6.6\times10^6$, we see that this constraint is never a significant limitation on the validity of the bounds.

[7] *Cf*. the equivalent special case result in [71].

[8] If $J\;\leq\;J_{\rm SEC}$ then for the sine function we have $\sin(\sqrt{\Omega_0-1}\;J)\;\leq\;\sin(\sqrt{\Omega_0-1}\;J_{\rm SEC})$, ***provided*** that $0\leq\sqrt{\Omega_0-1}\;J_{\rm SEC}\leq\pi/2$, thus leading to the condition that $z\;\leq\;z_{\rm max}=\exp\left(\frac{\pi}{2\sqrt{\Omega_0-1}}\right)-1$.

Note that as $\Omega_0 \to 1^+$, $z_{\max} \to +\infty$, and equation (4.48) can be developed in a Taylor series as

$$d_{P_{\rm SEC}}(z) \;=\; \frac{c}{H_0}\,\ln(1+z) - \frac{c}{6H_0}\,\left[\ln(1+z)\right]^3\,(\Omega_0-1) + O\left([\Omega_0-1]^2\right). \tag{4.53}$$

If instead one performs a low-redshift expansion, then for general Ω_0

$$d_{P_{\rm SEC}}(z) = \frac{cz}{H_0}\left\{1 - \frac{z}{2} + \mathcal{O}(z^2)\right\}. \tag{4.54}$$

4.3.4 DEC:

Remember that to satisfy the DEC, the Hubble function needs to satisfy both the NEC, inequality (4.17), ***and*** the second inequality (4.30). As a consequence, in order for the DEC to hold, Peebles' angular diameter distance must satisfy inequality (4.40), ***and*** a second inequality to be derived below. From equation (4.30), we obtain

$$J = \int_0^z \frac{H_0}{H(z)}\,\mathrm{d}z \;\;\geq\;\; J_{\rm DEC} \;\equiv\; \int_0^z \frac{\mathrm{d}z}{(1+z)\sqrt{1+\Omega_0\;[(1+z)^4-1]}}. \tag{4.55}$$

This integration is a bit more tricky than the previous integrations for the NEC and SEC. We obtain:

$$J_{\rm DEC} \;=\; \frac{1}{2\sqrt{1-\Omega_0}}\,\ln\left\{\frac{\left(1-\Omega_0+\sqrt{1-\Omega_0}\right)(1+z)^2}{1-\Omega_0+\sqrt{1-\Omega_0}\sqrt{1+\Omega_0\left[(1+z)^4-1\right]}}\right\}. \tag{4.56}$$

In the general case, this leads to the following algebraic lower bound on the angular diameter distance:

$$d_P(z) \;\;\geq\;\; d_{P_{\rm DEC}}(z) \equiv \frac{c}{H_0(1+z)}\sqrt{\frac{\sqrt{1+\Omega_0\left[(1+z)^4-1\right]}-\left(1+\Omega_0\left[(1+z)^2-1\right]\right)}{2\Omega_0\,(1-\Omega_0)}};$$

$$\begin{cases} \forall\,\Omega_0 \;\leq\; 1,\; \forall\; z \;\in [0,+\infty]; \\ \forall\,\Omega_0 \;>\; 1,\; \forall\; z \;\in (z_{\rm DEC},+\infty]. \end{cases} \tag{4.57}$$

This simplifies and makes explicit consequences that are implicit in the rather different-looking equations (19) and (20) of [72], which are presented in terms of (hyperbolic) trigonometric functions and their inverses, and which can be viewed as intermediate stages in deriving this much simpler algebraic result. Note that in contrast to the situation for the SEC, there is no constraint on a maximum value for z coming from the requirement that the sine function be monotonic.

The lower bound of the DEC in equation (4.57) can also be represented in a Taylor series as $\Omega_0 \to 1$,

$$d_{P_{\rm DEC}}(z) = \frac{c}{H_0}\frac{z\,(2+z)}{2\,(1+z)^2} + \frac{c}{H_0}\frac{z^2\,(2+z)^2\left(3z^2+6z+4\right)}{16\,(1+z)^6}\,(\Omega_0-1) + O\left([\Omega_0-1]^2\right). \tag{4.58}$$

If instead one performs a low-redshift expansion, then for general Ω_0

$$d_{P_{\rm DEC}}(z) = \frac{cz}{H_0}\left\{1 - \frac{(2\Omega_0+1)\,z}{2} + \mathcal{O}(z^2)\right\}. \tag{4.59}$$

4.3.5 Energy conditions and Supernovae data:

We can now plot the angular diameter distance bounds (NEC, WEC, SEC, and DEC) and compare them with the data from the supernova datasets. We have used data from the supernova legacy survey (legacy05) [28, 1] and the Riess *et al.* ***gold*** dataset of 2006 (gold06) [3]. See section 2.7 for detailed information on the supernovae datasets.

Figure 4.1 compares the upper bounds (NEC and SEC), and the lower bound (DEC), with the legacy05 dataset. In contrast figure 4.2 uses the gold06 dataset.

To satisfy the NEC, the data must lie under the ***red solid*** bound, and we can see that most of the data seem to satisfy this condition.

For the SEC to hold, the data must lie under the ***black dashdot*** bound. Visually it seems "obvious" that the data significantly violate the SEC.

Finally, the DEC is satisfied if both: (1) the NEC is satisfied, and (2) if the data lies above the ***magenta dashed*** lower bound. This latter condition is well satisfied for the bulk of the data, therefore satisfying the DEC is dependent on the NEC holding.

As is traditional for estimates of cosmological distance, we plot only one-sigma statistical uncertainties, without any allowance for systematic uncertainties. Any realistic attempt at more careful treatment of the systematics, and/or going to 3-sigma error bars, makes it clear that the interpretation of these plots is an extremely subtle matter fraught with uncertainties and unknowns [85, 86].

There seem to be noticeable visual differences when looking at figures 4.1, or 4.2, which make it tricky to conclude whether the classical energy conditions are satisfied or not by just looking at the supernova data in isolation. For example, there are a few supernovae data in the redshift range $0.8 < z < 1$ that appear to violate the NEC in an obvious manner for the legacy05 dataset in figure (4.1) . However, the violation does not appear to be as dramatic when looking at the same range of data in the gold06 dataset in figure (4.2). Another example is that the NEC naively seems to be violated for data in the redshift range $0.4 < z < 0.6$ in the gold06 dataset. On the contrary, even if there are less data in the legacy05 dataset, one cannot draw the same conclusion.

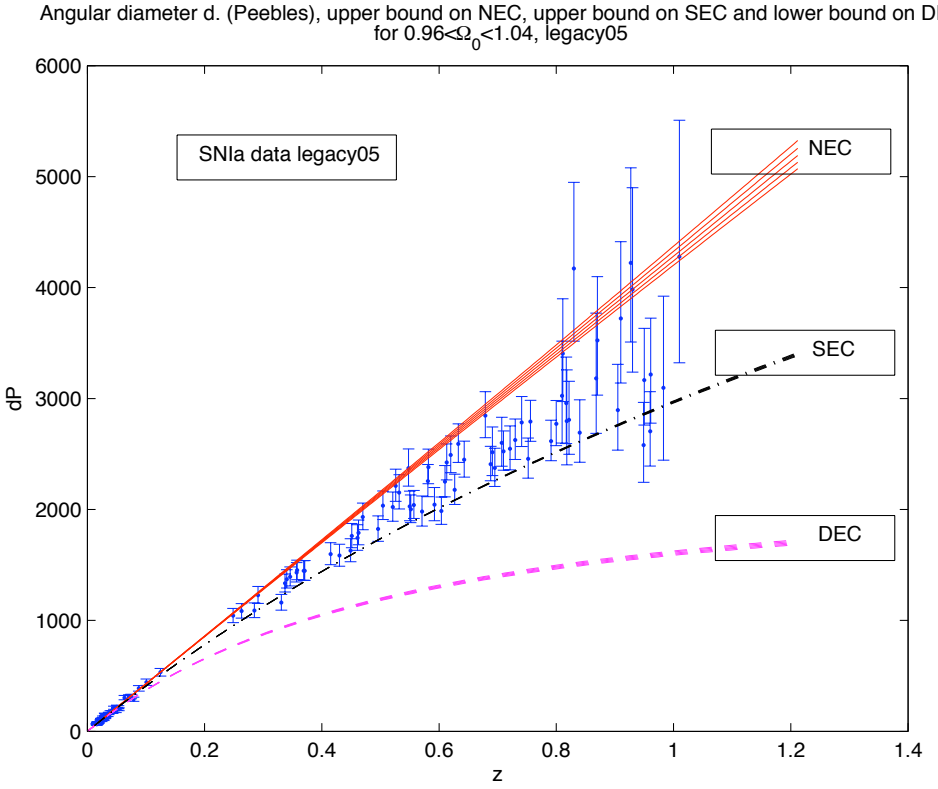


Figure 4.1: This figure shows Peebles' angular diameter distance $d_P(z)$ as a function of the z-redshift from the nearby and legacy survey, legacy05 dataset [28, 1]. Data under the ***red solid lines*** satisfy the NEC/WEC, data under the ***black dashdot lines*** satisfy the SEC, data under the ***red solid lines*** and above the ***magenta dashed lines*** satisfy the DEC. The 5 lines for each energy conditions correspond to varying values of the parameter $\Omega_0 = \{0.96, 0.98, 1.00, 1.02, 1.04\}$. The value of the Hubble constant is taken to be $H_0 = 70$ km/s/Mpc.

In contrast to references [71, 72, 73], we believe we cannot draw any firm conclusions by using the low-redshift linear part of the distance scale curve; as in those articles the data has been scaled to ***enforce*** a particular value of H_0 that was chosen to be compatible with other (non-SNae) determinations of the Hubble parameter. It is important to realise that the ***slope*** of the bounds at $z = 0$ depends sensitively on the estimate of H_0 one adopts, and can further be affected by the value of the magnitude ***offset*** reported for the data.

More generally, note in particular that for low redshift the luminosity distance is bounded, ***both above and below***, by constraints of the form $cz/H_0 + \mathcal{O}(z^2)$. Thus for data with ***any*** statistical uncertainties whatsoever, at low enough z, one would ***expect*** roughly half the supernovae to ***violate*** one or more of these bounds.

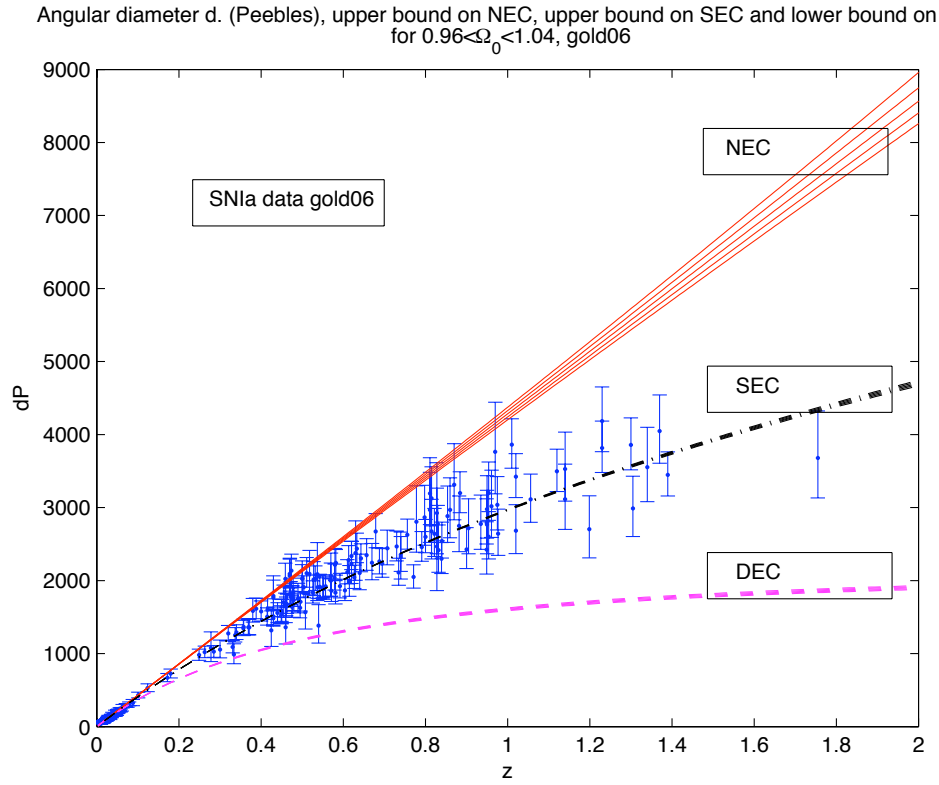


Figure 4.2: This figure shows Peebles' angular diameter distance $d_P(z)$ as a function of the z-redshift from the gold06 dataset [3]. Data under the ***red solid lines*** satisfy the NEC/WEC, data under the black dashdot lines satisfy the SEC, data under the red solid lines and above the ***magenta dashed lines*** satisfy the DEC. The 5 lines for each energy conditions correspond to varying values of the parameter $\Omega_0 = \{0.96, 0.98, 1.00, 1.02, 1.04\}$. The value of the Hubble constant is taken to be $H_0 = 70$ km/s/Mpc.

4.4 Energy conditions and the lookback time $\mathrm{T}(\mathrm{z})$

The lookback time is defined as [63, 64, 65]

$$\begin{aligned} T(z) &= \int_a^{a_0} \mathrm{d}t = \int \frac{\mathrm{d}t}{\mathrm{d}a}\,\mathrm{d}a = \int \frac{a}{\dot{a}}\,\frac{\mathrm{d}a}{a} \\ &= \int \frac{1}{H}\,\frac{\mathrm{d}[a_0/(1+z)]}{a_0/(1+z)} = -\int \frac{1}{H}\,\frac{\mathrm{d}z/(1+z)^2}{1/(1+z)}. \end{aligned} \tag{4.60}$$

That is

$$T(z) = \int_0^z \frac{1}{(1+z)\,H(z)}\,\mathrm{d}z. \tag{4.61}$$

In order to obtain bounds on the lookback time for the different energy conditions NEC, WEC, SEC, and DEC, we again use the bounds on the Hubble parameter $H(z)$.

4.4.1 NEC:

The null energy condition gives a bound on $H(z)$ in equation (4.17), leading to the inequality

$$T(z) = \int_0^z \frac{1}{(1+z)\,H(z)}\,\mathrm{d}z \;\leq\; T_{\mathrm{NEC}}(z) \equiv \int_0^z \frac{1}{(1+z)\,H_{\mathrm{NEC}}(z)}\,\mathrm{d}z \tag{4.62}$$

We integrate, and see that in this case the lookback time is bounded by an elementary function (of an algebraic argument) [1]

$$T(z) \;\leq\; T_{\mathrm{NEC}}(z) = \frac{1}{H_0\sqrt{\Omega_0}}\,\ln\left[\frac{(1+z)\left(1+\sqrt{\Omega_0}\right)}{\sqrt{\Omega_0+(1-\Omega_0)\,(1+z)^2}+\sqrt{\Omega_0}}\right], \tag{4.63}$$

this bound is valid $\forall\,\Omega_0 \;\leq\; 1,\ \forall\,z>0$ and $\forall\,\Omega_0 \;>\; 1,\ \forall\,z \;\in [0, z_{\mathrm{NEC}}]$.
Alternatively, this bound can be rewritten as

$$T(z) \leq T_{\mathrm{NEC}}(z) \;=\; \frac{1}{H_0\sqrt{\Omega_0}}\ln\left[\frac{\sqrt{\Omega_0+(1-\Omega_0)\,(1+z)^2}-\sqrt{\Omega_0}}{\left(1-\sqrt{\Omega_0}\right)(1+z)}\right]. \tag{4.64}$$

Using the standard result that $\sinh^{-1} x = \ln(x+\sqrt{x^2+1})$, we can for $k=-1$ (that is, $\Omega_0<1$) also re-cast this as

$$T(z) \leq T_{\mathrm{NEC}}(z) \;=\; \frac{1}{H_0\sqrt{\Omega_0}}\left\{\sinh^{-1}\left(\sqrt{\frac{\Omega_0}{1-\Omega_0}}\right) - \sinh^{-1}\left(\frac{1}{1+z}\,\sqrt{\frac{\Omega_0}{1-\Omega_0}}\right)\right\}. \tag{4.65}$$

Equivalent formulae can be found in reference [64]. Similarly for $k=+1$ (that is, $\Omega_0>1$) we can use the fact that $\cosh^{-1} x = \ln(x+\sqrt{x^2-1})$ to obtain [2]

$$T(z) \leq T_{\mathrm{NEC}}(z) \;=\; \frac{1}{H_0\sqrt{\Omega_0}}\left\{\cosh^{-1}\left(\sqrt{\frac{\Omega_0}{\Omega_0-1}}\right) - \cosh^{-1}\left(\frac{1}{1+z}\,\sqrt{\frac{\Omega_0}{\Omega_0-1}}\right)\right\}. \tag{4.66}$$

Finally, the upper bound derived from the NEC in equation (4.63), or equivalently any of equations (4.64)–(4.65)–(4.66), can also be represented in a Taylor series as $\Omega_0 \to 1$:

$$T_{\mathrm{NEC}}(z) \;=\; \frac{\ln(1+z)}{H_0} + \frac{z^2+2z-2\ln(1+z)}{4H_0}\,(\Omega_0-1) + O\left([\Omega_0-1]^2\right). \tag{4.67}$$

In particular for the special case $\Omega_0=1$ we recover the results of references [64] and [73].

[1] Equivalent formulae can be found in reference [64], see also equation (14) of [73].

[2] Note that this formula is valid only for $z \leq z_{\mathrm{NEC}}$, since otherwise the argument of the $\cosh^{-1}$ is less than unity.

4.4.2 WEC:

This energy conditions gives a bound on $H(z)$ in (4.22) for $\Omega_0 \in (0,1)$ only. Thereby it can be deduced that

$$T(z) = \int_0^z \frac{1}{(1+z)\,H(z)}\,\mathrm{d}z \;\leq\; T_{\mathrm{WEC}}(z) \equiv \int_0^z \frac{1}{(1+z)\,H_{\mathrm{WEC}}(z)}\,\mathrm{d}z, \tag{4.68}$$

providing the (weak) bound

$$T(z) \;\leq\; T_{\mathrm{WEC}}(z) \;\;=\;\; \frac{z}{H_0\,\sqrt{1-\Omega_0}\;(1+z)}. \tag{4.69}$$

Again, this provided no additional useful information beyond the NEC-derived bound.

4.4.3 SEC:

This energy condition gives a bound on $H(z)$ in (4.25). Thereby it can be deduced (as in the articles [63, 64, 65, 73]) that,

$$T(z) = \int_0^z \frac{1}{(1+z)\,H(z)}\,\mathrm{d}z \;\leq\; T_{\mathrm{SEC}}(z) \equiv \int_0^z \frac{1}{(1+z)\,H_{\mathrm{SEC}}(z)}\,\mathrm{d}z, \tag{4.70}$$

that is

$$T_{\mathrm{SEC}}(z) \;\;=\;\; \frac{z}{H_0\;(1+z)}. \tag{4.71}$$

The above result equation (4.71) is completely independent of the value of the parameter Ω_0. Note that this result was first introduced by Visser in [63, 64, 65].

4.4.4 DEC:

Remember that to satisfy the DEC, the Hubble function needs to satisfy the NEC, inequality (4.17), ***and*** inequality (4.30). As a consequence, in order for the DEC to hold, the lookback time must satisfy inequality (4.63), ***and*** a second inequality that we shall derive below. From equation (4.30), we obtain

$$T(z) = \int_0^z \frac{1}{(1+z)\,H(z)}\,\mathrm{d}z \;\geq\; T_{\mathrm{DEC}}(z) \equiv \int_0^z \frac{1}{(1+z)\,H_{\mathrm{DEC}}(z)}\,\mathrm{d}z, \tag{4.72}$$

that is

$$T_{\mathrm{DEC}}(z) = \frac{1}{H_0}\int_0^z \frac{1}{(1+z)^2\,\sqrt{1+\Omega_0\,((1+z)^4-1)}}\,\mathrm{d}z, \tag{4.73}$$

The integration of this bound is considerably harder than for the other energy conditions, and will require us to use hypergeometric functions. Let us first write

$$T_{\mathrm{DEC}}(z) = \frac{1}{H_0\,\sqrt{\Omega_0}}\int_0^z \frac{1}{(1+z)^4\,\sqrt{1-(1-\Omega_0^{-1})\,(1+z)^{-4}}}\,\mathrm{d}z, \tag{4.74}$$

and then, (following the procedure of [64]), apply the binomial theorem

$$\left[1-(1-\Omega_0^{-1})\,(1+z)^{-4}\right]^{-1/2}=\sum_{n=0}^{\infty}\binom{-1/2}{n}\,(-1)^n\,(1-\Omega_0^{-1})^n\,(1+z)^{-4n}. \tag{4.75}$$

Now the binomial series will converge provided

$$\left|(1-\Omega_0^{-1})\,(1+z)^{-4}\right|<1, \tag{4.76}$$

and in view of the region we are integrating over, this means that the series for the lookback time will converge provided

$$\left|1-\Omega_0^{-1}\right|<1, \qquad \text{that is} \qquad \Omega_0\in(1/2,\infty). \tag{4.77}$$

Subject to this condition we can integrate, and obtain the convergent series

$$T_{\rm DEC}(z)=\frac{1}{H_0\,\sqrt{\Omega_0}}\sum_{n=0}^{\infty}\binom{-1/2}{n}(-1)^n\frac{1}{4n+3}(1-\Omega_0^{-1})^n\,\left[1-(1+z)^{-4n-3}\right]. \tag{4.78}$$

As a practical matter, for many purposes this series representation may be enough, but we can tidy things up somewhat by first defining

$$S(x)=\sum_{n=0}^{\infty}\binom{-1/2}{n}\,\frac{(-x)^n}{4n+3}=\frac{1}{3}+\frac{x}{14}+\frac{3x^2}{88}+\mathcal{O}(x^3), \tag{4.79}$$

in which case

$$T_{\rm DEC}(z)=\frac{1}{H_0\,\sqrt{\Omega_0}}\left\{S\left(1-\Omega_0^{-1}\right)-(1+z)^{-3}\,S\left(\frac{(1-\Omega_0^{-1})}{(1+z)^4}\right)\right\}. \tag{4.80}$$

Explicitly, the first two terms in the "near–spatially-flat" Taylor series expansion is

$$\begin{aligned}T_{\rm DEC}(z) \;\;=\;\; & \frac{1}{H_0\,\sqrt{\Omega_0}}\left\{\frac{1-(1+z)^{-3}}{3}+\left(1-\Omega_0^{-1}\right)\frac{1-(1+z)^{-7}}{14}\right.\\ & \left.+\mathcal{O}\left(\left[1-\Omega_0^{-1}\right]^2\right)\right\}.\end{aligned} \tag{4.81}$$

For the restricted special case $k=0$, that is $\Omega_0=1$, the leading term reduces to equation (20) in [73].

Returning to the exact series result, we finally recognize that $S(x)$ is a particular example of hypergeometric series, [3] and so write

$$S(x)=\sum_{n=0}^{\infty}\binom{-1/2}{n}\,\frac{(-x)^n}{4n+3}=\frac{1}{3}\,{}_2F_1\left(\frac{1}{2},\frac{3}{4};\frac{7}{4};x\right). \tag{4.83}$$

[3] The classical hypergeometric series is given by

$${}_2F_1\,(a,b;c;x)=\sum_{n=0}^{\infty}\frac{(a)_n(b)_n}{(c)_n}\,\frac{x^n}{n!}, \tag{4.82}$$

where $(a)_n=a(a+1)(a+2)(a+n-1)$ is the rising factorial, or Pochhammer symbol. The series is in general a convergent power series for values of x such that $|x|<1$.

Therefore

$$T_{\rm DEC}(z) = \frac{1}{3\,H_0\,\sqrt{\Omega_0}} \times \left\{ {}_2F_1\left(\frac{1}{2},\frac{3}{4};\frac{7}{4};1-\Omega_0^{-1}\right) - (1+z)^{-3}\ {}_2F_1\left(\frac{1}{2},\frac{3}{4};\frac{7}{4};\frac{(1-\Omega_0^{-1})}{(1+z)^4}\right)\right\}. \tag{4.84}$$

Again this agrees with and generalizes the results reported in [64]. Of course writing the result in terms of hypergeometric functions does not necessarily give one much additional physical insight — for physical insight the series $S(x)$ is sufficient, and the realization that one is in fact dealing with a hypergeometric function is likely to be useful only if for some reason one wishes to numerically programme the bound into a computer. [4]

4.5 Energy conditions and the Omega parameter $\Omega(z)$

We have the following identity

$$\Omega - 1 = \frac{k\,c^2}{a^2\,H^2} = \frac{k\,c^2}{a_0^2\,H_0^2}\,\frac{a_0^2}{a^2}\,\frac{H_0^2}{H^2} = (\Omega_0-1)\,(1+z)^2\,\frac{H_0^2}{H^2}. \tag{4.85}$$

That is

$$\Omega(z) = 1 + (\Omega_0-1)\,(1+z)^2\,\frac{H_0^2}{H(z)^2}. \tag{4.86}$$

Therefore, a bound on $H(z)$ automatically implies a bound on $\Omega(z)$.

4.5.1 NEC:

The null energy condition gives a bound on $H(z)$, as in equation (4.17), leading to a simple rational polynomial bound

$$\Omega_{\rm NEC} = \frac{\Omega_0}{\Omega_0 + (1-\Omega_0)\,(1+z)^2}, \tag{4.87}$$

and the inequalities

$$\text{if } \Omega_0 < 1, \quad \forall\, z > 0, \quad \text{then } \Omega(z) \geq \Omega_{\rm NEC}; \tag{4.88}$$

$$\text{if } \Omega_0 = 1, \quad \forall\, z > 0, \quad \text{then } \Omega(z) = \Omega_{\rm NEC} = 1; \tag{4.89}$$

$$\text{if } \Omega_0 > 1, \quad \forall\, z > 0, \quad \text{then } \Omega(z) \leq \Omega_{\rm NEC}. \tag{4.90}$$

Note that as $\Omega_0 \to 1$, equation (4.87) can be developed in a Taylor series as

$$\Omega_{\rm NEC} = 1 + (1+z)^2\,(\Omega_0-1) + O\left([\Omega_0-1]^2\right). \tag{4.91}$$

[4] The result can also be cast in terms of elliptic integrals, as mentioned in [64] and [73], but this does not appear to be particularly illuminating.

4.5.2 WEC:

The weak energy condition gives a bound on $H(z)$, as in equation (4.22), but for $\Omega_0 \in (0,1)$ only, leading to the trivial result $\Omega_{\rm WEC} = 0$, and the trivial inequality

$$\text{if } \Omega_0 < 1, \quad \forall\, z > 0, \quad \Omega \;\geq\; \Omega_{\rm WEC} = 0. \tag{4.92}$$

That is, this bound is not useful, except as a consistency check.

4.5.3 SEC:

The strong energy condition gives a bound on $H(z)$, as in equation (4.25), leading to a trivial constant bound:

$$\Omega_{\rm SEC} \equiv \Omega_0, \tag{4.93}$$

and the inequalities

$$\text{if } \Omega_0 < 1, \quad \forall\, z > 0, \quad \text{then } \Omega(z) \;\geq\; \Omega_{\rm SEC} = \Omega_0; \tag{4.94}$$
$$\text{if } \Omega_0 = 1, \quad \forall\, z > 0, \quad \text{then } \Omega(z) \;=\; \Omega_{\rm SEC} = \Omega_0 = 1; \tag{4.95}$$
$$\text{if } \Omega_0 > 1, \quad \forall\, z > 0, \quad \text{then } \Omega(z) \;\leq\; \Omega_{\rm SEC} = \Omega_0. \tag{4.96}$$

4.5.4 DEC:

The dominant energy condition gives a bound on $H(z)$, as in equation (4.30), again leading to a rational polynomial bound:

$$\Omega_{\rm DEC} = \frac{\Omega_0\,(1+z)^4}{1+\Omega_0\left[(1+z)^4-1\right]}, \tag{4.97}$$

and the inequalities

$$\text{if } \Omega_0 < 1, \quad \forall\, z > 0, \quad \text{then } \Omega(z) \;\leq\; \Omega_{\rm DEC}; \tag{4.98}$$
$$\text{if } \Omega_0 = 1, \quad \forall\, z > 0, \quad \text{then } \Omega(z) \;=\; \Omega_{\rm DEC} = 1; \tag{4.99}$$
$$\text{if } \Omega_0 > 1, \quad \forall\, z \in [0, z_{\rm DEC}], \quad \text{then } \Omega(z) \;\geq\; \Omega_{\rm DEC}. \tag{4.100}$$

Note that as $\Omega_0 \to 1$, equation (4.97) can be developed in a Taylor series as

$$\Omega_{\rm DEC} \;=\; 1 + \frac{(\Omega_0 - 1)}{(1+z)^4} + O\left([\Omega_0 - 1]^2\right). \tag{4.101}$$

These bounds on $\Omega(z)$ are potentially of interest with regard to cosmological nucleosynthesis, which is effectively sensitive to $\Omega(z_{\rm nucleosynthesis})$. More generally, any bound on the number of relativistic particle species at any particular epoch can be converted, with a little work and some technical assumptions, into a bound on the Omega parameter at that epoch.

4.6 Energy conditions and the density $\rho(z)$

We have the following identity

$$\rho \;=\; 3\,\Omega\,H^2 = 3\,\left[1+(\Omega_0-1)\,(1+z)^2\,\frac{H_0^2}{H^2}\right]\,H^2. \tag{4.102}$$

That is

$$\rho(z) = 3H(z)^2 + 3\,(\Omega_0-1)\;\;(1+z)^2\;\;H_0^2, \tag{4.103}$$

showing that a bound on $H(z)$ automatically implies a bound on $\rho(z)$.

Alternatively, we can also write the following identity

$$\begin{aligned}\rho &= 3\left(H^2+\frac{kc^2}{a^2}\right) = 3H_0^2\,\frac{H^2}{H_0^2}+\frac{3kc^2}{a_0^2}\,\frac{a_0^2}{a^2}\\ &= \frac{\rho_0}{\Omega_0}\,\frac{H^2}{H_0^2}+\rho_0\left(1-\frac{1}{\Omega_0}\right)(1+z)^2.\end{aligned} \tag{4.104}$$

That is

$$\rho(z) = \rho_0\left[\frac{1}{\Omega_0}\,\frac{H(z)^2}{H_0^2}+\left(1-\frac{1}{\Omega_0}\right)\,(1+z)^2\right]. \tag{4.105}$$

Again, a bound on $H(z)$ automatically implies a bound on $\rho(z)$.

4.6.1 NEC:

The null energy condition gives a bound on $H(z)$, as in equation (4.17), leading to

$$\rho_{\rm NEC} = 3\Omega_0\;H_0^2 = \rho_0, \tag{4.106}$$

and the inequality,

$$\forall\;\Omega_0,\quad \forall\;z>0,\quad \rho(z)\;\geq\;\rho_{\rm NEC}=\rho_0. \tag{4.107}$$

This inequality was also derived by more direct means in [63, 64].

4.6.2 WEC:

The weak energy condition gives a bound on $H(z)$, as in equation (4.22), leading to

$$\rho_{\rm WEC} = 0, \tag{4.108}$$

and the inequality,

$$\forall\;\Omega_0<1,\quad \forall\;z>0,\quad \rho\;\geq\;\rho_{\rm WEC}=0. \tag{4.109}$$

Of course, since by assuming the WEC we have already assumed that $\rho>0$, this bound is not very useful (and is at best a consistency check on the formalism).

4.6.3 SEC:

The strong energy condition gives a bound on $H(z)$, as in equation (4.25), leading to the polynomial bound

$$\rho_{\rm SEC} = 3\Omega_0\ H_0^2\,(1+z)^2 = \rho_0\ (1+z)^2\,, \tag{4.110}$$

and the inequality,

$$\forall\,\Omega_0,\quad \forall\,z>0,\quad \rho(z)\ \geq\ \rho_{\rm SEC} = \rho_0\,(1+z)^2. \tag{4.111}$$

This inequality was also derived by more direct means in [63, 64].

4.6.4 DEC:

The dominant energy condition gives a bound on $H(z)$ in equation (4.30), leading to

$$\rho_{\rm DEC} = 3\Omega_0\ H_0^2\ (1+z)^6 = \rho_0\ (1+z)^6\,, \tag{4.112}$$

and the inequality

$$\forall\,\Omega_0,\quad \forall\,z>0,\quad \rho\ \leq\ \rho_{\rm DEC} = \rho_0\ (1+z)^6\,. \tag{4.113}$$

This inequality was also derived by more direct means in [63, 64].

Note that bounds on the density and the Hubble function are intimately related. A bound on one will automatically provide a bound on the other, and comments made regarding the Hubble bounds can be carried over to this situation as well.

4.7 Energy conditions and the pressure $\mathbf{p(z)}$

For the pressure $p(z)$ things are a little different; we have the following identity involving the second time derivative of the scale factor:

$$p\ \ =\ \ -\frac{\dot a^2}{a^2} - \frac{k\,c^2}{a^2} - 2\frac{\ddot a}{a}. \tag{4.114}$$

But

$$\frac{\dot a^2}{a^2} + 2\frac{\ddot a}{a}\ \ =\ \ \frac{1}{\dot a a^2}\frac{{\rm d}(\dot a^2 a)}{{\rm d}t} = \frac{1}{a^2}\frac{{\rm d}(\dot a^2 a)}{{\rm d}a} = \frac{1}{a^2}\frac{{\rm d}[H^2 a^3]}{{\rm d}a}, \tag{4.115}$$

implying

$$p = -\frac{1}{a^2}\left(\frac{{\rm d}[H^2a^3]}{{\rm d}a} + k\,c^2\right). \tag{4.116}$$

Here the point is that one would need a bound on the ***derivative*** of $H(z)$ in order to get a direct bound on the pressure $p(z)$. This does not appear to lead to anything useful.

However, if one has a bound on $H(z)$ and hence $\rho(z)$, one can indirectly get a constraint on $p(z)$ via the classical energy conditions. Again, this does not appear to lead to anything useful.

4.8 Strategy for general bounds with the w-parameter

In the current section we shall derive some very general bounds in terms of assumptions about the w-parameter, where as usual $w = p/\rho$. Specifically, we shall ask the question: If we know for theoretical reasons, or can observationally determine, that w lies in some restricted range

$$w(z) \in [w_-, w_+], \tag{4.117}$$

between redshift zero and redshift z, what constraint does that place on the cosmological expansion?

Our strategy will be to adopt a standard FLRW cosmology

$$\mathrm{d}s^2 = -c^2\,\mathrm{d}t^2 + a(t)^2\left\{\frac{\mathrm{d}r^2}{1-kr^2} + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\right\}, \tag{4.118}$$

then, (setting $8\pi G_N \to 1$, but explicitly retaining the speed of light c), we have the two Friedmann equations:

$$\rho = 3\left[\frac{\dot{a}^2}{a^2} + \frac{kc^2}{a^2}\right], \qquad \text{and} \qquad p = -\frac{\dot{a}^2}{a^2} - \frac{kc^2}{a^2} - 2\frac{\ddot{a}}{a}. \tag{4.119}$$

Together, these two equations imply the standard conservation law:

$$\dot{\rho} = -3(\rho + p)\frac{\dot{a}}{a}. \tag{4.120}$$

We also have the fundamental ***definitions*** [5]

$$\rho_{\mathrm{Hubble}} = 3\left[\frac{\dot{a}^2}{a^2}\right] = 3H^2, \tag{4.121}$$

and

$$\Omega = \frac{\rho}{\rho_{\mathrm{Hubble}}} = \frac{\rho}{3H^2} = \frac{H^2 + kc^2/a^2}{H^2} = 1 + \frac{kc^2}{a^2H^2}. \tag{4.122}$$

For intermediate steps of the calculation we shall work with the very simple linear equation of state

$$p = w_*\,\rho, \tag{4.123}$$

where w_* is taken to be a constant. Picking some generic cosmological parameter $X(z)$, we shall first calculate $X_{w_*}(z)$, and then (by assuming that $w(z) \in [w_-, w_+]$ from redshift zero out to redshift z, and depending on the direction of the relevant inequality) use this to derive bounds of the form

$$X_{w_-}(z) \leq X(z) \leq X_{w_+}(z), \tag{4.124}$$

or

$$X_{w_+}(z) \leq X(z) \leq X_{w_-}(z). \tag{4.125}$$

[5] Historically it was common to refer to this quantity as the ***critical density***, ρ_{critical}, but with the advent of widespread acceptance of a nonzero cosmological constant, or more generally ***dark energy***, the logical connection between this ***critical*** density and possible re-collapse of the universe has been severed. In a modern context then, it is inappropriate to refer to this as a ***critical*** density, and the considerably more neutral phrase ***Hubble density*** is preferable.

We shall also make the extremely mild assumption that the density is positive

$$\rho > 0. \tag{4.126}$$

This is certainly a completely redundant assumption for $k = 0$ and $k = +1$ FLRW universes. Only for $k = -1$ universes does this provide the ***extremely mild*** additional constraint $H > c/a$, that is, $H(z) > (c/a_0)\,(1+z)$. [6]

4.9 General bounds and the Density $\rho(z)$

We now apply this strategy to the density. From

$$\dot{\rho} = -3(\rho + p)\frac{\dot{a}}{a} = -3\rho(1 + w_*)\frac{\dot{a}}{a}, \tag{4.127}$$

we have

$$\frac{\dot{\rho}}{\rho} = -3(1 + w_*)\frac{\dot{a}}{a}. \tag{4.128}$$

So integrating, for constant w_* we obtain the well-known result

$$\rho_{w_*} = \rho_0(a/a_0)^{-3(1+w_*)} = \rho_0\,(1+z)^{3(1+w_*)}. \tag{4.129}$$

But now ask what happens if we only know that $w_- \leq w(z) \leq w_+$? (Where in the real observable universe $w(z)$ certainly need not be a constant.) Following the above analysis, we find that we must replace equalities by inequalities and so deduce

$$\rho_0\,(1+z)^{3(1+w_-)} \leq \rho(z) \leq \rho_0\,(1+z)^{3(1+w_+)}. \qquad (z \geq 0). \tag{4.130}$$

Note that for $z > 0$ we are looking into the past; in contrast for $-1 < z < 0$, we are looking into the future [86], and the inequality reverses to [7]

$$\rho_0\,(1+z)^{3(1+w_+)} \leq \rho(z) \leq \rho_0\,(1+z)^{3(1+w_-)}; \qquad (-1 < z \leq 0). \tag{4.131}$$

Of course, these simple constraints on the density are by far the most elementary of the inequalities we shall deduce — some of the other inequalities derived below will prove to be much more subtle.

If we now in addition relax our initial constraint on ρ_0, by assuming we only know that the present epoch density lies in some bounded interval

$$\rho_0 \in [\rho_{0_-}, \rho_{0_+}], \qquad \text{that is,} \qquad \rho_{0_-} \leq \rho_0 \leq \rho_{0_+}, \tag{4.132}$$

then these two bounds generalize to

$$\rho_{0_-}(1+z)^{3(1+w_-)} \leq \rho(z) \leq \rho_{0_+}(1+z)^{3(1+w_+)}; \qquad (z \geq 0); \tag{4.133}$$

$$\rho_{0_-}(1+z)^{3(1+w_+)} \leq \rho(z) \leq \rho_{0_+}(1+z)^{3(1+w_-)}; \qquad (-1 < z \leq 0). \tag{4.134}$$

[6] This is equivalent to enforcing $\dot{a} > c$, for a $k = -1$ FLRW universe, noting that $\dot{a}$ is ***not*** a physical velocity, so that $\dot{a} > c$ is a perfectly acceptable physical statement.

[7] Furthermore, note that there is no reason to ever go below $z = -1$, as $z = -1$ corresponds to infinite expansion. Also, note that the ***sign*** of $1 + w_-$ and $1 + w_+$ does not affect these inequalities.

4.10 General bounds and the Density parameter $\Omega(\mathbf{z})$

We have the following ***identity***

$$\Omega - 1 \equiv \frac{k\,c^2}{a^2\,H^2} = \frac{k\,c^2}{a_0^2\,H_0^2}\,\frac{a_0^2}{a^2}\,\frac{H_0^2}{H^2} = (\Omega_0 - 1)\,\frac{\Omega}{\Omega_0}\,\frac{\rho_0}{\rho}. \tag{4.135}$$

This leads to the useful result

$$\frac{\Omega(z) - 1}{\Omega(z)} = \left(\frac{\Omega_0 - 1}{\Omega_0}\right)\,\frac{\rho_0}{\rho(z)}. \tag{4.136}$$

Therefore, a bound on $\rho(z)$ automatically implies a bound on $\Omega(z)$. From the result for $\rho_{w_*}(z)$ presented above, we deduce that bounds on $\Omega(z)$ can be given in terms of

$$\frac{\Omega_{w_*}(z) - 1}{\Omega_{w_*}(z)} = \left(\frac{\Omega_0 - 1}{\Omega_0}\right)\,(1+z)^{-(3w_*+1)}, \tag{4.137}$$

which we can equivalently recast as

$$\Omega_{w_*}(z) = \frac{\Omega_0\,(1+z)^{3w_*+1}}{(1-\Omega_0) + \Omega_0\,(1+z)^{3w_*+1}}. \tag{4.138}$$

We can now use this quantity, which was derived for strictly constant w_*, to bound the density parameter $\Omega(z)$ for a more realistic matter model satisfying the milder condition $w_- \leq w(z) \leq w_+$. We obtain:

- If $\Omega_0 < 1$ (but remember that by assumption $\Omega_0 > 0$) then

$$\Omega_{w_-}(z) \leq \Omega(z) \leq \Omega_{w_+}(z); \qquad (z > 0), \tag{4.139}$$

$$\Omega_{w_+}(z) \leq \Omega(z) \leq \Omega_{w_-}(z); \qquad (-1 < z < 0). \tag{4.140}$$

- If $\Omega_0 = 1$ then $\forall z : \Omega(z) = 1$.

- If $\Omega_0 > 1$,

$$\Omega_{w_+}(z) \leq \Omega(z) \leq \Omega_{w_-}(z); \qquad (z > 0), \tag{4.141}$$

$$\Omega_{w_-}(z) \leq \Omega(z) \leq \Omega_{w_+}(z); \qquad (-1 < z < 0), \tag{4.142}$$

  but note that the bound can break down when the denominator of $\Omega_{w_*}(z)$ equals zero — this occurs at

$$z_\Omega(w_*, \Omega_0) = \left(\frac{\Omega_0 - 1}{\Omega_0}\right)^{1/(3w_*+1)} - 1. \tag{4.143}$$

  The failure of the bound might occur either in the past or the future depending on the value of w_*.

  - If $3w_* + 1 > 0$ then the bound is useful only for $z > z_\Omega(w_*, \Omega_0) < 0$, implying a limitation in the past.
  - If $3w_* + 1 = 0$ then the bound is valid for all z.

– If $3w_* + 1 < 0$ then the bound is useful only for $z < z_\Omega(w_*, \Omega_0) > 0$, implying a limitation in the future.

Note that nothing unusual need happen to the universe itself at $z_\Omega(w_*, \Omega_0)$, it is only the ***bound*** that loses its predictive usefulness. Combining these observations we see that for $\Omega_0 > 1$ it is better (in the sense of reducing the amount of special case exceptions to the general rule) to recast the bounds in the form:

$$\left(\frac{\Omega_0 - 1}{\Omega_0}\right)\,(1+z)^{-(3w_+ +1)} \leq \frac{\Omega(z) - 1}{\Omega(z)} \leq \left(\frac{\Omega_0 - 1}{\Omega_0}\right)\,(1+z)^{-(3w_- +1)} \qquad (4.144)$$

for $z > 0$, and

$$\left(\frac{\Omega_0 - 1}{\Omega_0}\right)\,(1+z)^{-(3w_- +1)} \leq \frac{\Omega(z) - 1}{\Omega(z)} \leq \left(\frac{\Omega_0 - 1}{\Omega_0}\right)\,(1+z)^{-(3w_+ +1)} \qquad (4.145)$$

for $-1 < z < 0$.

4.11 General bounds and the Hubble parameter $\mathrm{H}(\mathrm{z})$

Let us now use the density equation (the first Friedmann equation) and the definition of the density parameter Ω to write

$$H^2 = \frac{\rho}{3} - \frac{kc^2}{a^2} = \frac{\rho}{\rho_0}\frac{\rho_0}{3} - \frac{a_0^2}{a^2}\frac{kc^2}{a_0^2} = \frac{\rho}{\rho_0}\Omega_0 H_0^2 - \frac{a_0^2}{a^2}(\Omega_0 - 1)H_0^2. \qquad (4.146)$$

That is, as an ***identity***:

$$H^2 = H_0^2\left\{\Omega_0\frac{\rho}{\rho_0} - (\Omega_0 - 1)\frac{a_0^2}{a^2}\right\} = H_0^2\left\{\Omega_0\frac{\rho}{\rho_0} - (\Omega_0 - 1)(1+z)^2\right\}. \qquad (4.147)$$

But we have already derived a formula for $\rho_{w_*}(z)$, whence

$$H_{w_*}^2(z) = H_0^2\left\{\Omega_0(1+z)^{3(1+w_*)} - (\Omega_0 - 1)(1+z)^2\right\}, \qquad (4.148)$$

which we can recast as

$$H_{w_*}(z) = H_0(1+z)\sqrt{1 - \Omega_0 + \Omega_0(1+z)^{3w_*+1}}. \qquad (4.149)$$

For realistic matter, satisfying some constraint $w_- \leq w(z) \leq w_+$, we then deduce

$$H_{w_-}(z) \leq H(z) \leq H_{w_+}(z); \qquad (z > 0); \qquad (4.150)$$

$$H_{w_+}(z) \leq H(z) \leq H_{w_-}(z); \qquad (-1 < z < 0). \qquad (4.151)$$

Note that the Hubble bound ceases to provide useful information once the argument of the square root occurring in $H_{w_*}(z)$ becomes negative.

- For $\Omega_0 \leq 1$ there is no limitation in the physical region $z \in (-1, \infty)$.

- For $\Omega_0 > 1$ this limitation manifests itself at $z_H(w_*, \Omega_0) = z_\Omega(w_*, \Omega_0)$, the same place that the bound on $\Omega(z)$ ran into difficulties. (Some numerical estimates of where the bounds fail, based on current consensus observational data, are discussed in [90].)

Finally, suppose that we do not have precise information regarding H_0 and Ω_0, and only have the more limited information

$$H_0 \in [H_{0_-}, H_{0_+}], \qquad \Omega_0 \in [\Omega_{0_-}, \Omega_{0_+}], \tag{4.152}$$

then these two Hubble bounds further generalize to

$$\begin{aligned} H_{0_-}(1+z)\sqrt{1-\Omega_{0_-}+\Omega_{0_-}(1+z)^{3w_-+1}} &\le H(z) \\ &\le H_{0_+}(1+z)\sqrt{1-\Omega_{0_+}+\Omega_{0_+}(1+z)^{3w_++1}}; \quad (z>0); \end{aligned} \tag{4.153}$$

$$\begin{aligned} H_{0_-}(1+z)\sqrt{1-\Omega_{0_+}+\Omega_{0_+}(1+z)^{3w_++1}} &\le H(z) \\ &\le H_{0_+}(1+z)\sqrt{1-\Omega_{0_-}+\Omega_{0_-}(1+z)^{3w_-+1}}; \qquad (-1<z<0); \end{aligned} \tag{4.154}$$

subject to the caveat that for $\Omega_0 > 1$ we should not push the bound past $z_H(w_*, \Omega_0)$.

4.12 General bounds and distance scales

Let us for the time being focus on Peebles' definition of ***angular diameter distance***. This is what Weinberg calls the ***proper motion distance*** [21, 20], for more definitions and a discussion regarding the physical interpretation of the cosmological distance scales see [26], see also [86, 85]. We make this choice to minimize the number of factors of $1+z$ in subsequent formulae. Then the standard definition is

$$d_P = a_0 \ \sin_k\left(\frac{c}{a_0 H_0}\int \frac{H_0}{H(z)}\mathrm{d}z\right). \tag{4.155}$$

But since

$$\frac{c}{a_0 H_0} = \sqrt{k(\Omega_0 - 1)}, \tag{4.156}$$

this can be rewritten more suggestively as

$$d_P = \frac{c}{H_0}\frac{1}{\sqrt{1-\Omega_0}}\sinh\left(\sqrt{1-\Omega_0}\int \frac{H_0}{H(z)}\mathrm{d}z\right). \tag{4.157}$$

When interpreting this last formula for $\Omega_0 > 1$ we make use of the fact that $\sinh(i\Theta) = i\sin(\Theta)$. Substituting $H(z) \to H_{w_*}(z)$ and performing the integral, after considerable effort both Mathematica and Maple yield

$$\int \frac{H_0}{H_{w_*}(z)}\mathrm{d}z = \frac{2}{\sqrt{1-\Omega_0}\,(3w_*+1)}\ln\left\{\frac{(\sqrt{1-\Omega_0}+1)\,(1+z)^{(3w_*+1)/2}}{\sqrt{1-\Omega_0}+\sqrt{1-\Omega_0+\Omega_0(1+z)^{(3w_*+1)}}}\right\}, \tag{4.158}$$

whence

$$d_{P_{w_*}}(z) = \frac{c}{2H_0\sqrt{1-\Omega_0}}\left[\left\{\frac{(\sqrt{1-\Omega_0}+1)\,(1+z)^{(3w_*+1)/2}}{\sqrt{1-\Omega_0}+\sqrt{1-\Omega_0+\Omega_0(1+z)^{(3w_*+1)}}}\right\}^{2/(3w_*+1)} - \left\{\frac{(\sqrt{1-\Omega_0}+1)\,(1+z)^{(3w_*+1)/2}}{\sqrt{1-\Omega_0}+\sqrt{1-\Omega_0+\Omega_0(1+z)^{(3w_*+1)}}}\right\}^{-2/(3w_*+1)}\right]. \tag{4.159}$$

This simplifies slightly

$$d_{P_{w_*}}(z) = \frac{c}{2H_0\sqrt{1-\Omega_0}}\left[(1+z)\left\{\frac{(\sqrt{1-\Omega_0}+1)}{\sqrt{1-\Omega_0}+\sqrt{1-\Omega_0+\Omega_0(1+z)^{(3w_*+1)}}}\right\}^{2/(3w_*+1)} - (1+z)^{-1}\left\{\frac{(\sqrt{1-\Omega_0}+1)}{\sqrt{1-\Omega_0}+\sqrt{1-\Omega_0+\Omega_0(1+z)^{(3w_*+1)}}}\right\}^{-2/(3w_*+1)}\right]. \tag{4.160}$$

We now note

$$\frac{(\sqrt{1-\Omega_0}+1)}{\sqrt{1-\Omega_0}+\sqrt{1-\Omega_0+\Omega_0(1+z)^{(3w_*+1)}}} = \frac{\sqrt{1-\Omega_0+\Omega_0(1+z)^{(3w_*+1)}}-\sqrt{1-\Omega_0}}{(1-\sqrt{1-\Omega_0})\,(1+z)^{3w_*+1}}, \tag{4.161}$$

(cross multiply top and bottom), which finally permits us to write the most tractable form of our ***exact*** result for Peebles' angular diameter distance (in a constant $w(z) = w_*$ FLRW universe):

$$d_{P_{w_*}}(z) = \frac{c}{2H_0\sqrt{1-\Omega_0}\,(1+z)}\left[\left\{\frac{\sqrt{1-\Omega_0+\Omega_0(1+z)^{(3w_*+1)}}-\sqrt{1-\Omega_0}}{(1-\sqrt{1-\Omega_0})}\right\}^{2/(3w_*+1)} - \left\{\frac{\sqrt{1-\Omega_0+\Omega_0(1+z)^{(3w_*+1)}}+\sqrt{1-\Omega_0}}{(1+\sqrt{1-\Omega_0})}\right\}^{2/(3w_*+1)}\right]. \tag{4.162}$$

In this final expression we are always raising quantities to the same power, and the difference between the two terms is just in the placement of $+$ and $-$ signs. (Note that this expression is guaranteed to be real whatever the value of Ω_0; for $\Omega_0 > 1$ the two terms are complex conjugates of each other and after taking the pre-factor $\sqrt{1-\Omega_0}$ into account, the overall combination is guaranteed to be real.)

Note that once we have an explicit formula for the (Peebles) angular diameter distance d_P, any of the other standard cosmological distances can easily be obtained by multiplying by suitable powers of $(1+z)$ [21, 20, 26], see also [86, 85]. In particular the luminosity

distance is

$$d_{L_{w_*}}(z) = \frac{c}{2H_0\sqrt{1-\Omega_0}}\left[\left\{\frac{\sqrt{1-\Omega_0+\Omega_0(1+z)^{(3w_*+1)}}-\sqrt{1-\Omega_0}}{(1-\sqrt{1-\Omega_0})}\right\}^{2/(3w_*+1)} - \left\{\frac{\sqrt{1-\Omega_0+\Omega_0(1+z)^{(3w_*+1)}}+\sqrt{1-\Omega_0}}{(1+\sqrt{1-\Omega_0})}\right\}^{2/(3w_*+1)}\right]. \tag{4.163}$$

Returning to Peebles' angular diameter distance, the Taylor series expansion in z can be computed as

$$d_{P_{w_*}}(z) = \frac{c}{H_0}\left\{z - \frac{2+\Omega_0+3w_*\Omega_0}{4}\,z^2 + \frac{4+\Omega_0^2+w_*(2\Omega_0+6\Omega_0^2)+w_*^2(-6\Omega_0+9\Omega_0^2)}{8}\,z^3+\mathcal{O}(z^4)\right\}. \tag{4.164}$$

Perhaps of more interest is the Taylor series expansion in Ω_0 (since observationally we have good reasons for expecting $\Omega_0 \approx 1$). The leading term is easy to calculate

$$d_{P_{w_*}}(z) = \frac{2c}{H_0(3w_*+1)}\left\{1-(1+z)^{-(3w_*+1)/2}\right\} + \mathcal{O}[\Omega_0-1]. \tag{4.165}$$

Extracting the next $\mathcal{O}[\Omega_0-1]$ term is not too difficult, but is somewhat tedious

$$\begin{aligned} d_{P_{w_*}}(z) = {} & \frac{2c}{H_0(3w_*+1)}\left\{1-(1+z)^{-(3w_*+1)/2}\right\} \\ & -\frac{[\Omega_0-1]\,c}{H_0}\left\{\left[\frac{1-(1+z)^{-(3w_*+1)/2}}{(3w_*+1)}-\frac{1-(1+z)^{3(3w_*+1)/2}}{3(3w_*+1)}\right]\right. \\ & \left.-\frac{1}{6}\left[\frac{1-(1+z)^{-(3w_*+1)/2}}{(3w_*+1)/2}\right]^3\right\} \\ & +\mathcal{O}\left([\Omega_0-1]^2\right). \end{aligned} \tag{4.166}$$

In any realistic situation (provided you accept the standard consensus cosmology) the uncertainties in w will completely dwarf any possible effect due to uncertainties in Ω_0, so carrying the expansion to higher order is not warranted.

As usual, $d_{P_{w_*}}(z)$ [or $d_{L_{w_*}}(z)$] can be used to bound $d_P(z)$ [or $d_L(z)$]. Specifically, let w lie in the range $[w_-, w_+]$ then independent of $\Omega_0 \lesseqgtr 1$:

- $d_{P_{w_+}}(z) \leq d_P(z) \leq d_{P_{w_-}}(z)$; $\qquad (z>0)$,
- $d_{P_{w_-}}(z) \leq d_P(z) \leq d_{P_{w_+}}(z)$; $\qquad (-1<z<0)$.

Here $d_{P_{w_\pm}}(z)$ is given by the rather formidable equation (4.162).

4.13 General bounds and the Lookback time $\mathbf{T(z)}$

Finally, consider the ***lookback time*** defined by:

$$T(z) = \int_a^{a_0} \mathrm{d}t = \int \frac{\mathrm{d}t}{\mathrm{d}a}\mathrm{d}a = \int \frac{a}{\dot{a}}\frac{\mathrm{d}a}{a} = \int \frac{1}{H}\frac{\mathrm{d}(a_0/(1+z))}{a_0/(1+z)} = -\int \frac{1}{H}\frac{\mathrm{d}z/(1+z)^2}{1/(1+z)}. \quad (4.167)$$

That is:

$$T(z) = \int_0^z \frac{1}{(1+z)\,H(z)}\,\mathrm{d}z. \quad (4.168)$$

Using the known form of $H_{w_*}(z)$ we define

$$T_{w_*}(z) \equiv \frac{1}{H_0}\int_0^z \frac{1}{(1+z)^2\,\sqrt{1+\Omega_0\left((1+z)^{3w_*+1}-1\right)}}\,\mathrm{d}z, \quad (4.169)$$

and shall use this quantity to place bounds on the actual lookback time $T(z)$.

It is easy to obtain the leading term for $\Omega_0 \approx 1$:

$$T_{w_*}(z) = \frac{2}{3\,H_0\,(1+w_*)}\left\{1-(1+z)^{-(3w_*-1)/2}\right\} + \mathcal{O}[\Omega_0 - 1]. \quad (4.170)$$

The next sub-leading term again is trickier. We eventually obtain

$$\begin{aligned} T_{w_*}(z) \;=\; & \frac{2}{3\,H_0\,(1+w_*)}\left\{1-(1+z)^{-(3w_*-1)/2}\right\} \\ & -\frac{[\Omega_0-1]}{H_0}\left[\frac{1-(1+z)^{-3(w_*+1)/2}}{3(w_*+1)} - \frac{1-(1+z)^{-(9w_*+5)/2}}{9w_*+5}\right] \\ & +\mathcal{O}\left([\Omega_0-1]^2\right). \end{aligned} \quad (4.171)$$

Again, in any realistic situation (provided you accept the standard consensus cosmology) the uncertainties in w will completely dwarf any possible effect due to uncertainties in Ω_0. Exact integration and subsequent evaluation of the result for $T_{w_*}(z)$ can only be performed in terms of hypergeometric functions. Let us first be a little more careful about the use of the dummy variable in the integration and write

$$T_{w_*}(z) = \frac{1}{H_0\,\sqrt{\Omega_0}}\int_0^z \frac{1}{(1+\tilde{z})^{2+1/2(3w_*+1)}\,\sqrt{1-(1-\Omega_0^{-1})\,(1+\tilde{z})^{-(3w_*+1)}}}\,\mathrm{d}\tilde{z}, \quad (4.172)$$

and then, (following the procedure of [64, 90]), apply the binomial theorem

$$\left[1-(1-\Omega_0^{-1})\,(1+\tilde{z})^{-(3w_*+1)}\right]^{-1/2} = \sum_{n=0}^{\infty}\binom{-1/2}{n}(-1)^n\,(1-\Omega_0^{-1})^n\,(1+\tilde{z})^{-(3w_*+1)n}. \quad (4.173)$$

Now this particular binomial series will converge provided [8]

$$\left|(1-\Omega_0^{-1})\,(1+\tilde{z})^{-(3w_*+1)}\right| < 1. \quad (4.174)$$

[8] Note that for $z<0$ one is actually calculating the ***lookforward time*** — the time until the universe expands by an additional factor of $\frac{1}{1-|z|}$ in each direction.

That is, provided

$$\left|1-\Omega_0^{-1}\right| < (1+\tilde{z})^{3w_*+1}. \tag{4.175}$$

More explicitly, the ***integral*** will make sense provided

$$\left|\frac{1-\Omega_0}{\Omega_0}\right| < (1+\tilde{z})^{3w_*+1}; \qquad \forall\ \tilde{z}\in(0,z) \text{ or } \tilde{z}\in(z,0), \tag{4.176}$$

which is equivalent to

$$\left|\frac{1-\Omega_0}{\Omega_0}\right| < \min\{1,(1+z)^{3w_*+1}\}. \tag{4.177}$$

- In all cases, to ensure convergence at redshift zero, we must certainly have

$$\left|1-\Omega_0^{-1}\right| < 1, \qquad \text{that is} \qquad \Omega_0\in(1/2,\infty). \tag{4.178}$$

- If $z>0$ and $(3w_*+1)\geqslant 0$, $(w_*\geqslant -1/3)$, or if $z<0$ and $(3w_*+1)\leqslant 0$, $(w_*\leqslant -1/3)$: Then $(1+z)^{(3w_*+1)}\geq 1$, and no additional limitation is imposed.

- If $z>0$ and $(3w_*+1)<0$, $(w_*<-1/3)$, or if $z<0$ and $(3w_*+1)>0$, $(w_*>-1/3)$: In this situation $(1+z)^{(3w_*+1)}<1$, therefore we now obtain an additional limitation on z that is necessary to ensure convergence:

  - If $z>0$, then we need

$$z < \left|\frac{\Omega_0-1}{\Omega_0}\right|^{-1/|3w_*+1|} - 1 > 0. \tag{4.179}$$

  - If $z<0$ then we need

$$z > \left|\frac{1-\Omega_0}{\Omega_0}\right|^{1/(3w_*+1)} - 1 < 0. \tag{4.180}$$

- In view of equation (4.137) these last conditions can also be interpreted as constraints on the Ω parameter at the redshift one wishes to probe:

$$\left|1-\Omega_{w_*}(z)^{-1}\right| < 1, \qquad \text{that is} \qquad \Omega_{w_*}(z)\in(1/2,\infty). \tag{4.181}$$

Subject to this convergence condition we can integrate term by term, and obtain the convergent series

$$T_{w_*}(z) = \frac{1}{H_0\sqrt{\Omega_0}}\sum_{n=0}^{\infty}\binom{-1/2}{n}(-1)^n\frac{(1-\Omega_0^{-1})^n\ \left[1-(1+z)^{-(3w_*+1)n-3/2(w_*+1)}\right]}{(3w_*+1)n+3/2(w_*+1)}. \tag{4.182}$$

As a practical matter, for many purposes this series representation may be enough, but we can tidy things up somewhat by first defining

$$S_{w_*}(x) = \sum_{n=0}^{\infty}\binom{-1/2}{n}\frac{(-x)^n}{(3w_*+1)n+3/2(w_*+1)}, \tag{4.183}$$

in which case

$$T_{w_*}(z) = \frac{1}{H_0\,\sqrt{\Omega_0}}\left\{S_{w_*}\left(1-\Omega_0^{-1}\right) - (1+z)^{-3/2(w_*+1)}\; S_{w_*}\left(\frac{(1-\Omega_0^{-1})}{(1+z)^{3w_*+1}}\right)\right\}. \qquad (4.184)$$

Finally we can recognize that $S_{w_*}(x)$ is itself a particular example of a hypergeometric function, [9] and so we can write

$$S_{w_*}(x) \;=\; \frac{1}{3/2(w_*+1)}\;{}_2F_1\left(\frac{1}{2},\frac{3}{2}\left[\frac{w_*+1}{3w_*+1}\right];\frac{1}{2}\left[\frac{9w_*+5}{3w_*+1}\right];x\right). \qquad (4.185)$$

Therefore

$$\begin{aligned} T_{w_*}(z) \;\equiv\; & \frac{1}{3/2(w_*+1)\,H_0\,\sqrt{\Omega_0}} \times \left\{{}_2F_1\left(\frac{1}{2},\frac{3}{2}\left[\frac{w_*+1}{3w_*+1}\right];\frac{1}{2}\left[\frac{9w_*+5}{3w_*+1}\right];1-\Omega_0^{-1}\right)\right. \\ & \left. -(1+z)^{-3/2(w_*+1)}\;{}_2F_1\left(\frac{1}{2},\frac{3}{2}\left[\frac{w_*+1}{3w_*+1}\right];\frac{1}{2}\left[\frac{9w_*+5}{3w_*+1}\right];\frac{1-\Omega_0^{-1}}{(1+z)^{3w_*+1}}\right)\right\}. \end{aligned} \qquad (4.186)$$

As usual, $T_{w_*}(z)$ can be used to bound $T(z)$. Specifically, let $w(z)$ lie in the bounded range $w(z) \in [w_-, w_+]$, then independent of $\Omega_0 \lesseqgtr 1$:

- $T_{w_+}(z) \leq T(z) \leq T_{w_-}(z)$; $\qquad (z > 0)$,
- $T_{w_-}(z) \leq T(z) \leq T_{w_+}(z)$; $\qquad (-1 < z < 0)$.

Here $T_{w_\pm}(z)$ is given by the rather formidable equation (4.186), based on the use of hypergeometric functions.

4.14 Special cases and consistency checks

Useful special cases, and consistency checks we can perform on the formalism, include:

Dust: For pure dust, $w_+ = w_- = 0$, we have simple exact results

$$H_{\rm dust}(z) = H_0(1+z)\sqrt{1+\Omega_0 z}. \qquad (4.187)$$

$$\Omega_{\rm dust}(z) = \frac{\Omega_0\,(1+z)}{1+\Omega_0\,z}. \qquad (4.188)$$

$$\rho_{\rm dust}(z) = \rho_0\,(1+z)^3\,. \qquad (4.189)$$

[9] The classical hypergeometric series is given by

$${}_2F_1\,(a,b;c;x) = \sum_{n=0}^{\infty}\frac{(a)_n(b)_n}{(c)_n}\;\frac{x^n}{n!},$$

where $(a)_n = a(a+1)(a+2)(a+n-1)$ is the rising factorial, or Pochhammer symbol. This series is convergent for $|x| < 1$.

$$d_{P_{\rm dust}}(z) = \frac{2c}{H_0}\left\{\frac{(\sqrt{1+\Omega_0 z}-1)(\sqrt{1+\Omega_0 z}-1+\Omega_0)}{(1+z)\Omega_0^2}\right\}. \tag{4.190}$$

$$\begin{aligned} T_{\rm dust}(z) &= \frac{1}{H_0(1-\Omega_0)}\left\{1-\frac{\sqrt{1+\Omega_0 z}}{1+z}\right\} \\ &+ \frac{\Omega_0}{H_0(1-\Omega_0)^{3/2}}\left\{\tanh^{-1}\frac{\sqrt{1+\Omega_0 z}}{\sqrt{1-\Omega_0}} - \tanh^{-1}\frac{1}{\sqrt{1-\Omega_0}}\right\}. \end{aligned} \tag{4.191}$$

The only one of these equations for which the $\Omega_0 \to 1$ limit is even remotely subtle is the lookback time, for which

$$T_{\rm dust,\,\Omega_0=1}(z) = \frac{2}{3H_0}\left\{1-\frac{1}{(1+z)^{3/2}}\right\}. \tag{4.192}$$

Radiation: For pure radiation, $w_+ = w_- = 1/3$, we have

$$H_{\rm radiation}(z) = H_0(1+z)\sqrt{1+\Omega_0[(1+z)^2-1]}. \tag{4.193}$$

$$\Omega_{\rm radiation}(z) = \frac{\Omega_0\,(1+z)^2}{1-\Omega_0+\Omega_0\,(1+z)^2}. \tag{4.194}$$

$$\rho_{\rm radiation}(z) = \rho_0\,(1+z)^4\,. \tag{4.195}$$

$$d_{P_{\rm radiation}}(z) = \frac{c}{H_0}\left\{\frac{\sqrt{1+\Omega_0[(1+z)^2-1]}-1}{(1+z)\Omega_0}\right\}. \tag{4.196}$$

$$T_{\rm radiation}(z) = \frac{1}{H_0(1-\Omega_0)}\left[1-\frac{\sqrt{1+\Omega_0((1+z)^2-1)}}{1+z}\right]. \tag{4.197}$$

The only one of these equations for which the $\Omega_0 \to 1$ limit is even remotely subtle is the lookback time, for which

$$T_{\rm radiation,\,\Omega_0=1}(z) = \frac{1}{2H_0}\left\{1-\frac{1}{(1+z)^2}\right\}. \tag{4.198}$$

Cosmological constant: For pure cosmological constant $w_+ = w_- = -1$. We then obtain (now as equalities rather than inequalities) what would for the NEC have been a set of bounds, such as those presented in [63, 64, 65, 90] . (That is, a nonzero cosmological constant is right on the verge of violating the NEC.)

Furthermore, comparing with previous results in this chapter with bounds with the energy conditions:

- For $w_- = -1/3$ one has
$$H(z) \geq H_0\,(1+z). \tag{4.199}$$
This reproduces the SEC lower bound previously investigated in [63, 64, 65, 90].

- For $w_- = -1$ one has

$$H(z) \geq H_0\ (1+z)\sqrt{(1+z)^{-2} + [\Omega_0 - 1]\ [(1+z)^{-2} - 1]}, \tag{4.200}$$

whence

$$H(z) \geq H_0\ \sqrt{1 + [\Omega_0 - 1]\ [1 - (1+z)^2]} = H_0\ \sqrt{\Omega_0 + [1 - \Omega_0]\ (1+z)^2}. \tag{4.201}$$

This reproduces the NEC lower bound previously investigated in [63, 64, 65, 90].

- For $w_+ = +1$ we have

$$H(z) \leq H_0\ (1+z)\sqrt{(1+z)^4 + [\Omega_0 - 1]\ [(1+z)^4 - 1]}, \tag{4.202}$$

that is

$$H(z) \leq H_0\ (1+z)\sqrt{1 + \Omega_0\ [(1+z)^4 - 1]}. \tag{4.203}$$

This reproduces the DEC upper bound previously investigated in [63, 64, 65, 90].

4.15 Conclusions

In this chapter we have extended and generalized the discussion of the original articles [63, 64, 65], and more recently of [71, 72, 73], to develop a number of rugged and general energy-condition-induced bounds on various cosmological parameters, bounds which have all taken the form

$$X(z) \gtrless X_{\text{bound}} \equiv X_0\ f(\Omega_0, z), \tag{4.204}$$

where $X(z)$ is some cosmological parameter, X_0 is its present-day value, and $f(\Omega_0, z)$ is some dimensionless function depending on the particular bound under consideration. The bounds we have considered can be derived by ***elementary*** means, and are typically expressed in terms of polynomial, rational, algebraic, and elementary functions — though in one particular instance we had to resort to hypergeometric functions. Several of these bounds are completely new [such as the explicit bounds on $H(z)$ and $\Omega(z)$, and the physically important Taylor series expansions for $\Omega_0 \approx 1$], several are significant extensions of previously known partial results [see especially $d_{P_{\text{NEC}}}$ (4.40), $d_{P_{\text{DEC}}}$ (4.57), T_{DEC} (4.85)], and all of these bounds are now valid for arbitrary spatial curvature [see especially the explicit bounds on $\rho(z)$ for $k \neq 0$]. Additionally, since the analysis is now systematic and exhaustive, it is clear how the various energy conditions and their associated bounds are inter-related.

Furthermore, in the absence of any detailed understanding of the precise nature of the cosmological equation of state $\rho(p)$ it is useful to examine the question of just how much can be deduced with limited information. In the second part of this chapter we have also worked in terms of the w-parameter $w(z) = p/\rho$, and we have used the idealized case of constant w_* as a "***template***" for comparison purposes with more realistic $w(z)$. Specifically:

- For constant w_* the explicit results for the density $\rho_{w_*}(z)$ and Hubble parameter $H_{w_*}(z)$ are well-known. The explicit result for the Ω parameter $\Omega_{w_*}(z)$ is less well-known, and the explicit results we have obtained for the angular diameter distance $d_{P_{w_*}}(z)$ and lookback time $T_{w_*}(z)$ appear to be both novel and significant.

- More importantly we have seen that these idealized results for constant w_* can be used as the basis for general comparison results that bound the various features of the Hubble flow in the following sense: If we know that $w(z) \in [w_-, w_+]$ between redshift zero and redshift z, then for monotonically evolving generic cosmological quantities $X(z)$ we have derived a number of rigorous bounds of the form

$$X_{w_\pm}(z) \leq X(z) \leq X_{w_\mp}(z), \tag{4.205}$$

  where we have explicitly seen that the direction of the inequality depends both on the precise details of the evolution of $X(z)$, and on the redshift range of interest.

Finally we point out that all of our bounds have been explicitly calculated for ***all*** signs of the spatial curvature $k \in [-1, 0, +1]$, that is for all Ω_0 (though we have restricted ourselves to the physically very plausible $\Omega_0 > 0$). We have briefly sketched how to use these bounds and the energy conditions to confront the supernova data, but have not yet performed any detailed analysis of this point.

The bounds presented in this chapter may appear to be valid for a ***single*** fluid component. However, they can also be used for ***multi***-fluid components as follows. Consider that the total density ρ is given by the linear combination

$$\rho = \sum_i^n \rho_i, \tag{4.206}$$

and that the total pressure is given by the linear combination

$$p = \sum_i^n p_i, \tag{4.207}$$

with the equation of state $p_i = w_i \rho_i$ for each value of i. If we can determine the biggest and smallest of the w_i-values, that is $w_+ = \max_i(w_i)$ and $w_- = \min_i(w_i)$, we can then sum over the number of fluid components (as long as $\rho_i \geq 0$) and obtain the relation

$$p \lesseqgtr w_\pm \, \rho. \tag{4.208}$$

The bounds we have derived will then apply in this case. Thus all the bounds we have derived are both very general and very powerful.

Part II

Numerical Relativity

Nomenclature

We present here all the various notations we use throughout the numerical relativity part of this thesis. An attempt has been made to keep the basic notation as standard as possible, however, the following list will hopefully help clarify any potential ambiguities.

Greek Letters

$\alpha, \beta, ...$	Index for 4D space-time dimensions

Latin Letters

$a, b, ...$	Index for 3D space dimensions

Subscripts and Superscripts

a, b, c	Summation indices for the test functions and local nodes
$\mathbf{k}$	Superscript index for the $\mathbf{k}$-th element
h	Subscript index used to represent discrete quantities, e.g u_h
m, n, p	Summation indices for the unknown variables
q, r, s	Summation indices for the GLL (Gauss–Lobatto–Legendre) quadrature

Solution domains

$\mathbf{\Omega}$	Solution domain
$\mathbf{\Omega}_h$	Discrete solution domain
$\mathbf{\Omega^k}$	Domain of the $\mathbf{k}$-th element
$\mathbf{\Gamma}$	Boundary of the domain $\mathbf{\Omega}$
$\mathbf{\Gamma}_h$	Discrete boundary of the domain $\mathbf{\Omega}$

Various constants

N	Polynomial order for the spectral elements
N_E	Number of elements dividing the domain $\mathbf{\Omega}$
N_g	Total number of points in the domain $\mathbf{\Omega}$
N_{GLL}	Number of GLL (Gauss–Lobatto–Legendre) points per element $N_{GLL} = N+1$

Variables

u	Variable used for the wave equation
g_{ij}	Physical metric of the BSSN system
ψ	Conformal factor of the BSSN system
$\tilde{g}_{ij}$	Conformal metric of the BSSN system
ϕ	Variable of the BSSN system for the ϕ-method, $\phi = \ln \psi$
χ	Variable of the BSSN system for the χ-method, $\chi = \psi^{-4}$
A_{ij}	Extrinsic curvature of the BSSN system
$\tilde{A}_{ij}$	Conformal extrinsic curvature of the BSSN system

K	Trace of the extrinsic curvature of the BSSN system
$\tilde{\Gamma}^i$	Auxiliary variable of the BSSN system
Γ^i_{jk}	Connection of the physical metric of the BSSN system
$\tilde{\Gamma}^i_{jk}$	Conformal connection of the BSSN system
α	Lapse function of the BSSN system
β^i	Shift function of the BSSN system

Expansion basis notation

$h_i^n(x)$	Lagrange-Legendre polynomial basis functions of order n used in the SEM
u_{abc}	Expansion coefficients
x, y, z	Global Cartesian coordinates (physical coordinates)
ξ, η, ζ	Local Cartesian coordinates (computational coordinates in the master element)
$L_n(x)$	Legendre polynomials of order n
H_{ij}	First differentiation elemental matrix of the Lagrange-Legendre basis functions
$J^{\mathbf{k}}$	Jacobian of the elemental mapping for on the element $\mathbf{\Omega^k}$
$\mathcal{I}(a, b, c, \mathbf{k})$	Global numbering function that maps the local numbering of the computational nodes to their global numbering

Spaces

$\mathcal{L}^2(\mathbf{\Omega})$	Lebesgue space
$\mathcal{H}^1(\mathbf{\Omega})$	Sobolev space
$\mathcal{U}, \mathcal{W}$	Space of trial (unknowns) and test functions
$\mathcal{U}_h, \mathcal{W}_h$	Discrete space of trial (unknowns) and test functions
$\mathbb{P}_N(\mathbf{\Omega^k})$	Space of polynomials of degree less that or equal to N on the element $\mathbf{\Omega^k}$

Elemental matrices

$\mathbf{M^k}$	Elemental mass matrix for the **k**-th element
$\mathbf{A}_i^{\mathbf{k}}$	Elemental advection matrix type 1 for the **k**-th element, with respect to the i-th coordinate
$\mathbf{A}_{ij}^{\mathbf{k}}$	Elemental second derivative matrix type 1, for the **k**-th element, with respect to the i-th and j-th coordinates
$\mathbf{D}_i^{\mathbf{k}}$	Elemental advection matrix type 2 for the **k**-th element, for the i-th coordinate
$\mathbf{D}_{ij}^{\mathbf{k}}$	Elemental second derivative matrix type 2, for the **k**-th element, with respect to the i-th and j-th coordinates
$\mathbf{K}_{ii}^{\mathbf{k}}$	Elemental stiffness matrix, for the **k**-th element, with respect to the i-th coordinate
$\mathbf{B^k}$	Elemental boundary matrix for the **k**-th element
$(\mathbf{\Lambda}^a_{bc})^{\mathbf{k}}$	Elemental Christoffel symbol matrix for the **k**-th element
$\mathbf{F^k}$	Elemental force vector for the **k**-th element
$\mathbb{D}_{ij}^{\mathbf{k}}$	Elemental second covariant derivative matrix, for the **k**-th element, with respect to the i-th and j-th coordinates

$(\mathbb{R}_{ij})^{\mathbf{k}}$	Elemental Ricci tensor matrix, for the **k**-th element
$(\mathbb{X}_{ij}^{TF})^{\mathbf{k}}$	Elemental X_{ij}^{TF} matrix term, for the **k**-th element

Global matrices

$\mathbf{M}$	Mass matrix
$\mathbf{A}_i$	Advection matrix type 1 with respect to the i-th coordinate
$\mathbf{A}_{ij}$	Second derivative matrix type 1 with respect to the i-th and j-th coordinates
$\mathbf{D}_i$	Advection matrix type 2 with respect to the i-th coordinate
$\mathbf{D}_{ij}$	Second derivative matrix type 2 with respect to the i-th and j-th coordinates
$\mathbf{K}_{ii}$	Stiffness matrix with respect to the i-th coordinate
$\mathbf{B}$	Boundary matrix
$\mathbf{\Lambda}_{bc}^{a}$	Christoffel symbol matrix
$\mathbf{F}$	Force vector
$\mathbb{D}_{ij}$	Second covariant derivative matrix with respect to the i-th and j-th coordinates
$\mathbb{R}_{ij}$	Ricci tensor matrix
$\mathbb{X}_{ij}^{TF}$	Matrix term X_{ij}^{TF}, (Trace-free part of X_{ij})

Operators

∇^2	Laplacian in 3D space
$\mathcal{L}_\beta$	Lie derivative in 3D space with respect to the shift β
D	Covariant or contravariant derivative in 3D space
$\tilde{D}$	Conformal covariant or contravariant derivative in 3D space
:	Scalar matrix product or Hadamard matrix product with $(A : B)_{ijl} = a_{ijl}\, b_{ijl}$
$\otimes$	Matrix multiplication operator used in the SEM notation
$\cdot_{xy}$	3D Matrix multiplication operator, regular matrix product in the xy direction for each z dimension $(A \cdot_{xy} B)_{abc} = \sum_i A_{aic} B_{ib}$
$\cdot_{yz}$	3D Matrix multiplication operator, regular matrix product in the yz direction for each x dimension $(A \cdot_{yz} B)_{abc} = \sum_i A_{abi} B_{ic}$

Acronyms

SEM	Spectral Element Method
FEM	Finite Element Method
FD	Finite Difference Method
SM	Spectral Method
BSSN	Shibata-Nakamura-Baumgarte-Shapiro formulation of Einstein equations
GLL	Gauss–Lobatto-Legendre

"That's not right.
That's not even wrong."

Wolfgang Pauli (1900–1958)

5

Introduction to Numerical Relativity

One of the predictions of Einstein's theory of gravity is the existence of ***gravitational waves***. Perturbations of spacetime propagate as waves at the speed of light. Any accelerating object will produce gravitational waves but only astrophysical objects produce enough gravitational wave energy to be detected. One of the strongest predicted sources is the merger of two black holes. However, gravitational waves are weak and have not yet been directly detected. The only gravitational wave sources that we expect to be strong enough to be detected in the near future are astrophysical ones. One of the main goal of numerical relativity is to calculate gravitational wave forms from promising astrophysical sources, in order to provide theoretical templates both for the new ground-based gravitational wave laser interferometers like LIGO in the US, VIRGO in Italy, GEO in Germany, TAMA in Japan, and the proposed AIGO detector in Australia, as well as for space-based interferometers such as LISA.

The first attempts to simulate black holes were in 1964. For decades, the quest for numerically stable black-hole inspiral simulations has been very challenging. Many questions had to be addressed:

- How do we represent black-hole singularities in a computer code?
- How do we set up initial conditions for two black holes in orbit?
- What coordinate (gauge) conditions should we use?
- How do we accurately describe tiny black holes and huge waves?
- What if the equations are not numerically stable?

In numerical relativity, we need to set up the Einstein system as a Cauchy problem, that is, a problem with initial data that are evolved in time. Determination of initial data is highly non-trivial due to initial data constraints that have to be solved. In solving them we need to disentangle the gauge and physical degrees of freedom, as well as find solutions that describe the physical system one is interested in (astrophysically realistic conditions). This finally leads on to the question of whether we continue with free or constrained evolution. The tensorial nature of the field equations, the constraints and the coordinate freedom result in a development of multitude of formalisms. Indeed, one must choose the dynamical variables (the quantity advanced in time with evolution equations); one must choose the specific form of the field equations, multiple constraints can be added to the evolution

equations; furthermore, one must make a choice of coordinates or classes of coordinate systems. Constraints and coordinate freedom lead to many options for advancing the discrete solution from one time step to the next.

There are several techniques employed in numerical relativity for evolution problems: ***free evolution***, ***partially constrained evolution*** and ***fully constrained evolution***. In a ***free evolution***, the constraints are solved at the initial time only and then all the dynamical variables are advanced in time using evolution equations. In a ***partially constrained evolution***, some or even all of the constraints are solved at each time step for specific dynamical variables, instead of using the corresponding evolution equations. In a ***fully constrained evolution***, all of the constraints are solved at each time step and all 4 degrees of coordinate freedom are used to eliminate the dynamical variables, leaving exactly 2 dynamical degrees of freedom to be advanced with the evolution equations. Typically, in many problems of interest in numerical relativity, one should expect a large dynamic range, e.g for binary black hole collisions, one must resolve the dynamics on the scale of the black hole, and many wavelengths of characteristic gravitational radiation. Gravitational waves tend to be a small effect, however, they must be computed very accurately for maximal utility in the context of gravitational earth-based detections.

The first pioneering numerical simulation of colliding black holes were implemented by Hahn and Lindquist [91]. There was progress on all of these issues, but full inspiral-merger-ringdown simulations were not possible until 2005 with Pretorius' breakthrough simulation [92] based on a harmonic code. The method of the ***moving punctures*** was to follow developed in parallel by two independent groups [93, 94].

Recent developments in numerical relativity consists of the following topics, see reference [95] for a review: Binary black holes (BBHs); Binary neutron stars (BNSs); Binary black hole-neutron stars (BBHNSs); Rotating relativistic stars; Collisionless clusters; Scalar fields; Critical phenomena; Cosmic censorship; and General relativistic magnetohydrodynamics (GRMHD).

5.1 Einstein's legacy

In 1916, Einstein published his general theory of relativity, gravity appears as curvature of spacetime. General relativity explains many phenomena, amongst them, the perihelion advance of Mercury, and predicts the gravitational deflection of light by the sun, which was verified by Eddington in 1919. There is one more very important prediction: ***gravitational waves***. If we start with flat space,

$$g_{\mu\nu} = \eta_{\mu\nu}, \tag{5.1}$$

and then perturb the metric, we obtain:

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}, \tag{5.2}$$

where $|h_{\mu\nu}| \ll 1$ and

$$\bar{h}_{\mu\nu} = h_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}h. \tag{5.3}$$

The Einstein equations now reduce to,

$$\Box \bar{h}_{\mu\nu} = 16\pi T_{\mu\nu}, \tag{5.4}$$

that is, a ***wave equation*** for the perturbation. Note that in vacuum the Einstein equations further reduce to

$$\Box h_{\mu\nu} = 0. \tag{5.5}$$

Equation (5.4) is crucial as it implies that the perturbations of spacetime propagate as waves at the speed of light. These waves are the ***gravitational waves***. Any accelerating object will produce gravitational waves, however, gravity can be very weak (see Table 5.1 for some illustrative figures). Only astrophysical objects produce enough gravitational wave energy to be detected (one of the strongest predicted sources is the merger of two black holes.).

Table 5.1: Gravitational wave energy comparisons between astrophysical objects

Weak-field:	(radiation power) $\sim$ $\frac{32}{5}\frac{G}{c^5}$(moment of inertia)(frequency)^6
iron bar (1000 t, 100m, 3 Hz)	10^{-26} W
Earth around Sun	200 W
close binary stars	10^{15} – 10^{30} W
close neutron star binary (100 km, 100 Hz)	10^{45} W

Even if gravitational waves are relatively strong, by the time the waves reach Earth, they are extremely weak, hence the need for large detectors, such as LIGO and LISA. At merger of two black holes, the signal becomes an order of magnitude stronger, around 3% of the binarys mass is emitted as gravitational waves. Ground-based detectors have arms from 600 m up to 4 km long (see figure 5.1).

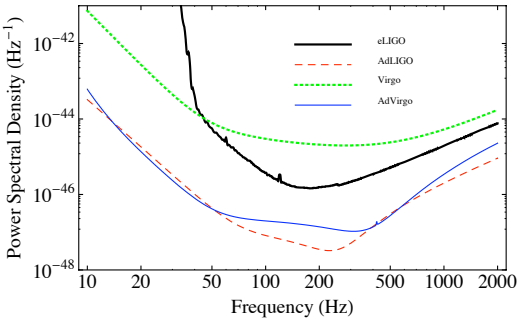
5.2 The 3+1 formalism

General Relativity is based on a covariant approach of describing spacetime geometry. It is not the most intuitive concept from everyday life experience, where we rather experience spacetime as a temporal succession of spatial geometries. The 3+1 formalism is closer to our intuitive experience, and relies on the slicing of the 4-dimensional spacetime by 3-dimensional surfaces (hypersurfaces). These hypersurfaces have to be spacelike, so that the metric induced on them by the Lorentzian spacetime metric $[signature(-,+,+,+)]$ is Riemannian $[signature(+,+,+)]$. This decomposition of Einstein equations can then be formulated as a Cauchy problem with constraints. One manipulates only time-varying tensor fields in the 3-dimensional space, where the standard scalar product is Riemannian.

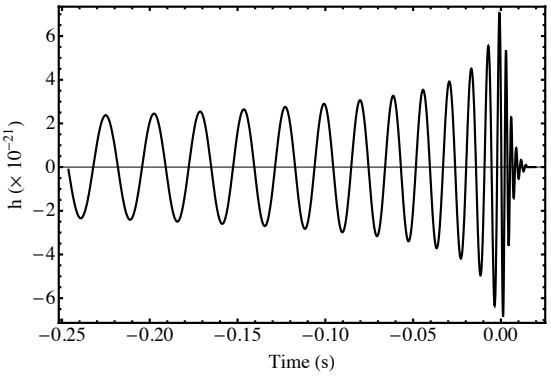
In order to do numerical simulations, one needs to decompose 4-dimensional objects into time and the 3-dimensional space components:

- by selecting a specific time coordinate;

(a) LIGO: 4km detector arms



(b) Theoretical noise curves (power spectral density) for the detectors Enhanced LIGO, Advanced LIGO, Virgo and Advanced Virgo, from [96].



(c) The gravitational-wave strain from an optimally-oriented 60 $M\odot$ equal-mass nonspinning black-hole binary located 100 Mpc away from the detector. The waveform covers about six orbits, or twelve GW cycles, before merger [96].

Figure 5.1: Gravitational wave detector LIGO in Hanford and Livingstone (US) and gravitational waves at merger.

- by decomposing every 4-dimensional object (metric, Ricci and stress-energy tensors) into 3-dimensional components in order to produce a Cauchy problem;
- by writing down the 3-dimensional field equations that translate the covariant ones into terms of the newly defined 3-dimensional objects.

General covariance is still preserved but becomes a hidden feature of the resulting $3+1$ equations. However, these equations themselves will not be covariant under a general coordinate transformation.

Let us introduce a global time function t whose level sets are the hypersurfaces defining the foliation. We then define the 3-dimensional metric g_{ij} for $i,j = 1,2,3)$ that measures distances within a given hypersurface. The ***lapse*** function α measures the proper time between adjacent hypersurfaces, that is, between the slice at time t and the next slice at $t+dt$. The ***shift*** vector β^i measures the relative speed between observers moving along the normal direction to the hypersurfaces, and those keeping constant spatial coordinates, in other words, β^i prescribes how the coordinates shift between the two slices. The four dimensional metric can then be written as:

$$ds^2 = \left(-\alpha^2 + \beta^i\beta_i\right)dt^2 + 2\beta_i dt dx^i + g_{ij}dx^i dx^j, \tag{5.6}$$

where $\beta_i = g_{ij}\beta^j$. The spacetime metric gives the invariant interval between neighboring points A and B on the two slices represented in Figure 5.2.

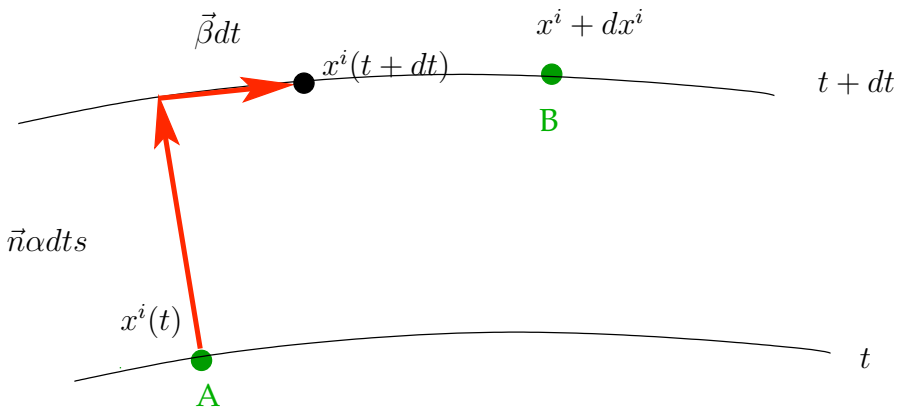


Figure 5.2: Illustration of the $3+1$ ADM decomposition.

To measure how the spatial hypersurfaces are immersed in spacetime, one introduces the ***extrinsic curvature tensor*** K_{ij}, given by the Lie derivative of g_{ij} along the time lines:

$$\partial_t g_{ij} = -2\alpha K_{ij} + D_i\beta_j + D_j\beta_i, \tag{5.7}$$

where D_i is the covariant derivative associated with g_{ij}. The Einstein equations are then spilt into two groups: the ***Hamiltonian*** and ***Momentum constraints*** and the ***evolution equations***. The first group involves no time derivatives and represents constraints that must be verified at all times.

The Hamiltonian constraints: In geometrized units ($G = c = 1$), it is:

$$R + (K)^2 - K_{ij}K^{ij} = 16\pi\rho, \tag{5.8}$$

where R is the scalar curvature of the spatial geometry, $K = \mathrm{tr}K = g^{ij}K_{ij}$ is the trace of the extrinsic curvature, and ρ is the energy density of matter measured by the normal observers.

The Momentum constraints: They are of the form

$$D_j\left(K^{ij} - g^{ij}K\right) = 8\pi J^i, \tag{5.9}$$

where J is the momentum flux of matter measured by the normal observers.

The evolution equations: They include the remaining 6 Einstein equations and contain the dynamics of the system:

$$\begin{aligned}\partial_t K_{ij} = {} & \beta^k D_k K_{ij} + K_{ik}D_j\beta^k + K_{jk}D_i\beta^k - D_iD_j\alpha \\ & +\alpha\left(R_{ij} - 2K_{ik}K^k_{\ j} + K_{ij}\ K\right) \\ & +4\pi\alpha\left[g_{ij}\left(\mathrm{tr}S - \rho\right) - 2S_{ij}\right],\end{aligned} \tag{5.10}$$

where S_{ij} is the stress-energy tensor of matter.

Note that the existence of the constraints implies that the 12 dynamical quantities g_{ij}, K_{ij} cannot be specified as arbitrary initial conditions. The Bianchi identities imply that the evolution equations preserve the constraints, consequently, if they are satisfied initially, they will remain satisfied at subsequent times. Also the $3 + 1$ formalism prescribes no equations whatsoever for the lapse α and the shift β^i. These four functions represent the gauge (coordinate) freedom inherent in general relativity. One of the main challenges of numerical relativity is to choose them appropriately, especially in the presence of black holes. It is very difficult to just specify the lapse and shift as known functions of spacetime, therefore they are often chosen dynamically as functions of the evolving geometry, in that sense, the coordinates are chosen as we go.

Equations (5.8), (5.9) and (5.10) are known as the Arnowitt-Deser-Misner equations (ADM). They represent the starting point of practically all of 3+1 numerical relativity. These equations are derived in more details in the original ADM article [97]. Note that the notation used here follows York's article [98].

The ADM equations are, however, not ideal for direct numerical simulations, because it turns out that the evolution system is only weakly hyperbolic and thus not well posed.

5.3 Hyperbolic systems

The ADM evolution equations previously introduced are highly non-unique. Indeed, arbitrary multiples of the constraints (multiples of zero) can be added to the equations without

affecting the physical solutions. Violating the constraints is inevitable in numerical simulations since truncation errors imply that the constraints are never satisfied exactly. This non-uniqueness of the evolution equations is well known. The original equations of ADM [97] differ from those of York [98] just by the addition of a multiple of the Hamiltonian constraint. It turns out that York's formulation is better behaved mathematically [99] and it has become the standard form used in numerical relativity.

An important key feature when studying the Cauchy problem is the well-posedness of the system of evolution equations: solutions exist (at least locally) and are stable (small changes in the initial data produce small changes in the solution). Hyperbolic systems are well posed under very general conditions [100]. In light of the disadvantages of the ADM system, a large number of 3 +1 hyperbolic formulations of general relativity have been developed (see the recent review article by Reula [101] for an extensive survey and a more complete list of references). Bona and Massó started studying hyperbolic formulations for numerical relativity [102, 103, 104] in the early 1990s. Baumgarte and Shapiro [105] showed that a reformulation of the ADM equations proposed by Nakamura, Oohara and Kojima [30], and Shibata and Nakamura [106], had far superior numerical stability properties than ADM. The ADM system is only weakly hyperbolic, whereas this new reformulation, now known as the BSSN formulation, is strongly hyperbolic [107, 108]. We will focus on the system based on the work by Baumgarte and Shapiro, and Shibata and Nakamura, and the work by Bona and Massó on slicing conditions.

5.4 The BSSN formulation

The key ideas of the BSSN (***Shibata-Nakamura-Baumgarte-Shapiro***) formalism are, to eliminate the mixed second derivatives in the Ricci tensor by introducing some auxiliary variables, and to evolve a conformal factor ψ and K separately in the spirit of the ***spin decomposition*** of geometric quantities. Thereby, the physical metric and extrinsic curvature variables are replaced by a ***conformal metric*** and ***extrinsic curvature***, in a similar fashion as the "York-Lichnerowicz" split [109, 110]:

$$g_{ij} = \psi^4 \tilde{g}_{ij}, \tag{5.11}$$

The variable ψ is the conformal factor used to provide conformally rescaled quantities $\{\psi,\ K,\ \tilde{g}_{ij},\ \tilde{A}_{ij}\}$ for the evolutions equations. Note that the tilde refers to conformal quantities. It proves convenient to split the extrinsic curvature into its trace K and tracefree part A_{ij} as

$$K_{ij} = A_{ij} + \frac{1}{3} g_{ij} K. \tag{5.12}$$

After decomposing the variables with respect to the conformal metric $\tilde{g}_{ij}$, we obtain:

$$A_{ij} = \psi^{-p} \tilde{A}_{ij}, \tag{5.13}$$
$$K = \tilde{K}, \tag{5.14}$$
$$\beta^i = \tilde{\beta}^i. \tag{5.15}$$

In the standard conformal decomposition often used to solve the constraint equations, one chooses $p = -2$. We will see in the next section, that the BSSN decomposition used in the

moving-puncture method has $p = 4$. It is important to notice that the trace of the extrinsic curvature and the contravariant components of the shift are unchanged by this conformal rescaling.

A new variable $\tilde{\Gamma}^i$ is introduced by the following relation:

$$\tilde{\Gamma}^i = \tilde{g}^{jk}\tilde{\Gamma}^i{}_{jk} = -\partial_j \tilde{g}^{ij}, \tag{5.16}$$

where $\tilde{\Gamma}^i{}_{jk}$ is the conformal connection. This is the original BSSN hyperbolic evolution system.

5.4.1 The puncture approach

One approach to constructing initial data for black hole simulation, is to introduce inner boundaries around each hole, with some imposed mixed (Robin) conditions to guarantee that the final solution did indeed describe one or more black holes, that is, that the solution contained ***apparent horizons***. Excision first appeared in an article by Jonathan Thornburg [111], following a suggestion by Unruh who pointed out that, given that black hole interiors are causally disconnected from the exterior universe, one can excise the inside of a black hole from the computational domain. Since event horizons require the knowledge of the complete spacetime, one should use the apparent horizon as surfaces within which to excise. Also, note that Robin conditions are used in a particular construction of initial data. They are not used in an evolution.

The key idea of the puncture approach, is that the singularities in the Hamiltonian constraint can be absorbed in an analytic expression. Black holes can be represented by a Brill-Lindquist two-sheeted topology at $t = 0$ (see figure 5.3):

$$\psi_{BL} = 1 + \sum_{i=1}^{N} \frac{M_i}{2|\vec{r} - \vec{r}_i|}, \tag{5.17}$$

where M_i are the bare masses of the black holes and $\vec{r}_i$ are the locations of the punctures. Now we can factor out the singular behaviour of the conformal factor ψ using the following ansatz with N black holes [112, 113]:

$$\psi = \psi_{BL} + u. \tag{5.18}$$

We now need to solve the Hamiltonian constraints for u everywhere on $\mathcal{R}^3$ with N punctures:

$$\triangle_{\text{flat}} u = -b\left(1 + \psi_{BL}^{-1} u\right)^{-7}; \tag{5.19}$$

$$b = \frac{1}{8}\psi_{BL}^{-7} \tilde{A}_{ij}\tilde{A}^{ij}. \tag{5.20}$$

There is now no need to excise; the poles at the centre of the black holes have been absorbed into the analytical terms and the ansatz (5.17) effectively takes the place of inner boundary conditions. The corrections u are regular everywhere. A slight disadvantage of this method is that one can locate the apparent horizon only after the data are constructed. One must also adjust the parameters M_i in order to achieve the desired black-hole masses (as defined by the area of the apparent horizon).

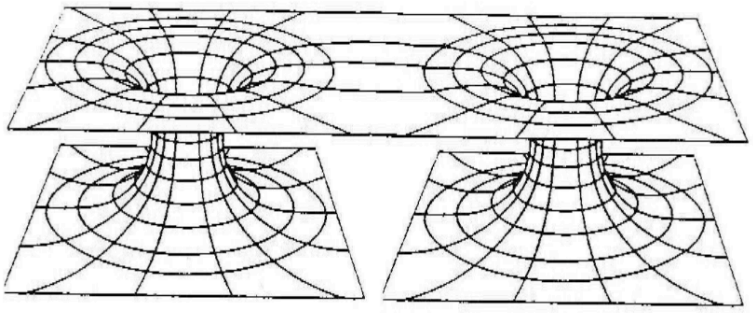


Figure 5.3: Brill-Lindquist two-sheeted topology to represent black holes at $t = 0$ numerically.

Originally, the ***static*** puncture evolution method was introduced in [112, 113], consisting in factoring out the singular part, then evolving the regular part and choosing singularity avoiding slices with vanishing shift to fix the punctures in the grid. Unfortunately, for orbiting binaries this involves a co-rotating gauge to keep the black hole horizons around the punctures, and the code crashed after one orbit. A later paper by the AEI group [114] achieved evolutions that lasted a bit longer, and it is not clear how far the method could be pushed. Two years later, the ***moving*** puncture evolution method was introduced by Campanelli *et al.* [93] and the Goddard group. The singular part of the metric is not factored out, both the regular and singular parts are evolved together. Gauge conditions are chosen so that the punctures move freely in the grid, this results in a stable and accurate code.

The initial data in a typical moving-puncture simulation represents black holes using a wormhole topology. When following the coordinates towards one of the black holes, one does not reach the black hole's singularity but instead passes through a wormhole to another exterior space eventually reaching an asymptotically flat region. For example, data containing N black holes consist of $N + 1$ asymptotically flat regions connected by N wormholes (figure 5.3). Each ***unphysical*** asymptotically flat region is compactified so that its spatial infinity is transformed to a single point, the ***puncture***. Note that in this construction, all of the black hole singularities are conveniently avoided, and no region needs to be ***excised***. Standard puncture data are smooth over the entire space except for the conformal factor ψ which diverges as $1/r$ near each puncture (only in the initial data, it diverges as $1/\sqrt{r}$ for evolutions).

5.4.2 The BSSN system and the moving-punctures

The moving-puncture extension of the BSSN system deals with puncture data, and involves introducing yet another variable, either $\phi = \ln \psi$ (see [115]) or $\chi = \psi^{-4}$ (see [93]). This newly introduced variable is evolved instead of the conformal factor ψ. Although ϕ diverges logarithmically at the puncture, the method appears to be stable. Furthermore, one needs to specify gauge conditions that allow the punctures to move across the numerical

grid. The physical metric and extrinsic curvature variables are now given by:

$$\tilde{g}_{ij} = e^{-4\phi} g_{ij} \quad (5.21)$$

$$\tilde{A}_{ij} = \tilde{K}_{ij} - \frac{1}{3}\tilde{g}_{ij} K. \quad (5.22)$$

The transformations between the ***physical*** metric g_{ij} and the ***conformal*** metric $\tilde{g}_{ij}$ are crucial and are described below:

$$\tilde{g}_{ij} = \psi^{-4} g_{ij} = e^{(-4\phi)} g_{ij}; \quad (5.23)$$

$$\tilde{g}^{ij} = \psi^{4} g^{ij} = e^{(4\phi)} g^{ij}. \quad (5.24)$$

The quantity $\tilde{A}_{ij}$ is rescaled like the metric itself:

$$\tilde{A}_{ij} = e^{-4\phi} A_{ij}, \quad (5.25)$$

The indices of $\tilde{A}_{ij}$ are raised and lowered with the conformal metric $\tilde{g}_{ij}$, so that

$$\tilde{A}^{ij} = e^{4\phi} A^{ij}. \quad (5.26)$$

The new variables of the evolution system are now ϕ, $\tilde{g}_{ij}$, $\tilde{A}_{ij}$, K and $\tilde{\Gamma}^i$. These variables are evolved using the following evolution system

$$\partial_0 \phi = -\frac{1}{6}\alpha K, \quad (5.27)$$

$$\partial_0 \tilde{g}_{ij} = -2\alpha \tilde{A}_{ij}, \quad (5.28)$$

$$\partial_0 \tilde{A}_{ij} = e^{-4\phi}[-D_i D_j \alpha + \alpha R_{ij}]^{TF} + \alpha(K\tilde{A}_{ij} - 2\tilde{A}_{ik}\tilde{A}^k{}_j), \quad (5.29)$$

$$\partial_0 K = -D^i D_i \alpha + \alpha(\tilde{A}_{ij}\tilde{A}^{ij} + \frac{1}{3}K^2), \quad (5.30)$$

$$\partial_t \tilde{\Gamma}^i = \tilde{g}^{jk}\partial_j\partial_k\beta^i + \frac{1}{3}\tilde{g}^{ij}\partial_j\partial_k\beta^k + \beta^j\partial_j\tilde{\Gamma}^i - \tilde{\Gamma}^j\partial_j\beta^i + \frac{2}{3}\tilde{\Gamma}^i\partial_j\beta^j - 2\tilde{A}^{ij}\partial_j\alpha + 2\alpha\left(\tilde{\Gamma}^i{}_{jk}\tilde{A}^{jk} + 6\tilde{A}^{ij}\partial_j\phi - \frac{2}{3}\tilde{g}^{ij}\partial_j K\right), \quad (5.31)$$

where $\partial_0 = \partial_t - \mathcal{L}_\beta$, $\tilde{D}_i$ is the covariant derivative with respect to the ***conformal metric*** $\tilde{g}_{ij}$, D_i is the covariant derivative with respect to the ***physical metric*** g_{ij}, and "TF" denotes the trace-free part of the expression with respect to the ***physical metric***[1], $X^{TF}_{ij} = X_{ij} - \frac{1}{3}g_{ij}X^k_k$.

[1] Note that the trace-free part expression given with the conformal metric is the same, as the conformal factor cancels out.

The Ricci tensor R_{ij} is given by

$$\begin{aligned} R_{ij} &= \tilde{R}_{ij} + R^{\phi}_{ij} && (5.32)\\ \tilde{R}_{ij} &= -\frac{1}{2}\tilde{g}^{lm}\partial_l\partial_m\tilde{g}_{ij} + \tilde{g}_{k(i}\partial_{j)}\tilde{\Gamma}^k + \tilde{\Gamma}^k\tilde{\Gamma}_{(ij)k} + \\ &\quad \tilde{g}^{lm}\left(2\tilde{\Gamma}^k_{l(i}\tilde{\Gamma}_{j)km} + \tilde{\Gamma}^k_{im}\tilde{\Gamma}_{klj}\right), && (5.33)\\ R^{\phi}_{ij} &= -2\tilde{D}_i\tilde{D}_j\phi - 2\tilde{g}_{ij}\tilde{D}^k\tilde{D}_k\phi + 4\tilde{D}_i\phi\tilde{D}_j\phi - \\ &\quad 4\tilde{g}_{ij}\tilde{D}^k\phi\tilde{D}_k\phi. && (5.34) \end{aligned}$$

The Lie derivatives of the tensor densities ϕ, $\tilde{g}_{ij}$ and $\tilde{A}_{ij}$ (with weights $1/6$, $-2/3$ and $-2/3$) are

$$\begin{aligned} \mathcal{L}_\beta\phi &= \beta^k\partial_k\phi + \frac{1}{6}\partial_k\beta^k\phi,\\ \mathcal{L}_\beta\tilde{g}_{ij} &= \beta^k\partial_k\tilde{g}_{ij} + \tilde{g}_{ik}\partial_j\beta^k + \tilde{g}_{jk}\partial_i\beta^k - \frac{2}{3}\tilde{g}_{ij}\partial_k\beta^k,\\ \mathcal{L}_\beta\tilde{A}_{ij} &= \beta^k\partial_k\tilde{A}_{ij} + \tilde{A}_{ik}\partial_j\beta^k + \tilde{A}_{jk}\partial_i\beta^k - \frac{2}{3}\tilde{A}_{ij}\partial_k\beta^k. \end{aligned}$$

In addition, the Lie derivative of a scalar field is given by

$$\mathcal{L}_\beta K = \beta^k\partial_k K. \qquad (5.35)$$

The covariant derivatives of the lapse are with respect with the ***physical metric*** and are defined by

$$D_iD_j\alpha = \partial_i\partial_j\alpha - 4\partial_{(i}\phi\,\partial_{j)}\alpha - \tilde{\Gamma}^k_{ij}\partial_k\alpha + 2g_{ij}g^{kl}\partial_k\phi\,\partial_l\alpha, \qquad (5.36)$$

Furthermore, the trace is given by

$$\begin{aligned} D^iD_i\alpha &= \exp(-4\phi)\;\tilde{g}_{il}\tilde{D}_l\tilde{D}_i\alpha; && (5.37)\\ &= \exp(-4\phi)\;\left(\tilde{g}^{ij}\partial_i\partial_j\alpha - \tilde{\Gamma}^k\partial_k\alpha + 2\tilde{g}^{ij}\partial_i\phi\partial_j\alpha\right). && (5.38) \end{aligned}$$

On the other hand the covariant derivative of ϕ is with respect to the ***physical metric*** and is defined by

$$\tilde{D}_i\tilde{D}_j\phi = \partial_i\partial_j\phi - \tilde{\Gamma}^k_{ij}\partial_k\phi. \qquad (5.39)$$

The BSSN evolution system is hyperbolic [114], first order in time and second order in space.

In practice, the punctures orbit each other and spiral inwards, as if the black holes were being represented by point particles, see figure 5.4. The plots of the punctures tracks easily match our intuitive picture of objects in orbit. However, a word of clarification is in order here, the orbiting punctures are not point particles but asymptotic infinities of wormholes.

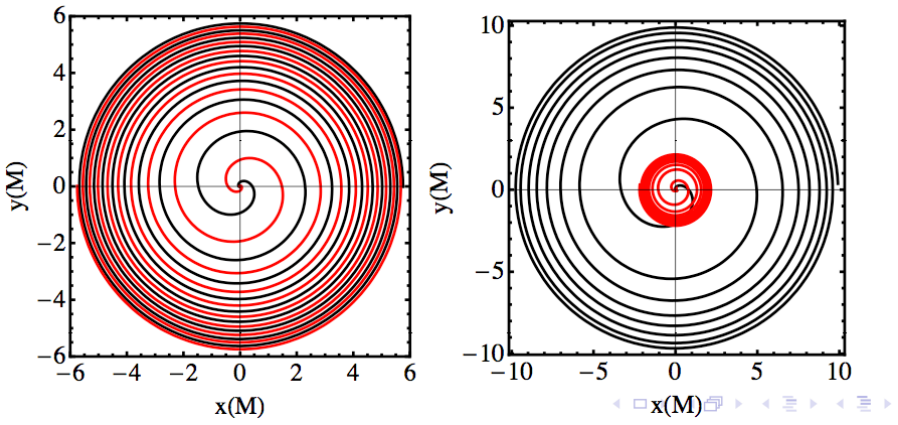


Figure 5.4: The punctures orbit each other and spiral inwards, as if the black holes were being represented by point particles. The plots are for equal-mass and mass-ratio 1:4 non-spinning binaries. The equal-mass data are published in [4]. The 1:4 data are from M. Hannam *et al* unpublished.

5.4.3 The ϕ method versus the χ method

In the ϕ-method, one works directly with the original BSSN variable ϕ,

$$\phi = \ln \psi, \tag{5.40}$$

and the evolution system remains as Eqs (5.27)-(5.31). The purely experimental result is that finite differencing across the $\ln(r)$ singularity at $r = 0$ leads to stable evolutions.

In the χ-method, a new conformal factor is defined, that is finite at the puncture,

$$\chi = \psi^{-4}, \tag{5.41}$$

$$\partial_0 \chi = \frac{2}{3}\chi\left(\alpha K - \partial_k \beta^k\right) \qquad \mathcal{L}_\beta \chi = \beta^k \partial_k \chi \tag{5.42}$$

Now Eq. (5.42) replaces Eq. (5.27) in the BSSN evolution system.

5.5 Coordinate conditions or choices for the gauge

Determining a good coordinate system for use in numerical relativity is not an easy task. First of all, the coordinate system must cover the regions of spacetime of interest, avoid physical singularities and furthermore, remain non-singular and non-pathological itself. A good choice of coordinates can simplify the physics, for example, spherical coordinates for spherical problems. One also wants a choice that will be computationally efficient, and most importantly compatible with hyperbolicity, well-posedness and stability. There are several traditional gauge choices (coordinate conditions) for the lapse α and the shift β:

Geodesic (Gaussian-normal) Coordinates:

$$\alpha = 1 \tag{5.43}$$
$$\beta^i = 0. \tag{5.44}$$

Unfortunately, this coordinate condition is ***singularity seeking***, but it does provide substantial simplification of the $3+1$ equations.

Normal Coordinates:

$$\beta^i = 0. \tag{5.45}$$

This has been widely used in initial phases of code development due to the simplification of the evolution equations.

Maximal Slicing:

$$K = 0. \tag{5.46}$$

The volume of the hypersurfaces are maximized with respect to the continuous deformations within spacetime. This gauge choice is ***singularity avoiding***, which makes it quite popular, however, it requires an elliptic solver at every time step, which makes it computationally expensive.

Note that these different gauge choices can be used in combinations. For example, maximal slicing determines the lapse function, but not the shift.

Harmonic Coordinates:

$$\nabla^a \nabla_a x^\alpha = 0. \tag{5.47}$$

This leads to the lapse and the shift, in a $3+1$ formalism:

$$\left(\partial_t - \beta^j \partial_j\right) \alpha = -\alpha^2 K \tag{5.48}$$
$$\left(\partial_t - \beta^j \partial_j\right) \beta^i = -\alpha^2 \left(g^{ij} \partial_j \ln \alpha + g^{jk} \Gamma^i{}_{jk}\right). \tag{5.49}$$

The field equations reduce to non-linear wave equations, which is very appealing and was therefore widely used in early hyperbolic formulations. However, harmonic slices may tend to be singularity seeking instead of ***singularity avoiding***, and also, harmonic coordinates may be susceptible to ***coordinate singularities***, see [116, 117] for more details on that matter.

Bona-Massó Slicing:

$$\left(\partial_t - \beta^j \partial_j\right) \ln \alpha = -\alpha f(\alpha) K, \tag{5.50}$$

with $f(\alpha) \geq 0$. This slicing condition is invariant under coordinate transformation on each hypersurface. This condition must be expressed in terms of slicing scalars (here first order) and their proper time derivatives. For specific values of $f(\alpha)$, one can recover some of the previously defined gauge conditions:

- $f = 0$: ***Geodesic slicing*** with $\alpha = 1$ initially;
- $f \to \infty$: ***Maximal slicing***;
- $f = 1$: ***Harmonic slicing***;
- $f = 2/\alpha$: $1 + \log$ ***slicing***.

The latter slicing condition $1 + \log$ has ***singularity avoidance*** properties very similar to the maximal slicing condition but is ***inexpensive computationally***. This is the gauge condition that we will use when applying the spectral element method to the BSSN system with moving punctures. This gauge choice relies on the ***covariant*** form of $1 + \log$ slicing,

$$(\partial_t - \beta^i \partial_i)\alpha = -2\alpha K. \tag{5.51}$$

For the shift, we use a gamma-freezing condition [118].

$$\partial_0 \beta^i = \frac{3}{4} B^i, \quad \partial_0 B^i = \partial_t \tilde{\Gamma}^i - \eta B^i. \tag{5.52}$$

Note that here $\partial_0 = \partial_t - \beta^k \partial_k$, but another variant would be to make the replacement $\partial_0 \to \partial_t$ everywhere. The gamma-freezing condition allows the puncture to move across the numerical grid. Thereby, the punctures orbit each other and spiral inwards, as if the black holes were represented by point particles, matching one's intuitive picture of objects in orbit. This condition was originally developed to provide a time-evolution analogue of the "***minimal distortion***" shift (which is given by an elliptic equation, as with maximal slicing), and is meant to minimize the dynamics of the shift. Also, this shift condition has the effect of causing the punctures to orbit, but that was not the intention of the shift condition.

5.6 Numerical Approximations in NR

Considering symmetries when possible improves the computational cost immensely. Numerical approximations consist of the discretization of a continuous set of arbitrary functions. Any function u is replaced by a finite set of discrete values :

$$u(t) \longrightarrow \{u_n\} \quad n = 0..N. \tag{5.53}$$

The continuous set of values of u is replaced by a discrete set of $N+1$ values. The discrete set of values of $\{u_n\}$ can be constructed in many different ways, which depend on the specific numerical method used. Common methods include:

Finite Difference (FD): In this approach, the continuous spacetime is replaced by a lattice of points, that is the numerical grid. The values u_n are the values of u at the grid points.

Spectral Methods (SM): In this approach, the values u_n correspond to the coefficients of the development of the function u in a series with a particular set of basis functions over the ***entire*** domain. Typically, u is approximated by ***global*** basis functions by:

$$u(r,t) = \sum_0^N u_n(t)\phi_n(r). \tag{5.54}$$

The order of the polynomial basis functions in Spectral Methods is usually of the order of $N = 50, 100, 200$. See [119] for a review of spectral methods in numerical relativity and see [120] for a general description of spectral methods.

Finite Element Method: (FEM) In this approach, the entire domain is divided into K elements. The values u_n correspond to the coefficient of the development of the function u in a series with a particular set of basis functions over each element. Typically, u is approximated by a superposition of ***local*** basis functions by:

$$u(r,t) = \sum_{0}^{N} u_n(t)\phi_n(r). \tag{5.55}$$

Moreover, the equations to solve necessarily need to be rewritten in a ***weak form***. A discrete weak formulation is derived from the continuous variational problem, and the latter is formulated by multiplying each side of the equations by a ***test function*** and integrating over the whole domain. The weak formulation is then obtained by integration by parts lowering the differentiability requirements of the approximate solution. Typically the basis functions are polynomials of order $N = 1, 2$ or 3 at most. When more accuracy is needed, there are 3 different strategies:

- Subdivide each element to improve resolution uniformly over the whole domain. This is usually called ***h-refinement*** because h is the common symbol for the size or average size of a subdomain.
- Subdivide only in regions of steep gradients where high resolution is needed. This is called ***r-refinement***.
- Keep the subdomains fixed while increasing N the degree of the polynomials in each subdomain. This strategy is referred to as ***p-refinement***, this is also a technique employed for Spectral Methods.

Spectral Element Method (SEM): This approach is a generalization of the finite element method, it combines the theory of spectral and pseudo-spectral methods for high order polynomials and the variational formulation of finite elements and the associated geometric flexibility. Similar to the FEM, the values u^n correspond to the coefficients of the development of the function u in a series with a particular set of basis functions over ***each*** element. Typically, u is approximated by a superposition of ***local*** basis functions by:

$$u(r,t) = \sum_{0}^{N} u_n(t)\phi_n(r). \tag{5.56}$$

Typically the basis functions are polynomials of order $N = 5, .., 20$ in the SEM case.

Comment on the above methods: Figure 5.5 shows the main differences between the space discretization methods. Note that FD method (second order) can formally be interpreted as the limit case of the FEM and SEM approach for a polynomial order $P = 1$. Spectral methods generate algebraic equations with full matrices. However, the high order

of the basis functions gives high accuracy for a given N. FEM and SEM have two advantages over the SM:

1. The resulting matrix equations from the variational formulation are sparse because only a couple of basis functions are non-zero in a given element;

2. In multi-dimensional problems, the elements become little triangles (FEM) or tetrahedra (FEM and SEM) which can be fitted to very irregular geometries. The disadvantage of FEM is low accuracy because each basis function is a polynomial of low degree.

However, the SEM gains advantages from both FD and SM: The domain is subdivided into elements, to gain the flexibility and matrix sparsity of finite elements, at the same time, the degree of the polynomial N in each subdomain is sufficiently high to retain the high accuracy and low storage of spectral methods.

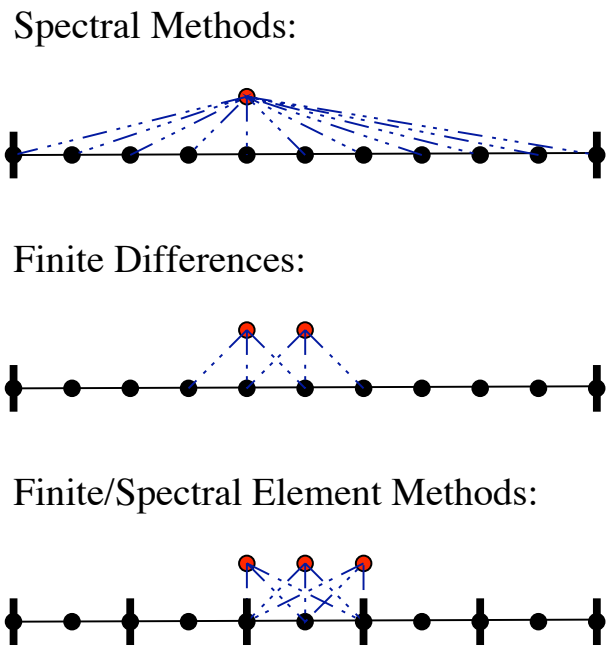


Figure 5.5: Illustration of the main differences between the space discretization of the Spectral Method, the Finite Difference method, and the Spectral Element Method.

In numerical relativity both the SM and FD are used to discretize space only, and the discretization of time is usually dealt with using other methods linked to the FD approach. The most commonly used time discretization method is the Runge-Kutta method. See [120] for a detailed discussion on the comparisons of the aforementioned numerical methods.

5.7 Yet another numerical method?

One of the main goals of Numerical Relativity, is to provide very accurate templates of gravitational waves for ground-based and space-based interferometers to detect. There are now robust and stable numerical methods for Numerical Relativity that work well, for instance finite differences and the moving punctures, and spectral methods with excision. Why is there need for yet another numerical implementation? Current simulations are certainly good enough for ground-based detection, however, the scientific community is not yet sure for LISA, the space-based interferometer. There is also a computational difficulty for high mass ratio simulations, run times can be extremely long. An order-of-magnitude improvement in code efficiency would change the situation tremendously.

What is the potential of the Spectral Element Method for Numerical Relativity? Would this method allow for better accuracy and efficiency? And possibly contribute to gravitational wave detection?

Anyone who considers arithmetical methods of producing random digits is, of course, in a state of sin.

John Von Neumann (1903–1957)

6 Introduction to the Spectral Element Method

The spectral element method (SEM) is a generalization of the finite element method (FEM). Space is divided into a number of subdomains (mesh), and the solution is written with local basis functions that are non-zero over a couple of sub-interval. Typically the basis functions are polynomials of order 1 or 2 in the finite element case, and high order Lagrange-Legendre type polynomials in the spectral element case. Spectral elements combine the theory of spectral and pseudo-spectral methods for high order polynomials and the variational formulation of finite elements and the associated geometric flexibility. The spectral element method may use any type of Jacobi polynomial to define the basis functions but typically either Chebyshev or Legendre polynomials are used. In this thesis we use Legendre polynomials and therefore, the local basis functions are the Lagrange-Legendre interpolants.

The spectral element method was first introduced by Maday and Patera [121] for engineering fluid flow problems. Since then, this method has been used extensively in large scale simulations of incompressible fluid flow, Stokes flows, fluid-structure interactions, the shallow-water equations, seismic wave propagation, oceanic models and many other applications, see [122, 123, 124, 125].

A word of caution is in order, this chapter contains a very technical and theoretical overview of the spectral element method, whereas Chapter 7 contains a more practical overview of the method, where one example (the wave equation) is treated in 1D and 3D.

6.1 Overview of the spectral element method

The spectral element method consists of writing a variational formulation of a specified problem with boundary conditions. Existence and uniqueness to a solution of the weak formulation obtained from this approach can be proved with some version of the Lax-Milgram theorem [126, 127, 128] which will be presented in section 6.4.2. One of its most valuable consequences is an error estimate on the solution when some conditions are met. A discrete weak formulation is derived from the continuous variational problem and its discrete solution is obtained with spectral convergence properties. The layout of the method is as follows:

1. ***The strong formulation*** of a problem is considered with boundary and initial conditions;

2. ***The variational approach*** to this problem is formulated by multiplying each side of the equations by a ***test function*** and integrating over the whole domain;
3. ***The weak formulation*** is obtained by integration by parts lowering the differentiability requirements of the approximate solution;
4. ***The domain discretization*** consists of dividing the domain into subdomains (elements);
5. ***The element discretization*** requires the choice of basis functions for the approximate solution and test functions. The weak formulation of the problem is discretized for each element.
6. ***The elemental matrix form*** of the problem results from the spectral element discretization of the weak form on each element.
7. ***The assembly*** process consists of assembling elemental matrices for each element to form a global system of algebraic equations (typically sparse matrices for conforming elements).

One huge advantage of the method is the fact that any order polynomial can be generated automatically, concurrently with its numerical integration rule. If we select the Gauss quadrature points for the integration rules to be the collocation points we get orthogonal basis functions which means that the mass matrices are then diagonal. There is also no need to define the basis functions explicitly because we can define implicit relations a priori for the inner products of the functions and their derivatives. Since the collocation points are not equi-spaced, staggered grids can be generated automatically by using varying order polynomials for the different variables (say the pressure and velocity in Navier–Stokes avoiding the development of any spurious pressure modes).

There are two paths to convergence with the spectral element method: ***algebraic*** through element refinement (h-refinement) and ***exponential*** (when the solution is smooth) through increasing the interpolation polynomial order (p-refinement). The optimal allocation between the h versus p-type discretization is ***very*** problem dependent. Smooth solutions in regular geometries are most efficiently computed with high-order polynomial order. Complicated geometries and localized features, such as fronts, require using more elements and lower order polynomials. In practice, polynomial orders $N = 7$ or 8 are common because they seem to be a good compromise between accuracy and computational efficiency. An element with a polynomial order N requires $N_{GLL} = N + 1$ points in each space dimension.

6.2 Strong formulation

For clarity of explanation, we will refer to two different problems throughout this section. The first problem is an elliptic stationary homogeneous Dirichlet problem in 3D defined by

$$\begin{cases} -\nabla^2 u(x) &= f(x), \quad \forall\, x \in \boldsymbol{\Omega} \\ u(x) &= 0, \quad \forall\, x \in \partial\boldsymbol{\Omega}. \end{cases} \tag{6.1}$$

This type of problem is a simple case-study, and all the theoretical and numerical analysis of the spectral/finite element method have been extensively explored for this problem [129].

It is therefore a good example to refer to when introducing the main philosophy of the SEM method, which is based on the theory of Sobolev spaces and relatively advanced functional-analytic concepts. This example is also treated in [123].

We will also consider a second more complex problem, treated in a similar way as the example in [130]. Let us consider a 3D advection-diffusion equation for the velocity $u(x,t)$ with $x = (x_1, x_2, x_3) \in \mathbb{R}^3$ with boundary and initial conditions:

$$\left\{ \begin{array}{rcll} \partial_t u + c \cdot \nabla u & = & \nu \nabla^2 u, & \forall\ (x,t) \in \mathbf{\Omega} \times \mathbb{T} \\ u(x,t) & = & b(x,t), & \forall\ (x,t) \in \partial\mathbf{\Omega} \times \mathbb{T}, \\ u(x,t_0) & = & u_i(x), & \forall\ x \in \mathbf{\Omega}, \end{array} \right. \tag{6.2}$$

where ν is the kinematic viscosity. We have the Burger's equation if $c = u$ and if $c = c(t) = (c_1(t), c_2(t), c_3(t))$ we have a linear advection equation with velocity c. Note that in problem (6.2) the boundary conditions are non-homogeneous Dirichlet conditions. Applying the SEM analysis in this case, is equivalent to applying the analysis in the homogeneous case with the use of the auxiliary function $U(x,t) = u(x,t) - U_0(x,t)$, with $U_{0|\partial\mathbf{\Omega}} = b$ and a slightly different right hand side. There are different techniques to deal with various boundary conditions (Dirichlet, Neumann, Fourier conditions) but the philosophy in the analysis is very similar.

Existence and uniqueness of a solution in an evolution problem is a bit more subtle than in a stationary problem. Typically the weak formulation can be formulated to hold on the time interval $\forall t \in [0,T]$, and the test functions v are time independent. In many textbooks, the existence and uniqueness of a solution of a non-stationary problem is simply "***assumed***".

6.3 Variational formulation

A variational formulation reduces the order of the partial derivatives by integrating by parts in the 1D case, or by using Green's formula for higher dimensions. This feature enlarges the space for the numerical solution in the sense that we are able to lower the differentiability requirements of the solution to roughly half those in the original equation. The variational approach requires the use of Lebesgue, Hilbert and Sobolev spaces from functional analysis. We recall that the Lebesgue space $\mathcal{L}^2(\mathbf{\Omega})$ is defined as:

$$\mathcal{L}^2(\mathbf{\Omega}) = \left\{ u : \mathbf{\Omega} \to \mathbb{R} \mid \int_{\mathbf{\Omega}} |u|^2 \, dx < \infty \right\}.$$

In other words, the function u is ***measurable*** over the domain $\mathbf{\Omega}$ if $u \in \mathcal{L}^2(\mathbf{\Omega})$. It is a Hilbert space when equipped with the scalar product

$$(u,v)_{\mathcal{L}^2(\mathbf{\Omega})} = \int_{\mathbf{\Omega}} u\, v \, dx,$$

and an induced norm given by

$$\|u\|_{\mathcal{L}^2(\mathbf{\Omega})} = \sqrt{(u,u)_{\mathcal{L}^2(\mathbf{\Omega})}} = \left(\int_{\mathbf{\Omega}} |u|^2 \, dx \right)^{\frac{1}{2}}.$$

Also the Sobolev space $\mathcal{H}^1(\mathbf{\Omega})$ is defined by

$$\mathcal{H}^1(\mathbf{\Omega}) = \left\{ u \in \mathcal{L}^2(\mathbf{\Omega}) \text{ and } \frac{\partial u}{\partial x_i} \in \mathcal{L}^2(\mathbf{\Omega}),\ i = 1, 2, 3 \right\},$$

and its corresponding norm is

$$\|u\|_{\mathcal{H}^1(\mathbf{\Omega})} = \left(\int_{\mathbf{\Omega}} (|u|^2 + \left| \frac{\partial u}{\partial x_i} \right|^2) \, dx \right)^{\frac{1}{2}}.$$

Furthermore we define the Sobolev space $\mathcal{H}^1_0(\mathbf{\Omega})$ which is the space of $\mathcal{H}^1(\mathbf{\Omega})$ containing all functions in $\mathcal{H}^1(\mathbf{\Omega})$ that vanish at the boundary $\partial\mathbf{\Omega}$,

$$\mathcal{H}^1_0(\mathbf{\Omega}) = \left\{ u \in \mathcal{L}^2(\mathbf{\Omega}) \text{ and } \frac{\partial u}{\partial x_i} \in \mathcal{L}^2(\mathbf{\Omega}),\ i = 1, 2, 3, \text{ and } u_{|\partial\mathbf{\Omega}} = 0 \right\}.$$

In a non-homogeneous Dirichlet problem as in (6.2) we define the Sobolev space

$$\mathcal{H}^1_b(\mathbf{\Omega}) = \left\{ u \in \mathcal{L}^2(\mathbf{\Omega}) \text{ and } \frac{\partial u}{\partial x_i} \in \mathcal{L}^2(\mathbf{\Omega}),\ i = 1, 2, 3, \text{ and } u_{|\partial\mathbf{\Omega}} = b(x, t) \right\}.$$

The first step in writing the variational formulation of the first problem (6.1) is to multiply each side of the equations by a test function $w \in \mathcal{H}^1_0(\mathbf{\Omega})$, integrate over the domain $\mathbf{\Omega}$, and then look for a solution $u \in \mathcal{H}^1_0(\mathbf{\Omega})$. Problem (6.1) is now defined by

$$\left\{ \begin{array}{l} \text{Find } u \in \mathcal{H}^1_0(\mathbf{\Omega}),\ \forall\ w \in \mathcal{H}^1_0(\mathbf{\Omega}) \\ -\displaystyle\int_{\mathbf{\Omega}} \nabla^2 u\ w\ dx = \int_{\mathbf{\Omega}} f\ v\ dx. \end{array} \right. \tag{6.3}$$

For the non-homogeneous case (6.2), the test function is still defined in the Sobolev space that vanishes at the boundary $\mathcal{H}^1_0(\mathbf{\Omega})$, but the space of the solution is different due to the non-vanishing boundary condition. Hence, problem (6.2) is now defined by

$$\left\{ \begin{array}{rcll} & \text{Find } u(.,t) \in \mathcal{H}^1_b(\mathbf{\Omega}) & & \\ \displaystyle\int_{\mathbf{\Omega}} \partial_t u\ w dx + \int_{\mathbf{\Omega}} c \cdot \nabla u\ w dx & = & \displaystyle\int_{\mathbf{\Omega}} \nu \nabla^2 u\ w dx, & \forall\ w \in \mathcal{H}^1_0(\mathbf{\Omega}) \\ u(x, t_0) & = & u_i(x), & \forall\ x \in \mathbf{\Omega}. \end{array} \right. \tag{6.4}$$

6.4 Weak formulation

Problems (6.1) and (6.2) are not quite in the form of a weak formulation yet, they need to be manipulated in order for the second derivative terms to disappear using some combination of integration by parts. In the multi-dimensional case, the most useful formulae are the divergence theorem

$$\begin{array}{rcl} \displaystyle\int_{\mathbf{\Omega}} \nabla \left[a(x) \nabla u(x) \right] w(x) dx & = & \displaystyle\int_{\mathbf{\Gamma}} a(S) \frac{\partial u(S)}{\partial n} w(S) dS \\ & & \displaystyle - \int_{\mathbf{\Omega}} a(x) \left[\nabla u(x) \right] \cdot \left[\nabla w(x) \right] dx, \end{array} \tag{6.5}$$

and Green's formula

$$\int_{\mathbf{\Omega}} \left[\nabla^2 u(x)\right] w(x) dx + \int_{\mathbf{\Omega}} \left[\nabla u(x)\right] \cdot \left[\nabla w(x)\right] dx = \int_{\mathbf{\Gamma}} \frac{\partial u(S)}{\partial \mathbf{n}} w(S) dS, \tag{6.6}$$

where $\partial/\partial\mathbf{n}$ is the derivative in the direction of the outward normal to the boundary $\mathbf{\Gamma}$.

We apply Green's formula to problem (6.1) and use the fact that the test function w vanishes at the boundary to write the equivalent problem in the weak formulation

$$\begin{cases} \text{Find } u \in \mathcal{H}_0^1(\mathbf{\Omega}),\ \forall w \in \mathcal{H}_0^1(\mathbf{\Omega}) \\ \displaystyle\int_{\mathbf{\Omega}} \nabla u \cdot \nabla w \, dx = \int_{\mathbf{\Omega}} f \, w \, dx. \end{cases} \tag{6.7}$$

Note that if u belongs to $\mathcal{C}^2(\mathbf{\Omega})$ then the weak solution is also a solution in the space of $\mathcal{L}^2(\mathbf{\Omega})$, and by reversing the integration by parts, u is also solution of the strong formulation.

There are several variational formulations possible for a given differential problem. They depend on the choice of unknown which is usually driven by computational costs, and depend on the choice of integral transformations which result in the order of differentiation of the unknown. The latest has consequences in the choice of the interpolating polynomials: for the FEM if we choose interpolating polynomials of order N=1 (for example hat functions) then we cannot use second derivatives in the weak form.

The weak formulation for problem (6.2) is

$$\begin{cases} \text{Find } u(.,t) \in \mathcal{H}_b^1(\mathbf{\Omega}), & \forall\ w \in \mathcal{H}_0^1(\mathbf{\Omega}) \\ \displaystyle\int_{\mathbf{\Omega}} \partial_t u \, w \, dx + \int_{\mathbf{\Omega}} C \, u \, w \, dx & = & -\int_{\mathbf{\Omega}} \nu \nabla u \, \nabla w \, dx, \end{cases} \tag{6.8}$$

with the initial condition,

$$u(x, t_0) = u_i(x),\ \forall\ x \in \mathbf{\Omega}. \tag{6.9}$$

and where $C = c \cdot \nabla$ is the advection operator.

For a system of equations, each equation can be multiplied by a different test function, for example, let us consider the Stokes system and nearly incompressible elasticity. Given a viscosity $\nu > 0$, a function f in the dual space of $\mathcal{H}^1(\mathbf{\Omega})$ and g in the dual space of $\mathcal{L}^2(\mathbf{\Omega})$ such that

$$\int_{\mathbf{\Gamma}} g \cdot n \, dS = 0, \tag{6.10}$$

we consider the problem,

$$\begin{cases} \text{Find } (u, p) \in \mathcal{H}_b^1(\mathbf{\Omega}) \times \mathcal{L}_0^2(\mathbf{\Omega}), \\ \quad -\nu\nabla^2 u + \nabla p & = & f, \\ \quad \nabla \cdot u & = & 0, \\ \quad u_{|\mathbf{\Gamma}} & = & g. \end{cases} \tag{6.11}$$

Note that the velocity is continuous whereas the pressure is discontinuous in this formulation. The weak formulation of this problem becomes,

$$\left\{\begin{array}{rcll} \text{Find } (u,p) \in \mathcal{H}_b^1(\mathbf{\Omega}) \times \mathcal{L}_0^2(\mathbf{\Omega}), & & & \\ \nu \displaystyle\int_{\mathbf{\Omega}} \text{trace}(\nabla u \nabla w^T)\, dx - \int_{\mathbf{\Omega}} \nabla \cdot w\, p\, dx & = & \displaystyle\int_{\mathbf{\Omega}} f\, dx, & \forall w \in \mathcal{H}_0^1(\mathbf{\Omega}) \\ - \displaystyle\int_{\mathbf{\Omega}} \nabla \cdot u\, q & = & 0, & \forall q \in \mathcal{L}_0^2(\mathbf{\Omega}), \end{array}\right. \tag{6.12}$$

where the term $\text{trace}(\nabla u \nabla w^T)$ is the Frobenius inner product[1] defined by

$$\text{trace}(\nabla u \nabla w^T) = \sum_{i,j=1}^{n} \frac{\partial u_i}{\partial x_j} \frac{\partial w_i}{\partial x_j}. \tag{6.13}$$

6.4.1 General Boundary conditions

Writing a problem in its variational formulation also depends on the specified boundary conditions. The following section is a short synopsis of examples on how boundary conditions are dealt with.

Non-homogeneous Dirichlet conditions: Consider the following problem,

$$\left\{\begin{array}{rcll} -\nabla^2 u & = & f, & \forall\, x \in \mathbf{\Omega} \\ u_{|\mathbf{\Gamma}} & = & u_b, & \end{array}\right. \tag{6.14}$$

and introduce an auxiliary variable $\tilde{u} = u + u_0$ such that $u_{0|\mathbf{\Gamma}} = u_b$. The problem written for the new unknown $\tilde{u}$ is a homogeneous Dirichlet boundary problem,

$$\left\{\begin{array}{rcll} -\nabla^2 \tilde{u} & = & f + \nabla^2 u_0, & \forall\, x \in \mathbf{\Omega} \\ \tilde{u}_{|\mathbf{\Gamma}} & = & 0, & \end{array}\right. \tag{6.15}$$

Finally, we associate the weak formulation

$$\left\{\begin{array}{rcl} \text{Find } \tilde{u} \in \mathcal{H}_0^1(\mathbf{\Omega}), & & \forall w \in \mathcal{H}_0^1(\mathbf{\Omega}), \\ \displaystyle\int_{\mathbf{\Omega}} \nabla \tilde{u} \cdot \nabla w\, dx & = & \displaystyle\int_{\mathbf{\Omega}} f\, v\, dx - \int_{\mathbf{\Omega}} \nabla u_0 \cdot \nabla w\, dx. \end{array}\right. \tag{6.16}$$

Without going into details, there are practical ways of choosing u_0.

[1] The Forbenius inner product between 2 matrices A and B is sometimes noted $A : B = \text{trace}(AB^T) = \text{trace}(A^T B)$. However, in this thesis we have chosen to use the notation ":" for the Hadamard prodcut that will be defined further on.

Neumann conditions:

Consider the following problem,

$$\left\{\begin{array}{rcl} -\nabla^2 u & = & f, \quad \forall\, x \in \mathbf{\Omega} \\ \dfrac{\partial u}{\partial \mathbf{n}_{|\mathbf{\Gamma}}} & = & g, \end{array}\right. \tag{6.17}$$

where values of u on $\mathbf{\Gamma}$ are unknown and therefore we look for test functions v and the solution u in the space $\mathcal{H}^1(\mathbf{\Omega})$. The weak formulation is given by

$$\left\{\begin{array}{rcl} \text{Find } u \in \mathcal{H}^1(\mathbf{\Omega}), & & \forall w \in \mathcal{H}^1(\mathbf{\Omega}), \\ \displaystyle\int_{\mathbf{\Omega}} \nabla u \cdot \nabla w \; dx & = & \displaystyle\int_{\mathbf{\Omega}} f \; w \; dx + \int_{\mathbf{\Gamma}} g \; w \; ds. \end{array}\right. \tag{6.18}$$

Note that in that case, another condition needs to be satisfied in order to obtain a solution, using Stokes' formula, we get the constraint

$$\int_{\mathbf{\Omega}} f \; dx + \int_{\mathbf{\Gamma}} g \; ds = 0. \tag{6.19}$$

Fourier conditions:

Consider the following problem,

$$\left\{\begin{array}{rcl} -\nabla^2 u & = & f, \qquad\qquad \forall\, x \in \mathbf{\Omega} \\ \dfrac{\partial u}{\partial \mathbf{n}_{|\mathbf{\Gamma}}} & = & -k(u - u_0) + \beta, \end{array}\right. \tag{6.20}$$

with $k > 0$. The corresponding weak formulation is of the form

$$\left\{\begin{array}{rcl} \text{Find } u \in \mathcal{H}^1(\mathbf{\Omega}), & & \forall w \in \mathcal{H}^1(\mathbf{\Omega}), \\ \displaystyle\int_{\mathbf{\Omega}} \nabla u \cdot \nabla w \; dx + \int_{\mathbf{\Gamma}} k \; u \; w \; ds & = & \displaystyle\int_{\mathbf{\Omega}} f \; w \; dx + \int_{\mathbf{\Gamma}} (k u_0 + \beta) \; w \; ds. \end{array}\right. \tag{6.21}$$

6.4.2 Existence and uniqueness of a solution

This is perhaps the most difficult and highly theoretical part of the method. Simple elliptic problems with various boundary conditions have been studied extensively, and have been put into formulations that satisfy certain conditions. More complex problems have been studied case by case, like the Navier-Stokes equations, Burger's equations, and other linear/nonlinear stationary or time-dependent problems, and/or are still active research topics [131].

Without too much detail, we introduce a few definitions that are required for the theorems of existence and uniqueness.

Let $\mathcal{W}$ be a general Hilbert space with scalar product $(\cdot,\cdot)_{\mathcal{W}}$ and corresponding induced norm $\|\cdot\|_{\mathcal{W}}$. A real bilinear form $a(\cdot,\cdot) \; : \; \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ is a function which is linear in each argument separately:

- $a(u_1 + u_2, w) = a(u_1, w) + a(u_2, w)$;
- $a(u, w_1 + w_2) = a(u, w_1) + a(u, w_2)$;
- $a(\lambda u, w) = a(u, \lambda w) = \lambda a(u, w)$.

A bilinear form can be

- ***symmetric*** if $a(u, w) = a(w, u)$, with $u,\ w \in \mathcal{W}$;
- ***continuous*** or ***bounded*** if $|a(u, w)| \leqslant \alpha \|u\|_{\mathcal{W}} \|w\|_{\mathcal{W}}$, $u,\ w \in \mathcal{W}$;
- ***coercive*** or ***elliptic*** if $a(u, u) \geqslant \beta \|u\|^2_{\mathcal{W}}$, $u \in \mathcal{W},\ \beta > 0$.

A sesquilinear form $a(\cdot, \cdot) \ : \ \mathcal{W} \times \mathcal{W} \to \mathbb{C}$ is a bilinear form that is linear in one argument and conjugate-linear in the other.

In a more concise abstract form the weak formulation (6.1) is equivalent to finding $u \in \mathcal{H}^1_0(\mathbf{\Omega})$ such that

$$a(u, w) = L(w), \quad \forall w \in \mathcal{H}^1_0(\mathbf{\Omega}), \tag{6.22}$$

where the symmetric, continuous and coercive bilinear form a is defined as

$$a(u, w) = \int_{\mathbf{\Omega}} \nabla u \cdot \nabla w \, dx, \quad \forall u,\ \forall w \in \mathcal{H}^1_0(\mathbf{\Omega}), \tag{6.23}$$

and the linear form L is also the scalar product

$$L(w) = (f, w) = \int_{\mathbf{\Omega}} f \, w \, dx, \quad \forall f \in \mathcal{L}^2(\mathbf{\Omega}),\ \forall w \in \mathcal{H}^1_0(\mathbf{\Omega}). \tag{6.24}$$

Note that as $a(\cdot, \cdot)$ is continuous, one can associate a linear operator $\mathcal{A}$ such that equation (6.22) can be rewritten as

$$\mathcal{A}u = \mathcal{F}, \tag{6.25}$$

where F is the linear operator associated to L. A lot of the theory relies on the concepts of distributions, for example, the linear operator A is defined in $\mathcal{H}'$ the space of distributions dual to $\mathcal{H}^1_0(\mathbf{\Omega})$.

The abstract weak form of problem (6.2) is

$$b(\dot{u}, w) + b(Cu, w) \quad = \quad -\nu a(u, w), \tag{6.26}$$

where

$$\begin{aligned} a(u, w) = & \int_{\mathbf{\Omega}} \nabla u \cdot \nabla w \, dx, \quad \forall u,\ \forall w \in \mathcal{H}^1_0(\mathbf{\Omega}) \\ b(u, w) = & \int_{\mathbf{\Omega}} u \, w \, dx \qquad \forall u,\ \forall w \in \mathcal{H}^1_0(\mathbf{\Omega}) \end{aligned} \tag{6.27}$$

There are different theorems of existence and uniqueness. The first theorem relies on the properties of the symmetric part of the bilinear form and it is not the most general version.

Theorem 1 *The Lax-Milgram theorem:*
*Let $\mathcal{H}$ be a Hilbert space and $a(\cdot,\cdot)$ a real bilinear form that is **symmetric, continuous,** and **coercive**. Moreover, let a linear functional $L(\cdot)$ be **continuous**. Note that $L(V) = \int_{\mathcal{H}} f\ w\ dx$ with $f \in \mathcal{H}'$. Then, the weak formulation*

$$\begin{cases} \text{Find } u \in \mathcal{H} \\ \quad a(u,w) \quad = \quad L(w), \quad \forall w \in \mathcal{H}, \end{cases} \tag{6.28}$$

admits a unique solution u in $\mathcal{H}$. Additionally,we have the estimate,

$$\|u\|_{\mathcal{H}} \leqslant \frac{1}{\beta}\|f\|_{H'}, \tag{6.29}$$

where β is the coercivity constant.

The Lax-Milgram theorem can be generalized to more complex problems that are non-symmetric and indefinite.

Theorem 2 *The generalized Lax-Milgram theorem:*
*Let $\mathcal{W}$ be a Hilbert space and let us introduce a complex sesquilinear form $a(\cdot,\cdot) : \mathcal{W} \times \mathcal{W} \to \mathbb{C}$. Assume that $a(\cdot,\cdot)$ is **continuous,** and **coercive**. Moreover, let a linear functional $L(\cdot)$ be **continuous**. Note that $L(V) = \int_{\mathcal{W}} f\ w\ dx$ with $f \in \mathcal{W}'$.*

Then, the weak formulation

$$\begin{cases} \text{Find } u \in \mathcal{W} \\ \quad a(u,w) \quad = \quad L(w), \quad \forall v \in \mathcal{W}, \end{cases} \tag{6.30}$$

admits a unique solution u in $\mathcal{W}$, satisfying,

$$\|u\|_{\mathcal{W}} \leqslant \frac{1}{\beta}\|f\|_{\mathcal{W}'}, \tag{6.31}$$

where β is the coercivity constant.

In practice, it is not always easy to prove the coercivity of a problem, but there are a few inequalities that can be very useful in this regard, see for instance some inequalities in appendix C. There is another theorem that uses the Babuška-Brezzi condition that makes explicit conditions under which a problem is well-posed.

Theorem 3 *Babuška-Brezzi theorem:*
*Let $\mathcal{W}$ and $\mathcal{Q}$ be Hilbert spaces with the associated norms $\|\cdot\|_{\mathcal{W}}$ and $\|\cdot\|_{\mathcal{Q}}$. Let $a(\cdot,\cdot)$ be a **continuous bilinear form** on $\mathcal{W} \times \mathcal{W}$, let $b(\cdot,\cdot)$ a **continuous bilinear form** on $\mathcal{W} \times \mathcal{Q}$.*

- *Assume that we can associate to $b(\cdot,\cdot)$ a continuous linear operator $B : \mathcal{W} \to \mathcal{Q}'$[2], defined by $(Bw, q) = b(w,q),\ u \in \mathcal{W},\ q \in \mathcal{Q}$.*

[2] $\mathcal{Q}'$ is the dual space of $\mathcal{Q}$, it is sometimes written as $\mathcal{Q}^*$.

- *Assume that* $b(\cdot,\cdot)$ *satisfies the inf-sup condition*

$$\inf_{q\neq 0\,\in\mathcal{Q}}\sup_{w\neq 0\,\in\mathcal{W}}\frac{b(w,q)}{\|v\|_{\mathcal{W}}\|q\|_{\mathcal{Q}}}\geqslant\gamma>0 \tag{6.32}$$

- *Assume that* $a(\cdot,\cdot)$ *is coercive on* $\ker B=\{w\in\mathcal{W}\mid b(w,q)=0,q\in\mathcal{Q}\}$.

Then the problem

$$\begin{cases} \text{Find } (u,p) & \in & \mathcal{W}\times\mathcal{Q}, & & \\ a(u,w) & + & b(w,p) & = & F(w), \quad \forall w\in\mathcal{W} \\ b(u,q) & & & = & G(q), \quad \forall q\in\mathcal{Q}, \end{cases} \tag{6.33}$$

has a solution (u,p)*. The first component* u *is unique while* p *is defined up to an element in* $\ker B$*. Additionally,*

$$\|u\|_{\mathcal{W}}\leqslant c_1\left(\|F\|_{\mathcal{W}'}+\|G\|_{\mathcal{Q}'}\right), \tag{6.34}$$

and

$$\|p\|_{\mathcal{Q}}\leqslant c_2\left(\|F\|_{\mathcal{W}'}+\|G\|_{\mathcal{Q}'}\right), \tag{6.35}$$

where c_1 *and* c_2 *are constants that depend only on* α_1, α_2 *(the continuity coefficients of* $a(\cdot,\cdot)$ *and* $b(\cdot,\cdot)$*),* β *(coercivity parameter), and* γ *(inf-sup condition).*

It is relatively easy to prove that the symmetric Lax-Milgram theorem holds in the case of the first elliptic stationary Dirichlet problem and that problem (6.1) has a unique solution.

The weak formulation (6.12) of the Stokes system and nearly incompressible elasticity problem (6.11) is of the same form as of problem (6.33). We can therefore apply the Babuška-Brezzi theorem 6.4.2 to prove existence and uniqueness of a solution. In that particular case we can define,

$$a(u,w)=\nu\int_{\mathbf{\Omega}}\text{trace}(\nabla u\nabla w^T)\,dx,\quad b(u,p)=-\int_{\mathbf{\Omega}}\nabla\cdot u\,p\,dx, \tag{6.36}$$

$$F(w)=\int_{\mathbf{\Omega}}f\,dx,\qquad G(q)=0, \tag{6.37}$$

and

$$\mathcal{W}=\mathcal{H}_0^1(\mathbf{\Omega}),\quad \mathcal{Q}=\mathcal{L}_0^2(\mathbf{\Omega}). \tag{6.38}$$

6.4.3 Nonlinear problems

Note that all the previous theorems deal with linear equations. There are several ways to deal with nonlinearity:

Linearization: It is common practice to linearize some unknowns in the Navier-Stokes equations for incompressible viscous fluids. Reference [132] has considered the finite element methods in general relativity using perturbation theory for a non-rotating black hole as their master equations. Proving the existence and uniqueness of a linear problem is a lot easier than its nonlinear counterpart.

Trilinear forms: In [133] a stationary nonlinear fluid-solid interaction is considered. A weak formulation is presented with linear, bilinear and trilinear functionals, for example the nonlinearity appears as

$$a(u,u,w) = \int_{\mathbf{\Omega}} \rho(u\cdot\nabla)u\cdot w\; dx. \tag{6.39}$$

However, there are no standard theorems or proofs of existence and uniqueness using those types of functionals to our knowledge.

Newton iteration: The argument presented in [134] is that by using a Newton iteration to a nonlinear problem, it boils down to solving a linear problem at each step. If this corresponding linear problem is well-posed then the original nonlinear one will be well-posed too. The same type of argument is discussed in [135].

However, in practice it is not automatic for every problem to obtain the required conditions of all the previous existence and uniqueness theorems. The complete analysis of some complex problems can be very difficult and is often still under active research. Some proofs are only possible numerically, see for example [131] for a numerical proof for solutions of nonlinear hyperbolic equations.

6.4.4 Summary on the weak formulation

The weak formulation and proof of existence and uniqueness of a solution can be very technical and require some insights in the theory of functional analysis. However there are some numerical techniques used to prove that the numerical solution obtained is not non-sense [131] and hence this part of the analysis is unfortunately overlooked most of the time. In practice, a reference to the general Lax-Milgram theorem is mentioned and the problem at hand is discretised in its weak formulation without any further lengthy technical arguments.

6.5 Domain discretization in space

The domain $\mathbf{\Omega}$ is decomposed into $N_E = N_{Ex}\times N_{Ey}\times N_{Ez}$ sub-domains $\mathbf{\Omega^k}$ such that

$$\bar{\mathbf{\Omega}} = \bigcup_{\mathbf{k}=0}^{\mathbf{N_E}} \bar{\mathbf{\Omega}}^{\mathbf{k}}, \quad \forall \mathbf{k},\, \mathbf{l} \quad \mathbf{\Omega}^{\mathbf{k}} \bigcap_{\mathbf{k}\neq\mathbf{l}} \mathbf{\Omega}^{\mathbf{l}} = 0, \tag{6.40}$$

where $\bar{\mathbf{\Omega}}$ is the closure of the domain $\mathbf{\Omega}$. The weak formulation is applied in each subdomain $\mathbf{\Omega^k}$ individually. It is common to use the notation u_h to represent the discrete variable resulting from the continuous variable u. In each subdomain, the ***generic*** variable $u_h^{\mathbf{k}}$ is expanded into cardinal basis functions. In higher dimensions, the formulation of the basis comes from the tensor product of one dimensional Lagrangian interpolant basis $h_i(x)$. So the Lagrangian interpolants are chosen as basis functions in each dimension. We expand

the unknowns as

$$\forall u_h^{\mathbf{k}} \in \mathcal{W}_h,\ u_h^{\mathbf{k}}(x,y,z,t) \quad = \quad \sum_{m=0}^{m=N} \sum_{n=0}^{n=N} \sum_{p=0}^{p=N} u_{mnp}^{\mathbf{k}}(t)\ h_m(x)\ h_n(y)\ h_p(z). \tag{6.41}$$

Variables/Coordinates	**x**	**y**	**z**
Unknowns	m	n	p
Tests functions	a	b	c
Local nodes	a	b	c
GLL quadrature	q	r	s
Master Element	ξ	η	ζ

Table 6.1: Index conventions in 3D

In 3D, the domain is decomposed into nonoverlapping hexahedral elements. Spectral elements are multi-element methods and have a weighted residual based implementation in the same way as the finite element method. There are two types of domain decompositions, one that uses ***conforming*** elements and another that is based on ***non-conforming*** elements (see figure 6.1). Conformity means that the decomposition satisfies the constraint that the intersection of two adjacent elements is either an entire edge or a vertex and the order of approximation is equal for adjacent elements. In other words, it means that the collocation points match at element interfaces. It is very difficult to keep conforming interfaces when using an adaptive mesh refinement procedure as it results in severe restrictions. Therefore the non-conforming elements need to be introduced for adaptive mesh refinement. In the variational approach, continuity across conforming element interfaces is naturally imposed. There are several ways to extend this classical conforming formulation to the non-conforming element case but that will not be described any further.

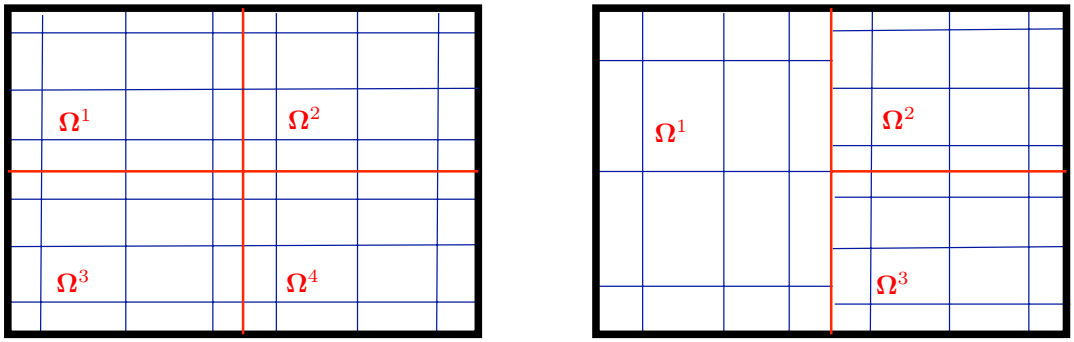
The particularity of the FEM and SEM is that the numerical solution of a problem is described by ***local basis functions***. Let $\mathbb{P}_N$ denote the space of polynomials of degree less that or equal to N in each spatial dimension. In the case of the FEM, the local basis functions are constructed from polynomials of order 0 to 3, for the SEM we use high degree polynomials based on the same philosophy as the spectral method (with Legendre or Chebychev polynomials). In the conforming SEM, $\mathcal{W}_h$ is a subspace of the Sobolev space $\mathcal{W} = \mathcal{H}_0^1(\mathbf{\Omega})$ that consists of all the piecewise high-order polynomials defined on $\mathbf{\Omega}^k$,

$$\mathcal{W}_h = \mathcal{H}_0^1(\mathbf{\Omega}) \bigcap \mathbb{P}_{N,\mathbf{k}}(\mathbf{\Omega}), \tag{6.42}$$

where the space $\mathbb{P}_{N,\mathbf{k}}(\mathbf{\Omega})$ is defined for each discretization parameter h over the domain $\mathbf{\Omega}$ such that

$$\mathbb{P}_{N,\mathbf{k}}(\mathbf{\Omega}) = \left\{ \theta \in \mathcal{L}^2(\mathbf{\Omega}), \quad \theta_{|_{\mathbf{\Omega}^k}} \in \mathbb{P}_N(\mathbf{\Omega}^k) \right\}. \tag{6.43}$$

$\mathbb{P}_N(\mathbf{\Omega}^k)$ denotes the space of polynomials of degree less than or equal to N on each subdomain $\mathbf{\Omega}^k$. Note that the space $\mathbb{P}_{N,\mathbf{k}}(\mathbf{\Omega})$ ensures that the solution is integrable over the domain $\mathbf{\Omega}$, whereas $\mathcal{H}_0^1(\mathbf{\Omega})$ ensures continuity over $\mathbf{\Omega}$.



Conforming Domain Ω Non-conforming Domain Ω

Figure 6.1: 2D conforming domain Ω on the left and 2D non-conforming domain Ω on the right

Here we use the index conventions in Table (6.5) and $u^{\mathbf{k}}_{mnp}(t) = u^{\mathbf{k}}_{mnp}(x,y,z,t)$ are the nodal basis coefficients. The space $\mathcal{W}_h = \mathcal{W} \bigcup \mathbb{P}_{N,\mathbf{k}}(\mathbf{\Omega}) \times \mathbb{P}_{N,\mathbf{k}}(\mathbf{\Omega}) \times \mathbb{P}_{N,\mathbf{k}}(\mathbf{\Omega})$ is taken to be a subspace of $\mathcal{W}$ and consisting of the tensor product of all piecewise high order polynomials of degree less than or equal to N defined on $\mathbf{\Omega^k}$. Furthermore, we have the definition

$$\mathbb{P}_{N,\mathbf{k}}(\mathbf{\Omega}) = \left\{ \theta \in \mathcal{L}^2(\mathbf{\Omega}), \quad \theta_{|_{\mathbf{\Omega^k}}} \in \mathbb{P}_N(\mathbf{\Omega^k}) \right\}. \tag{6.44}$$

For ***regular*** shaped elements (see Appendix F for ***general*** shaped elements), we can also derive the unknowns with respect to x, y, z or t in the following manner:

$$\partial_x u^{\mathbf{k}}_h(x,y,z,t) = \sum_{m=0}^{m=N} \sum_{n=0}^{n=N} \sum_{p=0}^{p=N} u^{\mathbf{k}}_{mnp}(t)\ \partial_x h_m(x)\ h_n(y)\ h_p(z), \tag{6.45}$$

$$\partial_y u^{\mathbf{k}}_h(x,y,z,t) = \sum_{m=0}^{m=N} \sum_{n=0}^{n=N} \sum_{p=0}^{p=N} u^{\mathbf{k}}_{mnp}(t)\ h_m(x)\ \partial_y h_n(y)\ h_p(z), \tag{6.46}$$

$$\partial_z u^{\mathbf{k}}_h(x,y,z,t) = \sum_{m=0}^{m=N} \sum_{n=0}^{n=N} \sum_{p=0}^{p=N} u^{\mathbf{k}}_{mnp}(t)\ h_m(x)\ h_n(y)\ \partial_z h_p(z), \tag{6.47}$$

$$\partial_t u^{\mathbf{k}}_h(x,y,z,t) = \sum_{m=0}^{m=N} \sum_{n=0}^{n=N} \sum_{p=0}^{p=N} \dot{u}^{\mathbf{k}}_{mnp}(t)\ h_m(x)\ h_n(y)\ h_p(z). \tag{6.48}$$

The test function w_h is selected to be the same as the basis functions $h(x) \times h(y) \times h(z)$ used for the generic unknowns u_h, and therefore using Einstein summation convention,

$$w^{\mathbf{k}}_h(x,y,z) = w^{\mathbf{k}}_{abc} h_a(x)\ h_b(y)\ h_c(z), \tag{6.49}$$

where $w^{\mathbf{k}}_{abc} = 1,\ \forall a,b,c$ for the test functions. Note that the same test functions are used for each variable here but they could be different if one was for example to choose a different polynomial order for each unknown. In the case of the Navier-Stokes equations the velocity

u can be written with a polynomial of order N and the pressure with order $N-2$. This avoids spurious modes, and hence the test functions associated to u and p respectively, are different.

Spectral methods are popular for their exponential convergence characteristics. However, in order to keep this order of the approximation error, one needs to numerically integrate equation (6.51) with sufficient accuracy. Hence, the quadrature errors need to be of the same order as the approximation error. The convergence of the numerical solution u_h to the solution u can be determined by stability and approximation theory.

If the conditions of the Lax-Milgram theorem were met for the weak formulation over the whole domain, then the theorem can also be applied in each sub-domain, and the discrete weak formulation admits a unique solution.

For a general Galerkin numerical approximation of the general problem on the continuous domain $\mathbf{\Omega}$,

$$\left\{ \begin{array}{lll} \text{Find } u \in \mathcal{W} & & \forall w \in \mathcal{W} \\ a(u,w)_{\mathcal{W}} & = & L(w)_{\mathcal{W}}, \end{array} \right. \tag{6.50}$$

the variational form applies to a family of discrete dimensional spaces $\mathcal{W}_h$,

$$\left\{ \begin{array}{lll} \text{Find } u_h \in \mathcal{W}_h & & \forall w_h \in \mathcal{W}_h \\ \sum\limits_{\mathbf{k}=1}^{N_E} a(u_h,w_h)_{\mathcal{W}_h} & = & \sum\limits_{\mathbf{k}=1}^{N_E} L(w_h)_{\mathcal{W}_h}, \end{array} \right. \tag{6.51}$$

where $h = (N, \mathbf{k})$ denotes a discretization parameter that depends on the number of elements N_E and the degree of the interpolating polynomials.

The partially discrete weak formulation of problem (6.1) can be written as

$$\left\{ \begin{array}{lll} \text{Find } u_h \in \mathcal{W}_h, & & \forall w_h \in \mathcal{W}_h \\ \sum\limits_{\mathbf{k}=1}^{N_E}{}' \int_{\mathbf{\Omega^k}} \nabla u_h \cdot \nabla w_h \, dx & = & \sum\limits_{\mathbf{k}=1}^{N_E}{}' \int_{\mathbf{\Omega^k}} f_h \, w_h \, dx, \end{array} \right. \tag{6.52}$$

where $\mathcal{W}_h$ is the same as in (6.42). Here $\sum\limits_{\mathbf{k}=1}^{N_E}{}'$ denotes elemental direct summation in which the continuity and boundary conditions are taken into account.

The partially discrete weak formulation for problem (6.2) is written as

$$\left\{ \begin{array}{lll} \text{Find } u_h(.,t) \in \mathcal{U}_h, & & \forall\ w_h \in \mathcal{W}_h \\ \sum\limits_{\mathbf{k}=1}^{N_E}{}' \int_{\mathbf{\Omega^k}} \partial_t u_h \, w_h \, dx + \sum\limits_{\mathbf{k}=1}^{N_E}{}' \int_{\mathbf{\Omega}^k} C \, u_h \, w_h \, dx & = & -\sum\limits_{\mathbf{k}=1}^{N_E}{}' \int_{\mathbf{\Omega}^k} \nu \nabla u_h \, \nabla w_h \, dx, \end{array} \right. \tag{6.53}$$

with the initial condition,

$$\forall \mathbf{k} \quad u_h(x,t_0) = \vec{u_i}(x), \ \forall\ x \in \mathbf{\Omega^k}. \tag{6.54}$$

and where $C = c \cdot \nabla$ is the advection operator. Note that here $\mathcal{U} = \mathcal{H}_b^1(\mathbf{\Omega})$ with $\mathcal{U}_h = \mathcal{H}_b^1(\mathbf{\Omega}) \bigcap \mathbb{P}_{N,\mathbf{k}}(\mathbf{\Omega})$.

On each element $\mathbf{\Omega^k}$ there are $N_{GLL}^3 = (N+1)^3$ nodal points but in total there are

$$N_g = [N_{Ex}N + 1] \times [N_{Ey}N + 1] \times [N_{Ez}N + 1] \tag{6.55}$$

global nodal points. One needs to create a global numbering function that keeps track of local and global nodes on the domain $\mathbf{\Omega}$. There are many ways to label the elements and element nodes. The different protocols of element and node numbering have no effect on the spectral element solution itself but they have a huge impact on the structure of the global mass and advection matrices and therefore on the efficiency of the spectral element code. Figure (6.2) illustrates an example of global numbering technique ***in 2D*** for $N_{Ex} = N_{Ey} = 2$ elements so $N_E = 4$, and polynomial order $N = 3$, that is $N_{GLL} = 4$ GLL points per space. Figure (6.3) shows the local numbering convention per subdomain $\mathbf{\Omega}^k$ and corresponding elemental matrix storage. Let $\mathcal{I}$ denote global indices which are functions of the element index $\mathbf{k}$ and the indexes a, b, c within each element,

$$\mathcal{I} = \mathcal{I}(a, b, c, \mathbf{k}). \tag{6.56}$$

The global numbering function maps the local numbering of the computational nodes to their global (non-redundant) numbering. $\mathcal{I}(a, b, c, \mathbf{k})$ is the global node index of the (a, b, c)-th GLL node internal to the $\mathbf{k}$-th element. Here, elements are numbered row by row from bottom-left to top-right. The table of indices $\mathcal{I}(a, b, c, \mathbf{k})$ is typically needed to build or assemble global data from local data (that is, assemble the contributions from each element).

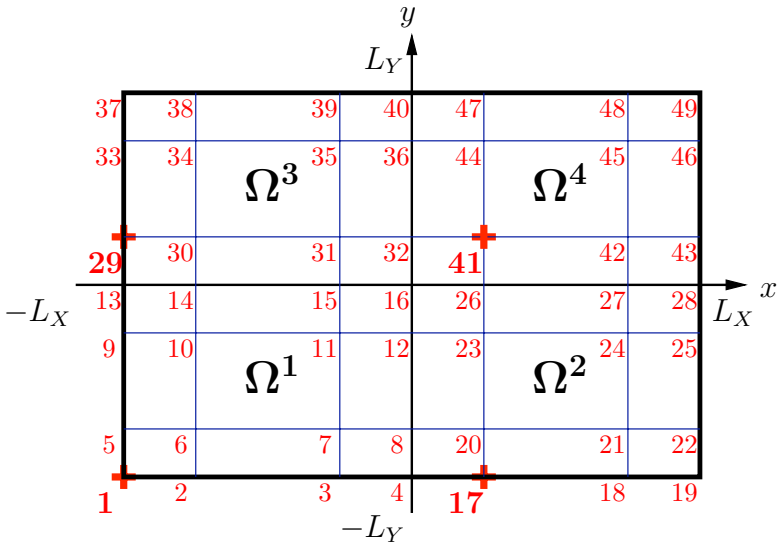


Figure 6.2: Global numbering conventions on a rectangular 2D domain $\mathbf{\Omega}$ in terms of global GLL nodes and subdomains $\mathbf{\Omega^k}$.

Note that in 2D, the elemental matrix that represents the unknown $u^{\mathbf{k}}$ on the $\mathbf{k}$-th element is a $N_{GLL} \times N_{GLL}$ matrix. Each nodal coefficient is noted $u^{\mathbf{k}}_{mn}$. In 3D, the elemental matrix that represents the unknown $u^{\mathbf{k}}$ on the $\mathbf{k}$-th element is a $N_{GLL} \times N_{GLL} \times N_{GLL}$ matrix. Each nodal coefficient is noted $u^{\mathbf{k}}_{mnp}$.

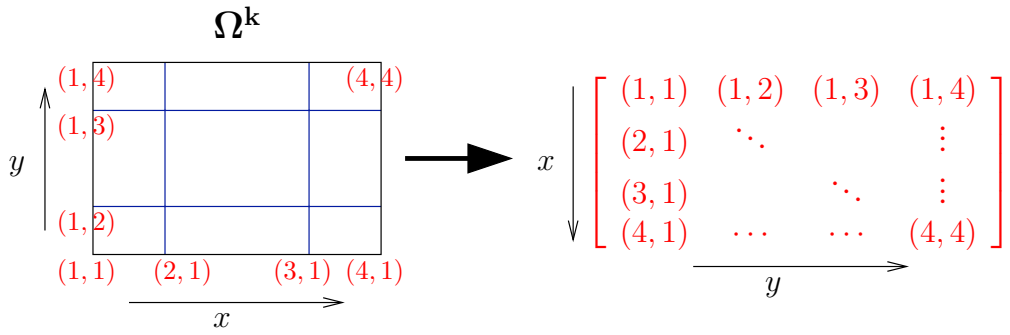


Figure 6.3: Local numbering conventions in 2D for any node (a, b) per subdomain $\mathbf{\Omega^k}$ and corresponding elemental matrix storage.

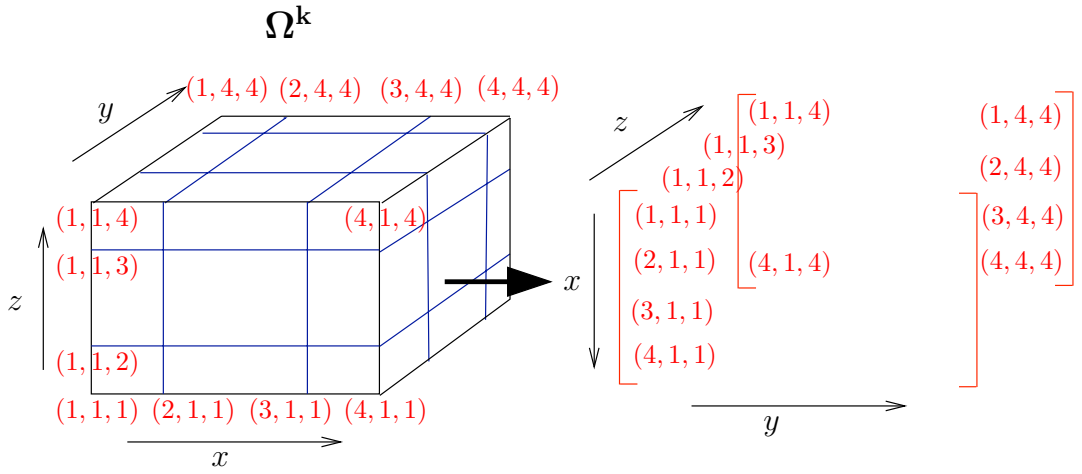


Figure 6.4: Local numbering conventions in 3D for any node (a, b, c) per subdomain $\mathbf{\Omega^k}$ and corresponding elemental matrix storage.

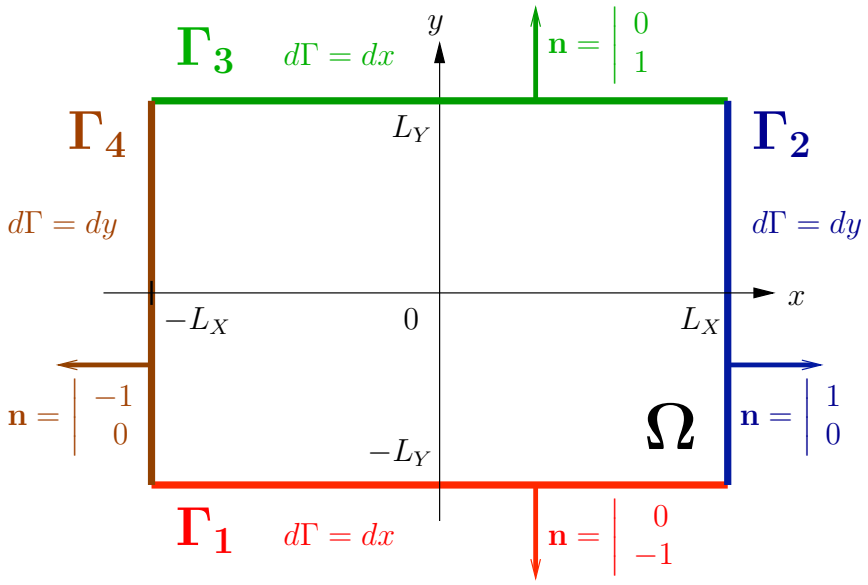


Figure 6.5: 2D domain $\mathbf{\Omega}$ with boundaries $\mathbf{\Gamma}_1$, $\mathbf{\Gamma}_2$, $\mathbf{\Gamma}_3$ and $\mathbf{\Gamma}_4$ and corresponding outward unit normal $\mathbf{n}$.

6.6 Element discretization

The elemental discretization is where the finite element method differs dramatically from the spectral element method. Generally the interpolating polynomials in the SEM are the Lagrange-Legendre basis functions of order N for one space direction. Each dimension does not need to have the same degree. Each element is discretized by $N^1 \times N^2 \times N^3$ collocation points $\xi_a^1, \xi_b^2, \xi_c^3$ in 3D. When considering a system with several unknowns it is not obvious that all the polynomial interpolants have the same degree. For example the Navier-Stokes problem formulated by Maday and Patera consider the mixed formulation $\mathbb{P}_N - \mathbb{P}_{N-2}$ for the velocity and pressure respectively. However, it is a lot more simple to consider $N^1 = N^2 = N^3 = N$ as is the case in this thesis, where $N_{GLL} = N + 1$ is the number of GLL points per element in each space direction.

6.6.1 Gauss–Lobatto–Legendre quadrature

There are several numerical techniques to compute integrals numerically, they are typically referred to as quadrature rule. Here the quadrature applied to the integrals is the ***Gauss–Lobatto–Legendre (GLL) quadrature***, it includes boundary points of the interval $\Lambda = [-1, 1]$ as collocation points. Refer to appendices D and E for more details on Lagrange and Legendre polynomials. The GLL quadrature is defined as follows

$$\int_{-1}^{1} \Phi(\xi) d\xi = \sum_{i=0}^{N} \rho_i \Phi(\xi_i) + \epsilon_N, \quad \forall \Phi \in \mathbb{P}_{2N-1}(-1, 1), \tag{6.57}$$

with the collocation points defined as

$$\xi_0 = -1, \ \xi_{N^1} = 1, \ L'_N(\xi_i) = 0 \ \forall i \in \{1, ..., N-1\}. \tag{6.58}$$

The quadrature weights are ρ_i given by,

$$\rho_i \quad = \quad \frac{2}{N(N+1)\Big(L_N(x_i)\Big)^2}. \tag{6.59}$$

L_N is the N^{th} order Legendre polynomial and the error is $\epsilon_N \sim \mathcal{O}(\Phi^{2N}(\xi))$ for some point $\xi \in (-1, 1)$.

A very important property of Gaussian quadratures (including the GLL quadrature), is that they are ***exact*** with $\epsilon_N = 0$ ***if*** the integrand $\Phi(\xi)$ is a polynomial of degree $2N - 1$ or less. For deformed elements there are additional errors related to curvature. In the spectral element method, each integration on the master element involves the product of two polynomials of degree N the unknown and the test function. The integration of the resulting polynomial of degree $2N$ is thus never exact, even in this simple case. In order to take advantage of efficient sum-factorization techniques, the basis points are taken to be the same as the quadrature points on each element. This results in a diagonal mass matrix obtained by a process of subintegration. Consequently, the mass matrix is always diagonal by construction. In this respect, the SEM is related to FEM in which mass lumping is used to avoid the costly resolution of the non-diagonal system resulting from the use of Gauss

quadrature. In other words, the GLL quadrature used in this way allows for fully explicit schemes.

The interpolants h_i are expressed as

$$h_i(\xi) = -\frac{\left(1-\xi^2\right) L'_N(\xi)}{N\,(N+1)\,L_N(\xi_i)(\xi-\xi_i)}, \quad \xi \in \Lambda,\ \forall i \in \{0, N\}, \tag{6.60}$$

with the following properties

$$h_i(\xi_j) \quad = \quad \delta_{ij},\ \forall i,j \in \{0,N\}^2,\ h_i \in \mathbb{P}_N(\Lambda). \tag{6.61}$$

Furthermore the derivative of the Lagrange-Legendre interpolants matrix is defined by

$$\partial_\xi h_j(\xi_i) = H_{ij} = \begin{cases} H_{00} = -H_{NN} = -\frac{N(N+1)}{4} & \\ H_{ii} = 0 & i \in \{1, N-1\} \\ H_{ij} = \frac{L_N(\xi_i)}{L_N(\xi_j)\,(\xi_i-\xi_j)} & i \neq j \end{cases} \tag{6.62}$$

On the other hand, the second derivative of the Legendre interpolants matrix is defined by

$$\partial_{\xi\xi} h_j(\xi_i) = W_{ij} = \begin{cases} W_{00} = \frac{(-1)^N}{3} L''_N(-1) & \\ W_{NN} = \frac{1}{3} L''_N(1) & \\ W_{ii} = \frac{1}{3}\frac{L''_N(\xi_i)}{L_N(\xi_i)} & i \in \{1, N-1\} \\ W_{ij} = -2\frac{L_N(\xi_i)}{L_N(\xi_j)\,(\xi_i-\xi_j)^2} & i \neq j \end{cases} \tag{6.63}$$

Refer to appendix D on how to derive the 2 preceding matrices.

The GLL quadrature formula in 3D is given for some function $f(\xi,\eta,\zeta)$ by

$$\int_{\Lambda^3} f(\xi,\eta,\zeta)\ d\xi\ d\eta\ d\zeta \quad \simeq \quad \sum_{q=0}^{N}\sum_{r=0}^{N}\sum_{s=0}^{N} \rho_{qrs} f(\xi_q,\eta_r,\zeta_s)\ d\xi\ d\eta\ d\zeta, \tag{6.64}$$

where $\rho_{qrs} = \rho_q\ \rho_r\ \rho_s$ are the weights in 3D.

6.6.2 Master Element

To apply the quadrature rule on each element, one needs to define an affine transformation to map each spectral element $\mathbf{\Omega^k}$ to the reference or master element $\Lambda \times \Lambda \times \Lambda = \Lambda^3$ (see Figure 6.6). This is a common feature to spectral methods, finite element methods and spectral element methods. Let us define the local elemental mappings:

$$(x,y,z)^{\mathbf{k}} \ = \ (x,y,z)^{\mathbf{k}}_{abc}\ h_a(\xi)\ h_b(\eta)\ h_c(\zeta), \tag{6.65}$$

we can now map the physical elements $(x,y,z)^{\mathbf{k}} \in \mathbf{\Omega^k}$ onto the computational domain $(\xi,\eta,\zeta) \in \Lambda^3$. We denote by $J^{\mathbf{k}}$ the Jacobian associated to this mapping such that

$$J^{\mathbf{k}} = \frac{\partial(x,y,z)^{\mathbf{k}}}{\partial(\xi,\eta,\zeta)} = \begin{pmatrix} \frac{\partial x^{\mathbf{k}}}{\partial\xi} & \frac{\partial x^{\mathbf{k}}}{\partial\eta} & \frac{\partial x^{\mathbf{k}}}{\partial\zeta} \\ \frac{\partial y^{\mathbf{k}}}{\partial\xi} & \frac{\partial y^{\mathbf{k}}}{\partial\eta} & \frac{\partial y^{\mathbf{k}}}{\partial\zeta} \\ \frac{\partial z^{\mathbf{k}}}{\partial\xi} & \frac{\partial z^{\mathbf{k}}}{\partial\eta} & \frac{\partial z^{\mathbf{k}}}{\partial\zeta} \end{pmatrix}. \tag{6.66}$$

By $\partial x^{\mathbf{k}}/\partial\xi$ we refer to $\partial x/\partial\xi$ for some point x in the **k**th element $\mathbf{\Omega^k}$. We refer to $|J^{\mathbf{k}}|$ as the determinant of the Jacobian $J^{\mathbf{k}}$. This change of variable is a key component of the method and $|J^{\mathbf{k}}|$ appears in the elemental matrix discretization,

$$\begin{aligned} |J^{\mathbf{k}}| \;=\; & \frac{\partial x^{\mathbf{k}}}{\partial\xi}\frac{\partial y^{\mathbf{k}}}{\partial\eta}\frac{\partial z^{\mathbf{k}}}{\partial\zeta} - \frac{\partial x^{\mathbf{k}}}{\partial\xi}\frac{\partial y^{\mathbf{k}}}{\partial\zeta}\frac{\partial z^{\mathbf{k}}}{\partial\eta} + \frac{\partial x^{\mathbf{k}}}{\partial\eta}\frac{\partial y^{\mathbf{k}}}{\partial\zeta}\frac{\partial z^{\mathbf{k}}}{\partial\xi} \\ & - \frac{\partial x^{\mathbf{k}}}{\partial\eta}\frac{\partial y^{\mathbf{k}}}{\partial\xi}\frac{\partial z^{\mathbf{k}}}{\partial\zeta} + \frac{\partial x^{\mathbf{k}}}{\partial\zeta}\frac{\partial y^{\mathbf{k}}}{\partial\xi}\frac{\partial z^{\mathbf{k}}}{\partial\eta} - \frac{\partial x^{\mathbf{k}}}{\partial\zeta}\frac{\partial y^{\mathbf{k}}}{\partial\eta}\frac{\partial z^{\mathbf{k}}}{\partial\xi}. \end{aligned} \tag{6.67}$$

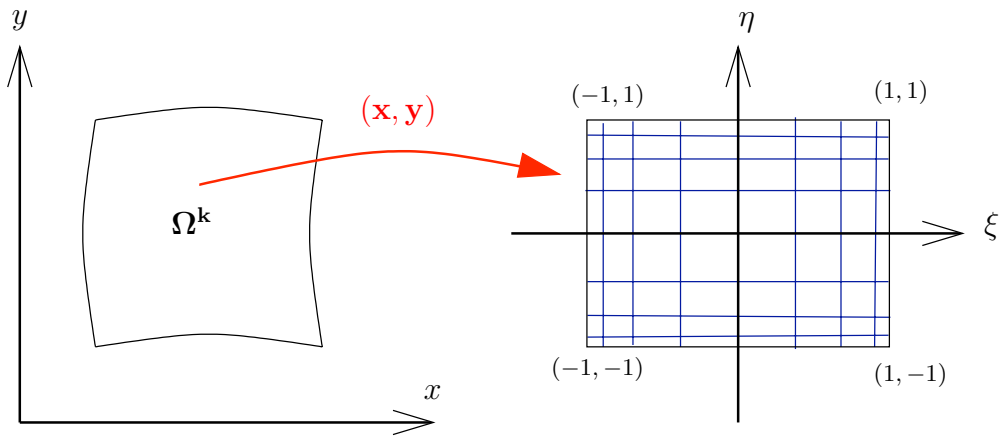


Figure 6.6: Coordinate mapping from a physical element to a master element in 2D.

Special case: 3D homogeneous element decomposition

Practically, derivatives with respect to the physical coordinate x are evaluated in terms of the computational coordinate ξ (respectively for y, η and z, ζ). The mapping from the element

$$(x,y,z)^{\mathbf{k}} \in \mathbf{\Omega^k} = [X_{\mathbf{k}}, X_{\mathbf{k}+1}] \times [Y_{\mathbf{k}}, Y_{\mathbf{k}+1}] \times [Z_{\mathbf{k}}, Z_{\mathbf{k}+1}] \tag{6.68}$$

to the computational space used is

$$\xi = \frac{2}{\Delta x^{\mathbf{k}}}(x^{\mathbf{k}} - X_{\mathbf{k}}) - 1, \tag{6.69}$$

$$\eta = \frac{2}{\Delta y^{\mathbf{k}}}(y^{\mathbf{k}} - Y_{\mathbf{k}}) - 1 \tag{6.70}$$

$$\zeta = \frac{2}{\Delta z^{\mathbf{k}}}(z^{\mathbf{k}} - Z_{\mathbf{k}}) - 1 \tag{6.71}$$

where $\Delta x^{\mathbf{k}} = X_{\mathbf{k}+1} - X_{\mathbf{k}}$, $\Delta y^{\mathbf{k}} = Y_{\mathbf{k}+1} - Y_{\mathbf{k}}$ and $\Delta z^{\mathbf{k}} = Z_{\mathbf{k}+1} - Z_{\mathbf{k}}$ so that

$$\begin{aligned} \frac{\partial x^{\mathbf{k}}}{\partial\xi} &= \frac{\Delta x^{\mathbf{k}}}{2}, \quad \frac{\partial x^{\mathbf{k}}}{\partial\eta} = 0, \quad \text{and} \quad \frac{\partial x^{\mathbf{k}}}{\partial\zeta} = 0, \\ \frac{\partial y^{\mathbf{k}}}{\partial\eta} &= \frac{\Delta y^{\mathbf{k}}}{2}, \quad \frac{\partial y^{\mathbf{k}}}{\partial\xi} = 0, \quad \text{and} \quad \frac{\partial y^{\mathbf{k}}}{\partial\zeta} = 0, \\ \frac{\partial z^{\mathbf{k}}}{\partial\zeta} &= \frac{\Delta z^{\mathbf{k}}}{2}, \quad \frac{\partial z^{\mathbf{k}}}{\partial\xi} = 0, \quad \text{and} \quad \frac{\partial z^{\mathbf{k}}}{\partial\eta} = 0, \end{aligned} \tag{6.72}$$

and hence, the determinant of the Jacobian simplifies drastically

$$|J^{\mathbf{k}}| \quad = \quad \frac{\Delta x^{\mathbf{k}}\ \Delta y^{\mathbf{k}}\ \Delta z^{\mathbf{k}}}{8}. \tag{6.73}$$

In particular, the Jacobian $|J^{\mathbf{k}}|$ becomes the same for all the elements $\forall\ \mathbf{k}$ in the case of a homogeneous (evenly decomposed) domain in the x, y and z directions.

Note that this is what we would intuitively expect: we have just provided a uniform scaling to our elements, without deforming them in any way, so the Jacobian here just gives us an appropriate scale factor.

6.6.3 Elemental matrix form

The elemental matrix forms can be very specific to a given problem. The following is a brief overview, the 1D and 3D wave equations are treated more explicitly and in details in chapter 7. For each element $\mathbf{\Omega^k}$, the weak formulation of problem 6.1 has a corresponding discretized equation in ***Local Matrix form*** as follows:

$$\begin{aligned} &\text{Weak form on } \mathbf{\Omega^k} && \int_{\mathbf{\Omega^k}} \nabla u_h \cdot \nabla w_h\ dx \quad = \quad \int_{\mathbf{\Omega^k}} f_h\ w_h\ dx \\ &\text{Local Matrix form} && \mathbf{M^k} \otimes u^{\mathbf{k}} \quad = \quad \mathbf{F^k}, \end{aligned} \tag{6.74}$$

where $\mathbf{M^k}$ is the ***local Mass*** matrix and $\mathbf{F}$ is the ***local Force*** vector. Here the notation "$\otimes$" refers to a matrix multiplication operator that depends on the matrix and the space dimension. See section 7.1.4 in Chapter 7 for more details. For problem 6.2, we have a different weak form and matrix discretization given by

$$\begin{aligned} &\text{Weak form on } \mathbf{\Omega^k} && \int_{\mathbf{\Omega}} \partial_t u_h\ w_h\ dx + \int_{\mathbf{\Omega}} C\ u_h\ w_h\ dx \quad = \quad -\int_{\mathbf{\Omega}} \nu \nabla u_h\ \nabla w_h\ dx \\ &\text{Local Matrix form} && \mathbf{M^k} \otimes \dot{u}^{\mathbf{k}} + \mathbf{C^k} \otimes u^{\mathbf{k}} \quad = \quad -\nu \mathbf{K^k} \otimes u^{\mathbf{k}}, \end{aligned} \tag{6.75}$$

where $\mathbf{C}$ is the linear or nonlinear ***local Advection*** matrix and $\mathbf{K}$ is the ***local Stiff*** matrix.

Local Mass matrix $\mathbf{M^k}$

In 1D, consider the scalar product on the element $\mathbf{\Omega^k}$, the mapping transformation and then the quadrature rule, we have,

$$\int_{\mathbf{\Omega^k}} u_h^{\mathbf{k}}(x)\ w_h^{\mathbf{k}}(x)\ dx \quad = \quad \int_{\Lambda^3} u_h^{\mathbf{k}}(\xi)\ w_h^{\mathbf{k}}(\xi)\ |J|^{\mathbf{k}}\ d\xi, \tag{6.76}$$

$$\left(\int_{\mathbf{\Omega^k}} u_h^{\mathbf{k}}(x)\ w_h^{\mathbf{k}}(x)\ dx\right)_{\text{GLL}} \quad \sim \quad \sum_{q=0}^{N}\sum_{i=0}^{N}\sum_{a=0}^{N} \rho_q^{\mathbf{k}}\ |J|^{\mathbf{k}}\ u_i^{\mathbf{k}} h_i(\xi)\ w_a^{\mathbf{k}} h_a(\xi), \tag{6.77}$$

where $\rho_q^{\mathbf{k}}$ are the GLL weights and ncp is a number combining the sum over the GLL points for the basis functions and the sum over the GLL points for the quadrature rule. The corresponding matrix to this elemental discretization is referred to as the ***local Mass*** matrix $\mathbf{M}^k$ in FEM and SEM. The bilinear form $a(h_i, h_a)$ is non zero only if h_i and h_a belong to the same element, this is due to the fact that each basis function is nonzero over a single element only.

Local Stiff matrix $\mathbf{K}^k$

Consider the first problem (6.1), the discretization of the left hand side on each element is

$$
\begin{aligned}
a(u_h, w_h)_{\mathbf{\Omega^k}} &= \int_{\mathbf{\Omega}^k} \nabla u_h^{\mathbf{k}}(x) \cdot \nabla w_h^{\mathbf{k}}(x)\, dx, && (6.78)\\
&\sim \int_{\Lambda^3} |J^{\mathbf{k}}|^{-1}\, \nabla u_h^{\mathbf{k}}(\xi) \cdot \nabla w_h^{\mathbf{k}}(\xi)\, d\xi, && (6.79)\\
&\sim \sum_{q=0}^{N}\sum_{i=0}^{N}\sum_{a=0}^{N} \rho_q^{\mathbf{k}}\, |J^{\mathbf{k}}|^{-1}\, \nabla\left(u_i^{\mathbf{k}} h_i(\xi)\right) \cdot \nabla\left(w_a^{\mathbf{k}} h_a(\xi)\right), && (6.80)
\end{aligned}
$$

where equation (6.80) can be further simplified with formulas of derivatives of the integrand polynomials.

The corresponding matrix to this elemental discretization is referred to as the ***local Stiff*** matrix $\mathbf{K}^k$ in FEM and SEM.

Local Force vector $\mathbf{F}^k$

Consider the first problem (6.1), the discretization of the right hand side on each element is

$$
\begin{aligned}
(f_h, w_h)_{\mathbf{\Omega^k}} &= \int_{\mathbf{\Omega^k}} f_h^{\mathbf{k}}(x)\, w_h^{\mathbf{k}}(x)\, dx, && (6.81)\\
&\sim \int_{\Lambda^3} |J^{\mathbf{k}}| f_h^{\mathbf{k}}(\xi)\, w_h^{\mathbf{k}}(\xi)\, d\xi, && (6.82)\\
&\sim \sum_{q=0}^{N}\sum_{i=0}^{N}\sum_{a=0}^{N} \rho_q^{\mathbf{k}}\, |J^{\mathbf{k}}|\, f_i^{\mathbf{k}} h_i(\xi)\, w_a^{\mathbf{k}} h_a(\xi). && (6.83)
\end{aligned}
$$

The corresponding vector to this elemental discretization is referred to as the ***local Force*** vector $\mathbf{F}^k$ in FEM and SEM.

Other Local matrices

The previous sections give a rough idea on how to discretize some integral terms that are present in a given problem. There are different types of integral terms but their discretization is accomplished in similar ways. For example, for the second problem we would obtain the linear or nonlinear ***local Advection*** matrix $\mathbf{C^k}$ coming from the integral term

$$
\int_{\mathbf{\Omega}^k} C\, u_h^k\, w_h^k\, dx. \qquad (6.84)
$$

Summary

The mesh corresponding to a spectral element method has to be a Gauss-Lobatto-based mesh. In elemental matrix notation, the first problem (6.1) can be written as

$$
\mathbf{M^k} \otimes u^{\mathbf{k}} = \mathbf{F^k}, \qquad (6.85)
$$

where the unknowns u^k correspond to the nodal basis coefficients on the approximate solution. Remember that the test functions w_h are chosen to be non-zero at only one global collocation point. The second problem (6.2) can be written in elemental matrix notation as

$$\mathbf{M^k} \otimes \dot{u}^{\mathbf{k}} + \mathbf{C^k} \otimes u^{\mathbf{k}} = -\nu \mathbf{K^k} \otimes u^{\mathbf{k}}. \tag{6.86}$$

6.7 Assembly

The assembly process uses the conformity between the element interfaces in an implicit manner, it involves a one-to-one matching between the unknowns of the elements sharing an interface. The ***global assembly*** operation is often referred to as ***direct stiffness summation***. This process constructs a continuous global expansion basis from the elemental basis functions. In practice, most operations are performed in a local fashion within each element and then the contributions are summed to form the global system. A mapping is needed in order to assemble the global system from the local system. It identifies the global node number of a local node within each element. Interior nodes may be independently numbered as global degrees of freedom. This assembly process depends tremendously on the topology of the mesh. When using an adaptive mesh refinement the assembly process needs to be modified for non-conforming elements.

All the elemental contributions need to be added together this is called ***the assembly of the global matrix***. In general, consider a local matrix $\mathbf{A^k}$, its global matrix $\mathbf{A}$ is

$$\mathbf{A} = \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \mathbf{A^k}, \tag{6.87}$$

where $\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}'$ represents the assembly summation. Figure 6.7 illustrates the process of direct summation to obtain a global system of algebraic equations.

As an example, we introduce the ***global Mass*** matrix $\mathbf{M}$. It is the result of the assembly process of the local Mass matrix $\mathbf{M^k}$ for every element. The discretization on the whole domain of the scalar product is

$$\int_{\mathbf{\Omega}} u_h^k(x)\, w_h^k(x)\, dx \sim \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \left(\int_{\mathbf{\Omega}^k} u_h^k(x)\, w_h^k(x)\, dx \right)_{\mathrm{GLL}} \tag{6.88}$$

$$\sim \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \mathbf{M}^k \otimes u^k. \tag{6.89}$$

All the local matrices of a problem are assembled to form algebraic systems to be solved. In parallel, this is an important aspect for processor communications.

The first problem (6.1) results in the following algebraic matrix system

$$\mathbf{M} \otimes u = \mathbf{F}. \tag{6.90}$$

The second problem (6.2) can be written as

$$\mathbf{M} \otimes \dot{u} + \mathbf{C} \otimes u = -\nu \mathbf{K} \otimes u. \tag{6.91}$$

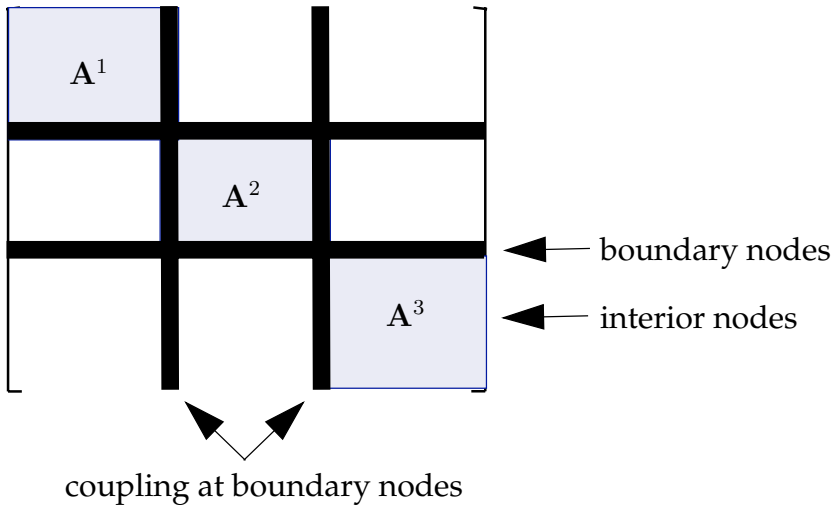


Figure 6.7: Schematic of the direct summation of local matrices $\mathbf{A^k}$ to form the global matrix A.

6.8 Why is the weak form important?

There are many bad and good reasons why the weak form is widely used in the FEM and SEM.

Historically, finite elements were developed by structural engineers in the sixties. It was very natural to use variational principles in analytical and earlier numerical methods, and hence the FEM should be discretizations of these variational principles. These variational formulas typically are expressed in terms of integral inner products. Although spectral elements are most commonly applied in fluid mechanics, and much has changed in the intervening forty years, it is nevertheless true that when a problem can be expressed in more than one way, the choice of representation is strongly steered by history.

Another bad reason is that the weak form simplifies convergence and uniqueness proofs for discrete algorithms.

A good reason for choosing the weak form is, as mentioned earlier in this chapter, it lowers the order of the highest derivative that must be computed. Thus, the weak form of a second order differential equation involves only first derivatives. This is advantageous in a couple of ways:

- In FEM, it becomes possible to solve second order differential equations using so-called tent or chapeau functions, which are piecewise linear, and thus the second derivative of these functions is zero. These basis functions are useless for solving a second order equation through any mean weighted residual method applied to the strong form. However, when applied to the weak form, piecewise linear basis functions give second order convergence.

- It is very helpful to reduce the order of derivatives in multiple space dimensions. Spectral elements are used primarily to cope with geometrical difficulty of some sort.

However, complicated geometry with curved elements have a nontrivial mapping (change-of-coordinates) that transforms the physical subdomain with its curvy sides into the unit square or cube where a tensor product spectral basis is applied. The coordinate transform implies that the differential equation in the computational coordinates has additional metric factors multiplying the derivatives. The crucial point is that the metric factors rise rapidly in number and complexity with the order of the derivative and the number of dimensions. Even the second derivative in 2D is a horrible mess and worse, it is a computationally expensive mess. The weak form implies that only first derivatives are needed for a second order differential equation and thus takes a lot of the pain out of the metric factors.

Another good reason to use the weak form is, that it is defined even when the strong form is not because of discontinuities.

The weak form simplifies the matching of subdomains. Indeed, the strong form of a second order differential equation requires explicitly matching both u and its first derivative at interdomain walls. In 1D, this is fairly easy, but in complicated geometry in 2D and 3D this matching is difficult and expensive. The weak form amazingly allows one to obtain an exponential rate of convergence as the number of unknowns in each element is increased even if only the function itself is explicitly matched.

Consequently, the weak forms and variational principles are very useful, and are used to the almost total exclusion of the strong form in finite elements and spectral elements.

6.9 Mesh generation

A first crucial step towards the accurate simulation of 3D problems with the SEM is the design of the mesh. There are many mesh generation techniques, such as advancing front, structured meshing and Delaunay; they all follow a ***bottom-up***[3] construction procedure [136]. Designing a good mesh can be a very difficult and demanding problem. In this section we will not discuss the specific details, but highlight some of the basic ingredients of mesh design, which are classical finite element results.

There are 2 possibilities of mesh for the spectral element method: triangle elements in 2D and tetrahedral elements in 3D, or, quadrilateral elements in 2D, and hexahedral elements in 3D. Tetrahedral elements are typical to the FEM, however they are not as popular with the SEM because of the tensorization product of the polynomial basis that is required to obtain a diagonal mass matrix.

The mesh is ***conforming*** if the 6 faces of each hexahedral elements match up exactly with the sides of neighboring elements.

A good mesh should take into account the major first and second order discontinuities in the problem, and the size of the elements should reflect the distribution of wave speeds if such waves are present in the model under study. The mapping between Cartesian points

[3]The ***bottom-up*** approach initially discretises the edges of the boundary representation into discrete segments which conform to the points of the boundary representation. Every surface of the boundary representation is then bounded by a set of discretised edges, and so the next step of the generation is to develop a surface discretisation in terms of quadrilateral elements. The generation process is finally completed by constructing elements in the interior of the domain which comply to the face and edge definitions constructed in the previous steps.

$X = (x, y, z)$ within a deformed hexahedral element $\mathbf{\Omega^k}$ and the master element are of the form

$$X(\xi, \eta, \zeta) = \sum_{a=1}^{n_a} N_a(\xi, \eta, \zeta) X_a. \tag{6.92}$$

The N_a shape functions are 3D tensorial products of Lagrange-Legendre basis functions of order n_a. The higher the order of the shape functions N_a, the more curved the physical elements can be. The final geometry of the curved hexahedral elements is also determined with the anchor points X_a. The behaviour of the Jacobian is controlled by the geometry of the mesh and is therefore a measure of the mesh quality.

To calculate the Jacobian, one needs to differentiate the mapping $X(\xi, \eta, \zeta)$ function with respect to ξ, η, ζ and obtain the respective partial derivatives:

$$\partial_\xi X(\xi, \eta, \zeta) = \sum_{a=1}^{n_a} \partial_\xi N_a(\xi, \eta, \zeta) X_a; \tag{6.93}$$

$$\partial_\eta X(\xi, \eta, \zeta) = \sum_{a=1}^{n_a} \partial_\eta N_a(\xi, \eta, \zeta) X_a; \tag{6.94}$$

$$\partial_\zeta X(\xi, \eta, \zeta) = \sum_{a=1}^{n_a} \partial_\zeta N_a(\xi, \eta, \zeta) X_a. \tag{6.95}$$

Finally, the Jacobian associated to this mapping is determined by

$$J = \frac{\partial X(\xi, \eta, \zeta)}{\partial(\xi, \eta, \zeta)} = \begin{pmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial x}{\partial \eta} & \frac{\partial x}{\partial \zeta} \\ \frac{\partial y}{\partial \xi} & \frac{\partial y}{\partial \eta} & \frac{\partial y}{\partial \zeta} \\ \frac{\partial z}{\partial \xi} & \frac{\partial z}{\partial \eta} & \frac{\partial z}{\partial \zeta} \end{pmatrix}. \tag{6.96}$$

Partial derivatives of the shape functions $\partial_\xi N_a$, $\partial_\eta N_a$ and $\partial_\zeta N_a$ are analytically determined in terms of the Lagrange-Legendre polynomials of degree 1 or 2 and their detivatives. One needs to ensure that the mapping from the physical to master element is unique and invertible. The mapping should be well defined and the Jacobian should ***never*** vanish.

Once the anchor points are specified, one can obtain the coordinate transformations, the first derivatives of the change of variables and the Jacobian as described above and perform all the elemental matrix calculations needed for the system to solve.

6.9.1 Quadrilateral elements

Each quadrilateral element is isomorphic to the square, its 4 corners are always used as anchors, but its side centres and its centre may be used as additional anchors. For simple elements with straight edges, only 4 control points (anchor points) suffice, whereas curved elements require 9 anchor points to describe the geometry accurately.

In 2D, the geometry of an element can be defined by its $n_a = 4$ anchor points X_a and shape functions $N_a(\xi, \eta)$ products of Lagrange polynomials h_{N,ξ_i} of degree $N = 1$ by

$$X(\xi, \eta) = \sum_{a=1}^{n_a} N_a(\xi, \eta) X_a. \tag{6.97}$$

Typically, the two Lagrange polynomials are of degree 1 for 2 anchor points or degree 2 for 3 anchor points, and are evaluated at $\xi_i = \pm 1$ and

$$h_{N,\xi_i=+1}(\xi) = \frac{1+\xi}{2} \qquad h_{N,\xi_i=-1}(\xi) = \frac{1-\xi}{2}. \tag{6.98}$$

Remember that the Lagrange-Legendre polynomials evaluated at a GLL point have a value of either 0 or 1. Any point in the physical element is given by the following relation:

$$\begin{aligned} X(\xi,\eta) &= \left(\frac{1-\xi}{2}\right)\left(\frac{1-\eta}{2}\right) X_1 + \left(\frac{1+\xi}{2}\right)\left(\frac{1-\eta}{2}\right) X_2 \\ &+ \left(\frac{1+\xi}{2}\right)\left(\frac{1+\eta}{2}\right) X_3 + \left(\frac{1-\xi}{2}\right)\left(\frac{1+\eta}{2}\right) X_4, \end{aligned} \tag{6.99}$$

where X_a are the four anchor points of coordinates (x_a, y_a).

Physical locations of the nodes. Consider a simple square element of length a. To find the points in the x and y direction, we need only know the maximum and minimum x values for that element in order to give the ξ values of the Gauss–Lobatto–Legendre points, i.e., the ith x value is

$$\begin{aligned} x_i &= \frac{1}{2}(x_{min} + x_{max}) + \frac{1}{2}(x_{max} - x_{min})\, \xi_{GLL}(i) \\ &= a\, \xi_{GLL}(i), \end{aligned} \tag{6.100}$$

$$\begin{aligned} y_j &= \frac{1}{2}(y_{min} + y_{max}) + \frac{1}{2}(y_{max} - y_{min})\, \eta_{GLL}(j) \\ &= a\, \eta_{GLL}(i). \end{aligned} \tag{6.101}$$

The first derivatives can easily be calculated in this case:

$$\begin{aligned} \frac{\partial x}{\partial \xi} &= a, & \frac{\partial x}{\partial \eta} &= 0, \\ \frac{\partial y}{\partial \eta} &= a, & \frac{\partial y}{\partial \xi} &= 0. \end{aligned} \tag{6.102}$$

The Jacobian is defined by the determinant of the first derivative functions, so for this type of simple square element of length a we have

$$|J| = \left|\frac{\partial(x,y)}{\partial(\xi,\eta)}\right| = a^2. \tag{6.103}$$

6.9.2 Hexahedral elements

The 3D generalization of the above is straightforward. Each hexahedral element is mapped into a reference cube. For a simple hexahedral element with straight faces, one needs eight corner nodes. By adding mid-side and centre nodes the number of anchors can become as large as 27-node hexahedral elements

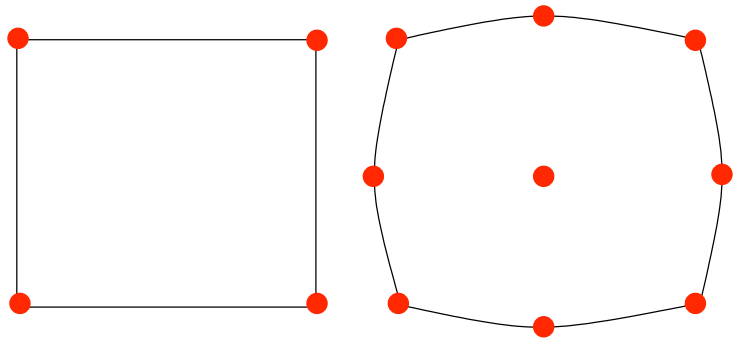


Figure 6.8: Schematic representations of a quadrilateral element with 4 anchor points and 9 anchor points.

In 3D, the geometry of an element can be defined by its $n_a = 8$ anchor points X_a and shape functions $N_a(\xi, \eta, \zeta)$ products of Lagrange polynomials h_{N,ξ_i} of degree $N = 1$ by

$$X(\xi, \eta, \zeta) = \sum_{a=1}^{n_a} N_a(\xi, \eta, \zeta) X_a. \tag{6.104}$$

Typically, the two Lagrange polynomials are of degree 1 for 2 anchor points or degree 2 for 3 anchor points, and are evaluated at $\xi_i = \pm 1$ and

$$h_{N,\xi_i=+1}(\xi) = \frac{1+\xi}{2} \qquad h_{N,\xi_i=-1}(\xi) = \frac{1-\xi}{2}. \tag{6.105}$$

Again, remember that the Lagrange-Legendre polynomials evaluated at a GLL point have a value of either 0 or 1. Any point in the physical element is given by the following relation:

$$\begin{aligned} X(\xi,\eta,\zeta) \;=\; & \left(\frac{1+\xi}{2}\right)\left(\frac{1+\eta}{2}\right)\left(\frac{1+\zeta}{2}\right) X_1 + \left(\frac{1+\xi}{2}\right)\left(\frac{1-\eta}{2}\right)\left(\frac{1+\zeta}{2}\right) X_2 \\ + \; & \left(\frac{1-\xi}{2}\right)\left(\frac{1-\eta}{2}\right)\left(\frac{1+\zeta}{2}\right) X_3 + \left(\frac{1-\xi}{2}\right)\left(\frac{1+\eta}{2}\right)\left(\frac{1+\zeta}{2}\right) X_4, \\ + \; & \left(\frac{1+\xi}{2}\right)\left(\frac{1+\eta}{2}\right)\left(\frac{1-\zeta}{2}\right) X_5 + \left(\frac{1+\xi}{2}\right)\left(\frac{1-\eta}{2}\right)\left(\frac{1-\zeta}{2}\right) X_6 \\ + \; & \left(\frac{1-\xi}{2}\right)\left(\frac{1-\eta}{2}\right)\left(\frac{1-\zeta}{2}\right) X_7 + \left(\frac{1-\xi}{2}\right)\left(\frac{1+\eta}{2}\right)\left(\frac{1-\zeta}{2}\right) X_8. \end{aligned} \tag{6.106}$$

where X_a are the four anchor points of coordinates (x_a, y_a).

Physical locations of the nodes. Consider a simple hexahedral element of length a. To find the points in the x, y and z direction, we need only know the maximum and minimum x values for that element in order to give the ξ values of the Gauss–Lobatto–Legendre points,

i.e., the ith x value is

$$\begin{aligned} x_i &= \frac{1}{2}(x_{min} + x_{max}) + \frac{1}{2}(x_{max} - x_{min})\ \xi_{GLL}(i) \\ &= a\ \xi_{GLL}(i), \end{aligned} \tag{6.107}$$

$$\begin{aligned} y_j &= \frac{1}{2}(y_{min} + y_{max}) + \frac{1}{2}(y_{max} - y_{min})\ \eta_{GLL}(j) \\ &= a\ \eta_{GLL}(i), \end{aligned} \tag{6.108}$$

$$\begin{aligned} z_l &= \frac{1}{2}(z_{min} + z_{max}) + \frac{1}{2}(z_{max} - z_{min})\ \zeta_{GLL}(l) \\ &= a\ \zeta_{GLL}(l). \end{aligned} \tag{6.109}$$

The first derivatives can easily be calculated in this case:

$$\begin{array}{lllllll} \dfrac{\partial x}{\partial \xi} = a, & \dfrac{\partial x}{\partial \eta} = 0, & \dfrac{\partial x}{\partial \zeta} = 0 \\ \dfrac{\partial y}{\partial \eta} = a, & \dfrac{\partial y}{\partial \xi} = 0, & \dfrac{\partial y}{\partial \zeta} = 0 \\ \dfrac{\partial z}{\partial \zeta} = a, & \dfrac{\partial z}{\partial \xi} = 0, & \dfrac{\partial z}{\partial \eta} = 0. \end{array} \tag{6.110}$$

The Jacobian is defined by the determinant of the first derivative functions, so for this type of simple square element of length a we have

$$|J|^{\mathbf{k}} = \left| \frac{\partial(x, y, z)}{\partial(\xi, \eta, \zeta)} \right| = a^3. \tag{6.111}$$

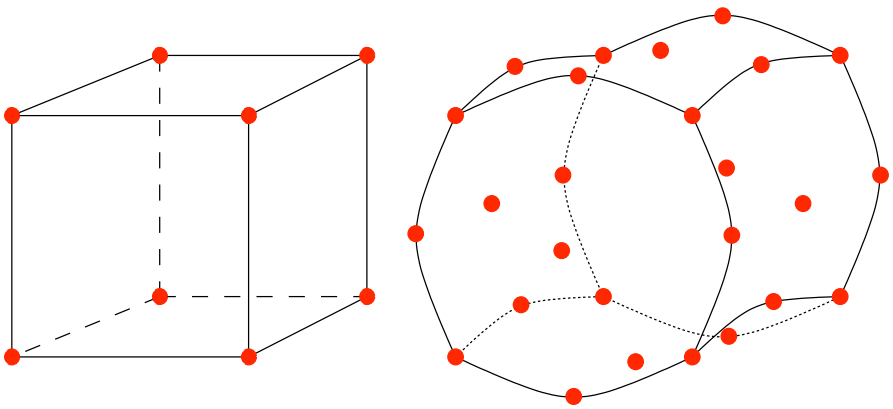


Figure 6.9: Schematic representations of a hexahedral element with 8 anchor points and 27 anchor points.

6.10 Time discretization for evolution problems

There are many time discretization schemes, explicit, implicit, semi-implicit multistep methods. Typically most packages have a few different methods available. Once a scheme has

been chosen, the problem can be discretized in time as well as in space. We have chosen a Runge–Kutta fourth order method for the time discretization.

The time discretization of the system

$$\dot{U} \;=\; \mathcal{A}U + \mathcal{F} \;=\; f(U,t), \tag{6.112}$$

is computed by an explicit ***fourth order*** Runge–Kutta method. Given an initial condition U_0, the solution U_{n+1} at time t_{n+1} is determined from the previous time t_n and solution U_n as follows:

$$\text{(RK4)} \quad \begin{cases} k_1 \;=\; f(U_n, t_n) \\ k_2 \;=\; f(U_n + a_{21}\Delta t\; k_1, t_n + c_2\Delta t) \\ k_3 \;=\; f(U_n + a_{31}\Delta t\; k_1 + a_{32}\Delta t\; k_2, t_n + c_3\Delta) \\ k_4 \;=\; f(U_n + a_{41}\Delta t\; k_1 + a_{42}\Delta t\; k_2 + a_{43}\Delta t\; k_3, t_n + c_4\Delta t) \\ \\ U_{n+1} \;=\; U_n + \Delta t\,(b_1k_1 + b_2k_2 + b_3k_3 + b_4k_4)\,. \end{cases} \tag{6.113}$$

To specify a particular method, one needs to provide the number of stages (here we use 4 stages and fourth order), and the coefficients a_{ij}, b_i and c_i. These data are usually arranged in a mnemonic device, known as a Butcher tableau:

0				
c_2	a_{21}			
c_3	a_{31}	a_{32}		
c_4	a_{41}	a_{42}	a_{43}	
	b_1	b_2	b_3	b_4

The Runge–Kutta method is consistent if

$$\sum_{j}^{i-1} a_{ij} \;=\; c_i \qquad \forall i = 2, \cdots, 4. \tag{6.114}$$

There are also other requirements, if we require the method to have a certain order p, meaning that the truncation error is $O(\Delta t^{p+1})$. These can be derived from the definition of the truncation error itself.

In all our numerical simulations, we have used the particular coefficients:

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

And hence, we have used in this thesis, the following fourth order Runge–Kutta method

with the spectral elements:

$$\text{(RK4)} \quad \begin{cases} k_1 &= f(U_n, t_n) \\ k_2 &= f(U_n + \dfrac{\Delta t}{2}\, k_1, t_n + \dfrac{\Delta t}{2}) \\ k_3 &= f(U_n + \dfrac{\Delta t}{2}\, k_2, t_n + \dfrac{\Delta t}{2}) \\ k_4 &= f(U_n + \Delta t\, k_3, t_n + \Delta t) \\ \\ U_{n+1} &= U_n + \dfrac{\Delta t}{6}\,(k_1 + 2k_2 + 2k_3 + k_4)\,. \end{cases} \qquad (6.115)$$

There are no large systems to solve or inversions of large matrices due to the explicit foward scheme used.

6.11 Filtering techniques

Many hyperbolic problems lead to discontinuous solutions, solutions of limited regularity or even solutions featuring sharp gradients. What happens with the SEM when discontinuities and shocks develop in the solution? What can we expect regarding accuracy and stability in such situations? In general, the accuracy of high order methods deteriorates in these situations, the pointwise error convergence of global approximations of discontinuous functions is at most first order. Furthermore, there is a loss of pointwise convergence at the point of discontinuity. On top of this, artificial and persistent oscillations are introduced around the point of discontinuity with possible propagation destroying the solution globally. In nonlinear equations, the situation can get much worse, the nonlinear interaction of the oscillations with the numerical solution will increase the energy of all the modes, thereby resulting in nonlinear instability, that is unbounded growth of high-frequency energy in time. This is due to the well known ***Gibbs phenomenon,*** which can potentially be treated by adding an artificial dissipation term that will stabilize the method; or, by filtering the high-frequency oscillations either in physical space or modal space. For a more complete analysis on filtering techniques and implementations , see the very informative books [137] and [138].

Filtering is a numerical technique more and more popular in spectral and spectral element methods for various reasons:

1. The numerical approximation can be stabilized and it results in a more robust method.
2. Filtering discontinuous functions can recover high-order accuracy in the smooth regions away from the discontinuity.

In modal space, the coefficients $\hat{u}$ of a spectral expansion for u can be multiplied by a filter function $\sigma(\eta)$. The filter function needs to have several properties: It is a infinitely differentiable function, and it should be equal to unity around the origin in order not to change the mean value of the filtered function u.
A filter is a real and even function $\sigma(\eta)$ ***of order*** p if:

1. $\sigma(0) = 1$, $\sigma^{(l)}(0) = 0$, with $1 \leq l \leq p-1$;

2. $\sigma(\eta) = 0$ for $|\eta| \geq 1$;

3. $\sigma(\eta) \in \mathcal{C}^{p-1}$, for $\eta \in (-\infty, \infty)$.

Let us introduce two commonly used filter functions:

The Exponential filter :

$$\sigma\left(\frac{n}{Nc}\right) = \begin{cases} 1, & 0 \leq n \leq N_c \\ \exp\left[-\alpha\left(\frac{n-N_c}{N-N_c}\right)^p\right], & N_c < n \leq N, \end{cases} \quad (6.116)$$

where p is the order of the filter, $\alpha = -\log\epsilon$ (ϵ is the machine zero), and N_c is the cutoff mode.

The Sharp cut-off filter

$$\sigma\left(\frac{n}{Nc}\right) = \begin{cases} 1, & 0 \leq n \leq N_c \\ 0, & N_c < n \leq N, \end{cases} \quad (6.117)$$

where N_c is the cutoff mode.

The numerical solution is approximated by

$$u(x) \sim u_N(x) = \sum_{n=1}^{N} u_n h_n(x), \quad (6.118)$$

where $h_n(x)$ are the Lagrange-Legendre polynomial basis functions. The nodal coefficients u_n can are also written as an expansion by

$$u_n = \sum_{i=1}^{N} u_i^* L_i(x), \quad (6.119)$$

where L_i are the Legendre polynomials obtained from the classical Jacobi polynomials. In terms of matrix notations, one has for the nodal coefficients

$$u = \mathbf{V}^T\mathbf{h}, \quad (6.120)$$

and the relation between the nodal coefficients u and modal coefficients u^*,

$$u = \mathbf{V}u^*, \quad (6.121)$$

where $\mathbf{V}$ is the Vandermonde matrix with respect to the Legendre polynomials (explicitly defined below)and $\mathbf{h}$ are the Lagrange-Legendre polynomials defined by $h_j(\xi_i) = \delta_{ij}$. Note that the modal coefficients u^* are recovered from the nodal coefficients $\bar{u}$ by taking the inverse of $\mathbf{V}$ by

$$u^* = \mathbf{V}^{-1}u. \quad (6.122)$$

For a nodal basis, a transformation to modal space employing the Vandermonde matrix $\mathbf{V}$ or a combination of the Legendre polynomial basis is required first. Then a filter can be applied on the new set of coefficients as a diagonal matrix. Finally, the coefficients are transformed back to nodal space. Let us define a filter vector f_i and filter diagonal $\mathbf{f} = \text{diag}(f_0, ..., f_N)$. The filtered modal values $\hat{u}^*$ are produced by the matrix operation

$$\hat{u}^* \quad = \quad \mathbf{f}u^* = \mathbf{f}\ \mathbf{V}^{-1}u, \tag{6.123}$$

thereby, the filtered ***nodal*** values can be recovered as

$$\hat{u} \quad = \quad \mathbf{V}\ \mathbf{f}\ \mathbf{V}^{-1}\bar{u} = \mathbf{F}u, \tag{6.124}$$

where $\mathbf{F}$ is called the filter matrix. In terms of sum notations it is equivalent to

$$\hat{u} \quad = \quad \sum_i \mathbf{F_{im}} u_m. \tag{6.125}$$

The Vandermonde matrix for the Legendre polynomials is

$$\mathbf{V_{ij}} \quad = \quad \begin{bmatrix} L_0(\xi_0) & L_1(\xi_0) & L_2(\xi_0) & \cdots & L_j(\xi_0) & \cdots & L_N(\xi_0) \\ \vdots & \vdots & \vdots & & \vdots & & \vdots \\ L_0(\xi_i) & L_1(\xi_i) & L_2(\xi_i) & \cdots & L_j(\xi_i) & \cdots & L_N(\xi_i) \\ \vdots & \vdots & \vdots & & \vdots & & \vdots \\ L_0(\xi_N) & L_1(\xi_N) & L_2(\xi_N) & \cdots & L_j(\xi_N) & \cdots & L_N(\xi_N) \end{bmatrix} \tag{6.126}$$

An alternative to the Vandermonde matrix $\mathbf{V}$ is the Legendre transform matrix derived in [139], given by

$$\mathbf{L_{ij}} = \begin{bmatrix} L_0(\xi_0) & L_1(\xi_0) & L_2(\xi_0) - L_0(\xi_0) & \cdots & L_j(\xi_0) - L_{j-2}(\xi_0) & \cdots & L_N(\xi_0) - L_{N-2}(\xi_0) \\ \vdots & \vdots & \vdots & & \vdots & & \vdots \\ L_0(\xi_i) & L_1(\xi_i) & L_2(\xi_i) - L_0(\xi_i) & \cdots & L_j(\xi_i) - L_{j-2}(\xi_i) & \cdots & L_N(\xi_i) - L_{N-2}(\xi_i) \\ \vdots & \vdots & \vdots & & \vdots & & \vdots \\ L_0(\xi_N) & L_1(\xi_N) & L_2(\xi_N) - L_0(\xi_N) & \cdots & L_j(\xi_N) - L_{j-2}(\xi_N) & \cdots & L_N(\xi_N) - L_{N-2}(\xi_N) \end{bmatrix} \tag{6.127}$$

The main reason for constructing the Legendre transform matrix as above instead of the Vandermonde matrix is simple. In this case, only the first two modes L_0 and L_1 affect the element boundary values, whereas the remainder of the modes only affect the element interior values. Therefore, the filter can be applied in an element-by-element sense without violating the $\mathcal{C}^0$ continuity condition at the element interfaces. These 2 filtering approaches preserve the end-values and thus the inter-element continuity is not affected.

In 2D, the filter is applied as follows:

$$\hat{\hat{u}} = \mathbf{F}u\mathbf{F}^T. \tag{6.128}$$

In terms of sum notations, it is equivalent to

$$\hat{u} = \sum_i \sum_j \mathbf{F_{im}}\mathbf{F_{nj}}u_{mn}. \tag{6.129}$$

In 3D, the filter is applied as follows:

$$\hat{u} = \mathbf{F}\Big((u\mathbf{F}^T) \cdot_{yz} \mathbf{F}^T\Big). \tag{6.130}$$

Again in terms of sum notations, this is equivalent to

$$\hat{u} = \sum_i \sum_j \sum_k \mathbf{F_{im}}\mathbf{F_{nj}}\mathbf{F_{pk}}u_{mnp}. \tag{6.131}$$

How does one choose a filter? Should the filter be applied once or more per time step, or once every several time steps? What is the effect of applying a filter repeatedly on the accuracy of the approximation?

Non-idempotent filters tend to zero out all but the first couple of modes, and the effect results in ***staircasing*** of the numerical solution. In comparison, an idempotent filter only zeros out the highest modes $n > N_c$.

The Gibbs oscillations may look bad, but surprisingly, they do not destroy the attractive properties of the scheme. The highly oscillatory result contains the information needed to reconstruct a spectrally accurate solution. Several model problems confirm Lax's statement [140]: information is contained in the oscillations associated with high order schemes, and high-order schemes retain more information than lower-order schemes.

Shu and Wang [141] recovered spectral accuracy for the nonlinear Burgers equation where discontinuity develops and moves around the domain. There is a very recent procedure that removes the Gibbs phenomenon completely, to obtain exponential accuracy in the maximum norm in any interval of analyticity, based on the Fourier or Gegenbauer series of a discontinuous but piecewise analytic function. For details, we refer to the review paper [142, 143].

6.12 Spectral elements and parallelization

One of the most attractive feature of the spectral element method, is its extremely good scalability on parallel computers. Clusters or grids of computers have a distributed memory architecture. The standard approach with parallel machines with distributed memory in a portable way is to use ***message passing interfaces*** (MPI). When using an explicit time discretization (Runge-Kutta fourth order in this project), the SEM algorithm consists of small local matrix products in each element. Therefore, processors spend most of their time doing actual calculations, and there is only a small amount of time in the communication step. In this light, we see that the SEM is not very sensitive to the speed of the network connecting

different processors, which make this method highly suitable to run on clusters or grids of computers.

Practically, one needs to split the mesh into as many domains as the number of available processors. Calculations can be performed locally on each processor on the elements contained in its corresponding domain. Then, one communication phase is required at each timestep for the assembly process. MPI communication tables that contain the sequence of messages that needs to be exchanged amongst the domains at each timestep need to be created only once and for all when the mesh is built.

In the spectral multi-domain method, the $\mathcal{C}^0$ and $\mathcal{C}^1$ boundary conditions at the interface of the elements have to be enforced explicitly. In contrast, the spectral element method uses the variational principle to guarantee $\mathcal{C}^0$ and $\mathcal{C}^1$ (weakly) continuity at the interface, which makes a parallel implementation more convenient.

In 2D, the computational cost grows proportionally to $N_E N_{GLL}^2$, remember that N_E is the number of elements and N_{GLL} is related to the interpolation polynomial order N by $N_{GLL} = N + 1$. The communication cost grows proportionally to $N_E^{1/2} N_{GLL}$. The speedup S is the ratio of the computational time between the serial and parallel codes, it can roughly be approximated (see [144]) by:

$$S = \frac{p}{(1 + N_E^{-1/2} N_{GLL}^{-2})}, \tag{6.132}$$

where p is the number of processors. This rough estimate shows that the communication cost increases only linearly with the method's order, whereas its computational cost increases cubically, which gives a quadratic ratio between the two costs. In comparison, high-order finite-difference methods have a quadratic increase of the communication cost with the order of the method, because of the number of neighboor points that must pass between processors increases. In the SEM, only edge points exchange information across elements.

6.13 Adaptive mesh refinement

Conforming elements require the implementation of interface conditions that are not too difficult. However, using non-conforming grids is very appealing as their use allows for parallel generation of meshes, adaptive mesh refinments and fast and independent solvers. There are various ways to extend the conforming formulation to include non-conforming elements, see [145]. Here is a of a few of them:

- ***Pointwise matching:*** This regains the lost $\mathcal{C}^0$ continuity at the non-conforming interfaces by enforcing pointwise projection of the unknowns.

- ***Mortar element method:*** In this method the lost $\mathcal{C}^0$ continuity is not regained but the jump at the non-conforming interfaces minimized by enforcing a weighted-integral matching.

- ***Robin interface conditions:*** This method is very recent [146] and is based on Schwarz type approaches that allow for the use of Robin interface conditions on non-conforming grids.

- ***The DARe technique:*** This procedure is based on an interpolating scheme to maintain continuity between elements. It is used in the package GASpAR in a fully dynamic adaptivity context [130].

6.14 Available SEM packages

There are now several spectral element packages available:

- in Matlab: ***SEMLAB*** available at
  http://www.gps.caltech.edu/~ampuero/software.html,
- in Fortran:
  SPECFEM3D available at
  http://www.geodynamics.org/cig/software/packages/seismo/specfem3d-globe,
  SEM2DPACK available at
  http://www.gps.caltech.edu/~ampuero/software.html,
  shallow_water, available at
  http://frederic.dupont8.free.fr/science/download.html#spoc,
  SEPRAN, available at
  http://ta.twi.tudelft.nl/sepran/sepran.html.
- in C++: ***GASpAR*** available at http://www.image.ucar.edu/TNT/Software/GASpAR/

SEMLAB: This package uses the spectral element method for 1D and 2D SH seismic wave propagation.

SPECFEM3D: This package simulates 3D global and regional seismic wave propagation in parallel but is very specific to this problem.

SEM2DPACK: This package uses the 2D Spectral Element Method for seismic wave propagation and earthquake dynamics. It is ideal for realistic 2D models (e.g. sedimentary basins, non-planar faults, heterogeneous or non-linear media).

shallow_water: This package is based on a PhD thesis by Frederic Dupont McGill University in Montreal "Comparison of numerical methods for modelling ocean circulation in basins with irregular coasts". It is written in parallel and is very specific to this problem and minimally documented.

SEPRAN: This package is a spectral element library written in Fortran 77, it does not have adaptive mesh refinement but can be used in parallel and has extensive user guides.

GASpAR: This is an adaptive spectral element code for geophysics and astrophysics. It has a lot of documentation and examples, see [130]. However, it does not seem to be very portable, it compiles with the Protland group compiler (PGI) (commercial) but does not seem to compile with the free GNU compiler collection (gcc). It would take a lot of work to correct all the errors and change the code to make it portable to gcc.

6.15 Conclusion

In this Chapter, we have presented an overview of the theory of the spectral element method. While the theory contains high levels of functional analysis and may be somewhat off-putting, this method has many successes in many different fields and offers great advantages over other numerical methods. The SEM combines the theory of spectral and pseudo-spectral methods for high order polynomials and the variational formulation of finite elements and the associated geometric flexibility.

The variational formulation is applied to the problem at hand and the weak formulation is then obtained. Space is divided into a number of elements, and the solution is written with local Lagrange–Legendre basis functions that are non-zero over a couple of elements. The spectral element discretization of the problem reduced to its weak form, results in elemental matrix forms of the problem. After the assembly process, one can obtain a global system of algebraic equations of the problem (typically sparse matrices for conforming elements) to solve. For explicit time stepping scheme, such a Runge–Kutta fourth order, there are no full matrix (non-sparse) to invert as the Mass matrix is diagonal due to the choice of the GLL quadrature.

We have introduced filtering techniques that can be useful for the SEM as ***stabilization*** techniques and as recovering high-accuracy in the smooth region away from any present discontinuity in the solution.

Finally, we have discussed the extremely good scalability of the SEM on parallel computers.

The art of doing mathematics consists in finding that special case which contains all the germs of generality.

David Hilbert (1862–1943)

7

The Spectral Element Method for the wave equation in 1D and 3D

This chapter applies the spectral element method (SEM) to a wave equation in 1D and 3D. The purpose here is to illustrate the application of the method described in a more theoretical framework in chapter 6 to a concrete problem, and show the consequent numerical results.

The wave equation is typically a second-order linear partial differential equation that describes the propagation of a variety of waves, such as sound waves, light waves and water waves. It arises in fields such as acoustics, electromagnetics, and fluid dynamics. The following wave equation belongs to the class of hyperbolic problems:

$$\partial_{tt} u = c^2 \nabla^2 u, \tag{7.1}$$

where ∇^2 is the Laplacian and where c is a fixed constant equal to the propagation speed of the wave. There is a very simple general solution to the 1 dimensional wave equation. If we now define 2 new variables by

$$v = r - ct \tag{7.2}$$

$$w = r + ct, \tag{7.3}$$

the wave equation is changed into

$$\frac{\partial^2 u}{\partial v \partial w} = 0. \tag{7.4}$$

General solutions of equation 7.4 are of the form

$$u(v, w) = F(v) + G(w), \tag{7.5}$$

which is equivalent to

$$u(x, t) = F(x - ct) + G(x + ct). \tag{7.6}$$

General solutions of the 1D wave equation are sums of a left traveling function F and a right traveling function G.

Spherical waves are waves whose amplitude depends only upon the radial distance r from a central point source. For such waves, the three-dimensional wave equation takes the form

$$\partial_{tt}u = c^2\left(\partial_{rr}u + \frac{2}{r}\partial_r u\right). \tag{7.7}$$

Note that we can rewrite equation 7.7 as

$$\partial_{tt}(ru) = c^2\Big(\partial_{rr}(ru)\Big), \tag{7.8}$$

so, the quantity ru satisfies the 1D wave equation. Therefore, the general solution for spherical wave equations takes the form

$$u(r,t) = \frac{F(r-ct)+G(r+ct)}{r}. \tag{7.9}$$

To illustrate the spectral element method, we choose a 3D exact spherical solution u_r with no source term of the form,

$$u_r(r,t) = \frac{e^{-(r-t)^2} - e^{-(r+t)^2}}{r}, \tag{7.10}$$

where $r = \sqrt{x^2+y^2+z^2}$. The speed of the wave has been set to $c = 1$ here. The motivation behind this particular solution is a numerically well-behaved solution everywhere on the domain $\mathbf{\Omega} = [-L, L]$. Note that for this particular solution, all the following limits are finite:

$$\lim_{r\to 0} u(r,t) = \frac{4}{e^{t^2}}; \tag{7.11}$$

$$\lim_{r\to \infty} u(r,t) = 0; \tag{7.12}$$

$$\lim_{t\to 0} u(r,t) = 0; \tag{7.13}$$

$$\lim_{t\to \infty} u(r,t) = 0. \tag{7.14}$$

We choose to implement the spectral element method to illustrate our 1D and 3D examples with Matlab. Matlab is an interpreted language and is ideal for developing and testing the SEM, in particular with fast and simple matrix vectorization calculations and relatively easy computer graphic visualization. See section 9.4.1 for more explanations on our choice of the Matlab language for this project.

7.1 Hyperbolic system first order in space and time in 1D

Remember from section 5.4.2 , that the BSSN system is hyperbolic and more importantly, first order in time. The wave equation presented in 7.1, however, is second order in time and in space. We therefore convert the wave equation into a hyperbolic system first order in time and space to deal with a simplified system as close as possible as the BSSN system.

7.1.1 Wave equation with source term

The wave equation on the domain $x \in [-L, L]$ is

$$\partial_{tt} u - \partial_{xx} u = S(x,t). \tag{7.15}$$

Ultimately we want to look at the 3D version of the problem with Sommerfeld-like boundary conditions, we therefore adapt the 1D wave equation to have this kind of solution. This is why we introduce a source term. This equation is second order in space and in time. One needs an initial condition $u_0 = u(x, t_0)$ and boundary conditions. We use the Sommerfeld-like absorbing or non-reflecting boundary conditions so that

$$\partial_x u(L,t) + \partial_t u(L,t) = 0; \tag{7.16}$$

$$\partial_x u(-L,t) - \partial_t u(-L,t) = 0. \tag{7.17}$$

Exact solution To evaluate the spectral element method accuracy we will compare the numerical solution to an exact solution. The following solution u_r, is an exact solution with no source term in 3D only, where $r = \sqrt{x^2 + y^2 + z^2}$:

$$u_r(r,t) = \frac{e^{-(r-t)^2} - e^{-(r+t)^2}}{r}. \tag{7.18}$$

If we take $r = x$ we obtain a 1D version,

$$u_x(x,t) = \frac{e^{-(x-t)^2} - e^{-(x+t)^2}}{x}. \tag{7.19}$$

However, equation 7.19 does not satisfy the wave equation, we therefore need to introduce a source term $S(x,t)$ so that:

$$\begin{aligned} S(x,t) &= \partial_{tt} u_x - \partial_{xx} u_x \\ &= \frac{2}{x^3}\Big\{2\Big[x e^{-(-x+t)^2}(-x+t) + x e^{-(x+t)^2}(x+t)\Big] \\ &\qquad - e^{-(-x+t)^2} + e^{-(x+t)^2}\Big\}. \end{aligned} \tag{7.20}$$

Now the function u_x is the solution of the inhomogeneous wave equation

$$S(x,t) = \partial_{tt} u_x - \partial_{xx} u_x. \tag{7.21}$$

From now on, we will denote $u_x = u$ to simplify notations.

Initial and boundary conditions

The initial condition of the solution for $t_0 = 0$ is

$$u(x,0) = u_0(x) = 0. \tag{7.22}$$

For the derivatives, we have

$$\partial_x u(x,0) = u_{x0}(x) = 0, \tag{7.23}$$
$$\partial_t u(x,0) = u_{t0}(x) = 4e^{-x^2}. \tag{7.24}$$

The absorbing boundary conditions are not exactly verified in 1D, therefore, we determine the function $b_1(x,t)$ and $b_2(x,t)$ so that:

$$\partial_x u(L,t) + \partial_t u(L,t) = b_1(L,t) \tag{7.25}$$
$$\partial_x u(-L,t) - \partial_t u(-L,t) = b_2(L,t) = -b_1(L,t). \tag{7.26}$$

The boundary function $b_1(L,t)$ has the following properties:

$$\lim_{L\to\infty} b_1(L,t) = 0; \tag{7.27}$$
$$\lim_{t\to\infty} b_1(L,t) = 0. \tag{7.28}$$

In view of the above limits, the right hand sides of equation (7.25) and (7.26) go to 0 exponentially for a large enough value of L. In other words, the boundary conditions are only approximate, but a sufficiently distant outer boundary can always be chosen such that they are correct to some required accuracy, for a certain time. Therefore the boundary conditions are ***numerically*** verified for this particular choice of source term and exact solution.

7.1.2 System of 3 unknowns: strong formulation

The BSSN system is a strongly hyperbolic system of order 1 in time and of order 2 in space. Thereby, we present the wave equation reformulated as a hyperbolic system of first order in time. Let's define $u_1 = u$ and then introduce 2 more variables,

$$u_2 = \partial_x u_1; \tag{7.29}$$
$$u_3 = \partial_t u_1. \tag{7.30}$$

The wave equation can be rewritten as a system

$$\begin{cases} \partial_t u_1 = u_3 \\ \partial_t u_2 = \partial_x u_3 \\ \partial_t u_3 = \partial_x u_2 + S(x,t) \end{cases}. \tag{7.31}$$

The system (7.31) can be written in matrix form

$$\partial_t U = \mathcal{A}U + F, \tag{7.32}$$

where $U = (u_1\ u_2\ u_3)^T$, $F = (0\ 0\ S(x,t))^T$, and

$$\mathcal{A} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & \partial_x \\ 0 & \partial_x & 0 \end{pmatrix}. \tag{7.33}$$

The initial conditions on u_1, u_2 and u_3 are given in terms of the initial condition u_0, its initial first space derivative $(\partial_x u)_0$ and its initial first time derivative $(\partial_t u)_0$. For the exact solution we are looking at we have:

$$u_2(x,0) \;=\; u_{x0}(x) \;=\; 0, \tag{7.34}$$

and

$$u_3(x,0) \;=\; u_{t0}(x) \;=\; -12e^{-x^2}. \tag{7.35}$$

The absorbing boundary conditions (7.25) and (7.26) translate into a relation between the two introduced variables,

$$u_2(L,t) + u_3(L,t) \;=\; b_1(x,t), \tag{7.36}$$
$$u_2(-L,t) - u_3(-L,t) \;=\; b_2(x,t). \tag{7.37}$$

For the exact solution we are looking at, these two relations (7.36) and (7.37) are not exact, but only approximately verified.

7.1.3 Weak formulation

We apply the variational formulation to the system (7.31) by multiplying the unknowns with a test function w, and integrating over the whole domain $\mathbf{\Omega}$. We obtain the weak formulation

$$\int_{\mathbf{\Omega}} \partial_t u_1 \; w \; dx \;=\; \int_{\mathbf{\Omega}} u_3 \; w \; dx, \tag{7.38}$$
$$\int_{\mathbf{\Omega}} \partial_t u_2 \; w \; dx \;=\; \underbrace{\int_{\mathbf{\Omega}} \partial_x u_3 \; w \; dx}_{\mathrm{I}_1}, \tag{7.39}$$
$$\int_{\mathbf{\Omega}} \partial_t u_3 \; w \; dx \;=\; \underbrace{\int_{\mathbf{\Omega}} \partial_x u_2 \; w \; dx}_{\mathrm{I}_2} + \int_{\mathbf{\Omega}} S(x,t) \; w \; dx \; . \tag{7.40}$$

Weak formulation version 1

The boundary conditions (7.36) and (7.37) are introduced when integrating I_1 or/and I_2 by parts,

$$\mathrm{I}_1 \;=\; \Big[u_3 \; w\Big]_{-L}^{L} - \int_{\mathbf{\Omega}} u_3 \; \partial_x w \; dx, \tag{7.41}$$
$$\;=\; -u_2(L,t)w(L) - u_2(-L,t)w(-L) - \int_{\mathbf{\Omega}} u_3 \; \partial_x w \; dx, \tag{7.42}$$

and

$$\mathrm{I}_2 \;=\; \Big[u_2 \; w\Big]_{-L}^{L} - \int_{\mathbf{\Omega}} u_2 \; \partial_x w \; dx, \tag{7.43}$$
$$\;=\; -u_3(L,t)w(L) + u_3(-L,t)w(-L) - \int_{\mathbf{\Omega}} u_2 \; \partial_x w \; dx. \tag{7.44}$$

In this version of the weak formulation we choose to integrate by parts both I_1 and I_2. In the next section we present an alternative weak formulation where we integrate by parts only I_2 in order to introduce the boundary conditions. The two versions differ numerically only slightly in amplitudes of the order of the numerical error. When we consider the 2D version of this system, it is not practical to integrate the 2D equivalent of I_1 as it would require the component in the x direction of the unknown u_3 rather than the unknown itself. In 1D both weak formulations are equivalent because the x-component of the unknown is the unknown itself.

When integrating by parts both integrals I_1 and I_2, the final weak formulation (***version 1***) of the system is

$$\int_{\mathbf{\Omega}} \partial_t u_1 \; w \; dx = \int_{\mathbf{\Omega}} u_3 \; w \; dx, \tag{7.45}$$

$$\int_{\mathbf{\Omega}} \partial_t u_2 \; w \; dx = -u_2(L,t)w(L) - u_2(-L,t)w(-L) - \int_{\mathbf{\Omega}} u_3 \; \partial_x w \; dx, \tag{7.46}$$

$$\int_{\mathbf{\Omega}} \partial_t u_3 \; w \; dx = -u_3(L,t)w(L) + u_3(-L,t)w(-L) - \int_{\mathbf{\Omega}} u_2 \; \partial_x w \; dx + \int_{\mathbf{\Omega}} S(x,t) \; w \; dx. \tag{7.47}$$

We need to define the space of solutions and test functions. Let's define the space of measurable functions $\mathcal{V} = \mathcal{L}^2(\mathbf{\Omega})$ and the Hilbert space

$$\mathcal{W} = \mathcal{H}^1(\mathbf{\Omega}) = \left\{ w \in \mathcal{L}^2(\mathbf{\Omega}) \text{ and } \partial_x w \in \mathcal{L}^2(\mathbf{\Omega}) \right\}. \tag{7.48}$$

All 3 unknowns are defined in the space $\mathcal{V}$, whereas the test functions w is defined in $\mathcal{W}$. Ultimately this means that the solution $u = u_1$ to the original wave equation belongs to the space

$$\mathcal{U} = \left\{ u \in \mathcal{L}^2(\mathbf{\Omega}), \; \partial_x u \in \mathcal{L}^2(\mathbf{\Omega}), \text{ and } \partial_t u \in \mathcal{L}^2(\mathbf{\Omega}) \right\}. \tag{7.49}$$

Weak formulation version 2

In preparation for solving this problem in higher dimensions, we present an alternative weak formulation that is more practical in 2D and 3D. The boundary conditions (7.36) and (7.37) are introduced when integrating I_2 only by parts. As mentioned in the previous subsection the motivation comes from the fact that integrating the 2D equivalent of I_1 would require the x-component of the unknown u_3 and would require some kind of projection operator.

When integrating by parts only integral I_2, the final weak formulation (***version 2***) of the system is

$$\int_{\mathbf{\Omega}} \partial_t u_1 \; w \; dx = \int_{\mathbf{\Omega}} u_3 \; w \; dx, \tag{7.50}$$

$$\int_{\mathbf{\Omega}} \partial_t u_2 \; w \; dx = \int_{\mathbf{\Omega}} \partial_x u_3 \; w \; dx \tag{7.51}$$

$$\int_{\mathbf{\Omega}} \partial_t u_3 \; w \; dx = -u_3(L,t)w(L) - u_3(-L,t)w(-L) - \int_{\mathbf{\Omega}} u_2 \; \partial_x w \; dx + \int_{\mathbf{\Omega}} S(x,t) \; w \; dx. \tag{7.52}$$

In this alternative weak formulation u_1, u_2 and u_3 are defined in the space $\mathcal{V}$, whereas the test function w is in $\mathcal{W}$.

7.1.4 Domain Discretization

The domain $\mathbf{\Omega}$ is decomposed into N_E sub-domains $\mathbf{\Omega^k}$ such that

$$\bar{\mathbf{\Omega}} = \bigcup_{\mathbf{k}=0}^{K} \bar{\mathbf{\Omega}}^{\mathbf{k}}, \quad \forall \mathbf{k},\ l \quad \mathbf{\Omega}^{\mathbf{k}} \bigcap_{\mathbf{k}\neq l} \mathbf{\Omega}^{l} = 0, \tag{7.53}$$

where $\bar{\mathbf{\Omega}}$ is the closure of the domain $\mathbf{\Omega}$. The weak formulation is applied in each subdomain $\mathbf{\Omega^k}$ individually, ***version 1*** becomes

$$\int_{\mathbf{\Omega^k}} \partial_t u_1^{\mathbf{k}}\, w_1^{\mathbf{k}}\, dx = \int_{\mathbf{\Omega^k}} u_3^{\mathbf{k}}\, w_1^{\mathbf{k}}\, dx, \tag{7.54}$$

$$\int_{\mathbf{\Omega^k}} \partial_t u_2^{\mathbf{k}}\, w_2^{\mathbf{k}}\, dx = \Big(- u_2^{\mathbf{k}}(L,t) w_2^{\mathbf{k}}(L) - u_2^{\mathbf{k}}(-L,t) w_2^{\mathbf{k}}(-L) \Big)_{\text{bdy}} - \int_{\mathbf{\Omega^k}} u_3^{\mathbf{k}}\, \partial_x w_2^{\mathbf{k}}\, dx, \tag{7.55}$$

$$\int_{\mathbf{\Omega^k}} \partial_t u_3^{\mathbf{k}}\, w_3^{\mathbf{k}}\, dx = \Big(- u_3^{\mathbf{k}}(L,t) w(L) - u_3^{\mathbf{k}}(-L,t) w_3^{\mathbf{k}}(-L) \Big)_{\text{bdy}} - \int_{\mathbf{\Omega^k}} u_2^{\mathbf{k}}\, \partial_x w_3^{\mathbf{k}}\, dx + \int_{\mathbf{\Omega^k}} S^{\mathbf{k}}\, w_3^{\mathbf{k}}\, dx \tag{7.56}$$

Note that the boundary terms in (7.55) and (7.56) are relevant only for the 2 elements that share a node with the boundary $x = -L$ and $x = L$. If the source term $S(x,t)$ is given as a function of x and t, then $S^{\mathbf{k}}$ is nothing but the restriction of $S(x,t)$ to the subdomain $\mathbf{\Omega^k}$.

In terms of the second version of the weak formulation, we write in each subdomain $\mathbf{\Omega^k}$ individually and ***version 2*** becomes

$$\int_{\mathbf{\Omega^k}} \partial_t u_1^{\mathbf{k}}\, w_1^{\mathbf{k}}\, dx = \int_{\mathbf{\Omega^k}} u_3^{\mathbf{k}}\, w_1^{\mathbf{k}}\, dx, \tag{7.57}$$

$$\int_{\mathbf{\Omega^k}} \partial_t u_2^{\mathbf{k}}\, w_2^{\mathbf{k}}\, dx = \int_{\mathbf{\Omega^k}} \partial_x u_3^{\mathbf{k}}\, w_2^{\mathbf{k}}\, dx, \tag{7.58}$$

$$\int_{\mathbf{\Omega^k}} \partial_t u_3^{\mathbf{k}}\, w_3^{\mathbf{k}}\, dx = \Big(- u_3^{\mathbf{k}}(L,t) w(L) - u_3^{\mathbf{k}}(-L,t) w_3^{\mathbf{k}}(-L) \Big)_{\text{bdy}} - \int_{\mathbf{\Omega^k}} u_2^{\mathbf{k}}\, \partial_x w_3^{\mathbf{k}}\, dx + \int_{\mathbf{\Omega^k}} S^{\mathbf{k}}\, w_3^{\mathbf{k}}\, dx \tag{7.59}$$

Important remark: Normally, if we integrate by parts on each element separately there would be a number of boundary terms at the interior boundaries. For the exact solutions these terms cancel in pairs, but the spectral elements are only $\mathcal{C}^0$ and the subdomain wall boundary terms do not cancel. This difference arises from the fact that discretization and differentiation of the weak form do not commute for $\mathcal{C}^0$ spectral elements. These extra terms go to zero in the limit so the spectral element strategy is to ignore them which is equivalent

to performing an integration by parts first and discretization second. This is referred to as the ***variational crime*** [120].

In each subdomain, the solutions $u_1^{\mathbf{k}}, u_2^{\mathbf{k}}$ and $u_3^{\mathbf{k}}$ are expanded into cardinal function series of polynomial order N,

$$\forall u_1^{\mathbf{k}} \in \mathcal{V}_h,\ u_1^{\mathbf{k}}(x,t) = \sum_{m=0}^{m=N} u_{1m}^{\mathbf{k}}(t)\ h_m(x), \tag{7.60}$$

$$\forall u_2^{\mathbf{k}} \in \mathcal{V}_h,\ u_2^{\mathbf{k}}(x,t) = \sum_{m=0}^{m=N} u_{2m}^{\mathbf{k}}(t)\ h_m(x), \tag{7.61}$$

$$\forall u_3^{\mathbf{k}} \in \mathcal{V}_h,\ u_3^{\mathbf{k}}(x,t) = \sum_{m=0}^{m=N} u_{3m}^{\mathbf{k}}(t)\ h_m(x). \tag{7.62}$$

The elemental Lagrangian interpolants $h_m(\xi)$ are chosen as basis functions and $u_{1m}^{\mathbf{k}}(t) = u_{1m}^{\mathbf{k}}(x,t)$ are the nodal basis coefficients. The space $\mathcal{V}_h = \mathcal{V} \bigcup \mathbb{P}_{N,k}(\mathbf{\Omega})$ is taken to be a subspace of $\mathcal{V}$ which consists of all piecewise high order polynomials of degree less than or equal to N defined on $\mathbf{\Omega^k}$. Furthermore, we have the definition

$$\mathbb{P}_{N,K}(\mathbf{\Omega}) = \left\{ \theta \in \mathcal{L}^2(\mathbf{\Omega}), \quad \theta_{|_{\mathbf{\Omega^k}}} \in \mathbb{P}_N(\mathbf{\Omega^k}) \right\}. \tag{7.63}$$

The test function w is selected to be the same as the shape functions $h_m(x)$ used for the unknowns $u_1^{\mathbf{k}}, u_2^{\mathbf{k}}$ and $u_3^{\mathbf{k}}$, and therefore

$$w_1^{\mathbf{k}}(x) = \sum_{a=0}^{a=N} h_a(x), \tag{7.64}$$

$$w_2^{\mathbf{k}}(x) = \sum_{a=0}^{a=N} h_a(x), \tag{7.65}$$

$$w_3^{\mathbf{k}}(x) = \sum_{a=0}^{a=N} h_a(x). \tag{7.66}$$

Note that here, the same test functions are used for each variable, but they could in principle be different if, for example, one was to choose a different polynomial order for each unknown. Recall from the discussion in the previous chapter, that, in the case of the Navier-Stokes equations the velocity u can be written with polynomial order N and the pressure with order $N-2$. This avoids spurious modes and hence the test functions associated with u and p respectively are different.

On each element $\mathbf{\Omega^k}$ there are $N_{GLL} = (N+1)$ nodal points but in total there are $N_g = N \times N_E + 1$ global nodal points. One needs to create a global numbering function that keeps track of local and global nodes on the domain $\mathbf{\Omega}$. Let $\mathcal{I}$ denote global indices which are functions of the element index $\mathbf{k}$ and the index a within each element, for example,

$$\mathcal{I}(a,k) = N(k-1) + a. \tag{7.67}$$

Master Element

To apply the quadrature rule on each element, one needs to define an affine transformation to map each spectral element $\mathbf{\Omega^k}$ to the reference or master element Λ. Let us define the local elemental mappings:

$$x^{\mathbf{k}}(\xi) = x^{\mathbf{k}}_m \, h_m(\xi), \tag{7.68}$$

where there is a summation on m. We can now map the physical elements $(x^{\mathbf{k}}) \in \mathbf{\Omega^k}$ onto the computational domain $(\xi) \in \Lambda$. We denote by $J^{\mathbf{k}}$ the Jacobian associated to this mapping such that

$$J^{\mathbf{k}} = \frac{\partial x^{\mathbf{k}}}{\partial \xi} = \left(\frac{\partial x}{\partial \xi}\right)^{\mathbf{k}}. \tag{7.69}$$

By $\partial x^{\mathbf{k}}/\partial\xi$ we refer to $\partial x/\partial\xi$ for some point x in the kth element $\mathbf{\Omega^k}$. We refer to $|J^{\mathbf{k}}|$ as the determinant of the Jacobian $J^{\mathbf{k}}$. This change of variable is a key component of the method and $|J^{\mathbf{k}}|$ appears in the elemental matrix discretization.

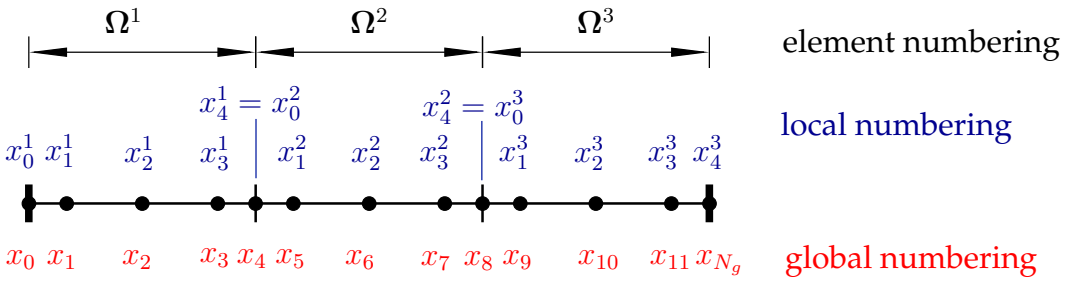


Figure 7.1: Illustration of a 1D SEM mesh with 3 elements of order $N = 4$ and $N_{GLL} = 5$ GLL (Gauss–Lobatto–Legendre) points per element.

Figure 7.1 illustrates a homogeneous 1D spectral element mesh with 3 elements of order $N = 4$ with 5 GLL points per element $\mathbf{\Omega^k}$. Practically, derivatives with respect to the physical coordinate x are evaluated in terms of the computational coordinate ξ. The mapping from the element $x^{\mathbf{k}} \in \mathbf{\Omega^k} = [X_{\mathbf{k}}, X_{\mathbf{k}+1}]$ to the computational space used is

$$\xi = \frac{2}{\Delta x^{\mathbf{k}}}(x^{\mathbf{k}} - X_{\mathbf{k}}) - 1, \tag{7.70}$$

where $\Delta x^{\mathbf{k}} = X_{k+1} - X_{\mathbf{k}}$ so that

$$\frac{\partial x^{\mathbf{k}}}{\partial \xi} = \frac{\Delta x^{\mathbf{k}}}{2}, \tag{7.71}$$

and hence, the determinant of the Jacobian is

$$|J^{\mathbf{k}}| = \frac{\Delta x^{\mathbf{k}}}{2}. \tag{7.72}$$

In particular, $|J^{\mathbf{k}}|$ is the same for all the elements $\forall$ $\mathbf{k}$ in the case of a homogeneous (evenly decomposed) domain.

Elemental matrix forms

On each subdomain, each integral is discretized in a similar fashion.

Elemental Mass matrix $\mathbf{M^k}$: The test functions are non zero for only one nodal point per element. We apply the method of weighted residuals to the integral $\int_{\mathbf{\Omega^k}} u_3^{\mathbf{k}}\ w^{\mathbf{k}}\ dx$ and write for each value $a \in \{0, N\}$

$$\int_{\mathbf{\Omega^k}} u_3^{\mathbf{k}}\ w_a^{\mathbf{k}}\ dx = \int_{\mathbf{\Omega^k}} \sum_{m=0}^{m=N} u_{3m}^{\mathbf{k}} h_m(x)\ h_a(x)\ dx. \tag{7.73}$$

The first step is to do a change of variable from the physical coordinate to the computational coordinate and then apply the GLL quadrature rule to the integral with weights $\rho_q^{\mathbf{k}}$.

$$\begin{aligned} \int_{\mathbf{\Omega^k}} \sum_{m=0}^{m=N} u_{3m}^{\mathbf{k}} h_m(x)\ h_a(x)\ dx &= \int_{\Lambda} \sum_{m=0}^{m=N} u_{3m}^{\mathbf{k}} h_m(\xi)\ h_a(\xi)\ |J^{\mathbf{k}}| d\xi \\ &= \sum_{q=0}^{q=N} \sum_{m=0}^{m=N} u_{3m}^{\mathbf{k}} h_m(\xi_q)\ h_a(\xi_q)\ \rho_q^{\mathbf{k}} |J^{\mathbf{k}}|. \end{aligned} \tag{7.74}$$

Now we use the properties of the Legendre interpolant. In particular equation (6.61) states that

$$h_i(\xi_j) = \delta_{ij}. \tag{7.75}$$

We can then write for each value $i \in \{0, N\}$,

$$\begin{aligned} \int_{\mathbf{\Omega^k}} \sum_{m=0}^{m=N} u_{3m}^{\mathbf{k}} h_m(x)\ h_a(x)\ dx &= \sum_{m=0}^{m=N} u_{3m}^{\mathbf{k}} \sum_{q=0}^{q=N} \delta_{qm}\ \delta_{qa}\ \rho_q^{\mathbf{k}} |J^{\mathbf{k}}| \\ &= \rho_a^{\mathbf{k}} |J^{\mathbf{k}}| u_{3a}^{\mathbf{k}}. \end{aligned} \tag{7.76}$$

To adopt a matrix form we can write for each element $\mathbf{\Omega^k}$ the system of N unknowns $u_{3i}^{\mathbf{k}}$

$$\begin{pmatrix} \int\limits_{\mathbf{\Omega^k}} \sum\limits_{m=0}^{m=N} u_{3m}^{\mathbf{k}} h_m(x) h_0(x) dx \\ \vdots \\ \int\limits_{\mathbf{\Omega^k}} \sum\limits_{m=0}^{m=N} u_{3m}^{\mathbf{k}} h_m(x) h_a(x) dx \\ \vdots \\ \int\limits_{\mathbf{\Omega^k}} \sum\limits_{m=0}^{m=N} u_{3m}^{\mathbf{k}} h_m(x) h_N(x) dx \end{pmatrix}^{\mathbf{k}} = \underbrace{\begin{pmatrix} \ddots & & 0 \\ & \rho_a^{\mathbf{k}} |J^{\mathbf{k}}| & \\ 0 & & \ddots \end{pmatrix}^{\mathbf{k}}}_{\mathbf{M^k}} \underbrace{\begin{pmatrix} u_{30}^{\mathbf{k}} \\ \vdots \\ u_{3a}^{\mathbf{k}} \\ \vdots \\ u_{3N}^{\mathbf{k}} \end{pmatrix}^{\mathbf{k}}}_{u_3^{\mathbf{k}}} \tag{7.77}$$

We define the ***local or elemental mass matrix*** $\mathbf{M^k}$ by

$$\mathbf{M}_{am}^{\mathbf{k}} = \delta_{am}\ \rho_a^{\mathbf{k}} |J^{\mathbf{k}}|. \tag{7.78}$$

Note that $\mathbf{M^k}$ is diagonal due to the choice of the inexact[1] GLL quadrature formula used in the SEM. Therefore, on each element we have the following discretization,

$$\int_{\mathbf{\Omega^k}} u_3^{\mathbf{k}}\, w_1^{\mathbf{k}}\, dx \;=\; \sum_{m=0}^{m=N} \mathbf{M}_{am}^{\mathbf{k}} u_{3m}^{\mathbf{k}} \;=\; \mathbf{M^k} \otimes u_3^{\mathbf{k}}. \tag{7.79}$$

The notation $\otimes$ is matrix multiplication operator that depends on the elemental matrix it is applied and on the spatial dimension. In the 1D case, and for the local mass matrix this $\otimes$ operator is a regular matrix multiplication. We will see more definitions of the $\otimes$ operator in the coming sections and in 3D. The following integrals have a similar discretization,

$$\int_{\mathbf{\Omega^k}} \partial_t u_1^{\mathbf{k}}\, w_1^{\mathbf{k}}\, dx \;=\; \mathbf{M^k} \otimes \dot{u}_1^{\mathbf{k}}, \tag{7.80}$$

$$\int_{\mathbf{\Omega^k}} \partial_t u_2^{\mathbf{k}}\, w_2^{\mathbf{k}}\, dx \;=\; \mathbf{M^k} \otimes \dot{u}_2^{\mathbf{k}}, \tag{7.81}$$

$$\int_{\mathbf{\Omega^k}} \partial_t u_3^{\mathbf{k}}\, w_3^{\mathbf{k}}\, dx \;=\; \mathbf{M^k} \otimes \dot{u}_3^{\mathbf{k}}, \tag{7.82}$$

where

$$\dot{u}_1^{\mathbf{k}} \;=\; \partial_t u_1^{\mathbf{k}}. \tag{7.83}$$

Elemental Force vector: To obtain the elemental force vector that arises from the integral on each subdomain $\mathbf{\Omega^k}$

$$\int_{\mathbf{\Omega^k}} S^{\mathbf{k}}\, w_3^{\mathbf{k}}\, dx, \tag{7.84}$$

we proceed in a similar fashion as for the elemental mass matrix.

For each value $i \in \{0, N\}$ we have,

$$\int_{\mathbf{\Omega^k}} S^{\mathbf{k}}\, w_{3a}^{\mathbf{k}}\, dx = \int_{\mathbf{\Omega^k}} S^{\mathbf{k}}(x)\, h_a(x)\, dx \tag{7.85}$$

Again, the first step is to do a change of variable from the physical coordinate to the computational coordinate and then apply the GLL quadrature rule to the integral with weights $\rho_q^{\mathbf{k}}$.

$$\int_{\mathbf{\Omega^k}} S^{\mathbf{k}}(x)\, h_a(x)\, dx \;=\; \int_{\Lambda} S^{\mathbf{k}}(\xi)\, h_a(\xi)\, |J^{\mathbf{k}}| d\xi \tag{7.86}$$

$$\;=\; \sum_{q=0}^{q=N} S^{\mathbf{k}}(\xi_q)\, h_a(\xi_q)\, \rho_q^{\mathbf{k}} |J^{\mathbf{k}}|, \tag{7.87}$$

$$\;=\; \sum_{q=0}^{q=N} S^{\mathbf{k}}(\xi_q)\, \delta_{qa}\, \rho_q^{\mathbf{k}} |J^{\mathbf{k}}| \tag{7.88}$$

[1] We recall that the GLL quadrature rule is exact for an integrand of order $2N-1$ or less. With the SEM, the integrand is a product of 2 basis functions or order N which results in a polynomial of order $2N$. However, by choosing an inexact quadrature rule, the mass matrix is by construction diagonal which is computationally a very important advantage.

Simplifying further, we can write for each value $a = \{0, N\}$,

$$\int_{\mathbf{\Omega^k}} S^{\mathbf{k}}(x)\ h_a(x)\ dx \quad = \quad S_a^{\mathbf{k}}\ \rho_a^{\mathbf{k}} |J^{\mathbf{k}}|. \tag{7.89}$$

Adopting a vector notation, we can write for each element $\mathbf{\Omega^k}$ the vector of N elements

$$\begin{pmatrix} \int\limits_{\mathbf{\Omega^k}} S^{\mathbf{k}}(x) h_0(x) dx \\ \vdots \\ \int\limits_{\mathbf{\Omega^k}} S^{\mathbf{k}}(x) h_a(x) dx \\ \vdots \\ \int\limits_{\mathbf{\Omega^k}} S^{\mathbf{k}}(x) h_N(x) dx \end{pmatrix}^{\mathbf{k}} = \underbrace{\begin{pmatrix} \rho_1^{\mathbf{k}} |J^{\mathbf{k}}|\ S_0^{\mathbf{k}} \\ \vdots \\ \rho_a^{\mathbf{k}} |J^{\mathbf{k}}|\ S_a^{\mathbf{k}} \\ \vdots \\ \rho_N^{\mathbf{k}} |J^{\mathbf{k}}|\ S_N^{\mathbf{k}} \end{pmatrix}^{\mathbf{k}}}_{\mathbf{F^k}} \tag{7.90}$$

We define the ***local or elemental force vector*** $\mathbf{F^k}$ by

$$\mathbf{F}_a^{\mathbf{k}} \quad = \quad \rho_a^{\mathbf{k}} |J^{\mathbf{k}}|\ S_a^{\mathbf{k}}, \tag{7.91}$$

where $S_a^{\mathbf{k}} = S^{\mathbf{k}}(\xi_a)$. Therefore, on each element we have the following discretization,

$$\int_{\mathbf{\Omega^k}} S^{\mathbf{k}}\ w_3^{\mathbf{k}}\ dx \ = \ \mathbf{F^k}. \tag{7.92}$$

Elemental Advection matrix: Typically the advection matrix arises from integral terms of this form

- in ***version 1***

$$\int_{\mathbf{\Omega^k}} u_3^{\mathbf{k}}\ \partial_x w_2^{\mathbf{k}}\ dx; \tag{7.93}$$

- in ***version 2***

$$\int_{\mathbf{\Omega^k}} \partial_x u_3^{\mathbf{k}}\ w_2^{\mathbf{k}}\ dx. \tag{7.94}$$

In both integrals one derivative term is present and they generate related elemental advection matrices respectively $\mathbf{D^k}$ and $\mathbf{A^k}$. Again the discretization process of these integrals is somewhat similar to the one of the elemental mass matrix.

For each element $\mathbf{\Omega^k}$ and each value $i = \{0, N\}$ that corresponds to each nodal point,

$$\int_{\mathbf{\Omega^k}} u_3^{\mathbf{k}}\ \partial_x w_2^{\mathbf{k}}\ dx = \int_{\mathbf{\Omega^k}} \sum_{m=0}^{m=N} u_{3m}^{\mathbf{k}} h_m(x)\ \partial_x h_a(x)\ dx. \tag{7.95}$$

Once again, we apply a change of variable from the physical coordinate to the computational coordinate and then apply the GLL quadrature rule to the integral with weights $\rho_q^{\mathbf{k}}$.

$$\begin{aligned} \int_{\mathbf{\Omega^k}} \sum_{m=0}^{m=N} u_{3m}^{\mathbf{k}} h_m(x)\ \partial_x h_a(x)\ dx \quad &= \quad \int_{\Lambda} \sum_{m=0}^{m=N} u_{3m}^{\mathbf{k}} h_m(\xi)\ \partial_\xi h_a(\xi)\ \frac{\partial \xi}{\partial x} |J^{\mathbf{k}}| d\xi \\ &= \quad \sum_{q=0}^{q=N} \sum_{m=0}^{m=N} u_{3m}^{\mathbf{k}} h_m(\xi_q)\ \partial_\xi h_a(\xi_q)\ \rho_q^{\mathbf{k}}. \end{aligned} \tag{7.96}$$

Remember that in 1D $\partial x^{\mathbf{k}}/\partial \xi = |J^{\mathbf{k}}|$, so that the determinant of the Jacobian disappears in the advection integral. The derivative of the Lagrange-Legendre interpolant is defined in equation (6.62)

$$\partial_\xi h_i(\xi_j) = H_{ji}. \tag{7.97}$$

Therefore, we can write for each nodal point,

$$\int_{\mathbf{\Omega^k}} \sum_{m=0}^{m=N} u_{3m}^{\mathbf{k}} h_m(x)\ \partial_x h_a(x)\ dx \ = \ \sum_{m=0}^{m=N} u_{3m}^{\mathbf{k}} \sum_{q=0}^{q=N} \delta_{qm}\ H_{qa}\ \rho_q^{\mathbf{k}}. \tag{7.98}$$

To adopt a matrix form we can write for each element $\mathbf{\Omega^k}$ the system of N unknowns $u_{3a}^{\mathbf{k}}$

$$\begin{pmatrix} \int\limits_{\mathbf{\Omega^k}} \sum\limits_{m=0}^{m=N} u_{3m}^{\mathbf{k}} h_m(x) \partial_x h_0(x) dx \\ \vdots \\ \int\limits_{\mathbf{\Omega^k}} \sum\limits_{m=0}^{m=N} u_{3m}^{\mathbf{k}} h_m(x) \partial_x h_a(x) dx \\ \vdots \\ \int\limits_{\mathbf{\Omega^k}} \sum\limits_{m=0}^{m=N} u_{3m}^{\mathbf{k}} h_m(x) \partial_x h_N(x) dx \end{pmatrix}^{\mathbf{k}} = \underbrace{\begin{pmatrix} \rho_0^{\mathbf{k}} H_{00} & \cdots & \rho_N^{\mathbf{k}} H_{N0} \\ \vdots & \rho_q^{\mathbf{k}} H_{qa} & \vdots \\ \rho_0^{\mathbf{k}} H_{0N} & \cdots & \rho_N^{\mathbf{k}} H_{NN} \end{pmatrix}^{\mathbf{k}}}_{\mathbf{D^k}} \underbrace{\begin{pmatrix} u_{30}^{\mathbf{k}} \\ \vdots \\ u_{3a}^{\mathbf{k}} \\ \vdots \\ u_{3N}^{\mathbf{k}} \end{pmatrix}^{\mathbf{k}}}_{u_3^{\mathbf{k}}} \tag{7.99}$$

We define the ***local or elemental advection matrix*** $\mathbf{D^k}$ by

$$\mathbf{D}_{am}^{\mathbf{k}} \ = \ \sum_{m=0}^{m=N} \delta_{qm}\ H_{qa}\ \rho_q^{\mathbf{k}} \ = \ \rho_m^{\mathbf{k}} H_{ma} \ = \ \left(\rho_m^{\mathbf{k}}\ H_{am}\right)^T. \tag{7.100}$$

Note that $\mathbf{D^k}$ is not diagonal. Finally, on each element we have the following discretization,

$$\int_{\mathbf{\Omega^k}} u_3^{\mathbf{k}}\ \partial_x w_2^{\mathbf{k}}\ dx \ = \ \sum_{m=0}^{m=N} \mathbf{D}_{am}^{\mathbf{k}} u_{3m}^{\mathbf{k}} \ = \ \mathbf{D^k} \otimes u_3^{\mathbf{k}}. \tag{7.101}$$

The following integral has a similar discretization,

$$\int_{\mathbf{\Omega^k}} u_2^{\mathbf{k}}\ \partial_x w_3^{\mathbf{k}}\ dx \ = \ \sum_{m=0}^{m=N} \mathbf{D}_{am}^{\mathbf{k}} u_{2m}^{\mathbf{k}} \ = \ \mathbf{D^k} \otimes u_2^{\mathbf{k}}. \tag{7.102}$$

When keeping the original integral in ***version 2***

$$I_1 \ = \ \int_{\mathbf{\Omega^k}} \partial_x u_3^{\mathbf{k}}\ w_2^{\mathbf{k}}\ dx \tag{7.103}$$

the discretization is slightly different. We obtain in this case

$$\int_{\mathbf{\Omega^k}} \partial_x u_3^{\mathbf{k}}\ w_2^{\mathbf{k}}\ dx \ = \ \sum_{m=0}^{m=N} \mathbf{A}_{am}^{\mathbf{k}} u_{3m}^{\mathbf{k}} \ = \ \mathbf{A^k} \otimes u_3^{\mathbf{k}}, \tag{7.104}$$

where

$$\mathbf{A}^{\mathbf{k}}_{am} \;=\; \sum_{m=0}^{m=N} H_{qm}\,\delta_{qa}\,\rho^{\mathbf{k}}_{q} \;=\; \rho^{\mathbf{k}}_{a} H_{am}. \tag{7.105}$$

We have the relation between those two types of matrices in ***1D only***,

$$\mathbf{A}^{\mathbf{k}} \;=\; \left(\mathbf{D}^{\mathbf{k}}\right)^{T}. \tag{7.106}$$

Elemental boundary terms: As mentioned earlier, in an ***Important remark***, the treatment of the elemental boundary terms is special. If we integrate by parts on each element separately there would be a number of boundary terms at the interior boundaries. The spectral element strategy is to ignore these extra interior terms and only take into account the terms on the elements at the boundary. In 1D, this means that we only look at the boundary terms at the first and last global nodes. Let's define the boundary matrix $\mathbf{B}^{\mathbf{k}}$ so that

$$\mathbf{B}^{\mathbf{k}} \otimes u_2^{\mathbf{k}} = \begin{pmatrix} -u_2^1(-L,t)\;\delta(-L) \\ 0 \\ \vdots \\ 0 \\ -u_2^{N_E}(L,t)\;\delta(L) \end{pmatrix}^{\mathbf{k}} = \begin{pmatrix} -u^1_{2\mathcal{I}(0,1)}\;\delta_{i\mathcal{I}(0,1)} \\ 0 \\ \vdots \\ 0 \\ -u^{N_E}_{2\mathcal{I}(N,N_E)}\;\delta_{i\mathcal{I}(N,N_E)} \end{pmatrix}^{\mathbf{k}}. \tag{7.107}$$

Furthermore we have the boundary term

$$\mathbf{B}^{\mathbf{k}} \otimes u_3^{\mathbf{k}} = \begin{pmatrix} -u^1_{3\mathcal{I}(0,1)}\;\delta_{i\mathcal{I}(0,1)} \\ 0 \\ \vdots \\ 0 \\ -u^{N_E}_{3\mathcal{I}(N,N_E)}\;\delta_{i\mathcal{I}(N,N_E)} \end{pmatrix}^{\mathbf{k}}. \tag{7.108}$$

7.1.5 Elemental matrix system

Version 1 In each subdomain $\mathbf{\Omega^k}$, we have the following discretization of the weak formulation (***version 1***) of equations (7.183), (7.184) and (7.185),

$$\mathbf{M}^{\mathbf{k}} \otimes \dot{u}_1^{\mathbf{k}} \;=\; \mathbf{M}^{\mathbf{k}} \otimes u_3^{\mathbf{k}}, \tag{7.109}$$

$$\mathbf{M}^{\mathbf{k}} \otimes \dot{u}_2^{\mathbf{k}} \;=\; \mathbf{B}^{\mathbf{k}} \otimes u_2^{\mathbf{k}} - \mathbf{D}^{\mathbf{k}} \otimes u_3^{\mathbf{k}}, \tag{7.110}$$

$$\mathbf{M}^{\mathbf{k}} \otimes \dot{u}_3^{\mathbf{k}} \;=\; \mathbf{B}^{\mathbf{k}} \otimes u_3^{\mathbf{k}} - \mathbf{D}^{\mathbf{k}} \otimes u_2^{\mathbf{k}} + \mathbf{F}^{\mathbf{k}} \tag{7.111}$$

Version 2 In each subdomain $\mathbf{\Omega^k}$, we have the following discretization of the weak formulation (***version 2***) of equations (7.57), (7.58) and (7.59),

$$\mathbf{M}^{\mathbf{k}} \otimes \dot{u}_1^{\mathbf{k}} \;=\; \mathbf{M}^{\mathbf{k}} \otimes u_3^{\mathbf{k}}, \tag{7.112}$$

$$\mathbf{M}^{\mathbf{k}} \otimes \dot{u}_2^{\mathbf{k}} \;=\; \mathbf{A}^{\mathbf{k}} \otimes u_3^{\mathbf{k}}, \tag{7.113}$$

$$\mathbf{M}^{\mathbf{k}} \otimes \dot{u}_3^{\mathbf{k}} \;=\; \mathbf{B}^{\mathbf{k}} \otimes u_3^{\mathbf{k}} - \mathbf{D}^{\mathbf{k}} \otimes u_2^{\mathbf{k}} + \mathbf{F}^{\mathbf{k}} \tag{7.114}$$

Remember that in the 1D case, for ***elemental matrix*** operations, the matrix operator $\otimes$ corresponds to a regular matrix vector multiplication.

7.1.6 Assembly of global discretization matrix

All the elemental contributions need to be added together and this is called ***the assembly of the global matrix***. In general, consider a local matrix $A^{\mathbf{k}}$, the global matrix A is noted

$$A = \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' A^{\mathbf{k}}, \tag{7.115}$$

where $\sum_{k=1}^{k=N_E}{}'$ represents the assembly process often called ***direct stiffness summation***. Technically, this is not quite a direct sum; each submatrix has one line and column in common (boundary nodes) with another submatrix. The $'$ is to point out the fact that this is not a typical summation. Figure 7.2 illustrates the process of direct stiffness summation to obtain a global system of algebraic equations.

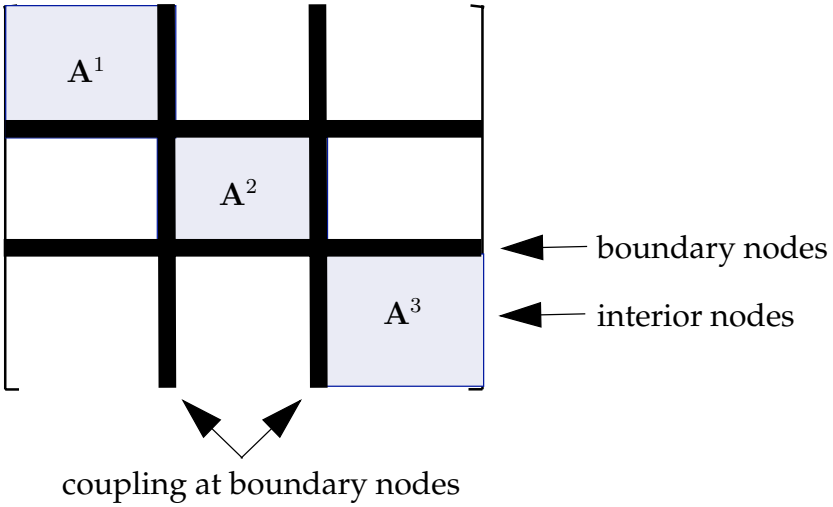


Figure 7.2: Schematic of the direct summation of local matrices $\mathbf{A}^{\mathbf{k}}$ to form the global matrix A.

The global system of our problem is therefore given for ***version 1*** by

$$\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}'\left(\mathbf{M}^{\mathbf{k}} \otimes \dot{u}_1^{\mathbf{k}}\right) = \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}'\left(\mathbf{M}^{\mathbf{k}} \otimes u_3^{\mathbf{k}}\right), \tag{7.116}$$

$$\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}'\left(\mathbf{M}^{\mathbf{k}} \otimes \dot{u}_2^{\mathbf{k}}\right) = \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}'\left(\mathbf{B}^{\mathbf{k}} \otimes u_2^{\mathbf{k}} - \mathbf{D}^{\mathbf{k}} \otimes u_3^{\mathbf{k}}\right), \tag{7.117}$$

$$\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}'\left(\mathbf{M}^{\mathbf{k}} \otimes \dot{u}_3^{\mathbf{k}}\right) = \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}'\left(\mathbf{B}^{\mathbf{k}} \otimes u_3^{\mathbf{k}} - \mathbf{D}^{\mathbf{k}} \otimes u_2^{\mathbf{k}} + \mathbf{F}^{\mathbf{k}}\right). \tag{7.118}$$

In ***version 2*** only equation (7.117) is modified and becomes

$$\sum_{k=1}^{k=N_E}{}' \left(\mathbf{M}^{\mathbf{k}} \otimes \dot{u}_2^{\mathbf{k}} \right) \;=\; \sum_{k=1}^{k=N_E}{}' \left(\mathbf{A}^{\mathbf{k}} \otimes u_3^{\mathbf{k}} \right). \tag{7.119}$$

Global system version 1

The final global system of the 3 unknowns in matrix form in ***version 1*** is

$$\mathbf{M} \otimes \dot{u}_1 \;=\; \mathbf{M} \otimes u_3, \tag{7.120}$$
$$\mathbf{M} \otimes \dot{u}_2 \;=\; \mathbf{B} \otimes u_2 - \mathbf{D} \otimes u_3, \tag{7.121}$$
$$\mathbf{M} \otimes \dot{u}_3 \;=\; \mathbf{B} \otimes u_3 - \mathbf{D} \otimes u_2 + \mathbf{F}. \tag{7.122}$$

Note that after the assembly, we still use the notation $\otimes$ as a matrix multiplication operator on the ***global*** matrices, but it is not a regular matrix-vector multiplication anymore as for the local matrix operations. To simplify the notations further, we introduce the system in a similar fashion as in (7.32)

$$\dot{U} \;=\; \mathcal{A}_{v1} U + \mathcal{F}, \tag{7.123}$$

that is

$$\underbrace{\begin{pmatrix} \dot{u}_1 \\ \dot{u}_2 \\ \dot{u}_3 \end{pmatrix}}_{\dot{U}} \;=\; \underbrace{\begin{pmatrix} 0 & 0 & 1 \\ 0 & \mathbf{B}/\mathbf{M} & -\mathbf{D}/\mathbf{M} \\ 0 & -\mathbf{D}/\mathbf{M} & \mathbf{B}/\mathbf{M} \end{pmatrix}}_{\mathcal{A}_{v1}} \underbrace{\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}}_{U} + \underbrace{\begin{pmatrix} 0 \\ 0 \\ S \end{pmatrix}}_{\mathcal{F}} \tag{7.124}$$

Global system version 2

The final global system of the 3 unknowns in matrix form in ***version 2*** is

$$\mathbf{M} \otimes \dot{u}_1 \;=\; \mathbf{M} \otimes u_3, \tag{7.125}$$
$$\mathbf{M} \otimes \dot{u}_2 \;=\; \mathbf{A} \otimes u_3, \tag{7.126}$$
$$\mathbf{M} \otimes \dot{u}_3 \;=\; \mathbf{B} \otimes u_3 - \mathbf{D} \otimes u_2 + \mathbf{F}. \tag{7.127}$$

In terms of a system notation we introduce

$$\dot{U} \;=\; \mathcal{A}_{v2} U + \mathcal{F}, \tag{7.128}$$

that is

$$\underbrace{\begin{pmatrix} \dot{u}_1 \\ \dot{u}_2 \\ \dot{u}_3 \end{pmatrix}}_{\dot{U}} \;=\; \underbrace{\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & \mathbf{A}/\mathbf{M} \\ 0 & -\mathbf{D}/\mathbf{M} & \mathbf{B}/\mathbf{M} \end{pmatrix}}_{\mathcal{A}_{v2}} \underbrace{\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}}_{U} + \underbrace{\begin{pmatrix} 0 \\ 0 \\ S \end{pmatrix}}_{\mathcal{F}} \tag{7.129}$$

In practice:

From a computational point of view the assembly of the local advection matrix

$$\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \left(\mathbf{D}^{\mathbf{k}} \otimes u_3^{\mathbf{k}} \right), \tag{7.130}$$

is constructed by first initializing the matrix $\mathbf{D}$ to zero, and then beginning an outer loop over the element index k. For each value k, there are two inner loops over the indices i and j to calculate all the rows and columns that are affected by the element $\mathbf{\Omega^k}$. For example

$$\mathbf{D}_{\mathcal{I}(i,k),\mathcal{I}(j,k)} := \mathbf{D}_{\mathcal{I}(i,k),\mathcal{I}(j,k)} + \mathbf{D}_{ij}^{\mathbf{k}} \qquad i,j \in \{1, N_{GLL}\} \tag{7.131}$$

$$:= \mathbf{D}_{\mathcal{I}(i,k),\mathcal{I}(j,k)} + \left(\rho_i^{\mathbf{k}} H_{ij} \right)^T. \tag{7.132}$$

Note that the coefficients i and j run from $\{1, N_{GLL}\}$ instead of $\{0, N\}$ for computational reasons. Once assembled the matrix $\mathbf{D}$ is sparse and "almost" block diagonal in 1D (boundary nodes in common between submatrices). So in order to limit memory storage it is useful to perform calculations directly while assembling. Further optmisation of the numerical procedure could be achieved by implementing a block-diagonal sparse matrix, but is not essential for the illustration presented here.

7.1.7 Time Discretization

The time discretization of the system

$$\dot{U} \;=\; \mathcal{A} \otimes U + \mathcal{F} \;=\; f(U,t), \tag{7.133}$$

is computed by an explicit fourth order Runge–Kutta method. Given an initial condition U_0, the solution U_{n+1} at time t_{n+1} is determined from the previous time t_n and the solution U_n. The details of the Runge–Kutta fourth order method can be found more explicitly in 6.10.

7.2 Numerical results for a hyperbolic system first order in space and time in 1D

In this section we show some numerical results obtained with the spectral element method with a 1D wave equation with a source term. The numerical solution is represented in figure 7.3 as a function of time t and space x. For our particular problem the timestep Δt is set by the Courant–Friedrichs–Lewy stability condition[2] CFL by the following relation:

$$\Delta t = \text{CFL} \times \min(\Delta x). \tag{7.134}$$

[2]The Courant–Friedrichs–Lewy stability condition ensures the stability of explicit time schemes.

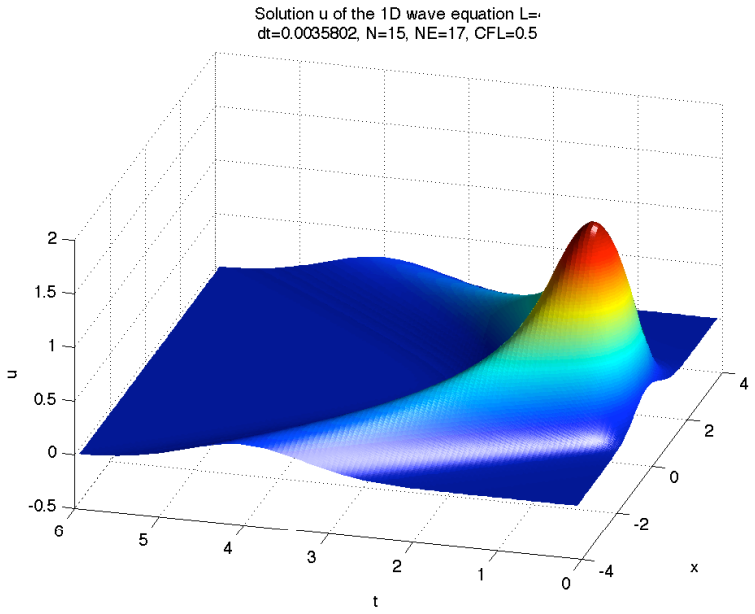


Figure 7.3: Numerical solution u_1 for $N = 15$, $N_E = 17$, $L = 4$ and a Courant–Friedrichs–Lewy condition $CFL = 0.5$.

To compare the exact and numerical solution we calculate the numerical norm

$$\left(\text{numerical } \mathcal{L}^2 \text{ norm }\right) = \sqrt{\frac{\sum_{j=1}^{N_g} \left(u_{\text{exact}}(x_j) - u_{\text{numerics}}(x_j) \right)^2}{N_g}} \tag{7.135}$$

7.2.1 $\mathcal{L}^2$ norm and hp-convergence in 1D

Figure (7.4) shows the numerical $\mathcal{L}^2$ norm for varying accuracies with $N_E = 9, 17$ number of elements, and polynomial orders $N = 5, 9, 15$. In Table 7.2.1, we give the number of degrees of freedom or total number of points N_g, for each of the different combinations of polynomial order N and number of elements N_E.

In the 1970s, the mathematical theory of FEM has established rigorously the convergence of the h-version of the finite element method. The error in the numerical solution decays ***algebraically*** by refining the mesh, that is introducing more elements while keeping the order of the interpolating polynomial N constant. An alternative approach is to keep the number of elements fixed and increase the order of the interpolating polynomials in order to reduce the error in the numerical solution. This is called p-type refinement and is typical of polynomial spectral methods. For infinitely smooth solutions, the p-refinement usually leads to an ***exponential*** decay of the numerical error.

A few basic key definitions are in order here, for more detailed definitions see [120]:

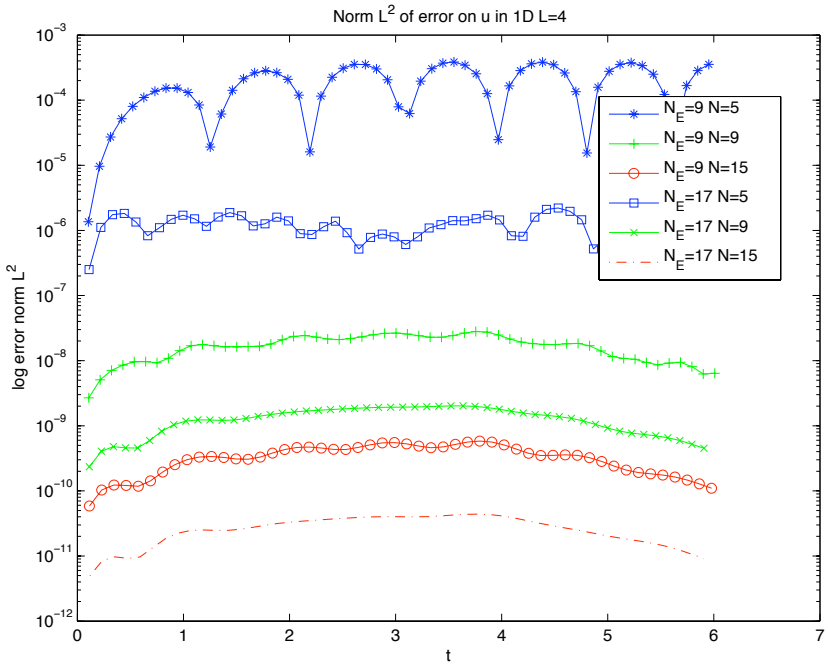


Figure 7.4: $\mathcal{L}^2$ norm of the numerical and exact solution u_1 for varying polynomial order $N = 5, 9, 15$ and number of elements $N_E = 9, 17$ for a domain $L = 4$, with $CFL = 0.5$.

N \ N_E	7	9	11	13	15	17
5	36	46	56	66	76	86
7	50	64	78	92	106	120
9	64	82	100	118	136	154
11	78	100	122	144	166	188
13	92	118	144	170	196	222
15	106	136	166	196	226	256

Table 7.1: Degrees of freedom N_g (total number of points) as a function of the polynomial order N and the number of elements N_E in 1D.

Algebraic convergence rate: The term a_n follows an algebraic convergence rate if

$$a_n \sim O(\frac{1}{n^k}) \qquad n \gg 1, \tag{7.136}$$

where k is the index of convergence.

Exponential convergence rate: The term a_n follows an exponential convergence rate if

$$a_n \sim O(\exp(-kn)) \qquad n \gg 1, \tag{7.137}$$

where k is the index of convergence.

Figure 7.5 shows the $\mathcal{L}^2$ norm as a function of the total degree of freedom (total number of points N_g), for both the h-refinement with a fixed polynomial order N, and a p-refinement based on an evenly decomposed mesh. In figure 7.5, we also illustrate the shapes of algebraic and exponential convergence rates for comparison purposes. Note that on a log-log axis, algebraic convergence asymptotes to a straight line whose slope is $-k$, whereas exponential convergence bends away with ever-increasing negative slopes.

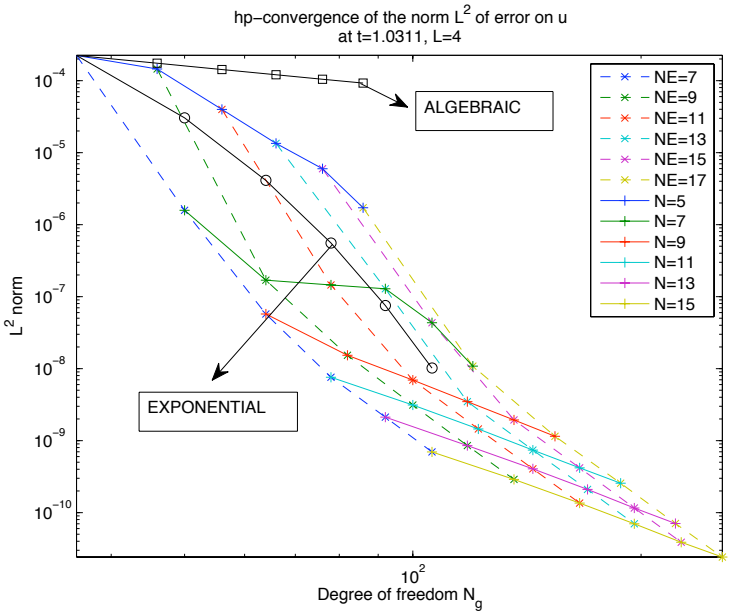


Figure 7.5: hp-convergence for the $\mathcal{L}^2$ norm of the numerical and exact solution u_1 as a function of the number of points N_g. We fix the polynomial order N and vary the number of elements N_E (h-convergence in solid lines), and we fix the number of elements N_E and vary the polynomial order N (p-convergence in dashed lines). See Table 7.2.1 for the values of N and N_E. The norms are taken at $t = 1$ for a domain $L = 4$, with $CFL = 0.5$

The h-refinement initially resolves the solution faster than the p-refinement, however, as the asymptotic exponential convergence is achieved the p-refinement overtakes the h-refinement process.

The optimum convergence path as a function of degrees of freedom N_g, involves using both h and p-refinement.

In general, we would like to know the error as a function of the computational cost, but this is much harder to measure. However, for smooth solutions, the concept of hp-refinement still provides the optimal convergence strategy.

7.2.2 Experiments on Sommerfeld-like Boundary conditions in 1D

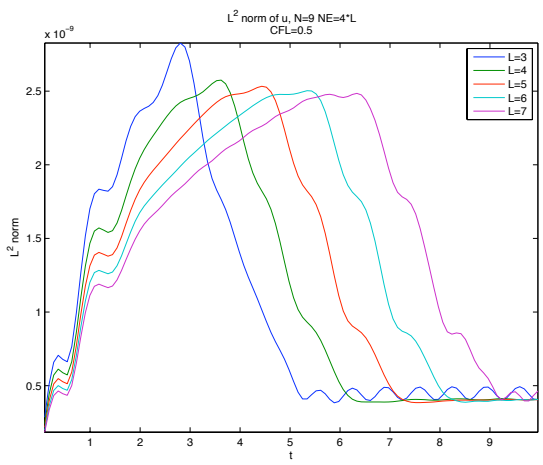
We now focus our attention to the numerical errors coming from the boundary conditions.

- Figure (7.6(a)) shows the numerical $\mathcal{L}^2$ norm for the same polynomial order $N = 9$ and the same ratio of elements $N_E = 4L$ for various values of length of domains $L = 3, 4, 5, 6, 7$.
- Figure (7.6(b)) shows the numerical $\mathcal{L}^2$ norm for the same polynomial order $N = 15$ and the same ratio of elements $N_E = 4L$.
- Figure (7.6(c)) shows the numerical $\mathcal{L}^2$ norm for the same polynomial order $N = 15$ and the same ratio of elements $N_E = 6L$.

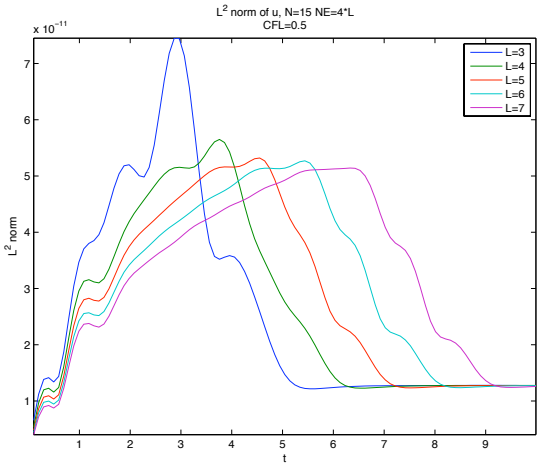
All norms clearly suggest that with the same resolution in space and time, the error decreases as the boundary is pushed further away.

7.2.3 Convergence in time

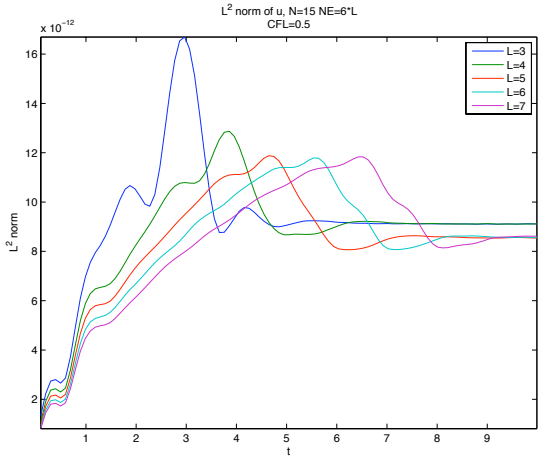
Figure 7.7 shows the fourth order convergence in time with the Runge–Kutta method. A scheme is fourth order convergent ***if*** the norm $\text{Norm}_{\Delta t}$ obtained with timestep Δt is$(\text{Norm}_{\Delta t}) = (2^4 \times \text{Norm}_{\Delta t/2})$ where $\text{Norm}_{\Delta t/2}$ is the norm obtained with a time-stepping of $\Delta t/2$ and the same spatial resolution. In the figure we see that both norms are on top of each other which shows a fourth order convergent scheme in time.



(a) Polynomial order $N = 9$ with $N_E = 4L$



(b) Polynomial order $N = 15$ with $N_E = 4L$



(c) Polynomial order $N = 15$ with $N_E = 6L$

Figure 7.6: Convergence test on the Sommerfeld-like boundary conditions in 1D. For the same accuracy in space and time, the domain is successively $L = 3, 4, 5, 6, 7$. The error decreases as the boundary is pushed further away.

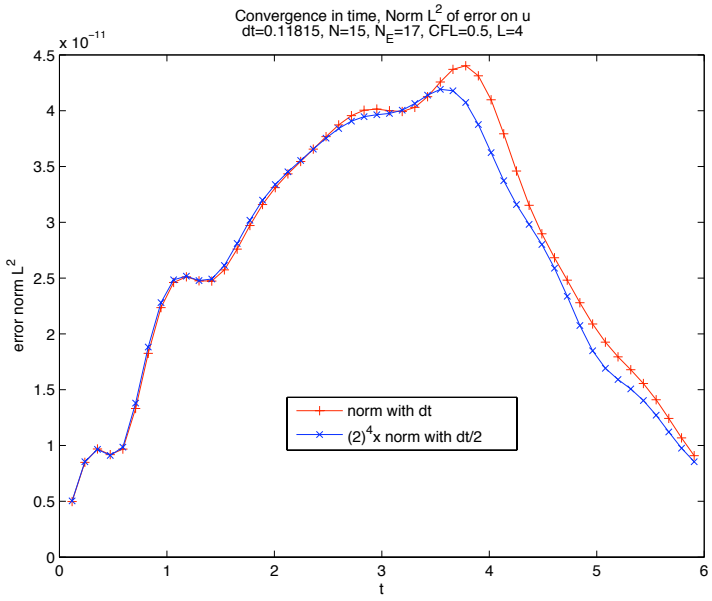


Figure 7.7: Fourth-order convergence in time for the Runge-Kutta method. The $\mathcal{L}^2$ norms are given for the same spatial accuracy but for Δt and $\Delta t/2$, we can see that $(\text{Norm}_{\Delta t}) = (2^4 \times \text{Norm}_{\Delta t/2})$ which shows a fourth order convergent scheme. The domain is $L = 4$ with a polynomial order $N = 15$ a number of elements $N_E = 17$ and with $CFL = 0.5$

7.3 Hyperbolic system first order in space and time in 3D

We now convert the 3D wave equation into a hyperbolic system first order in time and space to deal with a simplified system as close as possible as the BSSN system.

7.3.1 Wave equation with source term

The original wave equation is given on the domain $x, y, z \in [-L, L]$.

$$\partial_{tt} u - (\partial_{xx} u + \partial_{yy} u + \partial_{zz} u) \;=\; S(x, y, z, t). \tag{7.138}$$

This equation is in 3D and is second order in space and in time. One needs an initial condition $u_0 = u(x, y, z, t_0)$ and boundary conditions. We use the Sommerfeld-like absorbing or non-reflecting boundary conditions.

Exact solution in 3D: To evaluate the spectral element method accuracy we compare the numerical solution to an exact solution u given by,

$$u(r, t) \;=\; \frac{\exp(-(r-t)^2) - \exp(-(r+t)^2)}{r}, \tag{7.139}$$

where $r = \sqrt{x^2 + y^2 + z^2}$. There is no source term for this solution.

Initial and boundary conditions in 3D for a spherical solution: The initial condition of the solution for $t_0 = 0$ is

$$u(r, 0) \;=\; u_0(r). \tag{7.140}$$

For the derivatives, we have

$$\partial_x u(r, 0) \;=\; u_{x0}(r) \;=\; 0, \tag{7.141}$$

$$\partial_y u(r, 0) \;=\; u_{y0}(r) \;=\; 0, \tag{7.142}$$

$$\partial_y u(r, 0) \;=\; u_{z0}(r) \;=\; 0, \tag{7.143}$$

$$\partial_t u(0, t) \;=\; u_{t0}(t) \;=\; \frac{4 - 8t^2}{e^{t^2}}. \tag{7.144}$$

We use some absorbing boundary conditions on the 6 boundaries (faces) $\mathbf{\Gamma}_1$, $\mathbf{\Gamma}_2$, $\mathbf{\Gamma}_3$, $\mathbf{\Gamma}_4$, $\mathbf{\Gamma}_5$ and $\mathbf{\Gamma}_6$, so that

$$\text{On } \mathbf{\Gamma}_i : \frac{x_i}{R_i}\, \partial_t u\big|_{\mathbf{\Gamma}_i} + \partial_i u\big|_{\mathbf{\Gamma}_i} + \frac{x_i}{R_i^2}\, u\big|_{\mathbf{\Gamma}_i} \;=\; b_i, \tag{7.145}$$

where $R_i = \left(\sqrt{x^2 + y^2 + z^2}\right)\Big|_{\mathbf{\Gamma}_i}$. Specifically,

$$\text{On } \mathbf{\Gamma}_1 : \quad x_i = -L_Z, \;\; \partial_i = \partial_z, \;\; R_i = \sqrt{x^2 + y^2 + (-L_Z)^2}; \tag{7.146}$$

$$\text{On } \mathbf{\Gamma}_2 : \quad x_i = L_Z, \;\; \partial_i = \partial_z, \;\; R_i = \sqrt{x^2 + y^2 + (L_Z)^2}; \tag{7.147}$$

$$\text{On } \mathbf{\Gamma}_3 : \quad x_i = -L_Y, \;\; \partial_i = \partial_y, \;\; R_i = \sqrt{x^2 + (-L_Y)^2 + z^2}; \tag{7.148}$$

$$\text{On } \mathbf{\Gamma}_4 : \quad x_i = L_Y, \;\; \partial_i = \partial_y, \;\; R_i = \sqrt{x^2 + (L_Y)^2 + z^2}; \tag{7.149}$$

$$\text{On } \mathbf{\Gamma}_5 : \quad x_i = -L_X, \;\; \partial_i = \partial_x, \;\; R_i = \sqrt{(-L_X)^2 + y^2 + z^2}; \tag{7.150}$$

$$\text{On } \mathbf{\Gamma}_6 : \quad x_i = L_X, \;\; \partial_i = \partial_x, \;\; R_i = \sqrt{(L_X)^2 + y^2 + z^2}. \tag{7.151}$$

The boundary functions $b_i, i = 1, 6$ are not exactly zero but go to zero exponentially for a large enough value of L. They are given by

$$b_i = \frac{4(R_i + t)x_i e^{-(R_i+t)^2}}{R_i}. \tag{7.152}$$

Figure 7.8 represents a domain $\mathbf{\Omega}$ in 3D with the boundaries $\mathbf{\Gamma}_i$ and outward unit normals $\mathbf{n}_i$.

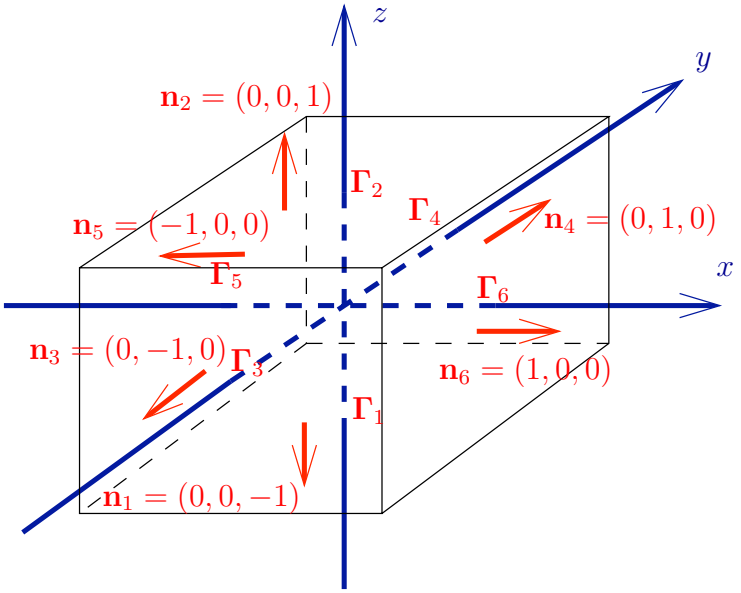


Figure 7.8: Representation of a 3D domain with boundary faces and outward unit normal vectors.

7.3.2 System of 5 unknowns: strong formulation

Let us define $u_1 = u$ and then introduce 4 more variables (coming from the partial derivatives in space and time),

$$u_2 = \partial_x u_1; \tag{7.153}$$
$$u_3 = \partial_y u_1; \tag{7.154}$$
$$u_4 = \partial_z u_1; \tag{7.155}$$
$$u_5 = \partial_t u_1. \tag{7.156}$$

The wave equation can be rewritten into a system

$$\begin{cases} \partial_t u_1 = u_5 \\ \partial_t u_2 = \partial_x u_5 \\ \partial_t u_3 = \partial_y u_5 \\ \partial_t u_4 = \partial_z u_5 \\ \partial_t u_5 = \partial_x u_2 + \partial_y u_3 + \partial_z u_4 + S(x, y, t) \end{cases}. \tag{7.157}$$

The system (7.157) can be written in matrix form

$$\partial_t U = AU + F, \tag{7.158}$$

where $U = (u_1\ u_2\ u_3\ u_4\ u_5)^T$, $F = (0\ 0\ 0\ 0\ S(x,y,z,t))^T$, and

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & \partial_x \\ 0 & 0 & 0 & 0 & \partial_y \\ 0 & 0 & 0 & 0 & \partial_z \\ 0 & \partial_x & \partial_y & \partial_z & 0 \end{pmatrix}. \tag{7.159}$$

The initial conditions on u_1, u_2, u_3, u_4 and u_5 are given in terms of the initial condition u_0, its initial first space derivatives $(\partial_x u)_0$, $(\partial_y u)_0$, $(\partial_z u)_0$ and its initial first time derivative $(\partial_t u)_0$.

7.3.3 Weak formulation

To simplify the notation, we use $\int_\Omega = \iiint_\Omega$. We apply the variational formulation to the system (7.157) and obtain the following weak formulation

$$\int_\Omega \partial_t u_1\ w\ d\Omega = \int_\Omega u_5\ w\ d\Omega, \tag{7.160}$$

$$\int_\Omega \partial_t u_2\ w\ d\Omega = \underbrace{\int_\Omega \partial_x u_5\ w\ d\Omega}_{\mathrm{I}_1}, \tag{7.161}$$

$$\int_\Omega \partial_t u_3\ w\ d\Omega = \underbrace{\int_\Omega \partial_y u_5\ w\ d\Omega}_{\mathrm{I}_2}, \tag{7.162}$$

$$\int_\Omega \partial_t u_4\ w\ d\Omega = \underbrace{\int_\Omega \partial_z u_5\ w\ d\Omega}_{\mathrm{I}_3}, \tag{7.163}$$

$$\int_\Omega \partial_t u_5\ w\ d\Omega = \underbrace{\int_\Omega \partial_x u_2\ w\ d\Omega + \int_\Omega \partial_y u_3\ w\ d\Omega + \int_\Omega \partial_z u_4\ w\ d\Omega}_{\mathrm{I}_4} + \int_\Omega S(x,y,z,t)\ w\ d\Omega\ . \tag{7.164}$$

The boundary conditions are generally introduced when integrating by parts. However, not all of the above integrals I_1, I_2, I_3 and I_4 should be integrated by parts in 3 dimensions.

7.3.4 Integration by parts in 3D

Let us recapitulate a few general formulae of integration by parts in 3 dimensions.

$$\begin{aligned}\int_{\mathbf{\Omega}} (\nabla \cdot u)\ w\ d\Omega &= \int_{\mathbf{\Omega}} (\partial_x u + \partial_y u + \partial_z u)\ w\ d\Omega,\\ &= \int_{\mathbf{\Gamma}} (\mathbf{n} \cdot u)\ w\ d\Gamma - \int_{\mathbf{\Omega}} (u \cdot \nabla w)\ d\Omega,\\ &= \int_{\mathbf{\Gamma}} (\mathbf{n}_x\ u_x + \mathbf{n}_y\ u_y + \mathbf{n}_z\ u_z)\ w\ d\Gamma\\ &- \int_{\mathbf{\Omega}} (u_x\ \partial_x w + u_y\ \partial_y w + u_z\ \partial_z w)\ d\Omega, \end{aligned} \tag{7.165}$$

where $\mathbf{n}$ is the outward unit normal vector to the boundary $\mathbf{\Gamma}$ and $\mathbf{n}_x, \mathbf{n}_y, \mathbf{n}_z$ are its x, y and z components. Furthermore u_x, u_y and u_z represent the x, y and z components of u. Now, integrals I_1, I_2, and I_3 correspond to the decomposition in x, y and z directions of the integral in the left hand side of formula (7.165). We can always have the following decomposition:

$$\int_{\mathbf{\Omega}} \partial_x u\ w\ d\Omega = \int_{\mathbf{\Gamma}} (\mathbf{n}_x\ u_x)\ w\ d\Gamma - \int_{\mathbf{\Omega}} u_x\ \partial_x w\ d\Omega, \tag{7.166}$$

$$\int_{\mathbf{\Omega}} \partial_y u\ w\ d\Omega = \int_{\mathbf{\Gamma}} (\mathbf{n}_y\ u_y)\ w\ d\Gamma - \int_{\mathbf{\Omega}} u_y\ \partial_y w\ d\Omega, \tag{7.167}$$

$$\int_{\mathbf{\Omega}} \partial_z u\ w\ d\Omega = \int_{\mathbf{\Gamma}} (\mathbf{n}_z\ u_z)\ w\ d\Gamma - \int_{\mathbf{\Omega}} u_z\ \partial_z w\ d\Omega. \tag{7.168}$$

We would still need to decompose the solution u into its x, y and z components which would require some projection operator. Hence, integrating these equations is far from ideal.

Let us introduce a second integration by parts formula,

$$\begin{aligned}\int_{\mathbf{\Omega}} \left(\nabla^2 u\ \right) w\ d\Omega &= \int_{\mathbf{\Omega}} (\nabla \cdot \nabla u\ )\, w\ d\Omega = \int_{\mathbf{\Omega}} \left(\partial_x^2 u + \partial_y^2 u + \partial_z^2 u\right)\ w\ d\Omega,\\ &= \int_{\mathbf{\Gamma}} (\mathbf{n} \cdot \nabla u)\ w\ d\Gamma - \int_{\mathbf{\Omega}} (\nabla u \cdot \nabla w)\ d\Omega,\\ &= \int_{\mathbf{\Gamma}} (\mathbf{n}_x\ \partial_x u + \mathbf{n}_y\ \partial_y u + \mathbf{n}_z\ \partial_z u)\ w\ d\Gamma\\ &- \int_{\mathbf{\Omega}} (\partial_x u\ \partial_x w + \partial_y u\ \partial_y w + \partial_z u\ \partial_z w)\ d\Omega, \end{aligned} \tag{7.169}$$

We use formula (7.169) in the weak formulation of the original wave equation

$$\int_{\mathbf{\Omega}} \partial_t^2 u_1\ w\ d\Omega = \int_{\mathbf{\Omega}} \left(\partial_x^2 u_1 + \partial_y^2 u_1 + p_z^2 u_1\right)\ w\ d\Omega + \int_{\mathbf{\Omega}} S\ w\ d\Omega, \tag{7.170}$$

and integrate by parts I_4:

$$\begin{aligned}\int_{\mathbf{\Omega}} \left(\partial_x^2 u_1 + \partial_y^2 u_1 + \partial_z^2 u_1\right)\ w\ d\Omega &= \int_{\mathbf{\Gamma}} (\mathbf{n}_x\ \partial_x u_2 + \mathbf{n}_y\ \partial_y u_3 + \mathbf{n}_y\ \partial_z u_4)\ w\ d\Gamma\\ &- \int_{\mathbf{\Omega}} (u_2\ \partial_x w + u_3\ \partial_y w + u_4\ \partial_z w)\ d\Omega. \end{aligned} \tag{7.171}$$

7.3.5 Final weak formulation

The final weak formulation is

$$\int_{\mathbf{\Omega}} \partial_t u_1 \; w \; d\Omega = \int_{\mathbf{\Omega}} u_5 \; w \; d\Omega, \tag{7.172}$$

$$\int_{\mathbf{\Omega}} \partial_t u_2 \; w \; d\Omega = \int_{\mathbf{\Omega}} \partial_x u_5 \; w \; d\Omega, \tag{7.173}$$

$$\int_{\mathbf{\Omega}} \partial_t u_3 \; w \; d\Omega = \int_{\mathbf{\Omega}} \partial_y u_5 \; w \; d\Omega, \tag{7.174}$$

$$\int_{\mathbf{\Omega}} \partial_t u_4 \; w \; d\Omega = \int_{\mathbf{\Omega}} \partial_z u_5 \; w \; d\Omega, \tag{7.175}$$

$$\begin{aligned} \int_{\mathbf{\Omega}} \partial_t u_5 \; w \; d\Omega &= \int_{\mathbf{\Gamma}} \left(\mathbf{n}_x \; \partial_x u_2 + \mathbf{n}_y \; \partial_y u_3 + \mathbf{n}_x \; \partial_x u_4 \right) \; w \; d\Gamma \\ &- \int_{\mathbf{\Omega}} \left(u_2 \; \partial_x w + u_3 \; \partial_y w \right) \; d\Omega + \int_{\mathbf{\Omega}} S(x,y,z,t) \; w \; d\Omega. \end{aligned} \tag{7.176}$$

We need to define the space of solutions and test functions. Let us define the space of measurable functions $\mathcal{V} = \mathcal{L}^2(\mathbf{\Omega})$ and the Hilbert spaces

$$\mathcal{W}_x = \mathcal{H}^1(\mathbf{\Omega}) = \left\{ w \in \mathcal{L}^2(\mathbf{\Omega}) \text{ and } \partial_x w \in \mathcal{L}^2(\mathbf{\Omega}) \right\}, \tag{7.177}$$

$$\mathcal{W}_y = \mathcal{H}^1(\mathbf{\Omega}) = \left\{ w \in \mathcal{L}^2(\mathbf{\Omega}) \text{ and } \partial_y w \in \mathcal{L}^2(\mathbf{\Omega}) \right\}, \tag{7.178}$$

$$\mathcal{W}_z = \mathcal{H}^1(\mathbf{\Omega}) = \left\{ w \in \mathcal{L}^2(\mathbf{\Omega}) \text{ and } \partial_z w \in \mathcal{L}^2(\mathbf{\Omega}) \right\}, \tag{7.179}$$

and

$$\mathcal{W}_{xyz} = \mathcal{H}^1(\mathbf{\Omega}) = \left\{ w \in \mathcal{L}^2(\mathbf{\Omega}) \text{ and } \left(\partial_x w + \partial_y w + \partial_z w \right) \in \mathcal{L}^2(\mathbf{\Omega}) \right\}. \tag{7.180}$$

Here, $\partial_t u_1$, $\partial_t u_2$, $\partial_t u_3$, $\partial_t u_4$ and $\partial_t u_5$ are all in $\mathcal{V}$, $u_2 \in \mathcal{W}_x$ and $u_3 \in \mathcal{W}_y$, and $u_4 \in \mathcal{W}_z$, whereas $u_5 \in \mathcal{W}_{xyz}$. Ultimately this means that the solution $u = u_1$ to the original wave equation belongs to the space

$$\mathcal{U} = \left\{ u \in \mathcal{L}^2(\mathbf{\Omega}), \; \left(\partial_x u + \partial_y u + \partial_z u \right) \in \mathcal{L}^2(\mathbf{\Omega}), \text{ and } \partial_t^2 u \in \mathcal{L}^2(\mathbf{\Omega}) \right\}. \tag{7.181}$$

7.3.6 Domain Discretization

The domain $\mathbf{\Omega}$ is decomposed into $N_E = N_{Ex} \times N_{Ey} \times N_{Ez}$ sub-domains $\mathbf{\Omega^k}$ such that

$$\bar{\mathbf{\Omega}} = \bigcup_{k=0}^{K} \bar{\mathbf{\Omega}}^{\mathbf{k}}, \quad \forall k, \; l \quad \mathbf{\Omega^k} \bigcap_{k \neq l} \mathbf{\Omega}^l = 0, \tag{7.182}$$

where $\bar{\mathbf{\Omega}}$ is the closure of the domain $\mathbf{\Omega}$. The weak formulation is applied in each subdomain $\mathbf{\Omega^k}$ individually,

$$\int_{\mathbf{\Omega^k}} \partial_t u_1^{\mathbf{k}}\, w^{\mathbf{k}}\, d\Omega = \int_{\mathbf{\Omega^k}} u_5^{\mathbf{k}}\, w^{\mathbf{k}}\, d\Omega, \tag{7.183}$$

$$\int_{\mathbf{\Omega^k}} \partial_t u_2^{\mathbf{k}}\, w^{\mathbf{k}}\, d\Omega = \int_{\mathbf{\Omega^k}} \partial_x u_5^{\mathbf{k}}\, w^{\mathbf{k}}\, d\Omega, \tag{7.184}$$

$$\int_{\mathbf{\Omega^k}} \partial_t u_3^{\mathbf{k}}\, w^{\mathbf{k}}\, d\Omega = \int_{\mathbf{\Omega^k}} \partial_y u_5^{\mathbf{k}}\, w^{\mathbf{k}}\, d\Omega, \tag{7.185}$$

$$\int_{\mathbf{\Omega^k}} \partial_t u_4^{\mathbf{k}}\, w^{\mathbf{k}}\, d\Omega = \int_{\mathbf{\Omega^k}} \partial_y u_5^{\mathbf{k}}\, w^{\mathbf{k}}\, d\Omega, \tag{7.186}$$

$$\begin{aligned}\int_{\mathbf{\Omega^k}} \partial_t u_5^{\mathbf{k}}\, w^{\mathbf{k}}\, d\Omega &= -\int_{\mathbf{\Omega^k}} \left(u_2^{\mathbf{k}}\, \partial_x w^{\mathbf{k}} + u_3^{\mathbf{k}}\, \partial_y w^{\mathbf{k}} + u_4^{\mathbf{k}}\, \partial_z w^{\mathbf{k}}\right)\, d\Omega + \int_{\mathbf{\Omega^k}} S^{\mathbf{k}}(x,y,z,t)\, w^{\mathbf{k}}\, d\Omega \\ &\quad + \int_{\mathbf{\Gamma^k}} \left(\mathbf{n}_x^{\mathbf{k}}\, \partial_x u_2^{\mathbf{k}} + \mathbf{n}_y^{\mathbf{k}}\, \partial_y u_3^{\mathbf{k}} + \mathbf{n}_z^{\mathbf{k}}\, \partial_z u_4^{\mathbf{k}}\right)\, w^{\mathbf{k}}\, d\Gamma.\end{aligned} \tag{7.187}$$

Note that the boundary terms in (7.187) are relevant only for all the elements that share at least one side with $\mathbf{\Gamma^k}$. If the source term $S(x,y,z,t)$ is given as a function of x, y, and z and t, then $S^{\mathbf{k}}$ is nothing else than the restriction of $S(x,y,z,t)$ to the subdomain $\mathbf{\Omega^k}$.

In each subdomain, the solutions $u_1^{\mathbf{k}}$, $u_2^{\mathbf{k}}$, $u_3^{\mathbf{k}}$, $u_4^{\mathbf{k}}$ and $u_5^{\mathbf{k}}$ are expanded into cardinal basis functions. In higher dimensions, the formulation of the basis comes from the tensor product of one dimensional Lagrangian interpolant basis $h_a(x)$. So the Lagrangian interpolants are chosen as basis functions in each dimension. We expand the unknowns as

$$\forall u_1^{\mathbf{k}} \in \mathcal{V}_h,\ u_1^{\mathbf{k}}(x,y,z,t) = \sum_{m=0}^{m=N}\sum_{n=0}^{n=N}\sum_{p=0}^{p=N} u_{1mnp}^{\mathbf{k}}(t)\, h_m(x)\, h_n(y)\, h_p(z), \tag{7.188}$$

$$\forall u_2^{\mathbf{k}} \in \mathcal{V}_h,\ u_2^{\mathbf{k}}(x,y,z,t) = \sum_{m=0}^{m=N}\sum_{n=0}^{n=N}\sum_{p=0}^{p=N} u_{2mnp}^{\mathbf{k}}(t)\, h_m(x)\, h_n(y)\, h_p(z), \tag{7.189}$$

$$\forall u_3^{\mathbf{k}} \in \mathcal{V}_h,\ u_3^{\mathbf{k}}(x,y,z,t) = \sum_{m=0}^{m=N}\sum_{n=0}^{n=N}\sum_{p=0}^{p=N} u_{3mnp}^{\mathbf{k}}(t)\, h_m(x)\, h_n(y)\, h_p(z), \tag{7.190}$$

$$\forall u_4^{\mathbf{k}} \in \mathcal{V}_h,\ u_4^{\mathbf{k}}(x,y,z,t) = \sum_{m=0}^{m=N}\sum_{n=0}^{n=N}\sum_{p=0}^{p=N} u_{4mnp}^{\mathbf{k}}(t)\, h_m(x)\, h_n(y)\, h_p(z), \tag{7.191}$$

$$\forall u_5^{\mathbf{k}} \in \mathcal{V}_h,\ u_5^{\mathbf{k}}(x,y,z,t) = \sum_{m=0}^{m=N}\sum_{n=0}^{n=N}\sum_{p=0}^{p=N} u_{5mnp}^{\mathbf{k}}(t)\, h_m(x)\, h_n(y)\, h_p(z). \tag{7.192}$$

Here we use the index conventions exposed in Table (6.5) and $u_{1mnp}^{\mathbf{k}}(t) = u_{1mnp}^{\mathbf{k}}(x,y,z,t)$ are the nodal basis coefficients. The space $\mathcal{V}_h = \mathcal{V} \bigcup \mathbb{P}_{N,k}(\mathbf{\Omega})$ is taken to be a subspace of $\mathcal{V}$ consisting of the tensor product of all piecewise high order polynomials of degree less than or equal to N defined on $\mathbf{\Omega^k}$. Furthermore, we have the definition

$$\mathbb{P}_{N,K}(\mathbf{\Omega}) = \left\{\theta \in \mathcal{L}^2(\mathbf{\Omega}),\quad \theta_{|_{\mathbf{\Omega^k}}} \in \mathbb{P}_N(\mathbf{\Omega^k}) \times \mathbb{P}_N(\mathbf{\Omega^k})\right\}. \tag{7.193}$$

We can also differentiate the unknowns with respect to x, y or t in the following manner:

$$\partial_x u_1^{\mathbf{k}}(x,y,z,t) = \sum_{m=0}^{m=N}\sum_{n=0}^{n=N}\sum_{p=0}^{p=N} u_{1mn}^{\mathbf{k}}(t)\ \partial_x h_m(x)\ h_n(y),\ h_n(z), \tag{7.194}$$

$$\partial_y u_1^{\mathbf{k}}(x,y,z,t) = \sum_{m=0}^{m=N}\sum_{n=0}^{n=N}\sum_{p=0}^{p=N} u_{1mn}^{\mathbf{k}}(t)\ \partial_y h_m(x)\ h_n(y),\ h_n(z), \tag{7.195}$$

$$\partial_z u_1^{\mathbf{k}}(x,y,z,t) = \sum_{m=0}^{m=N}\sum_{n=0}^{n=N}\sum_{p=0}^{p=N} u_{1mn}^{\mathbf{k}}(t)\ \partial_z h_m(x)\ h_n(y),\ h_n(z), \tag{7.196}$$

$$\partial_t u_1^{\mathbf{k}}(x,y,z,t) = \sum_{m=0}^{m=N}\sum_{n=0}^{n=N}\sum_{p=0}^{p=N} \dot{u}_{1mn}^{\mathbf{k}}(t)\ h_m(x)\ h_n(y),\ h_n(z). \tag{7.197}$$

The test function w is selected to be as the same as the basis functions $h(x) \times h(y) \times h(z)$ used for the unknowns $u_1^{\mathbf{k}}, u_2^{\mathbf{k}}$, $u_3^{\mathbf{k}}$ and $u_4^{\mathbf{k}}$, and therefore using Einstein summation convention,

$$w^{\mathbf{k}}(x,y,z) = w_{abc}^{\mathbf{k}} h_a(x)\ h_b(y)\ h_c(z), \tag{7.198}$$

where $w_{abc}^{\mathbf{k}} = 1$. As mentioned previously, the same test functions are used for each variable here but they could be different.

On each element $\mathbf{\Omega^k}$ there are $N_{GLL} = (N+1)^3$ nodal points but in total there are

$$N_g = [N_{Ex}\,(N_{GLL}-1)+1] \times [N_{Ey}\,(N_{GLL}-1)+1] \times [N_{Ez}\,(N_{GLL}-1)+1] \tag{7.199}$$

global nodal points. One needs to create a global numbering function that keeps track of local and global nodes on the domain $\mathbf{\Omega}$. There are many ways to label the elements and element nodes. The different protocols of element and node numbering have no effect on the spectral element solution itself but they have a huge impact on the structure of the global mass and advection matrices and therefore on the efficiency of the spectral element code. Figure (6.2) illustrates an example of global numbering technique in 2D for $N_E = 4$ and $N_{GLL} = 4$, the same process of global numbering in 3D is very similar. Figure (6.4) shows the local numbering convention per subdomain $\mathbf{\Omega^k}$ and corresponding elemental matrix storage in 3D. Let $\mathcal{I}$ denote global indices which are functions of the element index k and the indexes a, b, c within each element,

$$\mathcal{I} = \mathcal{I}(a,b,c,\mathbf{k}). \tag{7.200}$$

The global numbering function maps the local numbering of the computational nodes to their global (non-redundant) numbering. $\mathcal{I}(a,b,c,\mathbf{k})$ is the global node index of the (a,b,c)-th GLL node internal to the $\mathbf{k}$-th element. Elements are numbered row by row from bottom-left to top-right (see Figure(6.3)). The table of indices $\mathcal{I}(a,b,c,\mathbf{k})$ is typically needed to build or assemble global data from local data (contributions from each element).

7.3.7 Master Element

To apply the quadrature rule on each element, one needs to define an affine transformation to map each spectral element $\mathbf{\Omega^k}$ to the reference or master element $\Lambda \times \Lambda \times \Lambda = \Lambda^3$ in the

3D case. Let us define the local elemental mappings:

$$(x,y,z)^{\mathbf{k}} \;=\; (x,y,z)^{\mathbf{k}}_{abc}\; h_a(\xi)\; h_b(\eta)\; h_c(\zeta), \tag{7.201}$$

we can now map the physical elements $(x,y,z)^{\mathbf{k}} \in \mathbf{\Omega}^{\mathbf{k}}$ onto the computational domain $(\xi,\eta,\zeta) \in \Lambda^3$. We denote by $J^{\mathbf{k}}$ the Jacobian associated to this mapping such that

$$J^{\mathbf{k}} = \frac{\partial(x,y,z)^{\mathbf{k}}}{\partial(\xi,\eta,\zeta)} = \begin{pmatrix} \frac{\partial x^{\mathbf{k}}}{\partial \xi} & \frac{\partial x^{\mathbf{k}}}{\partial \eta} & \frac{\partial x^{\mathbf{k}}}{\partial \zeta} \\ \frac{\partial y^{\mathbf{k}}}{\partial \xi} & \frac{\partial y^{\mathbf{k}}}{\partial \eta} & \frac{\partial y^{\mathbf{k}}}{\partial \zeta} \\ \frac{\partial z^{\mathbf{k}}}{\partial \xi} & \frac{\partial z^{\mathbf{k}}}{\partial \eta} & \frac{\partial z^{\mathbf{k}}}{\partial \zeta} \end{pmatrix}. \tag{7.202}$$

By $\partial x^{\mathbf{k}}/\partial\xi$ we refer to $\partial x/\partial\xi$ for some point x in the kth element $\mathbf{\Omega}^{\mathbf{k}}$. We refer to $|J^{\mathbf{k}}|$ as the determinant of the Jacobian $J^{\mathbf{k}}$. This change of variable is a key component of the method and $|J^{\mathbf{k}}|$ appears in the elemental matrix discretization,

$$|J^{\mathbf{k}}| \;=\; \frac{\partial x^{\mathbf{k}}}{\partial \xi}\frac{\partial y^{\mathbf{k}}}{\partial \eta}\frac{\partial z^{\mathbf{k}}}{\partial \zeta} + \frac{\partial x^{\mathbf{k}}}{\partial \eta}\frac{\partial y^{\mathbf{k}}}{\partial \zeta}\frac{\partial z^{\mathbf{k}}}{\partial \xi} + \frac{\partial x^{\mathbf{k}}}{\partial \zeta}\frac{\partial y^{\mathbf{k}}}{\partial \xi}\frac{\partial z^{\mathbf{k}}}{\partial \eta} \tag{7.203}$$

$$- \;\frac{\partial x^{\mathbf{k}}}{\partial \zeta}\frac{\partial y^{\mathbf{k}}}{\partial \eta}\frac{\partial z^{\mathbf{k}}}{\partial \xi} - \frac{\partial x^{\mathbf{k}}}{\partial \xi}\frac{\partial y^{\mathbf{k}}}{\partial \zeta}\frac{\partial z^{\mathbf{k}}}{\partial \eta} - \frac{\partial x^{\mathbf{k}}}{\partial \eta}\frac{\partial y^{\mathbf{k}}}{\partial \xi}\frac{\partial z^{\mathbf{k}}}{\partial \zeta}. \tag{7.204}$$

Practically, derivatives with respect to the physical coordinate x are evaluated in terms of the computational coordinate ξ (respectively for y, η and z, ζ). The mapping from the element

$$(x,y,z)^{\mathbf{k}} \in \mathbf{\Omega}^{\mathbf{k}} = [X_{\mathbf{k}}, X_{\mathbf{k}+1}] \times [Y_{\mathbf{k}}, Y_{\mathbf{k}+1}] \times [Z_{\mathbf{k}}, Z_{\mathbf{k}+1}] \tag{7.205}$$

to the computational space used is

$$\xi \;=\; \frac{2}{\Delta x^{\mathbf{k}}}(x^{\mathbf{k}} - X_{\mathbf{k}}) - 1, \tag{7.206}$$

$$\eta \;=\; \frac{2}{\Delta y^{\mathbf{k}}}(y^{\mathbf{k}} - Y_{\mathbf{k}}) - 1 \tag{7.207}$$

$$\zeta \;=\; \frac{2}{\Delta z^{\mathbf{k}}}(z^{\mathbf{k}} - Z_{\mathbf{k}}) - 1 \tag{7.208}$$

where $\Delta x^{\mathbf{k}} = X_{k+1} - X_k$, $\Delta y^{\mathbf{k}} = Y_{k+1} - Y_k$ and $\Delta z^{\mathbf{k}} = Z_{k+1} - Z_k$ so that

$$\frac{\partial x^{\mathbf{k}}}{\partial \xi} = \frac{\Delta x^{\mathbf{k}}}{2} \quad \text{and} \quad \frac{\partial x^{\mathbf{k}}}{\partial \eta} = \frac{\partial x^{\mathbf{k}}}{\partial \zeta} = 0, \tag{7.209}$$

$$\frac{\partial y^{\mathbf{k}}}{\partial \eta} = \frac{\Delta y^{\mathbf{k}}}{2} \quad \text{and} \quad \frac{\partial y^{\mathbf{k}}}{\partial \xi} = \frac{\partial y^{\mathbf{k}}}{\partial \zeta} = 0, \tag{7.210}$$

$$\frac{\partial z^{\mathbf{k}}}{\partial \zeta} = \frac{\Delta z^{\mathbf{k}}}{2} \quad \text{and} \quad \frac{\partial z^{\mathbf{k}}}{\partial \xi} = \frac{\partial z^{\mathbf{k}}}{\partial \eta} = 0 \tag{7.211}$$

and hence, the determinant of the Jacobian is

$$|J^{\mathbf{k}}| \;=\; \frac{\Delta x^{\mathbf{k}}\,\Delta y^{\mathbf{k}}\,\Delta z^{\mathbf{k}}}{8}. \tag{7.212}$$

In particular, the $|J^{\mathbf{k}}|$ becomes the same for all the elements $\mathbf{\Omega}^{\mathbf{k}}$ in the case of a homogeneous (evenly decomposed) domain in the x, y, z directions.

7.3.8 Elemental matrix forms

On each subdomain, each integral is discretized in a similar fashion as for the 1D case. In higher dimensions, the number of subscripts and superscipts increase a great deal. ***In the following, we drop the superscript*** $\mathbf{k}$ ***that refers to the element*** $\mathbf{k}$ ***for clarity and simplification of the notations.***

Elemental Mass matrix M For the elemental mass matrix, we will go into more details and explanations. The Mass matrix $\mathbf{M}$ appears in the following type of integral

$$\int_{\mathbf{\Omega}} u\ w\ d\mathbf{\Omega} \quad = \quad \mathbf{M} \otimes u. \tag{7.213}$$

The test functions are non zero for only one nodal point (a, b, c) per element $\mathbf{\Omega^k}$ (see Figure (6.3)). In a general manner, we apply the method of weighted residuals to the integral and write for all three values $(a, b, c) \in \{0, N\}^3$

$$\int_{\mathbf{\Omega}} u\ w_{abc}\ d\mathbf{\Omega} = \int_{\mathbf{\Omega}} u_{mnp} h_m(x) h_n(y) h_p(z)\ h_a(x) h_b(y) h_c(z)\ dxdydz. \tag{7.214}$$

Note that we are now using the Einstein summation convention, equation (7.214) has three sums on m, n and p in the right hand side. The first step is to do a change of variables from the physical coordinate to the computational coordinate (isoparametric element) and then apply the GLL quadrature rule to the integral with weights ρ_{qrs}.

$$\int_{\mathbf{\Omega}} u\ w_{abc}\ d\mathbf{\Omega} \quad = \quad \int_{\Lambda^3} u_{mnp} h_m(\xi) h_n(\eta) h_p(\zeta)\ h_a(\xi) h_b(\eta) h_c(\zeta)\ |J| d\xi d\eta d\zeta \tag{7.215}$$

$$= \quad \rho_{qrs} u_{mnp} h_m(\xi_q) h_n(\eta_r) h_p(\zeta_s)\ h_a(\xi_q) h_b(\eta_r) h_c(\zeta_s)\ |J|. \tag{7.216}$$

Note that the right hand side of equation (7.216) has 6 sums over the indices q, r, s, m, n and p in this order from left to right for values in $\{0, N\}$. Remember that in this elemental representation (Figure (6.3) for 2D), u, w and ρ are $N_{GLL} \times N_{GLL} \times N_{GLL}$ matrices in 3D.

Now we use the properties of the Legendre interpolants, in particular equation (6.61) states that

$$h_i(\xi_j) = \delta_{ij}. \tag{7.217}$$

We can then write,

$$\int_{\mathbf{\Omega}} u\ w_{abc}\ d\mathbf{\Omega} \quad = \quad \rho_{qrs} u_{mnp} \delta_{qm} \delta_{rn} \delta_{sp}\ \delta_{qa} \delta_{rb} \delta_{sc}\ |J|. \tag{7.218}$$

To adopt a matrix notation, one of the key formulae are the following matrix products between any two matrices A and B:

$$(A \cdot_{xy} B)_{ijl} \quad = \quad \sum_{r=0}^{N} a_{irl}\ b_{rjl}; \tag{7.219}$$

$$(A \cdot_{yz} B)_{ijl} \quad = \quad \sum_{r=0}^{N} a_{ijr}\ b_{irl}. \tag{7.220}$$

These 3D matrix products are the natural extension to the well known 2D formula for each of the extra dimensions,

$$(A \cdot B)_{ij} \;=\; \sum_{r=0}^{N} a_{ir}\, b_{rj}. \tag{7.221}$$

Let us also define the scalar matrix product, also known as the Hadamard matrix product, by

$$(A : B)_{ijl} \;=\; a_{ijl}\, b_{ijl}. \tag{7.222}$$

To obtain the elemental mass matrix in 3D, we deal with index after index from right to left that is p, n, then m, then s,r and finally q. Let's reformulate equation (7.214) with this procedure in mind.

$$\begin{aligned}
\int_{\mathbf{\Omega}} u\; w_{abc}\; d\mathbf{\Omega} \;&=\; \rho_{qrs} u_{mnp} \delta_{qm} \delta_{rn} \delta_{sp}\; \delta_{qa} \delta_{rb} \delta_{sc}\; |J|, && (7.223)\\
&=\; \rho_{qrs} u_{qrs}\; \delta_{qa} \delta_{rb} \delta_{sc}\; |J|, && (7.224)\\
&=\; (\rho : u)_{qrs}\; \delta_{qa} \delta_{rb} \delta_{sc}\; |J|, && (7.225)\\
&=\; (\rho : u)_{qrc}\; \delta_{qa} \delta_{rb}\; |J|, && (7.226)\\
&=\; (\rho : u)_{abc}\; |J|. && (7.227)
\end{aligned}$$

So in terms of an elemental or local representation we have for each physical node (a, b, c) in each element $\mathbf{\Omega^k}$ the following relation for the 3D matrices

$$\begin{pmatrix} \cdots & & \\ \vdots & \int_{\mathbf{\Omega}} u\; w_{abc}\; d\mathbf{\Omega} & \vdots \\ \cdots & & \end{pmatrix}_{\mathbf{k}} \;=\; \underbrace{\begin{pmatrix} \cdots & & \\ \vdots & \rho_{abc} |J| & \vdots \\ \cdots & & \end{pmatrix}_{\mathbf{k}}}_{\mathbf{M}} : \underbrace{\begin{pmatrix} \cdots & & \\ \vdots & u_{abc} & \vdots \\ \cdots & & \end{pmatrix}_{\mathbf{k}}}_{u} \tag{7.228}$$

Elemental advection matrix $\mathbf{A}_k$ type 1 The advection matrix $\mathbf{A}_k$ appears in the following type of integral (see Appendix F for ***general*** shaped elements)

$$\int_{\mathbf{\Omega}} f \partial_k u\; w\; d\mathbf{\Omega} \;=\; f : (\mathbf{A}_k \otimes u)\,, \tag{7.229}$$

where, f, u and w are scalar functions and $k = x, y$, or z. The operator $\otimes$ will act on u in a different manner depending on the value of k as described below:

$$\begin{aligned}
\mathbf{A}_x \otimes u \;&=\; \rho : (H \cdot_{xy} u)\, |J| \frac{\partial \xi}{\partial x} && (7.230)\\
\mathbf{A}_y \otimes u \;&=\; \rho : \left(u \cdot_{xy} H^T\right) |J| \frac{\partial \eta}{\partial y} && (7.231)\\
\mathbf{A}_z \otimes u \;&=\; \rho : \left(u \cdot_{yz} H^T\right) |J| \frac{\partial \zeta}{\partial z}. && (7.232)
\end{aligned}$$

Remember that the elemental unknown u is represented by a $(N_{GLL})^3$ matrix. So the notation $\cdot_{xy}$ means a matrix product in the xy direction for each z dimension. And the notation $\cdot_{yz}$ means a matrix product in the yz direction for each x dimension. More explicitly,

$$\left(u \cdot_{xy} H^T\right)_{abc} = \sum_i u_{aic} H^T_{ib} \tag{7.233}$$

$$\left(u \cdot_{yz} H^T\right)_{abc} = \sum_i u_{abi} H^T_{ic}. \tag{7.234}$$

Furthermore, the H matrix represents the first derivative of the Legendre interpolants.

Elemental advection matrix $\mathbf{D}_k$ type 2 The advection matrix $\mathbf{D}_k$ appears in the following type of integral

$$\int_{\mathbf{\Omega}} f u \, \partial_k w \, d\mathbf{\Omega} = f : (\mathbf{D}_k \otimes u), \tag{7.235}$$

where, f, u and w are scalar functions and $k = x, y$, or z. The operator $\otimes$ will act on u in a different manner depending on the value of k as described below:

$$\mathbf{D}_x \otimes u = \left[H^T \cdot_{xy} \left(|J| \frac{\partial \xi}{\partial x} : \rho : u \right) \right] \tag{7.236}$$

$$\mathbf{D}_y \otimes u = \left[\left(|J| \frac{\partial \eta}{\partial y} : \rho : u \right) \cdot_{xy} H \right] \tag{7.237}$$

$$\mathbf{D}_z \otimes u = \left[\left(|J| \frac{\partial \zeta}{\partial z} : \rho : u \right) \cdot_{yz} H \right]. \tag{7.238}$$

Elemental boundary terms $\mathbf{B}$ The treatment of the elemental boundary terms are very special. If we integrate by parts on each element separately there would be lots of boundary terms at the interior boundaries. The spectral element strategy is to ignore these extra interior terms and only take into account the terms on the elements at the boundaries. In 3D and in a rectangular domain, this means that we look at six boundaries $\mathbf{\Gamma}_1$, $\mathbf{\Gamma}_2$, $\mathbf{\Gamma}_3$, $\mathbf{\Gamma}_4$, $\mathbf{\Gamma}_5$ and $\mathbf{\Gamma}_6$ corresponding to the six faces of the domain. The boundary integral in our problem is defined by

$$\int_{\mathbf{\Gamma^k}} \left(\mathbf{n}^{\mathbf{k}}_x \, \partial_x u^{\mathbf{k}}_2 + \mathbf{n}_y \, \partial_y u^{\mathbf{k}}_3 + \mathbf{n}_y \, \partial_z u^{\mathbf{k}}_4 \right) w^{\mathbf{k}} \, d\Gamma. \tag{7.239}$$

Figure 7.8 illustrates the six faces of the domain and the corresponding values of Γ_i, $d\Gamma$ and $\mathbf{n}$. We use some absorbing boundary conditions on the 6 boundaries so that

$$\text{On } \mathbf{\Gamma}_i : \frac{x_i}{R_i} \, \partial_t u\big|_{\mathbf{\Gamma}_i} + \partial_i u\big|_{\mathbf{\Gamma}_i} + \frac{x_i}{R_i^2} \, u\big|_{\mathbf{\Gamma}_i} = b_i, \tag{7.240}$$

where $R_i = \left(\sqrt{x^2 + y^2 + z^2} \right)\Big|_{\mathbf{\Gamma}_i}$. The boundary functions $b_i, i = 1, 6$ are not exactly zero but are very close to zero numerically. Each face has specific values for the outward unit

normal and the 2D jacobian $d\mathbf{\Gamma}$ of the the surface. Specifically,

$$\mathbf{\Gamma}_1: \quad (x,y,-L_Z) \quad d\mathbf{\Gamma} = dx\,dy \quad \mathbf{n}_1 = (0,0,-1); \tag{7.241}$$
$$\mathbf{\Gamma}_2: \quad (x,y,+L_Z) \quad d\mathbf{\Gamma} = dx\,dy \quad \mathbf{n}_2 = (0,0,+1); \tag{7.242}$$
$$\mathbf{\Gamma}_3: \quad (x,-L_Y,z) \quad d\mathbf{\Gamma} = dx\,dz \quad \mathbf{n}_3 = (0,-1,0); \tag{7.243}$$
$$\mathbf{\Gamma}_4: \quad (x,+L_Y,z) \quad d\mathbf{\Gamma} = dx\,dz \quad \mathbf{n}_4 = (0,+1,0); \tag{7.244}$$
$$\mathbf{\Gamma}_5: \quad (-L_X,y,z) \quad d\mathbf{\Gamma} = dy\,dz \quad \mathbf{n}_5 = (-1,0,0); \tag{7.245}$$
$$\mathbf{\Gamma}_6: \quad (+L_X,y,z) \quad d\mathbf{\Gamma} = dy\,dz \quad \mathbf{n}_6 = (+1,0,0). \tag{7.246}$$

The boundary term on $\mathbf{\Gamma}_1$ will be treated in more detail. First of all, applying the boundary conditions (7.240) to the face $\mathbf{\Gamma}_1$, we obtain:

$$\text{On } \mathbf{\Gamma}_1: \frac{-L_Z}{R_1}\,\partial_t u\big|_{\mathbf{\Gamma}_1} + \partial_z u\big|_{\mathbf{\Gamma}_1} + \frac{-L_Z}{R_1^2}\,u\big|_{\mathbf{\Gamma}_1} = b_1, \tag{7.247}$$

For the wave equation written as a hyperbolic system, equation (7.247) implies a relation between the variables u_2, u_3, u_4 and u_5:

$$\text{On } \mathbf{\Gamma}_1: \frac{-L_Z}{R_1}\,u_5\big|_{\mathbf{\Gamma}_1} + u_4\big|_{\mathbf{\Gamma}_1} + \frac{-L_Z}{R_1^2}\,u_1\big|_{\mathbf{\Gamma}_1} = b_1, \tag{7.248}$$

$$\int_{\mathbf{\Gamma}_1} (\mathbf{n}_x u_2 + \mathbf{n}_y u_3 + \mathbf{n}_z u_4) w_{abc} d\mathbf{\Gamma} = \int_{\mathbf{\Gamma}_1} (0 + 0 - u_4) h_a(x) h_b(y) h_c(-L_Z) dx dy. \tag{7.249}$$

We can now replace the right hand side of equation (7.249) by the relation between the variables in equation (7.247), and obtain, the following discretization

$$\int_{\mathbf{\Gamma}_1} (-u_4) w_{abc} d\mathbf{\Gamma} = -\int_{\mathbf{\Gamma}_1} \left(\frac{L_Z}{R_1} u_5 + \frac{L_Z}{R_1^2} + b_1 \right) h_a(x) h_b(y) h_c(-L_Z) dx dy. \tag{7.250}$$

Since $R_1 = R(x,y,-L_Z)$ in the x,y and z coordinates, it corresponds to $R(\xi_r,\eta_r,\zeta_0) = R_{qr0}$ in terms of elemental notations. After a change of variable to the 3D master element and applying the quadrature rule in 3D, we obtain:

$$\int_{\mathbf{\Gamma}_1} (-u_4) w_{abc} d\mathbf{\Gamma} = -\rho_{qr0} \left(\frac{L_Z}{R_{qr0}} u_{5mnp} + \frac{L_Z}{R_{1qr0}^2} u_{1mnp} + b_1 \right) \delta_{qm}\delta_{rn}\delta_{0p}\delta_{qa}\delta_{rb}\delta_{0c} \frac{\partial x}{\partial \xi}\frac{\partial y}{\partial \eta}|J|. \tag{7.251}$$

After simplification done in a very similar as was the 1D case, we define the elemental boundary matrix $\mathbf{B}_{\mathbf{\Gamma}_1}$ of size $N_{GLL} \times N_{GLL} \times N_{GLL}$ so that

$$\text{On } \mathbf{\Gamma}_1: \ (\mathbf{B}_{\mathbf{\Gamma}_1} \otimes (-u_4))_{abc} = -\frac{\Delta x \Delta y \Delta z}{8} \rho_{00} \left(\frac{L_Z}{R_{ab0}} u_{ab0} + \frac{L_Z}{R_{1ab0}^2} u_{1ab0} + b_1 \right). \tag{7.252}$$

Only the bottom face of this 3D elemental boundary matrix is non zero, it corresponds to the nodes of the bottom boundary in the domain $\mathbf{\Omega}$ in Figure 7.8. The other 5 elemental boundary matrices on $\mathbf{\Gamma}_2$, $\mathbf{\Gamma}_3$, $\mathbf{\Gamma}_4$ and $\mathbf{\Gamma}_5$ are determined very similarly.

Finally, the total ***elemental boundary matrix*** $\mathbf{B}$ is defined as follows

$$\mathbf{B} = \mathbf{B}_{\mathbf{\Gamma}_1} + \mathbf{B}_{\mathbf{\Gamma}_2} + \mathbf{B}_{\mathbf{\Gamma}_3} + \mathbf{B}_{\mathbf{\Gamma}_4} + \mathbf{B}_{\mathbf{\Gamma}_5} + \mathbf{B}_{\mathbf{\Gamma}_6}. \tag{7.253}$$

Elemental matrix system

In each subdomain $\mathbf{\Omega}^{\mathbf{k}}$, we have the following discretization of the weak formulation of equations (7.183), (7.184), (7.185), and (7.186),

$$\mathbf{M}^{\mathbf{k}} \otimes \dot{u}_1^{\mathbf{k}} = \mathbf{M}^{\mathbf{k}} \otimes u_4^{\mathbf{k}}, \tag{7.254}$$

$$\mathbf{M}^{\mathbf{k}} \otimes \dot{u}_2^{\mathbf{k}} = \mathbf{A}_x^{\mathbf{k}} \otimes u_4^{\mathbf{k}}, \tag{7.255}$$

$$\mathbf{M}^{\mathbf{k}} \otimes \dot{u}_3^{\mathbf{k}} = \mathbf{A}_y^{\mathbf{k}} \otimes u_4^{\mathbf{k}}, \tag{7.256}$$

$$\mathbf{M}^{\mathbf{k}} \otimes \dot{u}_4^{\mathbf{k}} = \mathbf{B}^{\mathbf{k}} \otimes u_4^{\mathbf{k}} - \left(\mathbf{D}_x^{\mathbf{k}} \otimes u_2^{\mathbf{k}} + \mathbf{D}_y^{\mathbf{k}} \otimes u_3^{\mathbf{k}} \right). \tag{7.257}$$

7.3.9 Assembly of global discretization matrix

All the elemental contributions need to be added together this is called ***the assembly of the global matrix***. In general, consider a local matrix $\mathbf{A}^{\mathbf{k}}$, where here the superscipt $\mathbf{k}$ represents the $\mathbf{k}$th-element, the global matrix $\mathbf{A}$ is noted

$$\mathbf{A} = \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \mathbf{A}^{\mathbf{k}}, \tag{7.258}$$

where $\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}'$ represents the assembly summation.

From a computational point of view the assembly of the local advection matrix in the x direction $\mathbf{A}_x$,

$$\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \left(\mathbf{A}_x^{\mathbf{k}} \otimes u^{\mathbf{k}} \right), \tag{7.259}$$

is constructed by first initializing the matrix $\mathbf{A}_x^{\mathbf{k}}$ to zero, and then beginning an outer loop over the element index $\mathbf{k}$. For each value $\mathbf{k}$, there are two inner loops over the indices a, b and c to calculate all the rows and columns that are affected by the element $\mathbf{\Omega}^{\mathbf{k}}$. For example

$$\begin{aligned} & a, b, c \in \{1, N_{GLL}\} \\ & \left(\mathbf{A}_x^{\mathbf{k}} \otimes u^{\mathbf{k}} \right)_{\mathcal{I}(a,b,c,\mathbf{k})} := \left(\mathbf{A}_x^{\mathbf{k}} \otimes u^{\mathbf{k}} \right)_{\mathcal{I}(a,b,c,\mathbf{k})} + \left(\mathbf{A}_x^{\mathbf{k}} \otimes u^{\mathbf{k}} \right)_{abc}. \end{aligned} \tag{7.260}$$

This can also be written in matrix language over a loop over the elements k (Matlab)

$$\left(\mathbf{A}_x^{\mathbf{k}} \otimes u^{\mathbf{k}} \right)_{\mathcal{I}(:,:,:,\mathbf{k})} := \left(\mathbf{A}_x^{\mathbf{k}} \otimes u^{\mathbf{k}} \right)_{\mathcal{I}(:,:,:,\mathbf{k})} + \left(\rho^{\mathbf{k}} : \left(H \cdot_{xy} u^{\mathbf{k}} \right) \right) |J| \frac{\partial \xi}{\partial x}. \tag{7.261}$$

Note that the indices a, b and c run from $\{1, N_{GLL}\}$ instead of $\{0, N\}$ for computational reasons. In order to limit memory storage it is useful to perform calculations directly while assembling.

The assembly process will result in a global system of n-***variables*** in matrix form.

The global system of our problem is therefore given by

$$\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E} {}'\left(\mathbf{M}^{\mathbf{k}} \otimes \dot{u}_1^{\mathbf{k}}\right) = \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E} {}'\left(\mathbf{M}^{\mathbf{k}} \otimes u_5^{\mathbf{k}}\right), \tag{7.262}$$

$$\sum_{k=1}^{k=N_E} {}'\left(\mathbf{M}^{\mathbf{k}} \otimes \dot{u}_2^{\mathbf{k}}\right) = \sum_{k=1}^{k=N_E} {}'\left(\mathbf{A}_x^{\mathbf{k}} \otimes u_5^{\mathbf{k}}\right), \tag{7.263}$$

$$\sum_{k=1}^{k=N_E} {}'\left(\mathbf{M}^{\mathbf{k}} \otimes \dot{u}_3^{\mathbf{k}}\right) = \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E} {}'\left(\mathbf{A}_y^{\mathbf{k}} \otimes u_5^{\mathbf{k}}\right), \tag{7.264}$$

$$\sum_{k=1}^{k=N_E} {}'\left(\mathbf{M}^{\mathbf{k}} \otimes \dot{u}_4^{\mathbf{k}}\right) = \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E} {}'\left(\mathbf{A}_z^{\mathbf{k}} \otimes u_5^{\mathbf{k}}\right), \tag{7.265}$$

$$\sum_{k=1}^{k=N_E} {}'\left(\mathbf{M}^{\mathbf{k}} \otimes \dot{u}_5^{\mathbf{k}}\right) = \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E} {}'\left(\mathbf{B}^{\mathbf{k}} u_4^{\mathbf{k}} - \left(\mathbf{D}_x^{\mathbf{k}} \otimes u_2^{\mathbf{k}} + \mathbf{D}_y^{\mathbf{k}} \otimes u_3^{\mathbf{k}} + \mathbf{D}_z^{\mathbf{k}} \otimes u_4^{\mathbf{k}}\right)\right). \tag{7.266}$$

In order to simplify the notations, we adopt the following convention

$$\sum_{k=1}^{k=N_E} {}'\left(\mathbf{A}_x^{\mathbf{k}} \otimes u_4^{\mathbf{k}}\right) = \mathbf{A}_x \otimes u_4. \tag{7.267}$$

The final global system of the 4 unknowns in matrix form is

$$\mathbf{M} \otimes \dot{u}_1 = \mathbf{M} \otimes u_4, \tag{7.268}$$

$$\mathbf{M} \otimes \dot{u}_2 = \mathbf{A}_x \otimes u_4, \tag{7.269}$$

$$\mathbf{M} \otimes \dot{u}_3 = \mathbf{A}_y \otimes u_4, \tag{7.270}$$

$$\mathbf{M} \otimes \dot{u}_4 = \mathbf{A}_y \otimes u_5, \tag{7.271}$$

$$\mathbf{M} \otimes \dot{u}_5 = \mathbf{B} \otimes u_4^{\mathbf{k}} - \left(\mathbf{D}_x \otimes u_2 + \mathbf{D}_y \otimes u_3 + \mathbf{D}_y \otimes u_4\right). \tag{7.272}$$

To simplify the notations further, we introduce the system in a similar fashion as in (7.158)

$$\dot{U} = \mathcal{A}U, \tag{7.273}$$

that is

$$\underbrace{\begin{pmatrix} \dot{u}_1 \\ \dot{u}_2 \\ \dot{u}_3 \\ \dot{u}_4 \\ \dot{u}_5 \end{pmatrix}}_{\dot{U}} = \underbrace{\begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & \frac{\mathbf{A}_x}{\mathbf{M}} \\ 0 & 0 & 0 & 0 & \frac{\mathbf{A}_y}{\mathbf{M}} \\ 0 & 0 & 0 & 0 & \frac{\mathbf{A}_z}{\mathbf{M}} \\ 0 & \frac{\mathbf{B}-\mathbf{D}_x}{\mathbf{M}} & \frac{\mathbf{B}-\mathbf{D}_y}{\mathbf{M}} & \frac{\mathbf{B}-\mathbf{D}_z}{\mathbf{M}} & 0 \end{pmatrix}}_{\mathcal{A}} \underbrace{\begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{pmatrix}}_{U} \tag{7.274}$$

7.3.10 Time Discretization

The time discretization of the system

$$\dot{U} = \mathcal{A}U = f(U,t), \tag{7.275}$$

is computed by an explicit fourth order Runge–Kutta method. Given an initial condition U_0, the solution U_{n+1} at time t_{n+1} is determined from the previous time t_n and the solution U_n. The details of the Runge–Kutta fourth order method can be found more explicitly in 6.10.

7.4 Numerical results in 3D

In this section we show some numerical results obtained with the spectral element method with a spherical 3D wave equation with no source term.

For our particular problem the timestep Δt is set by the stability condition CFL by the following relation in 3D:

$$\text{CFL} = \max\left(\frac{\Delta t}{\Delta s}\right), \tag{7.276}$$

where $\Delta s = \sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}$ so

$$\max(\Delta t) \leqslant \text{CFL} \times \min(\Delta s). \tag{7.277}$$

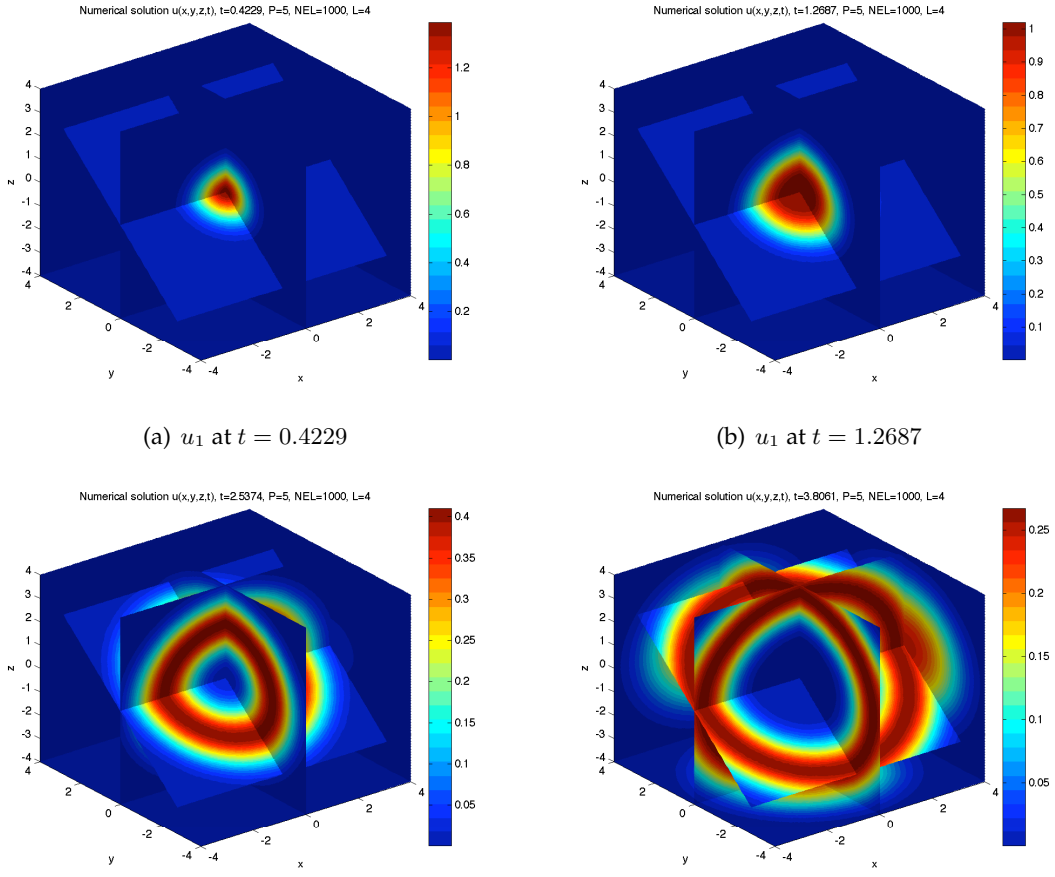
Note that this formula is slightly different than the 1D case.

To compare the exact and numerical solution we calculate the $\mathcal{L}^2$ norm given by equation 7.135.

The numerical solution is represented in figure 7.10 as a function of time t and radius r, whereas figure 7.9 shows the solution as specific time steps and in Cartesian coordinates x, y and z.

7.4.1 $\mathcal{L}^2$ norm and hp-convergence in 3D

Figure (7.11) show the numerical norm $\mathcal{L}^2$ for varying accuracies with $N_E = 9, 17$ number of elements, and polynomial orders $N = 5, 9, 15$ and for a domain where $L = 4$. In Table



(a) u_1 at $t = 0.4229$

(b) u_1 at $t = 1.2687$

(c) u_1 at $t = 2.5374$

(d) u_1 at $t = 3.8061$

Figure 7.9: Numerical solution u_1 at several time steps for $P = 5$, $N_E = 1000$, $L = 4$ and $CFL = 0.5$.

N \ N_E	7^3	9^3	11^3	13^3	15^3	17^3
5	46 656	97 336	175 616	287 496	438 976	636 056
7	125 000	262 144	474 552	778 688	1 191 016	1 728 000
9	262 144	551 368	1 000 000	1 643 032	2 515 456	3 652 2264
11	474 552	1 000 000	1 815 848	2 985 984	4 574 296	6 644 672
13	778 688	1 643 032	2 985 984	4 913 000	7 529 536	10 941 048
15	1 191 016	2 515 456	4 574 296	7 529 536	11 543 176	16 777 216

Table 7.2: Degrees of freedom N_g (total number of points) as a function of the polynomial order N and the number of elements N_E in 3D.

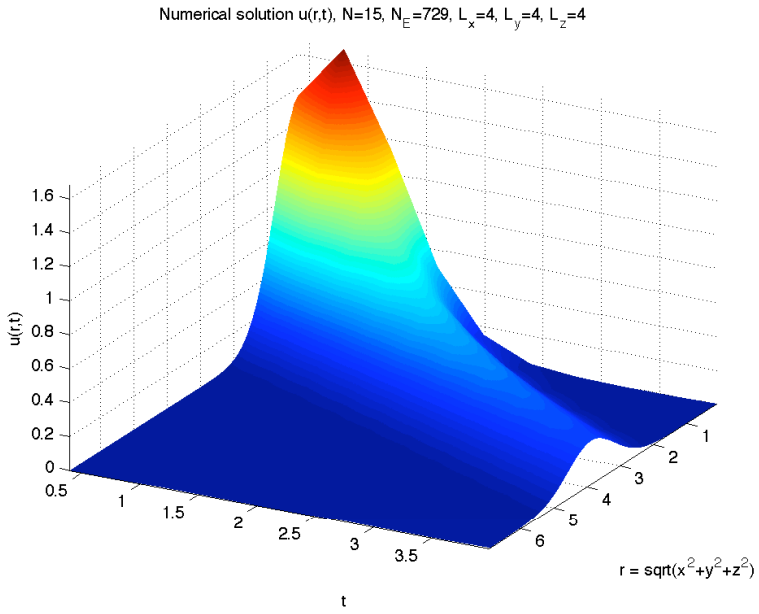


Figure 7.10: Numerical solution u_1 as a function of r and t for $P = 15$, $N_{Ex} = N_{Ey} = N_{Ez} = 9$ so $N_E = 729$, $L = 4$ and $CFL = 0.5$. The total number of space points is $N_g = 2515456$

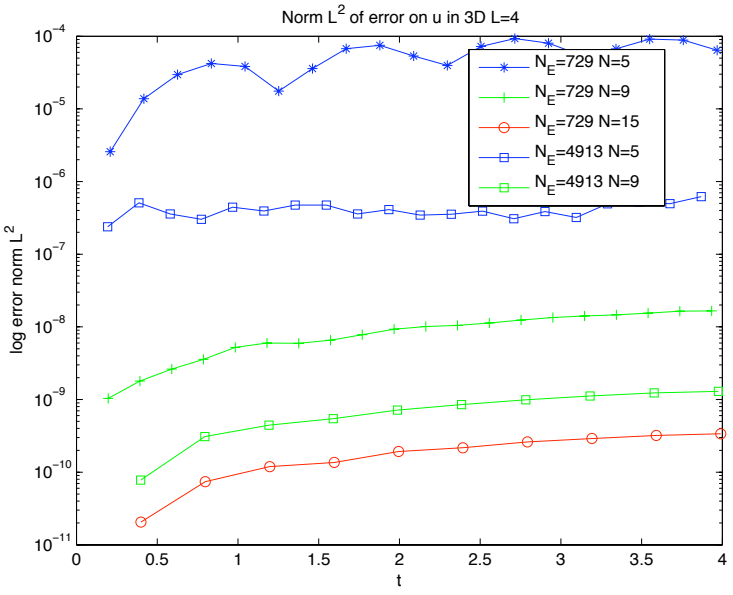


Figure 7.11: $\mathcal{L}^2$ norm of the numerical and exact solution u_1 for varying polynomial order $P = 5, 9, 15$ and number of elements $N_E = 9, 17$ for a domain $L = 4$, with $CFL = 0.5$.

7.4.1, we give the number of degree of freedom or total number of points N_g, for each of the different combinations of polynomial order N and number of elements N_E.

Figure 7.5 shows the $\mathcal{L}^2$ norm as a function of the total degree of freedom (total number of points N_g), for both the h-refinement with a fixed polynomial order N, and a p-refinement based on an evenly decomposed mesh.

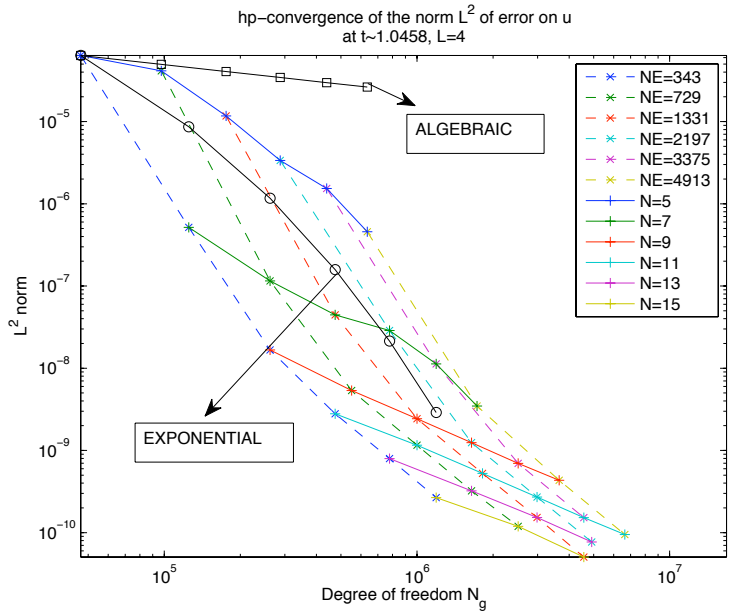


Figure 7.12: hp-convergence for the $\mathcal{L}^2$ norm of the numerical and exact solution u_1 as a function of the number of points N_g. We fix the polynomial order N and vary the number of elements N_E (h-convergence in solid lines), and we fix the number of elements N_E and vary the polynomial order N (p-convergence in dashed lines). See Table 7.4.1 for the values of N and N_E. The norms are taken at $t = 1$ for a domain $L = 4$, with $CFL = 0.5$

The h-refinement initially resolves the solution faster than the p-refinement, however, as the asymptotic exponential convergence is achieved the p-refinement takes over the h-refinement process.

The optimum convergence path as a function of degrees of freedom N_g, involves using both h and p-refinement.

7.4.2 Experiments on Sommerfeld Boundary conditions in 3D

We now focus our attention to the numerical errors coming from the boundary conditions.

- Figure (7.4.2) show the numerical $\mathcal{L}^2$ norm for the same polynomial order $N = 5$ and the same ratio of elements $N_E = 4L$ for various values of length of domains $L = 2, 3, 4, 5, 6$.

All norms clearly suggest that with the same resolution in space and time, the error decreases as the boundary is pushed further away.

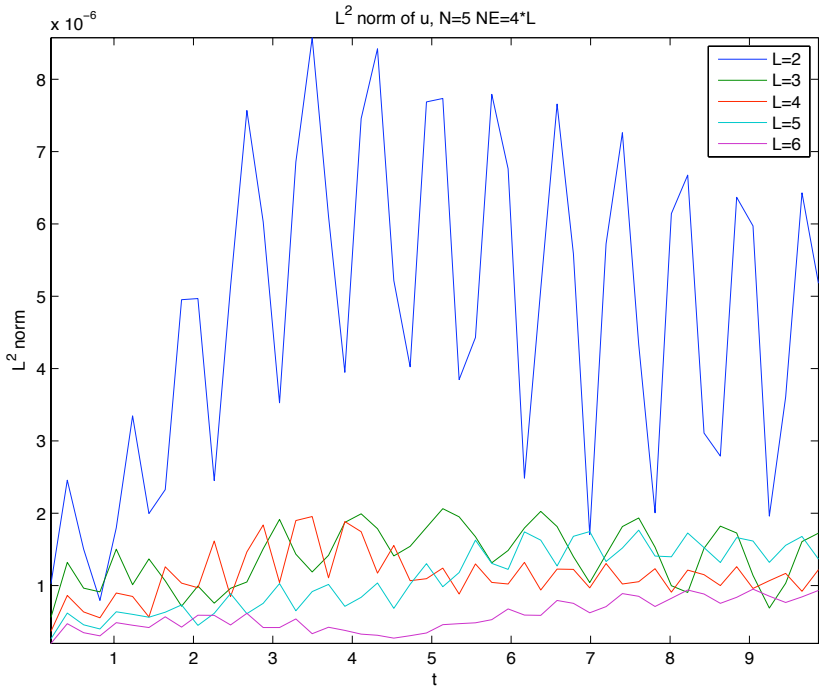


Figure 7.13: Convergence test on the Sommerfeld boundary conditions in 3D. For the same accuracy in space and time, the domain is successively $L = 2, 3, 4, 5, 6$. The error decreases as the boundary is pushed further away.

7.4.3 Convergence in time

Figure 7.14 shows the fourth order convergence in time with the Runge–Kutta method. Recall that a scheme is fourth order convergent if the norm $\text{Norm}_{\Delta t}$ obtained with timestep Δt is $(\text{Norm}_{\Delta t}) = (2^4 \times \text{Norm}_{\Delta t/2})$ where $\text{Norm}_{\Delta t/2}$ is the norm obtained with a time-stepping of $\Delta t/2$ and the same spatial resolution. In the figure we see that both norms are on top of each other which shows a fourth order convergent scheme in time.

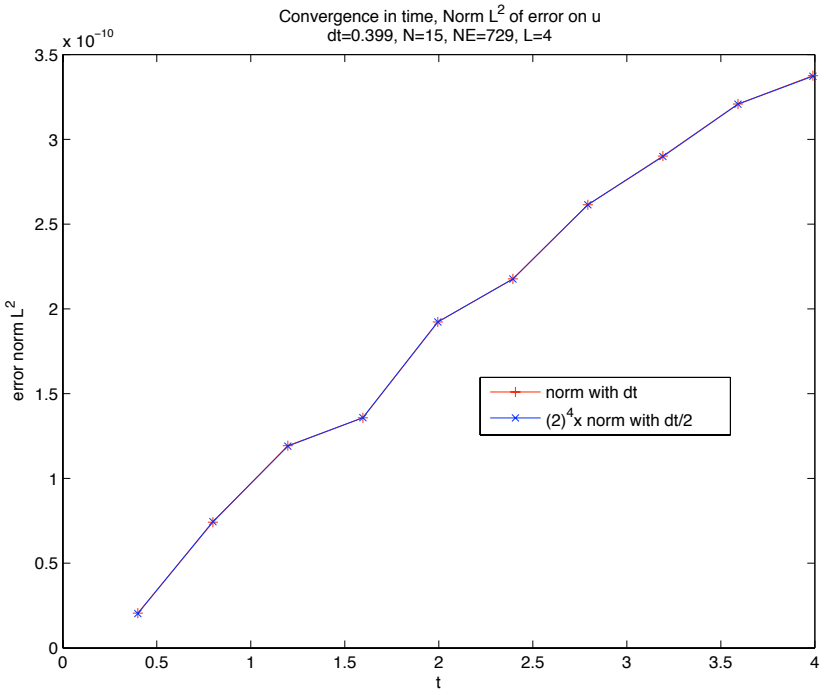


Figure 7.14: Fourth-order convergence in time for the Runge-Kutta method. The $\mathcal{L}^2$ norms are given for the same spatial accuracy but for Δt and $\Delta t/2$, we can see that $(\text{Norm}_{\Delta t}) = (2^4 \times \text{Norm}_{\Delta t/2})$ which shows a fourth order convergent scheme. The domain is $L = 4$ with a polynomial order $N = 15$ a number of elements $N_E = 9^3$.

7.5 Conclusion

In this Chapter, we have presented 2 applications of the spectral element method to the wave equation reformulated as a hyperbolic equation of first order in time and space in 1D and 3D. We have seen how to obtain a suitable weak formulation from the variational principle in the 1D case and 3D case.

We have derived the spectral element discretization of this 1D and 3D hyperbolic system in some detail, explaining how the most general elemental matrix forms of the system are calculated. We have also illustrated the local numbering convention in 3D that we have been using for this problem, and that we will be using for the BSSN system. We have also

presented the global system of algebraic equations of the problem to solve in 1D and 3D. For explicit time stepping schemes, such a Runge–Kutta fourth order, there are no full matrices (non-sparse) to invert as the Mass matrix is diagonal due to the choice of the GLL quadrature.

We have presented numerical results in both the 1D and 3D case, showing the advantages of the hp-convergence, and recovering the fourth order convergence in time of the Runge–Kutta method.

"As far as the laws of mathematics refer to reality, they are not certain, and as far as they are certain, they do not refer to reality."

Albert Einstein (1879-1955)

8 SEM for the BSSN puncture formulation

This chapter applies the spectral element method (SEM) to the BSSN system with the puncture method. Section 8.1 introduces the strong formulation of the BSSN system. We will then introduce the weak form of the BSSN system in section 8.2. The spectral element discretization will be explained in section 8.3. Finally, we will present the global assembled system in 8.4 and 8.5.

8.1 Strong form of the BSSN system

The standard BSSN variables are ϕ or χ, $\tilde{g}_{ij}$, $\tilde{A}_{ij}$, K, and $\tilde{\Gamma}^i$. We evolve the following system.

$$\partial_0 \phi = -\frac{1}{6}\alpha K, \tag{8.1}$$

$$\partial_0 \chi = \frac{2}{3}\chi\alpha K \tag{8.2}$$

$$\partial_0 \tilde{g}_{ij} = -2\alpha\tilde{A}_{ij}, \tag{8.3}$$

$$\begin{aligned}\partial_0 \tilde{A}_{ij} &= e^{-4\phi}[-D_iD_j\alpha + \alpha R_{ij}]^{TF} \\ &\quad +\alpha(K\tilde{A}_{ij} - 2\tilde{A}_{ik}\tilde{A}^k{}_j),\end{aligned} \tag{8.4}$$

$$\partial_0 K = -D^iD_i\alpha + \alpha(\tilde{A}_{ij}\tilde{A}^{ij} + \frac{1}{3}K^2), \tag{8.5}$$

$$\begin{aligned}\partial_t \tilde{\Gamma}^i &= \tilde{g}^{jk}\partial_j\partial_k\beta^i + \frac{1}{3}\tilde{g}^{ij}\partial_j\partial_k\beta^k + \beta^j\partial_j\tilde{\Gamma}^i \\ &\quad -\tilde{\Gamma}^j\partial_j\beta^i + \frac{2}{3}\tilde{\Gamma}^i\partial_j\beta^j - 2\tilde{A}^{ij}\partial_j\alpha \\ &\quad +2\alpha\left(\tilde{\Gamma}^i{}_{jk}\tilde{A}^{jk} + 6\tilde{A}^{ij}\partial_j\phi - \frac{2}{3}\tilde{g}^{ij}\partial_j K\right),\end{aligned} \tag{8.6}$$

$$(\partial_t - \beta^i\partial_i)\alpha = -2\alpha K \tag{8.7}$$

$$\partial_0\beta^i = \frac{3}{4}B^i, \tag{8.8}$$

$$\partial_0 B^i = \partial_t\tilde{\Gamma}^i - \eta B^i. \tag{8.9}$$

where $\partial_0 = \partial_t - \mathcal{L}_\beta$, $\tilde{D}_i$ is the covariant derivative with respect to the ***conformal metric*** $\tilde{g}_{ij}$, D_i is the covariant derivative with respect to the ***background metric*** g_{ij}, and "TF" denotes the trace-free part of the expression with respect to the *background* metric, $X^{TF}_{ij} = X_{ij} - \frac{1}{3}g_{ij}X^k_k$.

In particular, we have chosen a particular form of the system with "1+log" slicing and "gamma-driver" shift.

The Ricci tensor R_{ij} is given by

$$R_{ij} = \tilde{R}_{ij} + R^{\phi}_{ij} \tag{8.10}$$

$$\tilde{R}_{ij} = -\frac{1}{2}\tilde{g}^{lm}\partial_l\partial_m\tilde{g}_{ij} + \tilde{g}_{k(i}\partial_{j)}\tilde{\Gamma}^k + \tilde{\Gamma}^k\tilde{\Gamma}_{(ij)k} + \tilde{g}^{lm}\left(2\tilde{\Gamma}^k_{l(i}\tilde{\Gamma}_{j)km} + \tilde{\Gamma}^k_{im}\tilde{\Gamma}_{klj}\right), \tag{8.11}$$

$$R^{\phi}_{ij} = -2\tilde{D}_i\tilde{D}_j\phi - 2\tilde{g}_{ij}\tilde{D}^k\tilde{D}_k\phi + 4\tilde{D}_i\phi\tilde{D}_j\phi - 4\tilde{g}_{ij}\tilde{D}^k\phi\tilde{D}_k\phi. \tag{8.12}$$

The Lie derivatives of the tensor densities ϕ, $\tilde{g}_{ij}$ and $\tilde{A}_{ij}$ (with weights $1/6$, $-2/3$ and $-2/3$) are

$$\begin{aligned}
\mathcal{L}_\beta\phi &= \beta^k\partial_k\phi + \frac{1}{6}\partial_k\beta^k, \\
\mathcal{L}_\beta\chi &= \beta^k\partial_k\chi - \frac{2}{3}\partial_k\beta^k\chi, \\
\mathcal{L}_\beta K &= \beta^k\partial_k K, \\
\mathcal{L}_\beta\tilde{g}_{ij} &= \beta^k\partial_k\tilde{g}_{ij} + \tilde{g}_{ik}\partial_j\beta^k + \tilde{g}_{jk}\partial_i\beta^k - \frac{2}{3}\tilde{g}_{ij}\partial_k\beta^k, \\
\mathcal{L}_\beta\tilde{A}_{ij} &= \beta^k\partial_k\tilde{A}_{ij} + \tilde{A}_{ik}\partial_j\beta^k + \tilde{A}_{jk}\partial_i\beta^k - \frac{2}{3}\tilde{A}_{ij}\partial_k\beta^k.
\end{aligned}$$

The covariant derivatives of the lapse are with respect with the physical metric and are defined by

$$D_iD_j\alpha = \partial_i\partial_j\alpha - 4\partial_{(i}\phi\,\partial_{j)}\alpha - \tilde{\Gamma}^k_{ij}\partial_k\alpha + 2g_{ij}g^{kl}\partial_k\phi\,\partial_l\alpha, \tag{8.13}$$

Furthermore, the trace is given by

$$D^iD_i\alpha = \exp(-4\phi)\,\tilde{g}^{il}\tilde{D}_l\tilde{D}_i\alpha; \tag{8.14}$$

$$= \exp(-4\phi)\,\left(\tilde{g}^{ij}\partial_i\partial_j\alpha - \tilde{\Gamma}^k\partial_k\alpha + 2\tilde{g}^{ij}\partial_i\phi\partial_j\alpha\right). \tag{8.15}$$

The covariant derivative of ϕ is with respect to the background metric and is defined by

$$D_iD_j\phi = \partial_i\partial_j\phi - \tilde{\Gamma}^k_{ij}\partial_k\phi. \tag{8.16}$$

8.2 Weak form of the BSSN system

The weak form needs to be applied to the evolution system, equations (8.1)-(8.6), as well as the equations coming from the gauge choice, equations (8.7) and (8.8). Alternatively if the χ-method is chosen, equation (8.2) replaces equation (8.1).

It is important to remember that there might not be a unique weak formulation of the evolution system, depending on one's choice of which integrals are integrated by parts in order for the boundary conditions to appear in the system. In this section, we will present two different versions of weak formulation.

8.2.1 General integration by parts formulae in 3D

Let $\beta = (\beta^x, \beta^y, \beta^z)$ be a vector, ϕ be a scalar function of the coordinates (x, y, z), and w a scalar test function, we have the following equality:

$$\int_{\mathbf{\Omega}} \beta \cdot (\nabla \phi)\ w\ d\mathbf{\Omega} = \underbrace{\int_{\mathbf{\Omega}} \nabla \cdot (\beta\phi)\ w\ d\mathbf{\Omega}}_{A} - \underbrace{\int_{\mathbf{\Omega}} (\nabla \cdot \beta)\, \phi\ w\ d\mathbf{\Omega}}_{B}, \tag{8.17}$$

In terms of sums using the Einstein convention, this equality is equivalent to:

$$\int_{\mathbf{\Omega}} \beta^k \partial_k \phi\ w\ d\mathbf{\Omega} \ = \ \int_{\mathbf{\Omega}} \partial_k \left(\beta^k \phi\right)\ w\ d\mathbf{\Omega} - \int_{\mathbf{\Omega}} \left(\partial_k \beta^k\right) \phi\ w\ d\mathbf{\Omega}. \tag{8.18}$$

We have the following formulae of integration by parts ,

$$A \ = \ \int_{\mathbf{\Omega}} \nabla \cdot (\beta\phi)\ w\ d\mathbf{\Omega} \ = \ \int_{\mathbf{\Gamma}} [(\beta\phi)\, w] \cdot \mathbf{n}\ d\mathbf{\Gamma} - \int_{\mathbf{\Omega}} (\beta\phi) \cdot \nabla w\ d\mathbf{\Omega}\ ; \tag{8.19}$$

$$B \ = \ \int_{\mathbf{\Omega}} (\nabla \cdot \beta)\, \phi\ w\ d\mathbf{\Omega} \ = \ \int_{\mathbf{\Gamma}} (\mathbf{n} \cdot \beta)\, \phi w\ d\mathbf{\Gamma} - \int_{\mathbf{\Omega}} [\beta \cdot \nabla (\phi w)]\ d\mathbf{\Omega}\ , \tag{8.20}$$

where $\mathbf{n} = (\mathbf{n}^x, \mathbf{n}^y, \mathbf{n}^z)$ is the outward normal unit vector.

If one integrates by parts ***both*** A ***and*** B then the term that introduces boundary conditions disappears and the point of the exercise of integrating by parts is completely missed. So we have two choices there:

- integrate A by parts and leave B as it is:

$$\begin{aligned} \int_{\mathbf{\Omega}} \beta \cdot (\nabla \phi)\ w\ d\mathbf{\Omega} \ &= \ \int_{\mathbf{\Gamma}} [(\beta\phi)\, w] \cdot \mathbf{n}\ d\mathbf{\Gamma} - \int_{\mathbf{\Omega}} (\beta\phi) \cdot \nabla w\ d\mathbf{\Omega} \\ &\quad - \int_{\mathbf{\Omega}} (\nabla \cdot \beta)\, \phi\ w\ d\mathbf{\Omega}; \end{aligned} \tag{8.21}$$

- integrate B by parts and leave A as it is:

$$\begin{aligned} \int_{\mathbf{\Omega}} \beta \cdot (\nabla \phi)\ w\ d\mathbf{\Omega} \ &= \ \int_{\mathbf{\Omega}} \nabla \cdot (\beta\phi)\ w\ d\mathbf{\Omega} - \int_{\mathbf{\Gamma}} (\mathbf{n} \cdot \beta)\, \phi w\ d\mathbf{\Gamma} \\ &\quad + \int_{\mathbf{\Omega}} [\beta \cdot \nabla (\phi w)]\ d\mathbf{\Omega}. \end{aligned} \tag{8.22}$$

It turns out that the first option involves only $\nabla w, \nabla \beta$ and no $\nabla \phi$ and gives a nicer and shorter formula than the second one and therefore we make this choice for now.

There is an alternative choice of integration by parts that involves integrating second order space derivatives. First, remember the previous integration by parts formula

$$\int_{\mathbf{\Omega}} \nabla \cdot F\ w\ d\mathbf{\Omega} \ = \ \int_{\mathbf{\Gamma}} F \cdot \mathbf{n}\ w d\mathbf{\Gamma} - \int_{\mathbf{\Omega}} F \cdot \nabla w\ d\mathbf{\Omega}, \tag{8.23}$$

where $F = (f_1, f_2, f_3)$ holds true for all F and in particular for $F = (f, 0, 0)$ (or equivalently for the y and z components). Hence we have the following dimension specific integration by parts formula (with no summation):

$$\int_{\mathbf{\Omega}} \partial_i f \; w \; d\mathbf{\Omega} \;=\; \int_{\mathbf{\Gamma}} f \mathbf{n}^i \; w d\mathbf{\Gamma} - \int_{\mathbf{\Omega}} f \partial_i w \; d\mathbf{\Omega}. \tag{8.24}$$

Therefore, when integrating by parts we obtain

$$\begin{aligned} \int_{\mathbf{\Omega}} f \partial_i \partial_j u \; w \; d\mathbf{\Omega} &= \int_{\mathbf{\Gamma}} f \; \partial_j u \; \mathbf{n}^i \; w d\mathbf{\Gamma} - \int_{\mathbf{\Omega}} \partial_j u \; \partial_i (f w) \; d\mathbf{\Omega} \\ &= \int_{\mathbf{\Gamma}} f \; \partial_j u \; \mathbf{n}^i \; w d\mathbf{\Gamma} - \int_{\mathbf{\Omega}} \partial_j u \; \partial_i f \; w \; d\mathbf{\Omega} \\ &\quad - \int_{\mathbf{\Omega}} f \; \partial_j u \; \partial_i w \; d\mathbf{\Omega}. \end{aligned} \tag{8.25}$$

8.2.2 Weak form, version 1

A choice of weak form of the evolution system is as follows:

$$\begin{aligned} \int_{\mathbf{\Omega}} \partial_t \phi \; w \; d\mathbf{\Omega} &= \int_{\mathbf{\Gamma}} \beta^k \mathbf{n}^k \phi w \; d\mathbf{\Gamma} - \int_{\mathbf{\Omega}} \beta^k \phi \partial_k w \; d\mathbf{\Omega} - \int_{\mathbf{\Omega}} \partial_k \beta^k \phi w \; d\mathbf{\Omega} \\ &\quad + \frac{1}{6} \int_{\mathbf{\Omega}} \partial_k \beta^k w \; d\mathbf{\Omega} - \frac{1}{6} \int_{\mathbf{\Omega}} \alpha K w \; d\mathbf{\Omega}; \end{aligned} \tag{8.26}$$

or equivalently for the χ-method,

$$\begin{aligned} \int_{\mathbf{\Omega}} \partial_t \chi \; w \; d\mathbf{\Omega} &= \int_{\mathbf{\Gamma}} \beta^k \mathbf{n}^k \chi w \; d\mathbf{\Gamma} - \int_{\mathbf{\Omega}} \beta^k \chi \partial_k w \; d\mathbf{\Omega} - \frac{5}{3} \int_{\mathbf{\Omega}} \partial_k \beta^k \chi w \; d\mathbf{\Omega} \\ &\quad + \frac{2}{3} \int_{\mathbf{\Omega}} \chi \alpha K w \; d\mathbf{\Omega}; \end{aligned} \tag{8.27}$$

$$\begin{aligned} \int_{\mathbf{\Omega}} \partial_t \tilde{g}_{ij} \; w \; d\mathbf{\Omega} &= \int_{\mathbf{\Gamma}} \beta^k \mathbf{n}^k \tilde{g}_{ij} w \; d\mathbf{\Gamma} - \int_{\mathbf{\Omega}} \beta^k \tilde{g}_{ij} \partial_k w \; d\mathbf{\Omega} - \frac{5}{3} \int_{\mathbf{\Omega}} \partial_k \beta^k \tilde{g}_{ij} w \; d\mathbf{\Omega} \\ &\quad + \int_{\mathbf{\Omega}} \left(\tilde{g}_{ik} \partial_j \beta^k + \tilde{g}_{jk} \partial_i \beta^k \right) w \; d\mathbf{\Omega} - 2 \int_{\mathbf{\Omega}} \alpha \tilde{A}_{ij} w \; d\mathbf{\Omega}; \end{aligned} \tag{8.28}$$

$$\begin{aligned} \int_{\mathbf{\Omega}} \partial_t \tilde{A}_{ij} \; w \; d\mathbf{\Omega} &= \int_{\mathbf{\Gamma}} \beta^k \mathbf{n}^k \tilde{A}_{ij} w \; d\mathbf{\Gamma} - \int_{\mathbf{\Omega}} \beta^k \tilde{A}_{ij} \partial_k w \; d\mathbf{\Omega} - \frac{5}{3} \int_{\mathbf{\Omega}} \partial_k \beta^k \tilde{A}_{ij} w \; d\mathbf{\Omega} \\ &\quad + \int_{\mathbf{\Omega}} \left(\tilde{A}_{ik} \partial_j \beta^k + \tilde{A}_{jk} \partial_i \beta^k \right) w \; d\mathbf{\Omega} \\ &\quad + \int_{\mathbf{\Omega}} e^{-4\phi} \left[-D_i D_j \alpha + \alpha R_{ij} \right]^{TF} w \; d\mathbf{\Omega} \\ &\quad + \int_{\mathbf{\Omega}} \alpha (K \tilde{A}_{ij} - 2 \tilde{A}_{ik} \tilde{A^k}_j) \; w \; d\mathbf{\Omega}; \end{aligned} \tag{8.29}$$

$$\int_{\Omega} \partial_t K\ w\ d\mathbf{\Omega} = \int_{\Gamma} \beta^k \mathbf{n}^k K w\ d\mathbf{\Gamma} - \int_{\Omega} \beta^k K \partial_k w\ d\mathbf{\Omega} - \int_{\Omega} \partial_k \beta^k K w\ d\mathbf{\Omega} - \int_{\Omega} D^i D_i \alpha w\ d\mathbf{\Omega} + \int_{\Omega} \alpha \left(\tilde{A}_{ij}\tilde{A}^{ij} + \frac{1}{3}K^2 \right) w\ d\mathbf{\Omega}; \quad (8.30)$$

$$\int_{\Omega} \partial_t \tilde{\Gamma}^i\ w\ d\mathbf{\Omega} = \int_{\Omega} \tilde{g}^{jk}\partial_j\partial_k\beta^i w\ d\mathbf{\Omega} + \frac{1}{3}\int_{\Omega} \tilde{g}^{ij}\partial_j\partial_k\beta^k w\ d\mathbf{\Omega} + \int_{\Omega} \beta^j \partial_j \tilde{\Gamma}^i w\ d\mathbf{\Omega} - \int_{\Omega} \tilde{\Gamma}^j \partial_j \beta^i w\ d\mathbf{\Omega} + \frac{2}{3}\int_{\Omega} \tilde{\Gamma}^i \partial_j \beta^j w\ d\mathbf{\Omega} - 2\int_{\Omega} \tilde{A}^{ij}\partial_j \alpha w\ d\mathbf{\Omega} + 2\int_{\Omega} \alpha \left(\tilde{\Gamma}^i_{jk}\tilde{A}^{jk} + 6\tilde{A}^{ij}\partial_j\phi - \frac{2}{3}\tilde{g}^{ij}\partial_j K \right) w\ d\mathbf{\Omega}. \quad (8.31)$$

For the shift and the lapse we have the following weak form,

$$\int_{\Omega} \partial_t \alpha\ w\ d\mathbf{\Omega} = \int_{\Gamma} \beta^k \mathbf{n}^k \alpha w\ d\mathbf{\Gamma} - \int_{\Omega} \beta^k \alpha \partial_k w\ d\mathbf{\Omega} - \int_{\Omega} \partial_k \beta^k \alpha w\ d\mathbf{\Omega} - 2\int_{\Omega} \alpha K w\ d\mathbf{\Omega}; \quad (8.32)$$

$$\int_{\Omega} \partial_t \beta^i\ w\ d\mathbf{\Omega} = \int_{\Omega} \beta^i \partial_i \beta^i w\ d\mathbf{\Omega} + \frac{3}{4}\int_{\Omega} B^i w\ d\mathbf{\Omega} \quad (8.33)$$

$$\int_{\Omega} \partial_t B^i\ w\ d\mathbf{\Omega} = \int_{\Omega} \beta^i \partial_i B^i w\ d\mathbf{\Omega} + \int_{\Omega} \partial_t \tilde{\Gamma}^i w\ d\mathbf{\Omega} - \int_{\Omega} \eta B^i w\ d\mathbf{\Omega}. \quad (8.34)$$

A few comments:

- For the variables $\phi, \chi, \tilde{g}_{ij}, \tilde{A}_{ij}, K$ and α the boundary conditions appear in the same way, due to the integration by parts discussed in equation (8.21), from the integral

$$\int_{\Omega} \beta^k \partial_k \phi\ w\ d\mathbf{\Omega} = \int_{\Gamma} \beta^k \mathbf{n}^k \phi\ w\ d\mathbf{\Gamma} - \int_{\Omega} \phi \beta^k \partial_k w\ d\mathbf{\Omega} - \int_{\Omega} \phi \partial_k \beta^k\ w\ d\mathbf{\Omega}. \quad (8.35)$$

- For the moment there is no integral on the boundary for the variable $\tilde{\Gamma}^i$ because there is no obvious integral term appropriate for integration by parts. Since $\tilde{\Gamma}^i = -\partial_j \tilde{g}^{ij}$, then if there is a boundary condition introduced for $\tilde{g}_{ij}$ it should be implicitly introduced for the variable $\tilde{\Gamma}^i$ as well.

- For a similar reason, there is no boundary condition on β in the system (8.33)-(8.34). However, the variable β appears in the boundary integral in all the other evolution equations (8.26)-(8.30), so the boundary condition on β could be introduced in this manner.

8.2.3 Weak form, version 2

A second choice of weak form of the evolution system can be obtained by not integrating by parts the same type of integrals as in version 1, but to integrate second order space derivatives instead, only appearing with the terms denoted $(\text{Terms})^{\text{IP}}$ below. We use the test function w.

$$\int_{\mathbf{\Omega}} \partial_t \phi \; w \; d\mathbf{\Omega} \;=\; \int_{\mathbf{\Omega}} \mathcal{L}_\beta \phi \; w \; d\mathbf{\Omega} - \frac{1}{6}\int_{\mathbf{\Omega}} \alpha K w \; d\mathbf{\Omega}; \tag{8.36}$$

or equivalently for the χ-method,

$$\int_{\mathbf{\Omega}} \partial_t \chi \; w \; d\mathbf{\Omega} \;=\; \int_{\mathbf{\Omega}} \mathcal{L}_\beta \chi \; w \; d\mathbf{\Omega} + \frac{2}{3}\int_{\mathbf{\Omega}} \chi \alpha K w \; d\mathbf{\Omega}; \tag{8.37}$$

$$\int_{\mathbf{\Omega}} \partial_t \tilde{g}_{ij} \; w \; d\mathbf{\Omega} \;=\; \int_{\mathbf{\Omega}} \mathcal{L}_\beta \tilde{g}_{ij} \; w \; d\mathbf{\Omega} - 2\int_{\mathbf{\Omega}} \alpha \tilde{A}_{ij} \; w \; d\mathbf{\Omega}; \tag{8.38}$$

$$\begin{aligned}\int_{\mathbf{\Omega}} \partial_t \tilde{A}_{ij} \; w \; d\mathbf{\Omega} \;=\; & \int_{\mathbf{\Omega}} \mathcal{L}_\beta \tilde{A}_{ij} \; w \; d\mathbf{\Omega} + \int_{\mathbf{\Omega}} \alpha (K\tilde{A}_{ij} - 2\tilde{A}_{ik}\tilde{A}^k{}_j) \; w \; d\mathbf{\Omega} \\ & + \Big(\int_{\mathbf{\Omega}} e^{-4\phi} \left[-D_i D_j \alpha + \alpha R_{ij} \right]^{TF} w \; d\mathbf{\Omega} \Big)^{\text{IP}};\end{aligned} \tag{8.39}$$

$$\begin{aligned}\int_{\mathbf{\Omega}} \partial_t K \; w \; d\mathbf{\Omega} \;=\; & \int_{\mathbf{\Omega}} \mathcal{L}_\beta K \; w \; d\mathbf{\Omega} - \Big(\int_{\mathbf{\Omega}} D^i D_i \alpha w \; d\mathbf{\Omega} \Big)^{\text{IP}} \\ & + \int_{\mathbf{\Omega}} \alpha \left(\tilde{A}_{ij}\tilde{A}^{ij} + \frac{1}{3}K^2 \right) w \; d\mathbf{\Omega};\end{aligned} \tag{8.40}$$

$$\begin{aligned}\int_{\mathbf{\Omega}} \partial_t \tilde{\Gamma}^i \; w \; d\mathbf{\Omega} \;=\; & \Big(\int_{\mathbf{\Omega}} \tilde{g}^{jk} \partial_j \partial_k \beta^i w \; d\mathbf{\Omega} \Big)^{\text{IP}} + \Big(\frac{1}{3} \int_{\mathbf{\Omega}} \tilde{g}^{ij} \partial_j \partial_k \beta^k w \; d\mathbf{\Omega} \Big)^{\text{IP}} \\ & + \int_{\mathbf{\Omega}} \beta^j \partial_j \tilde{\Gamma}^i w \; d\mathbf{\Omega} - \int_{\mathbf{\Omega}} \tilde{\Gamma}^j \partial_j \beta^i w \; d\mathbf{\Omega} \\ & + \frac{2}{3}\int_{\mathbf{\Omega}} \tilde{\Gamma}^i \partial_j \beta^j w \; d\mathbf{\Omega} - 2\int_{\mathbf{\Omega}} \tilde{A}^{ij} \partial_j \alpha w \; d\mathbf{\Omega} \\ & + 2\int_{\mathbf{\Omega}} \alpha \left(\tilde{\Gamma}^i_{jk}\tilde{A}^{jk} + 6\tilde{A}^{ij}\partial_j \phi - \frac{2}{3}\tilde{g}^{ij}\partial_j K \right) w \; d\mathbf{\Omega}.\end{aligned} \tag{8.41}$$

For the shift and the lapse we have the following weak form,

$$\int_{\mathbf{\Omega}} \partial_t \alpha \; w \; d\mathbf{\Omega} \;=\; \int_{\mathbf{\Omega}} \mathcal{L}_\beta \alpha \; w \; d\mathbf{\Omega} - 2\int_{\mathbf{\Omega}} \alpha K w \; d\mathbf{\Omega}; \tag{8.42}$$

$$\int_{\mathbf{\Omega}} \partial_t \beta^i \; w \; d\mathbf{\Omega} \;=\; \int_{\mathbf{\Omega}} \mathcal{L}_\beta \beta^i \; w \; d\mathbf{\Omega} + \frac{3}{4}\int_{\mathbf{\Omega}} B^i w \; d\mathbf{\Omega}; \tag{8.43}$$

$$\begin{aligned}\int_{\mathbf{\Omega}} \partial_t B^i \; w \; d\mathbf{\Omega} \;=\; & \int_{\mathbf{\Omega}} \mathcal{L}_\beta B^i \; w \; d\mathbf{\Omega} + \int_{\mathbf{\Omega}} \partial_t \tilde{\Gamma}^i w \; d\mathbf{\Omega} \\ & - \int_{\mathbf{\Omega}} \eta B^i w \; d\mathbf{\Omega};\end{aligned} \tag{8.44}$$

The weak form of all the Lie derivatives are explicitly given by:

$$\int_{\Omega} \mathcal{L}_{\beta}\phi \; w \; d\mathbf{\Omega} = \int_{\Omega} \beta^k \partial_k \phi \; w \; d\mathbf{\Omega} + \frac{1}{6}\int_{\Omega} \partial_k \beta^k \; w \; d\mathbf{\Omega}; \tag{8.45}$$

$$\int_{\Omega} \mathcal{L}_{\beta}\chi \; w \; d\mathbf{\Omega} = \int_{\Omega} \beta^k \partial_k \chi w \; d\mathbf{\Omega} - \frac{2}{3}\int_{\Omega} \chi \partial_k \beta^k \; w \; d\mathbf{\Omega}; \tag{8.46}$$

$$\begin{aligned}\int_{\Omega} \mathcal{L}_{\beta}\tilde{g}_{ij} \; w \; d\mathbf{\Omega} &= \int_{\Omega} \beta^k \partial_k \tilde{g}_{ij} \; w \; d\mathbf{\Omega} - \frac{2}{3}\int_{\Omega} \tilde{g}_{ij} \partial_k \beta^k \; w \; d\mathbf{\Omega}; \\ &+ \int_{\Omega} \left(\tilde{g}_{ik}\partial_j \beta^k + \tilde{g}_{jk}\partial_i \beta^k\right) \; w \; d\mathbf{\Omega};\end{aligned} \tag{8.47}$$

$$\begin{aligned}\int_{\Omega} \mathcal{L}_{\beta}\tilde{A}_{ij} \; w \; d\mathbf{\Omega} &= \int_{\Omega} \beta^k \partial_k \tilde{A}_{ij} \; w \; d\mathbf{\Omega} - \frac{2}{3}\int_{\Omega} \tilde{A}_{ij} \partial_k \beta^k \; w \; d\mathbf{\Omega}; \\ &+ \int_{\Omega} \left(\tilde{A}_{ik}\partial_j \beta^k + \tilde{A}_{jk}\partial_i \beta^k\right) w \; d\mathbf{\Omega}\end{aligned} \tag{8.48}$$

$$\int_{\Omega} \mathcal{L}_{\beta}K \; w \; d\mathbf{\Omega} = \int_{\Omega} \beta^k \partial_k K \; w \; d\mathbf{\Omega}; \tag{8.49}$$

$$\int_{\Omega} \mathcal{L}_{\beta}\alpha \; w \; d\mathbf{\Omega} = \int_{\Omega} \beta^k \partial_k \alpha w \; d\mathbf{\Omega}; \tag{8.50}$$

$$\int_{\Omega} \mathcal{L}_{\beta}\beta^i \; w \; d\mathbf{\Omega} = \int_{\Omega} \beta^k \partial_k \beta^i \; w \; d\mathbf{\Omega}; \tag{8.51}$$

$$\int_{\Omega} \mathcal{L}_{\beta}B^i \; w \; d\mathbf{\Omega} = \int_{\Omega} \beta^k \partial_k B^i \; w \; d\mathbf{\Omega}. \tag{8.52}$$

Comments: The integration by parts here introduce boundary conditions for the variables α, β, $\tilde{g}_{ij}$, ϕ. Note that this way, there are no boundary conditions for χ, K, $\tilde{A}_{ij}$ and $\tilde{\Gamma}^i$.

8.2.4 Weak form, version 3

Another alternative weak form could be a combination of the two previous weak forms, versions 1 and 2. Moreover, a combination would allow for the introduction of boundary conditions for all the variables.

8.2.5 Abstract weak form of the BSSN

We can write the weak forms of BSSN in terms of bilinear forms. Terms that are not integrated by parts can be described by the following bilinear form:

$$a(u, w) = \int_{\Omega} u \; w d\mathbf{\Omega}. \tag{8.53}$$

The terms that are integrated by parts can be described by the following bilinear form

$$b(u, w) = \int u \; \nabla w d\mathbf{\Omega}. \tag{8.54}$$

8.3 Discretization of the weak form of the BSSN system

For each element, the weak form presented in the previous section 8.2 holds and can be discretized and written in matrix form before the assembly process.

8.3.1 Elemental matrix form of the BSSN system

In order to keep the number of subscripts and superscripts to a minimum, we will now not use explicitly the letter $\mathbf{k}$ ***corresponding to the*** $\mathbf{k}$***-th element for the elemental matrices.*** That is, for example, the elemental mass matrix $\mathbf{M^k}$ will be referred to as $\mathbf{M}$ in the following section.

Elemental matrix form of the BSSN system version 1

The elemental matrix form of the BSSN system of version 1 is given by,

$$\begin{aligned}\mathbf{M}\otimes\dot{\phi} &= \mathbf{B}\otimes\left(\beta^k:\phi\right)-\mathbf{D}_k\otimes\left(\beta^k:\phi\right)-\phi:\left(\mathbf{A}_k\otimes\beta^k\right)\\ &\quad+\frac{1}{6}\mathbf{A}_k\otimes\beta^k-\frac{1}{6}\mathbf{M}\otimes(\alpha:K)\,; \end{aligned}\tag{8.55}$$

alternatively, for the χ-method,

$$\begin{aligned}\mathbf{M}\otimes\dot{\chi} &= \mathbf{B}\otimes\left(\beta^k:\chi\right)-\mathbf{D}_k\otimes\left(\beta^k:\chi\right)-\frac{5}{3}\chi:\left(\mathbf{A}_k\otimes\beta^k\right)\\ &\quad+\frac{2}{3}\mathbf{M}\otimes(\chi:\alpha:K)\,; \end{aligned}\tag{8.56}$$

$$\begin{aligned}\mathbf{M}\otimes\dot{\tilde{g}}_{ij} &= \mathbf{B}\otimes\left(\beta^k:\tilde{g}_{ij}\right)-\mathbf{D}_k\otimes\left(\beta^k:\tilde{g}_{ij}\right)-\frac{5}{3}\tilde{g}_{ij}:\left(\mathbf{A}_k\otimes\beta^k\right)\\ &\quad+\tilde{g}_{ik}:\left(\mathbf{A}_j\otimes\beta^k\right)+\tilde{g}_{jk}:\left(\mathbf{A}_i\otimes\beta^k\right)\\ &\quad-2\mathbf{M}\otimes\left(\alpha:\tilde{A}_{ij}\right); \end{aligned}\tag{8.57}$$

$$\begin{aligned}\mathbf{M}\otimes\dot{\tilde{A}}_{ij} &= \mathbf{B}\otimes\left(\beta^k:\tilde{A}_{ij}\right)-\mathbf{D}_k\otimes\left(\beta^k:\tilde{A}_{ij}\right)-\frac{5}{3}\tilde{A}_{ij}:\left(\mathbf{A}_k\otimes\beta^k\right)\\ &\quad+\tilde{A}_{ik}:\left(\mathbf{A}_j\otimes\beta^k\right)+\tilde{A}_{jk}:\left(\mathbf{A}_i\otimes\beta^k\right)+\mathbb{X}^{TF}(e^{-4\phi})\\ &\quad+\mathbf{M}\otimes\left(\alpha:K:\tilde{A}_{ij}-2\alpha:\tilde{A}_{ik}:\tilde{A}^k_j\right); \end{aligned}\tag{8.58}$$

$$\begin{aligned}\mathbf{M}\otimes\dot{K} &= \mathbf{B}\otimes\left(\beta^k:K\right)-\mathbf{D}_k\otimes\left(\beta^k:K\right)-K:\left(\mathbf{A}_k\otimes\beta^k\right)\\ &\quad+\mathbf{M}\otimes\left[\alpha:\left(\tilde{A}_{ij}:\tilde{A}^{ij}+\frac{1}{3}K:K\right)\right]\\ &\quad-\left(\mathbb{D}_{li}(\tilde{g}_{il})\otimes\alpha\right); \end{aligned}\tag{8.59}$$

$$\begin{aligned}\mathbf{M}\otimes\dot{\tilde{\Gamma}}^i &= -\tilde{g}^{jk}:\left(\mathbf{A}_{jk}\otimes\beta^i\right)+\frac{1}{3}\tilde{g}^{ij}:\left(\mathbf{A}_{jk}\otimes\beta^k\right)-\beta^j:\left(\mathbf{A}_j\otimes\tilde{\Gamma}^i\right)\\ &\quad+2\alpha:\left[\mathbf{\Lambda}^i{}_{jk}:\tilde{A}^{jk}+6\tilde{A}^{ij}\left(\mathbf{A}_j\otimes\phi\right)-\frac{2}{3}\tilde{g}^{ij}:\left(\mathbf{A}_j\otimes K\right)\right]\\ &\quad-\tilde{\Gamma}^j:\left(\mathbf{A}_j\otimes\beta^i\right)+\frac{2}{3}\tilde{\Gamma}_i:\left(\mathbf{A}_j\otimes\beta^j\right)\\ &\quad-2\tilde{A}^{ij}:\left(\mathbf{A}_j\otimes\alpha\right). \end{aligned}\tag{8.60}$$

And finally for the lapse and the shift:

$$\begin{aligned}\mathbf{M}\otimes\dot{\alpha} &= \mathbf{B}\otimes\left(\beta^k:\alpha\right)-\mathbf{D}_k\otimes\left(\beta^k:\alpha\right)-\alpha:\left(\mathbf{A}_k\otimes\beta^k\right)\\ &\quad -2\mathbf{M}\otimes(\alpha:K)\,;\end{aligned} \tag{8.61}$$

$$\mathbf{M}\otimes\dot{\beta}^i = \beta^k:\left(\mathbf{A}_k\otimes\beta^i\right)-\frac{3}{4}\mathbf{M}\otimes B^i; \tag{8.62}$$

$$\mathbf{M}\otimes\dot{B}^i = \beta^k:\left(\mathbf{A}_k\otimes B^i\right)+\mathbf{M}\otimes\dot{\tilde{\Gamma}}^i-\eta\mathbf{M}\otimes B^i. \tag{8.63}$$

The dot notation $\dot{\beta}$ refers to $\partial_t\beta$ the time derivative of β. The symbol : refers to the term by term matrix multiplication (Hadamard product), that is $(A:B)_{ij}=a_{ij}*b_{ij}$. On the other hand, the symbol $\otimes$ is a multiplication operator that will be defined below for each elemental matrix type.

Elemental matrix form of the BSSN system version 2

The elemental matrix form of the BSSN system of version 2 is given by,

$$\mathbf{M}\otimes\dot{\phi} = \mathbf{L}_\beta\otimes\phi-\frac{1}{6}\mathbf{M}\otimes(\alpha:K)\,; \tag{8.64}$$

alternatively, for the χ-method,

$$\mathbf{M}\otimes\dot{\chi} = \mathbf{L}_\beta\otimes\chi-\frac{2}{3}\beta^k:(\mathbf{A}_k\otimes\chi)+\frac{2}{3}\mathbf{M}\otimes(\chi:\alpha:K)\,; \tag{8.65}$$

$$\mathbf{M}\otimes\dot{\tilde{g}}_{ij} = \mathbf{L}_\beta\otimes\tilde{g}_{ij}-2\mathbf{M}\otimes\left(\alpha:\tilde{A}_{ij}\right); \tag{8.66}$$

$$\begin{aligned}\mathbf{M}\otimes\dot{\tilde{A}}_{ij} &= \mathbf{L}_\beta\otimes\tilde{A}_{ij}+\mathbb{X}^{TF}(e^{-4\phi})\\ &\quad +\mathbf{M}\otimes\left(\alpha:K:\tilde{A}_{ij}-2\alpha:\tilde{A}_{ik}:\tilde{A}^k_j\right);\end{aligned} \tag{8.67}$$

$$\begin{aligned}\mathbf{M}\otimes\dot{K} &= \mathbf{L}_\beta\otimes K-(\mathbb{D}_{li}(\tilde{g}_{il})\otimes\alpha)\\ &\quad +\mathbf{M}\otimes\left[\alpha:\left(\tilde{A}_{ij}:\tilde{A}^{ij}+\frac{1}{3}K:K\right)\right];\end{aligned} \tag{8.68}$$

$$\begin{aligned}\mathbf{M}\otimes\dot{\tilde{\Gamma}}^i &= \tilde{g}^{jk}:\left(\mathbf{A}_{jk}\otimes\beta^i\right)+\frac{1}{3}\tilde{g}^{ij}:\left(\mathbf{A}_{jk}\otimes\beta^k\right)-\beta^j:\left(\mathbf{A}_j\otimes\tilde{\Gamma}^i\right)\\ &\quad +2\alpha:\left[\mathbf{\Lambda}^i{}_{jk}:\tilde{A}^{jk}+6\tilde{A}^{ij}\left(\mathbf{A}_j\otimes\phi\right)-\frac{2}{3}\tilde{g}^{ij}:\left(\mathbf{A}_j\otimes K\right)\right]\\ &\quad -\tilde{\Gamma}^j:\left(\mathbf{A}_j\otimes\beta^i\right)+\frac{2}{3}\tilde{\Gamma}_i:\left(\mathbf{A}_j\otimes\beta^j\right)\\ &\quad -2\tilde{A}^{ij}:\left(\mathbf{A}_j\otimes\alpha\right).\end{aligned} \tag{8.69}$$

And finally for the lapse and the shift:

$$\mathbf{M} \otimes \dot{\alpha} \;=\; \mathbf{L}_\beta \otimes \alpha - 2\mathbf{M} \otimes (\alpha : K)\,; \tag{8.70}$$

$$\mathbf{M} \otimes \dot{\beta^i} \;=\; \mathbf{L}_\beta \otimes \beta^i - \frac{3}{4}\mathbf{M} \otimes B^i; \tag{8.71}$$

$$\mathbf{M} \otimes \dot{B^i} \;=\; \mathbf{L}_\beta \otimes B^i + \mathbf{M} \otimes \dot{\tilde{\Gamma}}^i - \eta\mathbf{M} \otimes B^i. \tag{8.72}$$

8.3.2 Basic combinations of Elemental Matrices appearing in the BSSN system

The basic elemental matrices, the mass matrix $\mathbf{M}$, advection and diffusion matrices $\mathbf{A}$, $\mathbf{D}$ are fully described in chapter 7. The following matrices are combinations of those basic elemental matrices. See Appendix F for ***general*** shaped elements.

Elemental advection matrix product of advection matrix $\mathbf{A}_k$ type 1

The product of advection matrices$\mathbf{A}_k$ appears in the following type of integral

$$\int_{\mathbf{\Omega}} f\partial_i u \partial_j u \; w \; d\mathbf{\Omega} \;=\; f : (\mathbf{A}_i \otimes u) : (\mathbf{A}_j \otimes u) \; \frac{1}{|J|}. \tag{8.73}$$

Note that the factor $1/|J|$ present in the above equation results from the multiplication of 2 space derivatives within the same integral (one change of variables) and the current definition we are using for $\mathbf{A}_k$. Each time the product of 2 space derivatives appear in the same integral this factor of $1/|J|$ needs to be present accordingly to our definition of $\mathbf{A}_k$.

Elemental matrix for second derivative in space $\mathbf{A}_{ij}$ type 1

The advection matrix $\mathbf{A}_{ij}$ appears in the following type of integral

$$\int_{\mathbf{\Omega}} f\partial_i \partial_j u \; w \; d\mathbf{\Omega} \;=\; f : (\mathbf{A}_{ij} \otimes u)\,, \tag{8.74}$$

where, f, u and w are scalar functions and $i, j = x, y,$ or z. The operator $\otimes$ will act on u in a different manner depending on the value of i, j as described below:

- $i \neq j$

$$\mathbf{A}_{xy} \otimes u \;=\; \mathbf{A}_{yx} \otimes u \;=\; \rho : \left[(H \cdot_{xy} u) \cdot_{xy} H^T\right] |J| \frac{\partial \xi}{\partial x}\frac{\partial \eta}{\partial y}; \tag{8.75}$$

$$\mathbf{A}_{xz} \otimes u \;=\; \mathbf{A}_{zx} \otimes u \;=\; \rho : \left[(H \cdot_{xy} u) \cdot_{yz} H^T\right] |J| \frac{\partial \xi}{\partial x}\frac{\partial \zeta}{\partial z}; \tag{8.76}$$

$$\mathbf{A}_{yz} \otimes u \;=\; \mathbf{A}_{zy} \otimes u \;=\; \rho : \left[\left(u \cdot_{xy} H^T\right) \cdot_{yz} H^T\right] |J| \frac{\partial \eta}{\partial y}\frac{\partial \zeta}{\partial z}. \tag{8.77}$$

- $i = j$ then

$$\mathbf{A}_{xx} \otimes u = \mathbf{Q}_x \otimes u + (A_x \otimes u) \frac{\partial^2 \xi}{\partial x^2} \left(\frac{\partial x}{\partial \xi} \right)^2; \quad (8.78)$$

$$\mathbf{A}_{yy} \otimes u = \mathbf{Q}_y \otimes u + (A_y \otimes u) \frac{\partial^2 \eta}{\partial y^2} \left(\frac{\partial y}{\partial \xi} \right)^2; \quad (8.79)$$

$$\mathbf{A}_{zz} \otimes u = \mathbf{Q}_z \otimes u + (A_z \otimes u) \frac{\partial^2 \zeta}{\partial z^2} \left(\frac{\partial z}{\partial \zeta} \right)^2; \quad (8.80)$$

where

$$\mathbf{Q}_x \otimes u = \rho : (W \cdot_{xy} u) |J| \left(\frac{\partial \xi}{\partial x} \right)^2; \quad (8.81)$$

$$\mathbf{Q}_y \otimes u = \rho : \left(u \cdot_{xy} W^T \right) |J| \left(\frac{\partial \eta}{\partial y} \right)^2; \quad (8.82)$$

$$\mathbf{Q}_z \otimes u = \rho : \left(u \cdot_{yz} W^T \right) |J| \left(\frac{\partial \zeta}{\partial z} \right)^2. \quad (8.83)$$

Note that the W matrix represents the second derivative of the Legendre interpolants.

Comment: For our particular choice of domain, $\partial^2 \xi / \partial x^2 = 0$ (respectively for η, ζ), so $\mathbf{A}_{ii} \otimes u = \mathbf{Q}_i \otimes u$.

Elemental matrix for second derivative in space $\mathbf{D}_{ii}$ type 2

The advection matrix $\mathbf{D}_{ii}$ results from the integration by parts of

$$\int_{\mathbf{\Omega}} f \partial_i \partial_i u \; w \; d\mathbf{\Omega}, \quad (8.84)$$

and appears in the following type of integral

$$\int_{\mathbf{\Omega}} \partial_i u \; \partial_i (f w) \; d\mathbf{\Omega} = \mathbf{D}_{ii}(f) \otimes u. \quad (8.85)$$

This type 2 matrix $\mathbf{D}_{ii}$ is an alternative version to $\mathbf{A}_{ii}$, to avoid the use of the second node differentiation matrix W.

In terms of elemental matrices, we have (for i, j equal)

$$\mathbf{D}_{ii}(f) \otimes u = -(\mathbf{A}_i \otimes u) : (\mathbf{A}_i \otimes f) \frac{1}{|J|} - \mathbf{K}_{ii}(f) \otimes u, \quad (8.86)$$

where $\mathbf{K}_{ii}$ is the stiffness matrix. Including the boundary term from the integration by parts, we have the elemental discretization

$$\int_{\mathbf{\Omega}} f \partial_i \partial_i u \; w \; d\mathbf{\Omega} = \mathbf{B} \otimes (\partial_i u) - (\mathbf{A}_i \otimes u) : (\mathbf{A}_i \otimes f) \frac{1}{|J|} - \mathbf{K}_{ii}(f) \otimes u. \quad (8.87)$$

Furthermore, in the case of 2 multiplying functions f and g, we have

$$\begin{aligned}\mathbf{D}_{ii}(f:g)\otimes u &= -g:(\mathbf{A}_i\otimes u):(\mathbf{A}_i\otimes f)\,\frac{1}{|J|}-f:(\mathbf{A}_i\otimes u):(\mathbf{A}_i\otimes g)\,\frac{1}{|J|}\\ &\quad -\mathbf{K}_{ii}(f:g)\otimes u. \end{aligned} \tag{8.88}$$

Elemental stiffness matrix $\mathbf{K}_{ii}$

The elemental stiffness matrix $\mathbf{K}_{ii}$ appears in the following type of integral

$$\int_{\mathbf{\Omega}} f\partial_i u\;\partial_i w\;d\mathbf{\Omega} \;=\; \mathbf{K}_{ii}(f)\otimes u, \tag{8.89}$$

where, f, u and w are scalar functions and $i = x, y$, or z. This type of integral results from the integration by parts of

$$\int_{\mathbf{\Omega}} f\partial_i\partial_i u\;w\;d\mathbf{\Omega}. \tag{8.90}$$

The operator $\otimes$ will act on u in a different manner depending on the value of i as described below:

- $i = x$

$$\mathbf{K}_{xx}(f)\otimes u \;=\; \left[H^T\cdot_{xy}\left(f:|J|\,\left(\frac{\partial\xi}{\partial x}\right)^2:\rho:(H\cdot_{xy}u)\right)\right]; \tag{8.91}$$

- $i = y$

$$\mathbf{K}_{yy}(f)\otimes u \;=\; \left[\left(f:|J|\,\left(\frac{\partial\eta}{\partial y}\right)^2:\rho:\left(u\cdot_{xy}H^T\right)\right)\cdot_{xy}H\right]; \tag{8.92}$$

- $i = z$

$$\mathbf{K}_{zz}(f)\otimes u \;=\; \left[\left(f:|J|\,\left(\frac{\partial\zeta}{\partial z}\right)^2:\rho:\left(u\cdot_{yz}H^T\right)\right)\cdot_{yz}H\right]; \tag{8.93}$$

8.3.3 Specific Elemental matrices to the BSSN system

Elemental matrix for the Christoffel symbol $\mathbf{\Lambda}^a_{bc}$ associated with the metric $\tilde{g}_{ij}$

The elemental matrix for the Christoffel symbol $\mathbf{\Lambda}^a_{bc}$ appears in

$$\int_{\mathbf{\Omega}} f\tilde{\Gamma}^a_{bc}\;w\;d\mathbf{\Omega} \;=\; f:\mathbf{\Lambda}^a_{bc}, \tag{8.94}$$

where

$$\mathbf{\Lambda}^a_{bc} \;=\; \frac{1}{2}\tilde{g}^{al}:\left(\mathbf{A}_b\otimes\tilde{g}_{lc}+\mathbf{A}_c\otimes\tilde{g}_{lb}-\mathbf{A}_l\otimes\tilde{g}_{bc}\right). \tag{8.95}$$

Elemental matrix for the contracted Christoffel symbol $\mathbf{\Lambda}_{abc}$ associated with the metric $\tilde{g}_{ij}$

The elemental matrix for the Christoffel symbol $\mathbf{\Lambda}_{abc}$ appears in

$$\int_{\mathbf{\Omega}} f\tilde{\Gamma}_{abc}\, w\, d\mathbf{\Omega} \;=\; \int_{\mathbf{\Omega}} f\tilde{g}_{am}\tilde{\Gamma}^{m}_{\ bc}\, w\, d\mathbf{\Omega} \;=\; f : \mathbf{\Lambda}_{abc}, \tag{8.96}$$

where

$$\mathbf{\Lambda}_{abc} \;=\; \frac{1}{2}\tilde{g}_{am} : \tilde{g}^{ml} : \left(\mathbf{A}_b \otimes \tilde{g}_{lc} + \mathbf{A}_c \otimes \tilde{g}_{lb} - \mathbf{A}_l \otimes \tilde{g}_{bc}\right). \tag{8.97}$$

Elemental matrix for the product of Christoffel symbols and contracted Christoffel symbols $\mathbf{\Lambda}^{a}_{\ bc}\mathbf{\Lambda}_{def}$

The elemental matrix for the product of the Christoffel symbol $\tilde{\Gamma}^{a}_{\ bc}$ and the contracted form $\tilde{\Gamma}_{def}$ appears in

$$\int_{\mathbf{\Omega}} f\tilde{\Gamma}^{a}_{\ bc}\tilde{\Gamma}_{def}\, w\, d\mathbf{\Omega} \;=\; f : \left(\mathbf{\Lambda}^{a}_{\ bc}\mathbf{\Lambda}_{def}\right), \tag{8.98}$$

where

$$\begin{aligned}\mathbf{\Lambda}^{a}_{\ bc}\mathbf{\Lambda}_{def} \;=\; & \frac{1}{4}\tilde{g}^{al} : \left(\mathbf{A}_b \otimes \tilde{g}_{lc} + \mathbf{A}_c \otimes \tilde{g}_{lb} - \mathbf{A}_l \otimes \tilde{g}_{bc}\right) : \\ & \tilde{g}_{dm} : \tilde{g}^{mo} : \left(\mathbf{A}_e \otimes \tilde{g}_{of} + \mathbf{A}_f \otimes \tilde{g}_{oe} - \mathbf{A}_o \otimes \tilde{g}_{ef}\right)\, \frac{1}{|J|}.\end{aligned} \tag{8.99}$$

Elemental matrix for the covariant derivatives in space $\mathbb{D}_{ij}$

The second covariant derivative $\mathbb{D}_{ij}$ appears in the following term

$$\int_{\mathbf{\Omega}} f D_i D_j u\, w\, d\mathbf{\Omega} \;=\; \mathbb{D}_{ij}(f) \otimes u, \tag{8.100}$$

where $u = \alpha$ or $u = \phi$ and the definition of $D_i D_j \alpha$ is given in equation (8.13).

Version 1 If we do not integrate by parts the second order space derivative terms, we have

$$\mathbb{D}_{ij}(f) \otimes u \;=\; f : \left(\mathbb{D}_{ij} \otimes u\right). \tag{8.101}$$

And for the case $u = \alpha$, we have the following elemental matrix definitions

$$\begin{aligned}\mathbb{D}_{ij} \otimes u \;=\; & \mathbf{A}_{ij} \otimes u - 4\left(\mathbf{A}_{(i} \otimes \phi\right) : \left(\mathbf{A}_{j)} \otimes u\right)\, \frac{1}{|J|} - \mathbf{\Lambda}^{k}_{\ ij} : \left(\mathbf{A}_k \otimes u\right)\, \frac{1}{|J|} \\ & + 2g_{ij}g^{kl} : \left[\left(\mathbf{A}_k \otimes \phi\right) : \left(\mathbf{A}_l \otimes u\right)\right]\, \frac{1}{|J|}.\end{aligned} \tag{8.102}$$

For the case $u = \phi$, we have

$$\mathbb{D}_{ij} \otimes u \;=\; \mathbf{A}_{ij} \otimes u - \mathbf{\Lambda}^{k}_{\ ij} : \left(\mathbf{A}_k \otimes u\right)\, \frac{1}{|J|}. \tag{8.103}$$

Version 2 If we do integrate by parts the second order space derivative terms, we have for the case $u = \alpha$,

$$\begin{aligned}\mathbb{D}_{ij}(f) \otimes u &= \mathbf{D}_{ij}(f) \otimes u + \Big[-4 \left(\mathbf{A}_{(i} \otimes \phi\right) : \left(\mathbf{A}_{j)} \otimes u\right) \frac{1}{|J|} \\ &\quad - \mathbf{\Lambda}^k_{ij} : (\mathbf{A}_k \otimes u) \frac{1}{|J|} + 2\tilde{g}_{ij}\tilde{g}^{kl} : [(\mathbf{A}_k \otimes \phi) : (\mathbf{A}_l \otimes u)] \frac{1}{|J|} \Big].\end{aligned} \tag{8.104}$$

For the case $u = \phi$, we have

$$\mathbb{D}_{ij} \otimes u = \mathbf{D}_{ij} \otimes u - \mathbf{\Lambda}^k_{ij} : (\mathbf{A}_k \otimes u) \frac{1}{|J|}. \tag{8.105}$$

Note that we use the definition of symmetry

$$\left(\mathbf{A}_{(i} \otimes \phi\right) : \left(\mathbf{A}_{j)} \otimes u\right) = \frac{1}{2}\Big[(\mathbf{A}_i \otimes \phi) : (\mathbf{A}_j \otimes u) + (\mathbf{A}_j \otimes \phi) : (\mathbf{A}_i \otimes u) \Big]. \tag{8.106}$$

Elemental matrix for the Ricci tensor $\mathbb{R}_{ij}$

The elemental matrix for the Ricci tensor appears in the following

$$\int_{\mathbf{\Omega}} f R_{ij}\, w\, d\mathbf{\Omega} = \mathbb{R}_{ij}(f) = \left(\tilde{\mathbb{R}}_{ij}(f) + \mathbb{R}^{\phi}_{ij}(f)\right), \tag{8.107}$$

Version 1 If we do not integrate by parts the second order space derivative terms, we have

$$\mathbb{R}_{ij}(f) = \left(\tilde{\mathbb{R}}_{ij}(f) + \mathbb{R}^{\phi}_{ij}(f)\right) = f : \left(\tilde{\mathbb{R}}_{ij} + \mathbb{R}^{\phi}_{ij}\right) \tag{8.108}$$

where

$$\begin{aligned}\tilde{\mathbb{R}}_{ij} &= -\frac{1}{2}\tilde{g}^{lm} : (\mathbf{A}_{lm} \otimes \tilde{g}_{ij}) + \tilde{g}_{k(i} : \left(\mathbf{A}_{j)} \otimes \tilde{\Gamma}^k\right) \\ &\quad + \tilde{\Gamma^k} : \mathbf{\Lambda}_{(ij)k} + \tilde{g}^{lm} : \left(2\mathbf{\Lambda}^k_{l(i}\mathbf{\Lambda}_{j)km} + \mathbf{\Lambda}^k_{jm}\mathbf{\Lambda}_{klj}\right);\end{aligned} \tag{8.109}$$

and

$$\begin{aligned}\mathbb{R}^{\phi}_{ij} &= -2\mathbb{D}_{ij} \otimes \phi - 2\tilde{g}_{ij} : \tilde{g}^{kl} : (\mathbb{D}_{lk} \otimes \phi) + 4 (\mathbf{A}_i \otimes \phi) : (\mathbf{A}_j \otimes \phi) \frac{1}{|J|} \\ &\quad -4\tilde{g}_{ij} : \tilde{g}^{kl} : (\mathbf{A}_l \otimes \phi) : (\mathbf{A}_k \otimes \phi) \frac{1}{|J|}.\end{aligned} \tag{8.110}$$

Version 2 If we do integrate by parts the second order space derivative terms, we have

$$\begin{aligned}\tilde{\mathbb{R}}_{ij}(f) &= -\frac{1}{2}\left(\mathbf{D}_{lm}(f:\tilde{g}^{lm})\otimes\tilde{g}_{ij}\right)+\tilde{g}_{k(i}:\left(\mathbf{A}_{j)}\otimes\tilde{\Gamma}^{k}\right)\\ &\quad +\tilde{\Gamma^{k}}:\mathbf{\Lambda}_{(ij)k}+\tilde{g}^{lm}:\left(2\mathbf{\Lambda}^{k}_{l(i}\mathbf{\Lambda}_{j)km}+\mathbf{\Lambda}^{k}_{jm}\mathbf{\Lambda}_{klj}\right);\end{aligned} \tag{8.111}$$

and

$$\begin{aligned}\mathbb{R}^{\phi}_{ij}(f) &= -2\mathbb{D}_{ij}(f)\otimes\phi-2\left(\mathbb{D}_{lk}(f:\tilde{g}_{ij}:\tilde{g}^{kl})\otimes\phi\right)+4\left(\mathbf{A}_{i}\otimes\phi\right):\left(\mathbf{A}_{j}\otimes\phi\right)\frac{1}{|J|}\\ &\quad -4\tilde{g}_{ij}:\tilde{g}^{kl}:\left(\mathbf{A}_{l}\otimes\phi\right):\left(\mathbf{A}_{k}\otimes\phi\right)\frac{1}{|J|}.\end{aligned} \tag{8.112}$$

Elemental matrix $\mathbb{X}^{TF}$ for the X^{TF}_{ij} term

The integral that contains the trace free part TF term

$$\int_{\mathbf{\Omega}} f\left[-D_iD_j\alpha+\alpha R_{ij}\right]^{TF}\ w\ d\mathbf{\Omega} \;=\; \mathbb{X}^{TF}(f), \tag{8.113}$$

is discretized by

$$\begin{aligned}\mathbb{X}^{TF}(f) &= f:\left[-\mathbb{D}_{ij}\otimes\alpha+\alpha:\mathbb{R}_{ij}\right]^{TF}\\ &= \left(-\mathbb{D}_{ij}(f)\otimes\alpha+\mathbb{R}_{ij}(f,\alpha)\right)\\ &\quad -\frac{1}{3}\left(-\mathbb{D}_{lk}(f:g_{ij}:g^{kl})\otimes\alpha+\mathbb{R}_{lk}(f:\alpha:g_{ij}:g^{kl})\right).\end{aligned} \tag{8.114}$$

Elemental boundary terms B

The elemental boundary matrix $\mathbf{B}$ appears in the following integrals

$$\int_{\mathbf{\Gamma}}\mathbf{n}^{k}u\ w\ d\mathbf{\Gamma} \;=\; \mathbf{B}\otimes u. \tag{8.115}$$

The term $\mathbf{B}\otimes u$ will depend strongly on the choice of the boundary conditions on u on $\mathbf{\Gamma}$.

As mentioned in section 7.1.4 , if we integrate by parts on each element separately there would be many boundary terms at the interior boundaries. For the exact solutions these terms cancel in pairs, but the spectral elements are only $\mathcal{C}^0$ and the subdomain wall boundary terms do not cancel. This difference arises from the fact that discretization and derivation of the weak form do not commute for $\mathcal{C}^0$ spectral elements. These extra terms go to zero in the limit so the spectral element strategy is to ignore them which is equivalent to performing an integration by parts first and discretization second. As previously mentioned, this is referred to as the ***variational crime*** [120].

In 3D and in a ***rectangular*** domain, this means that we look at 6 boundaries $\mathbf{\Gamma}_1$, $\mathbf{\Gamma}_2$, $\mathbf{\Gamma}_3$, $\mathbf{\Gamma}_4$, $\mathbf{\Gamma}_5$ and $\mathbf{\Gamma}_6$ corresponding to the 6 faces of the domain. See figure 7.8 for a clearer picture of the domain $\mathbf{\Omega}$ and boundaries $\mathbf{\Gamma}$.

Terms containing second order space derivatives The following equations contain terms with second order space derivatives that need to be integrated by parts:

$$\partial_t \tilde{A}_{ij} \sim \exp(-4\phi)\left[-D_i D_j \alpha + \alpha R_{ij}\right]^{TF} \quad (8.116)$$

$$\exp(-4\phi) X_{ij} - \frac{1}{3}\exp(-4\phi)\tilde{g}^{kl} X_{lk} \quad (8.117)$$

$$\partial_t K \sim -D^i D_i \alpha; \quad (8.118)$$

$$\partial_t \tilde{\Gamma}^i \sim \tilde{g}^{jk}\partial_j\partial_k\beta^i + \frac{1}{3}\tilde{g}^{ij}\partial_j\partial_k\beta^k; \quad (8.119)$$

in particular with,

$$\tilde{R}_{ij} \sim -\frac{1}{2}\tilde{g}^{lm}\partial_l\partial_m\tilde{g}_{ij}; \quad (8.120)$$

$$R^{\phi}_{ij} \sim -2\tilde{D}_i\tilde{D}_j\phi - 2\tilde{g}_{ij}\tilde{D}^k\tilde{D}_k\phi; \quad (8.121)$$

$$D_i D_j \alpha \sim \partial_i\partial_j\alpha; \quad (8.122)$$

$$D^i D_i \alpha \sim \exp(-4\phi)\tilde{g}^{ij}\partial_i\partial_j\alpha; \quad (8.123)$$

$$\tilde{D}_i\tilde{D}_j\phi \sim \partial_i\partial_j\phi. \quad (8.124)$$

There is an integration by parts only for the terms $\partial_i\partial_j u = \partial_i\partial_i u$ with obviously $i = j$.

Formulae for integration by parts There are two cases:

$$\int_\Omega \partial_i\partial_i u \; w d\Omega = \int_\Gamma \left(\partial_i u \cdot \mathbf{n}\right) w d\Gamma - \int_\Omega \partial_i u \; \partial_i w d\Omega \quad (8.125)$$

and with a multiplying function

$$\int_\Omega f\partial_i\partial_i u \; w d\Omega = \int_\Gamma \left((f\partial_i u)\cdot \mathbf{n}\right) w d\Gamma - \int_\Omega \partial_i f \partial_i u \; w d\Omega - \int_\Omega f\partial_i u \partial_i w d\Omega. \quad (8.126)$$

Note that $\mathbf{n}$ is the normal unit vector for each surface Γ

$$\mathbf{n} = \begin{cases} \mathbf{n}^x \\ \mathbf{n}^y \\ \mathbf{n}^z \end{cases} \quad (8.127)$$

and $\mathbf{n}^x, \mathbf{n}^y, \mathbf{n}^z$ have values (± 1 or 0) depending on the surface Γ (6 surfaces in 3D). The term $\cdot$ is the scalar product and hence for some function u we have the following

$$u \cdot \mathbf{n} = u^x\mathbf{n}^x + u^y\mathbf{n}^y + u^z\mathbf{n}^z. \quad (8.128)$$

Terms that need boundary conditions in the standard puncture data After a few calculations, here is the summary of the terms that will need values or relations on the boundaries $\Gamma_i,\ i = 1..6$:

- Terms that appear in $\partial_t \tilde{A}_{ij}$, $\partial_t K$ and more specifically in $D^i D_i \alpha$:

$$\int_\Gamma \Big(\exp(-4\phi)\ \tilde{g}^{ii} \partial_i \alpha \Big) \cdot \mathbf{n}\ w d\Gamma, \qquad \text{with summation on } i. \tag{8.129}$$

- Terms that appear in $\partial_t \tilde{\Gamma}^i$, and more specifically in $\tilde{g}^{jk} \partial_j \partial_k \beta^i$:

$$\int_\Gamma \Big(\tilde{g}^{jj} \partial_j \beta^i \Big) \cdot \mathbf{n}\ w d\Gamma, \qquad \text{with summation on } j. \tag{8.130}$$

- Terms that appear in $\partial_t \tilde{\Gamma}^i$, and more specifically in $\frac{1}{3} \tilde{g}^{ij} \partial_j \partial_k \beta^k$:

$$\int_\Gamma \Big(\tilde{g}^{ii} \partial_i \beta^i \Big) \cdot \mathbf{n}\ w d\Gamma, \qquad \text{with NO summation on } i. \tag{8.131}$$

- Terms that appear in $\partial_t \tilde{A}_{ij}$, and more specifically in X_{ij} that is $\alpha \tilde{R}_{ij}$:

$$\int_\Gamma \Big(\alpha \tilde{g}^{ll} \partial_l \tilde{g}_{ij} \Big) \cdot \mathbf{n}\ w d\Gamma, \qquad \text{with summation on } l. \tag{8.132}$$

- Terms that appear in $\partial_t \tilde{A}_{ij}$, and more specifically in $-\frac{1}{3} \exp(-4\phi) \tilde{g}^{kl} X_{lk}$ (variant of the above):

$$\int_\Gamma \Big(\exp(-4\phi) \tilde{g}^{kl} \alpha \tilde{g}^{mm} \partial_m \tilde{g}_{lk} \Big) \cdot \mathbf{n}\ w d\Gamma, \qquad \text{with summation on } m, l, k; \tag{8.133}$$

  and

$$\int_\Gamma \Big(\exp(-4\phi) \tilde{g}^{ll} \alpha \partial_l \phi \Big) \cdot \mathbf{n}\ w d\Gamma, \qquad \text{with summation on } l. \tag{8.134}$$

- Terms that appear in $\partial_t \tilde{A}_{ij}$, and more specifically in αR^ϕ_{ij}:

$$\int_\Gamma \Big(\alpha \partial_i \phi \Big) \cdot \mathbf{n}\ w d\Gamma, \qquad \text{with NO summation on } i; \tag{8.135}$$

  and

$$\int_\Gamma \Big(\alpha \tilde{g}_{ij} \tilde{g}^{kk} \partial_k \phi \Big) \cdot \mathbf{n}\ w d\Gamma, \qquad \text{with summation on } k \text{ only.} \tag{8.136}$$

Terms that need boundary conditions in the stationary trumpet Schwarzschild puncture data The stationary trumpet Schwarzschild puncture data will be defined in detail in Chapter 9, it is the initial data we have used for all our numerical results for the BSSN system presented in this thesis. These calculations, for the stationary solution, are just simplification of the general case (non-stationary) for the boundaries $\Gamma_i,\ i = 1..6$:

- Terms that appear in $\partial_t \tilde{A}_{ij}$, $\partial_t K$ and more specifically in $D^i D_i \alpha$:

$$\int_\Gamma \Big(\exp(-4\phi)\ \partial_i \alpha \Big) \cdot \mathbf{n}\ w d\Gamma, \qquad \text{with summation on } i. \tag{8.137}$$

- Terms that appear in $\partial_t \tilde{\Gamma}^i$, and more specifically in $\tilde{g}^{jk}\partial_j\partial_k\beta^i$:

$$\int_\Gamma \Big(\partial_j \beta^i\Big) \cdot \mathbf{n}\, wd\Gamma, \qquad \text{with summation on } j. \tag{8.138}$$

- Terms that appear in $\partial_t \tilde{\Gamma}^i$, and more specifically in $\frac{1}{3}\tilde{g}^{ij}\partial_j\partial_k\beta^k$:

$$\int_\Gamma \Big(\partial_i \beta^i\Big) \cdot \mathbf{n}\, wd\Gamma, \qquad \text{with NO summation on } i. \tag{8.139}$$

- Terms that appear in $\partial_t \tilde{A}_{ij}$, and more specifically in X_{ij} that is $\alpha\tilde{R}_{ij}$:

$$\int_\Gamma \Big(\alpha\partial_l\tilde{\delta}_{ii}\Big) \cdot \mathbf{n}\, wd\Gamma, \qquad \text{with summation on } l. \tag{8.140}$$

  This term should be identically zero for this particular solution.

- Terms that appear in $\partial_t \tilde{A}_{ij}$, and more specifically in $-\frac{1}{3}\exp(-4\phi)\tilde{g}^{kl}X_{lk}$ (variant of the above):

$$\int_\Gamma \Big(\exp(-4\phi)\alpha\partial_m\tilde{\delta}_{kk}\Big) \cdot \mathbf{n}\, wd\Gamma, \qquad \text{with summation on } m, k. \tag{8.141}$$

  This term should be identically zero for this particular solution., and

$$\int_\Gamma \Big(\exp(-4\phi)\alpha\partial_l\phi\Big) \cdot \mathbf{n}\, wd\Gamma, \qquad \text{with summation on } l. \tag{8.142}$$

- Terms that appear in $\partial_t \tilde{A}_{ij}$, and more specifically in αR^ϕ_{ij}:

$$\int_\Gamma \Big(\alpha\partial_i\phi\Big) \cdot \mathbf{n}\, wd\Gamma, \qquad \text{with NO summation on } i; \tag{8.143}$$

  and

$$\int_\Gamma \Big(\alpha\tilde{\delta}_{ij}\partial_k\phi\Big) \cdot \mathbf{n}\, wd\Gamma, \qquad \text{with summation on } k \text{ only.} \tag{8.144}$$

To investigate the SEM, we simply impose the analytic solution at the boundaries instead of using Sommerfeld boundary conditions. When the boundaries are pushed far away, it simplifies the boundary conditions immensely. At spatial infinity, the analytic solution of most of the variables are zero or constant and therefore their spatial derivatives are zero. It is important to note that, there is no need to introduce different equations on the boundaries as is the case in FD.

We have briefly discussed a possible way of treating the boundary conditions, however, further work is needed to investigate the application of the Sommerfeld boundary conditions to the BSSN system with the SEM.

8.4 Assembly of global discretization matrix

All the elemental contributions now need to be "added together" this is called ***the assembly of the global matrix***. Now, reintroducing the superscript $\mathbf{k}$ for the $\mathbf{k}$th-element, we construct the general global matrix $\mathbf{A}$ from its associated elemental matrix $\mathbf{A}^{\mathbf{k}}$,

$$A = \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' A^{\mathbf{k}}, \tag{8.145}$$

where $\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}'$ represents the assembly process or direct stiffness summation.

8.4.1 Global assembled matrix system of the BSSN system version 1

The elemental matrix form of the BSSN system of version 1 is given by,

$$\begin{aligned}\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \mathbf{M}\otimes\dot{\phi} &= \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \Bigg[\mathbf{B}\otimes\Big(\beta^k:\phi\Big)-\mathbf{D}_k\otimes\Big(\beta^k:\phi\Big)-\phi:\Big(\mathbf{A}_k\otimes\beta^k\Big)\\ &\quad+\frac{1}{6}\mathbf{A}_k\otimes\beta^k-\frac{1}{6}\mathbf{M}\otimes(\alpha:K)\Bigg]^{\mathbf{k}};\end{aligned} \tag{8.146}$$

alternatively, for the χ-method,

$$\begin{aligned}\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \mathbf{M}\otimes\dot{\chi} &= \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \Bigg[\mathbf{B}\otimes\Big(\beta^k:\chi\Big)-\mathbf{D}_k\otimes\Big(\beta^k:\chi\Big)-\frac{5}{3}\chi:\Big(\mathbf{A}_k\otimes\beta^k\Big)\\ &\quad+\frac{2}{3}\mathbf{M}\otimes(\chi:\alpha:K)\Bigg]^{\mathbf{k}};\end{aligned} \tag{8.147}$$

$$\begin{aligned}\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \mathbf{M}\otimes\dot{\tilde{g}}_{ij} &= \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \Bigg[\mathbf{B}\otimes\Big(\beta^k:\tilde{g}_{ij}\Big)-\mathbf{D}_k\otimes\Big(\beta^k:\tilde{g}_{ij}\Big)-\frac{5}{3}\tilde{g}_{ij}:\Big(\mathbf{A}_k\otimes\beta^k\Big)\\ &\quad+\tilde{g}_{ik}:\Big(\mathbf{A}_j\otimes\beta^k\Big)+\tilde{g}_{jk}:\Big(\mathbf{A}_i\otimes\beta^k\Big)\\ &\quad-2\mathbf{M}\otimes\Big(\alpha:\tilde{A}_{ij}\Big)\Bigg]^{\mathbf{k}};\end{aligned} \tag{8.148}$$

$$\begin{aligned}\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \mathbf{M}\otimes\dot{\tilde{A}}_{ij} &= \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \Bigg[\mathbf{B}\otimes\Big(\beta^k:\tilde{A}_{ij}\Big)-\mathbf{D}_k\otimes\Big(\beta^k:\tilde{A}_{ij}\Big)-\frac{5}{3}\tilde{A}_{ij}:\Big(\mathbf{A}_k\otimes\beta^k\Big)\\ &\quad+\tilde{A}_{ik}:\Big(\mathbf{A}_j\otimes\beta^k\Big)+\tilde{A}_{jk}:\Big(\mathbf{A}_i\otimes\beta^k\Big)+\mathbb{X}^{TF}(e^{-4\phi})\\ &\quad+\mathbf{M}\otimes\Big(\alpha:K:\tilde{A}_{ij}-2\alpha:\tilde{A}_{ik}:\tilde{A}^k_j\Big)\Bigg]^{\mathbf{k}};\end{aligned} \tag{8.149}$$

$$\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \mathbf{M}\otimes\dot{K} = \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \Bigg[\mathbf{B}\otimes\Big(\beta^k : K\Big) - \mathbf{D}_k\otimes\Big(\beta^k : K\Big) - K : \Big(\mathbf{A}_k\otimes\beta^k\Big) + \mathbf{M}\otimes\left[\alpha : \left(\tilde{A}_{ij} : \tilde{A}^{ij} + \frac{1}{3}K : K\right)\right] - (\mathbb{D}_{li}(\tilde{g}_{il})\otimes\alpha)\Bigg]^{\mathbf{k}}; \tag{8.150}$$

$$\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \mathbf{M}\otimes\dot{\tilde{\Gamma}}^i = \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \Bigg[-\tilde{g}^{ij} : \big(\mathbf{A}_{jk}\otimes\beta^i\big) + \frac{1}{3}\tilde{g}^{ij} : \Big(\mathbf{A}_{jk}\otimes\beta^k\Big) + 2\alpha : \left[\mathbf{\Lambda}^i{}_{jk} : \tilde{A}^{jk} + 6\tilde{A}^{ij}\left(\mathbf{A}_j\otimes\phi\right) - \frac{2}{3}\tilde{g}^{ij} : \left(\mathbf{A}_j\otimes K\right)\right] - \tilde{\Gamma}^j : \big(\mathbf{A}_j\otimes\beta^i\big) + \frac{2}{3}\tilde{\Gamma}_i : \big(\mathbf{A}_j\otimes\beta^j\big) - \beta^j : \Big(\mathbf{A}_j\otimes\tilde{\Gamma}^i\Big) - 2\tilde{A}^{ij} : \left(\mathbf{A}_j\otimes\alpha\right)\Bigg]^{k}. \tag{8.151}$$

And finally for the lapse and the shift:

$$\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \mathbf{M}\otimes\dot{\alpha} = \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \Bigg[\mathbf{B}\otimes\Big(\beta^k : \alpha\Big) - \mathbf{D}_k\otimes\Big(\beta^k : \alpha\Big) - \alpha : \Big(\mathbf{A}_k\otimes\beta^k\Big) - 2\mathbf{M}\otimes(\alpha : K)\Bigg]^{\mathbf{k}}; \tag{8.152}$$

$$\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \mathbf{M}\otimes\dot{\beta}^i = \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \left[\beta^k : \big(\mathbf{A}_k\otimes\beta^i\big) - \frac{3}{4}\mathbf{M}\otimes B^i\right]^{\mathbf{k}}; \tag{8.153}$$

$$\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \mathbf{M}\otimes\dot{B}^i = \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \Bigg[\beta^k : \big(\mathbf{A}_k\otimes B^i\big) + \mathbf{M}\otimes\dot{\tilde{\Gamma}}^i - \eta\mathbf{M}\otimes B^i\Bigg]^{\mathbf{k}}. \tag{8.154}$$

Recall that the dot notation $\dot{\beta}$ refers to $\partial_t\beta$ the time derivative of β. The symbol : refers to the term by term matrix multiplication (Hadamard product). On the other hand, the symbol $\otimes$ is multiplication operator which exact definition depends on each elemental matrix type and space dimensions for differentiating elemental matrices.

8.4.2 Global assembled matrix system of the BSSN system version 2

The elemental matrix form of the BSSN system of version 2 is given by,

$$\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \mathbf{M}\otimes\dot{\phi} = \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \left[\mathbf{L}_\beta\otimes\phi - \frac{1}{6}\mathbf{M}\otimes(\alpha : K)\right]^{\mathbf{k}}; \tag{8.155}$$

alternatively, for the χ-method,

$$\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \mathbf{M}\otimes\dot{\chi} = \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \left[\mathbf{L}_\beta\otimes\chi - \frac{2}{3}\beta^k : (\mathbf{A}_k\otimes\chi) + \frac{2}{3}\mathbf{M}\otimes(\chi:\alpha:K)\right]^{\mathbf{k}}; \tag{8.156}$$

$$\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \mathbf{M}\otimes\dot{\tilde{g}}_{ij} = \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \left[\mathbf{L}_\beta\otimes\tilde{g}_{ij} - 2\mathbf{M}\otimes\left(\alpha:\tilde{A}_{ij}\right)\right]^{\mathbf{k}}; \tag{8.157}$$

$$\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \mathbf{M}\otimes\dot{\tilde{A}}_{ij} = \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \left[\mathbf{L}_\beta\otimes\tilde{A}_{ij} + \mathbb{X}^{TF}(e^{-4\phi}) + \mathbf{M}\otimes\left(\alpha:K:\tilde{A}_{ij} - 2\alpha:\tilde{A}_{ik}:\tilde{A}^k_j\right)\right]^k; \tag{8.158}$$

$$\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \mathbf{M}\otimes\dot{K} = \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \left[\mathbf{L}_\beta\otimes K - (\mathbb{D}_{li}(\tilde{g}_{il})\otimes\alpha) + \mathbf{M}\otimes\left[\alpha:\left(\tilde{A}_{ij}:\tilde{A}^{ij} + \frac{1}{3}K:K\right)\right]\right]^{\mathbf{k}}; \tag{8.159}$$

$$\begin{aligned}\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \mathbf{M}\otimes\dot{\tilde{\Gamma}}^i = \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \Bigg[&\tilde{g}^{ij}:\left(\mathbf{A}_{jk}\otimes\beta^i\right) + \frac{1}{3}\tilde{g}^{ij}:\left(\mathbf{A}_{jk}\otimes\beta^k\right)\\ &+2\alpha:\left[\mathbf{\Lambda}^i{}_{jk}:\tilde{A}^{jk} + 6\tilde{A}^{ij}\left(\mathbf{A}_j\otimes\phi\right) - \frac{2}{3}\tilde{g}^{ij}:\left(\mathbf{A}_j\otimes K\right)\right]\\ &-\tilde{\Gamma}^j:\left(\mathbf{A}_j\otimes\beta^i\right) + \frac{2}{3}\tilde{\Gamma}_i:\left(\mathbf{A}_j\otimes\beta^j\right) - \beta^j:\left(\mathbf{A}_j\otimes\tilde{\Gamma}^i\right)\\ &-2\tilde{A}^{ij}:\left(\mathbf{A}_j\otimes\alpha\right)\Bigg]^{\mathbf{k}}.\end{aligned} \tag{8.160}$$

And finally for the lapse and the shift:

$$\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \mathbf{M}\otimes\dot{\alpha} = \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \left[\mathbf{L}_\beta\otimes\alpha - 2\mathbf{M}\otimes(\alpha:K)\right]^{\mathbf{k}}; \tag{8.161}$$

$$\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \mathbf{M}\otimes\dot{\beta}^i = \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \left[\mathbf{L}_\beta\otimes\beta^i - \frac{3}{4}\mathbf{M}\otimes B^i\right]^{\mathbf{k}}; \tag{8.162}$$

$$\sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \mathbf{M}\otimes\dot{B}^i = \sum_{\mathbf{k}=1}^{\mathbf{k}=N_E}{}' \left[\mathbf{L}_\beta\otimes B^i + \mathbf{M}\otimes\dot{\tilde{\Gamma}}^i - \eta\mathbf{M}\otimes B^i\right]^{\mathbf{k}}; \tag{8.163}$$

8.5 Time Discretization

The time discretization of the system

$$\dot{U} \;=\; \mathcal{A}U + \mathcal{F} \;=\; f(U,t), \tag{8.164}$$

is computed by an explicit fourth order Runge–Kutta method. Given an initial condition U_0, the solution U_{n+1} at time t_{n+1} is determined from the previous time t_n and the solution U_n. The details of the Runge–Kutta fourth order method can be found more explicitly in 6.10.

8.6 Conclusion

In this Chapter, we have presented an overview of the spectral element method applied to the BSSN system, a hyperbolic space+time reformulation of the Einstein equations of general relativity.

We have applied the variational formulation to the BSSN system and presented several possible weak forms. From these weak forms, we have explained in detail how the elemental matrix forms specific to the BSSN system are calculated in light of the spectral element discretization.

We have briefly discussed a possible way of treating the boundary condition. However, further work is needed to investigate the application of the Sommerfeld-like boundary conditions to the BSSN system with the SEM.

Finally, we have also presented the global system of algebraic equations of the reformulated Einstein equations.

A computation is a temptation that should be resisted as long as possible.

J. P. Boyd, paraphrasing T. S. Eliot

9 Exploring the Spectral Element Method for moving puncture simulations

In this Chapter, we present numerical experiments and results from the applications of the spectral element method to the BSSN system, and by extension to the Einstein equations of general relativity.

The first implementation of any numerical method to a complex problem such as the Einstein equations is not an easy task. Many numerical experimentations are needed to understand the behaviour of the SEM applied to our specific problem: the BSSN system. How does the method handle the irregularities, steep gradients and discontinuities across the puncture? How much resolution is needed and in which part of the domain, in order to obtain accurate and stable simulations? Through numerous numerical experiments, we have tried to answer these questions and get an overall understanding and feeling of how the SEM works with our problem at hand.

We have studied a particular stationary solution using the ***Schwarzschild trumpet puncture data*** solution derived in [5, 6]. We will discuss our motivations behind the choice of the Matlab language to implement our numerical code. We will then illustrate the geometric flexibility of the method before presenting numerical results, showing how the SEM handles discontinuities at the puncture and further away in the smoother parts of the solution.

9.1 The puncture data for a Schwarzschild black hole

There exists a true stationary "$1 + \log$" slice[1] through the Schwarzschild spacetime called the ***trumpet*** solution [5, 6]. Refer to figure 9.1 to clearly see why the geometry is called ***trumpet***. This solution is flat at one end and cylindrical of radius $R \sim 1.31M$ at the other. When this solution is transformed to isotropic coordinates, it provides puncture trumpet data that are time independent in a moving-puncture simulation. This solution provides an excellent test-case for our numerical evolution of the SEM applied to the BSSN system. The moving-puncture approach is currently the most popular method for simulating black-hole binaries, but for such codes, there is no analytic black hole solution that can be used to test a new code, except the "$1 + \log$" stationary trumpet solution presented in [5, 6] and also

[1] See section 5.5 for a discussion on slicing and gauge conditions in numerical relativity.

in [147, 148]. Note that there also exists a "maximal" trumpet solution.

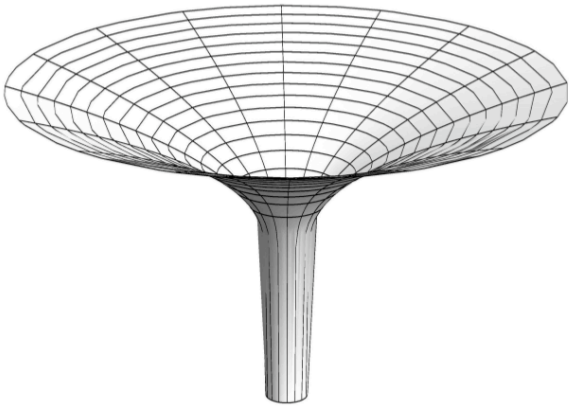


Figure 9.1: Embedding diagram of a 2 dimensional slice of a maximal puncture ***trumpet*** data solution [5, 6]

The Schwarzschild metric in Schwarzschild coordinates is

$$ds^2 = -f dT^2 + f^{-1} dR^2 + R^2 d\Omega^2, \tag{9.1}$$

where $f = 1 - 2M/R$. The quantities R and T denote the Schwarzschild radial coordinate and Schwarzschild time. The surface $R = 2M$ is the event horizon, $R = 0$ is a physical singularity, and $R \to \infty$ is spatial infinity (keeping T fixed). Now, to obtain the metric in isotropic coordinates, we apply the following coordinate transformation $R = \psi^2 r$, where we define ψ as

$$\psi = 1 + \frac{M}{2r}. \tag{9.2}$$

The Schwarzschild metric then becomes

$$ds^2 = -\left(\frac{1 - \frac{M}{2r}}{1 + \frac{M}{2r}}\right)^2 dT^2 + \psi^4 (dr^2 + r^2 d\Omega^2). \tag{9.3}$$

The Schwarzschild metric in isotropic coordinates is better adapted to the standard puncture method. The isotropic coordinate r does not reach the physical singularity at $R = 0$. For large r we have $R \to \infty$, but for small r we have again $R \to \infty$. There is a minimum of $R = 2M$ at $r = M/2$. We now have two copies of the space outside the event horizon $R > 2M$, and the two spaces are connected by a wormhole with a throat at $R = 2M$. The point $r = 0$ is referred to as the ***puncture***.

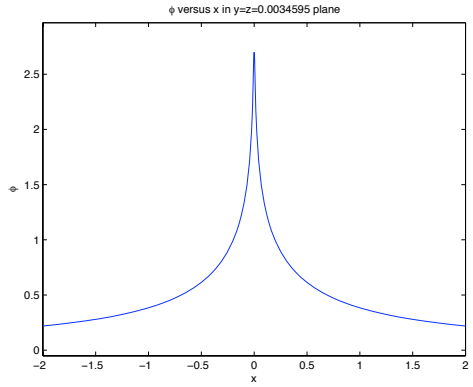
The Schwarzschild metric solution in isotropic coordinates is

$$\tilde{g}_{ij} = \delta_{ij}, \tag{9.4}$$

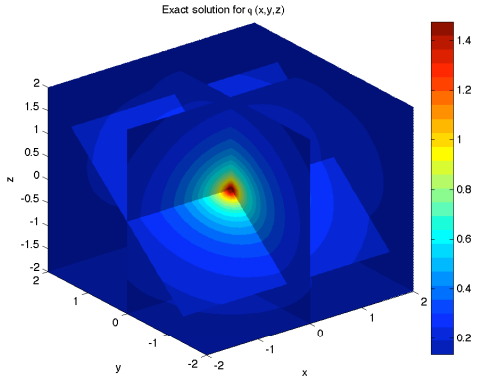
$$\psi = 1 + \frac{M}{2r}, \tag{9.5}$$

$$\tilde{A}_{ij} = 0, \tag{9.6}$$

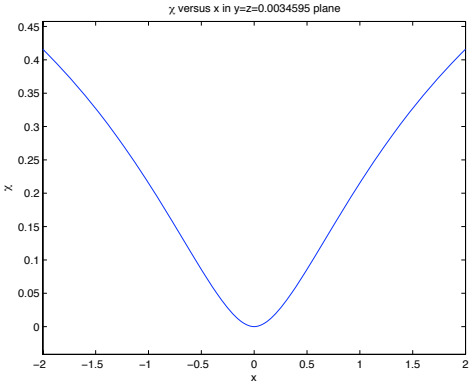
$$K = 0. \tag{9.7}$$



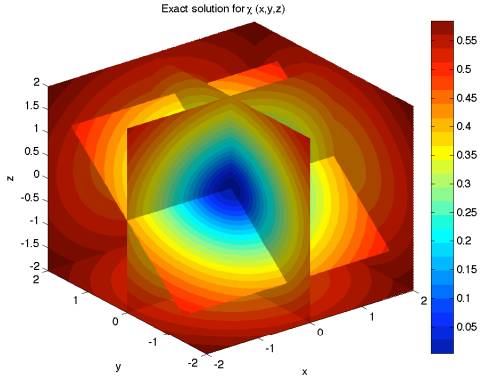
(a) ϕ versus x in the $y = z \sim 0$ plane



(b) $\phi(x, y, z)$ for a domain $L = 2$

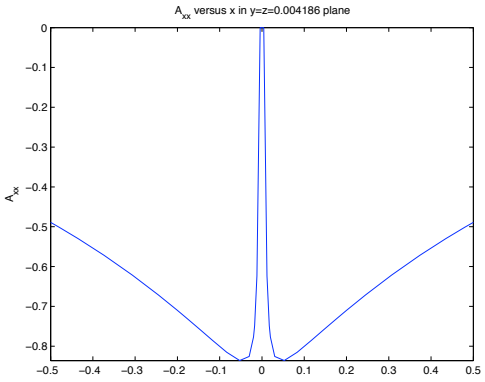


(c) χ versus x in the $y = z \sim 0$ plane

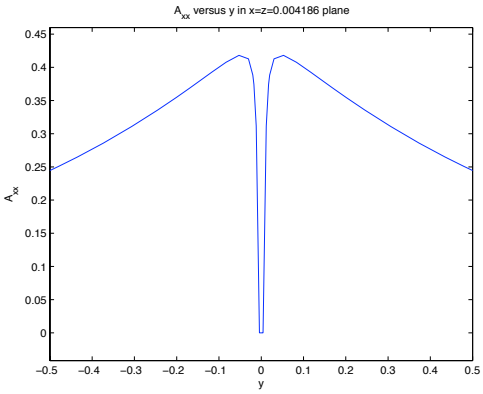


(d) $\chi(x, y, z)$ for a domain $L = 2$

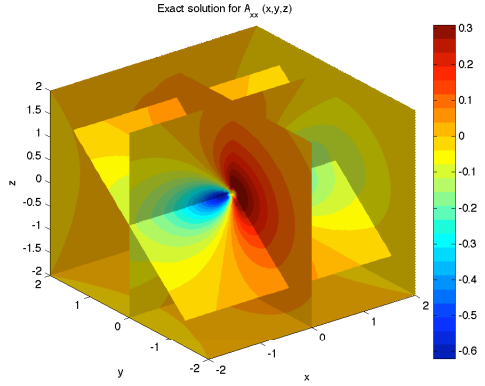
Figure 9.2: Exact solution for χ and ϕ for a trumpet Schwarzschild black hole



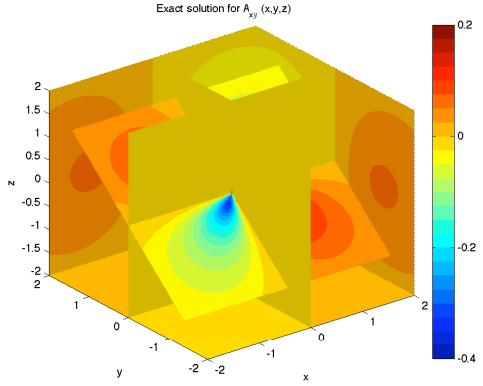
(a) $\tilde{A}_{xx}$ versus x in the $y = z \sim 0$ plane



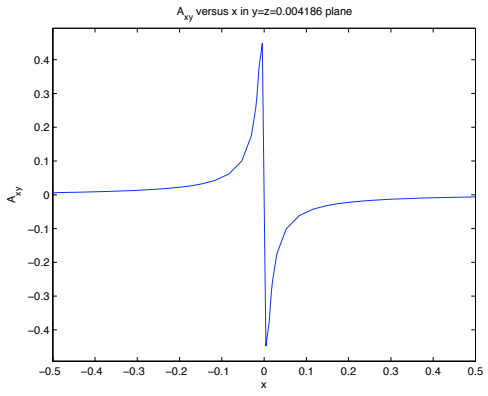
(b) $\tilde{A}_{xx}$ versus y in the $x = z \sim 0$ plane



(c) $\tilde{A}_{xx}(x, y, z)$ for a domain $L = 2$

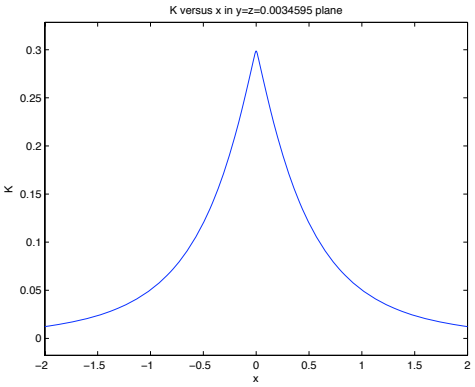


(d) $\tilde{A}_{xy}(x, y, z)$ for a domain $L = 2$

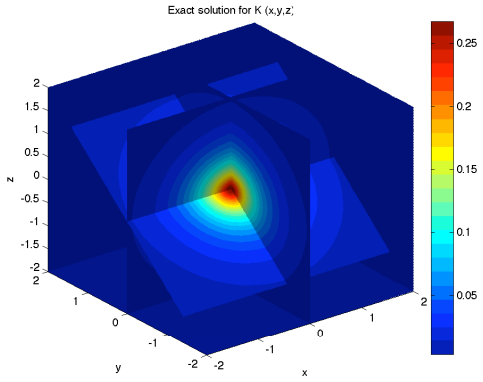


(e) $\tilde{A}_{xy}$ versus x in the $x = z \sim 0$ plane

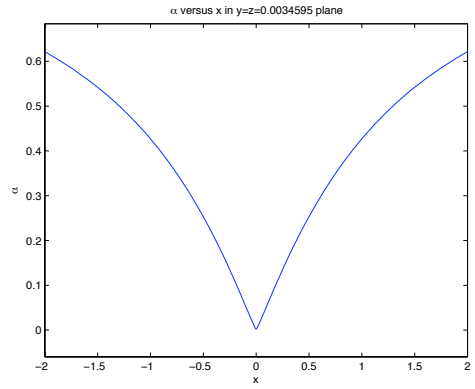
Figure 9.3: Exact solution for $\tilde{A}_{ij}$ for a trumpet Schwarzschild black hole



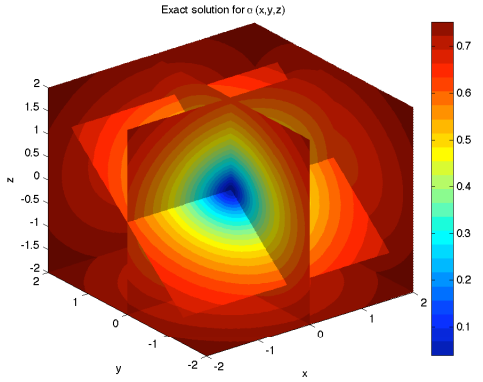
(a) K versus x in the $y = z \sim 0$ plane



(b) $K(x, y, z)$ for a domain $L = 2$

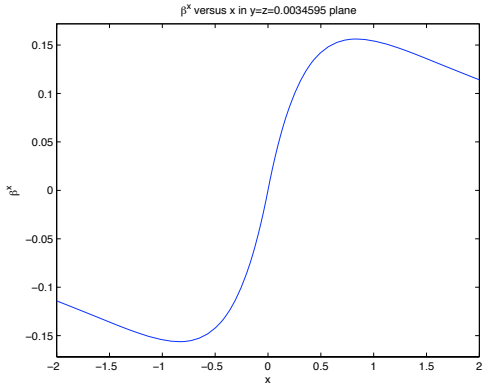


(c) α versus x in the $y = z \sim 0$ plane

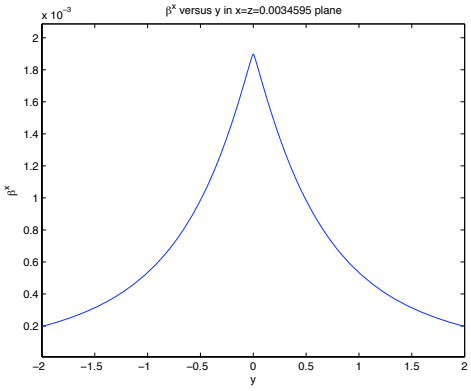


(d) $\alpha(x, y, z)$ for a domain $L = 2$

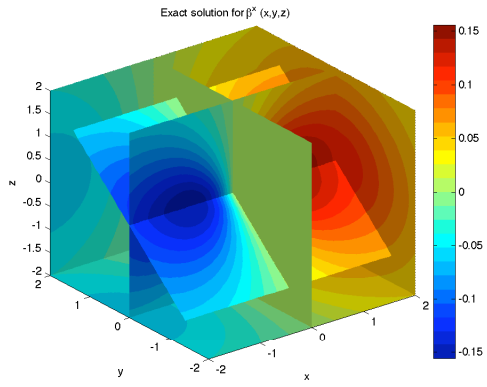
Figure 9.4: Exact solution for K and α for a trumpet Schwarzschild black hole



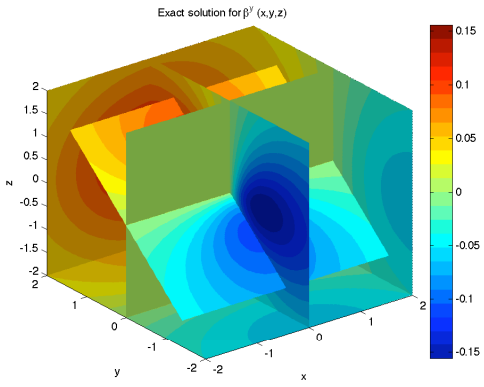
(a) β^x versus x in the $y = z \sim 0$ plane



(b) β^x versus y in the $x = z \sim 0$ plane



(c) $\beta^x(x, y, z)$ for a domain $L = 2$



(d) $\beta^y(x, y, z)$ for a domain $L = 2$

Figure 9.5: Exact solution for β^i for a trumpet Schwarzschild black hole

The lapse and shift are

$$\alpha = \frac{1-\frac{M}{2r}}{1+\frac{M}{2r}}, \tag{9.8}$$

$$\beta^i = 0. \tag{9.9}$$

If equations (9.4) – (9.7) are chosen as initial data, with the lapse (9.8) and shift (9.9), then the data will remain unchanged: this is a ***stationary*** solution. This solution might seem trivial to implement, however, it is difficult to reproduce numerically in a standard 3D black-hole evolution code. Indeed, most codes are not stable when the lapse is negative, which it is here for $r < M/2$. In a numerical code we prefer to use a lapse that is always positive, or at least non-negative. However, a "$1+\log$" or maximal slicing evolution with two asymptotically flat ends and with a non negative lapse cannot reach a stationary state. By giving up one of the flat ends, we can obtain a trumpet stationary solution with a "$1+\log$" slicing condition and a positive lapse. Note that there is also a solution with a maximally sliced condition $K = 0$, but one must solve an elliptic equation at each timestep to find the corresponding lapse function that maintains maximal slicing. This is computationally expensive and it is more practical to choose a slicing condition so that the lapse can be calculated from an evolution equation like the rest of the dynamical variables.

The "$1+\log$" slicing condition corresponds to the following evolution equation:

$$(\partial_t - \beta^i \partial_i)\alpha = -n\alpha K. \tag{9.10}$$

After analytical calculations, the corresponding time-independent Schwarzschild solution can be derived for this slicing condition, see [5, 6] for details. The lapse is now given by

$$\alpha^2 = 1 - \frac{2M}{R} + \frac{C(n)^2 e^{2\alpha/n}}{R^4}, \tag{9.11}$$

where the value of the constant $C(n)$ is given by

$$C^2(n) = \frac{(3n+\sqrt{4+9n^2})^3}{128n^3} e^{-2\alpha_c/n}, \tag{9.12}$$

with the specific value for α_c

$$\alpha_c^2 = \frac{\sqrt{4+9n^2}-3n}{\sqrt{4+9n^2}+3n}. \tag{9.13}$$

We can then calculate β^R (in Schwarzschild coordinates) with the relation:

$$\beta^R = \alpha\sqrt{\alpha^2 - f}. \tag{9.14}$$

The metric term g_{RR} is given by

$$g_{RR} = \frac{1}{\alpha^2}, \tag{9.15}$$

and finally, the extrinsic curvature is described by

$$K_{RR} = \frac{\beta'}{\alpha^2}, \quad (9.16)$$

$$K_{\theta\theta} = R\beta, \quad (9.17)$$

$$K_{\phi\phi} = R\beta \sin^2\theta, \quad (9.18)$$

where $\beta = \sqrt{\beta_i \beta^i}$ and $\beta' = \partial\beta/\partial R$. The trace of the extrinsic curvature $K = K^i_i$ is

$$K = \frac{2\beta}{R} + \beta'. \quad (9.19)$$

Equations (9.11-9.19) give the $1 + \log$ trumpet solution for a Schwarzschild black hole. The horizon is located at $\alpha(R = 2M) = 0.376179$.

Figures 9.2, 9.3, 9.4 and 9.5 show the exact solution for most of the BSSN variables. Note that $\tilde{g}_{ij} = \delta_{ij}$ and $\tilde{\Gamma}^i = 0$ are not represented. The solution is represented as a function of x in the plane $y = z \sim 0$ (near the puncture) and in 3 dimensions as a function of x, y and z. Note that the conformal metric $\tilde{g}_{ij}$ and the extrinsic curvature $\tilde{A}_{ij}$ are symmetric and therefore we have:

$$\tilde{g}_{xy} = \tilde{g}_{yx} \quad \tilde{g}_{xz} = \tilde{g}_{zx} \quad \tilde{g}_{yz} = \tilde{g}_{zy}; \quad (9.20)$$

$$\tilde{A}_{xy} = \tilde{A}_{yx} \quad \tilde{A}_{xz} = \tilde{A}_{zx} \quad \tilde{A}_{yz} = \tilde{A}_{zy}. \quad (9.21)$$

From the aforementioned figures, we can clearly see the properties of each variable. The variables χ, $\tilde{g}_{ij}$, $\tilde{\Gamma}^i$ are all smooth, the variables ϕ, K, α and β^i are continuous but only $\mathcal{C}^0$ so we can expect some of the derivatives to be discontinuous or have at least some ***kinks*** near the puncture. In particular, note that α behaves as $|x|$ close to the puncture. In contrast, the variables $\tilde{A}_{ij}$ are clearly discontinuous across the puncture and these terms will indeed be the most problematic.

Not only is this "$1 + \log$" stationary trumpet solution extremely useful for testing a new code, but it also offers the possibility to test each equation separately, evolving only one variable at a time and keeping all the other variables exact. It is therefore possible to work on each variable as part of an uncoupled system, as well as all the variables as part of a coupled system.

9.2 Behaviour of extrinsic curvature near the puncture

This is easiest to see in the maximal case, where we can write many relations in closed form. The physical extrinsic curvature is given in spherical coordinates by

$$K^i_j = \mathrm{diag}(-2C/R^3, C/R^3, C/R^3), \quad (9.22)$$

where $C^2 = 27M^4/16$. The BSSN extrinsic curvature is then given in Cartesian coordinates by

$$\tilde{A}_{ij} = \frac{C}{\psi^6 r^3}(1 - 3n_x n_y), \quad (9.23)$$

where $n_i = x^i/r$. We therefore have, for example, that

$$\tilde{A}_{xx} = \frac{C}{\psi^6}\frac{y^2+z^2-2x^2}{r^5}. \quad (9.24)$$

Near the puncture, the conformal factor behaves like

$$\psi \approx \sqrt{\frac{3M}{2r}}, \quad (9.25)$$

and so the conformal extrinsic curvature looks like

$$\tilde{A}_{xx} = A\,\frac{y^2+z^2-2x^2}{r^2}, \quad (9.26)$$

where A is a constant. We see that this quantity has direction-dependent limits. If we choose $y = z = 0$, then we have

$$\tilde{A}_{xx}{}_{|_{y=z=0}} = -2A. \quad (9.27)$$

If we instead choose $x = y = 0$, then we have

$$\tilde{A}_{xx}{}_{|_{x=y=0}} = A. \quad (9.28)$$

The same effect also shows up in the $1 + log$ solution. But the length scale is small, and so the effect is only clear with very high resolution near the puncture.

9.3 Experimenting with the SEM and BSSN system: Why? What? Where? How?

Why would a numerical method work or not with a particular system? What would be the best set-up to take full advantage of the SEM for BSSN simulations? Where would the most resolution be needed? How would the method behave overall with our problem?

Through numerous numerical experiments with various meshes and resolution setups, we have investigated the overall behaviour of the SEM with the nonlinear Einstein equations. We have tried to address each of the following:

From a computational point of view. Why did we choose Matlab for our implementation of the SEM? We will also present rough estimates of the memory requirements of the method for the BSSN system.

Geometric flexibility. The SEM is well-known for its geometric flexibility. How easy is it to create a mesh adapted to our problem?

Different versions of the weak form. This tests how much the behaviour depends on the weak form, does it make a big difference? Is this something that deserves more attention?

The ϕ-method versus the χ-method with the SEM. With finite differences, both methods can be used leading to similar results? Is there an advantage using one method over the other with the SEM?

Far from the puncture. This is where we expect the system to behave well. What is the verdict?

Puncture at the centre of an element. This puts the discontinuities in the center of one element, and allows us to look at how the errors behave. This also counts as a sort of "***base case***", from which we can compare cleverer setups. As we will see further on, it turns out that the discontinuity at the puncture is a very significant problem.

How can the problems at the puncture be dealt with, without the use of any stabilization technique (e.g. filtering)? With FD methods, the errors do not propagate away from the puncture, is it possible here, that the errors will not propagate beyond one element?

Offset mesh. To deal with the irregularities at the puncture, it might be a good idea to locate the discontinuity at the edge of an element. Indeed, in structural mechanics, cracks in the material introduce discontinuities in the solution. These are best dealt with by modifying the mesh to cover cracks across elements rather than inside an element. Can we do the same with the BSSN system?

Increasing the number of elements near the puncture. In some problems, this is a way to deal with discontinuities, is this a practical way to deal with the puncture?

Filtering "as much or as little as needed". If all else fails to deal with the discontinuities at the puncture, will filtering make a difference?

Long-term stable evolutions? How does the SEM applied to the BSSN system handle long-term evolutions?

9.4 From a computational point of view

9.4.1 Why Matlab?

The Matlab language was primarily chosen for its ability to integrate numerical computation in particular with fast and simple matrix vectorization calculations and relatively easy computer graphic visualization. Matlab programming language is in some ways superior and in some ways inferior to traditional upper-level languages such as Fortran, C and C++.

As an interpreted language, the instructions in the code are translated into machine language and executed in real time in contrast with a compiled source code therefore Matlab codes are not suitable for large-scale computations. In other words, this translates into a slower running speed for the code.

However, Matlab is ideal for developing and testing the SEM. This involves mostly elemental matrix calculations with the BSSN system, as well as implementing various meshes

and hence using many available graphics components. Matlab includes a great deal of infrastructure – matrix operations, input/output, visualisation, as well as freely available examples of SEM in 1D and 2D [149] – making it much quicker to develop and test new ideas, without having to spend time writing and debugging infrastructure that is irrelevant to the method.

Translation of a Matlab code to another computer language is straightforward and it will offer the possibility to run in parallel which works very well with the SEM since elemental calculations can be done faster on separate CPUs communicating results only during the assembly process.

9.4.2 Memory efficiency of the SEM

In the Matlab code, all the elemental matrices except (Mass and Boundary) are assembled while performing calculations at the same time because of memory storage advantages (many zeros), and hence only the non zero blocks **A**, **D** etc... are stored. The Runge-Kutta method requires such calculations ***4 times*** per timestep over a loop on the number of elements N_E. Remember that N_g is the global number of space points and $N = P + 1$ is the number of GLL points per element.

Memory requirement for the construction of a 3D SEM mesh

- $Iglob$ is the global index function and requires: $\sim N_g \times (8 \text{ Bytes})$;
- $iglob$ is the local to global index function $\mathcal{I}(a, b, c, k)$ and requires: $\sim (N_{GLL})^3 \times N_E \times (8 \text{ Bytes})$;
- Jacobian, first derivatives (e.g $\partial x/\partial \xi$) requires: $\sim (N_{GLL})^3 \times N_E \times n \times (8 \text{ Bytes})$, where n depends on the structure of the mesh $n = 10$ for a general deformed mesh or $n = 4$ when $\partial x/\partial \eta$ and other derivatives are zero.
- x, y, z are the physical coordinates and require: $\sim 3N_g \times (8 \text{ Bytes})$.

Total minimum memory requirement for a 3D mesh:

$$\text{memory} \quad \sim \quad 4N_g \times (8 \text{ Bytes}) + (n+1)(N_{GLL})^3 \times N_E \times (8 \text{ Bytes}) + O\,(8 \text{ Bytes})\,. \tag{9.29}$$

If we fix the memory to a maximum value then we have the following restrictions on the global number of points N_g:

$$N_g \quad < \quad \frac{\text{memory} - (n+1)(N_{GLL})^3 \times N_E \times (8 \text{ Bytes})}{4 \times (8 \text{ Bytes})}. \tag{9.30}$$

Memory requirement for the BSSN evolution equations

- $Iglob$ is the global index function and requires: $\sim N_g \times (8 \text{ Bytes})$;
- $iglob$ is the local to global index function $\mathcal{I}(a, b, c, k)$ and requires: $\sim (N_{GLL})^3 \times N_E \times (8 \text{ Bytes})$;

- x, y, z are the physical coordinates and require: $\sim 3N_g \times (8 \text{ Bytes})$;
- Jacobian, first derivatives (e.g $\partial x/\partial \xi$) requires: $\sim (N_{GLL})^3 \times N_E \times n \times (8 \text{ Bytes})$, where n depends on the structure of the mesh $n = 10$ for a general deformed mesh or $n = 4$ when $\partial x/\partial \eta$ and other derivatives are zero.
- $unknowns$, if we estimate roughly 40 variables the requirement is: ~ 40 var $\times N_g \times (8 \text{ Bytes})$;
  However, using a 4 time stepping scheme $RK4$, all the variables have to be stored 4 times for the Runge-Kutta method, the requirement would be: ~ 4 steps $\times 40$ var $\times N_g \times (8 \text{ Bytes})$;
  Fortunately, we can optimize this by assembling and calculating the right hand sides of each equation separately requiring only 4 auxiliary variables for all the unknown, the memory requirement is now:~ 44 var $\times N_g \times (8 \text{ Bytes})$;
- Elemental matrices, only $\mathbf{M}$ is pre-assembled, the rest (~ 10 to 20 depending on the variable under claculation) are calculated for each element but not stored, the rough memory requirement estimate is:
  $\sim N_g \times (8 \text{ Bytes}) + (\text{nb elemental matrices}) \times (N_{GLL})^3 \times (8 \text{ Bytes})$;
- Node differentiation matrices, H, H^T require: $\sim (2 \text{ node differentiation matrices}) \times (N_{GLL})^2 \times (8 \text{ Bytes})$;

Total minimum memory requirement for the evolution equation:

$$\begin{aligned} \text{memory} \quad \sim \quad & 49 \times N_g \times (8 \text{ Bytes}) + (n+1)(N_{GLL})^3 \times N_E \times (8 \text{ Bytes}) \\ & + (\text{nb elemental matrices}) \times (N_{GLL})^3 \times (8 \text{ Bytes}) \\ & + 2 \times (N_{GLL})^2 \times (8 \text{ Bytes}) + O\,(8 \text{ Bytes})\,. \end{aligned} \tag{9.31}$$

What do these memory requirement estimates tell us? These estimates are an indication of the minimum memory requirements for the application of the SEM to the BSSN system. Using the interpreted language Matlab will certainly imply slower simulations than with any other method implemented with a compiled language. Although we are not able to comment on the computational speed, these memory requirement estimates certainly show an advantage of the SEM over the finite difference method (FD): the derivatives of the variables do not need to be calculated and kept in memory globally over the whole domain, and all calculations are done on an elemental basis only requiring N^3_{GLL} type matrices at a time. Remember that for the SM and SEM, the derivatives of any function consist of the nodal coefficients with the derivatives of the Lagrange–Legendre basis functions. The derivatives of the interpolants only need calculating once and are stored in a N^2_{GLL} matrix H_{ij}.

9.5 Geometric flexibility

In Chapter 7, all the simulations have been performed on a uniform mesh: the "***even***" mesh, where the domain is decomposed evenly, and all the elements have the same size. In the simulations presented there, the solution of the wave equation was traveling across the domain L which was relatively small $L = 2, ..., 7$. The consequent number of elements and hence number of points was not extremely significant and problematic.

To solve the BSSN system, however, we need a large domain of at least $L = 80M$, the larger the domain, the less errors propagating from the boundary conditions spread in the solution. An even mesh is absolutely impractical for such large domains.

For example, to obtain a $\mathcal{L}^2$ norm of 1.10^{-6} on average, for the wave equation in 3D with a domain of $L = 4$, we need, a polynomial order $N = 5$, a total number of elements $N_E = 16^3$, and therefore a total number of points of $N_g = 531\ 441$. To obtain the same level of resolution with an even mesh for a domain of $L = 80$ however, we require the same polynomial order, a total number of elements $N_E = 320^3$, which makes the total number of points $N_g = 4\ 103\ 684\ 801$.

In light of the previous section, the memory requirements for this simulation would be of roughly 223 GigaBytes of RAM.

Since the most spatial resolution needed for the BSSN variables concentrate near the puncture, it is definitely in our interest to design a mesh that would take this property into account.

In this section, we present some "***distorted meshes***" that consist of small elements near the puncture that can get stretched out when moving outwards to the boundaries.

9.5.1 Distorted meshes

For a regular evenly decomposed mesh, the anchor points can be formulated as

$$X_a = -L_X + a\frac{2L_X}{N_{Ex}} \quad \forall a = \{0, ..., N_E\}. \tag{9.32}$$

We can now define a distorted ***square*** mesh in a similar fashion:

$$L_2 = (L_X)^{1/2}, \tag{9.33}$$

$$X_a = \text{sign}\left(-L_2 + a\frac{2L_2}{N_{Ex}}\right)\left(-L_2 + a\frac{2L_2}{N_{Ex}}\right)^2 \quad \forall a = \{0, ..., N_E\}. \tag{9.34}$$

We can also define a distorted ***cubic*** mesh as follows

$$L_3 = (L_X)^{1/3}, \tag{9.35}$$

$$X_a = \left(-L_3 + a\frac{2L_3}{N_{Ex}}\right)^3 \quad \forall a = \{0, ..., N_E\}. \tag{9.36}$$

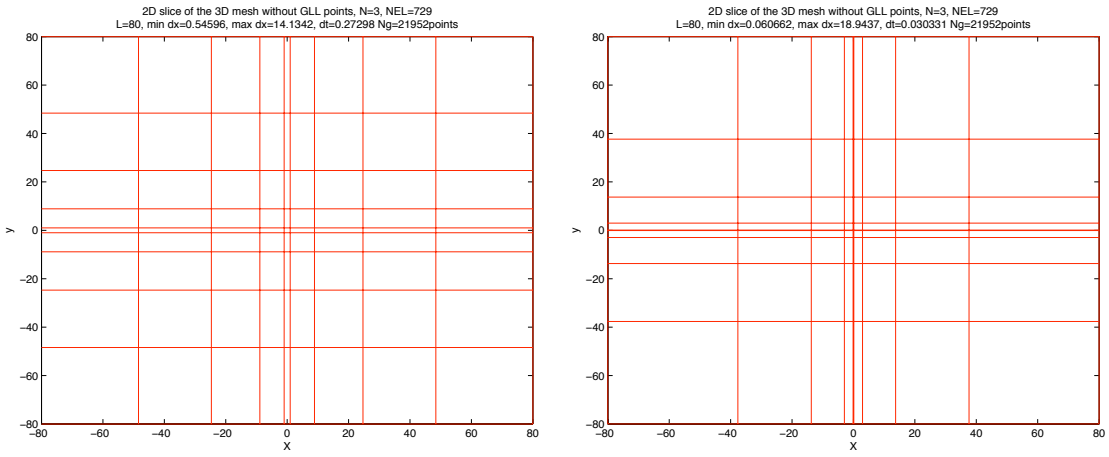
Many types of distorted meshes can be defined in this fashion. What do these meshes look like? These meshes distort an even mesh by squeezing the elements close to the centre and by enlarging elements further away from the centre. This allows for more points near the puncture and less points where we do not need as much resolution.

These types of meshes have some disadvantages: they provide high resolution near the coordinate planes, not just near the puncture, and this is a waste of resources. However, these meshes have the advantage of being simple to implement, and allow enough variation in resolution for testing purposes, and illustrate the basic flexibility of meshes in the SEM approach. The ideal mesh would mimic spherical coordinates at large distances.

In Figure 9.6, we represent 2 examples of distorted square and cubic meshes with the same number of elements N_E and domain $L = 80$. Note that the cubic mesh distorts the size of the elements the most, the elements at the centre are smaller and the elements close to the boundaries are larger than with a square mesh. In Appendix G we present these meshes with varying parameters

- Figure (G.1) present a 2D slice of a 3D distorted ***square*** mesh with anchor points without GLL points with $N_E = 5^3, 7^3, 9^3, 11^3$ respectively.
- Figure (G.2) present a 2D slice of a 3D distorted ***cubic*** mesh with anchor points without GLL points with $N_E = 5^3, 7^3, 9^3, 11^3$ respectively.

Tables (G.1) and (G.2) illustrate the requirements for distorted square meshes and distorted cubic meshes respectively as functions of the number of elements N_E and the polynomial order N and in terms of the minimum and maximum dx, timestep dt required and the total number of points N_g involved.



(a) Square mesh with $N_E = 9^3$

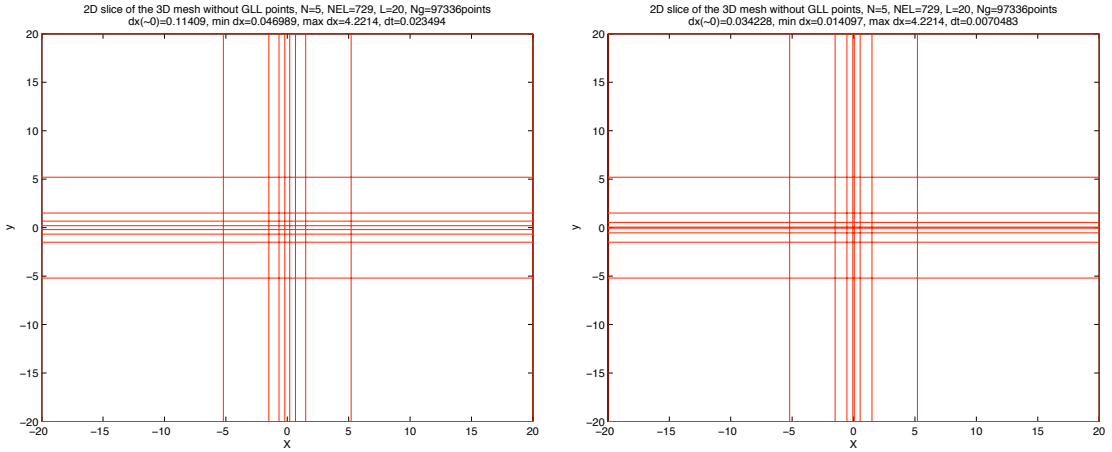
(b) Cubic mesh with $N_E = 9^3$

Figure 9.6: 3D distorted ***square*** and ***cubic*** meshes represented in a 2D slice for the same number of elements $N_E = 9^3$ and domain $L = 80$.

9.5.2 Mixed Distorted meshes

We can also combine the definitions of several mesh types to obtain one mixed distorted mesh in order to ensure a more precise density of points in certain areas. This allows for even more flexibility in the choice of the grid. For example, we can define an inside box to be an even mesh for $L_{Xin} = 1.5$ and a square distorted mesh on the outside up to $L_X = 20$ for example. We can also have a distorted square in the inside box and any other distorted mesh outside. Figure 9.7 illustrates 2 mixed distorted meshes. In Appendix G, we present more mixed distorted meshes for varying number of elements N_E, see figure G.3 for details.

We have illustrated the relative ease in implementing simple distorted meshes more adapted to our problem, taking advantage of the geometric flexibility of the SEM. These



(a) Even inside Box $N_{EIn} = 5^3$, square outside Box, $N_E = 7^3$.

(b) Square inside Box $N_{EIn} = 5^3$, square outside Box, $N_E = 7^3$.

Figure 9.7: 3D mixed distorted meshes represented in a 2D slice for number of elements $N_E = 7^3$ and for a domain $L = 20$. The outside area is a square mesh and the inside box is even for (a), whereas the inside box is square for (b).

distorted and mixed distorted meshes are not ideal. An ideal mesh would mimic spherical coordinates at large distances, they allow for more points near the puncture and less points where we do not need as much resolution.

9.6 Different versions of the weak form

In this section we present results for the evolution equation of χ with 2 types of weak forms introduced in Chapter 8. Remember that in weak form 1, we integrate by parts and obtain:

$$\int_{\mathbf{\Omega}} \partial_t \chi \; w \; d\mathbf{\Omega} = \int_{\mathbf{\Gamma}} \beta^k \mathbf{n}^k \chi w \; d\mathbf{\Gamma} - \int_{\mathbf{\Omega}} \beta^k \chi \partial_k w \; d\mathbf{\Omega} - \frac{5}{3} \int_{\mathbf{\Omega}} \partial_k \beta^k \chi w \; d\mathbf{\Omega} + \frac{2}{3} \int_{\mathbf{\Omega}} \chi \alpha K w \; d\mathbf{\Omega}. \tag{9.37}$$

However, if we do not integrate by parts we obtain the weak form 2:

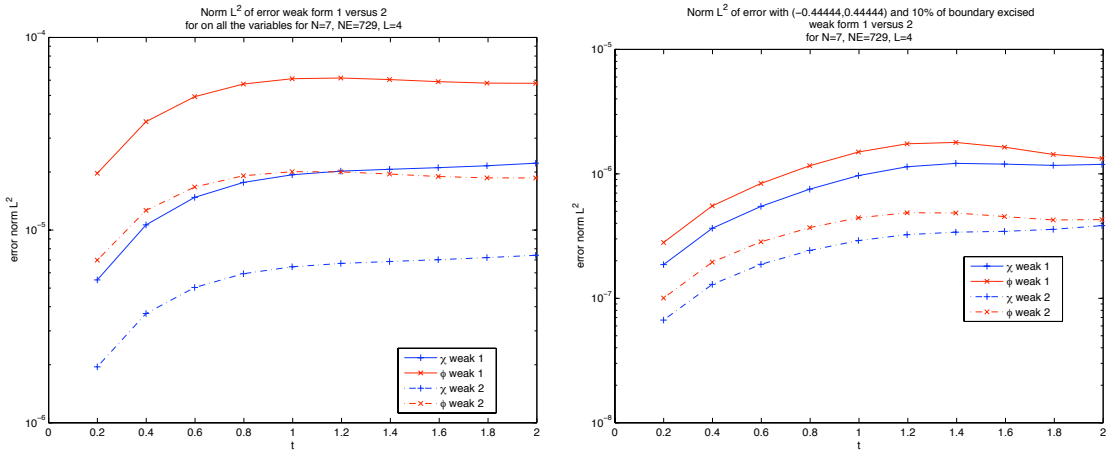
$$\int_{\mathbf{\Omega}} \partial_t \chi \; w \; d\mathbf{\Omega} = \int_{\mathbf{\Omega}} \mathcal{L}_\beta \chi \; w \; d\mathbf{\Omega} + \frac{2}{3} \int_{\mathbf{\Omega}} \chi \alpha K w \; d\mathbf{\Omega}, \tag{9.38}$$

with

$$\int_{\mathbf{\Omega}} \mathcal{L}_\beta \chi \; w \; d\mathbf{\Omega} = \int_{\mathbf{\Omega}} \beta^k \partial_k \chi w \; d\mathbf{\Omega} - \frac{2}{3} \int_{\mathbf{\Omega}} \chi \partial_k \beta^k \; w \; d\mathbf{\Omega}. \tag{9.39}$$

What is the difference numerically between implementing the first or second weak form? It turns out that the difference is roughly of the order of the numerical accuracy in the domain. Furthermore, the first weak form needs boundary conditions imposed whereas the second version does not. So in the end it turns out that the 2nd weak form offers slightly better

numerical results. We will use the second weak form in the following simulations. See figure 9.8 for visual comparisons of the $\mathcal{L}^2$ norms of χ and ϕ obtained from the evolution of χ.



(a) $\mathcal{L}^2$ norm over the entire domain

(b) $\mathcal{L}^2$ norm with the centre element and 10% of the boundary domain excised

Figure 9.8: $\mathcal{L}^2$ norms comparing the implementation of the weak form 1 (in solid lines) and weak form 2 (in dashed dot lines) for χ (in blue +) and ϕ (in red x).

9.7 The ϕ-method versus the χ-method with the SEM

The BSSN system is presented in detail in Chapters 5 and 8. We have mentioned that the conformal factor ψ could be evolved through the variable ϕ (with the ϕ-method) or through the variable χ (with the χ-method).

Recall that with the ϕ-method, one works directly with the original BSSN variable ϕ,

$$\phi = \ln \psi, \tag{9.40}$$

and the evolution equation for ϕ is

$$\partial_t \phi - \mathcal{L}_\beta \phi = -\frac{1}{6}\alpha K \qquad \mathcal{L}_\beta \phi = \beta^k \partial_k \phi + \frac{1}{6}\partial_k \beta^k. \tag{9.41}$$

The purely experimental result is that finite differencing across the $\ln(r)$ singularity at $r = 0$ leads to stable evolutions. Is this the case with the spectral element method?

In the χ-method, a new conformal factor is defined, that is finite at the puncture,

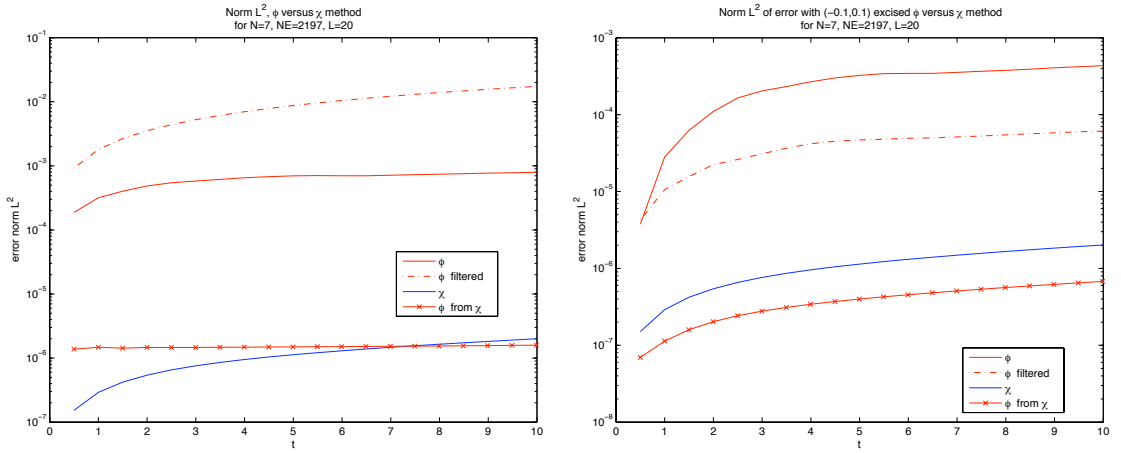
$$\chi = \psi^{-4}, \tag{9.42}$$

with the corresponding evolution equation

$$\partial_t \chi - \mathcal{L}_\beta \chi = \frac{2}{3}\chi \alpha K \qquad \mathcal{L}_\beta \chi = \beta^k \partial_k \chi - \frac{2}{3}\chi \partial_k \beta^k. \tag{9.43}$$

Since χ is initially finite at the puncture, it will be smooth across the entire domain, whereas ϕ will be discontinuous at the puncture. The disadvantage is that one has to make sure χ does not become negative in the evolution code, as $\phi = -1/4 \ln(\chi)$ would not be defined. It is common practice to set χ to a minimum cut-off value when $\chi < 0$.

A priori, it would make sense that the χ-method would exhibit better performance for the SEM. Figure 9.9 shows the comparison of the $\mathcal{L}^2$ norms of the evolution of ϕ, the evolution of ϕ with filtering, the evolution of χ and ϕ calculated from the evolution of χ. In this figure, the variables are evolved up to $t = 10M$, and it is clear that the norm of ϕ obtained from evolving χ gives the most desirable results. Hence, this experimental result confirms that using the χ-method is strongly recommended when using the SEM. All the following simulations in this thesis will be based on the χ-method.



(a) $\mathcal{L}^2$ norms over the entire domain $L = 20$ (b) $\mathcal{L}^2$ norms with the centre element excised $L = 20$

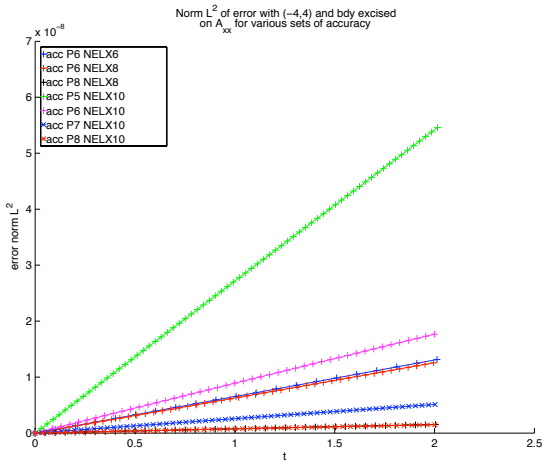
Figure 9.9: The ϕ-method versus the χ-method: comparison of the $\mathcal{L}^2$ norms of the evolution of ϕ, the evolution of ϕ with filtering, the evolution of χ and ϕ calculated form the evolution of χ.

9.8 Far from the puncture

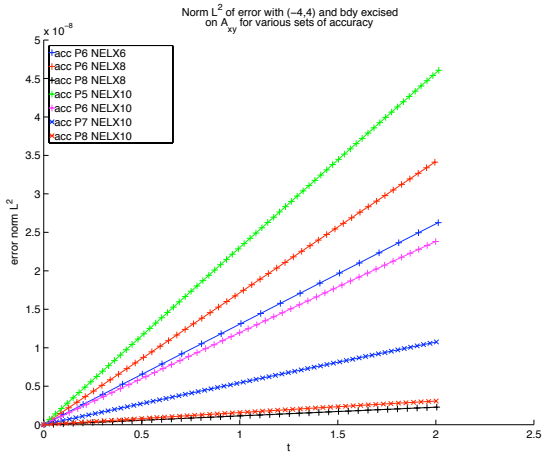
Sufficiently away from the puncture, all the variables of the BSSN system are smooth and we expect the system to behave well. This means that the SEM is able to obtain high-accuracy convergence in this part of the domain.

The variables that require the most attention are $\tilde{A}_{xx}$, $\tilde{A}_{xy}$, K and $\tilde{\Gamma}^x$. To illustrate the behaviour of the SEM far from the puncture, we show the norms of these variables for a domain $L = 64$ with a cubic mesh and varying polynomial order and number of elements in Figure 9.10. Note that these norms do not contain any values from the centre domain $(-4, 4)$, as we wish to only concentrate in the smooth parts of the variables for now.

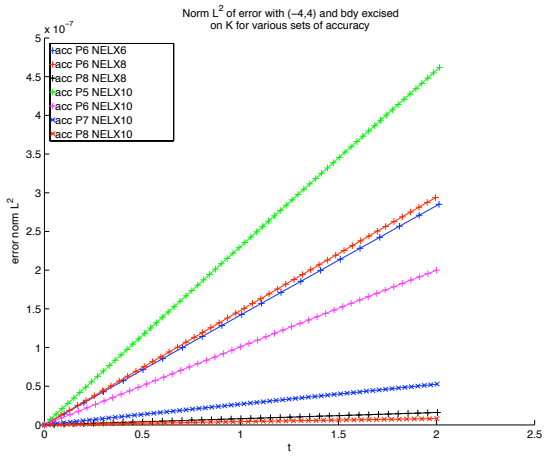
In appendix G, we show the pointwise errors and $\mathcal{L}^2$ norms of all the variables in detail, see Figures G.4, G.5, G.6, G.7, G.8, G.9, G.10, G.11, and G.12.



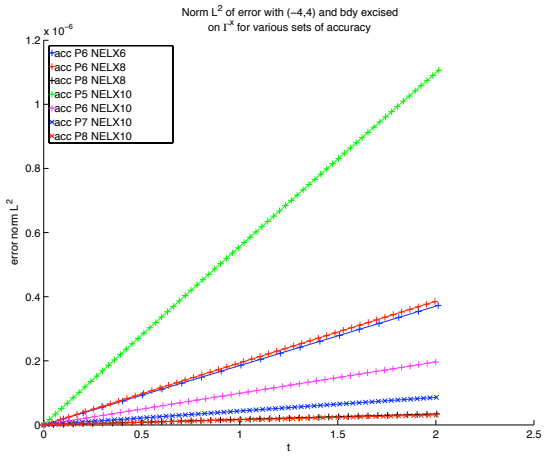
(a) Norm $\mathcal{L}^2$ of $\tilde{A}_{xx}$ with the centre excised and the boundary excised



(b) Norm $\mathcal{L}^2$ of $\tilde{A}_{xy}$ with the centre excised and the boundary excised



(c) Norm $\mathcal{L}^2$ of K with the centre excised and the boundary excised



(d) Norm $\mathcal{L}^2$ of $\tilde{\Gamma}^x$ with the centre excised and the boundary excised

Figure 9.10: Pointwise error and $\mathcal{L}^2$ norm for ϕ at the same time steps for varying accuracy and for different slices across a domain of $L = 64$ with a cubic mesh.

We can easily quantify the level of errors for $\tilde{g}_{ij} = \delta_{ij}$, as it is straightforward to interpret errors when the true value is one. In particular, let us concentrate on the results shown in figure G.6 for $\tilde{g}_{xx}$ for the lowest and highest resolution presented there. The lowest resolution contains $N_g = 50\ 653$ points, whereas the highest resolution contains more than 10 times as many points with $N_g = 531\ 441$ points. In the plane $y = z \sim 5.7$M the maximum pointwise error (absolute value) for the lowest resolution is of $1.5\ 10^{-5}$. For 10 times the resolution, the maximum error decreases to $2.5\ 10^{-7}$. Moving away from the puncture at roughly 50% of the domain away from the puncture (in the plane $y = z \sim 31.9$M), we obtain $1.5\ 10^{-8}$ versus $2.5\ 10^{-11}$. Closer to the boundaries, in the plane $y = z \sim 59$M, the maximum error is now $3.\ 10^{-10}$ for the lowest resolution and $7.5\ 10^{-13}$ for the highest res-

olution. To get an overall picture, we can also look at the percentage error of the $\mathcal{L}^2$ norms with the centre excised concentrating only in the smooth parts: on average, the lowest resolution gives an error of $1\ 10^{-6}$, with slightly more than 10 times the number of points, the error comes down to $5\ 10^{-9}$. By increasing the resolution by 10 times, on average, we have divided the error by $1\ 000$.

Although we have briefly discussed a possible way of treating the boundary conditions, further work is needed to investigate the application of the Sommerfeld-like boundary conditions to the BSSN system with the SEM. In these simulations, we have used analytic boundary conditions. We could also look at more physically appropriate boundary conditions [150, 151, 152, 153].

9.8.1 hp-convergence with χ

For ***infinitely smooth solutions*** and for an ***evenly*** decomposed domain, h-refinement usually leads to an ***algebraic*** decay of the numerical error, whereas, p-refinement usually leads to an ***exponential*** decay. In other words, the hp-convergence shows the rates of convergence when varying both the number of elements N_E and the polynomial order N.

Here we wish to show the results of the investigation of the hp-convergence for the variable χ in details for various setups of meshes: evenly decomposed mesh (***even*** mesh) and distorted meshes (***cubic*** and ***square*** meshes). Although χ is initially smooth even at the puncture, its evolution equation contains derivatives of β^i which are not completely regular at the puncture. Therefore, we look at hp-convergence norms after a very small evolution time $t \sim 0.1$M, before most oscillations (even small) have the potential to propagate from the puncture.

It is useful to look at the hp-convergence of χ on an evenly decomposed mesh and see how the results compare with the solution of the 3D wave equation.

- Figures 9.11(a) and 9.11(b), show the hp-convergence for an even domain of $L = 4$, which is a small domain but it allows us to work with reasonable amount points.

We also illustrate the shapes of algebraic and exponential convergence rates for comparison purposes. Recall that on a log-log axis, algebraic convergence asymptotes to a straight line whose slope is $-k$ (where k represents the index of convergence), whereas exponential convergence bends away with ever-increasing negative slopes. In the figures we have set $k = 1$. In figure 9.11(b), only the centre element is excised from the $\mathcal{L}^2$ norm, and we see that as the resolution increases around the puncture, the hp-convergence fades. This is due to the fact that some oscillations propagate in the neighbour elements very quickly, see figure G.13 in the appendix for the behaviour of χ for the highest resolution: oscillations are present outside the centre element extremely quickly.

- Figures 9.11(c) and 9.11(d), show the hp-convergence for a distorted ***cubic*** mesh of $L = 64$.

- Figures 9.11(e) and 9.11(f), show the hp-convergence for a distorted ***square*** mesh of $L = 64$.

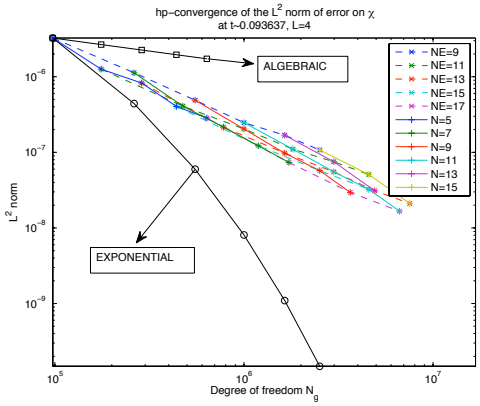
What conclusions can we draw from figure 9.11?

Over the entire domain, we do not obtain hp-convergence, this should be completely expected because even though χ is initially smooth, its evolution equation introduces discontinuities at the puncture.

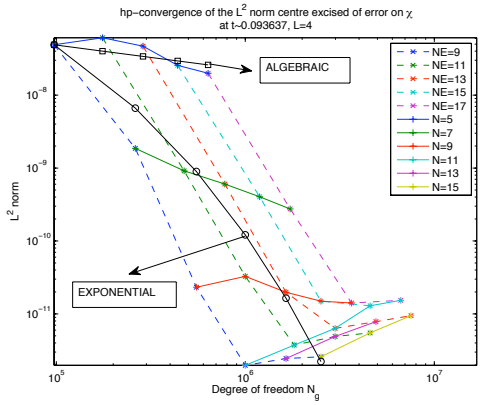
On the other hand, when excising the errors near the puncture from the $\mathcal{L}^2$ norm, we do obtain convergence rates that are very close to hp-convergence rates. Not only do we have hp-convergence in the smooth parts of the solution for an evenly decomposed domain (as expected), but it is also the case for distorted meshes. Although the rates of convergence are only approximate for a cubic mesh on the algebraic side (highly distorted domain), the rates of convergence are surprisingly clean for the square mesh.

These results do not just stand for the variable χ, all the other variables of the BSSN system present similar rates of convergence as the resolution is increased, in the smooth parts of the solutions. Hence, the convergence rates are very similar to the numerical results obtained with the 3D wave equation in Chapter 7.

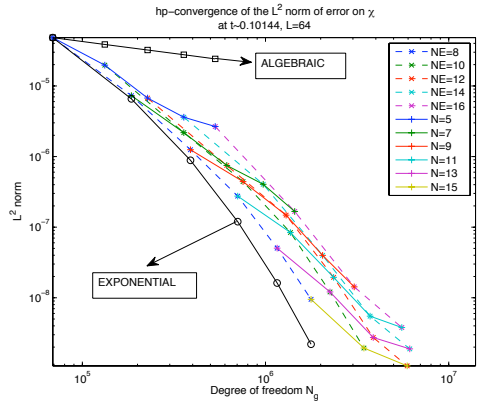
These numerical results demonstrate that the part of the code and system that we expect to behave well really do so. This section illustrates the power of the method, and more importantly, that the code is implemented correctly.



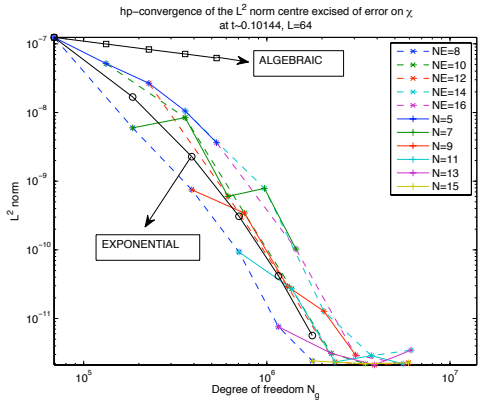
(a) hp-convergence with $\mathcal{L}^2$ norms over the entire domain, ***even*** mesh $L = 4$



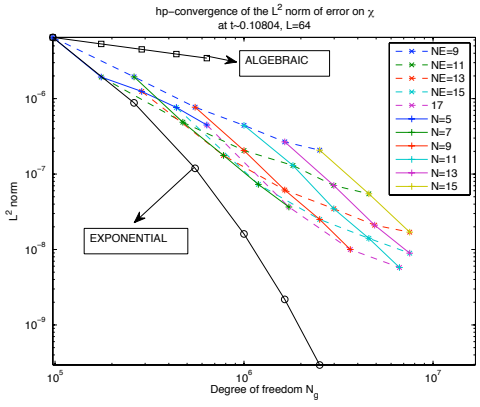
(b) hp-convergence with $\mathcal{L}^2$ norms with the centre element excised, ***even*** mesh $L = 4$



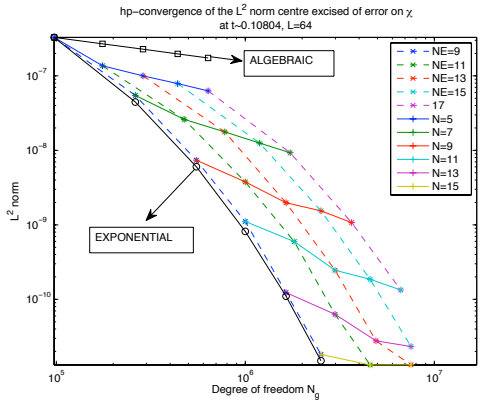
(c) Same as (a) but with a ***cubic*** mesh $L = 64$



(d) Same as (b) but with a ***cubic*** mesh $L = 64$



(e) Same as (a) but with a ***square*** mesh $L = 64$



(f) Same as (b) but with a ***square*** mesh $L = 64$

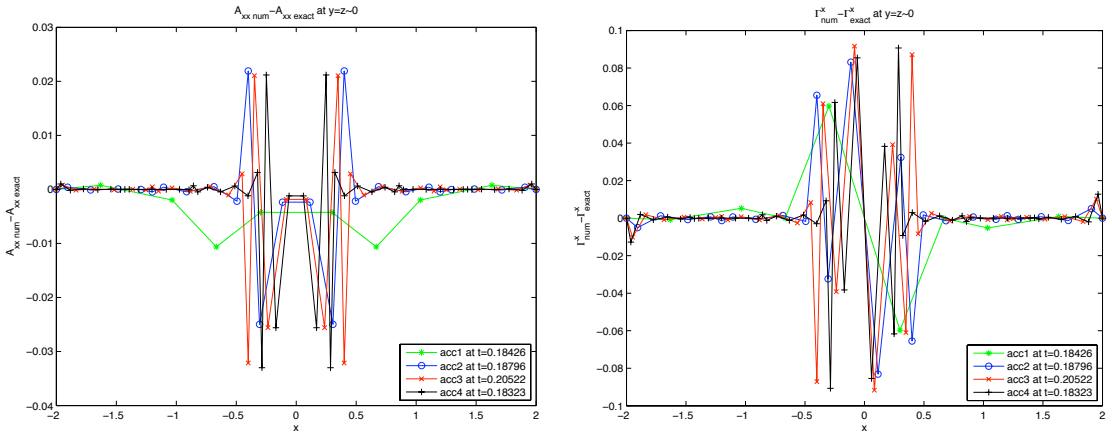
Figure 9.11: hp-convergence for χ on various types of meshes: evenly decomposed mesh of $L = 4$, cubic mesh and square mesh of $L = 64$.

9.9 Puncture at the centre of an element

We first present some results close to the discontinuity with an even mesh, that is, all the elements have the same size throughout the whole domain. Here, we wish to investigate the behaviour of the method near the puncture. Therefore, we run simulations on a very small domain $L = 2$ to be close to the point of irregularity, for short evolution times.

The 4 types of accuracy we use in our tests are defined below:

1. Accuracy 1): $N = 3$, $N_E = 3$, $N_g = 1000$ points (acc 1);
2. Accuracy 2): $N = 5$, $N_E = 5$, $N_g = 17576$ points (acc 2);
3. Accuracy 3): $N = 7$, $N_E = 5$, $N_g = 46656$ points (acc 3);
4. Accuracy 4): $N = 7$, $N_E = 7$, $N_g = 125000$ points (acc 4).



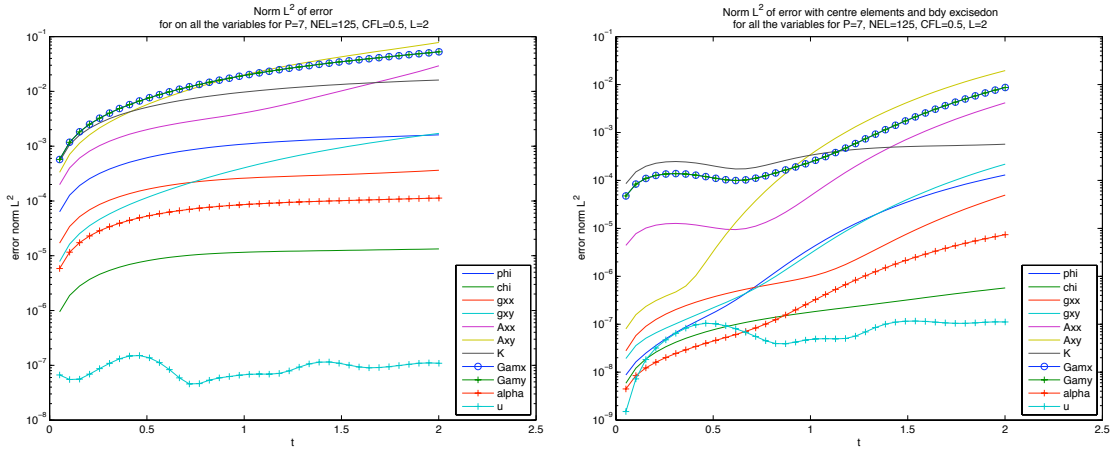
(a) Pointwise error of $\tilde{A}_{xx}$ near the puncture

(b) Pointwise error of $\tilde{\Gamma}^x$ near the puncture

Figure 9.12: Pointwise error of $\tilde{A}_{xx}$ and $\tilde{\Gamma}^x$ for 4 types of accuracy for $L = 2$ at $t \sim 0.2M$: acc1, acc2, acc3 and acc4.

Figure 9.12, shows the typical behaviour of discontinuous functions near the point of discontinuity: the appearance of Gibbs oscillations. In this figure, we concentrate on $\tilde{A}_{xx}$ and $\tilde{\Gamma}^x$, but we present figures of all the other variables in appendix G. We see the pointwise error of $\tilde{A}_{xx}$ and $\tilde{\Gamma}^x$ with 4 accuracies at a similar time of $t \sim 0.2M$. Remember that the pointwise error is just the difference between the exact and numerical solution.

In contrast, for acc3, figure 9.13 shows the $\mathcal{L}^2$ norm over the entire domain and the $\mathcal{L}^2$ norm with the region close to the puncture excised (centre element) and the boundary $(L - 0.5)$ excised. We can see that the $\mathcal{L}^2$ norm of the wave equation is doing much better than all the other variables over the entire domain. However, we can see from the excised norms that the norm of the solution of the wave equation is comparable to the BSSN variables until the oscillations propagate outside the centre element.



(a) $\mathcal{L}^2$ norm for accuracy 3)

(b) $\mathcal{L}^2$ norm for accuracy 3)

Figure 9.13: Comparison of the logarithmic norm $\mathcal{L}^2$ over the entire region with the centre element $(-0.67, 0.67)$ and the boundary $L - 0.5$ excised of all the variables for acc 1 with $CFL = 0.5$ and $L = 2$.

If we look at the behaviour of the method in more detail for all the BSSN variables, we can make several comments and conclusions:

- For most variables, we can see that the method completely fails in the element containing the puncture and thereby the discontinuities. As the number of points increase near the puncture, we observe that oscillations arise and get worse. Even the variables that are not discontinuous initially show this phenomenon due to the presence of discontinuous functions in their respective evolution equations.

- Figures G.14(c), G.15(c) and G.15(e) also present some oscillations near the boundary. This is due to the fact that there are some integration by parts in these evolution equations. Remember that the boundary terms go to zero as the boundary is pushed further away. Since we are looking at a very small domain here, these boundary terms are no longer negligible and the solution is not as accurate at the boundaries.

- In figure G.16, we present the norm over the entire domain of all the BSSN variables, including the norm obtained for the solution of the 3D wave equation u for comparison purposes. We can see that for a small accuracy (acc 1), the norms of the BSSN variables are slighlty worse than that of the wave solution. However, as the number of points increase around the puncture, the accuracy gets worse for the BSSN variables and the difference with u can be clearly seen. This difference is again due to the increasing oscillations and overshooting of the method near the puncture.

- Figure G.17, is the same as figure G.16 but with the region near the puncture and the boundary excised. We can now see that the norms of the BSSN are now comparable to the one of the wave solution at least up to a certain time. What happens is that the oscillations created by the discontinuities tend to propagate across the domain and start appearing outside the excised region. The variables $\tilde{A}_{xy}$, K, $\tilde{\Gamma}^i$ have steep

gradients or discontinuities close to the puncture, so there is a need for more points in this region to catch the rapidly changing values of the variables, more so than for the other variables. However, when a very steep gradient is present, increasing the number of points tends to introduce oscillations and over-shoot of the coefficients with the SEM. This is why the norms for these specific variables are worse than for the other variables.

9.10 The offset mesh: The puncture on an edge or face of an element

In the FEM or SEM, discontinuities in the numerical solutions are often treated by designing a mesh so that the discontinuity is at an edge or face of an element. For example, in structural mechanics, cracks in the material introduce discontinuities in the solution. Those are best dealt with by modifying the mesh to cover cracks across elements rather than inside an element. Indeed, the theory of the FEM and SEM tells us that the test functions have to be $\mathcal{C}^0$ inside each element, but they are not required to be $\mathcal{C}^0$ across elements.

In light of this property, we design an ***offset mesh***, that will put the puncture at an edge or a face of an element, and study the effects on the numerical results. Remember that some of the variables are not defined at the puncture, therefore we have a strong constraint on the mesh: we cannot define a grid point at the puncture. This means that we cannot decompose our mesh into even elements in each direction, otherwise, there would be a grid point at the puncture, the corner of 4 elements. To obtain an offset mesh, we need an even number of elements in each space direction. The offset is given by half the minimum distance between 2 GLL points:

$$\text{offset} = \frac{\min(x_i - x_j)}{2}, \qquad i, j = 1, N_{GLL}. \tag{9.44}$$

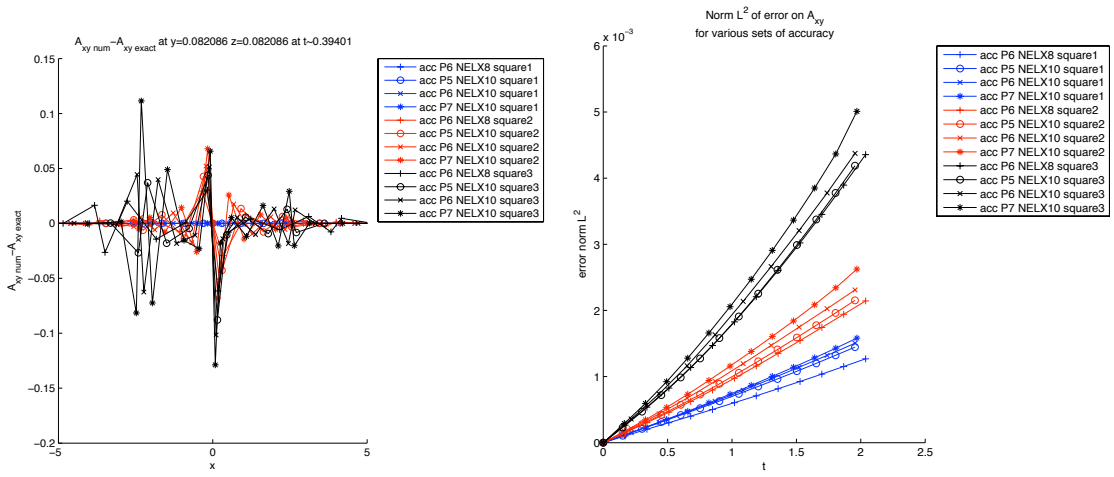
Since the GLL points are not evenly spaced, this means that the minimum distance will be between an anchor point (first GLL point) and the second GLL point, or the N-th GLL points and the $N+1$-th GLL points. It is very important to note that as the resolution increases, and also as the number of GLL points increases, the offset will get closer to zero. If the puncture has coordinates $(0,0,0)$, then:

1. ***Offset 1***: The centre of the mesh has coordinates $(\text{offset}, 0, 0)$. Therefore, the puncture sits on an edge of 3 elements. The computational domain will now be $[-L+\text{offset}, L+\text{offset}] \times [-L, L] \times [-L, L]$.

2. ***Offset 2***: The centre of the mesh has coordinates $(\text{offset}, \text{offset}, 0)$. Therefore, the puncture sits on a face of 2 elements. The computational domain will now be $[-L+\text{offset}, L+\text{offset}] \times [-L+\text{offset}, L+\text{offset}] \times [-L, L]$.

3. ***Offset 3***: The centre of the mesh has coordinates $(\text{offset}, \text{offset}, \text{offset})$. Therefore, the puncture sits inside one element. The computational domain will now be $[-L+\text{offset}, L+\text{offset}] \times [-L+\text{offset}, L+\text{offset}] \times [-L+\text{offset}, L+\text{offset}]$.

Note that $\tilde{A}_{ij}$, $\tilde{\Gamma}^i$ deserve closer scrutiny. We note that:



(a) Pointwise error of 3 offsets for $\tilde{A}_{xy}$

(b) Norm $\mathcal{L}^2$ of 3 offsets for $\tilde{A}_{xy}$

Figure 9.14: Comparison of 3 different types of offsets showing the pointwise error and L^2 norm with increasing accuracy for a domain $L = 64$ with a square mesh for the variable $\tilde{A}_{xy}$. In blue, the offset is 1, in red the offset is 2 and in black the offset is 3 (puncture inside the element in the negative values of x).

- Figure 9.14 shows the comparison of the 3 different types of offsets (offset 1 in blue, offset 2 in red and offset 3 in black) for the variable $\tilde{A}_{xy}$. The results for this variable are very indicative of what happens for all the other variables. We see the pointwise error and $\mathcal{L}^2$ norm with increasing accuracy for a domain $L = 64$ with a square mesh. Similar figures for the remainder of the variables are presented in Appendix G in G.18, G.19 and G.20.

- Note that offset 1 and 2 are much better than offset 3. In offset 3 the puncture is inside one element whereas for the other offsets the puncture is on a face or edge. Not only that, but the mesh with offset 3 results in significant oscillations that propagate in the whole domain. The method becomes unstable (see the $\mathcal{L}^2$ norms for offset 3 dramatically increasing after a certain amount of time).

These numerical experiments reveal that the overall error on the entire computational domain is affected by where the puncture is located.

We obtain much better results when discontinuities are placed between elements rather than inside an element. This would be fine for a stationary puncture, however when the puncture moves, the point of discontinuity will move, and eventually find itself inside an element. It could in principle be possible to adjust the mesh structure as the puncture moves, so that it is always located on an element face, but it would be preferable to be able to avoid this restriction. It is therefore more ideal, to know how to deal with a discontinuity inside an element with the spectral element method.

9.11 Increasing the number of elements

In some cases, increasing the resolution, and in particular the number of elements, near the area of irregularities, has proven to be a way of reducing the Gibbs phenomenon. In [154], numerical experiments were conducted on the inviscid Burgers equation with the SEM. By increasing the number of elements, the accuracy of the solution was improved significantly in terms of resolving the discontinuity more accurately ***as well as removing the oscillations without the use of a filter***. Fourier basis functions were used instead of the Lagrange-Legendre interpolants that we are using in this project. However, this type of behaviour is typical for ***all other*** spectral element discretizations tested in [154]: oscillations formed only for the fewer-element cases.

After numerous attempts, we have found that for $\tilde{A}_{xx}$, the oscillations in the puncture propagate beyond one element, even when the side of this element gets smaller and smaller. If this was going to work, we would probably need much smaller central elements, which would not be practical at all.

This is a useful result: it is not a very positive result, however, this is something important that we have learnt about the SEM and the BSSN system. Without filtering and with the puncture placed inside an element, we have been unable to simulate the BSSN system for any reasonable amount of time at any reasonable resolution with the SEM.

9.12 Filtering "as much or as little as needed"

Although the SEM applied to the BSSN system works very well sufficiently away from the puncture, we have seen that the discontinuous variables introduce Gibbs oscillations, eventually propagating across the whole domain, spoiling the accuracy of the numerical results.

In section 6.11, we have introduced a stabilization technique that can deal with discontinuous solutions: ***Filtering***.

If a solution $u(x)$ has a discontinuity in the computational domain, the Gibbs phenomenon appears with the SEM. This manifestation of the Gibbs phenomenon, or ultimately the lack of regularity, is a slow decay of the expansion coefficients. This suggests that we could attempt to modify the expansion coefficients to decay faster in the hope of recovering a more rapidly convergent expansion and more accurate approximations. In this light, we apply a ***filter*** to the modal coefficients of all the functions of the BSSN susceptible to irregularities or steep gradients near the puncture.

In the two books [137] and [138], one can find numerical examples with filtering techniques, both on the analytical and numerical levels. Numerical experiments show that using too low a filter order results in an overly dissipated solution. The characteristic of too strong a filter is the appearance of faceting[2] of the solution. We need to experiment numerically to know the right balance of the strength of the filter.

[2] In geometry, faceting is the process of removing parts of a polygon without creating any new vertices. In this context, it means that the numerical solution develops a staircase-like shape.

It is important to note that differentiation and filtering do ***not*** commute for many of the common spectral element filters, and in particular for the sharp cut-off filter we are using. Additional commutation error arises in addition to other numerical errors. This error does not appear to be significant however. In all our simulations, when applying filtering, we filter the modal coefficients after differentiation.

In this section, we have applied filtering in the centre element of the domain that contains the puncture. We use a sharp cut-off filter with a strong cut-off value $N_c = 1$. This means that for basis functions with a polynomial order N, all the modal values of the polynomial order with $N > N_c$ are set to zero. In the following figure, we present the behaviour of $\tilde{A}_{xy}$ initially and after $t = 2M$, without filtering and with filtering only discontinuous functions in the centre element:

- Figure 9.15, shows the pointwise error of $\tilde{A}_{xy}$ close to the puncture without filtering 9.15(a), 9.15(b), and with filtering 9.15(c), 9.15(d), for a very small domain of $L = 1$.

We also present similar results further away from the puncture for $\tilde{A}_{xy}$, and the same results obtained with $\tilde{A}_{xx}$ in Appendix G (see figures G.21, G.22, G.23 and G.24).

We wish to emphasize the different scale of the y-axis (pointwise error) on the plots with and without filtering.

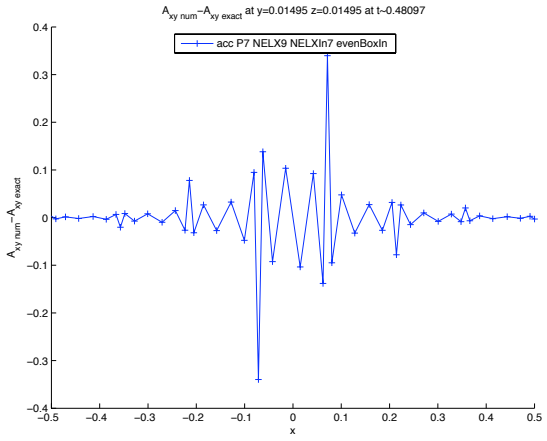
At $t \sim 0.5$M, the maximum error without filtering is roughly 20 times bigger than with filtering (0.4 versus 0.02). There is a significant difference in the maximum error which is placed in the centre element. If we look outside the centre element, the difference between the maximum error is even more impressive: the maximum pointwise error without filtering is 400 times bigger than with filtering (1.10^{-1} versus $2.5.10^{-3}$).

If we now look at the results at a later time of $t \sim 2.4$M, the results without filtering and with filtering show a clear drift. The maximum error in the centre element without filtering is roughly 25 times bigger than with filtering (1.2 versus 5.10^{-2}), and outside the centre element it is still roughly 400 times bigger (1 versus $2.5.10^{-3}$).

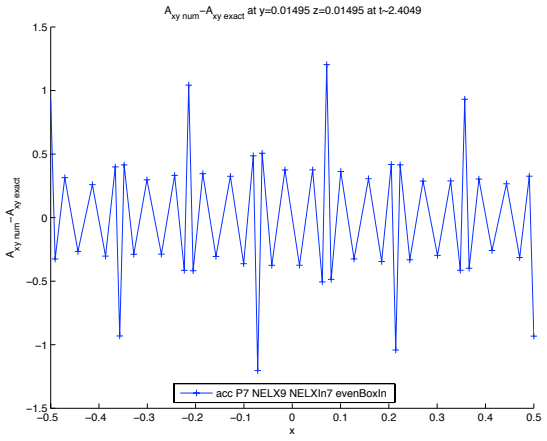
What happens with all the other variables in the coupled system?

- Figures 9.16 presents the results of a simulation of the coupled system for a small domain $L = 3$ up to $t = 3M$. In ***solid lines*** (Figure 9.16(a)), we see the $\mathcal{L}^2$ norms of the unfiltered variables, whereas in ***dashed dot lines*** (Figure 9.16(b)), we see the $\mathcal{L}^2$ norms of the filtered variables (only the centre element is filtered with $N_c = 1$). The filtered system shows more stability, as a matter of fact, the unfiltered system becomes completely unstable from $t \sim 2M$. Initially, the $\mathcal{L}^2$ norms of the unfiltered variables over the whole domain give better results than with the filtered variables. This is due to a loss of accuracy in the filtered element. However, this loss of accuracy is quickly compensated by the fact that the scheme is more stable, as there are no oscillations from the centre element propagating outwards as time advances.

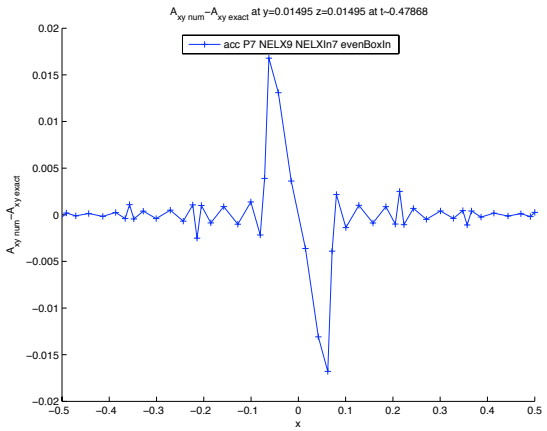
- Figure 9.17 presents the $\mathcal{L}^2$ norm of ϕ when varying the filter strength in the centre element. We set the cut off value of the filter to $N_c = 1, 2, ..., N$, where $N_c = N$ implies that there are no modal values affected and is equivalent to no filter. The $\mathcal{L}^2$ norms over the entire domain suggest that increasing the strength of the filter (low



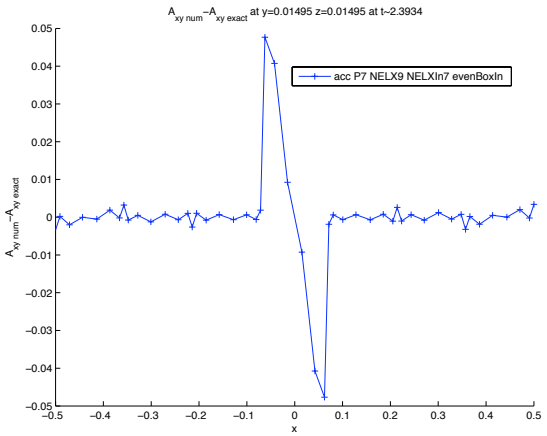
(a) $\tilde{A}_{xy}$ without filtering versus x at $t \sim 0.5$



(b) $\tilde{A}_{xy}$ without filtering versus x at $t \sim 2.4$

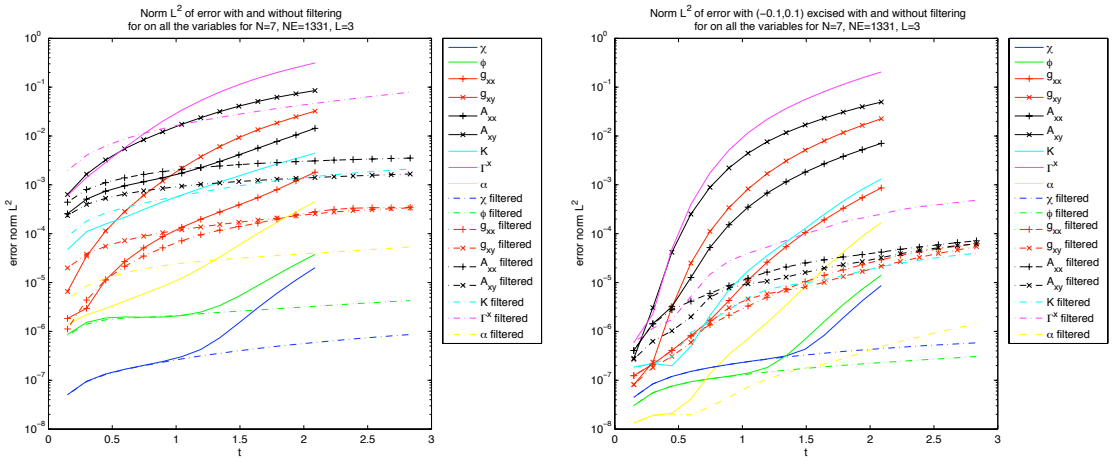


(c) $\tilde{A}_{xy}$ with filtering versus x at $t \sim 0.5$



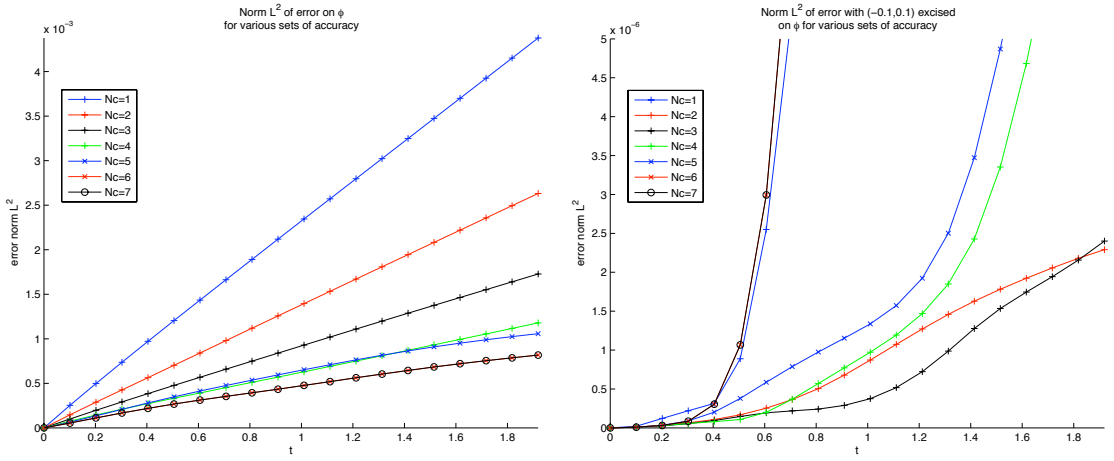
(d) $\tilde{A}_{xy}$ with filtering versus x at $t \sim 2.4$

Figure 9.15: Pointwise error of $\tilde{A}_{xy}$ close to the puncture (in a 2D slice for $y \sim z = 0.01$M), without filtering (a), (b) and with filtering (c), (d) for a very small domain of $L = 1$. Filtering makes a big difference in stopping the propagation of oscillations throughout the domain.



(a) $\mathcal{L}^2$ norms over the entire domain $L = 3$ (b) $\mathcal{L}^2$ norms with the centre element excised $L = 3$

Figure 9.16: The effect of filtering the centre element for a small domain $L = 3$ for most BSSN variables. In ***solid lines*** we see the $\mathcal{L}^2$ norms of the unfiltered variables, whereas in ***dashed dot lines*** we see the $\mathcal{L}^2$ norms of the filtered variables (only the centre element). The filtered system shows more stability.



(a) $\mathcal{L}^2$ norms over the entire domain $L = 2$, $N = 7$ and $N_E = 11$. (b) $\mathcal{L}^2$ norms with the centre element excised $L = 2$, $N = 7$ and $N_E = 11$.

Figure 9.17: Varying the strength of filtering at the centre element for ϕ with cut off values $N_c = 1, 2, 3, 4, 5, 6, N$ for a polynomial order $N = 7$ and for a small domain $L = 2$. Note that $N_c = N$ corresponds to no filtering.

values of N_c) results in less accuracy. However, when excising the values of the centre element, the $\mathcal{L}^2$ norm with the centre excised shows completely different results: for a very weak filter, the oscillations still appear and propagate, for a stronger filter, the results indicate more stable evolutions. Typically, for a polynomial order of $N = 7$, one should preferably set the cut-off value to $N_c \leq 3$ to eliminate the propagation of oscillations.

Filtering makes a big difference in stopping the propagation of oscillations throughout the domain. In fact, these numerical results seem to imply that filtering is essential to obtain stabilization for the application of the SEM in general relativity.

Although these results are obtained on a small domain, it is clear that filtering is one requirement for a working system. Since the filtering is required purely to deal with the discontinuities at the puncture, a large domain was not required for these tests.

9.13 Long-term stable evolutions?

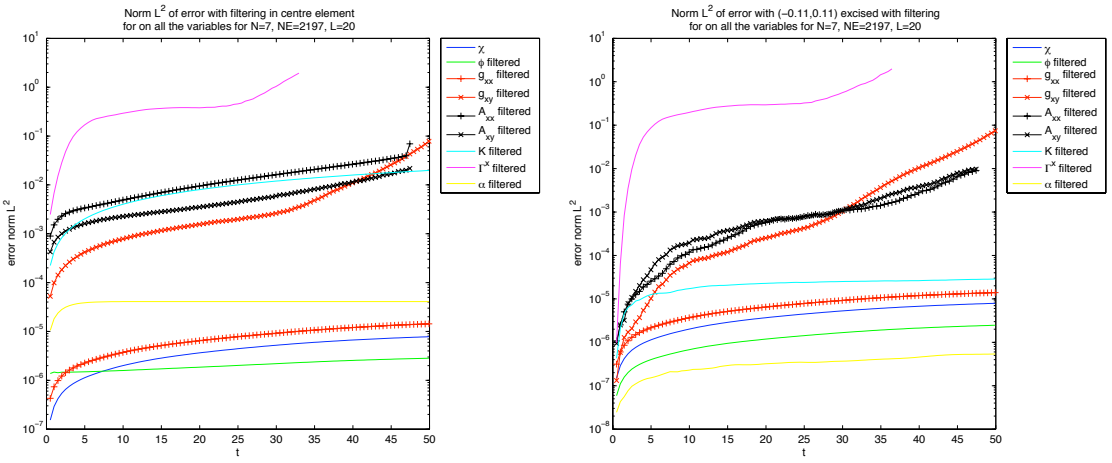
Long-term simulations are more computationally demanding for our SEM Matlab code. Remember that Matlab is an interpreted language and is therefore not as fast as other compiled languages. In terms of memory, the SEM applied to the BSSN system is quite efficient, since calculations of derivatives are executed on an elemental basis and assembled on the fly, we are not required to keep all these terms globally.

To minimize the computational time, we evolve each variable separately (uncoupled system) to investigate the long-term stability of the method in the best possible case. Obviously, if one variable in an uncoupled system leads to numerical instabilities, the full coupled system will suffer from the same fate.

The results in figure 9.18, show the $\mathcal{L}^2$ norm of each of the BSSN variables with the element containing the puncture filtered. We look at a domain of $L = 20$ and a simulation time of $t = 50$M. For a binary simulation, run times are typically for 1000s of M. So 50M is not considered a long time, however, recall that we have looked at one stationary Schwarzschild black hole only, so the runtime of 50M is sufficient to give some indication of the stability of each of the BSSN variables. Obviously the setup we are using here (distorted mesh) is not ideal and would not be used for binaries, and the mesh is the real key to efficiency with the SEM. The point of this test is thus to check whether all the evolution variables are stable.

Note that in these simulations, $\tilde{A}_{xx}$ and $\tilde{A}_{xy}$ are coupled, $\tilde{g}_{xy}$ is coupled to $\tilde{g}_{xx}$, and the norms are of the same order. However to show the effect of coupling we have plotted the norm of $\tilde{g}_{xx}$ uncoupled. When coupling variables, the errors are obviously worse but the behaviour is very similar.

The $\mathcal{L}^2$ norm of $\tilde{\Gamma}^x$ is by far the worst, because this mesh does not have enough resolution between 2M and 10M for this variable in particular. However, when using a higher resolution the norm is lower by roughly the rate we would expect, in particular, refer to the results with a cubic mesh presented in figure G.11.



(a) $\mathcal{L}^2$ norms over the entire domain

(b) $\mathcal{L}^2$ norms with the centre element excised

Figure 9.18: $\mathcal{L}^2$ norms of most variables of the BSSN system for a domain of $L = 20$ up to $t = 50$M, with a mesh ***softevenBoxIn***, $N = 7$, $N_E = 11^3$

None of these simulations were unstable, except the simulation for $\tilde{A}_{xx}$ and $\tilde{A}_{xy}$ which showed instabilities at $t = 47$M. With an outer boundary at 20M, however, this is not necessarily a bad sign, especially since the instabilities came from the under-resolved region between 2M and 10M and not the puncture. As a matter of fact, the filtered centre element containing the puncture shows stability for all the variables.

Although these results are very preliminary, they seem to indicate that the SEM has a good chance to be stable for long runs with filtering in the centre element. More experiments need to be done on a more ideal mesh for more efficient simulations. It is also possible that filtering more than the centre element might bring more stability especially for the quadratic nonlinear terms.

9.14 Conclusion

In this Chapter, we have studied the numerical results of a particular stationary solution using the ***Schwarzschild trumpet puncture data*** solution derived in [5, 6]. Not only is this "$1 + \log$" stationary trumpet solution extremely useful for testing a new code, but it also offers the possibility to test each equation separately, evolving only one variable at a time and keeping all the other variables exact. We have therefore worked on each variable as part of an uncoupled system as well as all the variables as part of a coupled system.

We have illustrated the relative ease in implementing simple distorted meshes more adapted to our problem, taking advantage of the geometric flexibility of the SEM. Although these meshes are not ideal, as an ideal mesh would mimic spherical coordinates at large distances, they allow for more points near the puncture and less points where we do not need as much resolution. These meshes allow enough variation in resolution for testing purposes.

We have shown experimentally that the numerical behaviour does not depend strongly on the weak form of the system. Experimental results also confirm that using the χ-method is strongly recommended when using the SEM.

When applying the SEM without the use of filtering, the method completely fails for most variables in the element containing the puncture and thereby the discontinuities. As the number of points increase near the puncture, we observe Gibbs oscillations that spread across the domain and eventually spoil the high accuracy. Indeed, the method is very powerful further away from the puncture, in the smooth parts of the solution. One can recover hp-convergence even in the case of distorted meshes.

Although we obtain much better results when discontinuities are placed between elements rather than inside an element, this is not a practical way to deal with moving punctures where the point of discontinuity moves. Without filtering, and with the puncture placed inside an element, we have been unable to simulate the BSSN system for a reasonable amount of time at any reasonable resolution with the SEM, even by increasing the number of elements near the puncture. This is not a positive result, but it is an important one, this also suggests that applying spectral methods with domain decomposition would be very difficult, since the oscillations near the puncture would be far worse than with the SEM as the polynomial order is typically more significant $N_{SM} \gg N_{SEM}$.

Filtering is significant in stopping the propagation of oscillations throughout the domain. In fact, our numerical results seem to imply that filtering is essential to obtain stabilization for the application of the SEM in general relativity. Although our long-term stability tests are preliminary, they seem to indicate that the SEM has a good chance to be stable for long runs with filtering in the centre element.

This is an important conclusion. It could have turned out that errors at the puncture propagate outwards, and that even filtering cannot cure the problem. There was a general feeling in the numerical relativity community that spectral-like methods may not be applicable to puncture evolutions, and the results here indicate that such evolution may indeed be quite possible, with the "***simple***" addition of some standard filtering methods.

Part III

Conclusion

"It is important to keep an open mind; just not so open that your brains fall out."

Albert Einstein (1879-1955)

10 Conclusion

In the first part of this thesis, we have addressed problems in General Relativity and Cosmology motivated by the simple key questions:

- How much information and how many constraints can one obtain from the Hubble flow in a FLRW universe?
- How general, precise, and useful, can results be under a minimum of theoretical assumptions?

In the second part of the work presented here, we have explored the use and benefits of the Spectral Element Method for Numerical Relativity.

- What is the potential of the Spectral Element Method for Numerical Relativity? How does this method work with the BSSN formulation and, with the method of moving punctures? Would this method allow for better accuracy and efficiency, and possibly contribute to gravitational wave detection?

Let us discuss and comment on the main results obtained from the above motivations.

10.1 General Relativity and Cosmology

In Cosmography, one keeps the geometry and symmetries of FLRW spacetime,

$$\mathrm{d}s^2 = -c^2\,\mathrm{d}t^2 + a(t)^2\left\{\frac{\mathrm{d}r^2}{1-k\,r^2} + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\right\}, \tag{10.1}$$

at least as a working hypothesis, but does not assume the Friedmann equations (Einstein equations), unless and until absolutely necessary.

Furthermore, it is quite common in cosmology to encounter physical quantities expanded as a Taylor series in the cosmological redshift z. Perhaps the most well-known exemplar of this phenomenon is the Hubble relation between distance and redshift. For instance, it is quite standard to phrase the investigation in terms of the luminosity distance versus redshift relation [20, 21]:

$$d_L(z) = \frac{c\,z}{H_0}\left\{1 + \frac{1}{2}\,[1-q_0]\,z + O(z^2)\right\}, \tag{10.2}$$

and its higher-order extension [22, 23, 24, 25]

$$d_L(z) = \frac{c\,z}{H_0}\left\{1+\frac{1}{2}\left[1-q_0\right]z+\frac{1}{6}\left[q_0+3q_0^2-(j_0+\Omega_0)\right]z^2 + \frac{1}{24}\left[2-2q_0-15q_0^2-15q_0^3+10q_0j_0+5j_0+s_0+2(1+3q_0)\Omega_0\right]z^3+O(z^4)\right\}. \tag{10.3}$$

However, we now have considerable high-z data available, for instance we have supernova data at least back to redshift $z \approx 1.75$. This opens up the theoretical question as to whether or not the Hubble series (or more generally any series expansion based on the z-redshift) actually converges for large redshift? Based on a combination of mathematical and physical reasoning, we have argued in Chapter 3, that the radius of convergence of any series expansion in z is less than or equal to 1, and that z-based expansions must break down for $z > 1$, corresponding to a universe less than half its current size.

Furthermore, we have argued on theoretical grounds for the utility of an improved parameterization $y = z/(1+z)$. In terms of the y-redshift we have shown that the radius of convergence of any series expansion in y is less than or equal to 1, so that y-based expansions are likely to be good all the way back to the big bang ($y = 1$), but that y-based expansions must break down for $y < -1$, now corresponding to a universe more than twice its current size. Our main conclusions of the latter are threefold:

- The use of the z-redshift for $z > 1$ is likely to lead to mathematical problems — specifically any Taylor series in z will be guaranteed to diverge for $z > 1$, and so finite truncations will be poor approximations to the underlying physical function. This is not all that early in the evolution of the universe — indeed many galaxies and supernovae are seen in the region $z \gtrsim 1$, so one ignores this issue at one's peril.

- The use of the y-redshift, where $y = z/(1+z)$, is very much to be encouraged for $z > 1$ (corresponding to $y > 1/2$). Taylor series in the y-redshift are likely to be well behaved all the way back to the big bang (corresponding to $y = 1$).

- By combining the notions of z-redshift, y-redshift, and the many reasonably standard notions of "cosmological distance" that have appeared in the literature, it is possible to extract ***many*** different versions of the Hubble law. Which version of the Hubble law one chooses to adopt for any specific purpose will depend on the specific question being addressed.

Is the expansion of the universe still accelerating in a Cosmographic framework? The "***big picture***" is best brought into focus by performing a global fit of all available supernova data to the Hubble relation, from the current epoch at least back to redshift $z \approx 1.75$. Indeed, all the discussion over acceleration versus deceleration, and the presence (or absence) of jerk (and snap) ultimately boils down, in a cosmographic setting, to doing a finite-polynomial truncated–Taylor series fit of the distance measurements (determined by supernovae and other means) to some suitable form of distance–redshift or distance–velocity relationship. Phrasing the question to be investigated in this way keeps it as close as possible to Hubble's

original statement of the problem, while minimizing the number of extraneous theoretical assumptions one is forced to adopt.

A central question thus has to do with the choice of the luminosity distance as the primary quantity of interest — there are several other notions of cosmological distance that can be used, some of which lead to simpler and more tractable versions of the Hubble relation. Why should the cosmology community be so fixated on using the luminosity distance d_L (or its logarithm, proportional to the distance modulus) and the redshift z as the relevant parameters? In principle, in place of luminosity distance $d_L(z)$ versus redshift z one could just as easily plot $f(d_L, z)$ versus $g(z)$, choosing $f(d_L, z)$ and $g(z)$ to be arbitrary locally invertible functions, and ***exactly the same physics*** would be encoded. Suitably choosing the quantities to be plotted and fit will not change the physics, ***but it might improve statistical properties and insight***.

By comparing cosmological parameters obtained using multiple different fits of the Hubble relation to different distance scales and different parameterizations of the redshift we have assessed, in Chapter 3, the robustness and reliability of the data fitting procedure. In performing this analysis we had hoped to verify the robustness of the Hubble relation, and to possibly obtain improved estimates of cosmological parameters such as the deceleration parameter and jerk parameter, thereby complementing other recent cosmographic and cosmokinetic analyses such as [12, 13, 14, 15, 16], as well as other analyses that take a sometimes skeptical view of the totality of the observational data [40, 41, 30, 42, 43]. The actual results of our current cosmographic fits to the data are considerably more ambiguous than we had initially expected, and there are many subtle issues hiding in the simple phrase "***fitting the data***".

There is a disturbingly strong model-dependence in the resulting estimates for the deceleration parameter. What happens when considering realistic estimates of systematic uncertainties (based on the published data)? Once realistic estimates of systematic uncertainties are budgeted for it becomes clear that purely statistical estimates of goodness of fit are dangerously misleading.

So, is the expansion of the universe still accelerating in this Cosmographic framework? While the "***preponderance of evidence***" certainly suggests an accelerating universe, we would argue that this conclusion is not currently supported "***beyond reasonable doubt***" — the supernova data (considered by itself) certainly ***suggests*** an accelerating universe, it is not sufficient to allow us to reliably conclude that the universe is accelerating. If one adds additional theoretical assumptions, such as by specifically fitting to a Λ-CDM model, the situation at first glance looks somewhat better — but this is then telling you as much about one's choice of theoretical model as it is about the observational situation.

Why do our conclusions seem to be so much at variance with currently perceived wisdom concerning the acceleration of the universe? The main reasons are twofold:

- Instead of simply picking a single model and fitting the data to it, we have tested the overall robustness of the scenario by encoding the same physics (H_0, q_0, j_0) in multiple different ways (d_L, d_F, d_P, d_Q, d_A; using both z and y) to test the robustness of the data fitting procedures.

- We have been much more explicit, and conservative, about the role of systematic uncertainties, and their effects on estimates of the cosmological parameters.

However, we are certainly not claiming that all is grim on the cosmological front — and do not wish our views to be misinterpreted in this regard — there are clearly parts of cosmology where there is plenty of high-quality data, and more coming in, constraining and helping refine our models. But regarding some specific cosmological questions the catch cry should still be "***Precision cosmology? Not just yet***" [62].

More recent data have now been released "***Union 07***" [155], and "***essence 09***" [156]. The combination of these datasets is referred to as the "Constitution dataset" [59]. Will these change our main conclusions? This remains to be seen.

In Chapter 4 we have extended and generalized the discussion of the original articles [63, 64, 65], and more recently of [71, 72, 73], to develop a number of rugged and general energy-condition-induced bounds on various cosmological parameters, bounds which have all taken the form

$$X(z) \gtrless X_{\rm bound} \equiv X_0 \; f(\Omega_0, z), \tag{10.4}$$

where $X(z)$ is some cosmological parameter, X_0 is its present-day value, and $f(\Omega_0, z)$ is some dimensionless function depending on the particular bound under consideration. The bounds we have considered can be derived by ***elementary*** means, and are typically expressed in terms of polynomial, rational, algebraic, and elementary functions — though in one particular instance we had to resort to hypergeometric functions. Several of these bounds are completely new [such as the explicit bounds on $H(z)$ and $\Omega(z)$, and the physically important Taylor series expansions for $\Omega_0 \approx 1$], several are significant extensions of previously known partial results, and all of these bounds are now valid for arbitrary spatial curvature. Additionally, since the analysis is now systematic and exhaustive, it is clear how the various energy conditions and their associated bounds are inter-related.

Furthermore, in the absence of any detailed understanding of the precise nature of the cosmological equation of state $\rho(p)$ it is useful to examine the question of just how much can be deduced with limited information. In the second part of Chapter 4, we have also worked in terms of the w-parameter $w(z) = p/\rho$, and we have used the idealized case of constant w_* as a "***template***" for comparison purposes with more realistic $w(z)$. Specifically:

- For constant w_* the explicit results for the density $\rho_{w_*}(z)$ and Hubble parameter $H_{w_*}(z)$ are well-known. The explicit result for the Ω parameter $\Omega_{w_*}(z)$ is less well-known, and the explicit results we have obtained for the angular diameter distance $d_{P_{w_*}}(z)$ and lookback time $T_{w_*}(z)$ appear to be both novel and significant.

- More importantly we have seen that these idealized results for constant w_* can be used as the basis for general comparison results that bound the various features of the Hubble flow in the following sense: If we know that $w(z) \in [w_-, w_+]$ between redshift zero and redshift z, then for monotonically evolving generic cosmological quantities $X(z)$ we have derived a number of rigorous bounds of the form

$$X_{w_\pm}(z) \leq X(z) \leq X_{w_\mp}(z), \tag{10.5}$$

  where we have explicitly seen that the direction of the inequality depends both on the precise details of the evolution of $X(z)$, and on the redshift range of interest.

All the bounds we have derived in Chapter 4 are thus both very general and very powerful.

Further developments and more realistic bounds could be achieved by looking at general linear combinations of w-matter. Consider that the density ρ is given by the linear combination

$$\rho = \sum_i \rho_i, \tag{10.6}$$

and that the pressure is given by the linear combination

$$p = \sum_i p_i, \tag{10.7}$$

with the equation of state $p_i = w_i \rho_i$ for each value of i. One could apply the same strategy discussed in Chapter 4 to the total linear combination of w-matter and in principle obtain even more realistic and tighter bounds.

10.2 Numerical Relativity

One of the main goals of numerical relativity is to provide very accurate templates of gravitational waves for ground-based and space-based interferometers. There are now robust and stable numerical methods for Numerical Relativity that work well, with finite differences and the moving punctures, and spectral methods with excision. Why the need for yet another numerical implementation? Current simulations are certainly good enough for ground-based detection, however, the scientific community is not yet sure for LISA, the space-based interferometer. There is also a computational difficulty for high mass ratio simulations; run timescales can be extremely long. An order-of-magnitude improvement in code efficiency would change the situation tremendously.

We have investigated the potential of the spectral element method applied to numerical relativity, and in particular, to the BSSN system. We have explored different options for a weak formulation, different mesh structures, and filtering options, and have highlighted the most useful directions to pursue.

First, we have presented an overview of the theory of the spectral element method. While the theory contains high levels of functional analysis and may be somewhat off-putting, this method has many successes in many different fields and offers great advantages over other numerical methods. The SEM combines the theory of spectral and pseudo-spectral methods for high order polynomials, and the variational formulation of finite elements, and the associated geometric flexibility.

The variational formulation is applied to the problem at hand and a weak formulation is then obtained. Space is divided into a number of elements, and the solution is written with local Lagrange–Legendre basis functions that are non-zero over a couple of elements. The spectral element discretization of the problem reduced to its weak form results in elemental matrix forms of the problem. After the assembly process, one can obtain a global system of

algebraic equations of the problem (typically sparse matrices for conforming elements). For explicit time stepping schemes, such as Runge–Kutta fourth order, there are no full matrices to invert as the Mass matrix is diagonal due to the choice of the GLL quadrature. This is a tremendous computational advantage.

We have applied the variational formulation to the BSSN system and presented several possible weak forms. From these weak forms, we have explained in detail how the elemental matrix forms specific to the BSSN system are calculated in light of the spectral element discretization.

The SEM is well-known for its geometric flexibility, we have illustrated the relative ease of designing a simple mesh specific to solving the BSSN system. Although these types of meshes have some disadvantages, providing high resolution near the coordinate planes, not just near the puncture, these meshes have the advantage of being very simple to implement, and allow enough variation in resolution for testing purposes. The "ideal" mesh, however, would mimic spherical coordinates at large distances.

When applying the SEM without the use of filtering, the method completely fails for most variables in the element containing the puncture and thereby the discontinuities. As the number of points increase near the puncture, we observe ***Gibbs oscillations*** that spread across the domain and eventually spoil the high accuracy.

Without filtering, and with the puncture placed inside an element, we have been unable to simulate the BSSN system for a reasonable amount of time at any reasonable resolution with the SEM, even by increasing the number of elements near the puncture. This is not a positive result, but it is an important one. On the other hand, the method is very powerful further away from the puncture, in the smooth parts of the solution. One can recover hp-convergence even in the case of distorted meshes.

Filtering is significant in stopping the propagation of oscillations throughout the domain. In fact, our numerical results seem to imply that filtering is essential to obtain stabilization for the application of the SEM in general relativity. Moreover, they seem to indicate that the SEM has a good chance to be stable for long runs with filtering in the centre element.

This is an important conclusion. It could have turned out that errors at the puncture propagate outwards, and that even filtering cannot cure the problem. There was a general feeling in the numerical relativity community that spectral-like methods may not be applicable to puncture evolutions, and the results here indicate that such evolution may indeed be quite possible, with the "***simple***" addition of some standard filtering methods.

We have briefly discussed a possible way of treating the boundary conditions, however, further work is needed to investigate the application of Sommerfeld-like boundary conditions to the BSSN system with the SEM. We could also look at more physically appropriate boundary conditions [150, 151, 152, 153].

The next obvious step is to implement and test non-stationary initial data and study the behaviour of the method in the case where the puncture will be moving in the grid. Shu and Wang [141] recovered spectral accuracy for the nonlinear Burgers equation where discontinuity develops and moves around the domain. This is a not a proof in itself, but

a good indication regarding the possibilities of the SEM with the BSSN system. Every problem is different. It is very difficult for complex problems to predict if the method will handle the moving discontinuities across the domain with the BSSN system until it is actually implemented and tested. Depending on the outcome, the next long-term step is, of course, the implementation of binary black holes with the SEM. Efforts should be invested in a mesh specifically designed to minimize the number of points needed. A similar method to the SEM would be the Mortar Element Method (MEM) which basically reduces to the SEM with domain decomposition that would allow for non-conformal elements in the mesh and henceforth mesh refinement implementation. Another possibility for BBH simulations would be to adopt the dual-coordinate method used by the Caltech and Cornell groups [157]: Changing the coordinate system at each time step, (or a number of time steps), so that the black holes effectively do not move in the numerical domain. Eventually one might want to move them a bit (for example, once an orbit, to capture the inspiral without having to warp the coordinates too much) but this would make for a significant computational saving in any spectral-like code.

The numerical analysis part of the SEM applied to the BSSN puncture method might be very useful in terms of bounding numerical errors (this is due in part to the finite element inheritance). Some important issues regarding the existence of solutions, and uniqueness of solutions could be explored analytically (as much as possible) and numerically. See ***Minamoto*** [131] for numerical methods using finite elements for proving the existence and uniqueness of numerical solutions of nonlinear hyperbolic problems.

A possible future development resides in the Gegenbauer reconstruction method, based on the Fourier or Gegenbauer series of a discontinuous but piecewise analytic function, to deal with discontinuities. This recently developed method removes the Gibbs phenomenon completely, and it is possible to obtain exponential accuracy in the maximum norm in any interval of analyticity.

Another possible development, would be to reformulate the BSSN systen using the Discontinuous Spectral Element Method using the Discontinuous Galerkin collocation method (DGSEM), instead of the Continuous Galerkin collocation method. A typical disadvantage of this method however is the fact that one needs to know very precisely the location of the discontinuity to implement this method. However, we always do know where the puncture is located. The puncture is tracked in a moving puncture code by noting that the speed of the puncture is given by $v^i = -\beta^i$, evaluated at the puncture. One can integrate up the velocity to give the puncture position. This works well, but would it give us the puncture location accurately enough for the DG method to work?

Finally, we have discussed the extremely good scalability of the SEM on parallel computers. This promises for efficient and fast computational simulations with the spectral element method. The SEM is not very sensitive to the speed of the network connecting different processors, which make this method highly suitable to run on clusters or grids of computers. A communication phase is required at each timestep for the assembly process. However, MPI communication tables that contain the sequence of messages that needs to be exchanged amongst the domains at each timestep need to be created only once and for

all when the mesh is built. Moreover, in the spectral multi-domain method, the $\mathcal{C}^0$ and $\mathcal{C}^1$ boundary conditions at the interface of the elements have to be enforced explicitly. In contrast, the spectral element method uses the variational principle to guarantee $\mathcal{C}^0$ and $\mathcal{C}^1$ (weak) continuity at the interface, which makes a parallel implementation more convenient. This is an important point, if conforming elements are used, the SEM can provide among the least computational overhead at processor boundaries.

The SEM is not just very interesting for binary black holes, but it would also be ideal for neutron stars, supernovae, and collapsing supernovae simulations because of the fact that it can easily handle complex geometries. In particular this would be a great advantage when dealing with different layers of discontinuous densities or pressures. In particular, we refer to the recent work by the Caltech-Cornell group [158], on using a mixture of finite-difference methods and spectral methods to deal with neutron stars and black hole binaries. Spectral methods are used in most places, but finite differences are used at the neutron star, because discontinuities prevent the use of the spectral method. SEM (or DGSEM) could potentially improve on this significantly.

10.3 Summary

In Chapter 3 we have discussed and presented results obtained in the context of Cosmography, that is without assuming the Einstein field equations, whereas in Chapter 4 we have derived powerful bounds in the context of Cosmodynamics, that is, assuming General Relativity.

In both frameworks, we considered how much information and how many constraints we could obtain from the Hubble flow in a FLRW universe. Indeed, the cosmological parameters contained in the Hubble relation between distance and redshift provide information on the behaviour of the universe (expansion, acceleration etc...).

In Cosmography, it is possible to concentrate more directly on the observational situation in a model-independent manner, because, it is possible to defer questions about the equation of state of the cosmological fluid, minimizing the number of theoretical assumptions one is bringing to the table. We have performed a number of inter-related cosmographic fits to supernova datasets, and paid particular attention to the extent to which the choice of distance scale and manner of representing the redshift scale affected the cosmological parameters.

In the context of Cosmodynamics, we have developed a number of rugged and general energy-condition-induced bounds on various cosmological parameters. We have also explored the extent to which a constraint on the w-parameter leads to useful and non-trivial constraints on the Hubble flow in terms of cosmological parameters $H(z)$, density $\rho(z)$, density parameter $\Omega(z)$, distance scales $d(z)$, and lookback time $T(z)$.

In the Numerical Relativity part of this thesis, we have explored the potential of a very recent and accurate numerical method, the Spectral Element Method (SEM), by treating a single Schwarszchild black hole evolution as a test case. The initial data we have implemented is a stationary solution called the ***Schwarzschild trumpet puncture data*** solution.

Spectral elements combine the theory of spectral and pseudo-spectral methods for high

order polynomials and the variational formulation of finite elements and the associated geometric flexibility. In Chapter 6 , we have summarized the theory of the SEM, and in Chapter 7, we have explained in details its practical implementation and the consequent numerical results for the wave equation formulated as a hyperbolic system in 1D and 3D.

In Chapter 8, we have formulated possible weak forms for the BSSN system and their corresponding spectral element discretization.

The accuracy of high order methods can deteriorate in the presence of discontinuities or sharp gradients. In Chapter 9, we have shown that we can treat the element that contain the puncture with a filtering method to avoid artificial and spurious oscillations (coming from discontinuous initial data from the BSSN system) forming and propagating into the domain .

Appendices

A

Some ambiguities in least-squares fitting

Let us suppose we have a function $f(x)$, and want to estimate $f(x)$ and its derivatives at zero via least squares. For any $g(x)$ we have a mathematical identity

$$f(x) = [f(x) - g(x)] + g(x), \tag{A.1}$$

and for the derivatives

$$f^{(m)}(0) = [f - g]^{(m)}(0) + g^{(m)}(0). \tag{A.2}$$

Adding and subtracting the same function $g(x)$ makes no difference to the underlying function $f(x)$, but it may modify the least squares estimate for that function. That is: Adding and subtracting a ***known*** function to the data ***does not commute*** with the process of performing a finite-polynomial least-squares fit. Indeed, let us approximate

$$[f(x) - g(x)] = \sum_{i=0}^{n} b_{f-g,i}\, x^i + \epsilon. \tag{A.3}$$

Then given a set of observations at points (f_I, x_I) we have (in the usual manner) the equations (for simplicity of the presentation all statistical uncertainties σ are set equal for now)

$$[f_I - g(x_I)] = \sum_{i=0}^{n} \hat{b}_{f-g,i}\, x_I^i + \epsilon_I, \tag{A.4}$$

where we want to minimize

$$\sum_I |\epsilon_I|^2. \tag{A.5}$$

This leads to

$$\sum_I [f_I - g(x_I)] x_I^j = \sum_{i=0}^{n} \hat{b}_{f-g,i} \sum_I x_I^{i+j}, \tag{A.6}$$

whence

$$\hat{b}_{f-g,i} = \left[\sum_I x_I^{i+j}\right]^{-1} \sum_I [f_I - g(x_I)] x_I^j, \tag{A.7}$$

where the square brackets now indicate an $(n+1)\times(n+1)$ matrix, and there is an implicit sum on the j index as per the Einstein summation convention. But we can re-write this as

$$\hat{b}_{f-g,i} = \hat{b}_{f,i} - \left[\sum_I x_I^{i+j}\right]^{-1} \sum_I [g(x_I)]x_I^j, \tag{A.8}$$

relating the least-squares estimates of $b_{f,i}$ and $b_{f-g,i}$. Note that by construction $i \leq n$. If we now use this to estimate $f^{(i)}(0)$, we see:

$$\hat{f}^{(i)}_{[f-g]+g}(0) = \hat{f}^{(i)}_{f-g}(0) + g^{(i)}(0), \tag{A.9}$$

whence

$$\hat{f}^{(i)}_{[f-g]+g}(0) = \hat{f}^{(i)}(0) - i!\left[\sum_I x_I^{i+j}\right]^{-1} \sum_I [g(x_I)]x_I^j + g^{(i)}(0), \tag{A.10}$$

where $\hat{f}^{(i)}(0)$ is the "naive" estimate of $f^{(i)}(0)$ obtained by simply fitting a polynomial to f itself, and $\hat{f}^{(i)}_{[f-g]+g}(0)$ is the "improved" estimate obtained by first subtracting $g(x)$, fitting $f(x)-g(x)$ to a polynomial, and then adding $g(x)$ back again. Note the formula for the shift of the estimate of the ith derivative of $f(x)$ is linear in the function $g(x)$ and its derivatives. In general this is the most precise statement we can make — the process of finding a truncated Taylor series simply does not commute with the process of performing a least squares fit.

We can gain some additional insight if we use Taylor's theorem to write

$$g(x) = \sum_{k=0}^{\infty} \frac{g^{(k)}(0)}{k!}x^k = \sum_{k=0}^{n} \frac{g^{(k)}(0)}{k!}x^k + \sum_{k=n+1}^{\infty} \frac{g^{(k)}(0)}{k!}x^k, \tag{A.11}$$

where we temporarily suspend concerns regarding convergence of the Taylor series. Then

$$\begin{aligned}\hat{f}^{(i)}_{[f-g]+g}(0) &= \hat{f}^{(i)}(0) + g^{(i)}(0) \\ &\quad -i!\left[\sum_I x_I^{i+j}\right]^{-1} \sum_I \left\{\sum_{k=0}^{n} \frac{g^{(k)}(0)}{j!}x_I^k + \sum_{k=n+1}^{\infty} \frac{g^{(k)}(0)}{j!}x_I^k\right\} x_I^j.\end{aligned} \tag{A.12}$$

So

$$\begin{aligned}\hat{f}^{(i)}_{[f-g]+g}(0) &= \hat{f}^{(i)}(0) + g^{(i)}(0) \\ &\quad -i!\left[\sum_I x_I^{i+j}\right]^{-1} \left\{\sum_{k=0}^{n} \frac{g^{(k)}(0)}{k!}\sum_I x_I^{j+k} + \sum_{k=n+1}^{\infty} \frac{g^{(k)}(0)}{k!}\sum_I x_I^{j+k}\right\},\end{aligned} \tag{A.13}$$

whence

$$\begin{aligned}\hat{f}^{(i)}_{[f-g]+g}(0) &= \hat{f}^{(i)}(0) + g^{(i)}(0) - i!\sum_{k=0}^{n} \frac{g^{(k)}(0)}{k!}\left[\sum_I x_I^{i+j}\right]^{-1}\left[\sum_I x_I^{j+k}\right] \\ &\quad -i!\sum_{k=n+1}^{\infty} \frac{g^{(k)}(0)}{k!}\left[\sum_I x_I^{i+j}\right]^{-1}\sum_I x_I^{j+k}.\end{aligned} \tag{A.14}$$

But two of these matrices are simply inverses of each other, so in terms of the Kronecker delta

$$\begin{aligned}\hat{f}^{(i)}_{[f-g]+g}(0) &= \hat{f}^{(i)}(0) + g^{(i)}(0) - i!\sum_{k=0}^{n}\frac{g^{(k)}(0)}{k!}\delta_{ik} \\ &\quad -i!\sum_{k=n+1}^{\infty}\frac{g^{(k)}(0)}{k!}\left[\sum_I x_I^{i+j}\right]^{-1}\sum_I x_I^{j+k},\end{aligned} \tag{A.15}$$

which now leads to significant cancellations

$$\hat{f}^{(i)}_{[f-g]+g}(0) = \hat{f}^{(i)}(0) - i!\sum_{k=n+1}^{\infty}\frac{g^{(k)}(0)}{k!}\left[\sum_I x_I^{i+j}\right]^{-1}\sum_I x_I^{j+k}. \tag{A.16}$$

This is the best (ignoring convergence issues) that one can do in the general case. Note the formula for the shift of the estimate of the ith derivative of $f(x)$ is linear in the derivatives of the function $g(x)$, and that it starts with the $(n+1)$th derivative. Consequently as the order n of the polynomial used to fit the data increases there are fewer terms included in the sum, so the difference between various estimates of the derivatives becomes smaller as more terms are added to the least squares fit.

In the particular situation we discuss in the body of the thesis

$$f(x) \to \tilde{\mu} = \ln\left(\frac{d(z)}{z\ \mathrm{Mpc}}\right); \quad g(x) \to \frac{K}{2}\ln(1+z); \quad K \in Z; \tag{A.17}$$

or a similar formula in terms of the y-redshift. Consequently, from equation (A.10), particularized to our case

$$\hat{\tilde{\mu}}^{(i)}_K(0) = \hat{\tilde{\mu}}^{(i)}(0) + \frac{K}{2}[\ln(1+z)]^{(i)}(0) - \frac{K\ i!}{2}\left[\sum_I z_I^{i+j}\right]^{-1}\left[\sum_I z_I^j\ \ln(1+z_I)\right]. \tag{A.18}$$

Then the "gap" between any two adjacent estimates for $\hat{\tilde{\mu}}^{(i)}_K(0)$ corresponds to taking $\Delta K = 1$ and so

$$\Delta\hat{\tilde{\mu}}^{(i)}(0) = \frac{(-1)^{i-1}\ (i-1)!}{2} - \frac{i!}{2}\left[\sum_I z_I^{i+j}\right]^{-1}\left[\sum_I z_I^j\ \ln(1+z_I)\right]. \tag{A.19}$$

But then for the particular case $i = 1$ which is of most interest to us

$$\hat{\tilde{\mu}}^{(1)}_K(0) = \hat{\tilde{\mu}}^{(1)}(0) + \frac{K}{2} - \frac{K}{2}\left[\sum_I z_I^{i+j}\right]^{-1}_{1j}\left[\sum_I z_I^j\ \ln(1+z_I)\right], \tag{A.20}$$

and

$$\Delta\hat{\tilde{\mu}}^{(1)}(0) = \frac{1}{2} - \frac{1}{2}\left[\sum_I z_I^{i+j}\right]^{-1}_{ij}\left[\sum_I z_I^j\ \ln(1+z_I)\right]. \tag{A.21}$$

By Taylor series expanding the logarithm, and reindexing the terms, this can also be recast as

$$\hat{\tilde{\mu}}_K^{(i)}(0) = \hat{\tilde{\mu}}^{(i)}(0) + \frac{K\ i!}{2} \sum_{k=n+1}^{\infty} \frac{(-1)^k}{k} \left[\sum_I z_I^{i+j}\right]^{-1} \sum_I z_I^{j+k}, \tag{A.22}$$

whence

$$\hat{\tilde{\mu}}_K^{(1)}(0) = \hat{\tilde{\mu}}^{(1)}(0) + \frac{K}{2} \sum_{k=n+1}^{\infty} \frac{(-1)^k}{k} \left[\sum_I z_I^{i+j}\right]^{-1}_{1j} \sum_I z_I^{j+k}, \tag{A.23}$$

and

$$\Delta\hat{\tilde{\mu}}^{(1)}(0) = \frac{1}{2} \sum_{k=n+1}^{\infty} \frac{(-1)^k}{k} \left[\sum_I z_I^{i+j}\right]^{-1}_{1j} \sum_I z_I^{j+k}, \tag{A.24}$$

(Because of convergence issues, if we work with z-redshift these last three formulae make sense only for supernovae datasets where we restrict ourselves to $z_I < 1$, working in y-redshift no such constraint need be imposed.) Now relating this to the modelling ambiguity in q_0, we have

$$[\Delta q_0]_{\text{modelling}} = -2\ \Delta\hat{\tilde{\mu}}^{(1)}(0), \tag{A.25}$$

so that

$$[\Delta q_0]_{\text{modelling}} = -1 + \left[\sum_I z_I^{i+j}\right]^{-1}_{1j} \left[\sum_I z_I^j\ \ln(1+z_I)\right]. \tag{A.26}$$

By Taylor-series expanding the logarithm, modulo convergence issues discussed above, this can also be expressed as:

$$[\Delta q_0]_{\text{modelling}} = -\sum_{k=n+1}^{\infty} \frac{(-1)^k}{k} \left[\sum_I z_I^{i+j}\right]^{-1}_{1j} \left[\sum_I z_I^{j+k}\right]. \tag{A.27}$$

In particular, without further calculation, these results collectively tell us that the different estimates for q_0 will always be evenly spaced, and it suggests that as $n \to \infty$ the differences will become smaller. This is actually what is seen in the data analysis we performed. ***If we were to have a good physics reason for choosing one particular definition of distance as being primary, we would use that for the least squares fit, and the other ways of estimating the derivatives would be "biased" — but in the current situation we have no physically preferred "best" choice of distance variable.***

B

Combining measurements from different models

Suppose one has a collection of measurements X_a, each of which is represented by a random variable $\hat{X}_a$ with mean $\mu_a = E(\hat{X}_a)$ and variance $\sigma_a^2 = E([\hat{X}_a - \mu_a]^2)$. How should one then combine these measurements into an overall "best estimate"?

If we have no good physics reason to reject one of the measurements then the best we can do is to describe the combined measurement process by a random variable $\hat{X}_{\hat{A}}$ where $\hat{A}$ is now a discrete random variable that picks one of the measurement techniques with some probability p_a. More precisely

$$\mathrm{Prob}(\hat{A} = a) = p_a, \tag{B.1}$$

where the values p_a are for now left arbitrary. Then

$$\mu = E(\hat{X}_{\hat{A}}) = \sum_a p_a \; E(\hat{X}_a) = \sum_a p_a \; \mu_a, \tag{B.2}$$

and

$$E(\hat{X}_{\hat{A}}^2) = \sum_a p_a \; E(\hat{X}_a)^2 = \sum_a p_a \; (\sigma_a^2 + \mu_a^2). \tag{B.3}$$

But equally well

$$E(\hat{X}_{\hat{A}}^2) = \sigma^2 + \mu^2, \tag{B.4}$$

so that overall

$$\mu = \sum_a p_a \; \mu_a, \tag{B.5}$$

and

$$\sigma^2 = \sum_a p_a \; \sigma_a^2 + \sum_a p_a \; (\mu_a - \mu)^2. \tag{B.6}$$

This lets us split the overall variance into the contribution from the purely statistical uncertanties on the individual measurements

$$\sigma_{\mathrm{statistical}} = \sqrt{\sum_a p_a \; \sigma_a^2}, \tag{B.7}$$

plus the "modelling ambiguity" arising from different ways of modelling the same physics

$$\sigma_{\mathrm{modelling}} = \sqrt{\sum_a p_a \; (\mu_a - \mu)^2}. \tag{B.8}$$

In the particular case we are interested in we have 5 different ways of modelling distance and no particular reason for choosing one definition of measurement over all the others so it is best to take $p_a = 1/5$.

Furthermore in the case of the estimates for the deceleration parameter, all individual estimates have the same statistical uncertainty, and the estimates are equally spaced with a gap Δ:

$$\sigma_a = \sigma_0; \qquad \mu_a = \mu_P + n\Delta; \quad n \in \{-2, -1, 0, 1, 2\}. \tag{B.9}$$

Therefore

$$\mu = \mu_P; \qquad \sigma_{\text{statistical}} = \sigma_0; \qquad \sigma_{\text{modelling}} = \sqrt{2}\,\Delta. \tag{B.10}$$

For estimates of the jerk, we no longer have the simple equal-spacing rule and equal statistical uncertainties rule, but there is still no good reason for preferring one distance surrogate over all the others so we still take $p_a = 1/5$ and the estimate obtained from the combined measurements satisfies

$$\mu = \frac{\sum_a \mu_a}{5}; \qquad \sigma_{\text{statistical}} = \sqrt{\frac{\sum_a \sigma_a^2}{5}}; \qquad \sigma_{\text{modelling}} = \sqrt{\frac{\sum_a (\mu_a - \mu)^2}{5}}. \tag{B.11}$$

These formulae are used to calculate the statistical and modelling uncertainties reported in tables 3.5–3.6 and 3.7–3.8 . Note that ***by definition*** the combined purely statistical and modelling uncertainties are to be added in quadrature

$$\sigma = \sqrt{\sigma_{\text{statistical}}^2 + \sigma_{\text{modelling}}^2}. \tag{B.12}$$

This discussion does not yet deal with the estimated systematic uncertainties ("known unknowns") or "historically estimated" systematic uncertainties ("unknown unknowns").

C Useful inequalities

The following inequalities are crucial tools for the analysis of variational problems, spectral element methods and also domain decomposition methods.

C.1 Cauchy-Schwarz inequality

Lemma 1 *Let $\mathcal{V}$ be a Hilbert space, with the inner product $(\cdot,\cdot)_{\mathcal{V}}$ and the norm $\|\cdot\|_{\mathcal{V}}$, then we have the Cauchy-Schwarz inequality,*

$$(u,v)_{\mathcal{V}} \leqslant \|u\|_V \, \|v\|_{\mathcal{V}}, \quad u,\ v \in \mathcal{V}. \tag{C.1}$$

C.2 Poincaré inequality:

Lemma 2 *Let $u \in \mathcal{H}^1(\mathbf{\Omega})$, then there exist constants, depending only on $\mathbf{\Omega}$, such that,*

$$\|u\|^2_{\mathcal{L}^2(\mathbf{\Omega})} \leqslant c_1 \, |u|^2_{\mathcal{H}^1(\mathbf{\Omega})} + c_2 \left(\int_{\mathbf{\Omega}} u \, d\vec{x} \right)^2. \tag{C.2}$$

Theorem 4 *Let $\mathbf{\Omega} \subset \mathbb{R}^n$ be a bounded Lipshitz domain and $\mathcal{V}$ be a closed subspace of $\mathcal{H}^1(\mathbf{\Omega})$ that contains $\mathbb{P}_0(\mathbf{\Omega})$, the space of constant functions on $\mathbf{\Omega}$. Let $\mathcal{W}$ be a Hilbert space with a norm $\|\cdot\|_{\mathcal{W}}$ and let $A:\ \mathcal{V} \to \mathcal{W}$ be a bounded linear operator, such that*

$$Av = 0, \quad v \in \mathbb{P}_0(\mathbf{\Omega}). \tag{C.3}$$

If

$$\|Au\|_{\mathcal{W}} \leqslant \|A\| \, \|u\|_{\mathcal{H}^1(\mathbf{\Omega})}, \quad u \in \mathcal{W}, \tag{C.4}$$

then

$$\|Au\|_{\mathcal{W}} \leqslant \|A\| \, c_{\mathbf{\Omega}} \ |u|_{\mathcal{H}^1(\mathbf{\Omega})}, \quad u \in \mathcal{W}, \tag{C.5}$$

where $c_{\mathbf{\Omega}}$ depends only on the domain $\mathbf{\Omega}$, but is independent of u, A and of the spaces $\mathcal{V}$ and $\mathcal{W}$.

In other words, it means that if the left hand side of the inequality does not change if one adds a constant to u, then the norm on the right hand side can be replaced by the semi-norm.

C.3 Friedrichs inequality

Lemma 3 *Let $\Gamma \subseteq \partial\mathbf{\Omega}$ have non-vanishing $(n-1)$-dimensional measure. Then there exists constants, depending only on $\mathbf{\Omega}$ and Γ, such that, for $u \in \mathcal{H}^1(\mathbf{\Omega})$,*

$$\|u\|^2_{\mathcal{L}^2(\mathbf{\Omega})} \leqslant c_1 \, |u|^2_{\mathcal{H}^1(\mathbf{\Omega})} + c_2 \|u\|^2_{\mathcal{L}^2(\Gamma)}. \tag{C.6}$$

In particular, if u vanishes on Γ,

$$\|u\|^2_{\mathcal{L}^2(\mathbf{\Omega})} \leqslant c_1 \, |u|^2_{\mathcal{H}^1(\mathbf{\Omega})} \,, \tag{C.7}$$

and therefore,

$$|u|^2_{\mathcal{H}^1(\mathbf{\Omega})} \leqslant \|u\|^2_{\mathcal{H}^1(\mathbf{\Omega})} \leqslant (c_1 + 1) \, |u|^2_{\mathcal{H}^1(\mathbf{\Omega})} \,. \tag{C.8}$$

D

General cardinal functions, Lagrange basis

Lagrange interpolation is a method of interpolation that introduces a family of Nth-degree Lagrange polynomials for a set of interpolation points x_i defined by the requirement that

$$h_{N,i}(x_j) \quad = \quad \delta_{ij} \qquad i,\ j \ \in \{0..N\}, \tag{D.1}$$

where δ_{ij} is the usual Kronecker delta symbol. The Lagrange interpolant is also defined below:

$$h_{N,i}(x_j) \quad = \quad \frac{(x-x_0)...(x-x_{i-1})(x-x_{i+1})...(x-x_N)}{(x_i-x_0)...(x_i-x_{i-1})(x_i-x_{i+1})...(x_i-x_N)}, \tag{D.2}$$

Note that the factor $(x - x_i)$ does not appear in the numerator which is a polynomial of degree N and the factor $(x_i - x_i)$ is missing in the denominator which is a constant.

An alternative approach is to define a Lagrange generating polynomial of degree $(N+1)$

$$\phi_{N+1}(x) \quad = \quad (x-x_0)(x-x_1)...(x-x_{N-1})(x-x_N). \tag{D.3}$$

The generating polynomial ϕ_{N+1} is zero at all the data points, that is

$$\phi_{N+1}(x_i) \quad = \quad 0 \qquad i \ \in \ \{0, N\}. \tag{D.4}$$

Furthermore, the derivative with respect to x at the collocation points x_i gives

$$\begin{aligned} & i \ \in \ \{0, N\} \\ \phi'_{N+1}(x_i) \quad &= \quad (x_i-x_0)...(x_i-x_{i-1})(x_i-x_{i+1})...(x_i-x_N). \end{aligned} \tag{D.5}$$

From the previous two definitions we can define the Lagrange interpolants in an alternative manner to equation (D.2) by

$$h_{N,i}(x) \quad = \quad \frac{\phi_{N+1}(x)}{\phi'_{N+1}(x_i)(x-x_i)}. \tag{D.6}$$

The term

$$c_i \quad = \quad \frac{1}{\phi'_{N+1}(x_i)} \tag{D.7}$$

is referred to as the barycentric weights.

The Lagrange interpolants are also called cardinal functions, cardinal basis or Lagrange basis. The Lagrange generating polynomial can be identified to (or some combination of) the Legendre, Lobatto, Chebychev, Hermite etc... polynomials.

Some desired interpolating polynomial can be expressed in terms of the Lagrange basis as

$$P_N(x) = \sum_{i=0}^{i=N} p_i h_{N,i}(x). \quad \text{D.8}$$

The cardinal functions in the SEM are Lagrange basis with the following Lagrange generating polynomials:

$$\phi_{N+1}(x) = -\frac{1}{c_{N-1}}(1-x^2)Lo_{N-1}(x), \quad \text{D.9}$$

where $Lo_N(x)$ are Lobatto polynomials and c_{N-1} are the highest power of the $N-1$-degree Lobatto polynomial. Since there is a relationship between Lobatto polynomials and Legendre polynomials

$$Lo_N(x) = L'_{N+1}(x), \quad \text{D.10}$$

the Lagrange generating polynomial can be written as

$$\phi_{N+1}(x) = -\frac{1}{c_N}(1-x^2)L'_N(x). \quad \text{D.11}$$

Substituting the spectral element Lagrange generating polynomial into the Lagrange basis definition, we obtain the cardinal basis functions for the spectral element method:

$$h_{N,i}(x) = -\frac{(1-x^2)L'_N(x)}{N(N+1)L_N(x_i)(x-x_i)}. \quad \text{D.12}$$

To derive equation (D.12), we have used one of the property of the Legendre polynomials described in Appendix E:

$$\Big(\big(1-x^2\big)\,L'_N(x)\Big)' = -N(N+1)L_N(x). \quad \text{D.13}$$

The grid points are

$$x_0 = -1 \quad x_N = 1 \quad \text{and the } (N-1) \text{ roots of } L'_N(x). \quad \text{D.14}$$

The quadrature weights are

$$\rho_i = \frac{2}{N(N+1)\Big(L_N(x_i)\Big)^2}. \quad \text{D.15}$$

Figure D.1 illustrates Lagrange-Legendre polynomials of order $P = 8$ in a $1D$ master element $\Lambda = [-1, 1]$.

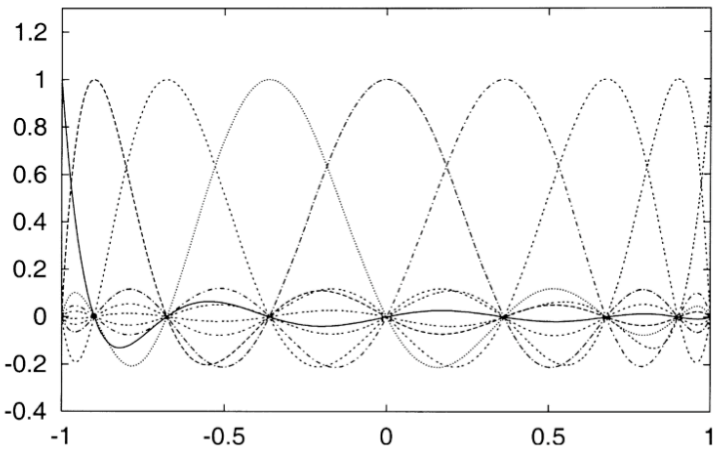


Figure D.1: Lagrange–Legendre interpolants of degree $P = 8$ at the Gauss–Lobatto–Legendre points on the reference segment $\Lambda = [-1, 1]$. The $N = P + 1 = 9$ GLL points can be distinguished along the horizontal axis. All Lagrange type polynomials are by definition 0 or 1 at each of these GLL points.

D.1 First derivative and the first node differentiation matrix H for Lagrange basis

We differentiate the desired interpolating polynomial expressed in terms of the Lagrange basis by

$$\frac{dP_N(x)}{dx} = \sum_{i=0}^{i=N} p_i \frac{dh_{N,i}(x)}{dx}. \tag{D.16}$$

Using the definition of $h_{N,i}$ with the Lagrange generating polynomial in equation (D.6), we obtain,

$$\frac{dh_{N,i}(x)}{dx} = \frac{\phi'_{N+1}(x)(x - x_i) - \phi_{N+1}(x)}{\phi'_{N+1}(x_i)(x - x_i)^2}. \tag{D.17}$$

We can then evaluate the right hand side at the interpolation nodes x_i with

$$d_x h_j(x_i) = H_{ij} = \begin{cases} H_{ij} = \dfrac{\phi'_{N+1}(x_i)}{\phi'_{N+1}(x_j)\,(x_i - x_j)} & i \neq j \\ H_{ii} = \dfrac{\phi''_{N+1}(x_i)}{2\phi'_{N+1}(x_i)} & i \in \{0, N\} \end{cases} \tag{D.18}$$

To obtain the first node differentiation matrix H we have used the property that $\phi_{N+1}(x_i) = 0$ at the nodes and in the case $i = j$ we have used Taylor series expansion around $x \to x_i$:

$$H_{ii} = \left(\frac{dh_i(x)}{dx}\right)_{x=x_i} = \frac{\phi''_{N+1}(x_i)}{2\phi'_{N+1}(x_i)}. \tag{D.19}$$

D.1.1 First derivative and the first node differentiation matrix H for SEM basis

We further substitute the Lagrange generating polynomial by the Legendre interpolants and obtain the first node differentiation matrix H:

$$\mathrm{d}_\xi h_j(\xi_i) = H_{ij} = \begin{cases} H_{00} = -H_{NN} = -\frac{N(N+1)}{4} & \\ H_{ii} = 0 & i \in \{1, N-1\} \\ H_{ij} = \frac{L_N(\xi_i)}{L_N(\xi_j)\,(\xi_i - \xi_j)} & i \neq j \end{cases} \tag{D.20}$$

D.2 Second derivative and the second node differentiation matrix W for Lagrange basis

We differentiate twice the desired interpolating polynomial expressed in terms of the Lagrange basis by

$$\frac{d^2 P_N(x)}{dx^2} = \sum_{i=0}^{i=N} p_i \frac{d^2 h_{N,i}(x)}{dx^2}. \tag{D.21}$$

Using the definition of $h_{N,i}$ with the Lagrange generating polynomial in equation (D.6), we obtain,

$$\frac{d^2 h_{N,i}(x)}{dx^2} = \frac{\phi''_{N+1}(x)(x-x_i)^2 - 2\phi'_{N+1}(x)(x-x_i) + 2\phi_{N+1}(x)}{\phi'_{N+1}(x_i)(x-x_i)^3}. \tag{D.22}$$

We can then evaluate the right hand side at the interpolation nodes x_i with

$$d_{xx} h_j(x_i) = W_{ij} = \begin{cases} W_{ij} = \dfrac{\phi''_{N+1}(x_i)(x_i - x_j) - 2\phi'_{N+1}(x_i)}{\phi'_{N+1}(x_j)\,(x_i - x_j)^2} & i \neq j \\ W_{ii} = \dfrac{\phi'''_{N+1}(x_i)}{3\phi'_{N+1}(x_i)} & i \in \{0, N\} \end{cases} \tag{D.23}$$

To obtain the second node differentiation matrix W we have used the property that $\phi_{N+1}(x_i) = 0$ at the nodes, and in the case $i = j$ we have used Taylor series expansions around $x \to x_i$ for $\phi_{N+1}(x)$, $\phi'_{N+1}(x)$ and $\phi''_{N+1}(x)$.

D.2.1 Second derivative and the second node differentiation matrix W for SEM basis

We further substitute the Lagrange generating polynomial by the Legendre interpolants and obtain the second node differentiation matrix W:

$$\mathrm{d}_{\xi\xi} h_j(\xi_i) = W_{ij} = \begin{cases} W_{00} = \frac{(-1)^N}{3} L''_N(-1) & \\ W_{NN} = \frac{1}{3} L''_N(1) & \\ W_{ii} = \frac{1}{3} \frac{L''_N(\xi_i)}{L_N(\xi_i)} & i \in \{1, N-1\} \\ W_{ij} = -2 \frac{L_N(\xi_i)}{L_N(\xi_j)\,(\xi_i - \xi_j)^2} & i \neq j \end{cases} \tag{D.24}$$

E

Legendre polynomial properties

In this section we present properties of the family of the Legendre polynomials $L_N(x)$ that have many applications. They can be used as an alternative to Chebychev polynomials and are very appropriate for non-periodic problems.

- Domain of definition: $x \in [-1, 1]$
- Weight $\rho(x) = 1$
- Inner product:

$$\int_{-1}^{1} L_i L_j \, dx = \frac{2}{2i+1}\delta_{ij} \tag{E.1}$$

- Endpoint values:

$$L_N(\pm 1) = (\pm 1)^N \qquad \frac{dL_N}{dx}(\pm 1) = (\pm 1)^{N-1}\frac{N(N+1)}{2} \tag{E.2}$$

- Explicit form

$$L_N(x) = \frac{1}{2^N}\frac{d^N}{dx^N}\left(x^2-1\right)^N. \tag{E.3}$$

- Three-Term Recurrence and starting values:
  Legendre polynomials recurrence:

$$\begin{aligned} L_0(x) &= 1, \qquad L_1(x) = x, \\ L_{N+1}(x) &= \frac{1}{N+1}\left[(2N+1)\,xL_N(x) - NL_{N-1}(x)\right]; \end{aligned} \tag{E.4}$$

  First derivative Legendre polynomials recurrence:

$$\begin{aligned} L'_0(x) &= 1, \quad L'_1(x) = 1, \quad L'_2(x) = 3x, \\ L'_{N+1}(x) &= \frac{1}{N}\left[(2N+1)\,xL'_N(x) - (N+1)\,L'_{N-1}(x)\right]; \end{aligned} \tag{E.5}$$

  Second derivative Legendre polynomials recurrence:

$$\begin{aligned} L''_0(x) &= 1, \qquad L''_1(x) = 0, \quad L''_2(x) = 3, \quad L''_3(x) = 15x \\ L''_{N+1}(x) &= \frac{1}{N-1}\left[(2N+1)\,xL''_N(x) - (N+2)\,L''_{N-1}(x)\right]. \end{aligned} \tag{E.6}$$

- General differentiation: There is a relationship between Legendre and Gegenbauer polynomials C_n^m

$$\frac{d^m L_N(x)}{dx^m} = 1 \cdot 3 \cdot 5 \cdots (2m-1) C_{N-m}^{m+1/2}(x). \quad \text{E.7}$$

- Important properties:

$$\left(\left(1-x^2\right) L_N'(x) \right)' = -N(N+1) L_N(x); \quad \text{E.8}$$

$$(N+1) L_{N+1}(x) = (2N+1) x L_N(x) - N L_{N-1}(x); \quad \text{E.9}$$

$$N L_N(x) = x L_N'(x) - L_{N-1}'(x); \quad \text{E.10}$$

$$L_{N+1}'(x) = x L_N'(x) + (n+1) L_N(x). \quad \text{E.11}$$

- Relation to the Lobatto Polynomials $Lo_N(x)$:

$$Lo_N(x) = L_{N+1}'(x). \quad \text{E.12}$$

- Relation to the Gegenbauer polynomials C_n^m

$$L_N(x) = C_N^{(1/2)}(x). \quad \text{E.13}$$

- Relation to the Jacobi polynomials $J_N^{(\alpha,\beta)}$

$$L_N(x) = J^{(0,0)_N}(x). \quad \text{E.14}$$

F

General shaped elements

In the problems treated in this thesis, we have dealt with meshes containing ***regular*** shaped elements. In that case, the linear mapping transformation is given by

$$x = x(\xi); \tag{F.1}$$
$$y = y(\eta); \tag{F.2}$$
$$z = z(\zeta). \tag{F.3}$$

Note that the Jacobian of this type of mapping is ***constant*** in each element. This will not be the case anymore when using ***general*** shaped elements, where the mapping transformations are:

$$x = x(\xi, \eta, \zeta); \tag{F.4}$$
$$y = y(\xi, \eta, \zeta); \tag{F.5}$$
$$z = z(\xi, \eta, \zeta). \tag{F.6}$$

Recall that in each subdomain, the ***generic*** variable $u_h^{\mathbf{k}}$ is expanded into cardinal basis functions. In higher dimensions, the formulation of the basis comes from the tensor product of one dimensional Lagrangian interpolant basis $h_i(x)$. So the Lagrangian interpolants are chosen as basis functions in each dimension. We expand the unknowns as

$$\forall u_h^{\mathbf{k}} \in \mathcal{W}_h,\ u_h^{\mathbf{k}}(x,y,z,t) = \sum_{m=0}^{m=N}\sum_{n=0}^{n=N}\sum_{p=0}^{p=N} u_{mnp}^{\mathbf{k}}(t)\ h_m(x)\ h_n(y)\ h_p(z). \tag{F.7}$$

However, for ***regular*** shaped elements, we can also differentiate the unknowns with respect to x, y, z or t in the following manner:

$$\partial_x u_h^{\mathbf{k}}(x,y,z,t) = \sum_{m=0}^{m=N}\sum_{n=0}^{n=N}\sum_{p=0}^{p=N} u_{mnp}^{\mathbf{k}}(t)\ \partial_x h_m(x)\ h_n(y)\ h_p(z), \tag{F.8}$$

$$\partial_y u_h^{\mathbf{k}}(x,y,z,t) = \sum_{m=0}^{m=N}\sum_{n=0}^{n=N}\sum_{p=0}^{p=N} u_{mnp}^{\mathbf{k}}(t)\ h_m(x)\ \partial_y h_n(y)\ h_p(z), \tag{F.9}$$

$$\partial_z u_h^{\mathbf{k}}(x,y,z,t) = \sum_{m=0}^{m=N}\sum_{n=0}^{n=N}\sum_{p=0}^{p=N} u_{mnp}^{\mathbf{k}}(t)\ h_m(x)\ h_n(y)\ \partial_z h_p(z), \tag{F.10}$$

$$\partial_t u_h^{\mathbf{k}}(x,y,z,t) = \sum_{m=0}^{m=N}\sum_{n=0}^{n=N}\sum_{p=0}^{p=N} \dot{u}_{mnp}^{\mathbf{k}}(t)\ h_m(x)\ h_n(y)\ h_p(z). \tag{F.11}$$

For ***general*** shaped elements, we have to use the ***chain rule***, and the derivative for x for example, becomes:

$$\partial_x u_h^{\mathbf{k}}(x,y,z,t) = \sum_{m=0}^{m=N}\sum_{n=0}^{n=N}\sum_{p=0}^{p=N} u_{mnp}^{\mathbf{k}}(t)\left[\partial_\xi h_m(\xi)\, h_n(\eta)\, h_p(\zeta)\frac{\partial \xi}{\partial x} + h_m(\xi)\, \partial_\eta h_n(\eta)\, h_p(\zeta)\frac{\partial \eta}{\partial x} + h_m(\xi)\, h_n(\eta)\, \partial_\zeta h_p(\zeta)\frac{\partial \zeta}{\partial x}\right]. \quad \text{F.12}$$

We show here, how this general case changes the formula for the Elemental advection matrix $\mathbf{A}_k$ type 1.

Recall that the advection matrix $\mathbf{A}_k$ appears in the following type of integral

$$\int_{\mathbf{\Omega}} f\partial_k u\; w\; d\mathbf{\Omega} = f : (\mathbf{A}_k \otimes u)\,, \quad \text{F.13}$$

where, f, u and w are scalar functions and $k = x, y$, or z. The operator $\otimes$ will act on u in a different manner depending on the value of k as described below:

$$\mathbf{A}_{k\xi} \otimes u = |J|\,\frac{\partial \xi}{\partial k} : \rho : (H \cdot_{xy} u) \quad \text{F.14}$$

$$\mathbf{A}_{k\eta} \otimes u = |J|\,\frac{\partial \eta}{\partial k} : \rho : \left(u \cdot_{xy} H^T\right) \quad \text{F.15}$$

$$\mathbf{A}_{k\zeta} \otimes u = |J|\,\frac{\partial \zeta}{\partial k} : \rho : \left(u \cdot_{yz} H^T\right). \quad \text{F.16}$$

And finally, in the ***general*** shaped element case, we have

$$\mathbf{A}_k \otimes u = \mathbf{A}_{k\xi} \otimes u + \mathbf{A}_{k\eta} \otimes u + \mathbf{A}_{k\zeta} \otimes u. \quad \text{F.17}$$

Calculations for the other Elemental matrices can be derived in a similar fashion.

G

Extended numerical results of the SEM and BSSN

This Appendix contains more detailed results and figures describing the numerical results discussed in Chapter 9.

G.1 Geometric flexibility

G.1.1 Distorted Meshes

We present distorted square and cubic meshes with varying parameters

- Figure (G.1) presents a 2D slice of a 3D distorted ***square*** mesh with anchor points without GLL points with $N_E = 5^3, 7^3, 9^3, 11^3$ respectively.
- Figure (G.2) presents a 2D slice of a 3D distorted ***cubic*** mesh with anchor points without GLL points with $N_E = 5^3, 7^3, 9^3, 11^3$ respectively.

Tables (G.1)and (G.2) illustrate the requirements for distorted meshes (square and cubic respectively). We show the mesh properties with varying parameters (number of elements N_E and polynomial order N): minimum and maximum dx, timestep dt required and the total number of points N_g

G.1.2 Mixed Distorted meshes

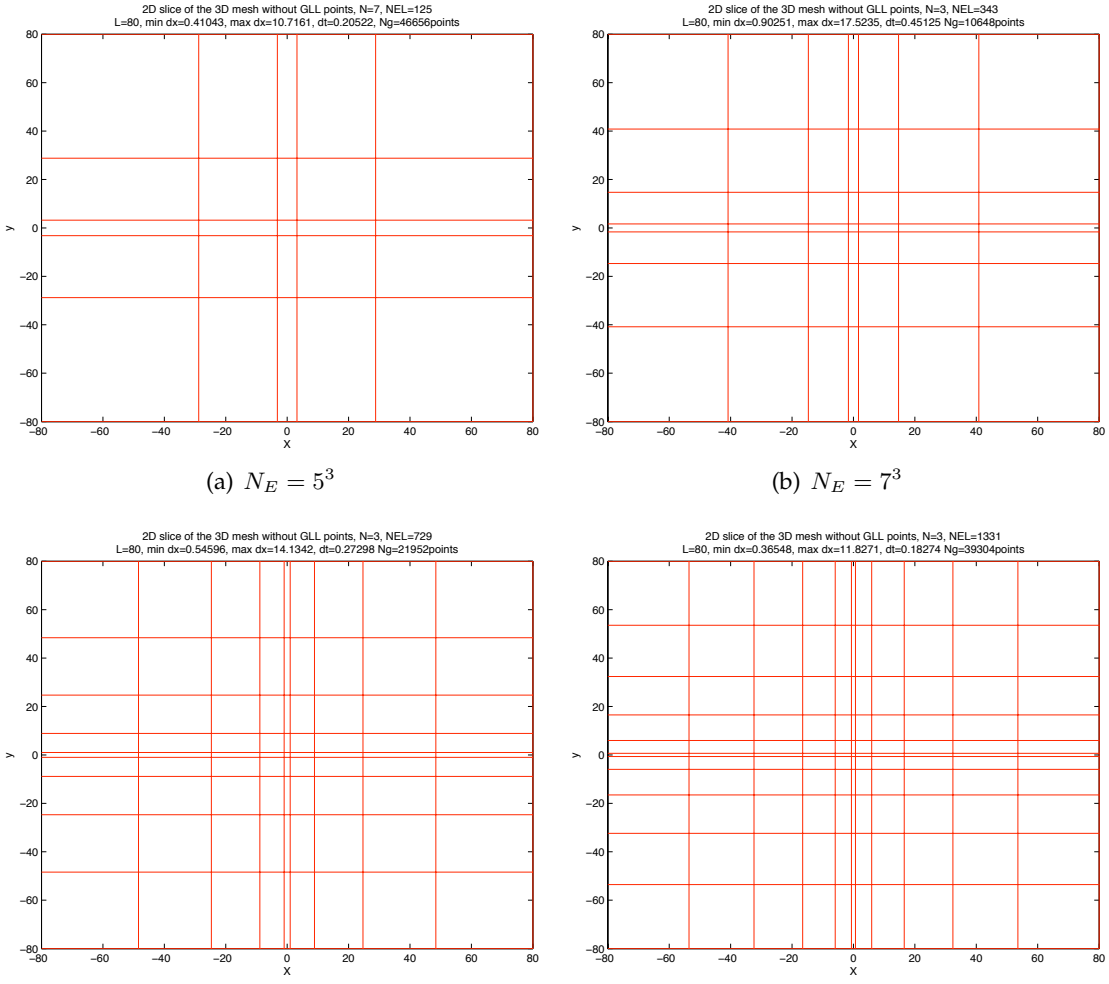
Figure 9.7 illustrates 2 mixed distorted meshes. We present here mixed distorted meshes for varying number of elements N_E, see Figure G.3 for details.

G.2 Far from the puncture

Sufficiently far from the puncture, all the variables of the BSSN system are smooth.

Figures G.4, G.5, G.6, G.7, G.8, G.9, G.10, G.11, and G.12, show the pointwise errors and $\mathcal{L}^2$ norms of most variables of the BSSN system for a domain $L = 64$ with a cubic mesh.

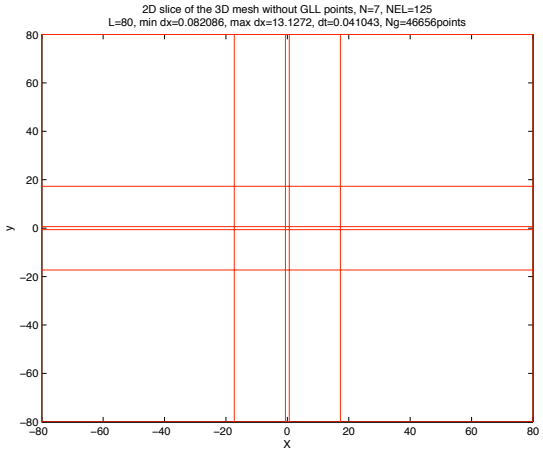
These numerical results demonstrate that the part of the code and system that we expect to behave well really do so.



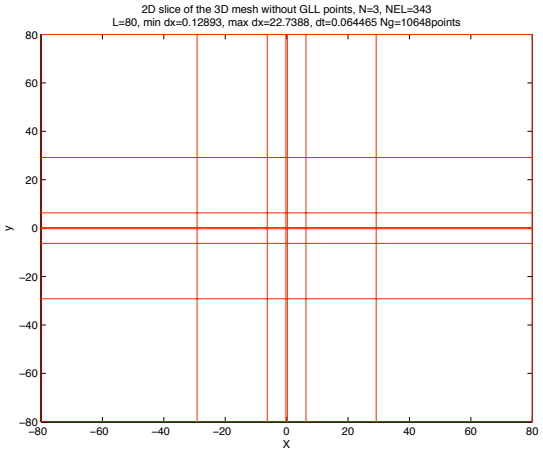
(a) $N_E = 5^3$ (b) $N_E = 7^3$

(c) $N_E = 9^3$ (d) $N_E = 11^3$

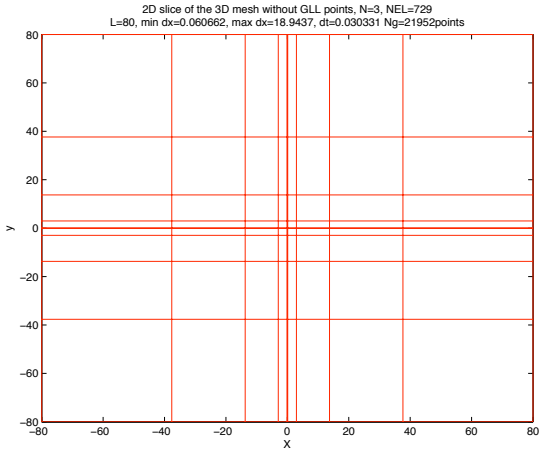
Figure G.1: 3D distorted ***square*** mesh represented in a 2D slice for varying number of elements N_E and for a domain $L = 80$.



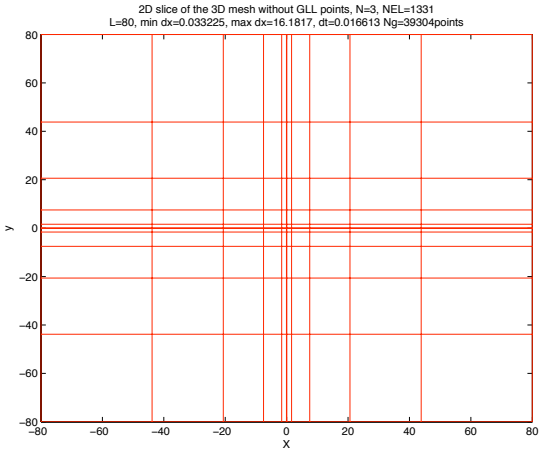
(a) $N_E = 5^3$



(b) $N_E = 7^3$



(c) $N_E = 9^3$



(d) $N_E = 11^3$

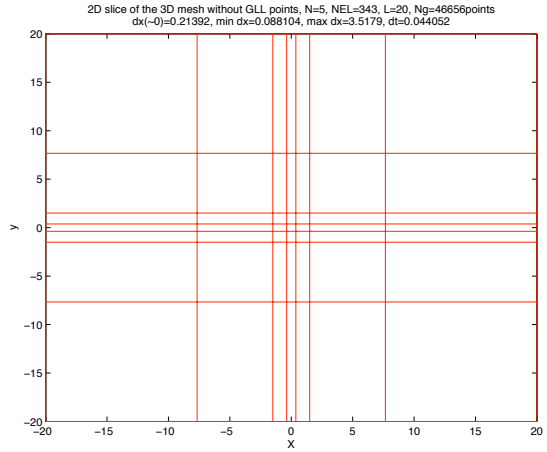
Figure G.2: Same as figure G.1 but for a 3D distorted ***cubic*** mesh.

Nb elements	poly order N	$dx(\sim 0)$	dx_{min}	dx_{max}	dt	Total points
5^3	$N = 7$	$dx = 1.34$	$dx = 0.41$	$dx = 10.71$	$dt = 0.21$	$N_g = 46656$
5^3	$N = 9$	$dx = 1.06$	$dx = 0.26$	$dx = 8.5$	$dt = 0.13$	$N_g = 97336$
5^3	$N = 15$	$dx = 0.67$	$dx = 0.097$	$dx = 5.2$	$dt = 0.05$	$N_g = 438976$
5^3	$N = 17$	$dx = 0.57$	$dx = 0.076$	$dx = 4.59$	$dt = 0.04$	$N_g = 636056$
7^3	$N = 3$	$dx = 1.46$	$dx = 0.9$	$dx = 17.5$	$dt = 0.45$	$N_g = 10648$
7^3	$N = 9$	$dx = 0.54$	$dx = 0.13$	$dx = 6.5$	$dt = 0.07$	$N_g = 262144$
7^3	$N = 11$	$dx = 0.45$	$dx = 0.09$	$dx = 5.3$	$dt = 0.045$	$N_g = 474552$
7^3	$N = 15$	$dx = 0.33$	$dx = 0.05$	$dx = 3.9$	$dt = 0.024$	$N_g = 1191016$
9^3	$N = 3$	$dx = 0.88$	$dx = 0.55$	$dx = 14.1$	$dt = 0.27$	$N_g = 21952$
9^3	$N = 5$	$dx = 0.56$	$dx = 0.23$	$dx = 9$	$dt = 0.12$	$N_g = 97336$
9^3	$N = 9$	$dx = 0.33$	$dx = 0.08$	$dx = 5.2$	$dt = 0.04$	$N_g = 551368$
9^3	$N = 11$	$dx = 0.27$	$dx = 0.054$	$dx = 4.3$	$dt = 0.027$	$N_g = 1000000$
11^3	$N = 3$	$dx = 0.59$	$dx = 0.37$	$dx = 11.8$	$dt = 0.18$	$N_g = 39304$
11^3	$N = 5$	$dx = 0.38$	$dx = 0.16$	$dx = 7.5$	$dt = 0.08$	$N_g = 175616$
11^3	$N = 9$	$dx = 0.22$	$dx = 0.05$	$dx = 4.4$	$dt = 0.027$	$N_g = 1000000$
11^3	$N = 11$	$dx = 0.18$	$dx = 0.036$	$dx = 3.6$	$dt = 0.018$	$N_g = 1815848$

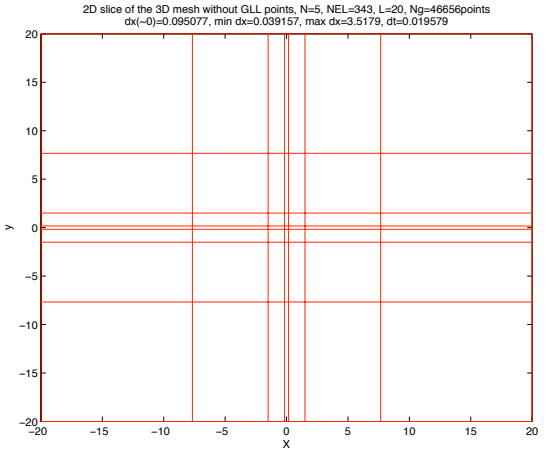
Table G.1: Properties for a distorted square mesh with a domain $L = 80$. The timestep is given by $dt = CFL \times dx_{min}$ with $CFL = 0.5$.

Nb elements	poly order N	$dx(\sim 0)$	dx_{min}	dx_{max}	dt	Total points
5^3	$N = 7$	$dx = 0.27$	$dx = 0.082$	$dx = 13.13$	$dt = 0.04$	$N_g = 46656$
5^3	$N = 9$	$dx = 0.21$	$dx = 0.051$	$dx = 10.36$	$dt = 0.017$	$N_g = 97336$
5^3	$N = 11$	$dx = 0.17$	$dx = 0.035$	$dx = 8.6$	$dt = 0.05$	$N_g = 175616$
5^3	$N = 15$	$dx = 0.13$	$dx = 0.019$	$dx = 6.35$	$dt = 0.0097$	$N_g = 438976$
5^3	$N = 17$	$dx = 0.11$	$dx = 0.015$	$dx = 5.62$	$dt = 0.0076$	$N_g = 636056$
7^3	$N = 3$	$dx = 0.21$	$dx = 0.13$	$dx = 22.7$	$dt = 0.064$	$N_g = 10648$
7^3	$N = 5$	$dx = 0.13$	$dx = 0.055$	$dx = 14.5$	$dt = 0.027$	$N_g = 46656$
7^3	$N = 9$	$dx = 0.077$	$dx = 0.019$	$dx = 8.4$	$dt = 0.009$	$N_g = 262144$
7^3	$N = 11$	$dx = 0.06$	$dx = 0.013$	$dx = 6.9$	$dt = 0.006$	$N_g = 474552$
7^3	$N = 15$	$dx = 0.047$	$dx = 0.007$	$dx = 5.1$	$dt = 0.0035$	$N_g = 1191016$
9^3	$N = 3$	$dx = 0.098$	$dx = 0.06$	$dx = 18.9$	$dt = 0.03$	$N_g = 21952$
9^3	$N = 5$	$dx = 0.06$	$dx = 0.026$	$dx = 12.08$	$dt = 0.013$	$N_g = 97336$
9^3	$N = 9$	$dx = 0.036$	$dx = 0.009$	$dx = 7$	$dt = 0.004$	$N_g = 551368$
9^3	$N = 11$	$dx = 0.03$	$dx = 0.006$	$dx = 5.8$	$dt = 0.003$	$N_g = 1000000$
11^3	$N = 3$	$dx = 0.053$	$dx = 0.033$	$dx = 10.3$	$dt = 0.017$	$N_g = 39304$
11^3	$N = 5$	$dx = 0.034$	$dx = 0.014$	$dx = 7.5$	$dt = 0.07$	$N_g = 175616$
11^3	$N = 9$	$dx = 0.02$	$dx = 0.0048$	$dx = 5.98$	$dt = 0.0024$	$N_g = 1000000$
11^3	$N = 11$	$dx = 0.016$	$dx = 0.0033$	$dx = 4.9$	$dt = 0.0017$	$N_g = 1815848$

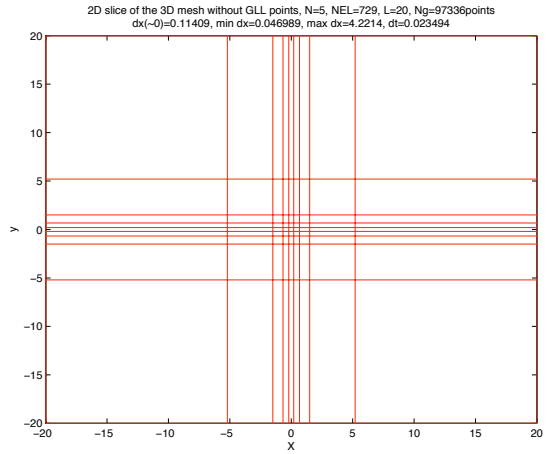
Table G.2: Same as table G.1 but for a distorted cubic mesh.



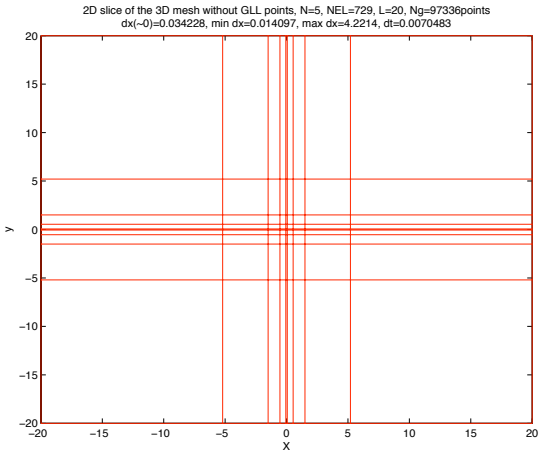
(a) Even inside Box $N_{EIn} = 3^3$, square outside Box, $N_E = 7^3$.



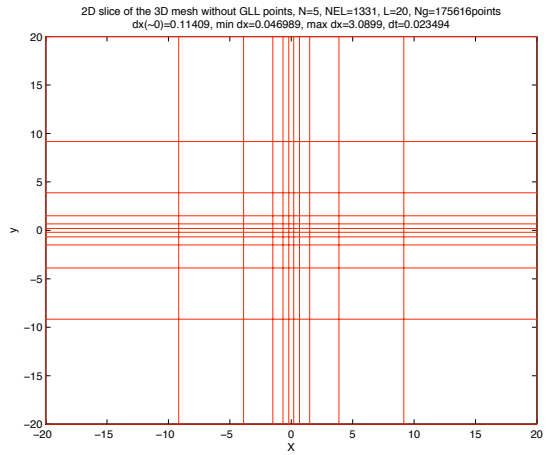
(b) Square inside Box $N_{EIn} = 3^3$, square outside Box, $N_E = 7^3$.



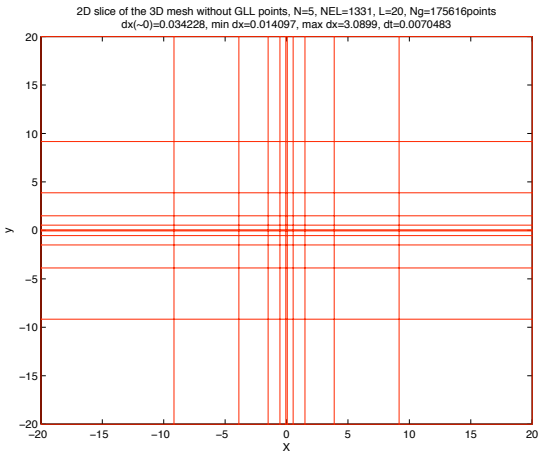
(c) Even inside Box $N_{EIn} = 5^3$, square outside Box, $N_E = 7^3$.



(d) Square inside Box $N_{EIn} = 5^3$, square outside Box, $N_E = 7^3$.

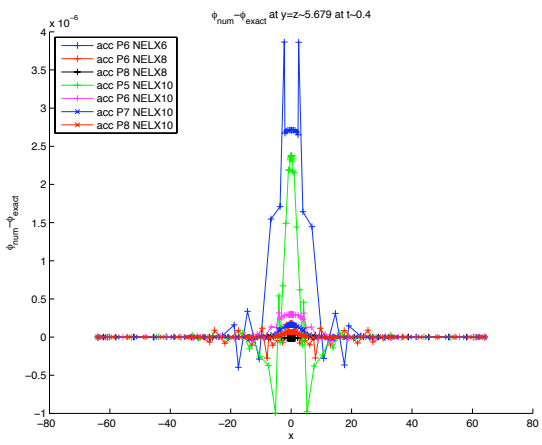


(e) Even inside Box $N_{EIn} = 5^3$, square outside Box, $N_E = 11^3$.

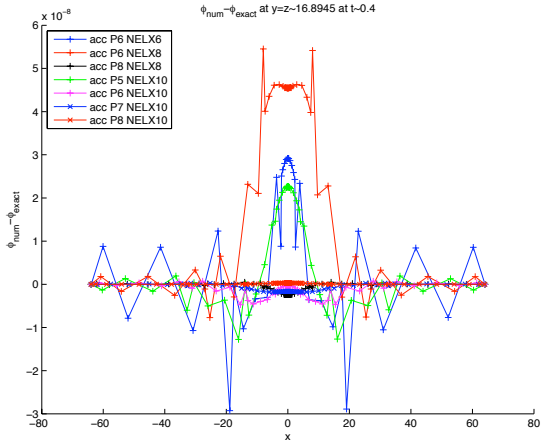


(f) Square inside Box $N_{EIn} = 5^3$, square outside Box, $N_E = 11^3$.

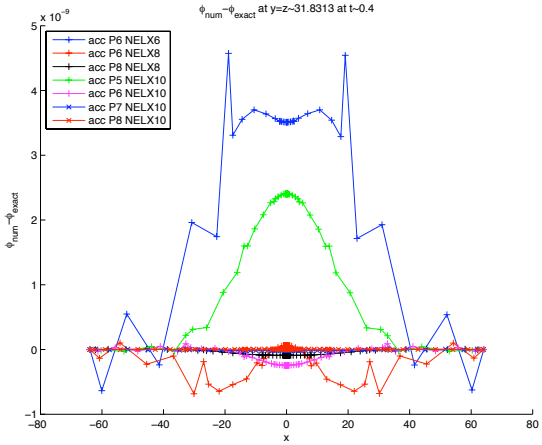
Figure G.3: 3D mixed distorted meshes represented in a 2D slice for a domain $L = 20$.



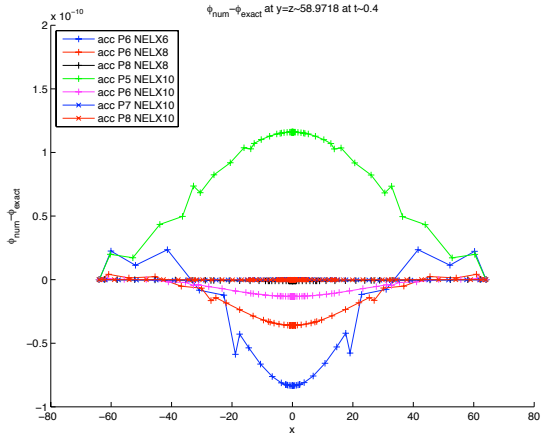
(a) ϕ versus x in a 2D slice for $y \sim z = 5.7$M



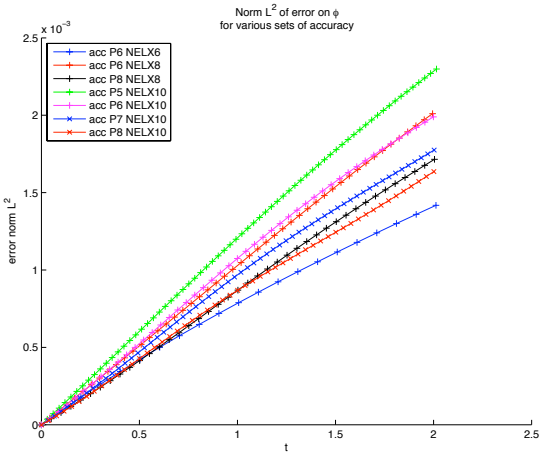
(b) ϕ versus x in a 2D slice for $y \sim z = 16.9$M



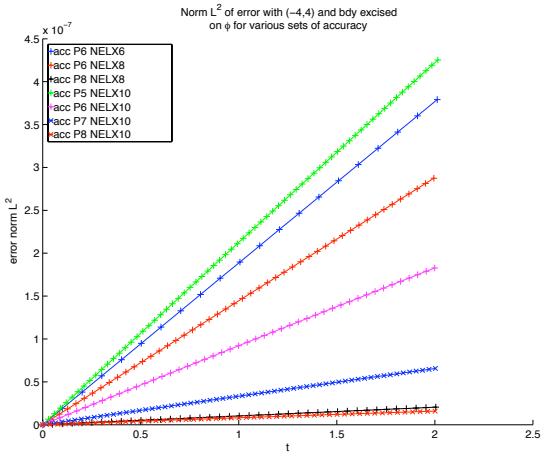
(c) ϕ versus x in a 2D slice for $y \sim z = 31.83$M



(d) ϕ versus x in a 2D slice for $y \sim z = 59$M

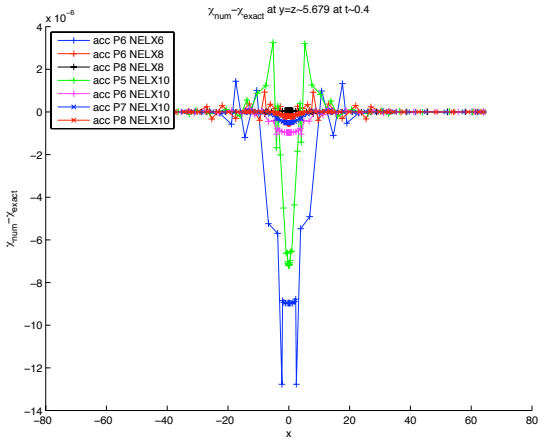


(e) Norm $\mathcal{L}^2$ of ϕ over the entire domain

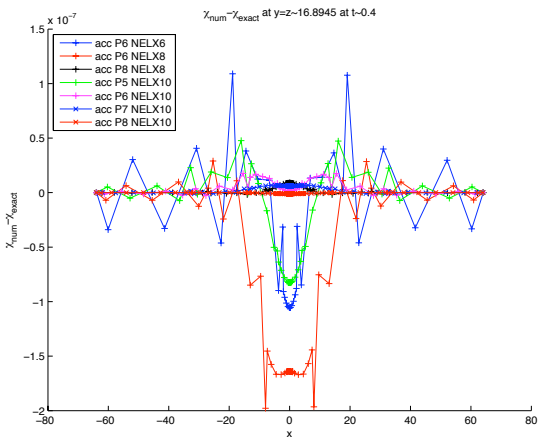


(f) Norm $\mathcal{L}^2$ of ϕ with the centre excised and the boundary excised

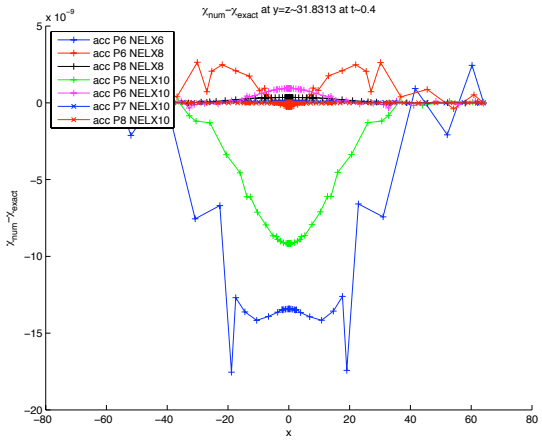
Figure G.4: Pointwise error and $\mathcal{L}^2$ norm for ϕ at the same time steps for varying accuracy and for different slices across a domain of $L = 64$ with a cubic mesh.



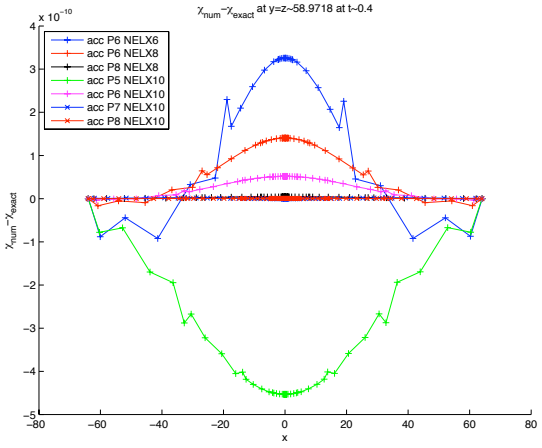
(a) χ versus x in a 2D slice for $y \sim z = 5.7$M



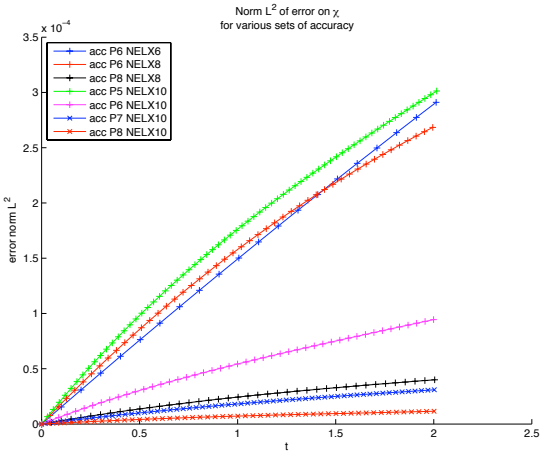
(b) χ versus x in a 2D slice for $y \sim z = 16.9$M



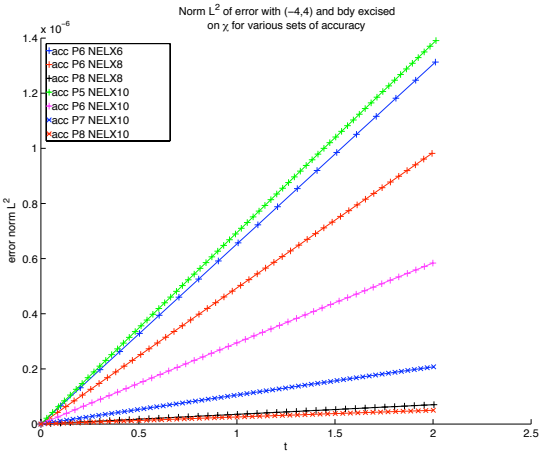
(c) χ versus x in a 2D slice for $y \sim z = 31.83$M



(d) χ versus x in a 2D slice for $y \sim z = 59$M

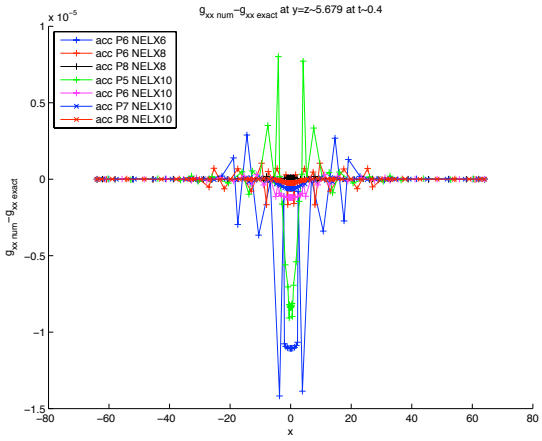


(e) Norm $\mathcal{L}^2$ of χ over the entire domain

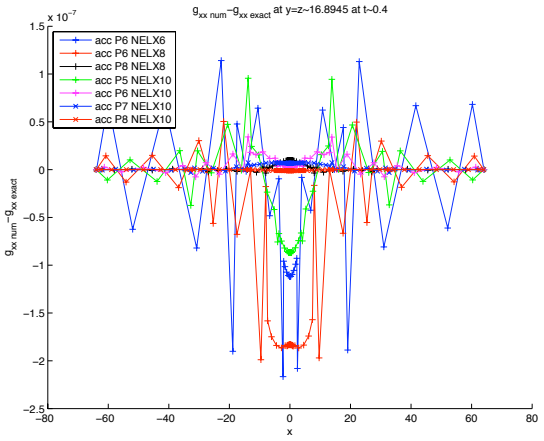


(f) Norm $\mathcal{L}^2$ of χ with the centre excised and the boundary excised

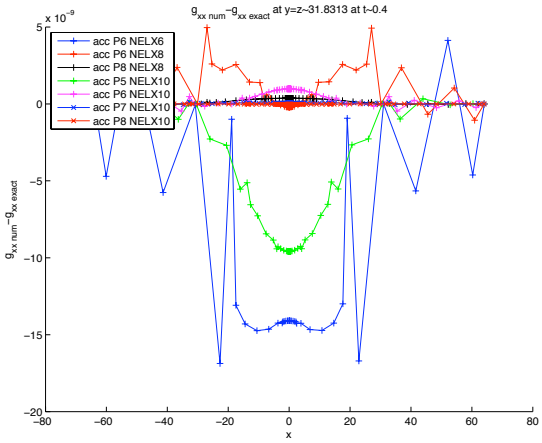
Figure G.5: Same as figure G.4 but for χ.



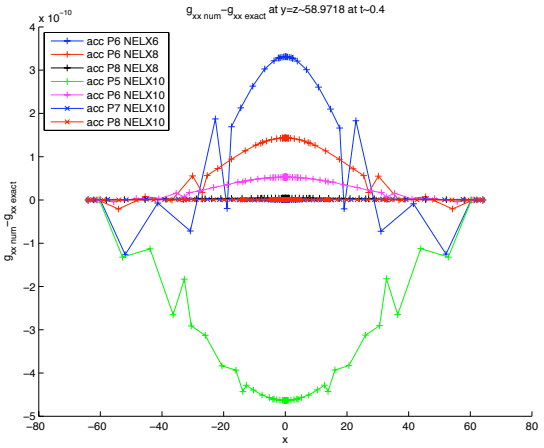
(a) $\tilde{g}_{xx}$ versus x in a 2D slice for $y \sim z = 5.7$M



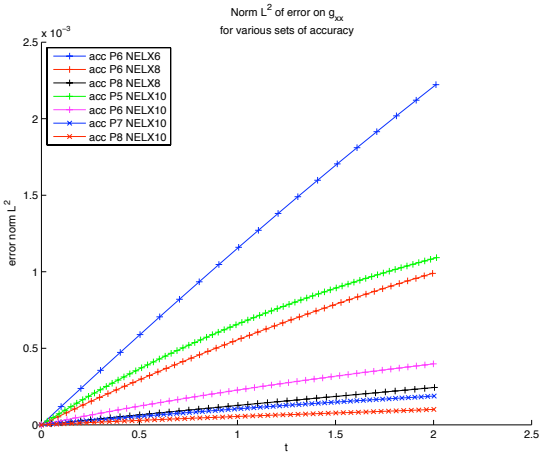
(b) $\tilde{g}_{xx}$ versus x in a 2D slice for $y \sim z = 16.9$M



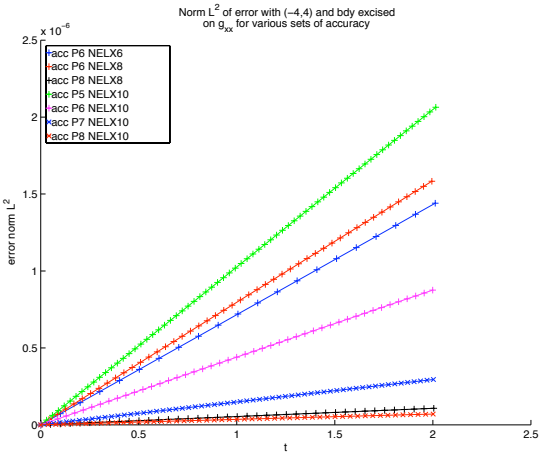
(c) $\tilde{g}_{xx}$ versus x in a 2D slice for $y \sim z = 31.83$M



(d) $\tilde{g}_{xx}$ versus x in a 2D slice for $y \sim z = 59$M

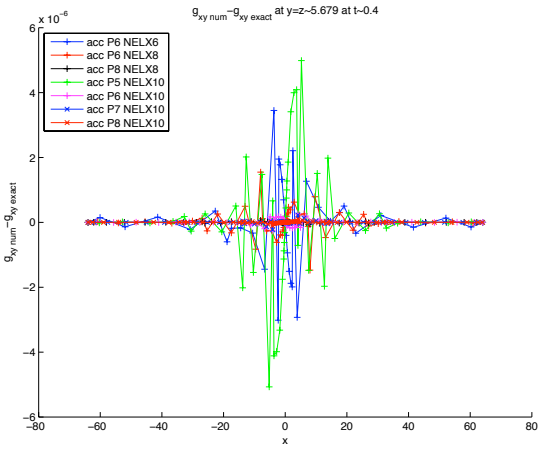


(e) Norm $\mathcal{L}^2$ of $\tilde{g}_{xx}$ over the entire domain

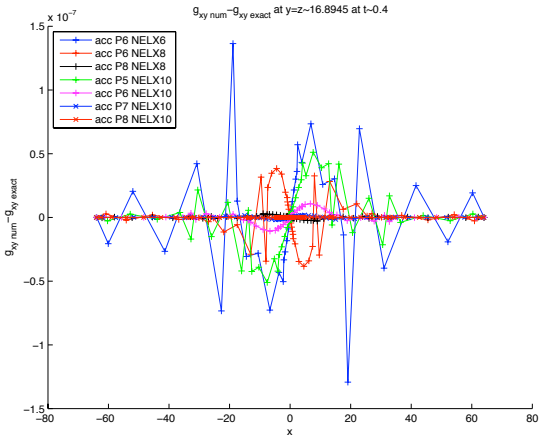


(f) Norm $\mathcal{L}^2$ of $\tilde{g}_{xx}$ with the centre excised and the boundary excised

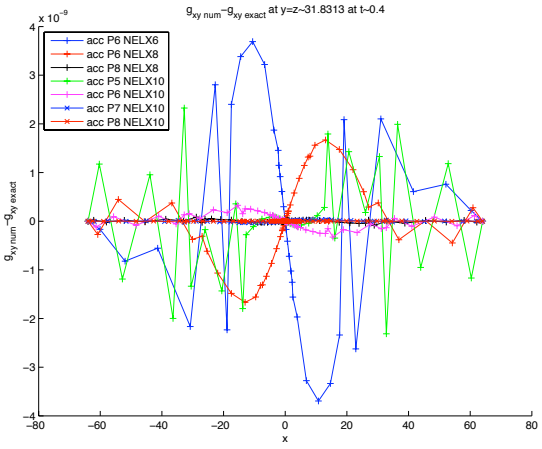
Figure G.6: Same as figure G.4 but for $\tilde{g}_{xx}$.



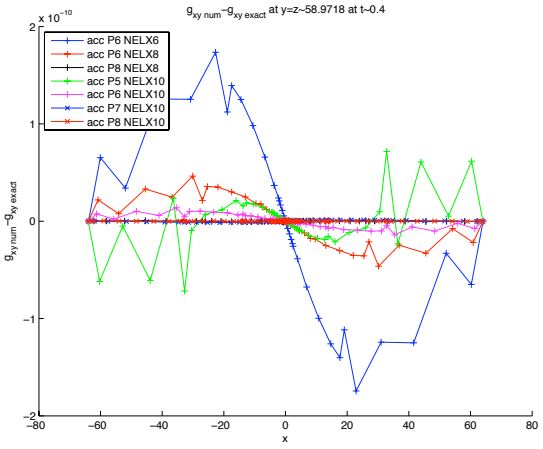
(a) $\tilde{g}_{xy}$ versus x in a 2D slice for $y \sim z = 5.7$M



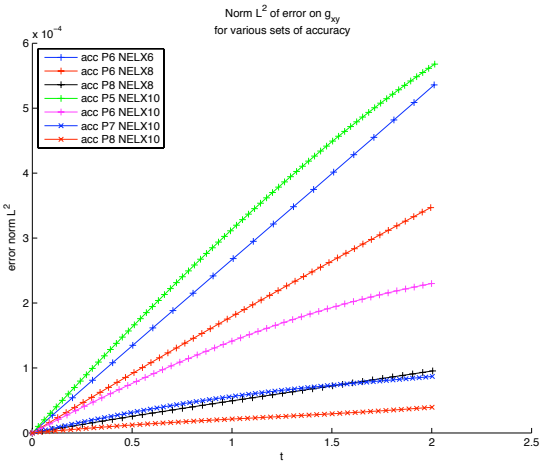
(b) $\tilde{g}_{xy}$ versus x in a 2D slice for $y \sim z = 16.9$M



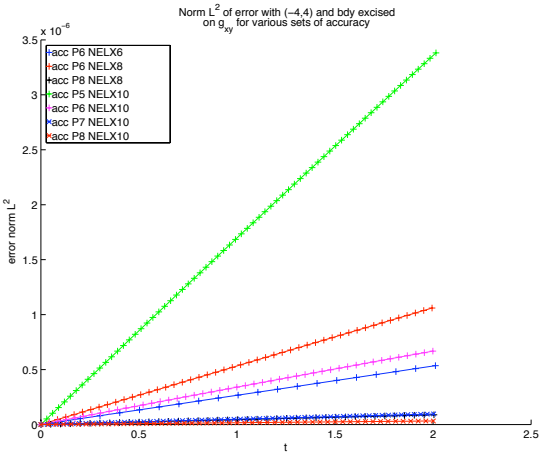
(c) $\tilde{g}_{xy}$ versus x in a 2D slice for $y \sim z = 31.83$M



(d) $\tilde{g}_{xy}$ versus x in a 2D slice for $y \sim z = 59$M

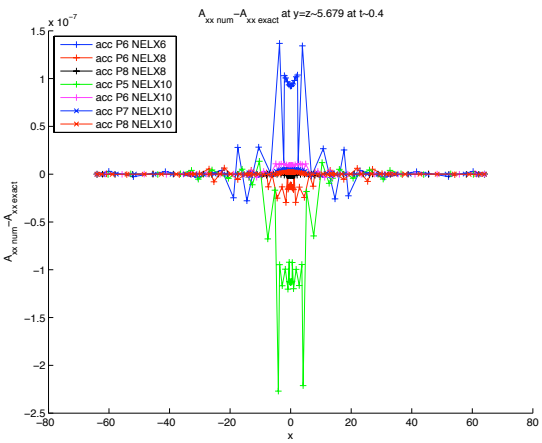


(e) Norm $\mathcal{L}^2$ of $\tilde{g}_{xy}$ over the entire domain

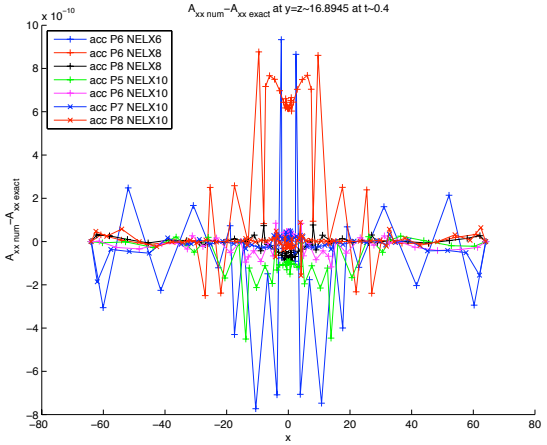


(f) Norm $\mathcal{L}^2$ of $\tilde{g}_{xy}$ with the centre excised and the boundary excised

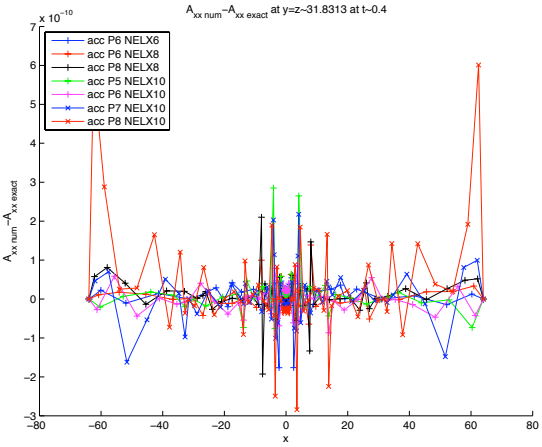
Figure G.7: Same as figure G.4 but for $\tilde{g}_{xy}$.



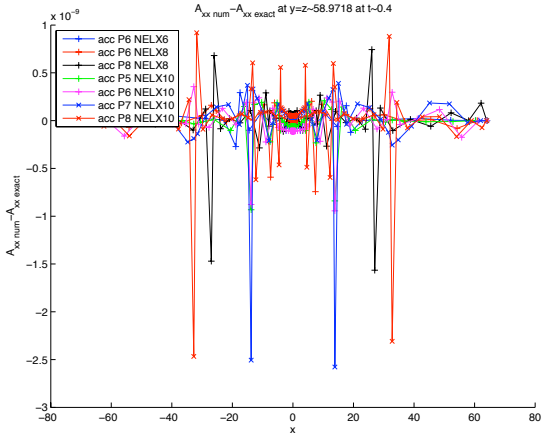
(a) $\tilde{A}_{xx}$ versus x in a 2D slice for $y \sim z = 5.7$M



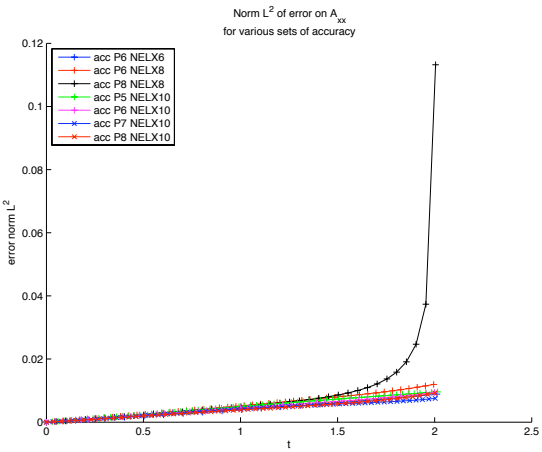
(b) $\tilde{A}_{xx}$ versus x in a 2D slice for $y \sim z = 16.9$M



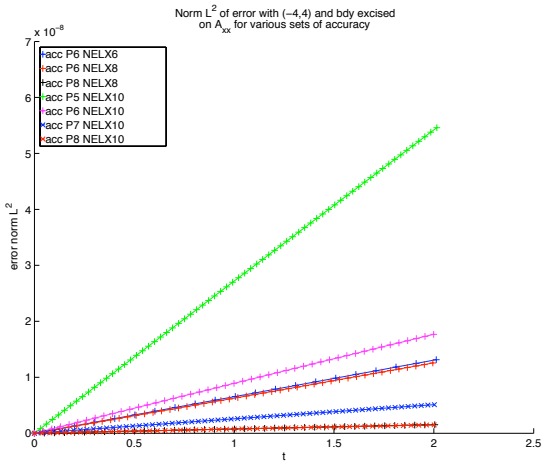
(c) $\tilde{A}_{xx}$ versus x in a 2D slice for $y \sim z = 31.83$M



(d) $\tilde{A}_{xx}$ versus x in a 2D slice for $y \sim z = 59$M

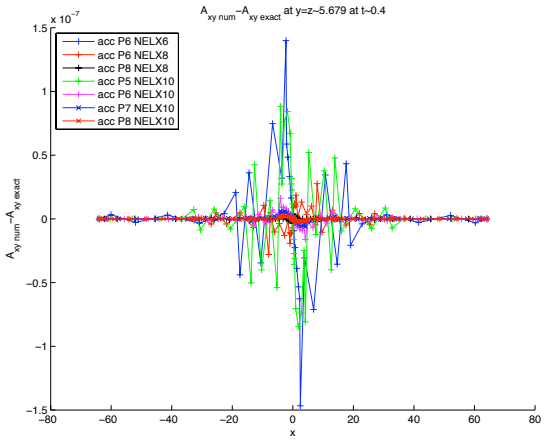


(e) Norm $\mathcal{L}^2$ of $\tilde{A}_{xx}$ over the entire domain

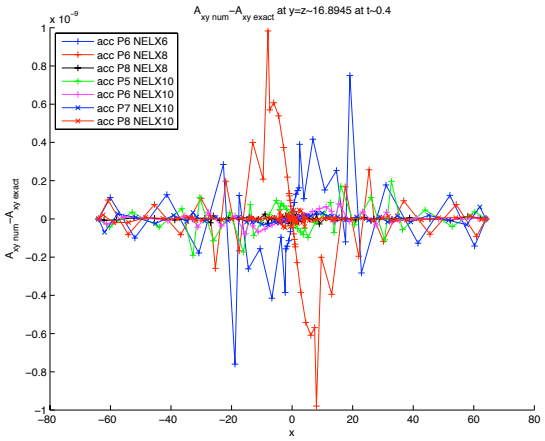


(f) Norm $\mathcal{L}^2$ of $\tilde{A}_{xx}$ with the centre excised and the boundary excised

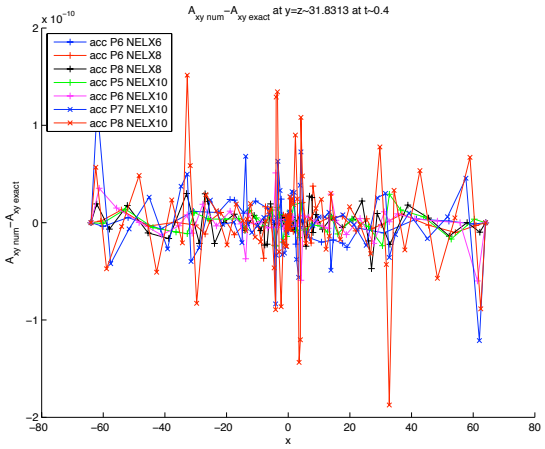
Figure G.8: Same as figure G.4 but for $\tilde{A}_{xx}$.



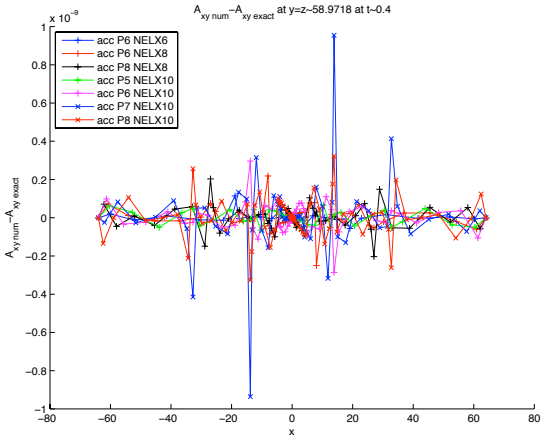
(a) $\tilde{A}_{xy}$ versus x in a 2D slice for $y \sim z = 5.7$M



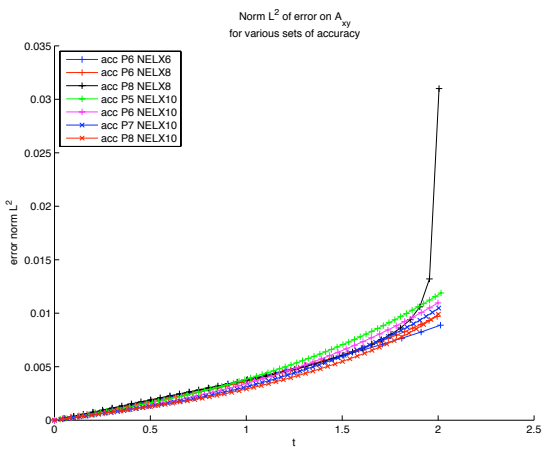
(b) $\tilde{A}_{xy}$ versus x in a 2D slice for $y \sim z = 16.9$M



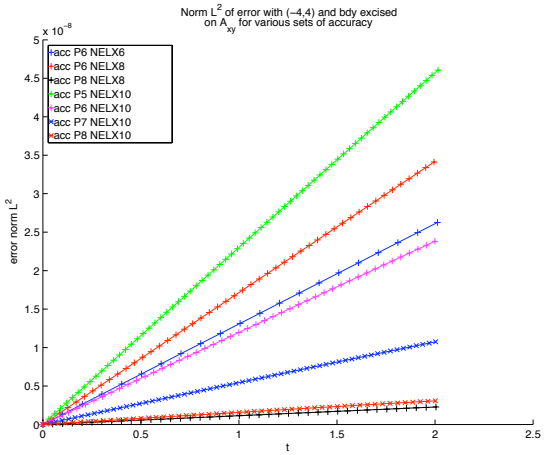
(c) $\tilde{A}_{xy}$ versus x in a 2D slice for $y \sim z = 31.83$M



(d) $\tilde{A}_{xy}$ versus x in a 2D slice for $y \sim z = 59$M

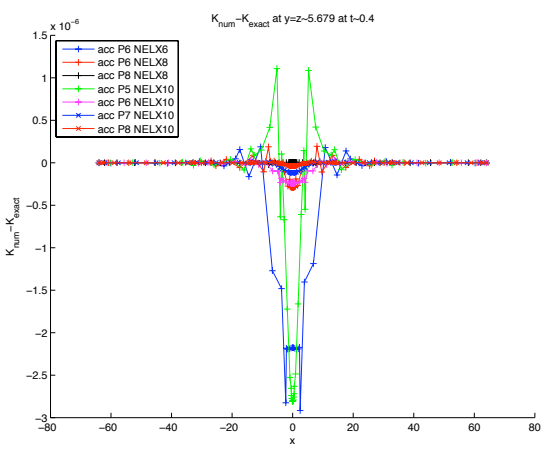


(e) Norm $\mathcal{L}^2$ of $\tilde{A}_{xy}$ over the entire domain

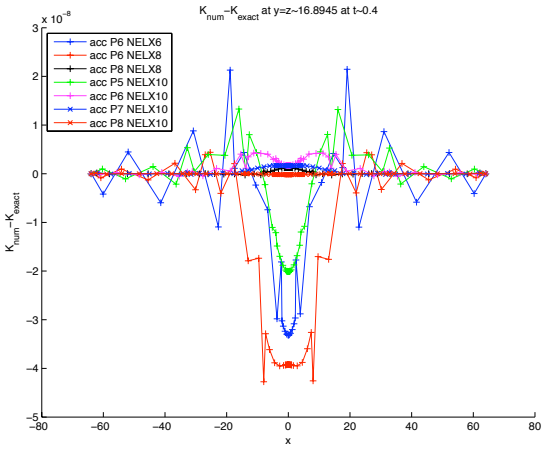


(f) Norm $\mathcal{L}^2$ of $\tilde{A}_{xy}$ with the centre excised and the boundary excised

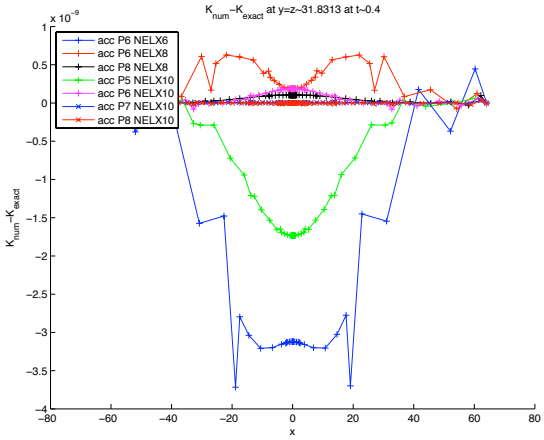
Figure G.9: Same as figure G.4 but for $\tilde{A}_{xy}$.



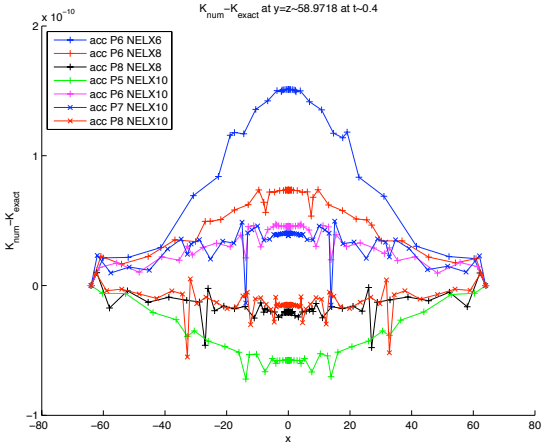
(a) K versus x in a 2D slice for $y \sim z = 5.7$M



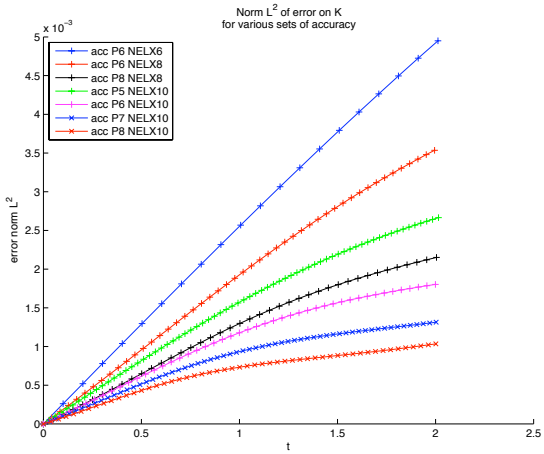
(b) K versus x in a 2D slice for $y \sim z = 16.9$M



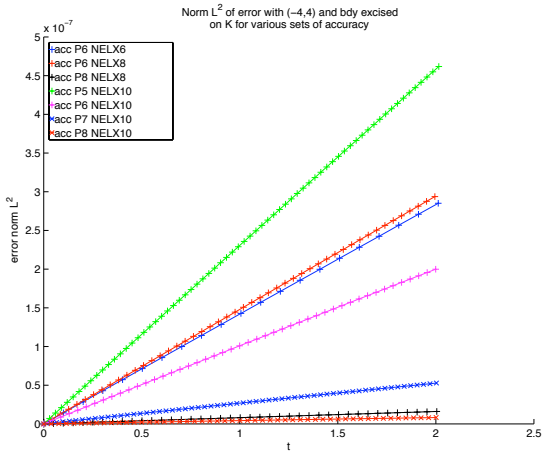
(c) K versus x in a 2D slice for $y \sim z = 31.83$M



(d) K versus x in a 2D slice for $y \sim z = 59$M

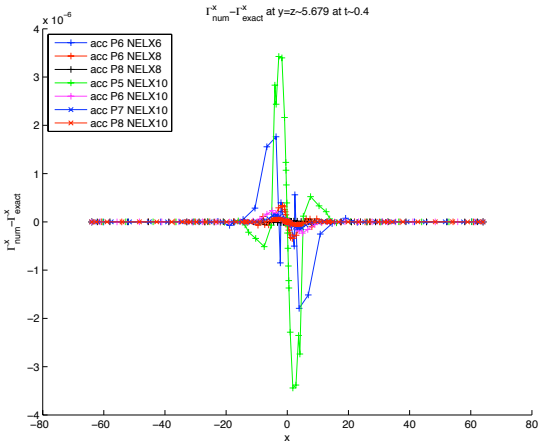


(e) Norm $\mathcal{L}^2$ of K over the entire domain

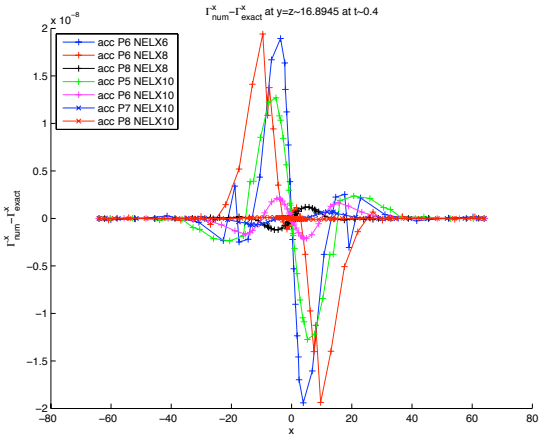


(f) Norm $\mathcal{L}^2$ of K with the centre excised and the boundary excised

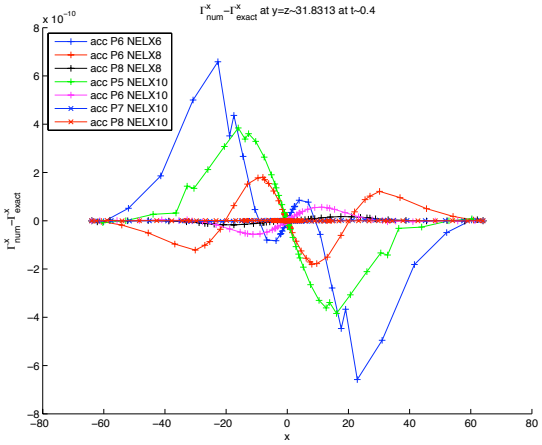
Figure G.10: Same as figure G.4 but for K.



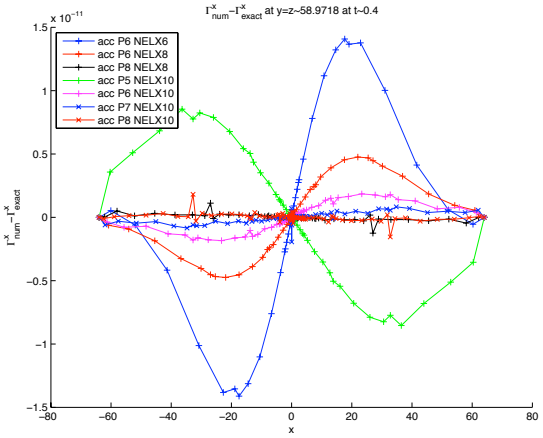
(a) $\tilde{\Gamma}^x$ versus x in a 2D slice for $y \sim z = 5.7$M



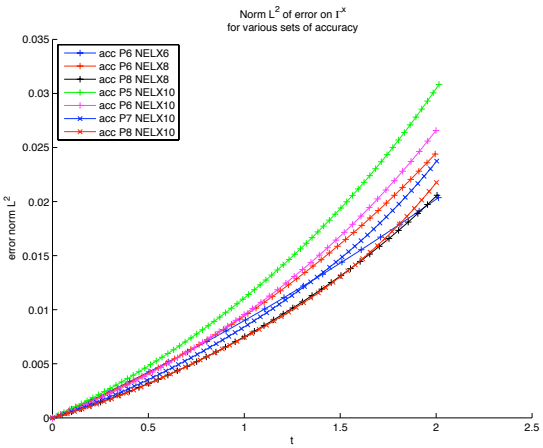
(b) $\tilde{\Gamma}^x$ versus x in a 2D slice for $y \sim z = 16.9$M



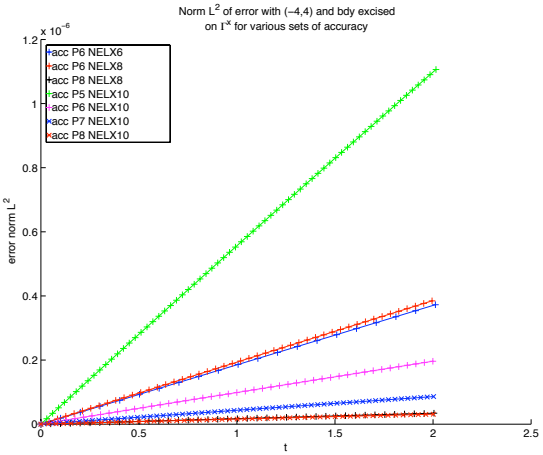
(c) $\tilde{\Gamma}^x$ versus x in a 2D slice for $y \sim z = 31.83$M



(d) $\tilde{\Gamma}^x$ versus x in a 2D slice for $y \sim z = 59$M

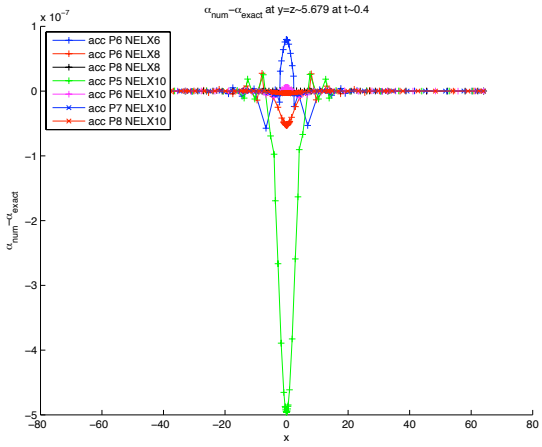


(e) Norm $\mathcal{L}^2$ of $\tilde{\Gamma}^x$ over the entire domain

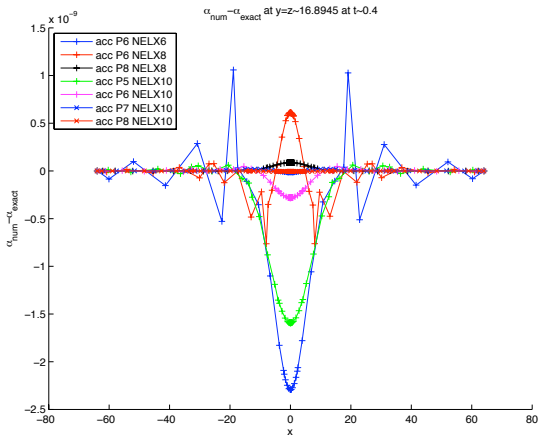


(f) Norm $\mathcal{L}^2$ of $\tilde{\Gamma}^x$ with the centre excised and the boundary excised

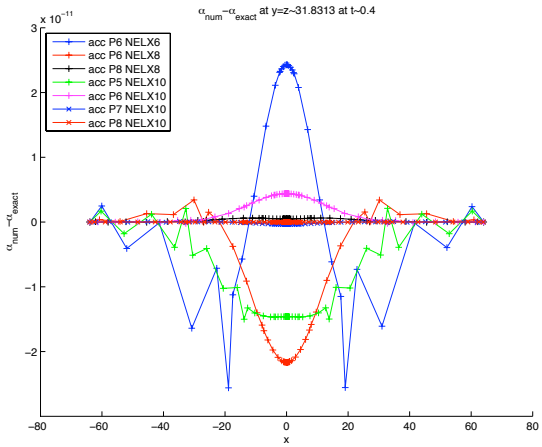
Figure G.11: Same as figure G.4 but for $\tilde{\Gamma}^x$.



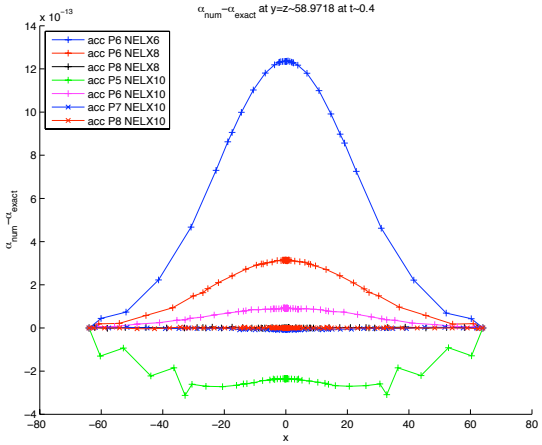
(a) α versus x in a 2D slice for $y \sim z = 5.7$M



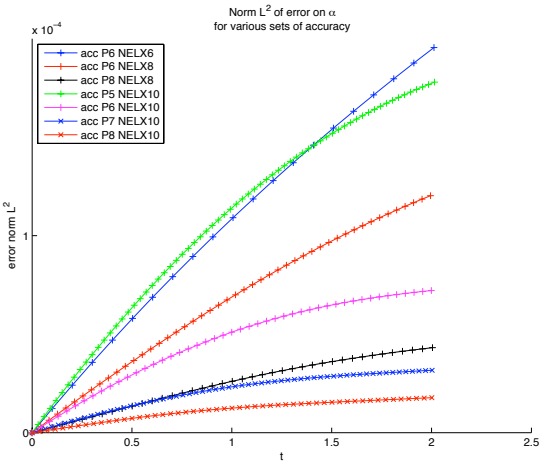
(b) α versus x in a 2D slice for $y \sim z = 16.9$M



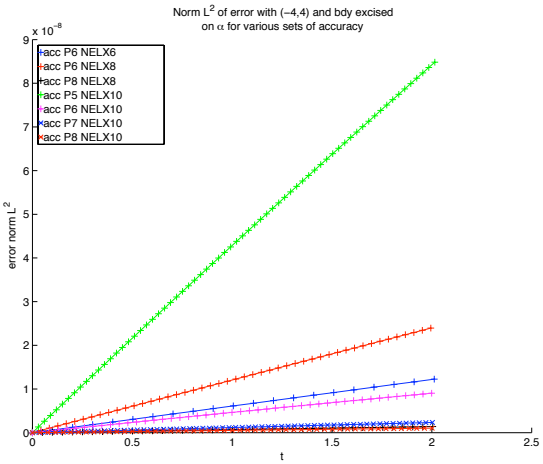
(c) α versus x in a 2D slice for $y \sim z = 31.83$M



(d) α versus x in a 2D slice for $y \sim z = 59$M



(e) Norm $\mathcal{L}^2$ of α over the entire domain

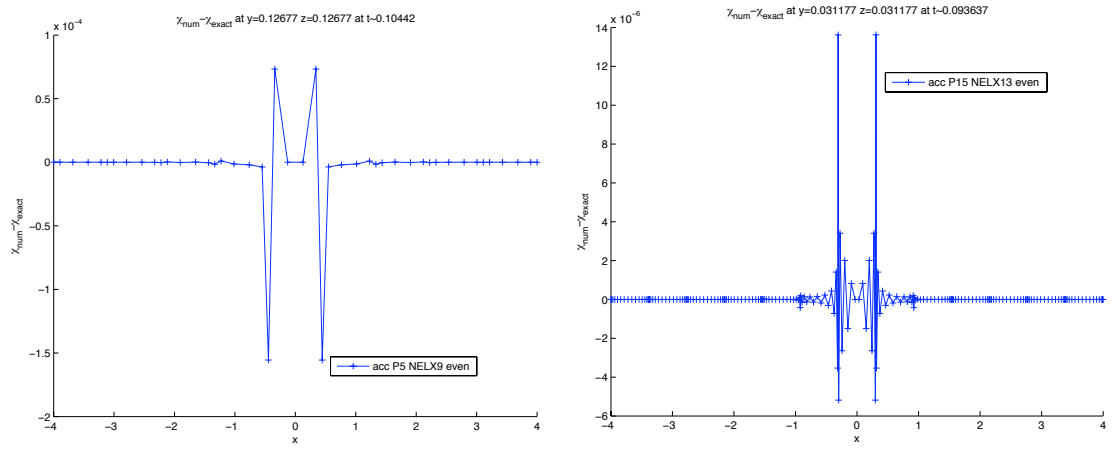


(f) Norm $\mathcal{L}^2$ of α with the centre excised and the boundary excised

Figure G.12: Same as figure G.4 but for α.

G.2.1 hp-convergence with χ

We obtain hp-convergence only away from the point of discontinuity.



(a) Pointwise error for χ near the puncture for low resolution, we see no oscillation in the neighbour elements of the puncture

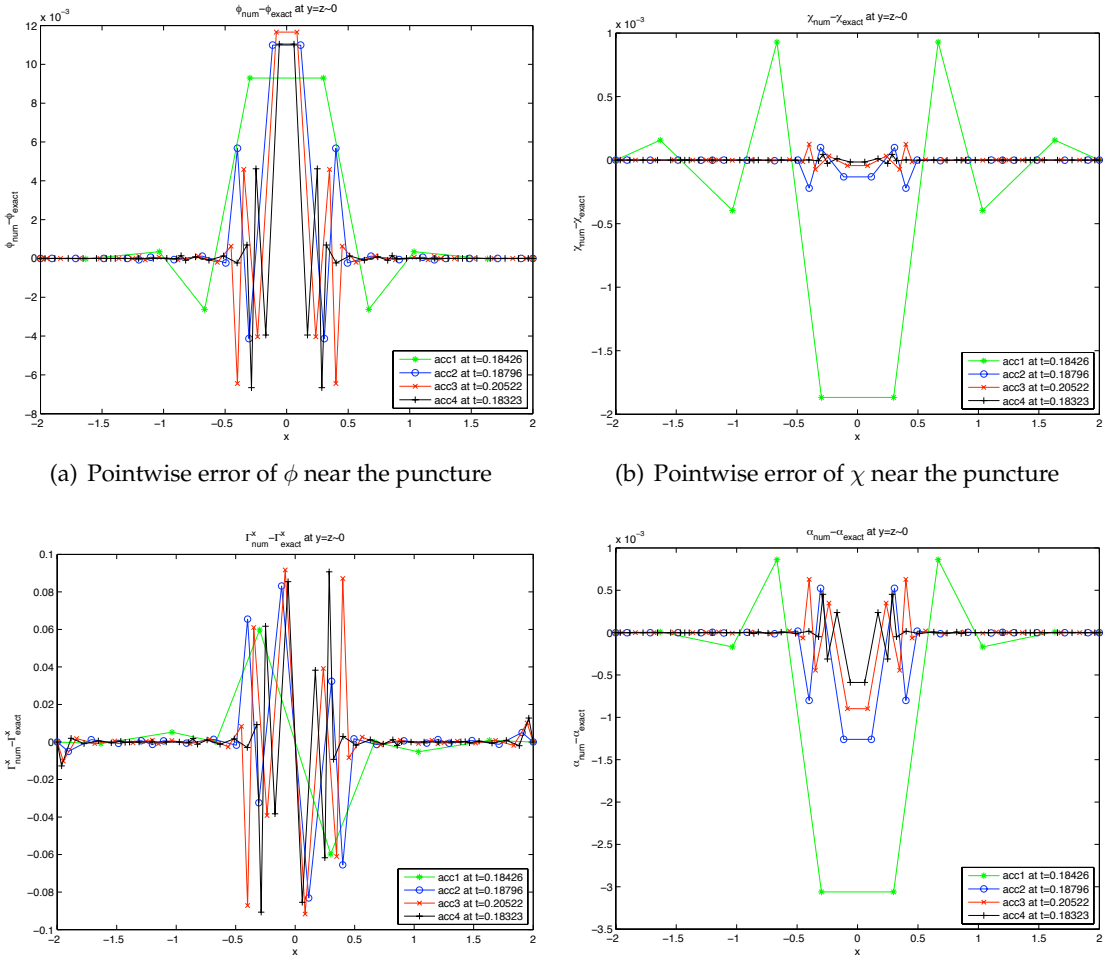
(b) Pointwise error for χ near the puncture for high resolution, we see oscillations in the neighbour elements of the puncture

Figure G.13: Pointwise error for χ near the puncture for low and high resolution on an ***even*** mesh $L = 4$: the effect of increasing the resolution near the puncture very quickly leads to the propagation of oscillations.

G.3 Puncture at the centre of an element

Figures G.14 and G.15 show the typical behaviour of discontinuous functions near the point of discontinuity: the appearance of Gibbs oscillations. We see the pointwise errors of most variables with 4 accuracies at a similar time of $t \sim 0.2M$.

In contrast, figures G.16 and G.17 show the $\mathcal{L}^2$ norm over the entire domain, and the $\mathcal{L}^2$ norm with the region close to the puncture being excised $(-0.67, 0.67)$, and the boundary $(L - 0.5)$ being excised. We can see that the $\mathcal{L}^2$ norm of the wave equation is doing much better than all the other variables over the entire domain. However, we can see from the excised norms that the norm of the solution of the wave equation is comparable to the BSSN variables until the oscillations propagate outside the centre element.



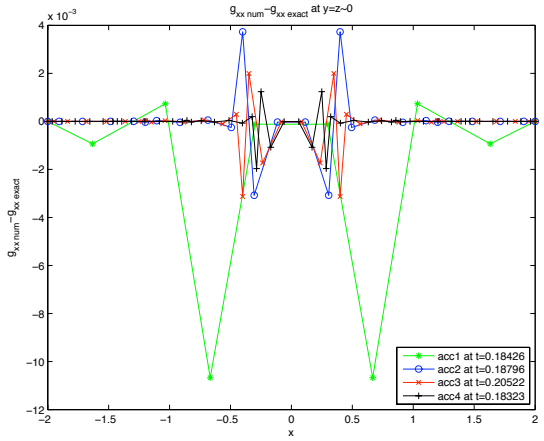
(a) Pointwise error of ϕ near the puncture

(b) Pointwise error of χ near the puncture

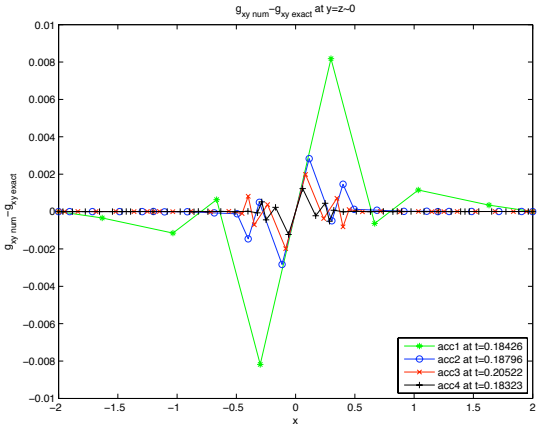
(c) Pointwise error of $\tilde{\Gamma}^x$ near the puncture

(d) Pointwise error of α near the puncture

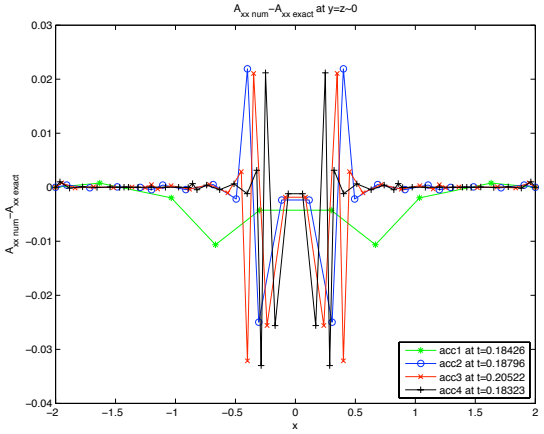
Figure G.14: Pointwise error of ϕ, χ, $\tilde{\Gamma}^x$ and α for 4 types of accuracy for $L = 2$ at $t \sim 0.2M$: 1) acc1 $P = 3,\ N_E = 3^3$; 2) acc2 $P = 5,\ N_E = 5^3$; 3) acc3 $P = 7,\ N_E = 5^3$; 4) acc4 $P = 7,\ N_E = 7^3$



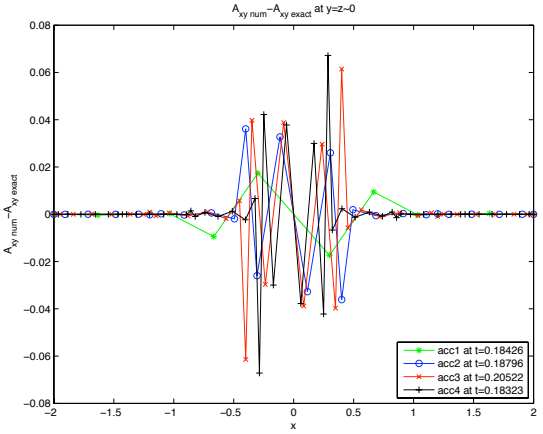
(a) Pointwise error of $\tilde{g}_{xx}$ near the puncture



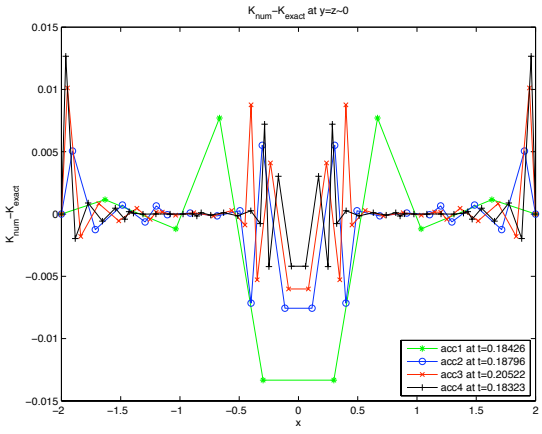
(b) Pointwise error of $\tilde{g}_{xy}$ near the puncture



(c) Pointwise error of $\tilde{A}_{xx}$ near the puncture

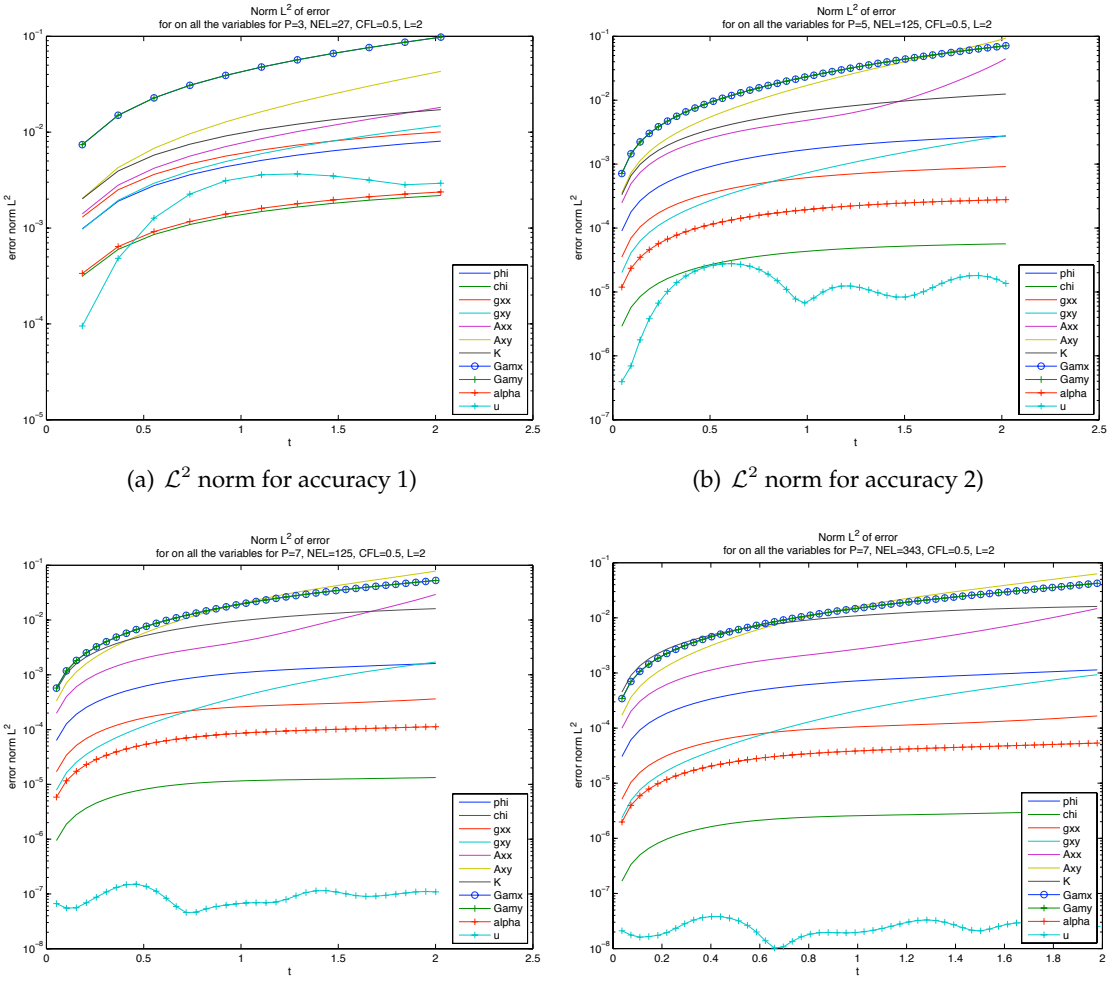


(d) Pointwise error of $\tilde{A}_{xy}$ near the puncture



(e) Pointwise error of K near the puncture

Figure G.15: Same as figure G.14 but for $\tilde{g}_{xx}$, $\tilde{g}_{xy}$, $\tilde{A}_{xx}$ and $\tilde{A}_{xy}$.



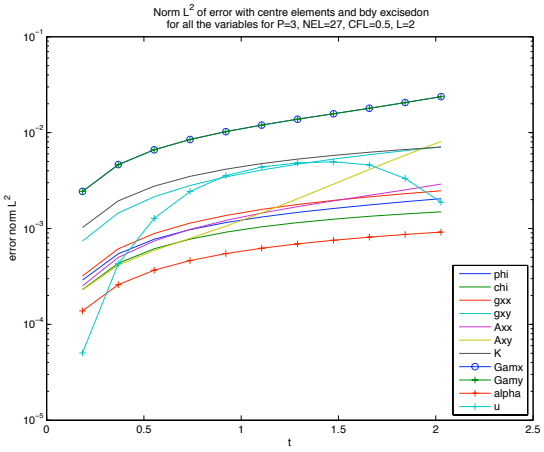
(a) $\mathcal{L}^2$ norm for accuracy 1)

(b) $\mathcal{L}^2$ norm for accuracy 2)

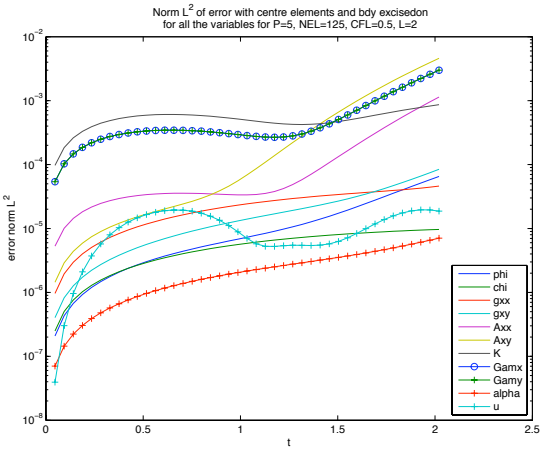
(c) $\mathcal{L}^2$ norm for accuracy 3)

(d) $\mathcal{L}^2$ norm for accuracy 4)

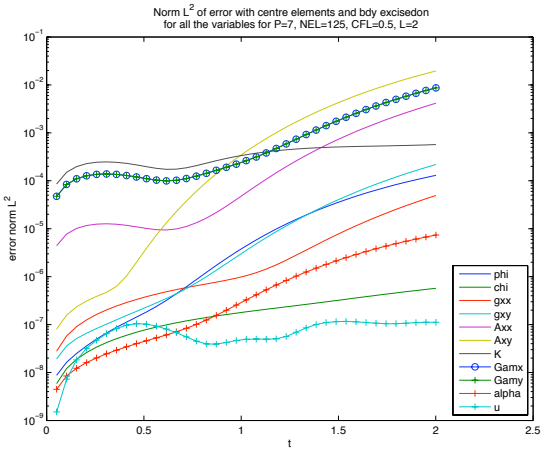
Figure G.16: Comparison of the logarithmic norm $\mathcal{L}^2$ over the entire region of all the variables for 4 types of accuracy with $CFL = 0.5$ and $L = 2$.



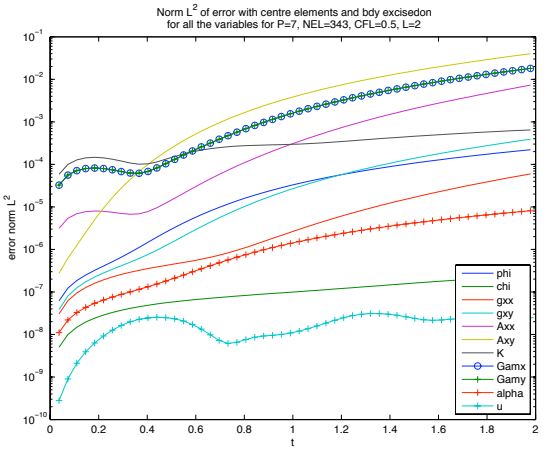
(a) $\mathcal{L}^2$ norm for accuracy 1)



(b) $\mathcal{L}^2$ norm for accuracy 2)



(c) $\mathcal{L}^2$ norm for accuracy 3)

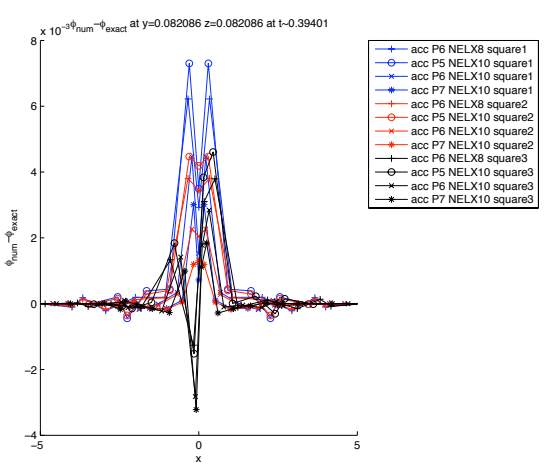


(d) $\mathcal{L}^2$ norm for accuracy 4)

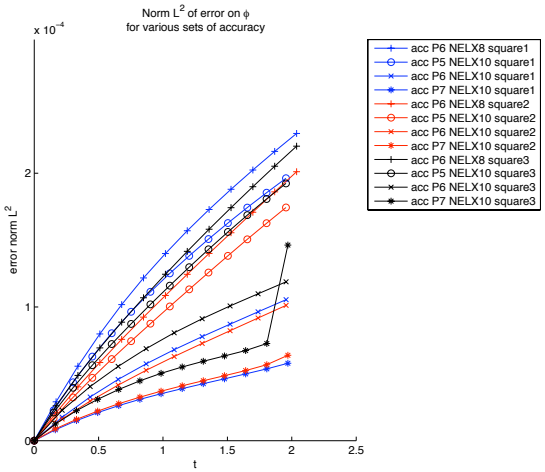
Figure G.17: Same as figure G.16 but with the centre element $(-0.67, 0.67)$ and the boundary $L - 0.5$ excised.

G.4 The offset mesh: The puncture on an edge or face of an element

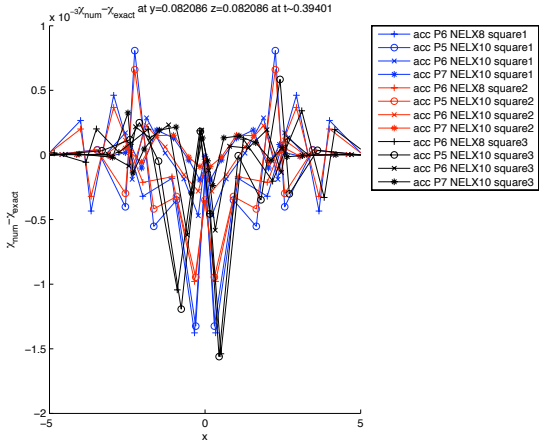
Figures G.18, G.19 and G.20 show the comparison of the 3 different types of offsets (offset 1 in blue, offset 2 in red and offset 3 in black) for the BSSN variables. We see the pointwise error and $\mathcal{L}^2$ norm with increasing accuracy for a domain $L = 64$ with a square mesh.



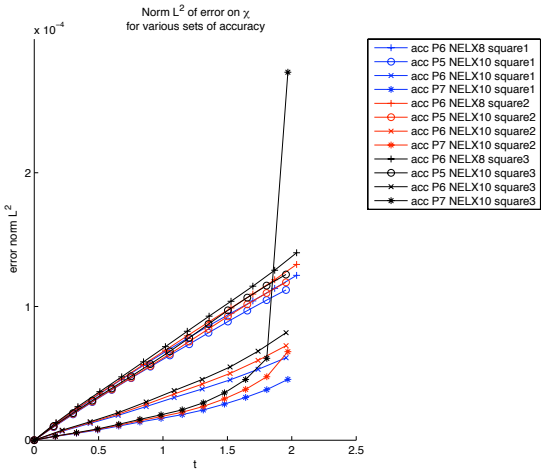
(a) Pointwise error of 3 offsets for ϕ



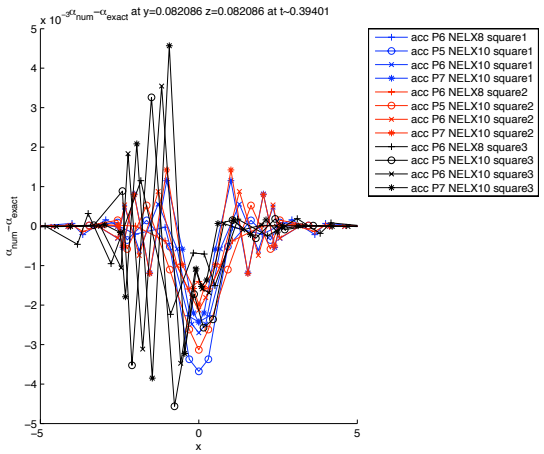
(b) Norm $\mathcal{L}^2$ of 3 offsets for ϕ



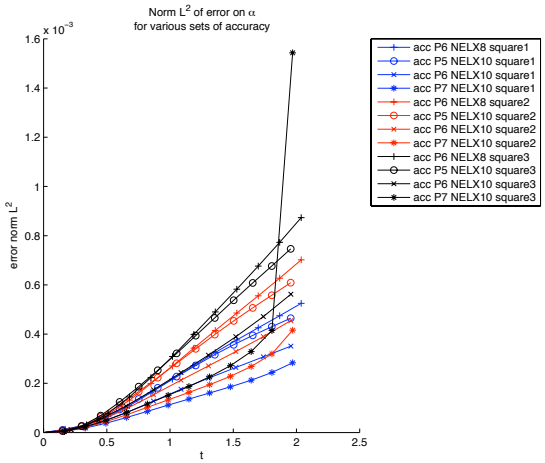
(c) Pointwise error of 3 offsets for χ



(d) Norm $\mathcal{L}^2$ of 3 offsets for χ

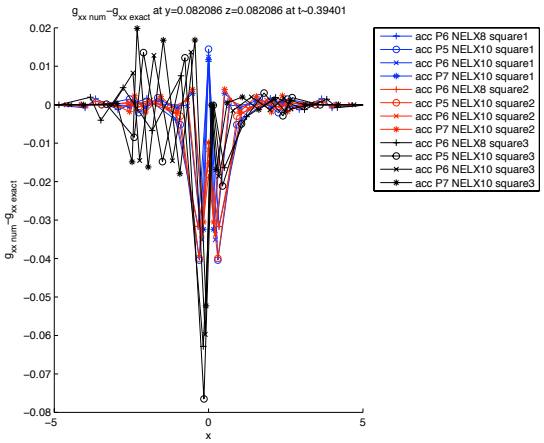


(e) Pointwise error of 3 offsets for α

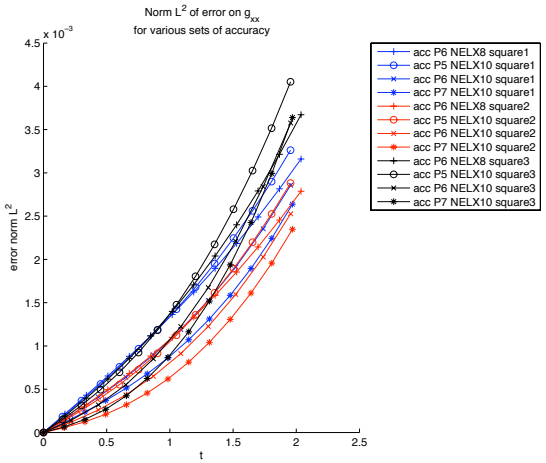


(f) Norm $\mathcal{L}^2$ of 3 offsets for α

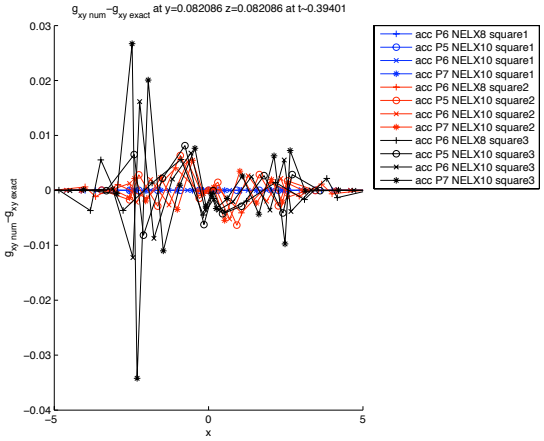
Figure G.18: Comparison of 3 different types of offsets showing the pointwise error and $\mathcal{L}^2$ norm for χ, ϕ, and α, with increasing accuracy for a domain $L = 64$ with a square mesh.



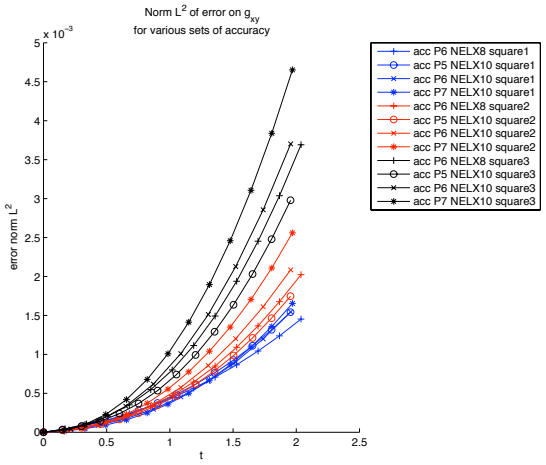
(a) Pointwise error of 3 offsets for $\tilde{g}_{xx}$



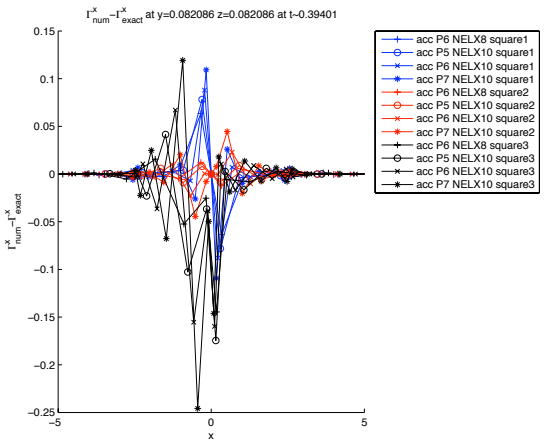
(b) Norm $\mathcal{L}^2$ of 3 offsets for $\tilde{g}_{xx}$



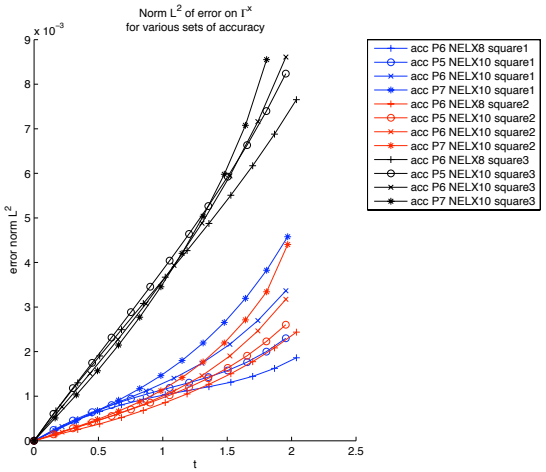
(c) Pointwise error of 3 offsets for $\tilde{g}_{xy}$



(d) Norm $\mathcal{L}^2$ of 3 offsets for $\tilde{g}_{xy}$

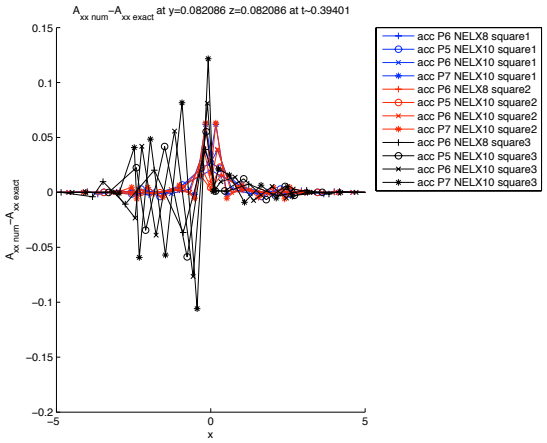


(e) Pointwise error of 3 offsets for $\tilde{\Gamma}^x$

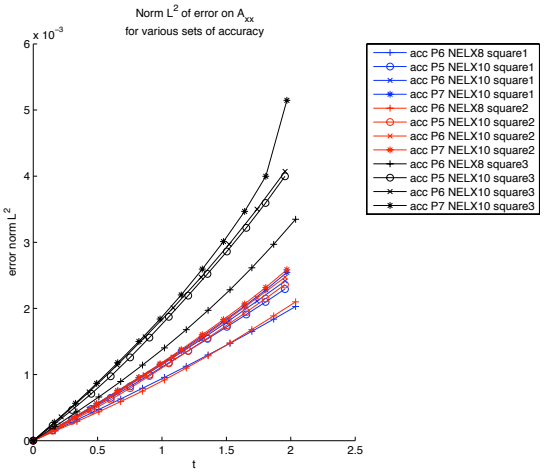


(f) Norm $\mathcal{L}^2$ of 3 offsets for $\tilde{\Gamma}^x$

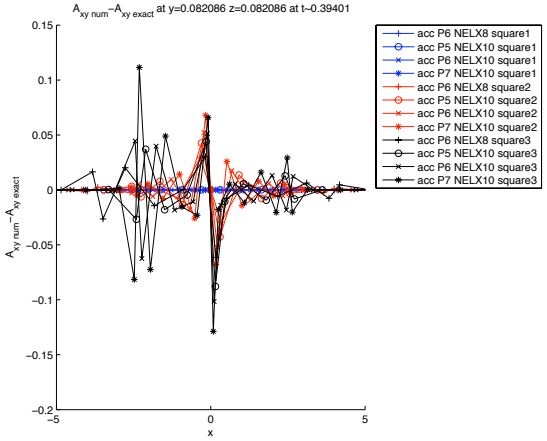
Figure G.19: Same as figure G.18 but for $\tilde{g}_{xx}$, $\tilde{g}_{xy}$ and $\tilde{\Gamma}^x$.



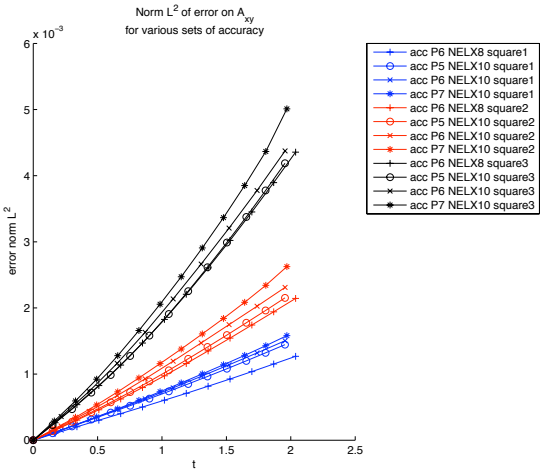
(a) Pointwise error of 3 offsets for $\tilde{A}_{xx}$



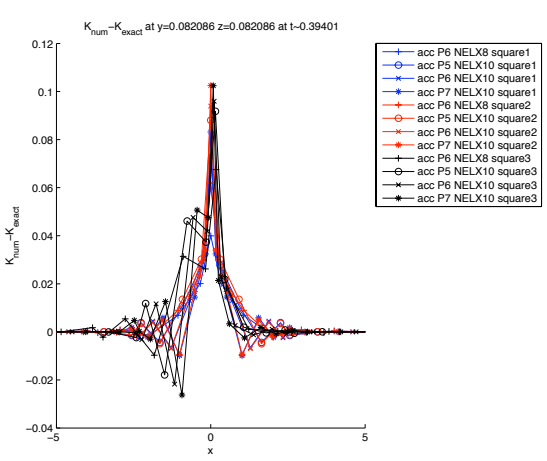
(b) Norm $\mathcal{L}^2$ of 3 offsets for $\tilde{A}_{xx}$



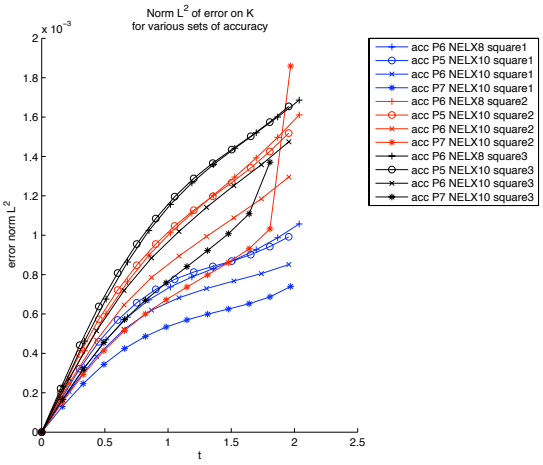
(c) Pointwise error of 3 offsets for $\tilde{A}_{xy}$



(d) Norm $\mathcal{L}^2$ of 3 offsets for $\tilde{A}_{xy}$



(e) Pointwise error of 3 offsets for K



(f) Norm $\mathcal{L}^2$ of 3 offsets for K

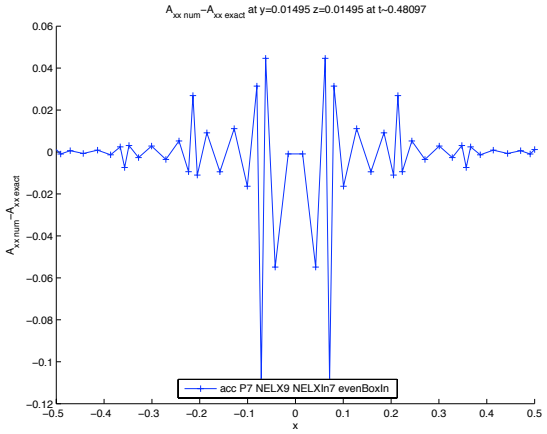
Figure G.20: Same as figure G.18 but for $\tilde{A}_{xx}$, $\tilde{A}_{xy}$ and K.

G.5 Filtering "as much or as little as needed"

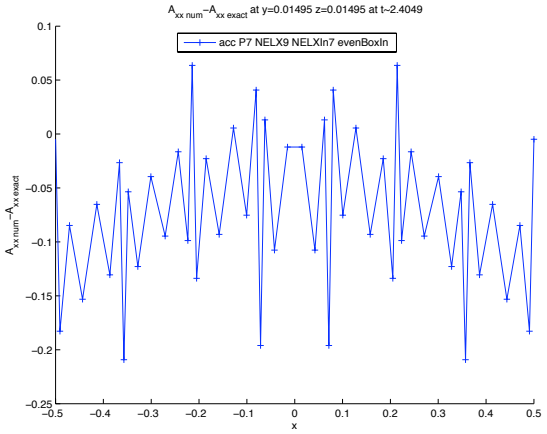
In the following figures, we present the behaviour of $\tilde{A}_{xx}$ and $\tilde{A}_{xy}$ initially and after $t = 2M$, without filtering and with filtering with a cut off value of $N_c = 1$ for the filter in the centre element:

- Figure G.21, shows the pointwise error of $\tilde{A}_{xx}$ close to the puncture without filtering G.21(a), G.21(b), and with filtering G.21(c), G.21(d), for a very small domain of $L = 1$.
- Figure G.22, shows the pointwise error of $\tilde{A}_{xy}$ close to the puncture without filtering G.22(a), G.22(b), and with filtering G.22(c), G.22(d), for a very small domain of $L = 1$.

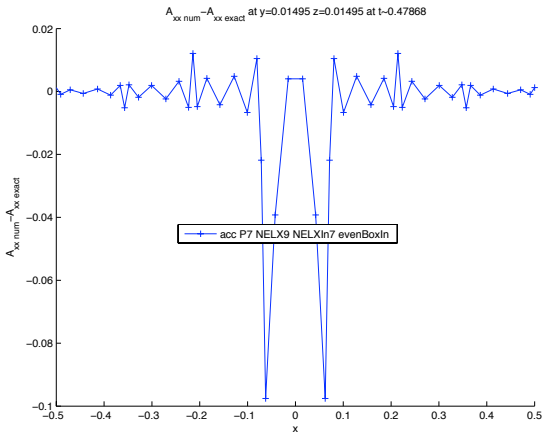
Notice the different scale of the y-axis (errors) on the plots with and without filtering.



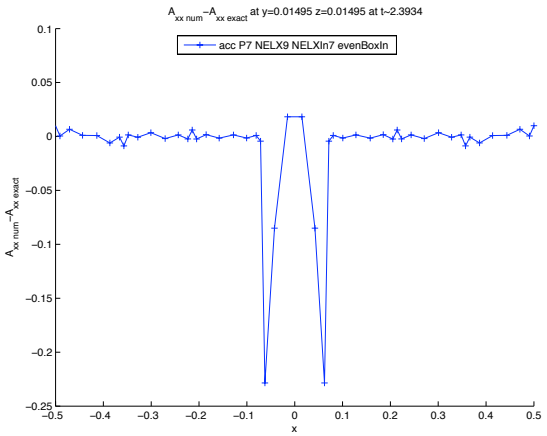
(a) $\tilde{A}_{xx}$ without filtering versus x at $t \sim 0.5$



(b) $\tilde{A}_{xx}$ without filtering versus x at $t \sim 2.4$

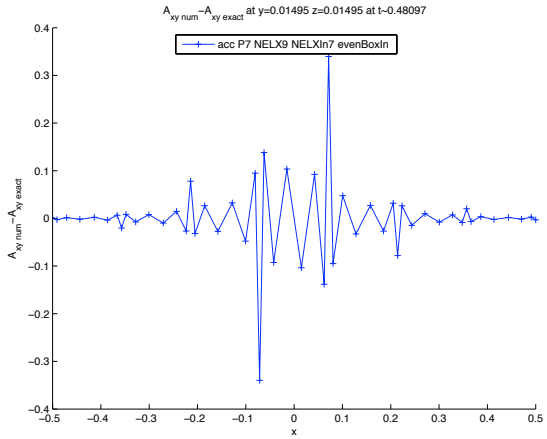


(c) $\tilde{A}_{xx}$ with filtering versus x at $t \sim 0.5$

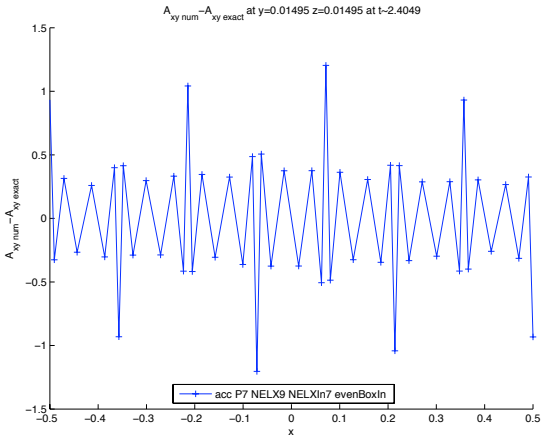


(d) $\tilde{A}_{xx}$ with filtering versus x at $t \sim 2.4$

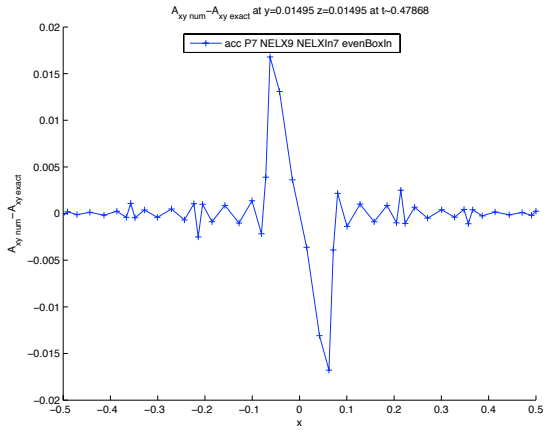
Figure G.21: Pointwise error of $\tilde{A}_{xx}$ close to the puncture (in a 2D slice for $y \sim z = 0.01$M), without filtering (a), (b) and with filtering (c), (d) for a very small domain of $L = 1$. Filtering makes a big difference in stopping the propagation of oscillations throughout the domain.



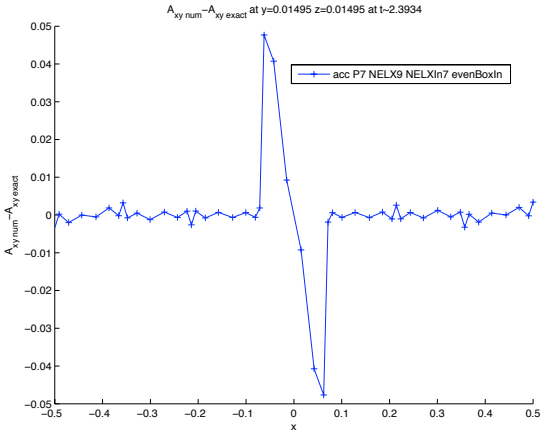
(a) $\tilde{A}_{xy}$ without filtering versus x at $t \sim 0.5$



(b) $\tilde{A}_{xy}$ without filtering versus x at $t \sim 2.4$

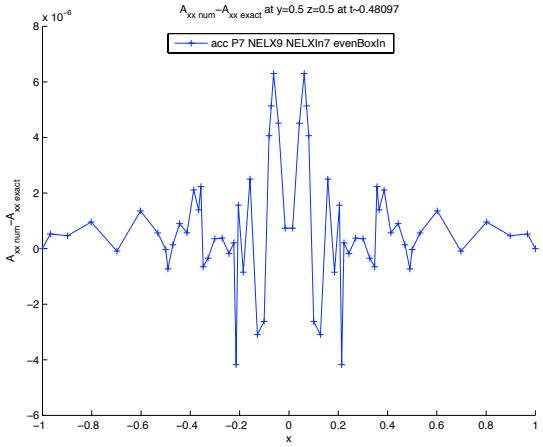


(c) $\tilde{A}_{xy}$ with filtering versus x at $t \sim 0.5$

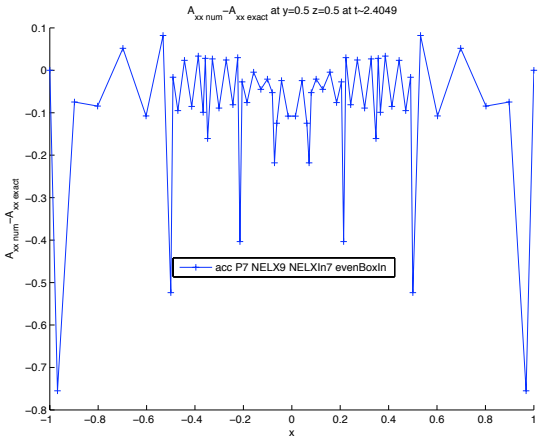


(d) $\tilde{A}_{xy}$ with filtering versus x at $t \sim 2.4$

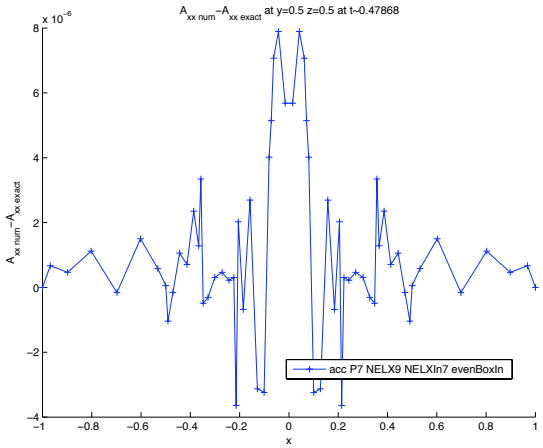
Figure G.22: Same as figure G.21 but for A_{xy}.



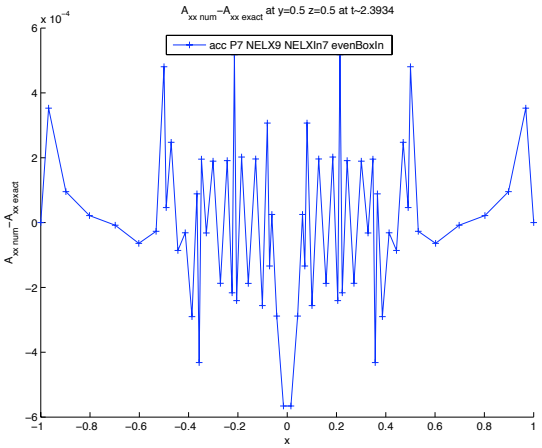
(a) $\tilde{A}_{xx}$ without filtering versus x at $t \sim 0.5$



(b) $\tilde{A}_{xx}$ without filtering versus x at $t \sim 2.4$

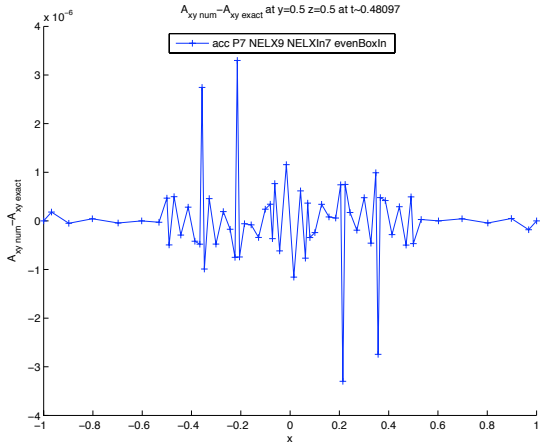


(c) $\tilde{A}_{xx}$ with filtering versus x at $t \sim 0.5$

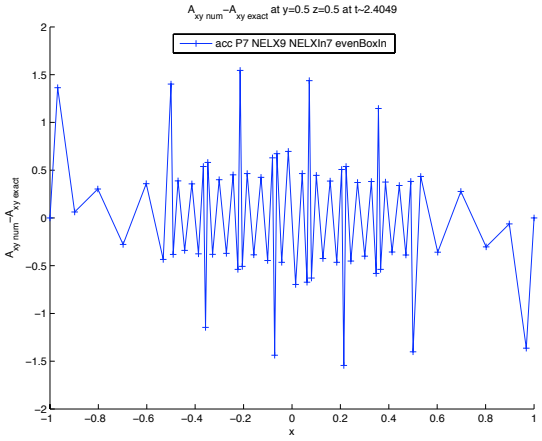


(d) $\tilde{A}_{xx}$ with filtering versus x at $t \sim 2.4$

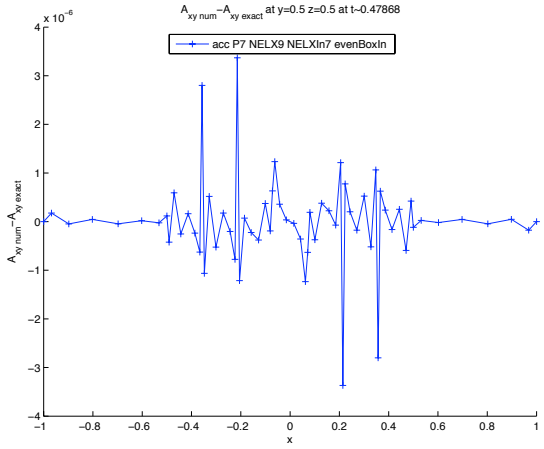
Figure G.23: Same as figure G.21 but further away from the puncture for $y \sim z = 0.5$M.



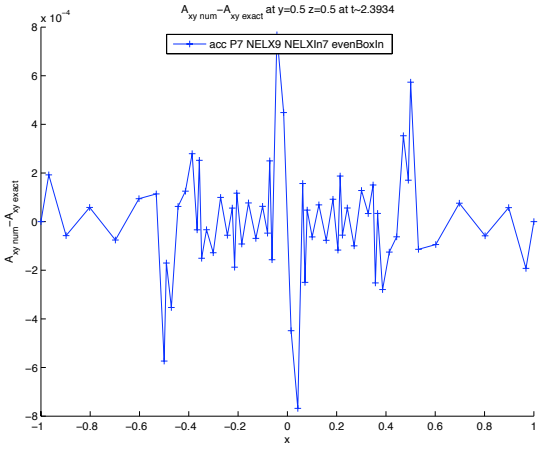
(a) $\tilde{A}_{xy}$ without filtering versus x at $t \sim 0.5$



(b) $\tilde{A}_{xy}$ without filtering versus x at $t \sim 2.4$



(c) $\tilde{A}_{xy}$ with filtering versus x at $t \sim 0.5$



(d) $\tilde{A}_{xy}$ with filtering versus x at $t \sim 2.4$

Figure G.24: Same as figure G.23 but for $\tilde{A}_{xy}$

Céline Nathalie Cattoën

OFFICE ADDRESS:
Victoria University of Wellington
School of Mathematics, Statistics
and Operations Research
P. O. Box 600, Wellington
New Zealand

CONTACT INFORMATION
Phone: +64 4635233 ext 8314
Email: celine.cattoen@mcs.vuw.ac.nz
url:
www.mcs.vuw.ac.nz/research/gg/Celine_Cattoen

PERSONAL INFORMATION

DATE OF BIRTH: January 31, 1981, Muret (31), France
NATIONALITY: French
MARITAL STATUS: Unmarried

EDUCATION

PHD in Mathematics 2006-2009
Victoria University of Wellington, New Zealand
THESIS TITLE: *Applied Mathematics of space-time & space+time: Problems in General Relativity and Cosmology*
SUPERVISOR: Prof. Matt Visser

MASTERS OF SCIENCE in Mathematics with Distinction (A^+) 2005-2006
Victoria University of Wellington, New Zealand
THESIS TITLE: *Cosmological milestones and gravastars – topics in general relativity*
SUPERVISOR: Prof. Matt Visser

MASTERS OF SCIENCE in Mathematical and Modelling Engineering 1999-2004
Institut National des Sciences Appliquées (INSA),
département de génie Mathématique et Modélisation, Toulouse (France)
Five-year French engineering degree, speciality Mathematical Modelling.
EXCHANGE STUDENT in the first half of the fifth year

Massey University, Palmerston North, New Zealand
PROJECT TITLE: *Bacteria-bacteriophage modelling in the cheese industry*
SUPERVISOR: Dr. Geoff Barnes

SIX MONTH INTERNSHIP in the second half of the fifth year

Industrial Research Limited, New Zealand
PROJECT TITLE: *Cryogenic pulse tube modelling*
SUPERVISOR: Dr. Graham Weir

AWARDS & SCHOLARSHIPS

STUDENT PRIZE at the AMSI Workshop on mathematical general relativity 2008

HARTLE PRIZE from the International Society on General Relativity and Gravitation 2007

For one of the best student presentations at the 18th international conference on General Relativity and Gravitation (Sydney, Australia, July 2007)

2006-2009 **VICTORIA UNIVERSITY PHD SCHOLARSHIP AWARD**

2006 **VUW SCIENCE FACULTY STRATEGIC RESEARCH GRANT**
Conference travel support

PUBLICATIONS & PAPERS

Journals

Bounding the Hubble flow in terms of the w parameter.
Céline Cattoën, Matt Visser
Published in **JCAP11(2008) 024**
EPRINT: arXiv:0806.2186

Cosmographic Hubble fits to the supernova data.
Céline Cattoën, Matt Visser
Published in **Phys.Rev.D78:063501,2008**
EPRINT: arXiv:0809.0537 [gr-qc]

Cosmodynamics: Energy conditions, Hubble bounds, density bounds, time and distance bounds.
Céline Cattoën, Matt Visser
Published in **Class.Quant.Grav.25:165013,2008**
EPRINT: arXiv:0712.1619 [gr-qc]

The Hubble series: Convergence properties and redshift variables.
Céline Cattoën, Matt Visser
Published in **Class.Quant.Grav.24:5985-5998,2007**
EPRINT: arXiv:0710.1887 [gr-qc]

Necessary and sufficient conditions for big bangs, bounces, crunches, rips, sudden singularities, and extremality events.
Céline Cattoën, Matt Visser
Published in **Class.Quant.Grav.22:4913-4930,2005**
EPRINT: arXiv:gr-qc/0508045

Gravastars must have anisotropic pressures.
Céline Cattoën, Matt Visser
Published in **Class.Quant.Grav.22:4189-4202,2005**
EPRINT: arXiv:gr-qc/0505137

Effective refractive index tensor for weak field gravity.
Petarpa Boonserm, Céline Cattoën, Tristan Faber, Matt Visser,
Silke Weinfurtner
Published in **Class.Quant.Grav.22:1905-1916,2005**
EPRINT: arXiv:gr-qc/0411034

Conference Proceedings

Generalized Puiseux series expansion for cosmological milestones.
Céline Cattoën, Matt Visser
Published in the **Proceedings of Eleventh Marcel Grossmann Meeting on General Relativity**, (World Scientific Singapore, 2008), pp 2057–2059
Edited by H. Kleinert, R. T. Jantzen, and R. Ruffini.
EPRINT: arXiv:gr-qc/0609073

Cosmological milestones and energy conditions.
Céline Cattoën, Matt Visser
Published in **J.Phys.Conf.Ser.68:012011,2007**
EPRINT: arXiv:gr-qc/0609064

Masters Thesis

Cosmological milestones and gravastars: Topics in general relativity.
Céline Cattoën
EPRINT: arXiv:gr-qc/0606011

Technical report

Cosmography: Extracting the Hubble series from the supernova data.
Céline Cattoën, Matt Visser
EPRINT: arXiv:gr-qc/0703122

Article in preparation

Black Hole puncture simulations with the spectral element method. Céline Cattoën, Mark Hannam

CONFERENCES & WORKSHOPS

AMSI WORKSHOP ON MATHEMATICAL GENERAL RELATIVITY Jul 7-9, 2008
Australian Mathematical Sciences Institute, Melbourne
CONTRIBUTED TALK: *Cosmodynamics: The Hubble flow in a FLRW universe*

JOINT MEETING OF THE AMS - NZMS 2007 Dec 12-15, 2007
The first Joint Meeting of the American Mathematical Society and the New Zealand Mathematical Society
Victoria University of Wellington, New Zealand
CONTRIBUTED TALK:
Cosmography: Extracting the Hubble Series from the Supernova Data

NEW ZEALAND MATHEMATICS AND STATISTICS POSTGRADUATE CONFERENCE (NZ-MASP) Nov 22-23, 2007
Queenstown, New Zealand
CONTRIBUTED TALK:
Cosmography: Extracting the Hubble Series from the Supernova Data

Jul 8-14, 2007 **18TH INTERNATIONAL CONFERENCE ON GENERAL RELATIVITY AND GRAVITATION (GR18)**
Sydney, Australia
CONTRIBUTED TALK:
New versions of the Hubble law

Jun 2007 **NEW ZEALAND INSTITUTE OF PHYSICS CONFERENCE**
University of Otago, Dunedin, New Zealand
CONTRIBUTED TALK:
New versions of the Hubble law

Feb 7-8, 2007 **NEW ZEALAND KOREA GRAVITY WORKSHOP**
Christchurch, New Zealand
CONTRIBUTED TALK:
Cosmography: Extracting the Hubble Series from the Supernova Data

Jul 23-29 2006 **11TH MARCEL GROSSMANN MEETING (MG11)**
On Recent Developments in Theoretical and Experimental General Relativity, Gravitation, and Relativistic Field Theories
Berlin, Germany
CONTRIBUTED TALK:
Necessary and sufficient conditions for big bangs, bounces, crunches, rips, sudden singularities and more

Jul 17-2, 2006 **NEW FRONTIERS IN NUMERICAL RELATIVITY CONFERENCE**
Albert-Einstein-Institut, Potsdam, Germany

Jun 29-2, 2006 **12TH CONFERENCE ON RECENT DEVELOPMENTS IN GRAVITY (NEB XII)**
Nafplio, Greece
CONTRIBUTED TALK:
Necessary and sufficient conditions for big bangs, bounces, crunches, rips, sudden singularities and more

Bibliography

[1] SNLS, "Super Nova Legacy Survey,". http://snls.in2p3.fr/conf/papers/cosmo1/. Electronic data available at given url.

[2] **Supernova Search Team** Collaboration, A. G. Riess *et al.*, "Type Ia Supernova Discoveries at $z > 1$ From the Hubble Space Telescope: Evidence for Past Deceleration and Constraints on Dark Energy Evolution," *Astrophys. J.* **607** (2004) 665–687, arXiv:astro-ph/0402512. http://braeburn.pha.jhu.edu//~ariess/R06/. Electronic data available at given url.

[3] A. G. Riess *et al.*, "New Hubble Space Telescope Discoveries of Type Ia Supernovae at $z > 1$: Narrowing Constraints on the Early Behavior of Dark Energy," *Astrophys. J.* **659** (2007) 98–121, arXiv:astro-ph/0611572.

[4] M. Hannam, S. Husa, U. Sperhake, B. Bruegmann, and J. A. Gonzalez, "Where post-Newtonian and numerical-relativity waveforms meet," *Phys. Rev.* **D77** (2008) 044020, arXiv:0706.1305 [gr-qc].

[5] M. Hannam, S. Husa, D. Pollney, B. Bruegmann, and N. O'Murchadha, "Geometry and Regularity of Moving Punctures," *Phys. Rev. Lett.* **99** (2007) 241102, arXiv:gr-qc/0606099.

[6] M. Hannam, S. Husa, F. Ohme, B. Bruegmann, and N. O'Murchadha, "Wormholes and trumpets: the Schwarzschild spacetime for the moving-puncture generation," *Phys. Rev.* **D78** (2008) 064020, arXiv:0804.0628 [gr-qc].

[7] R. D'Inverno, *Introducing Einstein's relativity*. Oxford University Press, 1992.

[8] F. Hoyle and M. S. Vogeley, "Voids in the PSCz Survey and the Updated Zwicky Catalog," *Astrophys. J.* **566** (2002) 641. [arXiv:astro-ph/0109357].

[9] F. Hoyle and M. S. Vogeley, "Voids in the 2dF Galaxy Redshift Survey," *Astrophys. J.* **607** (2004) 751–764. [arXiv:astro-ph/0312533].

[10] J. R. I. Gott, M. Jurić, D. Schegel, F. Hoyle, M. Vogeley, M. Tegmark, N. Bahcall, and J. Brinkmann, "A Map of the Universe," *Astrophys. J.* **624** (2005) 463. [arXiv:astro-ph/0310571].

[11] **WMAP** Collaboration, E. Komatsu *et al.*, "Five-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations:Cosmological Interpretation," *Astrophys. J. Suppl.* **180** (2009) 330–376, arXiv:0803.0547 [astro-ph].

[12] D. Rapetti, S. W. Allen, M. A. Amin, and R. D. Blandford, "A kinematical approach to dark energy studies," *Mon. Not. Roy. Astron. Soc.* **375** (2007) 1510–1520, arXiv:astro-ph/0605683.

[13] R. D. Blandford, M. A. Amin, E. A. Baltz, K. Mandel, and P. J. Marshall, "Cosmokinetics," arXiv:astro-ph/0408279.

[14] C. Shapiro and M. S. Turner, "What Do We Really Know About Cosmic Acceleration?," *Astrophys. J.* **649** (2006) 563–569, arXiv:astro-ph/0512586.

[15] R. R. Caldwell and M. Kamionkowski, "Expansion, Geometry, and Gravity," *JCAP* **0409** (2004) 009, arXiv:astro-ph/0403003.

[16] O. Elgaroy and T. Multamaki, "Bayesian analysis of Friedmannless cosmologies," *JCAP* **0609** (2006) 002, arXiv:astro-ph/0603053.

[17] W.-M. Yao *et al.*, "Review of Particle Properties," *Journal of Physics G* **33** (2006) no. 1, . http://pdg.lbl.gov/. Online version given at given url.

[18] J. M. Virey *et al.*, "On the determination of the deceleration parameter from Supernovae data," *Phys. Rev.* **D72** (2005) 061302. [arXiv:astro-ph/0502163].

[19] S. Eidelman *et al.*, " Review of Particle Physics," *Physics Letters B* **592** (2004) . http://pdg.lbl.gov. Url.

[20] S. Weinberg, "Gravitation and cosmology: Principles and applications of the general theory of relativity," *Wiley, New York* (1972) .

[21] P. J. E. Peebles, "Principles of physical cosmology," *Princeton University Press* (1993) .

[22] T. Chiba and T. Nakamura, "The luminosity distance, the equation of state, and the geometry of the universe," *Prog. Theor. Phys.* **100** (1998) 1077–1082, arXiv:astro-ph/9808022.

[23] V. Sahni, T. D. Saini, A. A. Starobinsky, and U. Alam, "Statefinder – a new geometrical diagnostic of dark energy," *JETP Lett.* **77** (2003) 201–206, arXiv:astro-ph/0201498.

[24] M. Visser, "Jerk and the cosmological equation of state," *Class. Quant. Grav.* **21** (2004) 2603–2616, arXiv:gr-qc/0309109.

[25] M. Visser, "Cosmography: Cosmology without the Einstein equations," *Gen. Rel. Grav.* **37** (2005) 1541–1548, arXiv:gr-qc/0411131.

[26] D. W. Hogg, "Distance measures in cosmology," arXiv:astro-ph/9905116.

[27] G. F. R. Ellis and T. Rothman, "Lost horizons," *American Journal of Physics* **61** (1993) 883–893.

[28] **The SNLS** Collaboration, P. Astier *et al.*, "The Supernova Legacy Survey: Measurement of Ω_M, Ω_Λ and w from the First Year Data Set," *Astron. Astrophys.* **447** (2006) 31–48, arXiv:astro-ph/0510447.

[29] S. Nesseris and L. Perivolaropoulos, "Tension and Systematics in the Gold06 SnIa Dataset," *JCAP* **0702** (2007) 025, arXiv:astro-ph/0612653.

[30] T. R. Choudhury and T. Padmanabhan, "A theoretician's analysis of the supernova data and the limitations in determining the nature of dark energy II: Results for latest data," *Astron. Astrophys.* **429** (2005) 807, arXiv:astro-ph/0311622.

[31] E. P. Hubble, "A relation between distance and radial velocity among extra-galactic nebulae," *Proc. Natl. Acad. Sci. USA* **15** (1929) 168–173.

[32] R. P. Kirshner, "Hubble's diagram and cosmic expansion," *Proc. Natl. Acad. Sci. USA* **101** (2004) 8–13.

[33] M. Visser, *Lorentzian wormholes: From Einstein to Hawking*. No. 412 p. Woodbury, USA, 1995.

[34] S. W. Hawking and G. F. R. Ellis, "The Large scale structure of space-time,". Cambridge University Press, Cambridge, 1973.

[35] R. M. Wald, *General Relativity*. Chicago, Usa: Univ. Pr., 1984.

[36] B. Rose, "A matter model violating the strong energy conditions. The influence of temperature," *Class. Quant. Grav.* **4** (1987) 1019–1030.

[37] B. Rose, "Construction of matter models which violate the strong energy condition and may avoid the initial singularity," *Class. Quant. Grav.* **3** (1986) 975–995.

[38] S. Jha, A. G. Riess, and R. P. Kirshner, "Improved Distances to Type Ia Supernovae with Multicolor Light Curve Shapes: MLCS2k2," *Astrophys. J.* **659** (2007) 122–148, arXiv:astro-ph/0612666. http://astro.berkeley.edu//~saurabh/mlcs2k2. Electronic data available at given url.

[39] **ESSENCE** Collaboration, W. M. Wood-Vasey *et al.*, "Observational Constraints on the Nature of the Dark Energy: First Cosmological Results from the ESSENCE Supernova Survey," *Astrophys. J.* **666** (2007) 694–715, arXiv:astro-ph/0701041.

[40] H. K. Jassal, J. S. Bagla, and T. Padmanabhan, "Observational constraints on low redshift evolution of dark energy: How consistent are different observations?," *Phys. Rev.* **D72** (2005) 103503, arXiv:astro-ph/0506748.

[41] H. K. Jassal, J. S. Bagla, and T. Padmanabhan, "The vanishing phantom menace," arXiv:astro-ph/0601389.

[42] T. Padmanabhan and T. R. Choudhury, "A theoretician's analysis of the supernova data and the limitations in determining the nature of dark energy," *Mon. Not. Roy. Astron. Soc.* **344** (2003) 823–834, arXiv:astro-ph/0212573.

[43] V. Barger, Y. Gao, and D. Marfatia, "Accelerating cosmologies tested by distance measures," *Phys. Lett.* **B648** (2007) 127–132, arXiv:astro-ph/0611775.

[44] M. Chevallier and D. Polarski, "Accelerating universes with scaling dark matter," *Int. J. Mod. Phys.* **D10** (2001) 213–224, arXiv:gr-qc/0009008.

[45] E. V. Linder, "Probing gravitation, dark energy, and acceleration," *Phys. Rev.* **D70** (2004) 023511, arXiv:astro-ph/0402503.

[46] E. V. Linder, "Biased Cosmology: Pivots, Parameters, and Figures of Merit," *Astropart. Phys.* **26** (2006) 102–110, arXiv:astro-ph/0604280.

[47] B. A. Bassett, P. S. Corasaniti, and M. Kunz, "The essence of quintessence and the cost of compression," *Astrophys. J.* **617** (2004) L1–L4, arXiv:astro-ph/0407364.

[48] D. Martin and A. Albrecht, "Talk about pivots," arXiv:astro-ph/0604401.

[49] I. Zehavi, A. G. Riess, R. P. Kirshner, and A. Dekel, "A Local Hubble Bubble from SNe Ia?," *Astrophys. J.* **503** (1998) 483, arXiv:astro-ph/9802252.

[50] R. Giovanelli, D. Dale, M. Haynes, E. Hardy, and L. Campusano, "No Hubble Bubble in the Local Universe," arXiv:astro-ph/9906362.

[51] P. R. Bevington, *Data reduction and analysis in the physical sciences*. McGraw–Hill, New York,, 1969.

[52] N. R. Draper and H. Smith, *Applied regression analysis*. Wiley, New York, 1998.

[53] E. A. P. D. C. Montgomery and G. G. Vining, *Introduction to linear regression analysis*. Wiley, New York, 2001.

[54] R. D. Cook and S. Weisberg, *Applied regression including computing and graphics*. Wiley, New York, 1999.

[55] J. a. J. E. G. W. J. Kennedy, *Statistical computing*. Marcel Dekker, New York, 1980.

[56] R. J. Carroll and D. Rupert, *Transformation and weighting in regression*. Chapman and Hall, London, 1998.

[57] G. J. S. Ross, *Nonlinear estimation*. Spinger–Verlag, New York, 1990.

[58] B. N. Taylor and C. E. Kuyatt, "Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results," *NIST Technical Note* no. 1297, . http://physics.nist.gov/cuu/Uncertainty/index.html. Online version given at given url.

[59] M. Hicken *et al.*, "Improved Dark Energy Constraints from 100 New CfA Supernova Type Ia Light Curves," *Astrophys. J.* **700** (2009) 1097–1140, arXiv:0901.4804 [astro-ph.CO].

[60] A. Signal Private communication.

[61] J. Jonsson, A. Goobar, R. Amanullah, and L. Bergstrom, "No evidence for Dark Energy Metamorphosis ?," *JCAP* **0409** (2004) 007, arXiv:astro-ph/0404468.

[62] S. L. Bridle, O. Lahav, J. P. Ostriker, and P. J. Steinhardt, "Precision Cosmology? Not Just Yet," *Science.* **299** (2003) 1532, arXiv:astro-ph/0303180.

[63] M. Visser, "Energy conditions in the epoch of galaxy formation," *Science* **276** (1997) 88–90.

[64] M. Visser, "General Relativistic Energy Conditions: The Hubble expansion in the epoch of galaxy formation," *Phys. Rev.* **D56** (1997) 7578–7587, arXiv:gr-qc/9705070.

[65] M. Visser, "Energy conditions and galaxy formation," arXiv:gr-qc/9710010.

[66] C. Molina-Paris and M. Visser, "Minimal conditions for the creation of a Friedman-Robertson-Walker universe from a 'bounce'," *Phys. Lett.* **B455** (1999) 90–95, arXiv:gr-qc/9810023.

[67] D. Hochberg, C. Molina-Paris, and M. Visser, "Tolman wormholes violate the strong energy condition," *Phys. Rev.* **D59** (1999) 044011, arXiv:gr-qc/9810029.

[68] C. Cattoën and M. Visser, "Necessary and sufficient conditions for big bangs, bounces, crunches, rips, sudden singularities, and extremality events," *Class. Quant. Grav.* **22** (2005) 4913–4930, arXiv:gr-qc/0508045.

[69] C. Cattoën and M. Visser, "Cosmological milestones and energy conditions," *J. Phys. Conf. Ser.* **68** (2007) 012011, arXiv:gr-qc/0609064.

[70] C. Cattoën and M. Visser, "Generalized Puisieux series expansion for cosmological milestones," arXiv:gr-qc/0609073.

[71] J. Santos, J. S. Alcaniz, and M. J. Reboucas, "Energy Conditions and Supernovae Observations," *Phys. Rev.* **D74** (2006) 067301, arXiv:astro-ph/0608031.

[72] J. Santos, J. S. Alcaniz, M. J. Reboucas, and N. Pires, "Lookback time bounds from energy conditions," *Phys. Rev.* **D76** (2007) 043519, arXiv:0706.1779 [astro-ph].

[73] J. Santos, J. S. Alcaniz, N. Pires, and M. J. Reboucas, "Energy Conditions and Cosmic Acceleration," *Phys. Rev.* **D75** (2007) 083523, arXiv:astro-ph/0702728.

[74] M. Visser, "Scale anomalies imply violation of the averaged null energy condition," *Phys. Lett.* **B349** (1995) 443–447, arXiv:gr-qc/9409043.

[75] M. Visser, "Gravitational vacuum polarization I: Energy conditions in the Hartle–Hawking vacuum," *Phys. Rev.* **D54** (1996) 5103–5115, arXiv:gr-qc/9604007.

[76] M. Visser, "Gravitational vacuum polarization II: Energy conditions in the Boulware vacuum," *Phys. Rev.* **D54** (1996) 5116–5122, arXiv:gr-qc/9604008.

[77] M. Visser, "Gravitational vacuum polarization. IV: Energy conditions in the Unruh vacuum," *Phys. Rev.* **D56** (1997) 936–952, arXiv:gr-qc/9703001.

[78] M. Visser, "Gravitational vacuum polarization," arXiv:gr-qc/9710034.

[79] M. Visser and C. Barcelo, "Energy conditions and their cosmological implications," arXiv:gr-qc/0001099.

[80] C. Barcelo and M. Visser, "Twilight for the energy conditions?," *Int. J. Mod. Phys.* **D11** (2002) 1553–1560, arXiv:gr-qc/0205066.

[81] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation*. Freeman, San Francisco, 1972.

[82] R. M. Wald, *General relativity*. Chicago University Press, 1984.

[83] S. M. Carroll, *Spacetime and geometry: An introduction to general relativity*. Addison–Wesley, San Francisco, 2004.

[84] J. B. Hartle, *Gravity: An introduction to Einstein's general relativity*. Addison–Wesley, San Francisco, 2003.

[85] C. Cattoën and M. Visser, "Cosmography: Extracting the Hubble series from the supernova data," arXiv:gr-qc/0703122.

[86] C. Cattoën and M. Visser, "The Hubble series: Convergence properties and redshift variables," *Class. Quant. Grav.* **24** (2007) 5985–5998, arXiv:0710.1887 [gr-qc].

[87] M. Seikel and D. J. Schwarz, "How strong is the evidence for accelerated expansion?," *JCAP* **0802** (2008) 007, arXiv:0711.3180 [astro-ph].

[88] M. Visser, *Lorentzian wormholes: From Einstein to Hawking*. AIP Press/Springer Verlag, New York, 1995.

[89] **Particle Data Group** Collaboration, S. Eidelman *et al.*, "Review of particle physics," *Phys. Lett.* **B592** (2004) 1.

[90] C. Cattoën and M. Visser, "Cosmodynamics: Energy conditions, Hubble bounds, density bounds, time and distance bounds," *Class. Quant. Grav.* **25** (2008) 165013, arXiv:0712.1619 [gr-qc].

[91] S. G. Hahn and R. W. Lindquist, "The two body problem in geometrodynamics.," *Ann. Phys.* **29** (1964) 304–331.

[92] F. Pretorius, "Evolution of Binary Black Hole Spacetimes," *Phys. Rev. Lett.* **95** (2005) 121101, arXiv:gr-qc/0507014.

[93] M. Campanelli, C. O. Lousto, P. Marronetti, and Y. Zlochower, "Accurate Evolutions of Orbiting Black-Hole Binaries Without Excision," *Phys. Rev. Lett.* **96** (2006) 111101, arXiv:gr-qc/0511048.

[94] J. G. Baker, J. Centrella, D.-I. Choi, M. Koppitz, and J. van Meter, "Gravitational wave extraction from an inspiraling configuration of merging black holes," *Phys. Rev. Lett.* **96** (2006) 111102, arXiv:gr-qc/0511103.

[95] S. L. Shapiro, "Numerical Relativity at the Frontier," *Prog. Theor. Phys. Suppl.* **163** (2006) 100–119, arXiv:gr-qc/0509094.

[96] M. Hannam *et al.*, "The Samurai Project: verifying the consistency of black-hole-binary waveforms for gravitational-wave detection," arXiv:0901.2437 [gr-qc].

[97] R. Arnowitt, S. Deser, and C. W. Misner, *Gravitation: An Introduction to Current Research*. L. Witten, New York, 1962.

[98] J. W. York, *Sources of Gravitational Radiation*. Cambridge University Press, Cambridge, UK, 1979.

[99] S. Frittelli, "Note on the propagation of the constraints in standard (3+1) general relativity," *Phys. Rev.* **D55** (1997) 5992–5996.

[100] J. L. H.-O. Kreiss, "Initial-Boundary Value Problems and the Navier-Stokes Equations," *Academic Press, Boston* (1989) .

[101] O. A. Reula, "Hyperbolic methods for Einstein's equations," *Living Rev. Rel.* **1** (1998) 3.

[102] C. Bona and J. Massó, "Einstein's evolution equations as a system of balance laws,"*Phys. Rev. D* **40** (Aug, 1989) 1022.

[103] C. Bona and J. Massó, "Hyperbolic system for Numerical Relativity," *Phys. Rev. Lett* **68** (1992) 1097.

[104] E. S. C. Bona, Massó J. and J. Stela, "A new formlaism for numerical relativity," *Phys. Rev. Lett* **75** (1995) 600.

[105] T. W. Baumgarte and S. L. Shapiro, "On the numerical integration of Einstein's field equations," *Phys. Rev.* **D59** (1999) 024007, arXiv:gr-qc/9810065.

[106] M. Shibata and T. Nakamura, "Evolution of three-dimensional gravitational waves: Harmonic slicing case," *Phys. Rev.* **D52** (1995) 5428–5444.

[107] O. Sarbach, G. Calabrese, J. Pullin, and M. Tiglio, "Hyperbolicity of the BSSN system of Einstein evolution equations," *Phys. Rev.* **D66** (2002) 064002, arXiv:gr-qc/0205064.

[108] H. R. Beyer and O. Sarbach, "On the well posedness of the Baumgarte-Shapiro-Shibata-Nakamura formulation of Einstein's field equations," *Phys. Rev.* **D70** (2004) 104004, arXiv:gr-qc/0406003.

[109] J. York, James W., "Gravitational degrees of freedom and the initial-value problem," *Phys. Rev. Lett.* **26** (1971) 1656–1658.

[110] A. Lichnerowicz *J. Math. Pure Appl* **23** (1944) 37.

[111] J. Thornburg, "Coordinates and boundary conditions for the general relativistic initial data problem," *Classical and Quantum Gravity* **4** (1987) no. 5, 1119–1131. http://stacks.iop.org/0264-9381/4/1119.

[112] S. Brandt and B. Bruegmann, "Black hole punctures as initial data for general relativity," *Phys. Rev. Lett.* **78** (1997) 3606–3609, arXiv:gr-qc/9703066.

[113] S. R. Brandt and B. Bruegmann, "BH punctures as initial data for general relativity," arXiv:gr-qc/9711015.

[114] P. Diener *et al.*, "Accurate evolution of orbiting binary black holes," *Phys. Rev. Lett.* **96** (2006) 121101, arXiv:gr-qc/0512108.

[115] J. G. Baker, J. Centrella, D.-I. Choi, M. Koppitz, and J. van Meter, "Gravitational wave extraction from an inspiraling configuration of merging black holes," *Physical Review Letters* **96** (2006) 111102. doi:10.1103/PhysRevLett.96.111102.

[116] M. Alcubierre and J. Massó, "Pathologies of hyperbolic gauges in general relativity and other field theories,"*Phys. Rev. D* **57** (Apr, 1998) R4511–R4515.

[117] M. Alcubierre, "Appearance of coordinate shocks in hyperbolic formalisms of general relativity,"*Phys. Rev. D* **55** (May, 1997) 5981–5991.

[118] M. Alcubierre *et al.*, "Gauge conditions for long-term numerical black hole evolutions without excision," *Phys. Rev.* **D67** (2003) 084023, arXiv:gr-qc/0206072.

[119] P. Grandclement and J. Novak, "Spectral Methods for Numerical Relativity," arXiv:0706.2286 [gr-qc].

[120] J. P. Boyd, *Chebyshev and Fourier spectral methods*. Courier Dover Publications, Second Edition, Illustrated, Revised ed., 2001.

[121] Maday and Patera, "Spectral element methods for the incompressible Navier-Stokes equations," *New York, American Society of Mechanical Engineers* **A90-47176 21-64** (1989) . http://adsabs.harvard.edu/abs/1989sasc.proc...71M.

[122] N. Bodard, "Interaction fluide-structure par la méthode des éléments spectraux," *Thèse EPFL* **no 3503** (2006) , http://library.epfl.ch/theses/?nr=3503. published thesis.

[123] I. Barosan, "Adaptive Spectral Elements for Diffuse Interface Multi-Fluid Flow," *Technische Universiteit Eindhoven* (2003) . PhD thesis, Advisors: H.E.H. Meijer, P.A.J. Hilbers, Co-advisor: P.D. Anderson.

[124] E. Chaljub, "Modélisation numérique de la propagation d'ondes sismiques en géométrie sphérique: application à la sismologie globale (Numerical modeling of the propagation of seismic waves in spherical geometry: applications to global seismology),". PhD thesis, Université Paris VII Denis Diderot, Paris, France.

[125] M. A. J. Casarin, "Schwarz Preconditioners for Spectral and Mortar Finite Element Methods with Applications to Incompressible Fluids," http://lciteseer.ist.psu.edu/casarin96schwarz.html. PhD thesis, Université Paris VII Denis Diderot, Paris, France.

[126] P. D. Lax and A. N. Milgram, ""Parabolic equations". Contributions to the theory of partial differential equations.," *Annals of Mathematics Studies* **33** (1954) 167–190.

[127] I. Babuška, "Error-bounds for finite element method," *Numerische Mathematik* **16** (January, 1971) no. 4, 322–333.

[128] D. Drivaliaris and N. Yannakakis, "Generalizations of the Lax-Milgram theorem,". doi:10.1155/2007/87104.

[129] A. Iserles, *A first course in the numerical analysis of differential equations*. isbn 0-521-55655-4. Cambridge University Press, New York, USA., (1996).

[130] P. F. Duane Rosenberg, Aime' Fournier and A. Pouquet, "Geophysical-astrophysical spectral-element adaptive refinement (GASpAR): Object-oriented h-adaptive code for geophysical fluid dynamics simulation,".

[131] T. Minamoto, "Numerical Existence and Uniqueness Proof for Solutions of Nonlinear Hyperbolic Equations,". http://citeseer.ist.psu.edu/article/minamoto99numerical.html.

[132] C. F. Sopuerta and P. Laguna, "A finite element computation of the gravitational radiation emitted by a point-like object orbiting a non-rotating black hole," *Phys. Rev.* **D73** (2006) 044028, arXiv:gr-qc/0512028.

[133] O. Ghattas and X. Li, "A variational finite element method for stationary nonlinear fluid-solid interaction," *Journal of Computational Physics* **121** (1995) pp. 347–356.

[134] M. Holst, "Finite element method in numerical relativity,". http:bh0.physics.ubc.ca/BIRS05/Talks/holst.pdf. Online version given at given url.

[135] H. Banks and N. Lybeck, "A nonlinear Lax-Milgram lemma arising in the modeling of elastomers,". http://citeseer.ist.psu.edu/188099.html.

[136] K. S. Joe F. Thompson Bharat and N. P. Weatherill, *Description Contents Handbook of Grid Generation*. CRC Press, 1998.

[137] J. S. Hesthaven and T. Warburton, *Nodal Discontinuous Galerkin Methods. Algorithms, Analysis, and Applications.*, vol. Vol. 54 of *Texts in Applied Mathematics*. Spinger–Verlag, New York, xvi, ed., 2008.

[138] G. E. Karniadakis and S. J. Sherwin, *Spectral/hp Element Methods for Computational Fluid Dynamics*. Oxford University Press, London, oxford science publications, second edition ed., (1999).

[139] J. P. Boyd, "Two comments on filtering (artificial viscosity) for chebyshev and legendre spectral and spectral element methods: preserving boundary conditions and interpretation of the filter as a diffusion," *J. Comput. Phys.* **143** (1998) no. 1, 283–288.

[140] P. D. Lax, *Selected Papers Volume I: Accuracy and Resolution in the Computation of Solutions of Linear and Nonlinear Equations*, vol. Volume 1. 2005. http://www.springerlink.com/content/h712716667306511.

[141] C. Shu and P. Wong, "A note on the accuracy of spectral methods applied to nonlinear conservation laws," *Journal of Scientific Computing* **10** (1995) 357–369.

[142] D. Gottlieb and C.-W. Shu, "On the Gibbs Phenomenon and Its Resolution," *SIAM Rev.* **39** (1997) no. 4, 644–668.

[143] J. P. Boyd, "Trouble with Gegenbauer reconstruction for defeating Gibbs' phenomenon: Runge phenomenon in the diagonal limit of Gegenbauer polynomial approximations," *J. Comput. Phys.* **204** (2005) no. 1, 253–264.

[144] M. Iskandarani, D. B. Haidvogel, J. C. Levin, E. Curchitser, and C. A. Edwards, "Multiscale Geophysical Modeling Using the Spectral Element Method," *Computing in Science and Engg.* **4** (2002) no. 5, 42–48.

[145] C. Sert and A. Beskok, *Spectral element formulations on non-conforming grids: a comparative study of pointwise matching and integral projection methods*, vol. 211. Academic Press Professional, Inc. San Diego, CA, USA, 2006. http://dx.doi.org/10.1016/j.jcp.2005.05.019.

[146] Y. M. Caroline Japhet and F. Nataf, "A new Cement to Glue non-conforming Grids with Robin interface conditions: the finite element case,".

[147] M. Hannam *et al.*, "Where do moving punctures go?," *J. Phys. Conf. Ser.* **66** (2007) 012047, arXiv:gr-qc/0612097.

[148] T. W. Baumgarte and S. G. Naculich, "Analytical Representation of a Black Hole Puncture Solution," *Phys. Rev.* **D75** (2007) 067502, arXiv:gr-qc/0701037.

[149] J.-P. Ampuero, "Available SEM packages, SEM2DPACK and SEMLAB." http://www.gps.caltech.edu/~ampuero/software.html.

[150] O. Rinne, L. T. Buchman, M. A. Scheel, and H. P. Pfeiffer, "Implementation of higher-order absorbing boundary conditions for the Einstein equations," *Class. Quant. Grav.* **26** (2009) 075009, arXiv:0811.3593 [gr-qc].

[151] M. Ruiz, O. Rinne, and O. Sarbach, "Outer boundary conditions for Einstein's field equations in harmonic coordinates," *Class. Quant. Grav.* **24** (2007) 6349–6378, arXiv:0707.2797 [gr-qc].

[152] O. Rinne, L. Lindblom, and M. A. Scheel, "Testing outer boundary treatments for the Einstein equations," *Class. Quant. Grav.* **24** (2007) 4053–4078, arXiv:0704.0782 [gr-qc].

[153] O. Rinne, "Stable radiation-controlling boundary conditions for the generalized harmonic Einstein equations," *Class. Quant. Grav.* **23** (2006) 6275–6300, arXiv:gr-qc/0606053.

[154] G. E. K. John Giannakouros, "Spectral element-FCT method for scalar hyperbolic conservation laws," *International Journal for Numerical Methods in Fluids* **14** (1992) no. 6, 707–727.

[155] **Supernova Cosmology Project** Collaboration, M. Kowalski *et al.*, "Improved Cosmological Constraints from New, Old and Combined Supernova Datasets," *Astrophys. J.* **686** (2008) 749–778, arXiv:0804.4142 [astro-ph].

[156] M. Hicken *et al.*, "CfA3: 185 Type Ia Supernova Light Curves from the CfA," *Astrophys. J.* **700** (2009) 331–357, arXiv:0901.4787 [astro-ph.CO].

[157] M. A. Scheel *et al.*, "Solving Einstein's equations with dual coordinate frames," *Phys. Rev.* **D74** (2006) 104006, arXiv:gr-qc/0607056.

[158] M. D. Duez *et al.*, "Evolving black hole-neutron star binaries in general relativity using pseudospectral and finite difference methods," *Phys. Rev.* **D78** (2008) 104015, arXiv:0809.0002 [gr-qc].